Distant Speech Recognition of Natural Spontaneous Multi-party Conversations

Yulan Liu

Department of Computer Science

The University of Sheffield

This dissertation is submitted for the degree of

Doctor of Philosophy

December 2016

I would like to dedicate this thesis to my loving parents, my caring friends, and my special Tristan.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

Yulan Liu
December 2016

Acknowledgements

I would like to acknowledge the support from my supervisor Prof. Thomas Hain for his patient guide and advice, and the colleagues and other students in the Speech and Hearing (SpandH) lab for their help, throughout my PhD study. In particular, I would like to acknowledge Dr. Raymond Ng for his kind support in my early days in Sheffield, and Dr. Oscar Sas for his guide on many technical details. I would like to thank all the PhD students in SpandH lab for the happy moments spent together. Especially, I would like to thank Rosanna and Ghada for their caring words and kind support all the time.

I would like to acknowledge the Marie Curie initial training network "speech communication with adaptive learning" (SCALE), the EPSRC project "natural speech technology" (NST) and the Department of Computer Science in the University of Sheffield which funded me throughout my PhD. Without their financial help, I will not be able to have such a research experience in the very popular area of machine learning. What's more, the SCALE project and NST project created a network of professional researchers in this field which has largely speeded up my academic growth and will carry on to benefit my future career.

In addition, I would like to acknowledge Dr. Charles Fox, Prof. Pengyuan Zhang and Dr. Penny Karanasou for their kind help. I have enjoyed a lot and learned a lot from the collaboration work with them. They have also provided very helpful advice and support for my post-PhD career which I will be grateful for a life time. I would also like to acknowledge Dr. Kenichi Kumatani for his kind and patient help with the implementation of beamforming algorithms based on BTK.

Last but not least, I would like to thank my parents and my friends for the kind support and tolerance with me. Particularly, I would like to thank Dr. Tristan Whitmarsh for his patience and accompany throughout my PhD, as a senior in both life and career, and as an intimate friend with whom I would like to share a lot more happiness in my life journey onwards.

Abstract

Distant speech recognition (DSR) has gained wide interest recently. While deep networks keep improving ASR overall, the performance gap remains between using close-talking recordings and distant recordings. Therefore the work in this thesis aims at providing some insights for further improvement of DSR performance.

The investigation starts with collecting the first multi-microphone and multi-media corpus of natural spontaneous multi-party conversations in native English with the speaker location tracked, *i.e.* the Sheffield Wargame Corpus (SWC). The state-of-the-art recognition systems with the acoustic models trained standalone and adapted both show word error rates (WERs) above 40% on headset recordings and above 70% on distant recordings. A comparison between SWC and AMI corpus suggests a few unique properties in the real natural spontaneous conversations, *e.g.* the very short utterances and the emotional speech. Further experimental analysis based on simulated data and real data quantifies the impact of such influence factors on DSR performance, and illustrates the complex interaction among multiple factors which makes the treatment of each influence factor much more difficult.

The reverberation factor is studied further. It is shown that the reverberation effect on speech features could be accurately modelled with a temporal convolution in the complex spectrogram domain. Based on that a polynomial reverberation score is proposed to measure the distortion level of short utterances. Compared to existing reverberation metrics like C_{50}, it avoids a rigid early-late-reverberation partition without compromising the performance on ranking the reverberation level of recording environments and channels. Furthermore, the existing reverberation measurement is signal independent thus unable to accurately estimate the reverberation distortion level in short recordings. Inspired by the phonetic analysis on the reverberation distortion via self-masking and overlap-masking, a novel partition of reverberation distortion into the intra-phone smearing and the inter-phone smearing is proposed, so that the reverberation distortion level is first estimated on each part and then combined.

Table of contents

List of figures xiii

List of tables xvii

Nomenclature xix

1 Introduction 1

1.1 Motivation . . . 1
1.2 Research Questions and Objectives . . . 3
1.3 Contributions . . . 3
1.4 Structure of Thesis . . . 5

2 Background 7

2.1 Deep Neural Networks as Classifier . . . 8
2.2 Speech Recognition and DNN . . . 13
2.3 Robustness in Distant Speech Recognition . . . 19
2.4 Reverberation Metric and Reverberation Measurement . . . 31

3 Motivation 37

3.1 Natural Spontaneous Speech Recordings with Rich Information . . . 37
3.2 Real Natural Spontaneous Speech Recognition: from Headset Recordings to Distant Recordings . . . 41
3.3 Reverberation Modelling for Distant Speech Recognition . . . 42
3.4 Signal Aware Reverberation Measurement . . . 44

4 The Sheffield Wargame Corpora 49

4.1 Data Collection Design and Recording Setup . . . 50
4.2 Data Statistics, Annotation and Transcribing . . . 55
4.3 Blog Data and Language Model . . . 59
4.4 Baseline Systems . . . 62

4.5 Summary 69

5 Challenges in Real Natural Spontaneous Speech 73
5.1 Speech Recognition of Headset Recordings 74
5.2 DSR: Factor Analysis with Simulated Data 84
5.3 DSR: Factor Analysis with Real Data 92
5.4 Summary and Discussion 103

6 Reverberation Modelling for Distant Speech Recognition 111
6.1 Complex Spectrogram Based Reverberation Modelling 112
6.2 The Local Phase and Magnitude Assumptions 117
6.3 Experimental Evaluation 119
6.4 Summary and Discussion 131

7 Reverberation Measurement 135
7.1 Motivation 136
7.2 Polynomial Reverberation Measurement 145
7.3 Phonetic Analysis Inspired Reverberation Measurement 147
7.4 Fisher Ratio Based Discrimative Analysis 154
7.5 Experiment Results 156
7.6 Summary and Discussion 172

8 Summary, Discussion and Future Work 177
8.1 Distant Speech Recognition of Real Natural Spontaneous Conversations . 177
8.2 Reverberation Modelling and Measurement 179
8.3 Future Work 180

References 183

List of figures

2.1 *Illustration of a 3-layer DNN. Circle: perceptron or neuron or unit; dashed lines: connections among neurons; arrow: direction of the information flow due to the mathematical dependence.* 8
2.2 *Illustration of different DNN topology.* 10
2.3 *Illustration of HMM-GMM.* 13
2.4 *A variety of strategies in generating representations with DNN.* 16
2.5 *Illustration of DNN-HMM hybrid system.* 17

4.1 *SWC1 recording.* 51
4.2 *SWC - 20 shared distant microphones among three recording days (top-down view).* 52
4.3 *SWC2 recording.* 53
4.4 *Transcribing SWC recordings with multi-channel audio using XTrans.* . . 55
4.5 *SWC statistic analysis on the histogram of the speech utterance duration, the number of words in each utterance, the speech duration per speaker, the number of competing utterances, the duration of overlapped speech and the percentage of the overlapped speech in each utterance.* 57
4.6 *Speaker head location tracking based on the Ubisense system: the coordinate system and the tracked speaker head locations in the first recording session of SWC2.* 58
4.7 *SWC baseline system: acoustic model adaptation in a DNN-HMM-GMM structure.* 64

5.1 *The average utterance duration histogram in the AMI corpus.* 75
5.2 *The average utterance level WER given different number of words in one utterance. "AMI acftest-1": the 6.1 hours evaluation data from AMI corpus with dataset defined by (Liu et al., 2014); "SWC eval": the 5.6 hours evaluation data from the "SA1" dataset definition of SWC data shown in Table. 4.7.* 76

5.3 *Relationship between speaking rate and the number of words in one utterance.* 78
5.4 *Average WER per utterance at a given speaking rate per utterance.* 79
5.5 *WER per speaker (male speaker IDs start with "m" and female speaker IDs start with "f").* 81
5.6 *Overall WER per session in SWC using different microphones channels.* . . 82
5.7 *Average WER per utterance of given level of speech overlapped.* 83
5.8 *Average utterance level WER with different amount of overlapped speech on SWC dev and eval dataset.* 86
5.9 *Loudspeaker location configuration to measure RIRs for simulating the speaker movement in the room.* 89
5.10 *All physical factors in real distant speech recordings that impact speech recognition performance directly or indirectly (blue square box: the factor or attribute can be quantified; red box with round corner: the factor or attribute cannot be quantified; orange box round corner: an interaction, it is a special category as it is not a single concrete factor; box with grey background: the factor or attribute is omitted in this work; box with white background: the factor or attribute will be investigated in this work; orange dash line: the two factors or attributes are related; blue arrow: the factor is involved in an interaction).* 93
5.11 *DSR WER and speaker dependent attributes analysis based on SDM recordings.* 95
5.12 *C_{50} statistics: reverberation level variation with speaker location and microphone location.* 97
5.13 *T_{60} statistics: reverberation level variation with speaker location and microphone location.* 98
5.14 *WER variation caused by microphone difference and speaker movement.* . . 99
5.15 *Utterance level WER as speaker-microphone distance changes - TBL1 microphones.* 105
5.16 *Utterance level WER as speaker-microphone distance changes - GRID microphones.* 106
5.17 *Utterance level WER as speaker-microphone distance changes - WALL microphones.* 107
5.18 *Average WER as speaker-microphone distance changes - TBL1 microphones.* 108
5.19 *Average WER as speaker-microphone distance changes - GRID microphones.* 109
5.20 *Average WER as speaker-microphone distance changes - WALL microphones.* 110

6.1 *Validation of local linear phase assumption with 200 utterances randomly selected from "SA1" evaluation dataset of SWC headset recordings.* . . . 120
6.2 *Magnitude (dB) and unwrapped phase of the complex spectrogram from a small proportion of speech sound (sampling rate: 16 kHz).* 121
6.3 *STFT magnitude change measurement via overall* e_{mag}. 121
6.4 *STFT amplitude change of speech signal (sampling rate: 16kHz) over 10 ms time span, with frame width being 25 ms (value in dB).* 122
6.5 *The magnitude spectrogram with STFT updated at the signal sampling rate, corresponding to speech sentence: "Are you just having one warlord, what's that?" (dB)* 123
6.6 *The magnitude spectrogram from reverberant signal and from complex spectrogram reconstructed based on the same speech utterance: "Are you just having one warlord, what's that?" (dB)* 125
6.7 *Residual error in magnitude spectrogram using different time resolution when construction the complex spectrogram, based on the same speech utterance: "Are you just having one warlord, what's that?" (dB)* 126
6.8 *Magnitude spectrogram of RIRs measured with the same speaker location in SWC recording room.* 128
6.9 *Zeros of equivalent complex valued filter based on reverberation modelling in Eq. (6.10) for the TBL1 microphones and the GRID microphones at different frequencies.* 129
6.10 *Magnitude spectrogram of RIRs measured with slight variation in speaker height at the same location using the same microphone (TBL1-01).* 131
6.11 *Magnitude spectrogram of RIRs measured with the same microphone (TBL1-01) with speaker of different distances to the table in the same direction, i.e. D3 shown in Fig. 5.9. "r1": 0.15 m to the table; "r2": 0.45 m to the table; "r3": 0.75 m to the table.* 132

7.1 *Early reverberation distortion via RIR truncation experiments.* 138
7.2 *The Spearman rank correlation between the PER and the* C_{50} *at different data scales, and the Spearman rank correlation among PERs calculated at different scale, regarding the microphone difference on simulated reverberant data.* 142
7.3 *Spearman rank correlation between the WER and the reverberation score based on* C_{50} *at different scale: overall, per speaker and per utterance, based on simulated reverberant speech (SWC hybrid system evaluation dataset).* 143

7.4 *An illustration on how phonetic power spectrum pattern changes in a hyper space due to the linear effect of inter-phone smearing.* 150
7.5 *Overall polynomial reverberation score as the length of truncated RIRs increases.* 158
7.6 *Correlation between the polynomial reverberation score and the WER regarding the channel difference.* 159
7.7 *Average phoneme duration in the TIMIT corpus (/pau/ refers to silence and pause).* 161
7.8 *Based on the average phoneme duration for each phoneme in TIMIT.* . . . 162
7.9 *Examples of speech dynamic index and average power spectrum.* 165
7.10 *Polynomial reverberation score for intra-phone smearing and PER change due to channel difference and phonetic difference, on the WSJCAM0 evaluation set of simulated reverberant data.* 166
7.11 *Average PER increase in comparison with magnitude spectrum variance in terms of phoneme difference.* 167
7.12 *Average PER increase in comparison with average power spectrum vector rotation caused by reverberation (3 preceding phones taken into consideration).* 168
7.13 *Average PER increase in comparison with average power spectrum vector rotation caused by reverberation (3 preceding phonemes taken into consideration).* 168
7.14 *Correlation change when tuning the combination weight in intra-phone smearing index (black line: Spearman rank correlation; green line: p-value).* 169
7.15 *Experimental examination of the correlation between reverberation score and the overall PER on WSJCAM0 based simulated data.* 170
7.16 *The change of Fisher ratio due to reverberation.* 171

List of tables

4.1 *SWC statistics of transcribed recordings.* 55
4.2 *SWC statistics.* 56
4.3 *Special words better covered by blog data - an investigation on the word occurrence percentage (%) in different text data components.* 60
4.4 *The SWC manual transcription and the text data components ($\log_2$ used).* 61
4.5 *LM components and text data statistics.* 61
4.6 *Perplexity of interpolated LM and conversational web data only based LM on SWC manual transcripts.* 62
4.7 *Datasets for SWC.* 63
4.8 *SWC adaptation baseline performance with "AD2" dataset definition using source acoustic models trained on the AMI corpus.* 65
4.9 *SWC standalone training system performance with "SA1" dataset definition.* 67
4.10 *WER for different beamforming and multi-microphone based dereverberation algorithms.* 69

5.1 *Emotional speech analysis: number of laughs, average number of word per utterance and average WER per utterance.* 80
5.2 *WERs with and without overlapped speech (%). "IHM": the original individual headset microphone recordings; "IHM.OL": simulated data with overlapped speech.* 85
5.3 *Analysis of the impact of overlapped speech and reverberation speech on WER with simulated data based on RIR from microphone "TBL1-01" at the center of table.* 87
5.4 *WER improvement from multi-microphone based dereverberation and beamforming on simulated reverberant speech based on RIRs from "TBL1" array.* 88
5.5 *WER comparison between static reverberation and changing reverberation due to speaker movement.* 90

5.6 *Dereverberation performance comparison using RIRs from different microphone arrays for simulated data.* 91
5.7 *WER on SDM with emotional speech: laugh.* 96
5.8 *WER based on distributed microphone selection using different strategies.* 101

6.1 *WER and pattern approximation accuracy with different N_f value.* 127

7.1 *Speech recognition performance comparison of simulated reverberant speech with full or truncated RIRs.* 138
7.2 *Phoneme category definition in the TIMIT corpus ("DPP": average duration per phoneme; "Per.": percentage in overall duration without silence).* 160
7.3 *Duration and duration percentage of different sounds in conversational English speech. "Dur.": overall duration (hours); "DPP": the average duration per phoneme (ms); "Pct.": percentage in overall duration without silence (%).* 161

Nomenclature

Roman Symbols

a The pre-emphasis weight in speech recognition front-end.

$b_{i,j}$ The bias in calculating the output of the j-th neuron in the i-th layer.

C_{50} Speech clarity

$D_\alpha(\triangle\tau, k)$ The average speech dynamic index for phoneme indexed with α.

$\bar{d}_\alpha$ The average duration of the phoneme indexed with α.

$D(\triangle\tau, k)$ Speech dynamic index, i.e. the average speech magnitude difference given a shift $\triangle\tau$ in the k-th frequency bin.

$E_\alpha(k)$ The overall energy of the phoneme indexed with α in the k-th frequency bin.

$\bar{e}_\alpha(k)$ The statistic average power spectrum vector based on all phonemes except for phoneme α that corrupt phoneme α via inter-phone smearing.

$\hat{\mathbf{e}}_{\bar{\alpha},\beta}$ The source power spectrum vector weighted by RIR indexed by β for the inter-phone smearing on the phoneme indexed by α.

$\hat{e}_{\bar{\alpha},\beta}(k)$ The k-th frequency bin of the source power spectrum vector weighted by RIR indexed by β for the inter-phone smearing on the phoneme indexed by α.

$e_{\bar{\alpha}}(k)$ The k-th frequency bin of the statistic average power spectrum vector based on all phonemes except for phoneme α that corrupt phoneme α via inter-phone smearing.

$\mathbf{e}_{\alpha,\beta}$ The equivalent power spectrum vector that causes inter-phone smearing on phoneme α by RIR indexed with β.

$e_{\alpha,l}(k)$ The average STFT energy in the k-th frequency bin for the l-th example recording of the phoneme indexed by α.

$\bar{e}_\alpha(k)$ The ensemble average STFT energy in the k-th frequency bin for the phoneme indexed with α.

$\mathbf{e}_\alpha$ Vector of the ensemble average STFT energy across frequency for the phoneme indexed with α.

$\bar{e}(k)$ The average power spectrum over all phonemes on the k-th frequency bin.

$E_\beta^{(\mathsf{d})}$ Direct sound energy in the RIR indexed with β.

$\bar{e}_{\beta,\alpha}^{(\mathsf{II})}(N_\mathrm{f},k)$ The k-th frequency of the average spectrum of RIR power spectrogram corresponding to inter-phone smearing considering the duration of phoneme α.

$e_\mathrm{mag}(m,k,N_\mathrm{f})$ Error introduced by local stationary magnitude assumption represented by the ratio between average local STFT magnitude variance and average local STFT energy.

e_j The error in the j-th class between the classification hypothesis based on DNN output posteriors and the reference labelling.

$E_X(k)$ Overall energy of speech signal spectrogram in the k-th frequency bin based on the the whole piece of clean speech recording.

F A symbolic representation of the cost function for DNN optimization.

$f(\cdot)$ DNN activation function.

$f_\mathrm{r}(\cdot)$ Rectified linear unit function.

$f_\mathrm{s}(\cdot)$ Sigmoid function.

$f_\mathrm{t}(\cdot)$ Hyperbolic tangent function.

G Number of samples in one piece of speech recording.

$g(\cdot)$ Softmax function.

$G_{\alpha,l}$ Number of samples in the l-th clean recording example of the phoneme indexed with α.

$\mathbf{h}$ Room impulse response (in vector form).

$H_\beta(\tau,N_\mathrm{f},k)$ The complex spectrogram of RIR indexed with β.

$H(m, N_{\mathrm{f}}, k)$ The raw STFT of room impulse response. Note that different from the STFT for signal, this STFT adopts a square window.

$h(n)$ Room impulse response (in FIR filter form).

$I_{\alpha,\beta}(N_{\mathrm{f}})$ Polynomial score based intra-phoneme smearing index for the phoneme indexed with α given the reverberation condition indexed with β.

$I^{(\mathrm{a})}_{\alpha,\beta}(N_{\mathrm{f}})$ The overall smearing index on phoneme α from RIR indexed with β when inter-phone smearing index is based on angular distance and the speech magnitude spectrogram is based on a time resolution of N_{f} samples.

$I^{(\mathrm{c})}_{\alpha,\beta}(N_{\mathrm{f}})$ The overall smearing index on phoneme α from RIR indexed with β when inter-phone smearing index is based on cosine distance and the speech magnitude spectrogram is based on a time resolution of N_{f} samples.

$I^{(\mathrm{l})}_{\alpha,\beta}(N_{\mathrm{f}})$ Intra-phone smearing index for the phoneme indexed with α given the reverberation condition indexed with β, calculated with a time resolution in spectrogram of N_{f} samples.

$I^{(\mathrm{ll,\,a})}_{\alpha,\beta}$ Inter-phone smearing index for the phoneme indexed with α given the reverberation condition indexed with β based on angle rotation.

$I^{(\mathrm{ll,\,c})}_{\alpha,\beta}$ Inter-phone smearing index for the phoneme indexed with α given the reverberation condition indexed with β based on cosine distance.

$I_{\beta}(N_{\mathrm{f}})$ Polynomial reverberation score of the RIR indixed with β using a time resolution of N_{f} samples when analysing the temporal change of clean speech magnitude spectrogram.

$I^{(\mathrm{a})}_{\beta}(N_{\mathrm{f}})$ Overall reverberation smearing index for the RIR indexed with β, based on angular distance in inter-phone smearing and a time resolution of N_{f} when calculating speech signal magnitude spectrogram properties.

$I^{(\mathrm{c})}_{\beta}(N_{\mathrm{f}})$ Overall reverberation smearing index for the RIR indexed with β, based on cosine distance in inter-phone smearing and a time resolution of N_{f} when calculating speech signal magnitude spectrogram properties.

$I_{\beta}(N_{\mathrm{f}}, \triangle\tau, k)$ An estimation of the reverberation distortion level caused by the average magnitude spectrogram change in clean speech over a time shift of $\triangle\tau$ in the frequency bin k, given a time resolution of N_{f} when calculating the spectrogram.

$I_{\beta}^{(\mathrm{I})}(N_{\mathrm{f}})$ Intra-phone smearing caused by reverberation condition indexed with β on all phonemes, calculated with a time resolution in spectrogram of N_{f} samples.

$I_{\alpha,\beta}^{(\mathrm{II,\,a})}$ Inter-phone smearing caused by reverberation condition indexed with β on all phonemes based on angle rotation.

$I_{\alpha,\beta}^{(\mathrm{II,\,c})}$ Inter-phone smearing caused by reverberation condition indexed with β on all phonemes based on cosine distance.

j Imaginary unit.

$J_{\alpha,\beta,\mathrm{F}}(k)$ The Fisher ratio between hone α and phoneme β based on features corresponding to the k-th frequency bin.

$J_{\alpha,\mathrm{F}}(k)$ The discriminative score calculated by averaging the discriminative scores from Fisher Ratio based discriminative analysis conductedbetween phoneme α and all other phonemes in pair, using the features corresponding to the k-th frequency bin.

J_{F} Overall discriminative score based on average Fisher ratio across frequencies.

$J_{\mathrm{F}}(k)$ The Fisher ratio over all classes given the features from the k-th frequency bin.

k Frequency bin index.

L_{α} The number of examples for the phoneme indexed with α; The number of DNN layers.

M Number of samples in room impulse response.

M_{β} Number of samples in the RIR indexed as β.

N Number of samples for discrete Fourier transform.

n Discrete time index.

$n_{\alpha}^{(\mathrm{I})}$ The number of summation components in reverberation modelling that corresponds to intra-phone smearing considering the duration of the phoneme indexed with α.

N_{f} Number of samples where the local linear phase assumption and local stationary magnitude assumption hold with sufficiently low error.

N_i The number of neurons in the i-th layer of DNN.

$P(\alpha)$ The percentage of the phoneme indexed with α over all phonemes by duration.

Q Number of phonemes in total.

r_α Average magnitude spectrogram variance and spectrogram energy ratio for the phoneme indexed with α.

$r_\alpha^{(\mathrm{I})}$ Ratio of intra-phone smearing on the phoneme indexed with α.

$r_\alpha^{(\mathrm{I})}$ The ratio of intra-phone smearing for the whole phoneme α.

$p_\alpha^{(\mathrm{I})}(\tau)$ The chance of intra-phone smearing in the τ-th STFT of the spectrogram of phoneme α.

$r_\alpha^{(\mathrm{II})}$ Ratio of inter-phone smearing on the phoneme indexed with α.

$r_\alpha^{(\mathrm{II})}$ The ratio of inter-phone smearing for the whole phoneme α.

$r_\alpha^{(\mathrm{I})}(\tau)$ Ratio of intra-phone smearing on the τ-th STFT of phoneme α.

$r_\alpha^{(\mathrm{I})}(\tau)$ Ratio of inter-phone smearing on the τ-th STFT of phoneme α.

$r^{(\mathrm{I})}$ Ratio of intra-phone smearing.

$r^{(\mathrm{II})}$ Ratio of inter-phone smearing.

T Number of samples in the speech recording.

T_{60} Reverberation Time

t_j The binary reference for the output of the j-th neuron in DNN output layer.

$v_\alpha(k)$ The within-class variance based on the power spectrum of the phoneme indexed with α in the k-th frequency bin.

$v_{\mathrm{mag}}(m,k,N_{\mathrm{f}})$ Variance of STFT magnitude on the k-th frequency bin over a short period of time, with STFT updated every N_{f} samples.

$v_{\alpha,\beta,\mathrm{b}}(k)$ The between-class variance in the Fisher discriminative analysis on phoneme α and phoneme β based on features corresponding to the k-th frequency bin.

$v_{\mathrm{b}}(k)$ The between-class variance based on features from the k-th frequency bin.

$v_{\alpha,\beta,\mathrm{w}}(k)$ The within-class variance in the Fisher discriminatve analysis on phoneme α and phoneme β based on features corresponding to the k-th frequency bin.

$v_{\mathrm{w}}(k)$ The within-class variance based on features from the k-th frequency bin.

$w_{i,k,j}$ The weight applied on the k-th input when calculating the output of the j-th neuron from the i-th layer.

$w(n)$ Window function used in short time Fourier transform.

$x_{i,j}$ The output to the j-th neuron in the i-th layer of deep neural network, which is also the j-th input element to the $(i+1)$-th layer. Particularly, $x_{0,j}$ is the j-th input of the first layer.

$X(\tau,k)$ The short time Fourier transform of clean headset recordings at time τ and the k-th frequency bin.

$x(n)$ Clean headset recording of speech signal.

$Y(\tau,k)$ The short time Fourier transform of distant microphone recordings at time τ and the k-th frequency bin.

y_i The i-th dimension of DNN output vector.

$y(n)$ Distant microphone recording of speech signal.

$z_{i,j}$ The hidden linear output of the j-th neuron in the i-th layer of DNN.

Greek Symbols

α The momentum in DNN optmization.

δ Sample shift or the difference between two discrete time indices.

η The learning rate in DNN optimization.

λ The weighting parameter to tune when combining two elements additively together.

$\mu_{\alpha,l}(k)$ Average STFT magnitude on the k-th frequency bin for the l-th example recording of the phoneme indexed with α.

$\mu_{\text{mag}}(m,k,N_{\text{f}})$ Average STFT magnitude on the k-th frequency bin over a short period of time, with STFT updated every N_{f} samples.

$\rho_{\text{p}}(k,N_{\text{f}})$ Local Pearson correlation between time index and STFT phase unwrapped over time.

$\bar{\rho}_{\text{p}}(k,N_{\text{f}})$ Average Pearson correlation between time index and STFT phase unwrapped over time.

$\rho'_{\mathrm{p}}(m,k,N_{\mathrm{f}})$ Local Pearson correlation between time index and STFT phase unwrapped over time at a strict standard.

$\sigma^2_{\alpha,l}(k)$ Variance of the STFT magnitude on the k-th frequency bin for the l-th example recording of the phoneme indexed with α.

τ Discrete time index.

$\theta(\tau,k)$ Unwrapped phase of clean speech STFT $X(\tau,k)$.

Superscripts

(I) Intra-phone smearing.

(II) Inter-phone smearing.

* Conjungate transpose.

Subscripts

α, γ The index of phoneme class.

β The index of RIR, or the index of reverberation condition, or the index of reverberation channel.

Other Symbols

$\angle$ The phase of a complex value.

$\lceil\ \rceil$ The smallest integer larger than the quoted value.

$*$ Convolution operation

$\lfloor\ \rfloor$ The largest integer smaller than the quoted value.

$\triangle$ The difference in the value of the variable following it.

Acronyms / Abbreviations

3D Three Dimensional

AIR Acoustic Impulse Response

ANN Artificial Neural Network

ASR Automatic Speech Recognition

BLSTM Bidirectional Long Short Term Memory

BPTT Back Propagation Through Time

CART Classification and Regression Tree

CI Cochlear Implant

CNN Convolutional Neural Network

DCT Discrete Cosine Transform

DFT Discrete Fourier Transform

DNN Deep Neural Network

DRR Direct-to-Reverberation Ratio

DSB Delay and Sum Beamforming

DSR Distant Speech Recognition

ELR Early-to-Late Reverberation Ratio

EM Expectation Maximization

FIR Finite Impulse Response

GMM Gaussian Mixture Model

GPU Graphic Processing Unit

GSC Generalized Sidelobe Canceler

GWPE Generalized Weighted Prediction Error

HF Hadamard-Fischer

HLDA Heteroscedastic Linear Discriminant Analysis

HMM Hidden Markov Model

IHM Individual Headset Microphone

JUD Joint Uncertainty Decoding

LCMV Linearly Constrained Minimum Variance

LDA Linear Discriminant Analysis

LM Language Model

LSTM Long Short Term Memory

LTI Linear Time Invariant

LVCSR Large Vocabulary Continuous Speech Recognition

MAP Maximum-a-posterior

MBR Minimum Bayes Risk

MDM Multiple Distant Microphone

MFCC Mel Frequency Cepstral Coefficient

MIMO Multiple-Input Multiple-Output

MINT Multiple-Input/output Inverse Theorem

MLLT Maximum Likelihood Linear Transform

MLP Multi-layer Perceptrons

MMI Maximum Mutual Information

MMSE Minimum Mean Square Error

MPE Minimum Phone Error

MVDR Minimum Variance Distortionless Response

NTT Nippon Telegraph and Telephone

PER Phoneme Error Rate

PESQ Perceptual Evaluation of Speech Quality

PLP Perceptual Linear Prediction

PPL Perplexity

RBM Restricted Boltzmann Machine

ReLU Rectified Linear Unit

PESQ Perceptual Evaluation of Speech Quality

RIR Room Impulse Response

RNN Recurrent Neural Network

ROVER Recognizer Output Voting Error Reduction

SAT Speaker Adaptive Training

SDBF Super Directive Beamforming

SDM Single Distant Microphone

sMBR state-level Minimum Bayesian Risk

SNR Signal to Noise Ratio

SNR Signal-to-noise Ratio

SPLICE stereo based piecewise linear compensation for environment

STFT Short Time Fourier Transform

SWC Sheffield Wargame Corpora

TDOA Time Difference of Arrival

TF Time Frequency

VDCNN Very Deep Convolutional Neural Network

VTLN Vocal Tract Length Normalisation

VTS Vector Taylor Series

wDSB Weighted Delay and Sum Beamforming

WER Word Error Rate

WPE Weighted Prediction Error

Chapter 1

Introduction

Contents

1.1 Motivation 1
1.2 Research Questions and Objectives 3
1.3 Contributions 3
1.4 Structure of Thesis 5

1.1 Motivation

In recent years, automatic speech recognition (ASR) systems have embraced groundbreaking performance improvement brought by the deep neural network (DNN) and other types of deep networks applied for the front-end, acoustic modelling and language modelling. The improvement DNN brings is widely observed in various tasks and applications, reducing WER by 20% relatively on average (Liu et al., 2014). The recognition performance on close-talking recordings has reached a record high level (Sercu and Goel, 2016). However when comparing the recognition on close-talking recordings and on distant recordings, a large performance gap still remains.

Compared to the speech recognition based on the close-talking recordings, DSR is challenged by the diverse reverberant environments and background noises. The performance improvement in DSR so far mainly comes from the research progress in the front-end, particularly in the speech enhancement of distant recordings. There are enhancement techniques making use of the distant recordings from multiple microphones, *e.g.* beamforiming, multichannel dereverberation. There are also enhancement techniques employing the multi-media recordings which provide rich information besides the speech audio, *e.g.* speech enhancement based on the audio-visual speaker tracking (Wölfel and

McDonough, 2009). These enhancement techniques rely on the quantity and the quality of distant speech recordings from multiple microphones or multiple media. Unfortunately such research corpora is of much smaller amount compared to those based on headset recordings. When it comes to real recordings of natural spontaneous multi-party speech, even less data are available. On the other hand, DNN demands more training data than classical models. In addition, compared to GMM-HMMs based acoustic models, DNN based acoustic models can have larger relative performance degradation from the acoustic mismatch between training data and test data especially when the training data is clean while the test data is very reverberant or noisy. The data mismatch widely exists in DSR tasks and it is frequently caused by the different acoustic environments and background noises. The lack of research data, particularly the lack of distant recordings of real natural spontaneous multi-party conversations, has been one important factor limiting the progress in DSR.

Due to the lack of research data, there is limited understanding of the challenges in the state-of-the-art DSR systems when they are applied in the real natural spontaneous conversations. In distant speech recordings there are influence factors which might behave very differently in real recordings compared to in simulated recordings, and in real natural conversations compared to in conversations with controlled topic or content, *i.e.* the controlled recordings. As a result the assumptions of some research algorithms developed based on the simulated recordings and the controlled recordings may not hold for the real recordings of natural conversations. Without real recordings of natural spontaneous conversations, it is very difficult to justify and quantity the influence of such gap in data, or to prioritise the research focus on the real challenges for the state-of-the-art DSR systems, when they are applied in the wild. Therefore, it is of high demand to collect more distant recordings of real natural spontaneous conversations.

Among the influence factors in the real recordings for DSR, reverberation is particularly important because it widely exists in distant recordings and it varies by the acoustic change in the environment. Besides, the convolutional nature of the reverberation distortion makes it one of the main contributors to the speech recognition performance gap between using the distant recordings and using the close-talking recordings. In the state-of-the-art DSR systems based on DNNs, there is very limited progress regarding novel algorithms or structures that improves the robustness of DNN against reverberation. So far, the multi-condition training has been found effective to improve the overall robustness of DNN by using training data with diverse distortions (Kinoshita et al., 2016). In addition, research has been conducted on the model combination and the model selection for a balance between the overall performance robustness among diverse testing environment conditions and the best performance in each environment condition.

The multi-condition training, model combination and model selection have been three main practical strategies to improve the robustness of the state-of-the-art DSR systems against diverse reverberat environments. For these methods, reverberation measurement, *i.e.* the estimation of the reverberation level of distant recordings, is critical for both the training data selection and the model selection. Existing reverberation measurement is based on acoustic reverberation metrics that estimate the reverberation level of the recording environment. However there are a few issues with such reverberation metrics. For example for the early-to-late reverberation ratio, the optimal boundary between early and late reverberation is not fully addressed analytically and the configuration in existing systems are completely based on experimental experience. In addition, the existing reverberation measurement is signal independent, thus it does not take the reverberation sensitivity of different speech sounds into account, while in fact the same reverberation could cause different levels of distortion on different speech feature patterns, as pointed by (Kokkinakis and Loizou, 2011). Therefore, further research is required on the reverberation measurement strategies that are optimal to the speech recognition tasks.

1.2 Research Questions and Objectives

The work covered by this thesis tries to provide some insights from two aspects that are of critical importance in improving the state-of-the-art DSR systems.

The first aspect is about the performance of the state-of-the-art DSR systems on real distant recordings of natural spontaneous multi-party conversations. The objective of this research is to understand the challenging influence factors in the real natural spontaneous multi-party conversations for the state-of-the-art DSR systems. In particular, the impact each factor has on WER and the interaction among multiple influence factors. To achieve this objective the recording of the Sheffield Wargame Corpora is collected as the recordings of real natural spontaneous multi-party conversations.

The second aspect is about the reverberation modelling and the measurement of the reverberation distortion level in the context of DSR. The objective of this research is to understand how reverberation distorts the speech feature patterns, and how to measure the reverberation distortion level in speech features for DSR when the speech sound properties are taken into consideration.

1.3 Contributions

There are three major contributions from the work covered in this thesis.

The first contribution is the release of the Sheffield Wargame Corpora (SWC), a database of multi-microphone and multi-media real recordings of natural spontaneous spontaneous multi-party conversations (Chapter §4). This work has led to two conference publications in Interspeech 2013 and Interspeech 2016 respectively (Fox et al., 2013; Liu et al., 2016).

The second contribution is that the experimental analysis of the DSR on the natural spontaneous multi-party conversational speech reveals a few unique properties in the real natural spontaneous conversations, as well as the challenging influence factors for the state-of-the-art DSR. The impact of the influence factors on DSR in such real recordings on DSR performance is quantified by WERs (Chapter §5).

The third contribution is the study on reverberation in DSR (Chapter §6, Chapter §7). An accurate reverberation modelling based on complex spectrogram is studied. Based on that, a novel polynomial format reverberation score is proposed and it is shown to provide a high rank correlation with WER regarding the channel difference. In addition, for the first time the difference between the reverberation level on the recording environment and channel and the reverberation distortion level in the speech feature pattern is emphasised. The research effort is devoted to improving the estimation of the reverberation distortion level in short speech audio recordings, from a novel angle of estimating the reverberation distortion by phonetic smearing.

It is worth mentioning that the author has also worked on other aspects of speech recognition during the PhD which are not detailed in this thesis due to limited space. The research focus of these works is slightly mismatched with the theme of this thesis, however they have benefited a lot the work to be presented in this thesis, and they have led to a few publications. Therefore these work will be briefly mentioned when related research point is presented. A full list of publications from the author's PhD research is listed below:

1. **Y. Liu**, C. Fox, M. Hasan, T. Hain, "The Sheffield Wargame Corpus - Day Two and Day Three". In *Interspeech*, San Francisco, USA, 2016.

2. T. Hain, J. Christian, O. Saz, S. Deena, M. Hasan, W. M. Ng, R. Milner, M. Doulaty, **Y. Liu**, "WebASR 2 - Improved Cloud Based Speech Technology". In *Interspeech*, San Francisco, USA, 2016.

3. O. Saz, M. Doulaty, S. Deena, R. Milner, W. M. Ng, M. Hasan, **Y. Liu**, T. Hain, "The 2015 Sheffield System for Transcription of Multi-Genre Broadcast Media". In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Arizona, USA, 2015.

4. **Y. Liu**, P. Karanasou and T. Hain, "An Investigation into Speaker Informed DNN Front-end for LVCSR". In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, 2015.

5. P. Zhang, **Y. Liu** and T. Hain, "Semi-Supervised DNN Training in Meeting Recognition". In *Spoken Language Technology Workshop (SLT)*, Nevado, USA, 2014.

6. **Y. Liu**, P. Zhang and T. Hain, "Using Neural Network Front-ends on Far Field Multiple Microphones Based Speech Recognition". In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014.

7. C. Fox, **Y. Liu**, E. Zwyssig and T. Hain, "The Sheffield Wargame Corpus". In *Interspeech*, Lyon, France, 2013.

1.4 Structure of Thesis

The thesis is structured in the following way. Chapter §2 reviews the existing work that are either related to or employed in the research work to present in this thesis. Based on the review, Chapter §3 highlights the unaddressed issues in the existing research and how these issues motivated the work in this thesis. Chapter §4 to Chapter §7 cover the original work by the author. Chapter §8 summarizes the major findings from the work presented in this thesis and proposes potential future extensions of current work.

Among the chapters that present the original work by the author, Chapter §4 details the collection of the Sheffield Wargame Corpora, a real multi-microphone and multi-media recording database of natural spontaneous multi-party conversations. This database is used in Chapter §5 for a case study to understand the challenges in DSR and to quantify the impact of the influence factors in the distant recordings in a comparison with the headset recordings. Chapter §6 and Chapter §7 focus on the reverberation. In particular, Chapter §6 performs analytic and experimental study on how reverberation effect could be accurately modelled by a convolution operation in the complex spectrogram domain. Based on the analytic analysis, Chapter §7 explores novel methods to measure the distortion level in the speech features caused by reverberation. It adopts a different mathematical strategy from existing reverberation measurement, in an effort to improving the estimation accuracy of the reverberation corruption level in speech feature pattern in DSR tasks.

Chapter 2

Background

Contents

2.1 Deep Neural Networks as Classifier 8

2.2 Speech Recognition and DNN 13

2.2.1 Acoustic Model in Speech Recognition 13

2.2.2 DNN in ASR 15

2.3 Robustness in Distant Speech Recognition 19

2.3.1 Room impulse response measurement 20

2.3.2 Noise Suppression and Robustness 22

2.3.3 Dereverberation 23

2.3.4 Beamforming 26

2.3.5 Environment Robustness with DNN 29

2.4 Reverberation Metric and Reverberation Measurement 31

This chapter reviews the existing work that are either related to or employed in the research work to present in later chapters. The review covers four aspects with four sections. Section 2.1 covers the fundamentals of deep neural network (DNN) and its implementation. Section 2.2 explains how DNN is used in a speech recognition system either as the front-end or as the acoustic model. Section 2.3 discusses the dereverberation and noise robustness algorithms for distant speech recognition (DSR). Section 2.4 focuses on the reverberation measurement and how it could benefit DSR.

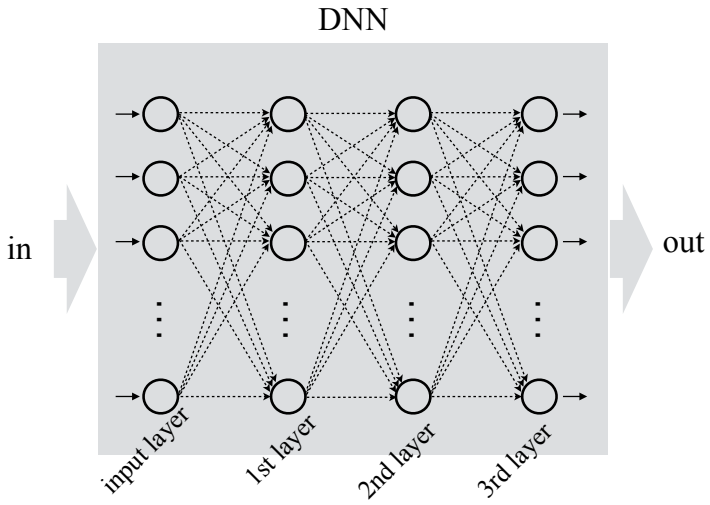


Fig. 2.1 *Illustration of a 3-layer DNN. Circle: perceptron or neuron or unit; dashed lines: connections among neurons; arrow: direction of the information flow due to the mathematical dependence.*

2.1 Deep Neural Networks as Classifier

The term "Deep neural network" (DNN) usually refers to the deep feedforward neural network. It is comprised of multiple-layer perceptrons (MLP) with connections among perceptrons in two consequent layers, as shown in Fig. 2.1. The information flows from the input layer to the last layer in one direction, thus being "feedforward". There is no connection among perceptrons in the same layer. The perceptron is also referred to as "neuron" or "unit" in the existing literature for historic reasons. To be consistent with the majority of the literature, the perceptron is referred to as "neuron" from now on in this thesis.

Denote the input to the j-th neuron in the input layer or the first layer of DNN as $x_{0,j}$. Denote the output of the j-th neuron in the i-th layer as $x_{i,j}$. When the i-th layer is not the last layer or the output layer of the DNN, $x_{i,j}$ is also the j-th input element to the $(i+1)$-th layer. The time index is dropped here to simplify the notation without causing confusions, while it is worth emphasizing that in many tasks the input and the output of each neuron in DNN are naturally time dependent. The connections among neurons in two consequent layers can be formulated as

$$x_{i,j} = f(z_{i,j}) \qquad (j = 0,1,...,N_i-1; i = 1,2,...,L) \tag{2.1}$$

$$z_{i,j} = \sum_{k=0}^{N_{i-1}-1} \sum_{j=0}^{N_i-1} x_{i-1,k} w_{i,k,j} + b_{i,j} \qquad (k = 0,1,...,N_{i-1}-1; j = 0,1,...,N_i-1; i = 1,2,...,L) \tag{2.2}$$

where N_i is the number of neurons in the i-th layer and L is the number of layers in the DNN. The neuron weight $w_{i,k,j}$ and the layer bias b_i are the DNN parameters to optimise. The function $f(\cdot)$ is a non-linear function to avoid magnitude explosion in a deep structure. An activation function processes each neuron independently, and a few frequent options for an activation function include the sigmoid function shown in Eq. (2.3), the hyperbolic tangent function shown in Eq. (2.4) and the rectified linear unit (ReLU) function shown in Eq. (2.5).

$$f_{\mathrm{s}}(z) = \frac{1}{1+e^{-z}} \tag{2.3}$$

$$f_{\mathrm{t}}(z) = \tanh(z) \tag{2.4}$$

$$f_{\mathrm{r}}(z) = \begin{cases} z & (z \geq 0) \\ 0 & (z < 0) \end{cases} \tag{2.5}$$

For the neurons in the output layer of DNN, the activation function should normalize the output magnitude of multiple perceptrons in the same layer, thus the softmax function is a good option (Eq. (2.6)).

$$g(z_{i,j}) = \frac{e^{z_{i,j}}}{\sum_{j=0}^{N_i-1} e^{z_{i,j}}} \tag{2.6}$$

The implementation of DNN is completed in two steps. First, the DNN configuration is determined, namely the number of layers, the number of neurons in each layer and the activation function for each layer. Second, the DNN parameters are optimised, namely the neuron weights and the layer biases are optimised with training data. In some literature the layer biases are omitted because they can be equivalently implemented with the weights on an extra neuron in each layer whose input is a constant value "1".

The DNN configuration is usually adjusted empirically. The configuration of output layer is closely associated with the application task. In classification tasks, the softmax function is usually employed as the activation function in the DNN ouput layer. In comparison, in regression tasks, to achieve a wide target value range, the activation function can be skipped in the output layer. The number of layers and the number of neurons in hidden layers are empirically adjusted, and they are usually proportional to the amount of available training data and computation resource. The topology of DNN, *i.e.* how to distribute the neurons in multiple hidden layers, is determined purely empirically. The hidden layers here refer to all layers except for the input layer and the output layer. For the input layer, the number of neurons are determined by the dimension of input features. For the output layer, the number of neurons is associated with the number of target classes in a classification task or the dimension of target in a regression task. In

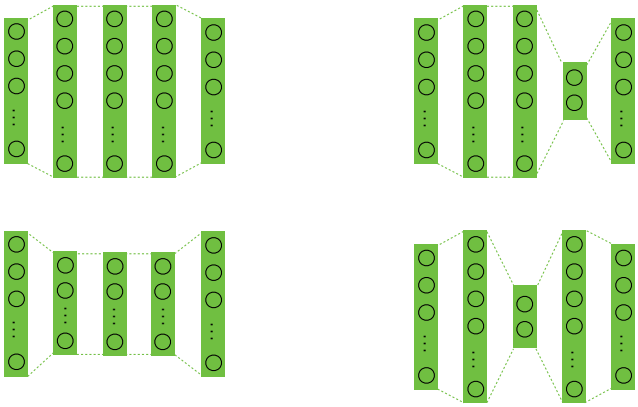


Fig. 2.2 *Illustration of different DNN topology.*

some structure the amount of neurons can be the same for all hidden layers, as illustrated in the left column of Fig. 2.2. Alternatively, one or a few layers can have significantly smaller amount of neurons compared to the other layers, thus shaping a bottleneck layer as illustrated in the right column of Fig. 2.2. The bottleneck layer structure forces DNN to reduce the effective dimension of the input feature and the intermediate presentations. A bottleneck layer is frequently adopted when the DNN is employed to produce a map from the high-dimensional input feature to a compressed representation which is also referred to as the "bottleneck feature" (Grezl et al., 2007).

The optimisation of DNN parameters could be performed with the gradient decent algorithm. Denote the final output of DNN as y_j, thus for DNN of L layers,

$$y_j = x_{L,j} \qquad (j = 0, 1, ..., N_L - 1). \tag{2.7}$$

The gradient descent based optimisation iteratively adjusts the DNN parameters in the direction of reducing the cost function. Denote the cost function as F, thus

$$w_{i,k,j} = \alpha \cdot \hat{w}_{i,k,j} - \eta \cdot \frac{\partial F}{\partial w_{i,k,j}}\Big|_{w_{i,k,j}=\hat{w}_{i,k,j}} \tag{2.8}$$

$$b_{i,j} = \alpha \cdot \hat{b}_{i,j} - \eta \cdot \frac{\partial F}{\partial b_{i,j}}\Big|_{b_{i,j}=\hat{b}_{i,j}} \tag{2.9}$$

where " $\hat{}$ " indicates one parameter is of the value from previous optimisation iteration. α is the "momentum" with a value between 0 and 1 to adjust the forgetting speed of the parameter value during optimisation, and η is the "learning rate" to adjust the updating speed of the parameter value. The partial derivative could be calculated in a backward

propagation manner from the $(i+1)$-th layer to the i-th layer recursively

$$\frac{\partial F}{\partial x_{i,j}} = \sum_{l=0}^{N_{i+1}} \frac{\partial F}{\partial x_{i+1,l}} \cdot \frac{\partial x_{i+1,l}}{\partial x_{i,j}} \tag{2.10}$$

so that

$$\frac{\partial F}{\partial w_{i,k,j}} = \frac{\partial F}{\partial x_{i,j}} \cdot \frac{\partial x_{i,j}}{\partial w_{i,k,j}} \tag{2.11}$$

$$\frac{\partial F}{\partial b_{i,j}} = \frac{\partial F}{\partial x_{i,j}} \cdot \frac{\partial x_{i,j}}{\partial b_{i,j}} \tag{2.12}$$

The partial derivatives $\frac{\partial x_{i,j}}{\partial w_{i,k,j}}$ and $\frac{\partial x_{i,j}}{\partial b_{i,j}}$ are dependent on the activation function used, *e.g.* sigmoid function (Eq. (2.3)), hyperbolic tangent function (Eq. (2.4)), rectified linear unit function (Eq. (2.5)) or softmax function (Eq. (2.6)). The detailed derivation for each activation function is skipped here.

The selection of cost function is dependent on the task. For a regression task, the DNN output is directly used to approximate the reference vector. In this case the cost function can be an Euclidean distance function that measures the error, between the DNN output and the ground-truth reference. For a classification task, the DNN serves directly as a multi-class classifier. Each neuron in the output layer corresponds to one candidate class. The output of that neuron approximates the posterior of corresponding class given the input features. In this case the cost function can be a cross entropy function of the ground-truth labels and the classification results based on the maximal posteriors. Denote the reference for the j-th neuron in the output layer as t_j, whose value is either 0 or 1 in a classification task. Then the cost function is

$$F = -\sum_{j=0}^{N_L-1} t_j \log y_j \tag{2.13}$$

Minimizing such a cross-entropy based cost function is equivalent to maximizing the probability of getting all input samples correctly labelled, *i.e.* maximizing the overall posterior of correct labelling. The details about such an equivalence have been discussed by Bishop (1995).

When using the softmax function as the activation function for DNN output layer along with the cross-entropy function as the cost function, with some derivation it could be found that for the output layer (L-th layer)

$$\frac{\partial F}{\partial w_{L,k,j}} = e_j \cdot x_{L-1,k} \tag{2.14}$$

$$\frac{\partial F}{\partial b_{L,j}} = e_j \tag{2.15}$$

where e_j is the error in the j-th class, *i.e.* the difference between the classification hypothesis based on the DNN output posteriors and the reference labelling,

$$e_j = y_j - t_j \tag{2.16}$$

More details about the derivation details can be found in the work by Bishop (1995) and are thus skipped here.

In practice, to achieve a balance between the computation cost and the robustness of optimised DNN, the implementation is usually based on the batch mode stochastic gradient decent, *i.e.* the parameters are updated once per data batch by the sum of gradients over all samples within this data batch.

The mathematical concept of DNN has been existing for a long time since 1960s (Bishop, 1995). It has been proved by Hornik et al. (1989) that a DNN with no less than two layers could potentially approximate any arbitrary functions given a sufficient number of perceptrons in each layer. However for a long time the application of DNN is very limited mainly for two reasons. First, unlike many other machine learning methods, the structure of DNN is highly independent from application. Such a structure on one hand ensures high flexibility of DNN in fitting into various applications from regression to classification, while on the other hand creates much more parameters to optimise compared to an alternative application dependent structure. Second, the large amount of parameters and the highly symmetric parameter space create a high level of analytic difficulty in optimisation. As a result the performance of the iterative optimisation in the practical implementation is highly dependent on the amount and diversity of available training data, as well as the computation resource. In addition, it is very difficult to interpret the role and the value of each parameter. As a result, for a very long time the DNN is unfavourable by many researchers. Fortunately, with the recent leap in the parallel computation based on the graphic processing unit (GPU) and the availability of big data, the parameter optimisation can be performed at a scale of millions of samples. With such recent advancement, the problems with DNN have been alleviated.

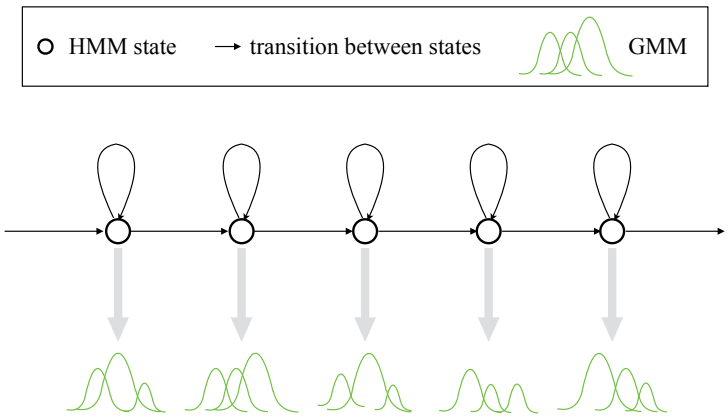


Fig. 2.3 *Illustration of HMM-GMM.*

2.2 Speech Recognition and DNN

2.2.1 Acoustic Model in Speech Recognition

In the speech recognition systems based on statistic models, a sequence of hypothesis words $\mathbf{w}$ is determined by maximizing the probability of hypothesis given the model parameters θ and the input observations $\mathbf{X}$.

$$\bar{\mathbf{w}} = \arg\max_{\mathbf{w}} P(\mathbf{w}|\mathbf{X},\theta) \tag{2.17}$$

with Bayesian rule,

$$\bar{\mathbf{w}} = \arg\max_{\mathbf{w}} \frac{p(\mathbf{X}|\mathbf{w},\theta)P(\mathbf{w}|\theta)}{p(\mathbf{X},\theta)} \tag{2.18}$$

$$= \arg\max_{\mathbf{w}} (\log p(\mathbf{X}|\mathbf{w},\theta) + \log P(\mathbf{w}|\theta) - \log p(\mathbf{X},\theta)) \tag{2.19}$$

where $p(\mathbf{X}|\mathbf{w},\theta)$ is acquired from the acoustic modelling and $P(\mathbf{w}|\theta)$ from the language modelling. The item $\log p(\mathbf{X},\theta)$ can be ignored because its value is independent from parameters $\mathbf{w}$,

$$\bar{\mathbf{w}} = \arg\max_{\mathbf{w}} (\log p(\mathbf{X}|\mathbf{w},\theta) + \log P(\mathbf{w}|\theta)) \tag{2.20}$$

Therefore, the name "acoustic modelling" corresponds to the fact that the optimization is based on the acoustic features $\mathbf{X}$.

In the large vocabulary continuous speech recognition (LVCSR), modelling word sequences directly is impractical due to the data sparsity issue, and modelling smaller units is usually preferred. The small unit could be phonetic such as monophone and triphone, or statistic such as the automatically clustered states. The conversion between the sequence

of small units to a word is achieved with the pronunciation model or the lexicon model (Lu et al., 2013). The most likely word sequence is acquired from the candidate words with highest overall probability given the acoustic observation, with the probability of each candidate word acquired from the probability of its component units. As shown in Fig. 2.3, the probability of each unit with given the acoustic observation could be modelled with a Gaussian mixture model (GMM). Therefore the acoustic modelling could be modified to estimate the likelihood of a hypothesis state sequence $\mathbf{q} = [q_0, q_1, q_2, ...]$, namely estimating $p(\mathbf{X}|\mathbf{q}, \theta)$.

One side effect of such hierarchical splitting is the complexity in finding the global maxima considering all possible sequences. Therefore the hidden Markov model (HMM) is introduced to simplify the sequence structure with the context modelling among hypothesis units. The simplification is made possible with a conditional independence assumption that the current state is only dependent on the previous state, and that the current observation is only dependent on the current state, as shown in Fig. 2.3. Denote the time sequence of observation feature vectors as $\mathbf{X} = [\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_N]$, where $\mathbf{x}_n$ is the observation feature vector at a discrete index n corresponding to the state q_n. Then the conditional independence assumption is

$$\begin{aligned} p(\mathbf{X}|\mathbf{q}, \theta) &= p\Big([\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_N]\Big|[q_0, q_1, ..., q_N], \theta\Big) \\ &\approx p(\mathbf{x}_0|q_0, \theta)\prod_{i=1}^{N} p(q_i|q_{i-1}, \theta)p(\mathbf{x}_i|q_i, \theta) \end{aligned} \tag{2.21}$$

where $p(q_i|q_{i-1}, \theta)$ is the transition probability from the state q_{i-1} to the state q_i given HMM parameters θ and $p(\mathbf{x}_i|q_i, \theta)$ is the observation probability which can be modelled with GMM, as illustrated in Fig. 2.3

$$p(\mathbf{x}_i|q_i = m) = \sum_{k=0}^{K} c_{mk}\mathcal{N}(\mathbf{x}_i; \mu_{mk}, \Sigma_{mk}) \tag{2.22}$$

where c_{mk} is the weight for the k-th Gaussian mixture when the state $q_i = m$.

The optimisation of HMM-GMM parameters is performed iteratively based on the expectation-maximization algorithm, namely the EM algorithm. At the expectation step, the GMM parameters are fixed and the state sequence is optimised in a backward order:

$$\hat{q}_i = \arg\max_{q_i} P(q_i|\mathbf{X}, \theta) \tag{2.23}$$

Therefore the expectation step only estimates the expected state sequence. At the maximization step, the state sequence is fixed and the GMM parameters are updated with corresponding observations based on the expected state sequence. Therefore the maximisa-

tion step only maximises the probability of the acoustic features given corresponding state. More details about the EM algorithm are skipped here, as they can be found in a variety of tutorials (Gales and Young, 2008; Renals and Hain, 2010; Young, 1996).

2.2.2 DNN in ASR

The work on using DNN to improve the acoustic modelling in the speech recognition system could be roughly categorized in two groups: those employ DNN in the front-end to produce improved acoustic features for HMM-GMM (Grezl et al., 2007; Hermansky et al., 2000; Liu et al., 2014; Rath et al., 2014), and those employ DNN as a part of acoustic model in place of GMM (Bengio et al., 1992; Hermansky et al., 2000; Seide et al., 2011). The first category is further referred to as the "DNN-HMM-GMM" system, and the second category is further referred to as the "DNN-HMM" hybrid system. In both categories the DNN is usually optimised as a multi-class classifier using labelled training data, with each neuron in DNN output layer corresponding to one class defined phonetically or statistically.

In the DNN-HMM-GMM system, DNN is employed in the front-end to generate improved representations or features, and there is no change in the HMM-GMM based acoustic modelling (Section 2.2.1). Therefore the DNN optimisation could be performed independently from the optimisation of the HMM-GMM based acoustic model. The representations produced by DNN can be combined with the traditional features such as the perceptual linear prediction (PLP) and Mel-frequency cepstral coefficient (MFCC) to build HMM-GMM (Bell et al., 2013; Liu et al., 2014; Rath et al., 2014), or they can completely replace the traditional features in training acoustic model (Grezl et al., 2007; Liu et al., 2014).

The representations could be generated from the final output of DNN, as well as the output of some hidden layer in DNN. As mentioned in Section 2.1, the final output of DNN approximates posteriors, thus the representations using the DNN final output are also referred to as the "probabilistic features" (Grezl et al., 2007). Grezl et al. (2007) has employed the probabilistic features independently in the acoustic modelling (Fig. 2.4 (a)), while Bell et al. (2013) concatenated the probabilistic features with the PLP features (Fig. 2.4 (b)). When the output of DNN hidden layer is used as representations (Hermansky et al., 2000), the number of neurons in that layer is usually constrained to control the feature dimension. As a consequence such a hidden layer looks like a "bottleneck layer" of the DNN in shape, and the corresponding representations are referred to as the "bottleneck features" (Grezl et al., 2007). Similar with the probabilistic features, the bottleneck features have been used both independently (Fig. 2.4 (c)) and in combination with the traditional

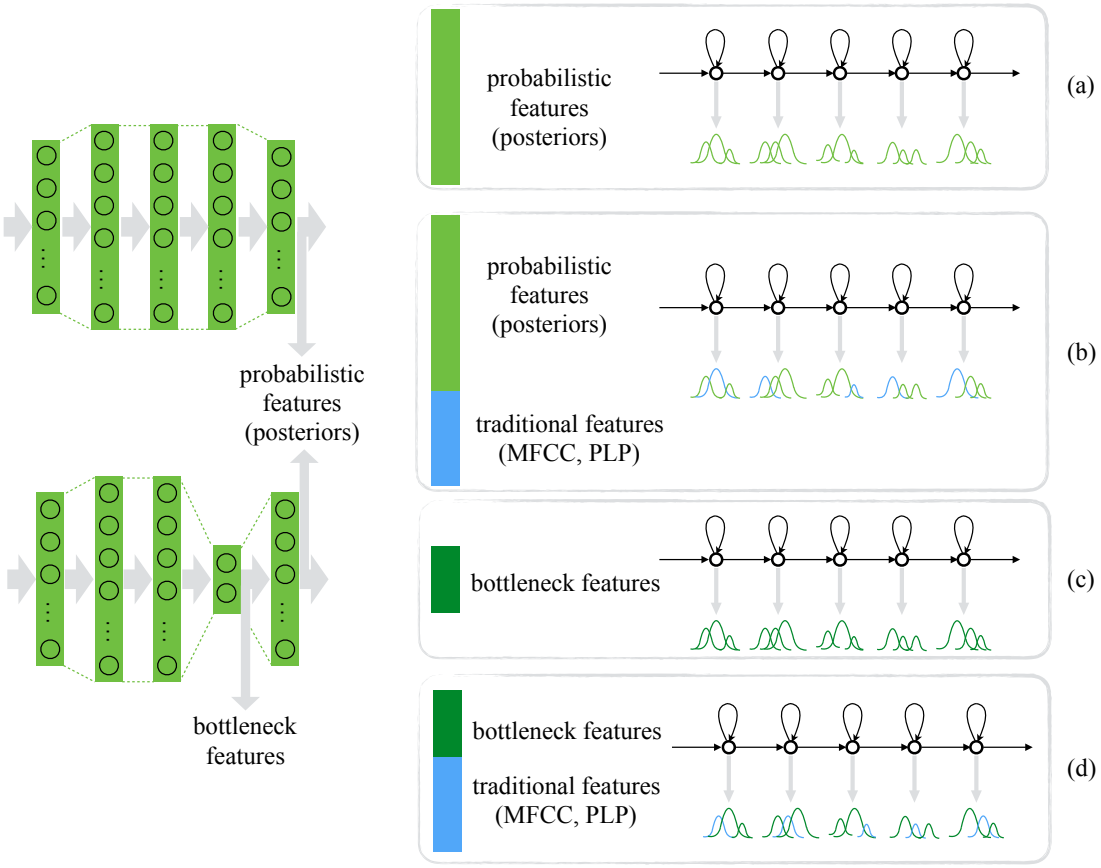


Fig. 2.4 *A variety of strategies in generating representations with DNN.*

features for acoustic modelling (Fig. 2.4 (d)) (Grezl et al., 2007; Hermansky et al., 2000; Liu et al., 2014; Rath et al., 2014).

In the DNN-HMM hybrid system, as shown in Fig. 2.5 the output of the m-th neuron in the DNN output layer approximates the posterior of corresponding class given the observation, namely $P(q_i = m|\mathbf{x}_i, \theta)$. With Bayesian rule the likelihood could be estimated from the posterior via

$$P(\mathbf{x}_i|q_i = m, \theta) = \frac{P(\mathbf{x}_i)P(q_i = m|\mathbf{x}_i, \theta)}{P(q_i = m)} \tag{2.24}$$

Therefore, DNN could replace GMM in estimating the likelihood in acoustic modelling. Since the probability of each class $P(q_i = m)$ is independent from both the current observation and the model parameters, it is a statistic prior that could be approximated by counting the occurrence frequency of each state class in a sufficiently large dataset. The probability $P(\mathbf{x}_i)$ could be neglected during the optimisation as it is independent from the model parameter value.

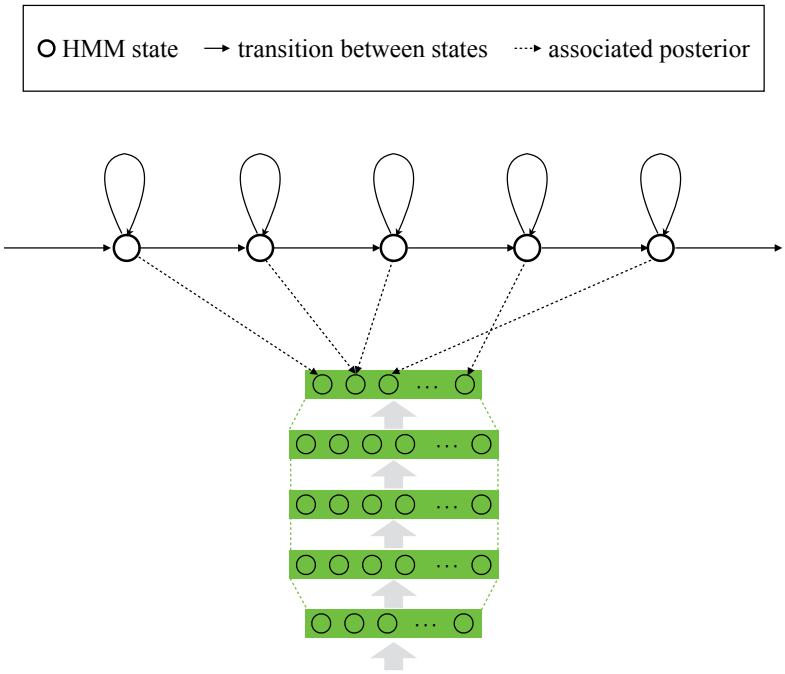


Fig. 2.5 *Illustration of DNN-HMM hybrid system.*

An early work by Hennebert et al. (1997) suggested that in the DNN-HMM hybrid system, DNN parameters can be optimised jointly with HMM parameters using a generalized EM algorithm, *i.e.* an unsupervised iterative optimisation similar to the case of HMM-GMM training. However in most of the state-of-the-art implementations, the HMM in hybrid system is directly inherited from a primary HMM-GMM system which also provides the alignment for the supervised training of DNN. Therefore, the DNN-HMM hybrid acoustic model in a state-of-the-art recognition system is actually trained in four steps:

- The HMM-GMM based acoustic model is trained with traditional features.
- The trained acoustic model is used to align the ground truth transcript on the training data to produce frame level labelling.
- The frame level labelling is used to train a DNN, and the DNN parameters are optimised to minimise a cost function, *e.g.* the cross-entropy.
- The HMM from the HMM-GMM system trained in the first step is combined with the DNN trained in the third step by converting the posteriors from the DNN output into the likelihood for HMM.

Recently, the dramatic increase in available computation resource and data resource has boosted various research on alternative deep networks and deep structures. In particular,

two types of network are attracting more and more interest. They are the convolutional neural network (CNN) and the long short term memory (LSTM) recurrent neural network (RNN). CNN was first found effective for image recognition by Krizhevsky et al. (2012). Later it is proved to outperform the DNN in the speech recognition systems of a hybrid CNN-HMM structure (Abdel-Hamid et al., 2012). Research work has been conducted to confirm the advantage of the convolution-based filtering along both the time axis (Lee et al., 2009) and the frequency axis (Abdel-Hamid et al., 2012). Compared with DNN, CNN is found to be more robust against the background noise, particularly when the CNN is implemented in a very deep structure (Qian et al., 2016; Sercu and Goel, 2016). The LSTM is an improved implementation of the original RNN and it alleviated the famous gradient vanish problem for RNN in the back-propagation through time (BPTT) (Hochreiter and Schmidhuber, 1997). For acoustic modelling, the LSTM has been reported to achieve better performance than the DNN and similar or better performance than the CNN (Graves et al., 2013; Qian et al., 2016; Sak et al., 2014).

There is also progress achieved in the optimisation of network parameters when the training data is very limited. A generative pretraining of network parameters based on the restricted Boltzmann machine (RBM) is found to outperform a random initialisation (Dahl et al., 2010), and the data augmentation is found helpful to improve the robustness of network by adding simulated training data (Cui et al., 2015; Ko et al., 2015; Tüske et al., 2014).

Similarly, progress has been made regarding the cost function for DNN training. The frequently adopted DNN optimisation objective, namely the cross-entropy minimization, is based on either the phonetic units or the statistic units rather than the words. As a result the DNN optimisation objective is not necessarily consistent with the overall objective of the speech recognition. Kingsbury (2009) proposed to optimise the DNN parameters based on the sequence classification criteria which weights the objective function of each sequence by the percentage of correct phonemes or correct words in given hypothesis utterance (Povey and Woodland, 2002). In addition, Kingsbury (2009) highlighted that the sequence training could be implemented using a similar computation structure to the cross-entropy based DNN training. Since the sequence training is in nature much easier to overfit to training data compared to the cross-entropy based training, in practice it is only used to finetune the parameters at the last stage. For example, the implementation by Veselý et al. (2013) finetunes DNN parameters based on the state-level minimum Bayesian risk (sMBR) objective or the minimum phone error (MPE) objective for a few iterations after the cross-entropy based DNN training has converged. It is observed that such a strategy well combines the merits of both methods and improves the speech recognition performance significantly (Veselý et al., 2013).

Regarding the DNN input, it is found that using different types of features can produce similar performance, e.g. PLP features, MFCC features and logarithmic Mel-frequency filter bank coefficients (FBANK) (Liu et al., 2014; Swietojanski et al., 2013). In comparison, using a concatenation of features from a few neighbouring frames generally improves the recognition performance with some context information (Grezl et al., 2007; Seide et al., 2011). It is also found that the DNN has a very high flexibility in modelling. With different types of features combined at input, DNN could learn multiple properties of data altogether. For example combining the normal DNN input of standard features such as PLP, MFCC and FBANK with the speaker code (Abdel-Hamid and Jiang, 2013) or the speaker i-vector (Liu et al., 2015; Saon et al., 2013) creates an effect of speaker adaptation of the DNN. Similar adaptation effect is observed when the standard input features are combined with noise information (Seltzer et al., 2013) or the room information (Giri et al., 2015). Furthermore, Liu et al. (2014) has proved that such feature combination strategy could be potentially extended to any auxiliary features that provides complimentary information to standard features for DNN to perform the acoustic modelling and the condition adaptation at the same time. This study corresponds to an earlier finding by Seide et al. (2011) that some traditional input transform which used to improve the performance of HMM-GMM system is not effective for DNN based system, particularly when a large amount of training data is available. One such example pointed by Seide et al. (2011) is the speaker adaptation algorithm vocal tract length normalisation (VTLN).

2.3 Robustness in Distant Speech Recognition

The speech recognition performance is partly determined by the recognition model used, and partly by the recordings from which the features are generated. Recording configuration such as the microphone installation can significant impact the quality of audio recordings and the extracted features. One current interest in speech recognition research and application is the occasion where the microphone is installed at a fixed location, in a scenario frequently referred to as the far-field recording or the distant recording. This is in contrast with a more advantageous setup to ASR system where the microphone is equipped on the interested speaker, and the headset microphone or the lapel microphone help to reduce the distance between the microphone and the speaker as well as the adverse environment effects in the speech recordings.

From the signal processing point of view, the environment effects distort clean speech signal in two aspects: the additive distortion and the convolutional distortion. Denote the clean speech signal as $x(n)$ where n is the discrete sampling index, and denote the distant

speech recording signal as $y(n)$, then

$$y(n) = x(n) * h(n) + v(n) \tag{2.25}$$

where $h(n)$ refers to the convolutional distortion and $v(n)$ refers to the additive distortion.

The additive distortion is represented by the background noise and the competing speech from other speakers if there is any. The electronic noise in the recording equipments is usually neglectable in the additive distortion. The convolutional distortion comes from the acoustic and electroacoustic components in the recording system. It is a joint outcome of the electroacoustic property of the microphone used for recording, and the acoustic property of the physical environment where the recording takes place. The overall effect of the convolutional distortion is usually approximated with a finite infinite response (FIR) filter $h(n)$, which is frequently referred as the acoustic impulse response (AIR). When the electroacoustic distortion caused by recording equipments is neglectable, the convolutional distortion is mainly determined by the room acoustics. Therefore the convolutional distortion is frequently referred to as the reverberation which could be approximated by another FIR filter. Such an FIR filter is usually referred to as the room impulse response (RIR), though it is not only dependent on the room acoustic properties but also on the installation of microphone, the speaker location in the room and the talking direction of the speaker.

There is one intrinsic difference between the distortion caused by additive distortion and convolutional distortion. In many cases the background noise and the competing speech can be assumed statistically uncorrelated with the targeted speech, thus the additive distortion could be modelled independently based on the statistic properties of the distortion source signal. On the contrary, the convolutional distortion is dependent on both the environment effect and the targeted clean signal itself. Due to this fundamental difference, different strategies have been used to treat the additive distortion and the convolutional distortion.

2.3.1 Room impulse response measurement

Before going into the details about the robustness algorithms for distant speech recognition, the estimation of RIR is briefly reviewed. RIR can be estimated with a synchronised recording of the signal produced by the sound source and the signal received by the sound receiver. Research has been conducted on the proper signal for such recording. It is found that for an occupied room the pseudo-random white noise is the most suitable signal for RIR estimation (Stan et al., 2002). However the pseudo-random white noise based RIR estimation is a non-linear method with limited SNR and the measurement requires tedious calibration. In comparison the swept sine signal, or the chirp signal, could avoid the non-

linear behaviour in electro-acoustic equipments (Stan et al., 2002). Thus it is frequently used as the RIR measurement signal for a high SNR in the estimation results when the room is unoccupied and quiet. The swept sine signal based method is very sensitive to the background noise during measurement recording, thus multiple recordings are usually conducted to increase the SNR against the potential background noise. These two types of signal are the most frequently used for RIR estimation. There are also other options such as the time-stretched pulses (Stan et al., 2002) and the maximum length sequence (Guidorzi et al., 2015), but they are not used as frequently in applications. One reason is that such methods usually rely on the assumption of a linear time-invariant (LTI) system and it causes distortion artefacts in the derived RIR when this assumption is not met.

For the RIR measurement based on the swept sine signal, the signal played by the loudspeaker is the sine wave with its frequency increasing linearly over time up to Nyquist frequency, *i.e.*

$$x(t) = A\sin(\omega(t)\cdot t) \tag{2.26}$$
$$\omega(t) = \alpha t \tag{2.27}$$

where A is a fixed amplitude to scale the sound volume, α is the increasing speed of instantaneous angular frequency of the swept sine signal. In discrete signal, equivalently

$$x(n) = A\sin(\omega(n)\cdot n) \tag{2.28}$$
$$\omega(n) = \alpha\cdot\frac{n}{f_s} \tag{2.29}$$

where f_s is the sampling rate. The room impulse response could be estimated from the measurement recordings with an inverse filter or a matching filter of the swept sine signal, and the matching filter is a simple time reverse flip of the swept sine signal itself (Kuttruff, 2000). Assuming that the transfer function of the loudspeaker to play the swept sine signal is flat enough to be ignored and that the recording environment is quiet, denote the corresponding microphone recording of the swept sine signal as $y(n)$, *i.e.*

$$y(n) \approx x(n) * h(n) \tag{2.30}$$

and the RIR can be estimated with

$$h(n) \approx y(n) * x(N-n) \tag{2.31}$$

where N is the length of measurement signals $x(n)$ and $y(n)$.

2.3.2 Noise Suppression and Robustness

Before the recent revival of DNN and the increasing popularity of other deep networks, research work on noise robustness in DSR has been mainly conducted in two aspects: the speech enhancement, *i.e.* directly enhancing the speech signal or the speech feature with noise suppression, and the noise robust acoustic modelling.

Among the speech enhancement techniques, some algorithms directly estimate the desired speech signal or feature, and some other algorithms perform the noise modelling first. The direct enhancement of speech signal is based on a reliable knowledge of the noise signal statistics, assuming that speech signal is not correlated with the noise signal. Examples are the spectral subtraction (Berouti et al., 1979; Boll, 1979; Lim and Oppenheim, 1979) and the Wiener filtering or minimum mean square error (MMSE) estimator (Chen et al., 2006; Lim and Oppenheim, 1979). The performance of such algorithms is however largely dependent on the availability of an accurate knowledge of the noise spectrum. In addition, the direct spectral subtraction creates the residual musical noise, and the Wiener filtering causes speech distortion which is proportional to the amoutn of noise suppressed (Chen et al., 2006). An alternative speech enhancement method is the mean and variance normalisation to remove the stationary component in the background noise (Furui, 1981). The normalisation is widely adopted for its simplicity and robustness. The segmental feature vector normalisation is even found by Viikki and Laurila (1998) to outperform some more advanced algorithm such as the parallel model combination. One typical noise modelling based speech feature enhancement algorithm is the stereo based piecewise linear compensation for environment (SPLICE) (Deng et al., 2001, 2004, 2005, May 2002). SPLICE assumes that the relation between the noisy speech and clean speech is piecewise linear, thus the speech component could be restored with an additive correction vector. As its name suggests, to estimate the correction vector, SPLICE requires both the clean data and the noisy data to be available in parallel.

To apply the noise suppression and compensation on the acoustic model directly, the vector Taylor series (VTS) can be used to approximate the function between the clean speech based acoustic model parameters and the noisy recording based acoustic model parameters (Moreno et al., 1996). This VTS function is also referred to as an environment function by Moreno et al. (1996), and it is based on the assumption that both the speech component and the noise component could be modelled with GMM. The training of VTS is performed iteratively with the EM algorithm to maximize the likelihood of acoustic model. In each iteration, the environment function is renewed with an VTS expansion around the updated mean vectors of GMMs for the clean speech component, followed by a re-estimation of the mean vectors and variance matrics for the additive noise component and

the mean vectors for the convolutional distortion. A simplified variant of VTS algorithm is the Jacobian adaptation which compensates at the first order for the additive noise that causes a mismatch between clean training data and noisy test data (Sagayama et al., 1997). Another algorithm based on the training-test mismatch compensation is the uncertainty decoding, as well as its variant joint uncertainty decoding (JUD) (Liao and Gales, 2008). There are many more noise robust techniques, such as the missing feature method (Cooke et al., 1997), and their details are skipped here.

2.3.3 Dereverberation

A lot of research effort has also been devoted to improving the reverberation robustness of the DSR performance. Based on the fact that reverberation is in nature a convolutional distortion which could be approximated with FIR filters, the earliest work on dereverberation was based on inverse filtering using recordings from one or multiple microphone channels (Miyoshi and Kaneda, 1988). It was proved theoretically possible to achieve an exact inverse of the room acoustics using multi-channel reverberant recordings when the RIRs corresponding to the multiple recording channels do not share any common zeros in the z-plane. The work by Miyoshi and Kaneda (1988) laid the theoretic foundation for one popular catetory of dereverberation: multiple-input/output inverse theorem (MINT). This is the first turning point in the dereverberation research, and since then using multiple channel recordings is widely observed to outperform the counterparts using single channel recordings only (Eneman and Moonen, 2007; Furuya et al., 2006; Gaubitch et al., 2008; Kodrasi and Doclo, 2014; Otsuka et al., 2014; Yoshioka and Nakatani, 2012).

The second turning point in the dereverberation research comes from the findings about different roles of the early reflections and the late reflections in reverberation. The early reflections, typically of 50-80 ms after the arrival of the direct sound, are strongly dependent on the speaker and microphone location (Yoshioka et al., 2012). The difference in speaker-microphone distance is found to cause a significant variation in early reflections (Bradley, 2011). In comparison, the late reflections have an exponentially decaying energy independent from the speaker location and the microphone location (Habets, 2005; Yoshioka et al., 2012). The early reflections introduce a perceptual colouration on the speech signal (Assmann and Summerfield, 2004; Gaubitch et al., 2008). The human speech perception research observes that the late reflections are much more harmful than the early reflections in reducing the speech intelligibility for both the impaired and the non-impaired listeners (Bradley et al., 2003; Hu and Kokkinakis, 2014). Therefore some dereverberation algorithms relax the treatment of the early reverberation so that the dereverberation parameters are optimised to suppress the late reverberation (Habets, 2005;

Hikichi et al., 2007; Kodrasi and Doclo, 2012; Yoshioka and Nakatani, 2012; Zhang et al., 2009). This strategy increases the overall performance of dereverberation as it focuses on the most harmful part of the reverberation, namely the late reverberation, which is also relatively insensitive to the location change of the talking speaker and the recording microphone.

Other research work also found that the robustness of dereverberation can be increased by introducing regularization (Hikichi et al., 2007), channel shortening (Zhang et al., 2010, 2009), spatial-temporal and spectral processing (Gaubitch et al., 2008), *etc.* In addition, research efforts have been devoted to modelling RIRs to cope with diverse environment conditions in test data via the acoustic model selection based on the reverberation level (Liu and Yang, 2015).

In an extensive experimental validation performed by Eneman and Moonen (2007) on multi-microphone based dereverberation algorithms in speech recognition tasks, some of the classical solutions obtained moderate benefit, e.g. beamforming, cepstral dereverberation and unnormalised matched filtering. In comparison, the more advanced subspace-based dereverberation techniques failed to enhance signals in the context of speech recognition task despite their high-computational load. Eneman and Moonen (2007) pointed out three main reasons for the poor performance of the more advanced techniques: the sensitivity to model order mismatch, the additive noise and the time varying acoustics in real life.

A multi-channel based dereverberation algorithm, the generalized weighted prediction error (GWPE) (Yoshioka and Nakatani, 2012), attracted much recent attention for its application in the best speech recognition system in ReverbChallenge 2014. ReverbChallenge 2014 is a research competition that provides a shared framework for the comparison of dereverbertion performance in the speech enhancement task and in the speech recognition task (Kinoshita et al., 2016). As an extension of the weighted prediction error (WPE) method Yoshioka and Nakatani (2012), the GWPE algorithm is a blind multiple-input multiple-output (MIMO) dereverberation algorithm based on a linear prediction of the late reverberation only. The philosophy of both WPE and GWPE is to predict the late reverberation in current speech recording samples based on previous speech recording samples with a linear autoregressive filter. Based on the prediction, the late reverberation could be suppressed or removed via a simple subtraction. In WPE the optimisation target of the autoregressive filter parameters is to minimised the reverberation prediction error assuming that the clean speech recordings are available in parallel with reverberant recordings. The GWPE improved the cost function to remove the dependence on the clean speech recordings available in parallel with reverberant recordings. The details of GWPE algorithm is quickly gone through below.

GWPE performs dereverberation in the spectrum domain as it is more computationally efficient than in the time domain. To further reduce the computation cost, the subband-counterparts are preferred to the raw spectrum. Denote the subband based spectrum vectors for the reverberant speech recordings as $\mathbf{y}(\tau)$, for the clean speech signal as $\mathbf{s}(\tau)$ and for the background noise as $\mathbf{v}(\tau)$, where τ is the frame index, *i.e.*

$$\mathbf{y}(\tau) = [y_0(\tau), y_1(\tau), ..., y_{L-1}(\tau)]^T \tag{2.32}$$
$$\mathbf{s}(\tau) = [s_0(\tau), s_1(\tau), ..., s_{L-1}(\tau)]^T \tag{2.33}$$
$$\mathbf{v}(\tau) = [v_0(\tau), v_1(\tau), ..., v_{L-1}(\tau)]^T \tag{2.34}$$

where L is the number of subbands. Assume that the background noise is statistically independent from the clean speech signals and that the following linear relationship holds

$$\mathbf{y}_l(\tau) = \sum_{n=0}^{J_l-1} \mathbf{H}_l^*(n)\mathbf{s}_l(\tau-n) + \mathbf{v}_l(\tau) \tag{2.35}$$

where "*" refers to conjugate transpose, $\mathbf{H}_l(n)$ is a matrix including a list of complex valued filters associated with the MIMO RIRs, and J_l is the order of the filter in the l-th subband. The auto-regressive filters used to estimate late reverberation from preceding samples is denoted as $\mathbf{G}_l(n)$, and the restored speech signal is denoted as $\hat{\mathbf{x}}_l(\tau)$, thus

$$\mathbf{r}_l(\tau) = \sum_{n=\triangle}^{\triangle+K_l-1} \mathbf{G}_l^*(n)\mathbf{y}_l(\tau-n) \tag{2.36}$$
$$\hat{\mathbf{x}}_l(\tau) = \mathbf{y}_l(\tau) - \mathbf{r}_l(\tau) \tag{2.37}$$

where K_l is the order of the late reverberation estimation filter in the l-th subband and $\triangle$ indicates the number of taps relaxed for early reflections. One consequence of reverberation as implied in Eq. (2.35) is the increase of auto-correlation in reverberant speech recordings. Therefore the cost function in GWPE to optimise the linear predictor $\mathbf{G}_l$ is chosen as a measurement of the auto-correlation level in the restored speech, namely the Hadamard-Fischer (HF) mutual correlation:

$$F = \frac{1}{|\mathscr{T}|} \sum_{\tau\in\mathscr{T}} \log(\det E(\hat{\mathbf{X}}_l(\tau)\hat{\mathbf{X}}_l^*(\tau))) - \log(\det E(\hat{\mathbf{X}}_l\hat{\mathbf{X}}_l^*)) \tag{2.38}$$

where $\hat{\mathbf{X}}_l = [\hat{\mathbf{X}}_l^*(T), ..., \hat{\mathbf{X}}_l^*(1)]^*$ and $\hat{\mathbf{X}}_l(\tau)$ is a vector of $\hat{\mathbf{x}}_l(\tau)$ from the recordings of multiple microphones.

Since speech is quasi-stationary within a short time span, there is a risk in minimizing the auto-correlation of the restored speech as it could potentially distort the temporal

structure in the speech spectrum as well. This problem is prevented in GWPE by the employment of channel shortening. This strategy ensures that the speech spectrum within the time span up to $\triangle$ preceding taps will be protected from distortion caused by dereverberation, as the auto-regressive filters for dereverberation are optimised to minimise the auto-correlation of signal. More details about the HF cost function and the analytic derivation for the dereverberation parameter optimisation are skipped here, and they could be found in the work by Yoshioka and Nakatani (2012).

GWPE has inherited a few general properties in existing blind MIMO algorithms, and such properties provide high flexibility for application no matter whether the dereverberation algorithm is used alone or jointly with other algorithms. First, the dereverberation operation produces the same number of channels in the output compared to the input. Second, it does not require any knowledge about the number of sound sources. Third, the dereverberation is a linear operation with channel shortening on the implicit RIRs. This prevents the non-linear distortion by a large degree. Fourth, the process conserves the time differences of arrival (TDOAs) in raw recordings from the multiple microphones. This allows an effective application of successive beamformers. In the best system of ReverbChallenge 2014 developed by researchers in Nippon Telegraph and Telephone (NTT), the GWPE achieved 30% relative WER reduction, and the minimum variance distortionless response (MVDR) beamforming on top achieved 30% further relative WER reduction (Kinoshita et al., 2016). The combination of GWPE and MVDR is later proved effective in the third CHiME challenge as well (Barker et al., 2016).

2.3.4 Beamforming

Beamforming is especially popular among the multi-channel based signal processing techniques. It has been found effective to improve the robustness of DSR against the background noise and reverberation. There are two shared points among all beamforming algorithms. First, they all perform spatial filtering, either explicitly or implicitly. With different spatial filtering, the beamformer response differs. Beamformer response refers to the amplitude and phase presented to a complex plane wave as a function of location and frequency. Second, most beamformers benefit from pre-steering in the implementation, *i.e.* aligning the time of recordings from multiple microphones beforehand. An accurate pre-steering could improve the efficiency and stability when optimising the beamforming parameters. To simplify the notation, the output of pre-steering, *i.e.* the time-delayed recording from the i-th microphone channel is denoted as

$$y_i^d(n) = y_i(n - \triangle_i) \tag{2.39}$$

where n is the sample index in the recording and $\triangle_i$ is the number of samples shifted during pre-steering.

The simplest while robustest beamforming algorithm is the delay and sum beamforming (DSB). Denote the i-th channel recording as $y_i(n)$, then the output of DSB is

$$\hat{x}(n) = \frac{1}{N}\sum_{i=1}^{N} y_i^d(n-\delta_i) \tag{2.40}$$

where δ_i is the delay in the i-th channel, and the time delays are the only parameters for DSB. DSB is effective on additive noise when the time alignment in pre-steering is accurate enough and the spatial average of the background noise approaches zero statistically. One example scenario suitable for DSB is the diffuse noise field (Jacobsen, 1979).

The weighted delay and sum beamforming (wDSB) makes one simple improvement over DSB, by introducing channel weights $w_i(n)$ to dynamically boost the advantageous microphone channels:

$$\hat{x}(n) = \sum_{i=1}^{N} w_i(n) y_i^d(n-\delta_i). \tag{2.41}$$

In application, the advantageous channels can be those geometrically close to the targeted speaker, those with largest speech magnitude in recordings, or those with the highest overall spatial correlation with any other channels (Anguera et al., 2007).

Both DSB and wDSB assume the background noise to be diffusive, as a result their effectiveness is limited when the background noise varies by location, frequency or sound arriving angle. This could be alleviated with spatial filtering implemented in the complex spectrogram domain. Denote the spectrum of the pre-steered recordings at the angular frequency ω as $\mathbf{y}(\omega)$. The time index is dropped here for notation simplicity without any causing any confusion. $\mathbf{y}(\omega) = [y_1(\omega), y_2(\omega), ..., y_N(\omega)]^T$ is a vector of complex spectrum values from all N microphone channels at the same angular frequency ω. Thus beamforming could be performed independently at each frequency bin as a simple matrix multiplication:

$$\hat{x}(\omega) = \mathbf{w}^*(\omega)\mathbf{y}(\omega) \tag{2.42}$$

where "*" refers to a conjugate transpose. Different beamformers optimise the spatial filtering parameters $\mathbf{w}(\omega)$ differently, and a few typical beamformers are detailed below.

The minimum variance distortionless response (MVDR) beamforming optimises the beamforming parameters $\mathbf{w}(\omega)$ by minimizing the overall variance of beamforming output. The signal distortion caused by this process is alleviated with a distortionless constraint

that

$$\mathbf{w}^*(\omega)\mathbf{v}(\omega) = 1 \tag{2.43}$$

where $\mathbf{v}(\omega)$ is the array manifold vector in the arriving direction of the targeted speech sound. In this case, the targeted speech is assumed to be a plane wave. The elements in vector $\mathbf{v}(\omega)$ indicate the phase shift in the recordings from different microphones, *i.e.*

$$\mathbf{v}(\omega) \triangleq [e^{-j\triangle\theta_1(\omega)}, e^{-j\triangle\theta_1(\omega)}, ..., e^{-j\triangle\theta_N(\omega)}]^T \tag{2.44}$$

As detailed by Wölfel and McDonough (2009) that the optimal solution to the beamfomring parameters for MVDR is

$$\mathbf{w}^*_{\text{mvdr}}(\omega) = \Lambda(\omega)\mathbf{v}^*(\omega)\Sigma_{\mathbf{N}}^{-1}(\omega) \tag{2.45}$$

$$\Lambda(\omega) \triangleq \left[\mathbf{v}^*(\omega)\Sigma_{\mathbf{N}}^{-1}(\omega)\mathbf{v}(\omega)\right]^{-1} \tag{2.46}$$

where $\Sigma_{\mathbf{N}}(\omega)$ is the spatial covariance of the background noise. Similar with the speech component, denote the spectrum of noise component recordings from multiple microphones as $\mathbf{n}(\omega)$, then the spatial covariance of the background noise is

$$\Sigma_{\mathbf{N}}(\omega) = E\left[\mathbf{n}(\omega)\mathbf{n}^*(\omega)\right] \tag{2.47}$$

where $E[\cdot]$ refers to expectation operation.

One variant MVDR beamformings is the super-directive beamforming (SDBF) which improves the directivity of MVDR at low-frequency. This is achieved by replacing the spatial spectral matrix $\Sigma_{\mathbf{N}}(\omega)$ with the coherence matrix $\Gamma_{\mathbf{N},m,n}(\omega)$ corresponding to a diffuse noise field:

$$\Gamma_{\mathbf{N},m,n}(\omega) = \text{sinc}\left(\frac{\omega d_{m,n}}{c}\right) \tag{2.48}$$

where $d_{m,n}$ is the distance between the m-th microphone and n-th microphone in the microphone array and c is the sound speed.

The performance of MVDR could also be improved with postfiltering, *i.e.* a frequency dependent weighting to MVDR beamforming output. This has been shown by Wölfel and McDonough (2009) to lead to another beamforming algorithm, namely the minimum mean square error (MMSE) beamforming, which optimises the beamforming parameters by minimizing the overall energy of beamforming output. Details about the derivation is skipped here and it could be found in the work by Wölfel and McDonough (2009). Assuming that the noise component and speech component are statistically uncorrelated in

the distant recordings, the solution to MMSE is

$$\mathbf{w}^*_{\text{mmse}}(\omega) = \Sigma_F(\omega)\mathbf{v}^*(\omega)\Sigma_{\mathbf{X}}^{-1}(\omega) \tag{2.49}$$

where $\Sigma_F(\omega)$ is the power spectral density of the distant recordings (Wölfel and McDonough, 2009).

There are many more classical beamforming algorithms than those mentioned above. They are skipped here and more details could be found in the work by Wölfel and McDonough (2009).

2.3.5 Environment Robustness with DNN

Recent progress in deep networks has brought new possibilities in improving the environment robustness of ASR system. One straightforward strategy is to combine the traditional noise and reverberation robust algorithms with the DNN based system in a pipeline structure, e.g. the DNN input features are based on the speech enhancement output in the front-end. The research work by Liu et al. (2014) and Swietojanski et al. (2013) proved that the wDSB is effective for both the DNN-HMM-GMM system and the DNN-HMM hybrid system in such a pipeline combination. In the summary of ReverbChallenge 2014 (Kinoshita et al., 2016), it is pointed out that the top recognition performance is achieved with a combination of speech enhancement, the DNN based acoustic model and advanced language model. In particular, the combination of GWPE based multi-channel dereverberation and the MVDR beamforming has played an important role in the NTT DSR system which achieved the best performance in both the ReverbChallenge 2014 (Kinoshita et al., 2016) and the third CHiME challenge (Barker et al., 2016).

An early research by Zuo et al. (2003) constructed a robust telephone speech recognition system with simulated data. When the system was evaluated on real data, it achieved the same performance with the system trained on real data using the same algorithms. However after that for a long time there is very limited report on similar applications in robust speech recognition. One breaking finding in ReverbChallenge 2014 is that an average DNN based system could achieve similar or much better performance than DNN free system equipped with multiple speech enhancement algorithms and with model combination (Kinoshita et al., 2016). In addition, compared to traditional front-end and GMM, DNN is found capable of adapting to a variety of training data properties, potentially due to its mathematical nature of being a universal function approximator (Hornik et al., 1989). These advantages of DNN brought back the research interest in using simulated data to enhance acoustic model in a way of multi-condition training, or multi-style training. To perform multi-condition training, the original training data is extended with a large amount

of extra training data simulated by distorting the original training data with different types and levels of noise and reverberation. The augmented training data and the original training data are mixed and randomized before optimising the DNN parameters in a standard way. The multi-condition training of DNN with the augmented simulated data is proved to be very effective in increasing the robustness of DNN against both the background noise and the environment reverberation in the third CHiME challenge (Barker et al., 2015).

One particular challenge in the application of DSR is the unpredictable diversity of user environment in terms of both the background noise and the environment reverberation. It is hardly possible to cover all test conditions even with the simulated data and with an unlimited computation resource. Compared to most traditional reverberation robust algorithms whose benefit is yet to be confirmed in the state-of-the-art DSR systems with deep networks, the multi-condition training is gaining more and more popularity due to its simplicity in implementation and its effectiveness widely observed in application. However this effectiveness in improving overall recognition performance is at a price of compromising the best performance in the conditions with little or no background noise or environment reverberation (Brutti and Matassoni, 2016). Therefore, a few research works propose to use an environment classifier, based on which the most proper acoustic model is selected for decoding (Brutti and Matassoni, 2014, 2016). In this way it is possible to maintain the best performance in each condition without losing the overall performance robustness. In such a system, a list of acoustic models are needed to cover a wide range of test environment conditions. In addition, the accuracy of the environment classification plays a key role in the whole system. Therefore, research effort has been devoted in the reverberation metrics and non-intrusive reverberation estimators which have a high correlation with speech recognition performance (Brutti and Matassoni, 2014, 2016; Parada et al., 2016; Parada, Sharma and Naylor, 2014).

In a few research work, it is found that feeding meta information at DNN input as auxiliary features concatenated with the standard features could also improve the robustness against particular types of noise and room reverberation. The noise aware training proposed by Seltzer et al. (2013) augmented the standard features with the noise modelling parameters in the DNN input to achieve better noise robustness, and the room aware training proposed by Giri et al. (2015) found that within limited test conditions even room ID could serve as a beneficial auxiliary feature.

Deep networks also make it possible to directly combine the multi-microphone recordings at feature level or signal level. It was first proved in the DNN-HMM hybrid system by Swietojanski et al. (2013) and in the DNN-HMM-GMM system by Liu et al. (2014) that simply concatenating features from multiple microphone channels at the DNN input could achieve a similar or better performance compared to wDSB, when the number of micro-

phone channel is not large. Since then, more research effort has been devoted to utilizing deep networks to perform beamforming-like function with multi-channel recordings. The advantage of such a structure is its potential in joint optimisation of the beamforming-like front-end and the acoustic model DNN, *i.e.* merging the optimisation of front-end with acoustic model which has been traditionally separated in two parts for engineering reasons. This philosophy leads to a variety of recent research work on the deep-beamforming (Xiao et al., 2016) and the end-to-end system (Sainath et al., 2015) for DSR.

Recent research progress also suggests the potential of further robustness improvement with novel deep networks and novel combination of deep networks. The convolutional neural network (CNN) is found effective for multi-channel based DSR. In the implementation by Swietojanski et al. (2014), the CNN filter weights are shared among multiple channels, and within each channel different filter weights and biases are allowed for convolutional bands. Yoshioka et al. (2015) examined the convolution in time and found that CNN could model the short time correlations in feature vectors with fewer parameters compared to DNN, thus achieving better generalization to unseen test environments. In addition, it is found by Yoshioka et al. (2015) that the performance improvement from CNN over DNN is complementary to the improvement from the traditional multi-channel dereverberation algorithm GWPE.

Both the system by Swietojanski et al. (2014) and Yoshioka et al. (2015) coupled CNN with follow-up DNN where all elements are fully connected. Recent research suggests that using multiple layers of CNN to make a very deep convolutional neural network structure (VDCNN) further improves the noise robustness compared to using shallow CNNs (Qian et al., 2016). Besides the increase in the overall number of CNN layers, two or more convolutional layers are adopted between every two pooling layers. Such a VDCNN-HMM hybrid acoustic model is found to provide similar performance with the LSTM-RNN based acoustic model in DSR of meeting data.

2.4 Reverberation Metric and Reverberation Measurement

Reverberation measurement was first used in the acoustic and physical research and applications. With recent progress in speech recognition, more and more interest is drawn to improving DSR performance in domestic applications which are typical for having reverberant environments. Since DNN is more sensitive to the mismatch between training data and test data compared to the traditional front-end and the GMM based acoustic model, research effort has been devoted to increasing the robustness of DNN with multi-condition training that covers a wide range of environment conditions (Barker et al., 2015). However

this is at a price of the best performance in data with little or no environment distortion. Therefore some researchers turn to model combination and model selection to maintain the best performance in each environment condition without compromising the overall robustness. For model selection to be effective on recordings from diverse reverberant environments, a reliable estimation of the reverberation level in given recordings is critical.

The early reverberation metrics used in speech recognition are from acoustics. The most popular example is the reverberation time, or T_{60}. It is the time it takes for the acoustic energy in the room to decay for 60 dB since an abrupt stop of the sound source from a steady acoustic status. A steady acoustic status refers to the status where the acoustic energy steadily produced by the sound source equals the acoustic energy absorbed or consumed in the concerned environment. Therefore T_{60} is an indication of the average room acoustic property (Bradley, 2011). Sabine's reverberation equation provides an engineering method to estimate the reverberation time of a room given the room geometry and the acoustic absorption coefficients of acoustic boundaries such as walls, ceiling, floor, window, *etc.* When using an RIR to approximate the reverberation effect of a room, T_{60} could also be estimated from the energy decay curve of RIR. With a higher level of room reverberation, it takes longer for the acoustic sound to vanish, resulting in a smaller tilt in the energy decay curve thus a larger T_{60}.

Later research on reverberation metrics paid more attention to the impact of sound reflections on speech perception in reverberant environment. Some research work pointed out that early reflections and late reflections play different roles in the reverberation perception by human as well as in the reverberation distortion on speech recognition. According to Haas (1972); Shankland (1977), the early reflections play an important role in human perception of room acoustic quality, as it introduces a colouration effect which is suggested to have a positive impact on the human intelligibility of speech with an effect similar to increasing the strength of direct-path sound, therefore increasing the effective signal-to-noise ratio (SNR) for both the impaired and non-impaired listeners (Bradley et al., 2003; Hu and Kokkinakis, 2014). This is particularly helpful in the scenarios where the direct sound is weak, for example when the talker's head is turned away from the listener or when the talker speaks at a position torwards the rear of the room (Bradley et al., 2003; Kuttruff, 2009). While the early sound reflections can be integrated with the direct sound, the late reflections cannot be integrated with the direct sound thus causing reverberation smearing (Hu and Kokkinakis, 2014). In addition, the early reverberation is strongly dependent on the talker and microphone positions as well as speaker-microphone distance (Bradley, 2011). In comparison, the magnitude of late reverberation decays approximately exponentially and the decaying rate is independent of talker or microphone positions (Yoshioka et al., 2012).

Reverberation time is a statistic metric which does not reflect the difference in the early reflections among multiple microphone channels placed at different locations in the room or at different distances to the talker, nor does it reflect the difference between early reflections and late reflections. Therefore a few reverberation metrics are proposed to highlight the different roles of early reflections and late reflections. Thiele (1953) proposed to use Deutlichkeit, the early to total sound energy ratio, to measure the clarity of speech. It is calculated as the energy ratio between the early reflections and all the reflections when the early reflections are defined to be the sound reflected within 50 ms after the arrival of direct sound, *i.e.*

$$D_{50} = \frac{\sum_{n=1}^{N_{50}} h(n)^2}{\sum_{n=1}^{N} h(n)^2} \tag{2.50}$$

where $\mathbf{h} = [h(0), h(1), ..., h(n), ..., h(N)]^T$ is the RIR and N_{50} refers to the discrete time index corresponding to 50 ms after the arrival of direct sound. In many literature the direct sound is included in the calculation of the early reflection energy, as in the RIRs measured from real rooms it is usually difficult to accurately separate the direct sound from the early reflections. Due to similar reasons, such energy ratio based speech clarity is also referred to as the direct-to-reverberation ratio (DRR).

Later C_{80} is proposed as a measure of clarity for musical sound and its variant C_{50} is preferred to measure the clarity for speech sound (Bradley, 2011):

$$C_{80} = 10\log\left(\frac{\sum_{n=1}^{N_{80}} h(n)^2}{\sum_{n=N_{80}+1}^{N} h(n)^2}\right) \tag{2.51}$$

$$C_{50} = 10\log\left(\frac{\sum_{n=1}^{N_{50}} h(n)^2}{\sum_{n=N_{50}+1}^{N} h(n)^2}\right) \tag{2.52}$$

where N_{80} refers to the discrete time index corresponding to 80 ms after the arrival of direct sound. In some literature, speech clarity C_{50} is also extended into a category of reverberation metric early-to-late reverberation ratio (ELR) (Brutti and Matassoni, 2014):

$$ELR_T = \frac{\sum_{n=1}^{N_T} h(n)^2}{\sum_{n=N_T+1}^{N} h(n)^2} \tag{2.53}$$

where T indicates the boundary between early reverberation and late reverberation. Thus C_{50} and C_{80} can also be extended into

$$C_T = 10\log ELR_T. \tag{2.54}$$

Both C_{50} and D_{50} have been listed as the recommended metrics for speech clarity by ISO 3382-1 (Bradley, 2011). Existing research observes that the speech clarity is sensitive to the location of both the sound source and the recording microphone. As the distance between the sound source and the recording microphone increases, both C_{50} and D_{50} decrease. In the measurement performed in real rooms by Harvie-Clark and Dobinson (2013), it is observed that C_{50} is slightly larger when the sound source is located at the corner of the room compared to being located next to the middle of one wall, and in both cases the sound source faces the center of the room.

A few recent research work has investigated the correlation between the reverberation metrics and the speech intelligence. Sehr et al. (2010) analysed the correlation between the attenuation in RIR coefficients and the recognition accuracy using simulated data. Their work provided an experimental support for using reverberation metric D_{50} to predict the speech recognition performance on data of various reverberation levels. Parada, Sharma and Naylor (2014) compared T_{60}, DRR, D_{50} and C_{50} regarding their correlation with human speech perception via the speech quality score "perceptual evaluation of speech quality" (PESQ), and the C_{50} is found to provide the highest correlation among all. In addition, a further investigation on the boundary between early reflections and late reflections in ELR confirmed that using 50 ms as the boundary provides the highest correlation between ELR and PESQ score.

The ELR based reverberation measurement results over a large amount of data have also been observed to provide a high correlation with the speech recognition performance, and are thus employed for acoustic model selection in DSR (Brutti and Matassoni, 2016; Parada et al., 2015). However there are different opinions regarding the optimal boundary between early and late reverberation in ELR that provides the highest correlation with speech recognition performance. The 50 ms has been confirmed a good boundary between early and late reflections for human speech perception (Bradley, 2011; Parada, Sharma and Naylor, 2014), however Brutti and Matassoni (2014) found that using 110 ms in ELR calculation, namely the C_{110}, provides a higher correlation with the word recognition accuracy than the C_{50}, independent of the recognizer complexity. In further research by Brutti and Matassoni (2016) on the state-of-the-art speech recognition system based on the DNN-HMM hybrid acoustic models, the C_{110} is observed to provide a high correlation with the performance of large vocabulary continuous speech recognition on different evaluation data from Aurora 4 (Parihar et al., 2004), the third CHiME challenge (Barker et al., 2015), and ReverbChallenge 2014 (Kinoshita et al., 2016).

The high correlation between reverberation the metric ELR and the speech recognition performance boosted further research on using estimated reverberation level to improve DSR performance. Parada et al. (2015) explored several methods that use C_{50} to improve

the speech recognition performance on the ReverbChallenge data. The C_{50} value has been appended directly to ASR features, with an optional feature dimension reduction via heteroscedastic linear discriminant analysis (HLDA). It is found that appending C_{50} slightly improves the speech recognition performance when HLDA is used, and the improvement is relatively robust against the C_{50} estimator. In addition, multiple acoustic models can be prepared with the simulated data corresponding to different C_{50} value ranges, so that at test stage acoustic model could be selected based on the C_{50} of test data. According to Brutti and Matassoni (2016), C_{110} is also effective for such acoustic models selection in large vocabulary continuous speech recognition systems with either HMM-GMM based acoustic models or DNN-HMM hybrid acoustic models. Compared to the multi-condition training, model selection achieves a better balance between the overall robustness and the best performance at each reverberation level.

Since the calculation of ELR is dependent on RIR, research has been conducted on blind reverberation level estimation without RIR. Parada, Sharma and Naylor (2014) applied a classification and regression tree (CART) on features concatenating utterance-level short term features and long term features for a non-intrusive estimation of C_{50}. The short term features include the mean, variance, skewness, and kurtosis of spectrum per utterance, as well as the zero crossing rate, speech variance, pitch period, the importance-weithed SNR (iSNR), variance and dynamic range of Hilbert envelope, *etc.* (Sharma et al., 2010). The long term features include 16 frequency bins of the long term average speech spectrum deviation and unwrapped Hilbert phase (Sharma et al., 2010). Later Parada, Sharma, Lainez, Barreda, Naylor and Waterschoot (2014) used a deep belief network (DBN) initialised with the sparse autoencoder as a regression function to learn the C_{50} from training data, and the DBN based estimator was found to outperform the CART and linear regression based estimators. Parada et al. (2016) further employed bidirectional long short-term memory (BLSTM) recurrent neural network (RNN) to provide a frame level blind estimation of C_{50}, and the performance is found to be better than the DBN based estimator on both the simulated data and real data. Since the estimation result per frame can be quite noisy, an average over the output from multiple frames helps to reduce the estimation error, and an average over 200 frames is found to achieve the same estimation accuracy with using all the data from the same test reverberant condition (Parada et al., 2016).

Chapter 3

Motivation

Contents

3.1 Natural Spontaneous Speech Recordings with Rich Information . . 37

3.2 Real Natural Spontaneous Speech Recognition: from Headset Recordings to Distant Recordings 41

3.3 Reverberation Modelling for Distant Speech Recognition 42

3.4 Signal Aware Reverberation Measurement 44

This chapter explains the motivation of the research work to present in the rest chapters. It starts with Section 3.1 which highlights the lack of natural spontaneous speech recorded with multi-microphones and multi-media to provide rich information for research. Section 3.2 revisits the importance of analysing the ASR performance degradation from using the headset recordings to using the distant recordings of human-to-human natural spontaneous speech. As reverberation is a very challenging factor in DSR, Section 3.3 emphasizes that the negative impact of reverberation on ASR should be quantified with a method that focuses more on the feature pattern distortion than the signal quality degradation. Section 3.4 further suggests an exploration work on estimating the negative impact of reverberation on ASR by taking into account both the room properties and the speech sound properties.

3.1 Natural Spontaneous Speech Recordings with Rich Information

With the progress in speech recognition in recent years, there are more and more commercial applications exploring the potential of ASR as an artificial-intelligent personal assistant, such as Apple Siri, Microsoft Cortana, Google voice, *etc.* Though the commercial products

such as Amazon Echo have proved users' interest in using the distant speech recognition (DSR) to facilitate daily life, for a long time the majority of commercial exploration has been limited to using close-talking recordings due to the limited research progress in DSR. Recently, the recognition performance based on close-talking recording has reached a record high level, as competitive as human performance (Saon et al., 2016). In comparison, the performance gap remains between using close-talking recordings and using distant recordings.

Compared to the speech recognition based on close-talking recordings, the challenges in DSR mainly come from three aspects: reverberation, background noise and overlapped speech. Recent improvement in DSR performance mainly comes from the overall performance boost by the various deep networks and deep structures, while the three main challenges in DSR are still not fully addressed. Therefore there are two major targets for DSR research. The first target is to improve the robustness of speech recognition models on distant recordings, thus shortening the gap of recognition performance between using close-talking recordings and using distant recordings. The second target is to improve the DSR performance on real natural data. The first target has always been the research focus of DSR in the past and current research. In comparison, the second target is only realistic very recently with the record low word error rate (WER) in the close-talking recording based speech recognition. For both targets, research progress has so far benefited and will carry on benefiting from the multi-microphone speech recordings with rich information.

As reviewed in Chapter 2, it has been observed that the multiple distant microphone (MDM) based recordings provide better performance than the single distant microphone (SDM) based recordings in various DSR research areas. For dereverberation, MDM recording has been proved by Miyoshi and Kaneda (1988); Nagata et al. (2004) to have a theoretic advantage than the SDM recording when there is no common zeros among the room impulse responses (RIRs) measured with the recording microphones. One typical multi-channel dereverberation algorithm, the generalized weighted prediction error (GWPE), utilizes an autoregressive model to predict and to remove the late reverberation based on truncated history complex spectrogram in the MDM recordings (Yoshioka and Nakatani, 2012). The dereverberation algorithm can be combined with beamforming algorithms which further improve the DSR robustness with the MDM recordings from three aspects: suppressing the background noise, suppressing part of the reverberation smearing distortion, and enhancing speech signal arriving from the specific direction. When the competing talkers have different geometric locations, e.g. in different angles to the microphones, the MDM recordings make it possible to realise the speaker tracking based on source localization (Marković and Petrović, 2010; Sturim et al., 1997; Valin et al., 2004, 2006; Vermaak and Blake, 2001), which has been shown to be more effective

than speaker tracking using the SDM recordings only. This is because the frequently adopted speaker tracking strategies on the SDM recordings are based on speaker dependent features such as pitch and speaker dependent time-frequency (TF) mask. Such features can also be implemented on the MDM recordings (Pavlidi et al., 2013) jointly with source localization for speaker tracking. With the MDM recordings, speaker tracking could be further combined with beamforming to improve the robustness against speaker movement (Maganti et al., 2007).

Besides the MDM recordings, the rich information collected with multiple microphone arrays and multiple recording media has also been proved beneficial in improving the DSR performance. For example, it has been proved that the speaker tracking performance could be improved with simultaneous recordings from multiple microphone arrays (Ma et al., 2006; Potamitis et al., 2004) and synchronised audio-video recordings (Checka et al., 2004; Gatica-Perez et al., 2007; Maganti et al., 2007; Nakadai et al., 2002; Strobel et al., 2001), particularly when the speakers are moving while talking.

Recent progress in the feed-forward DNN and other novel deep networks suggests that the performance of deep networks in DSR can also benefit from the MDM recordings. In one piece of the author's early PhD work, it is observed that concatenating the log-Mel filter bank features from the multiple channels of MDM recordings as the input to the feed-forward DNN front-end provides similar or even better recognition performance than the wDSB (Liu et al., 2014). Similar observation was reported by Swietojanski et al. (2013) in a hybrid structured DSR system where the DNN serves as the acoustic model and the likelihood is estimated from the DNN output posterior. In a follow-up work by Xiao et al. (2016), it is found that further improvement could be achieved by jointly optimising the beamforming coefficients and the DNN parameters. Furthermore in the research by Sainath et al. (2015), the beamforming and the standard feature pipeline are completely replaced with various deep networks combined to extract the speech pattern information directly from the raw signals. From the early work on multi-channel feature concatenation to the recent work on beamforming with deep networks, there is a large increase in the demand of data, from the scale of 100 hours to 2000 hours, which also brings an increased demand of computation resource. As deep networks are by nature data-driven, it can be expected that further research progress in DSR with deep networks will rely on more MDM recordings.

However there are very limited MDM recordings currently available in the corpora for research, and they are rarely based on the real natural spontaneous speech but mostly based on the artificial scenarios, read speech or re-recorded speech. Among the few available large MDM databases, the AMI corpus (McCowan et al., 2005) and the ICSI corpus (Janin et al., 2003) are mainly based on planned speech. Existing research corpora rarely allow natural speaker movement during recording either. The high cost of accurately annotating

the speaker location with natural speaker movement is one reason for that. In the AMI corpus and the ICSI corpus, the movement of speakers is largely limited due to the planned meeting scenario.

In some research, the speaker moving effect is simulated by convolving the clean headset recordings with the RIRs measured or simulated with different combinations of microphone locations and loudspeaker locations. The popular RIR corpora that can be used for such simulation include the multichannel acoustic reverberation databases at York (MARDY) (Wen et al., 2006), the Aachen impulse response (AIR) (Jeub et al., 2009), the database of the omnidirectional and B-format room impulse response (Stewart and Sandler, 2010) and the ACE Challenge corpus (Eaton et al., 2015).

Some MDM recording databases promote the speaker location change in the room by requesting the speaker to read the prompts at a few given locations or by requesting the speaker to move in a few planned trajectories. One example of such re-read speech corpus is the MC-WSJ-AV corpus (Lincoln et al., 2005) which has been used in the ReverbChallenge 2014 (Delcroix et al., 2014). Examples of such corpora include the DIRHA-GRID (Matassoni et al., 2014) and DIRHA-ENGLISH corpus (Ravanelli et al., 2015), both of which simulate various sound source locations in multiple rooms with RIRs measured at corresponding location. The additional background noise is also added optionally to the simulated data.

To further improve DSR, more research data, especially more MDM recordings of the real natural spontaneous speech, is of an urgent need. As previously mentioned, the MDM recordings have been found beneficial for solving all the three aspects of DSR challenges, namely the reverberation, the background noise, and the overlapped speech. In addition it is crucial to have the MDM recordings accompanied with the individual headset microphone (IHM) recordings. As shown by existing research, to provide rich information it would be even better if the MDM recordings are also accompanied with the video recordings, multiple microphone array audio recordings and the ground-truth speaker location tracking. However, there are very limited research corpora that provide the real natural speech recordings with all these information. In a natural set-up of human to human conversation, the speakers could move around in the room which causes significant change in room impulse response (RIR) and reverberation, which dramatically increases the difficulty in speaker tracking, dereverberation and beamforming.

Recently, the research effort has been devoted to preparing data for a shared platform to validate and to compare different algorithms, not only on simulated data but also on the real multi-channel recordings with daily life environment noise. Two examples are the ReverbChallenge 2014 (Kinoshita et al., 2016) and the third CHiME challenge (Barker et al., 2015). In the ReverbChallenge 2014, algorithms and systems are validated on both

the simulated data and the real recordings, in an effort to bridge the DSR performance gap between using the simulated data and using the real data. The third CHiME challenge provided the real recordings in a few typical noisy environments in daily life, and the recording is performed with both the headset microphone and the multiple distant microphones, in an effort to shorten the DSR performance gap on the research data and on the distant recordings acquired in a real application.

Such databases still fall short in two aspects to represent the real natural human-to-human spontaneous conversations well. First, it rarely includes a natural speaker movement even in a domestic environment, because the post-recording annotation of speaker location is very costly. Second, the existing data is rarely composed of the real natural spontaneous conversations. This is partly because of the privacy concern in the data collection and partly because for a long time the average DSR performance is not good enough to polish any research algorithms on the over-challenging data such as the recordings of real human-to-human conversations. Recently, with the recognition performance improvements brought by deep networks, more research interest has been shifted to improving DSR performance on real and natural data (Barker et al., 2015).

Therefore, the first major contribution of the work to be presented in this thesis is about collecting the multi-microphone and multi-media real recordings of natural spontaneous multi-party conversations among native English speakers. The recording of the Sheffield Wargame Corpora (Chapter 4) is performed in three days in total. The research work is conducted with the help from Dr. Charles Fox and Dr. Madina Hasan. Relevant research work has been previously published in Interspeech 2013 (Fox et al., 2013) and Interspeech 2016 (Liu et al., 2016). Dr. Fox has a major contribution in both the hardware and software design of the recording system. Dr. Hasan prepared the language model for the second data release in Interspeech 2016.

3.2 Real Natural Spontaneous Speech Recognition: from Headset Recordings to Distant Recordings

One obvious consequence caused by the lack of the real natural spontaneous speech recordings is a limited understanding of the weaknesses in the existing DSR systems on the real natural spontaneous conversations, as well as a good strategy to treat the multiple weaknesses in practice. With the fast progress in ASR performance on close-talking recordings thanks to various deep networks and deep structures, it is now of crucial importance and current interest to gain a better understanding of the weaknesses in the state-of-the-art DSR systems when they are applied to the real natural spontaneous speech

recordings. In the long term, this knowledge is very important for reducing the gap between the human performance and machine learning performance in DSR.

Therefore the second major contribution of the research work to be presented in this thesis is a thorough analysis of the weaknesses in the DSR of real natural spontaneous multi-party conversation. This is only made possible with the collected real recordings of natural spontaneous multi-party native English speech, namely the SWC data. The analysis covers the techniques based on both the single microphone recordings and multiple microphone recordings. In addition, the analysis is not only performed on DSR, but also on speech recognition of headset recordings in the context of real natural spontaneous multi-party conversation, to bridge the gap between recognition performance using distant recordings and headset recordings. In addition, the distributed microphones and microphone arrays at different locations in the room are compared and analysed in terms of recognition performance with and without multi-channel based enhancement. Details will be covered in Chapter 5.

3.3 Reverberation Modelling for Distant Speech Recognition

Reverberation is one key factor in the distant speech recordings that impacts the recognition performance. Therefore research has been conducted on modelling the distortions caused by reverberation. Different from the background noise which is additive to speech in the time domain, the reverberation effect is convolutional to the speech signal. Thus the reverberation effect is frequently approximated with a linear FIR filter in time which is usually referred to the room impulse response (RIR). This RIR based reverberation approximation at signal level is a widely adopted reverberation model for its high accuracy in reconstructing the reverberant signal.

The RIR is very sensitive to any changes in the room arrangement, microphone location and speaker movement. Therefore further research effort has been devoted to the RIR modelling. The early work on RIR modelling is conducted based on room acoustics. The image source method (Kuttruff, 2009) is such an example. It simulates the RIRs in an empty room by replicating the sound reflection process in a simplified way, given the room geometry information and the average acoustic absorption coefficients of the acoustic boundaries. The image source method is good for approximating the reverberation effect of shoe-box shaped rooms, however it cannot simulate any spatial asymmetric effects caused by the furniture arrangement, nor is it a good option for rooms of irregular shapes. This is because the increase in the number and the geometric complexity of acoustic reflection

surfaces will dramatically increase the complexity and the computation cost of the image source model. Therefore the RIRs simulated with the image source method are very different from those measured in the real occupied rooms. Due to the simplification in the room configuration, the simulated RIRs tend to be over-simplified compared to measured RIRs, particularly regarding the colouration of the early reflections, the impact of the talking direction and the microphone directivity. Later Polack (1988) proposed to model RIRs by its statistic properties. This method is further extended by Doire et al. (2015) using an exponentially decaying statistic process to approximate the early reflections, or early reverberation, along with a smooth transition between the modelling of early reflections and late reflections in RIR. In this way the statistic methods reflect the reverberation level and the actual RIR coefficients are generated by a stochastic process. As a consequence the RIRs modelled statistically can only reflect some properties of the target reverberant environment while it cannot accurately reconstruct the signal or the feature patterns of reverberant recordings. The RIRs constructed with RIR modelling methods can hardly be used directly for reverberation treatment, mostly because of their difference to the RIRs measured in real environments. Instead, the RIRs can be used to simulate diverse data for the multi-condition training of the DNN acoustic model or the DNN front-end to increase the reverberation robustness.

The RIR parameters are usually optimised to reconstruct the reverberant speech signal rather than the patterns in reverberant speech features. Compared to the patterns in the speech features, the speech signal is more sensitive to the acoustic change in the reverberant environment and in the recording channel. Therefore the RIRs are also more sensitive to any acoustic change than the patterns in speech features. This is because some signal level variation will be normalised during feature generation. Therefore further research on reverberation modelling has been devoted to directly model the impact of reverberation on speech recognition features. Sehr and Kellermann (2008, 2009) modelled reverberation as a linear distortion on Melspec features. In that work, the Melspec features of reverberant signal are approximated with a convolution of an FIR filter and the Melspec features from the corresponding headset recordings, and the convolution is performed independently in each Mel frequency band. This reverberation modelling is not accurate analytically. As shown in Fig. 1 from Sehr and Kellermann (2008) and Fig. 1 from Sehr and Kellermann (2009), such a reverberation model leads to an over-simplified feature pattern structure.

Theoretically there are three advantages in the feature level reverberation modelling compared to the signal level reverberation modelling. First, the parameter dimension is reduced. Second, the parameter value can be less sensitive to the acoustic variation. Third, the modelling performed on speech features can be more aligned with the recognition tasks. However the existing feature level reverberation modelling could not accurately

construct the feature patterns of reverberant recordings. In addition, there is very limited research on the temporal variation of reverberation caused by the acoustic environment change which is one important source of reverberation modelling errors in real applications. As a result, there is very limited advantages of the existing feature level reverberation modelling compared to the signal level reverberation modelling, in terms of improving DSR performance.

The research work to be presented in Chapter 6 will focus on improving the reverberation modelling accuracy for speech recognition tasks. To achieve that, the impact of reverberation on the speech complex spectrogram is first investigated for a better understanding of the reverberation distortions in spectrogram which further leads to the feature smearing. Based on the analysis, the reverberation modelling is proposed in the complex spectrogram domain and it is evaluated in speech recognition tasks. In addition, the reverberation modelling parameters are analysed regarding the reverberation variation caused by speaker movements and the change in microphone installation.

3.4 Signal Aware Reverberation Measurement

The DNNs and the relevant variants of neural networks are widely employed in state-of-the-art speech recognition systems. They share a common property that the topological structure of neurons inside the deep networks is highly symmetric and replicative. This provides a lot of flexibility to the overall deep networks in adapting to different tasks and input features. However one drawback of such data-driven models is that its dependence on training data leads to an increased sensitivity to the training-test mismatch, particularly when the test data has some properties not well covered by the training data. In DSR applications it is very likely that the test data has some environment properties different from training data in terms of both reverberation and background noise. As reviewed in Section 2.3, progress has been made in improving DNN robustness against the background noise via noise aware training (Seltzer et al., 2013). In comparison there is very limited progress in improving the robustness of DNN against reverberation and reverberation variation in real data. Instead, there is increased interest in using diverse data to improve the robustness of DNNs via multi-condition training (Barker et al., 2015). However the multi-condition training is not perfect and it improves the overall robustness at a price of degrading the performance on relatively clean data and the performance in conditions with little training data. To address this problem in multi-condition training, research has been conducted on the data selection at training stage and the model selection at test stage according to the reverberation level of the speech recordings (Brutti and Matassoni,

2014, 2016). For this purpose, reverberation measurement becomes critical in avoiding the training test mismatch for DNNs.

Brutti and Matassoni (2014, 2016) have shown that the model selection and the model combination based on an estimated reverberation level improve the recognition performance in the adverse conditions without compromising the recognition performance in the other conditions. However Parada et al. (2016) pointed out that there is large variation in reverberation level estimation results for short recordings, particularly those with less than 0.5 seconds of speech. Parada et al. (2016) proposed to improve the reliability of reverberation level estimation by averaging the results over more data. However this strategy will inevitably increase system latency in applications.

"Reverberation measurement", in this work, refers to estimating the level of reverberation distortion in the speech patterns. In contrast, in most of the literature the reverberation measurement refers to estimating the reverberation level of the reverberant environment and the distant recording channel. Existing methods for the reverberation measurement are rarely designed for pattern recognition task in the first place. Instead, most existing methods have been developed directly or indirectly from the reverberation metrics that reflect the behaviour of sound energy decay from an acoustic point of view. The reverberation time T_{60} and the speech clarity C_{50} are two such reverberation metrics that are frequently used.

In addition, the existing reverberation measurement targets on the reverberation level of the environment and the distant recording channel. While T_{60} and C_{50} can be estimated from the RIR, research has also been conducted on the non-intrusive reverberation level estimation. In comparison with the intrusive reverberation level estimation, the non-intrusive methods do not rely on RIR. Since the human perception of the reverberation level in a small piece of speech recording is quite inaccurate, the human annotation on the reverberation level varies a lot by individual difference and recording content. Therefore the non-intrusive methods usually employ the RIR based objective reverberation metrics as a reference during the parameter optimisation using the simulated reverberant speech data. As a consequence, the non-intrusive methods inherit one property from the acoustics based reverberation metrics, *i.e.* the measurement results tend to reflect the reverberation level of the environment. Therefore, the existing reverberation measurement hardly aims at estimating the reverberation distortion level in the speech feature pattern.

As mentioned, existing methods have been focusing on estimating the reverberation level rather than the reverberation distortion level, as the reverberation level is only environment dependent while the reverberation distortion level is both environment dependent and signal dependent (Kokkinakis and Loizou, 2011). The difference between the reverberation level and the reverberation distortion level is not obvious when the measurement is con-

ducted on a large amount of data. However there are some applications that favour short utterances for low latency reverberation measurement based model selection at utterance level. In such applications the difference between reverberation level and reverberation distortion level can be amplified by the limited types of sound patterns in the very short speech recordings. Such short recordings can be one phrase of less than 2 seconds or one speech utterance of less than 4 seconds. As pointed out by Assmann and Summerfield (2004), different speech sounds have different levels of robustness against the reverberation distortion and the background noise, mainly due to the phonetic pattern structure. Such signal dependent differences could cause a mismatch between the reverberation level and the reverberation distortion level, which is not considered in any all signal independent reverberation metrics.

For example, Assmann and Summerfield (2004) pointed out that the stop consonants are brief in duration and low in intensity, making them particularly susceptible to masking by noise and temporal smearing via reverberation compared to vowels. From the pattern recognition point of view, when the speech signal is distorted by reverberation during recording, the stop consonants have a higher level of reverberation distortion than the vowels. Such slight difference between reverberation level and reverberation distortion level can be amplified by the limited sound pattern types in very short speech recordings. However given the same RIR, signal independent reverberation metrics would suggest that the same reverberation level regardless of signal properties. As a result, the signal independent reverberation metrics and the non-intrusive reverberation level estimators could not provide a reliable estimation of the reverberation distortion level in the speech sound patterns. In speech recognition tasks, compared to the training data selection and the acoustic model selection based on reverberation level, a better selection strategy is based on the distortion level of speech sound patterns. Therefore the existing methods estimating the reverberation level could be suboptimal for the data and model selection, particularly when the selection is performed based on the reverberation level estimated over short utterances.

There is a further issue in the existing reverberation level estimators, as most of them are based on the early-to-late reverberation ratio (ELR): the dispute of the optimal boundary between early and late reverberation. So far the partition boundary has been empirically determined via experiments, and different optimal partition boundaries have been concluded from the experimental research work conducted on different data and in different tasks. Parada, Sharma and Naylor (2014) achieved the best correlation between the ELR and the phoneme recognition accuracy, as well as the best correlation between the the ELR and speech quality score in perceptual evaluation of speech quality (PESQ), using 50 ms as the boundary in ELR calculation, or C_{50}. This finding led to further work

on non-intrinsive estimation of reverberation score based on C_{50} by Parada et al. (2016). In comparison, Brutti and Matassoni (2016) later observed that using 110 ms as the boundary provides a higher correlation between the ELR and the word recognition accuracy, namely C_{110}. The potential difference in data and algorithms is not discussed by the mentioned work, and there is not theoretical support for the optimal boundary in partitioning the early and late reverberation in the context of speech recognition. In many applications, 50 ms is still widely adopted as the boundary between the early and late reverberation mostly out of empirical reasons (Bradley, 2011; Xiong et al., 2014; Yoshioka et al., 2012).

The research work to be presented in Chapter 7 covers the very first research effort on taking the signal properties into account when estimating the level of reverberation distortion on speech sound patterns in DSR tasks. Based on the analytic work of reverberation modelling in Chapter 6, the proposed novel method adopts a polynomial structure which is free from the early-late reverberation partition issue. In addition, inspired by the phonetic analysis on reverberation distortion by Assmann and Summerfield (2004), a novel reverberation partition based on the distortions is explored regarding its potential in improving the estimation accuracy of reverberation distortion level.

Chapter 4

The Sheffield Wargame Corpora

Contents

4.1 Data Collection Design and Recording Setup 50

4.2 Data Statistics, Annotation and Transcribing 55

4.3 Blog Data and Language Model 59

4.4 Baseline Systems 62

4.4.1 Task and dataset 62

4.4.2 Baseline adaptation system 64

4.4.3 Baseline standalone system 66

4.4.4 More beamforming and dereverberation 68

4.5 Summary 69

Distant speech recognition (DSR) has gained a wide research interest and industrial demand for its essential role in providing a natural interface for the advanced technology in the artificial intelligence and machine learning. Powered by big data and the computation speed-up from the graphic processing unit (GPU) in recent years, the DNNs have contributed significantly to the performance improvement of automatic speech recognition (ASR) in various configurations and tasks. While the overall ASR performance is still improving, the WER gap between using close-talking recordings and using distant recordings still remains. One important reason that hinders the research progress in DSR is the amount of available research data, namely the distant speech recordings, particularly the multiple distant microphone (MDM) based speech recordings accompanied with close-talking recordings. Existing research has widely observed that using the MDM recordings

in DSR outperforms using the single distant microphone (SDM) recordings. However the available real MDM recordings of natural speech for research is of a much smaller amount compared to the SDM recordings.

This chapter covers one major contribution of the author's work on the Sheffield Wargame Corpora (SWC), a database of real natural spontaneous conversational speech recorded with multiple distant microphones from multiple microphone arrays, accompanied with the headset speech recordings, the video recording and the speaker location tracking. The work on SWC recording is presented from four aspects: data collection design and recording configuration (Section 4.1), the statistic analysis of the data and data transcription (Section 4.2), the specially prepared text data for language model (LM) training (Section 4.3) and the baseline speech recognition performance using the state-of-the-art ASR architectures (Section 4.4).

4.1 Data Collection Design and Recording Setup

The ASR performance on real natural human-to-human conversational speech is always of wide interest. While the research progress with the deep network based acoustic model is highly dependent on the amount of available data, the existing recordings of real natural human-to-human conversational speech are very limited for open research. A major issue with collecting naturally occurring spontaneous conversational speech data is the concern on privacy. The cost of transcribing and annotation can be high, but the privacy involved inevitably in the natural conversational discussions is even more difficult to address. As a result, most speech data collection is based on the rehearsal speech or the read speech (Barker et al., 2015; Lincoln et al., 2005). In the SWC recording, the privacy infringement is minimised by setting the recording in a table-top strategic game: the Warhammer 40K (Fig. 4.1).

Even though table-top game is a rare topic for speech corpus collection, it has a few advantages besides the privacy protection. First, it involves a heated discussion among the members of one team about the strategies to beat the other team. Second, the game is played by participants moving figurines around on the table based on the rules and the dice-throwing results. As a consequence, there is a continuous body movement while the players are speaking naturally and spontaneously. This can represent a typical scenario of the domestic multi-party social conversation. Third, it is a topic sufficiently interesting to the recording participants and they could keep playing and producing the speech data for as long as 10 hours non-stop. With this topic, it is possible to collect a large amount of speech data very quickly. Fourth, this game has both the online version and the table-top version,

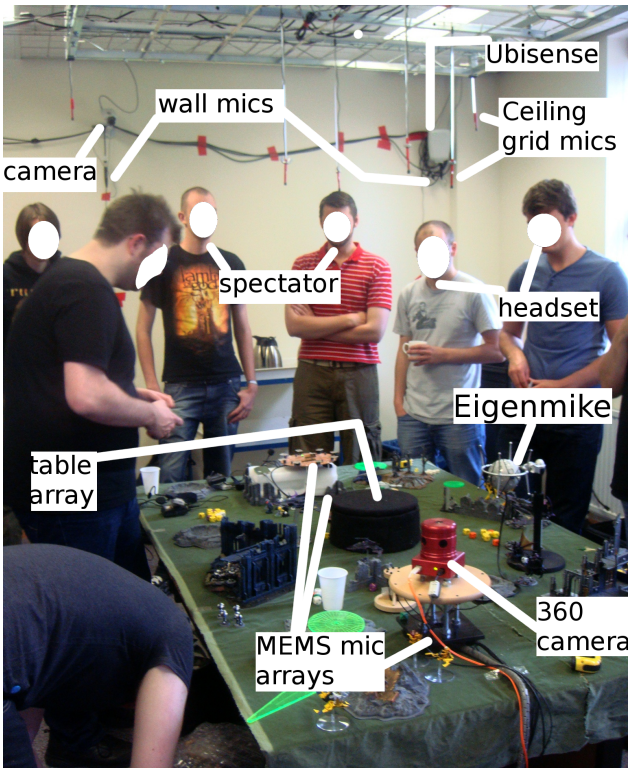


Fig. 4.1 *SWC1 recording.*

and the recording participants are used to both of them. The online version requires players to wear headsets. Therefore the participants are very used to speaking naturally with the headsets during recording. Unlike the rehearsal speech or the read speech where the speakers are subject to nervousness thus producing un-natural speech, this game ensures speech naturalness in recording. Therefore, the table-top game Warhammer 40K is chosen as the recording topic after comparing with other topics and evaluating many aspects.

The recording is conducted in three days in total. The first day recording (SWC1) was performed in 2012, the second day recording was in 2014 (SWC2) and the third day recording was in 2015 (SWC3). The recordings of the first two days contain native English speech from male speakers only, mainly due to the fact that most Warhamer 40K players are male. To address the gender bias issue, the third day recording is set up to be a tutorial scenario, where female players are trained to play the game under the guidance from one experienced player who is the tutor of each team. Therefore, each game of the first two day recordings, *i.e.* SWC1 and SWC2, has four male participants from two teams. In SWC3, there are six participants for each game in two teams, each team having one male tutor and two female players. Each game lasts for 1-2 hours. For long games, the recording pauses in the middle of the game to avoid overloading the recording hardware and software. Such

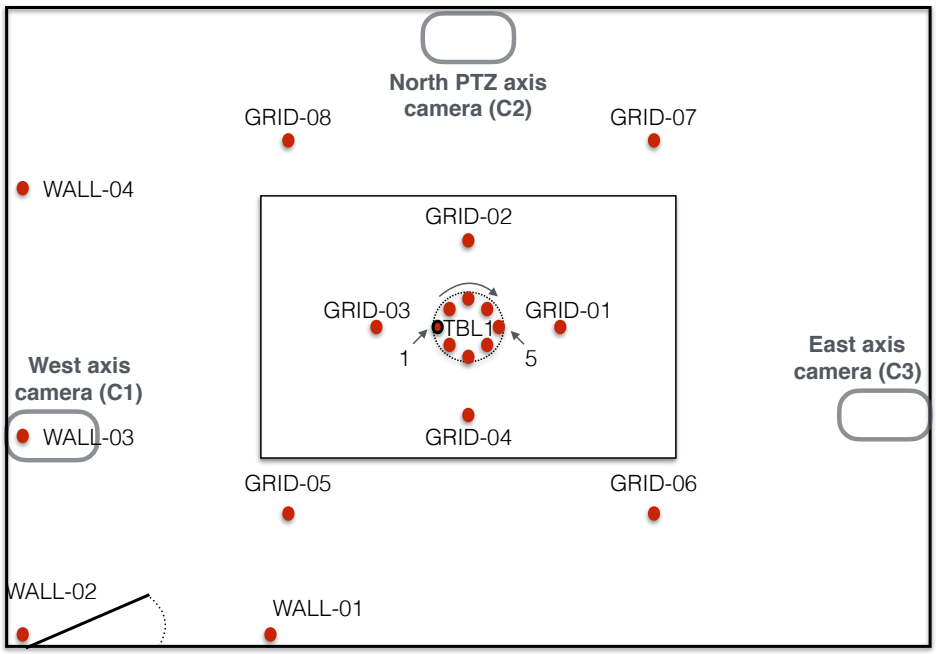


Fig. 4.2 *SWC - 20 shared distant microphones among three recording days (top-down view).*

a continuous recording is further referred to as a "session", and one game has one or two recording sessions.

At the beginning of each recording session, a clap board is used to provide a manual synchronisation for the audio recordings from multiple microphone arrays and the video recordings from multiple cameras. The participants are requested to stand at four corners of the table and provide one brief description in turn about their headset number, their dress and their roles in the game (Fig. 4.1). This helps to correct the unexpected problems in synchronising multi-channel and multi-media recordings in the post-recording data processing. Most sessions in SWC1 and SWC2 only have the four players in the recording area and each of them wears a headset microphone for close-talking speech recording. A few sessions have some invited viewers in the recording area, to watch the game and to interact with the players (Fig. 4.1). There is no close-talking speech recordings of such viewers. In SWC3, besides four female players all sessions also have two male tutors in the recording area. There is no close-talking speech recordings of the two extra male speakers either, thus their speech is not annotated or transcribed.

The recording system is composed of three parts: the audio recording, the video recording and the speaker location tracking. For audio recording, there are 24 microphones shared among the recordings of all three days - 4 headset microphones and 20 distant microphones. The geometry of the 20 distant microphones is shown in Fig. 4.2. There are

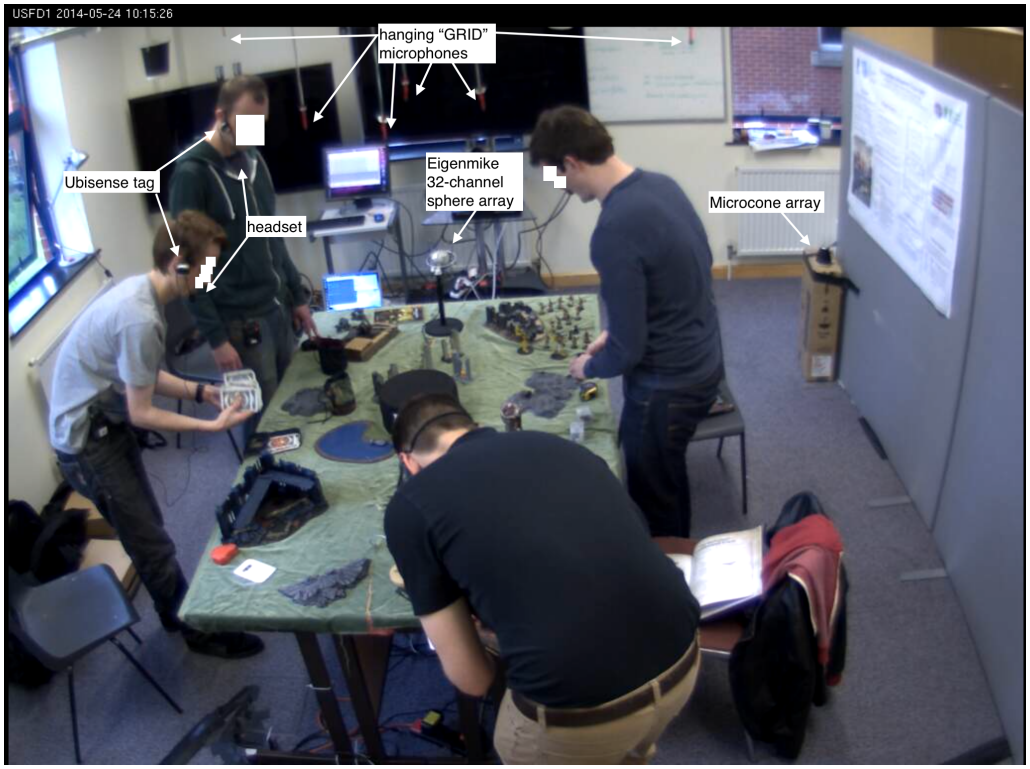


Fig. 4.3 *SWC2 recording.*

8 microphones placed on the table forming a circular array ("TBL1-*"), 8 microphones hanging from the grid forming a group of distributed microphones ("GRID-*"), and 4 microphones distributed in the room and adhered to the wall ("WALL-*"). All the 20 distant microphones are hyper-cardioid AKG C417/III vocal condenser microphones, and the headset microphones are all wireless with Sennheiser EW100 microphones of cardioid directivity. Overall the 24 microphone channels are sample synchronised using an all-Linux setup. The audio recording sampling rate is 48 kHz, and the 16 bit A/D conversion is realised by the MOTO 8 Pre's, linked by firewire 400 and FFADO drivers to the JACK middleware on a Ubuntu Studio desktop PC, streaming audio data direct to hard disc. Full details of the sample-synchronous recording setup can be found in the publication by Fox et al. (2012).

In SWC1 and SWC2, there are other distant microphone arrays but they are not equipped in SWC3. In SWC1 and SWC2, the audio recordings from such microphones are manually synchronised with above 24 shared microphone channels in the post-recording data processing. In SWC1, there is an omnidirectional 32-channel Eigenmike sphere array (diameter: 8.4cm), five 8-channel microphone circular arrays using analogue and digital MEMs microphones, all placed on the table along the central line. Among the five MEMs microphone arrays, there are two arrays of analogue MEMs microphones with a diameter

of 20 cm and 4 cm respectively, and three arrays of digital MEMs microphones, two of which have a diameter of 4 cm and the third has a diameter of 20 cm. In SWC2, extra audio recordings are available from an Eigenmike array and a Microcone array which is a circular digital MEMs microphone array with a diameter of 4 cm. The MEMs microphone array and the Eigenmike array are shared between SWC1 and SWC2. Only part of Session 1 in SWC2 has Eigenmike recordings available due to a software failure. The Microcone array has 6 microphones in a circular array (diameter: 8 cm), with the seventh microphone pointing right up to the ceiling. The MEMs microphone array is situated on the table while Microcone array is located next to a boundary panel at recording area, as shown in Fig. 4.3.

For video recordings, four cameras have been used in total. In SWC1, there are two cameras hanging overhead (C1 and C2 in Fig. 4.2) and a 360 degree panoramic PointGrey Ladybug2 camera on the table. In SWC2, the 360 degree panoramic camera is replaced with another CCTV camera hanging overhead (C3 in Fig. 4.2). Unfortunately there is no video recordings usable for SWC3 due to an unexpected technical problem that corrupted the recording data.

The speaker location tracking is realised with the Ubisense system. The system is composed of three parts: the tags, the sensors and the server. Four tags are adhered to the moving objects, *i.e.* the four headsets mounted on the head of the four players in SWC recording (Fig. 4.3). Four sensors are installed at the four corners of the recording area ceiling. The sensors communicate with the tags via radio signal. The radio signal is first sent by the sensor to the tag, and then the tag sent a unique feedback radio signal to the sensor. Both the sequences of the original radio signal and the feedback radio signal are recorded and processed by a computer server. A supervised calibration is conducted before recording where the Ubisense tags are placed at a few locations with known coordinates in the room. With a proper calibration, it is possible for the server to accurately estimate the 3D coordinates of each tag via the time difference of arrival (TDOA) of the feedback radio signal. This method is more robust than the speaker localization via multi-channel audio recordings, as there are fewer reflecting sources of radio signal in the recording area compared to the reflecting objects and surfaces for sound signal. In this way the Ubisense system could track the real time speaker location in an independent system simultaneously with the audio and video recording.

Table 4.1 summarizes the recording set-up. As discussed above, the SWC configuration provides a unique platform to collect the natural native English speech which features multiple aspects of challenges in DSR: the room reverberation, the naturally occurring background noise, the overlapped speech, the speaker movement while talking and the natural spontaneous speech with emotional speech and whisper speech.

Table 4.1 *SWC statistics of transcribed recordings.*

	SWC1	SWC2	SWC3	overall
#session	10	8	6	24
#game	4	4	3	11
#unique speaker	9	11	8	22
gender	M	M	F&M	F&M
#unique mic	96	71	24	103
#shared mic	-	-	-	24
video	√	√	-	√
location	√	√	√	√

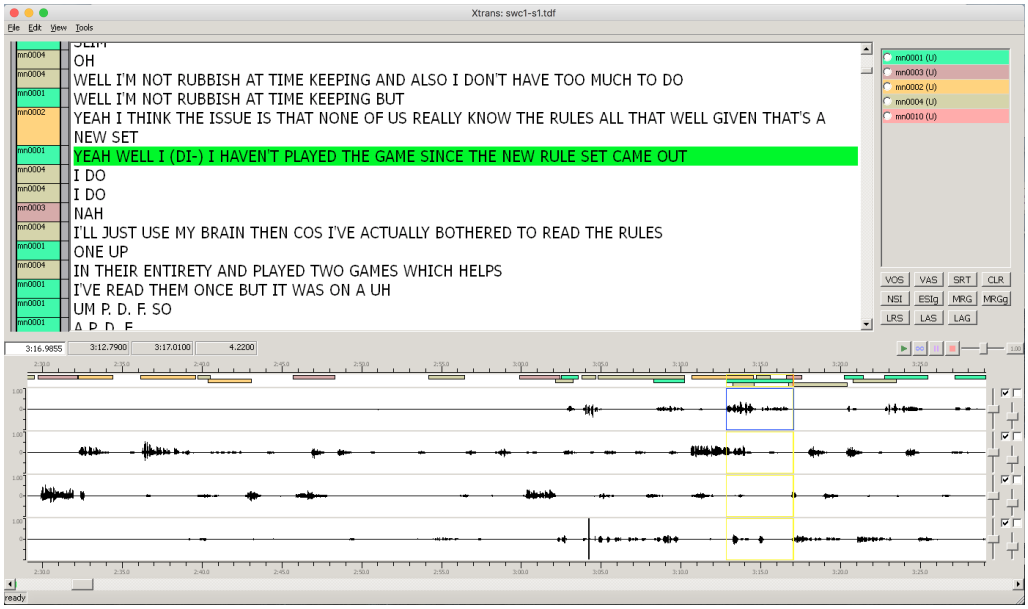


Fig. 4.4 *Transcribing SWC recordings with multi-channel audio using XTrans.*

4.2 Data Statistics, Annotation and Transcribing

The recordings are manually annotated and transcribed with the multi-channel transcribing tool XTrans[1]. This tool ensures the transcription quality with its easy access to the conversation context in the recordings of multiple synchronised headset microphones. A screen shot of using XTrans to transcribe the SWC recordings is shown in Fig. 4.4. The utterances are segmented to minimize the within utterance silence, without breaking the semantic completeness of an utterance based on human perception. The beginning and ending silence of an utterance are also minimized. The XTrans interface provides a nice visual and sound combination, so that the transcribers can accurately determine the time boundary of a speech sound.

[1] https://www.ldc.upenn.edu/language-resources/tools/xtrans

Table 4.2 *SWC statistics.*

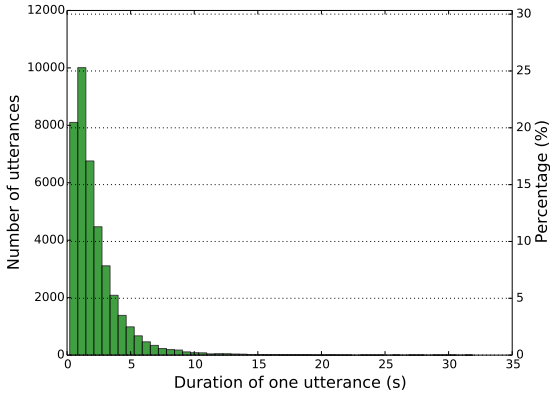
	SWC1	SWC2	SWC3	overall
transcribed speech	8.0h	10.5h	6.1h	24.6h
#speech utt.	14.0k	15.4k	10.2k	39.6k
average duration per utt.	2.1s	2.5s	2.2s	2.2s
#word per utt.	6.6	7.9	5.5	6.8
vocabulary	4.4k	5.7k	2.9k	8.5k

As mentioned in the previous section, the game topic used in SWC recording highly encourages discussions. After cleaning up, in total there are 24.6 hours transcribed speech from the raw recordings conducted in three days. Fig. 4.5c shows the amount of data per speaker. Due to the set-up, there is more male speech data than female speech data. For the same gender, there is a variation in the amount of data per speaker because some speakers participated in more recording sessions than others. For example this is the case with the female speaker "fn0017" who participated in all the recoding sessions in SWC3 due to a lack of further volunteers. More statistics are illustrated in Table 4.2 and Fig. 4.5 in details.

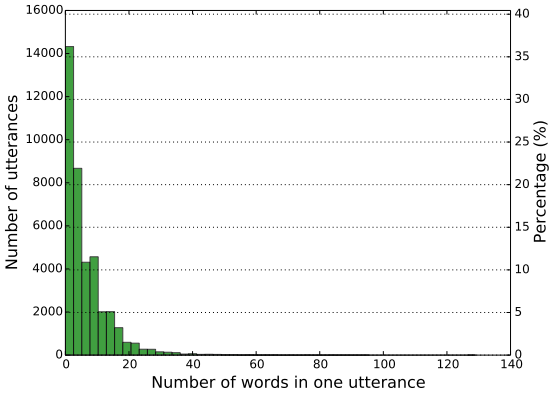
As shown in Table 4.2, on average the speech utterances are as short as 2.2 seconds, and the average number of words per utterance is 6.8. According to the histogram of the utterance duration and the number of words in each utterance shown in Fig. 4.5a and Fig. 4.5b respectively, the majority of the speech utterances last less than 5 seconds and have less than 15 words. In addition, Table 4.2 suggests that the vocabulary of SWC2 is much larger than the vocabulary of SWC3. This is because the players in SWC2 are more experienced in the game than the players in SWC3, thus providing more diverse terms in discussion.

The game topic used in SWC recording also encourages natural spontaneous speech where a large amount of overlapped speech takes place naturally. This is illustrated in Fig. 4.5d, Fig. 4.5e and Fig. 4.5f. As shown in Fig. 4.5f, around 50% of speech utterances have overlapped with competing speech utterances from a different speaker. Fig. 4.5f suggests that among the utterances with overlapped speech, the percentage of overlap within one utterance distributes approximately uniformly, with a slightly higher probability of one utterance having less than 50% being overlapped than of one utterance having more than 50% being overlapped. As shown in Fig. 4.5d, most of the utterances that are overlapped with other utterances have no more than 3 competing utterances. There are very few long utterances overlapped with more than 4 competing utterances.

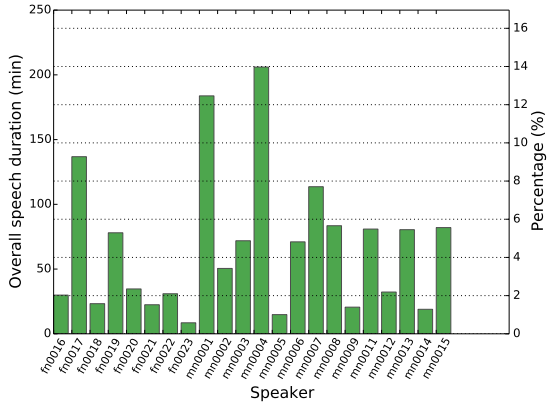
With the Ubisense system, the three dimensional (3D) coordinates of speakers' head location are recorded as well. The definition of the 3D coordinate system is shown in Fig. 4.6a. The corner to the back of the door is used as the origin in the 3D coordinate system.



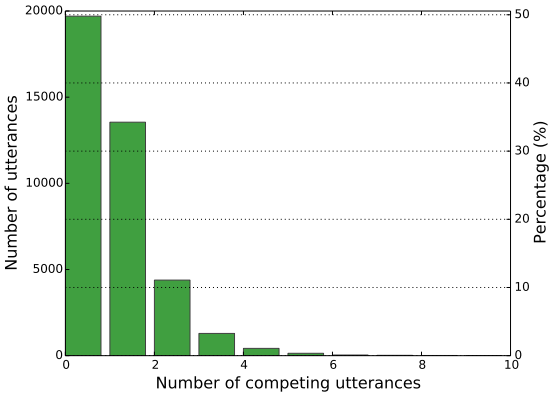
(a) *The duration of speech utterances.*



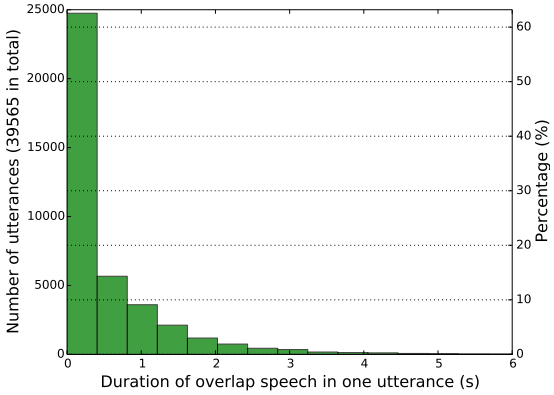
(b) *The number of words in each utterance.*



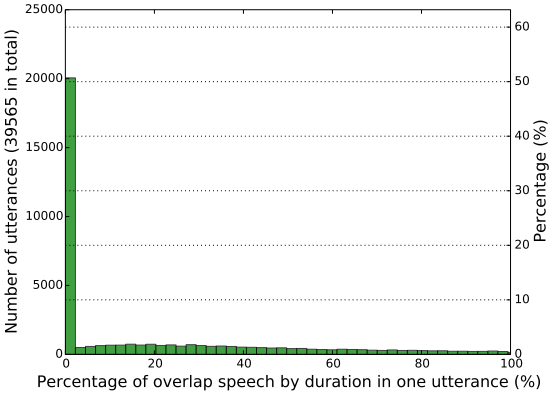
(c) *The speech duration per speaker, "m*" for male, "f*" for female.*



(d) *The number of competing utterances.*

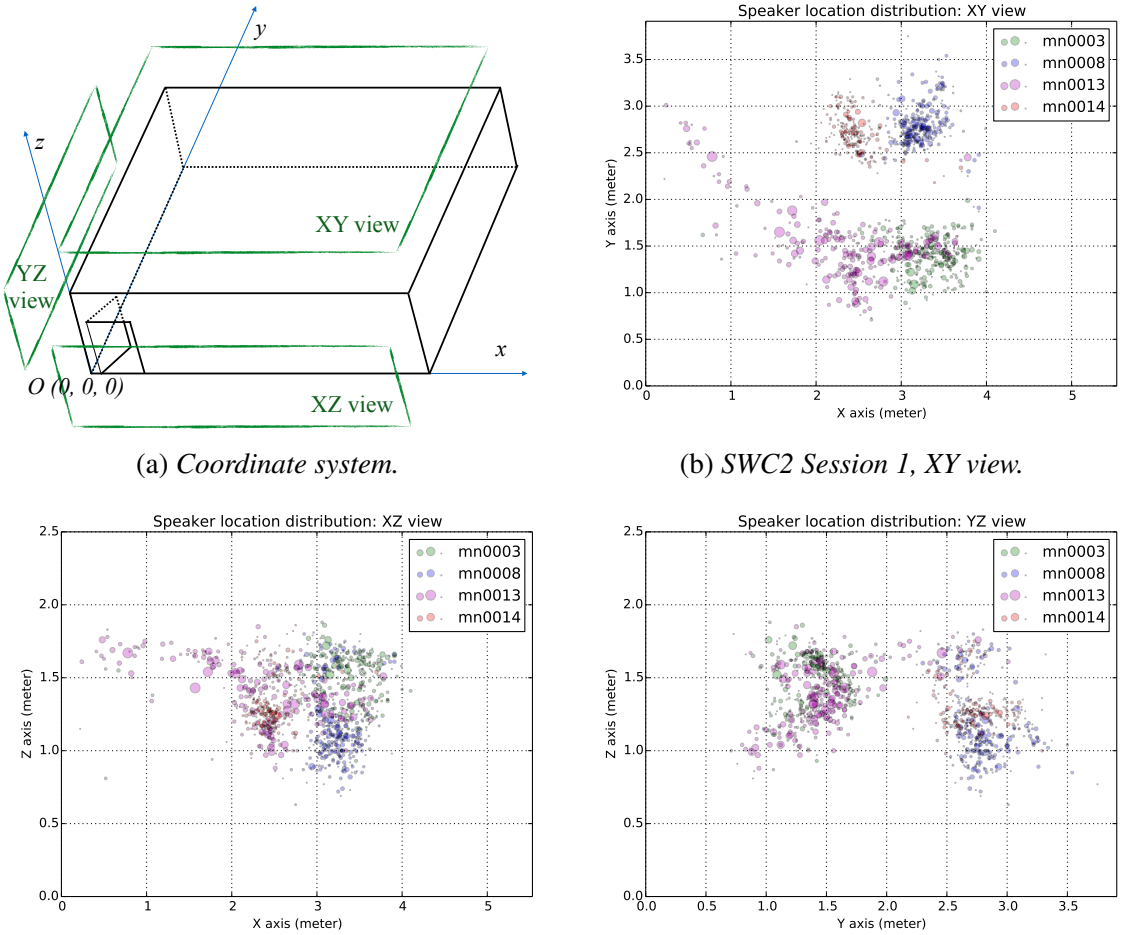


(e) *The duration of overlapped speech in each utterance.*



(f) *The percentage of overlapped speech in each utterance by duration.*

Fig. 4.5 *SWC statistic analysis on the histogram of the speech utterance duration, the number of words in each utterance, the speech duration per speaker, the number of competing utterances, the duration of overlapped speech and the percentage of the overlapped speech in each utterance.*



(a) *Coordinate system.*

(b) *SWC2 Session 1, XY view.*

(c) *SWC2 Session 1, XZ view.*

(d) *SWC2 Session 1, YZ view.*

Fig. 4.6 *Speaker head location tracking based on the Ubisense system: the coordinate system and the tracked speaker head locations in the first recording session of SWC2.*

Fig. 4.6b, Fig. 4.6c and Fig. 4.6d illustrates the speaker location distribution in the first recording session in SWC2. Each circle represents one speech utterance. The size of the circle is proportional to the number of words spoken in that utterance, and the center of the circle is the average speaker head location throughout that utterance. Compare Fig. 4.6b with Fig. 4.2, it can be observed that the players mostly move around the table. This session is special because it is the beginning of a day's gaming recording, hence the players need to take out the figurines from the bag on the ground and organize the figurines on the table. Since each player has one home battle field being one corner of the table, this recording session involves active speaker movement more vertically than horizontally. Fig. 4.6c and Fig. 4.6d illustrates the vertical head movement up to 1 meter. In comparison, Fig. 4.6b shows that there are three speakers having limited horizontal movement, while

the fourth speaker, “mn0013”, frequently moves between one wall and one corner of the table. This is because his bag is placed farthest to the table.

Another property of the SWC data due to its high level of naturalness is emotional speech. This is reflected by the proportion of utterances with laugher. In the raw manual annotation for SWC2 where emotion tags are kept along with the speech words, there are 8.1% of utterances with laugh. In SWC3 there are 25.1% of utterances with laugh. There is no emotion annotation for SWC1 unfortunately due to limited time and budget.

In summary, three recordings of SWC data collected 24.6 hours natural spontaneous English speech data from 14 males and 8 females, all being native English speakers. Having multi-microphone and multi-media simultaneous recordings, the SWC data has a few special properties that are representative in the natural spontaneous conversations but are uncommon in the rehearsal or read speech, such as very short utterances with a small number of words per utterance, a high proportion of overlapped speech, the natural and frequent body movement while speaking, the emotional speech and the vocal sound.

4.3 Blog Data and Language Model

The SWC recording is based on a special game topic for reasons discussed previously in Section 4.1. The acoustic challenges in SWC data are of the key research interest, as the unique properties help to push forward the ASR research, particularly DSR research, for the applications in the natural human-machine communication in a non-intrusive configuration using distant microphones. However since it is based on a topic rarely adopted by any existing speech research corpora, there is some concern about the language modelling (LM). Applying any LMs trained with other research data directly on the SWC data would introduce a large negative bias on the recognition performance due to the mismatch in vocabulary, topic and the text style. Therefore the SWC data requires an in-domain LM that better matches the properties of the SWC transcripts. Such a LM needs to be representative of both the spontaneous conversation and the game Warhammer 40K. Regarding the conversational speech aspect, Hain et al. (2007) have collected the text data for conversational speech in the work on meeting recognition system. However there is no existing text data that features the Warhammer 40K. To alleviate this problem, blog data is prepared to train an in-domain LM for the SWC data.

The blog data refers to a collection of the text from blogs of the Warhammer 40K players. Such blogs cover the reports of games the author participated in, the introduction to the figurines of Warhammer 40K characters the author created and the legend stories about the characters in the game. Though a forum discussion is better than a blog in terms

Table 4.3 *Special words better covered by blog data - an investigation on the word occurrence percentage (%) in different text data components.*

Word & phrase	SWC overall	Blog data overall	Warhammer wikipedia	Conversational web data
shoot	0.30	0.02	0.00	0.00
cover	0.23	0.07	0.04	0.01
roll	0.15	0.04	0.03	0.01
turn	0.15	0.14	0.03	0.02
squad	0.11	0.32	0.02	0.00
plasma	0.08	0.05	0.00	0.00
A. P.	0.07	0.05	0.00	0.00
points	0.07	0.23	0.10	0.01

of interactive conversational discussion, a forum discussion contains a lot of characters which will not appear in spoken speech, such as the emoji, some special characters and special abbreviations. In comparison, blogs provide a much larger amount of text data in a standard written format, with the downside that blogs can be written in a style that significantly differs from the speaking style in spontaneous speech. In particular, blogs tend to have much longer sentences, more accurate grammar and clearer logic without conversational interruption. Therefore the blog data is combined with the conversational web data (Hain et al., 2007) to train the LM. In this way the conversational web data provides a good compensation for the conversational speech style with a large vocabulary in diverse topics of spontaneous speech conversations.

In total, the text data of 260,000 words are collected from four blog websites which are further referred to as “cast”, “atomic”, “cadia” and “addict”. The text from Warhammer 40K wikipedia is also harvested which provides 26,000 additional words. Together with the conversational web data, in total there are six text data resources. Table 4.3 shows a few examples of the words that have significant more occurrences in the blog data than in the wikipedia data or in the conversational web data. It indicates that the blog data is complementary to both the Warhammer Wikipedia data and the conversational web data in providing some special words used frequently in Warhammer 40K games. Furthermore, Table 4.4 shows a comparison of the cross-entropy between the text data components and the SWC manual transcription.

Each of the six N-gram LMs is first trained based on the text from one data resource. Then the six LMs are interpolated with the weights optimised based on the manual transcripts of SWC1. The LM produced in this way incorporates the complimentary statistic properties from all six text data resources, and Table 4.5 shows the statistics of text data from each resource as well as the corresponding LM interpolation weights.

Table 4.4 *The SWC manual transcription and the text data components (*$\log_2$ *used).*

	SWC overall	Blog data overall	Warhammer wikipedia	Conversational web data
Vocabulary	7.0k	13.3k	4.1k	457.8k
Cross-entropy with SWC	197.9	288.4	257.6	302.6
Vocabulary in cross-entropy calculation	7.0k	4.4k	1.9k	6.3k

Table 4.5 *LM components and text data statistics.*

LM component	Number of words	Vocabulary	Interpolation weight
blog cast	71.2k	7.0k	0.06
blog atomic	126.8k	7.9k	0.05
blog cadia	40.4k	3.9k	0.19
blog addict	21.1k	3.3k	0.05
Warhammer wikipedia	26.2k	4.1k	0.003
conversational web data	165.9M	457.8k	0.65

The quality of interpolated LM could be evaluated by the number of out-of-vocabulary words (OOV) and the perplexity (PPL). Table 4.6 compares the 4-gram 30k vocabulary interpolated LM with the 4-gram 30k vocabulary LM component which is trained on conversational web data only. Compared to using the LM based on the conversational web data only, the interpolated LM improves the overall PPL by 34.5% relative, from 264.5 to 173.3.

Overall the PPLs shown in Table 4.6 are high compared to many other large scale conversational speech recognition tasks. For example, the meeting recognition system has a PPL below 100 on the AMI corpus (Hain et al., 2007; McCowan et al., 2005). The high PPL of LM on SWC data could be potentially improved by collecting more in-domain text data, for example more blog data. It is worth mentioning that SWC3 has higher OOV than SWC2 potentially because the conversations in SWC2 is more similar to the conversations due to repeated players and a similar game setup. As the LM interpolation is optimized with data from SWC1 only, it is potentially biased to the conversation style and vocabulary of SWC1 and SWC2.

In summary, this section has discussed the necessity of collecting blog data to train an in-domain LM, as well as how the LM is constructed by interpolating the LM components trained on the text data from a mixture of resources, namely the blog data, the Warhammer wikipedia data and the conversational web data. This interpolated 30k vocabulary 4-gram

Table 4.6 *Perplexity of interpolated LM and conversational web data only based LM on SWC manual transcripts.*

		SWC1	SWC2	SWC3	Overall
Number of word		88.5k	116.3k	56.3k	261.1k
Vocabulary		4.0k	4.6k	2.5k	7.0k
Interpolated LM	OOV	1.4k (1.6%)	2.8k (2.4%)	3.9k (6.9%)	8.1k (3.1%)
	PPL	173.4	195.9	135.0	173.3
Conversational web data based LM	OOV	1.3k (1.5%)	2.8k (2.4%)	3.9k (6.9%)	8.1k (3.1%)
	PPL	271.1	327.8	164.9	264.5

LM will be used in all experiments on SWC data in the following sections and the following chapters if there is no special explanation on exceptions.

4.4 Baseline Systems

Section 4.4.1 will first introduce the datasets for two strategies of preparing acoustic models: adapting an existing acoustic model and training a standalone acoustic model. A baseline system for each task is then introduced along with their performances in the following sections: Section 4.4.2 will focus on the baseline acoustic model adaptation system, and Section 4.4.3 will present the baseline standalone training system. Section 4.4.4 further illustrates the performance of the advanced algorithms in DSR based on the standalone system, such as multi-channel dereverberation and beamforming.

4.4.1 Task and dataset

Since SWC has in total 24.6 hours annotated speech, it is possible to train a small acoustic model with part of the SWC data so that the training-test mismatch is minimised. It is also possible to adapt an existing acoustic model trained on a much larger corpus to the SWC data. Therefore, multiple datasets are defined for different needs.

For the standalone training, three datasets are required: a training set (“train”), a development set (“dev”) and an evaluation set (“eval”). For the acoustic model adaptation, only two datasets are required: an adaptation set (“dev”) and an evaluation set (“eval”). This is because the acoustic model has been previously trained on corpora other than SWC. Since there is a limited number of volunteer participants in the recording, some players participated in multiple recording sessions. As a result it is impossible to have a complete separation in speaker among different sets with a balance on both the amount of data and speaker gender. Therefore, a compromise is made by defining two types of

Table 4.7 *Datasets for SWC.*

	Dataset	Strip	Duration (h)	#Utterance	#Speaker
adapt-1 (AD1)	dev	1.AB, 2.AB, 3.AB	16.3	26.2k	22
	eval	1.C, 2.C, 3.C	8.2	13.3k	22
adapt-2 (AD2)	dev	1	8.0	14.0k	9
	eval	2, 3	16.6	25.6k	18
standalone-1 (SA1)	train	1, 2.A, 3.A	13.5	22.6k	22
	dev	2.B, 3.B	5.5	8.5k	18
	eval	2.C, 3.C	5.6	8.4k	18
standalone-2 (SA2)	train	1	8.0	14.0k	9
	dev	2.A, 3.A	5.5	8.7k	18
	eval	2.BC, 3.BC	11.1	16.9k	18

datasets for each acoustic model strategy: one has the best speaker separation between training and test, and the other has the least speaker separation between training and test. This is to investigate the best and worst performance with each strategy of acoustic model preparation.

To achieve that, each recording session, or recording file, is split into three consequent parts with equal amount of annotated speech. These are referred to as the three strips in one session: strip A, strip B and strip C. For simplicity, the first strip of all recording sessions in SWC1 is notated as "1.A". Similarly, the second and third strips of all recording sessions in SWC2 combined together are notated as "2.BC". This notation is used in Table 4.7.

For the acoustic model adaptation, the first and second strips of every session in all three day recordings are used as the development data ("dev"), and the third strip of every session in all three day recordings is used as the evaluation data ("eval"). In this way there is the least separation between dev and eval datasets in gender, speaker and speaking style. This is referred to as the "adapt-1" or "AD1" dataset definition in Table 4.7. Alternatively, SWC1 can be used as development data ("dev") while SWC2 and SWC3 can be used as evaluation data ("eval"). In this way, the speakers are best separated between two datasets, and the female speakers only appear in the evaluation set. This is referred to as the "adapt-2" or "AD2" dataset definition, as shown in Table 4.7.

For the standalone training of the acoustic model, the whole SWC1 plus the first strip of every recording session in SWC2 and SWC3 are used as the training data ("train"). The second strip of every recording session in SWC2 and SWC3 is used as development data ("dev"), and the third strip of every recording session in SWC2 and SWC3 is used as evaluation data ("eval"). This dataset partition strategy has the least speaker separation, and it is referred to as the "standalone-1" or "SA1" in Table 4.7. In an alternative strategy, the whole SWC1 is used as training data ("train"). The first strip of every session in SWC2

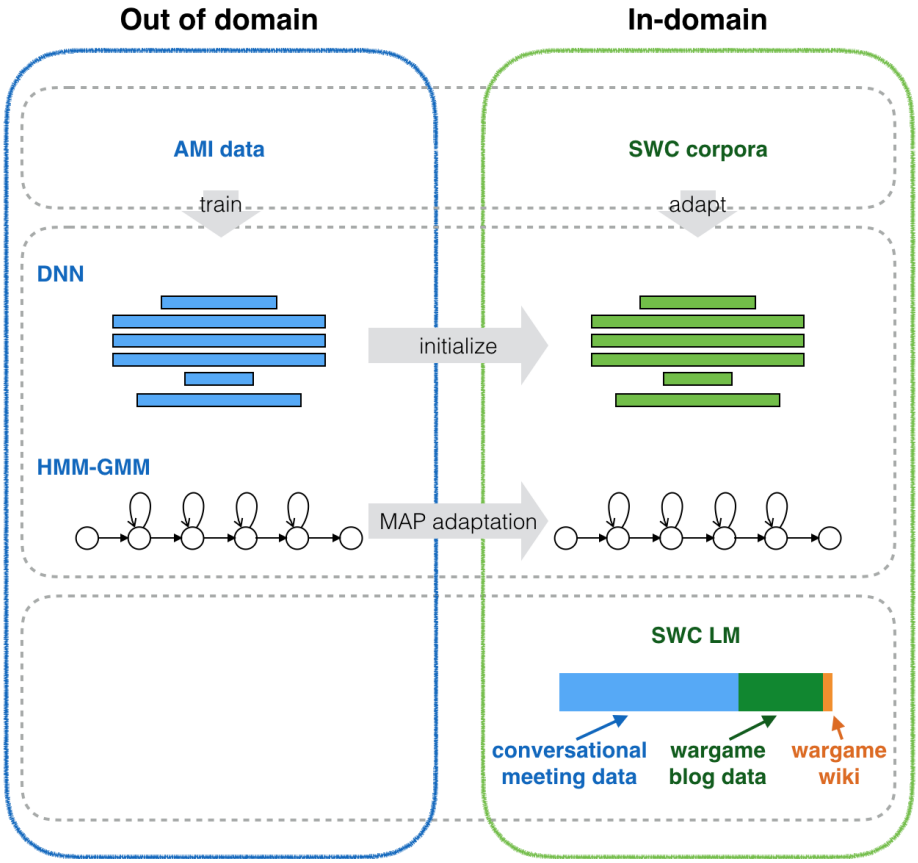


Fig. 4.7 *SWC baseline system: acoustic model adaptation in a DNN-HMM-GMM structure.*

and SWC3 are used as development data ("dev"). The second and the third strips of every session in SWC2 and SWC3 are used as evaluation data ("eval"). This data partition strategy has the best speaker separation between training set and evaluation set, and it is referred to as "standalone-2" or "SA2" in Table 4.7.

4.4.2 Baseline adaptation system

In the baseline system with acoustic model adaptation, the acoustic model to be adapted is first trained on the meeting corpus AMI (McCowan et al., 2005) in a DNN-HMM-GMM structure. A DNN is employed at the front-end which produces bottleneck features, and the HMM-GMMs are trained on these bottleneck features. The seed DNN trained from the AMI corpus is optimised layer by layer with around 80 hours audio from the AMI corpus with manual transcripts and the headset recording based alignment. The DNN parameter optimisation is performed to minimise the cross-entropy on the development dataset. This DNN training follows the configuration in an early work published by the

Table 4.8 *SWC adaptation baseline performance with "AD2" dataset definition using source acoustic models trained on the AMI corpus.*

	dev	eval					
	SWC1	SWC2	SWC3	overall			
	WER	WER	WER	Sub.	Del.	Ins.	WER
IHM	24.9	46.4	50.5	33.4	9.3	5.0	**47.7**
SDM	55.2	75.0	85.2	53.2	19.1	6.0	**78.2**
MDM	53.5	71.6	82.4	52.4	15.4	7.3	**75.0**

author (Liu et al., 2014). It is worth explaining that the AMI corpus is used to construct the seed acoustic model because it is a larger corpus with a similar recording setup to the SWC. In the AMI corpus, meeting style conversational speech is recorded with both individual headset microphones and an eight-channel distant circular microphone array. In addition, both the AMI corpus and the SWC are recorded in meeting rooms.

To adapt the source acoustic model to the SWC data, the DNN front-end is first fine-tuned for a few more iterations using the development data from SWC in the "AD2" dataset definition as shown in Table 4.7. The DNN fine-tuning is based on the alignment acquired with the SWC headset recordings and the DNN-HMM-GMMs previously trained on the AMI headset recordings (Liu et al., 2014). With the adapted DNN, bottleneck features are generated on the SWC development data to update the HMM and GMMs with maximum-a-posterior (MAP) adaptation for 8 iterations. The software employed for DNN training and bottleneck feature extraction is TNet[2], and the software employed for HMM-GMMs training, adaptation and decoding is HTK[3]. Neither speaker adaptation or normalisation is involved, but only the utterance level mean normalisation over the bottleneck features. The whole adaptation process is illustrated in Fig. 4.7.

As for the detailed configuration of DNN, the input to DNN are the 368 dimensional features compressed with the discrete cosine transform (DCT) from the 31 continuous frames (+/-15) of 23 dimensional log-Mel filter bank features. In primary experiments, it is found that using a smaller context window would degrade the recognition performance. The DNN topology is 368:2048$\times$3:26:1993. The 1993 output nodes of DNN correspond to the 1993 tied context dependent phoneme states. In primary experiments, it is found that using a wider output layer significantly increases the computation cost without a significant improvement in performance.

Results of the baseline adaptation system are reported on IHM, SDM and MDM in Table 4.8. For MDM, the weighted delay and sum beamforming (wDSB) is performed

[2] http://speech.fit.vutbr.cz/software/neural-network-trainer-tnet
[3] http://htk.eng.cam.ac.uk

using BeamformIt (Anguera et al., 2007), based on the audio recordings from the 8 channels circular microphone array placed at the center of table ("TBL1"). The scoring for IHM is performed with NIST tool "sclite", while the scoring for SDM and MDM is performed with another NIST tool "asclite" to allow the reference from up to 4 speakers when scoring the overlapped speech.

As shown in Table 4.8, even with the supervised adaptation using manual transcription and the headset recording based alignment on the SWC development data, the baseline adaptation system yields high WERs on the development dataset: 24.9% for IHM, 55.2% for SDM and 53.5% for MDM with 8 channel wDSB. The WER on the evaluation data is higher than on the development data, with 47.7% for IHM, 78.2% for SDM and 75.0% for MDM overall. The WER on SWC3 is higher than SWC2 due to the mismatch between development data and the SWC3 data in speaker gender and speaking style. Beamforming reduced the WER compared to the SDM baseline by 3.1% relative on SWC1, 4.5% relative on SWC2 and 3.3% relative on SWC3. The relative WER reduction by the wDSB is smaller compared to similar experiments on other corpora like the AMI corpus. On the AMI corpus, the wDSB introduced 7-9% relative WER reduction in a DNN-HMM-GMM ASR system (Liu et al., 2014) and around 8.5% relative WER reduction in a DNN-HMM hybrid ASR system (Swietojanski et al., 2013) using log Mel filter bank features. More investigation on the high WERs will be detailed in Chapter 5.

4.4.3 Baseline standalone system

As mentioned previously, the acoustic model can also be trained in a standalone manner using training data from SWC in the "SA1" dataset definition shown in Table 4.7. It is worth emphasizing that this "SA1" dataset has overlapped speakers between training and test, and the recognition performance is expected to degrade if a better speaker separation is used. The toolkit Kaldi[4] is used to construct a state-of-the-art acoustic model in the hybrid DNN-HMM structure. With the open and active community support to Kaldi development, one direct benefit from using Kaldi is that a Kaldi recipe for SWC data is shared in the research community, which ensures to replicate the original work.

The Kaldi recipe for SWC follows the existing Kaldi recipe for the AMI corpus[5] regarding the algorithms and configuration. The 13 dimensional MFCC features from 7 contextual frames (+/-3) are extracted and compressed with the linear discriminant analysis (LDA) to 40 dimensions. The output features from the compression will be further referred to as the "LDA features". The LDA features are used to train HMM-GMMs. The

[4] http://kaldi-asr.org
[5] https://github.com/kaldi-asr/kaldi/tree/master/egs/ami

Table 4.9 *SWC standalone training system performance with "SA1" dataset definition.*

		dev	eval	overall			
				Sub.	Del.	Ins.	WER
IHM	LDA+MLLT	50.9	51.8	35.9	8.9	6.4	51.3
	+SAT	48.7	**48.8**	34.4	8.1	6.3	**48.7**
	+MMI	48.8	49.1	34.4	8.8	5.7	48.9
	DNN	44.4	44.3	30.5	9.7	4.1	44.4
	+sMBR	42.0	**42.0**	29.5	7.6	5.0	**42.0**
	+fMLLR	48.1	48.1	32.9	11.4	3.8	48.1
	+sMBR	44.9	44.8	31.2	9.8	3.8	44.9
SDM	DNN	78.9	80.5	53.9	21.4	4.4	79.7
	+sMBR	76.4	**77.3**	39.1	35.5	2.2	**76.8**
MDM	DNN	76.0	77.9	53.3	18.2	5.5	76.9
	+sMBR	73.8	**74.9**	36.0	36.0	2.4	**74.3**

initial model training uses hypothesis timing where utterances are split into equal chunks. The alignment is updated each time the acoustic model significantly improves during the training process.

The HMM-GMMs based on monophones are first trained, followed by the HMM-GMMs trained on the clustered states. This is then followed by the LDA training and the maximum likelihood linear transform (MLLT), the speaker adaptive training (SAT), and the maximum mutual information (MMI) training. Alignments from the system with LDA+MLLT is used for DNN training. The input of the DNN are 520 dimensional feature vectors, comprised of 13 (+/-6) contextual 40 dimensional LDA features that were used for HMM-GMM training. DNN parameters are initialised with the restricted Boltzmann machines (RBMs), in a topology of 520:2048$\times$6:3804, and then fine-tuned to minimise the cross-entropy on the development dataset. This is further followed by 4 iterations of further fine-tuning for minimum phone error (MPE) or the state level variant of the minimum Bayes risk (sMBR), with the updated alignment. The configuration of the DNN topology and DNN training follows the existing Kaldi recipe for the AMI corpus. The configuration is not the optimized for the SWC data but it is a sound enough baseline setup.

For IHM, the speaker adaptation is also performed. The HMM-GMMs with LDA+MLLT+SAT provide the alignment and feature level maximum likelihood linear regression (fMLLR) per speaker for DNN training. In the speaker adaptation experiments on IHM, the DNN parameters are initialised with RBMs in a topology of 143:2048$\times$6:3710, because the DNN input features are comprised of 11 (+/-5) contextual 13 dimensional MFCC features with fMLLR applied.

To reduce the memory cost, the 30k 4-gram LM introduced in Section 4.3 is pruned. Table 4.9 shows the performance using different acoustic models and microphone channels. Compared to the LDA+MLLT based system on IHM recordings, the SAT reduces the overall WER of HMM-GMMs based system from 51.3% to 48.7% (5.1% relative), while MMI did not reduce WER further. For DNN-HMM hybrid system however, speaker adaptation via fMLLR degraded WER from 44.4% to 48.1%. Among the baseline experiments, the best overall WER of 42.0% on IHM is achieved with sMBR fine-tuning on DNN parameters without speaker adaptation. Therefore, fMLLR is not used in SDM or MDM experiments. However, fine-tuning DNN with sMBR is found to be effective for both SDM and MDM, achieving the best overall WER of 76.8% on SDM and 74.3% on MDM. The weighted delay and sum beamforming reduced the WER from SDM baseline by 3.3% relative.

4.4.4 More beamforming and dereverberation

With the multi-channel and multi-media recordings, it is possible to apply the multi-microphone based beamforming and dereverberation techniques to improve the performance of DSR on the SWC data. In addition, it is possible to use the speaker localization data to replace the blindly estimated TDOA which can have limited the beamforming performance with the potential TDOA estimation errors, considering the considerable amount of body movement and overlapped speech in the SWC recordings. Therefore, three more beamforming algorithms have been tried on SWC data: the simple delay and sum beamforming (DSB) using the speaker location from Ubisense system, the super-directive beamforming (SDBF) and minimum variance distortionless beamforming (MVDR). These beamforming algorithms are realized with BTK[6] with the support from Dr. Kenichi Kumatani (Wölfel and McDonough, 2009). The generalized weighted prediction error (GWPE) is used for the multi-microphone based dereverberation, and it is combined with beamforming by using the dereverberation output as the input for beamforming. The GWPE algorithm is realized with the codes provided by NTT[7]. Table 4.10 compares the performance of different multi-channel techniques when they are used alone or in combination in a standalone system based on the DNN-HMM hybrid acoustic model as described in Section 4.4.3. All the multi-channel algorithms employs the audio recordings from the 8 microphones in the circular array placed at the center of table ("TBL1").

The speaker head location tracked by the Ubisense system helps to improve the beamforming performance slightly over a blind estimation of TDOA, reducing the overall

[6] http://distantspeechrecognition.sourceforge.net
[7] http://www.kecl.ntt.co.jp/icl/signal/wpe/

Table 4.10 *WER for different beamforming and multi-microphone based dereverberation algorithms.*

	Use Ubisense based speaker location?	dev	eval	Overall			
				Sub.	Del.	Ins.	WER
SDM	-	76.4	77.3	39.1	35.5	2.2	76.8
wDSB	-	73.8	74.9	36.0	36.0	2.4	74.3
GWPE+wDSB	-	72.5	74.6	50.1	17.7	5.8	73.5
DSB	Y	72.7	74.5	50.1	19.0	4.4	73.6
SDBF	Y	73.2	74.7	50.5	18.7	4.7	73.9
MVDR	Y	72.4	74.2	50.0	18.6	4.7	73.3
GWPE+DSB	Y	71.9	73.8	49.0	18.9	4.7	72.7
GWPE+MVDR	Y	**70.5**	**72.1**	43.5	24.4	3.4	**71.3**

WER from 74.3% to 73.6%. When the TDOA is estimated from the tracked speaker head location, the MVDR provides the best performance among all tried beamformers, with an overall WER of 73.3%. The SDBF does not perform better than DSB potentially because the noise in SWC data is not stationary or diffusive. Combining beamforming with multi-microphone dereverberation algorithm GWPE brings down the WER further, from 73.6% to 72.7% when the DSB is used and from 73.3% to 71.3% when the MVDR beamforming is used. The combination of GWPE and MVDR provides the best performance on MDM with an overall WER of 71.3%, this is 5.5% absolutely lower or 7.2% relatively lower compared to SDM overall WER of 76.8%.

It is worth emphasizing that in the MVDR implementation, the noise spatial covariance matrix is estimated with a randomly selected piece of the background noise from SWC1. The noise covariance matrix is is not updated based on the background noise in a local context of the concerned speech utterance. As a result, such implementation has potentially limited the performance of the MVDR beamformers. As for other beamformers, primary experiments have been conducted to tune the setup based on the default configuration in BTK. In the tuning, some default configurations by BTK which are useful for other tasks have to be disabled for SWC. For example, it is found that the Kalman filter based speaker location tracking has to be disabled because of the fast change in the location of active speaker due to the high speaker switch rate, the large proportion of overlapped speech, and the existing of four speakers in four directions in the SWC data.

4.5 Summary

This chapter has covered the details about SWC from four aspects: the data collection design and recording configuration in Section 4.1, the post-recording data processing and

analysis in Section 4.2, the language model preparation in Section 4.3 and the baseline ASR systems as well as their performance in Section 4.4.

As mentioned in Section 3.1, the SWC recording was conducted to collect more data of natural native and spontaneous English speech with multi-microphone and multi-media recordings. With an unusual topic of a desk-top game Warhammer 40K, the SWC data collection minimises the privacy infringement in real recordings of natural spontaneous conversations while encouraging the natural speaker movement and frequent overlapped speech in heated discussion. The recording was performed over three days and the recording system is comprised of three parts: the audio recording, the video recording and the Ubisense based speaker location tracking. The audio recording system involves synchronised audio recordings from both the individual headset microphones and the multiple distant microphone arrays, and the video recordings using multiple cameras. The video recording is conducted from different angles with multiple cameras. With the Ubisense system[8] providing three dimensional speaker location tracking using the radio signal, SWC is the first speech database with natural speaker movement along with speaker location tracking.

The post-recording analysis on SWC data reveals some unique properties of the SWC data compared to existing corpora, namely very short utterances with an average duration of 2.2 seconds, a high proportion (around 50%) of utterances with part being overlapped with competing speech, the emotional speech with a big gender difference and the speaker movement while talking. The audio recording is annotated and transcribed manually, leading to 24.6 hours annotated speech from 14 male speakers and 8 female speakers.

Considering the unusual topic of desktop-game in SWC compared to existing speech corpora, the text from four Warhammer 40K blog websites are collected. It is combined with the text from the Warhammer wikipedia and the conversational web data for language model training. The Warhammer wikipedia provides a larger game related vocabulary, and the conversational web data provides much larger text data of a conversational speech style with a wider range of topics. The LM components are first trained on the text data from each resource and then interpolated by minimising the perplexity on SWC1 manual transcripts. The examples words, cross-entropy and perplexity have been compared in Section 4.3 to show the complementary properties of the multiple components in the text data recourse as well as the trained LM components. The LM based on the combination of all components is used in all following experiments on SWC data.

Two configurations of the baseline ASR systems for SWC data have been chosen, namely the adaptation configuration and the stand-alone training configuration. The

[8] http://ubisense.net/en

adaptation configuration takes the DNN-HMM-GMMs based acoustic model from AMI corpus and fine-tunes it with the SWC development dataset. The stand-alone training configuration trains a DNN-HMM hybrid acoustic model based on the SWC training dataset. The recognition is performed on individual headset recordings, single distant microphone recordings and multiple distant microphone recordings. The overall WERs are very high compared to existing speech corpora. The WERs on close-talking recordings are above 40% and the WERs on far-field recordings are above 70%. Using multiple distant microphone recordings achieved better performance than using only single distant microphone. The 8 channel wDSB which has been reported to bring effective WER reduction only improves the WER on SWC data by a very small proportion. The best performance on distant recordings is 71.3% in the overall WER. It is achieved with a combination of 8 channel dereverberation GWPE and MVDR beamforming with TDOA from the speaker location tracked by Ubisense system.

This chapter has focused on reporting SWC data recording and baseline systems. More analysis on the reasons for high WERs on SWC data compared to existing speech corpora will be covered in the next Chapter 5, where the SWC data is used as a case study to highlight the remaining challenges for DSR in the daily application when the machine learning encounters natural human-to-human multi-party spontaneous and conversational speech.

The work on SWC is accomplished jointly with Dr. Charles Fox who organised the recording events and set up the recording hardware and software (Fox et al., 2012), and with Dr. Madina Hasan who trained the in-domain LM using the blog text data and conversational web data. The first day recordings of SWC1 is released in Interspeech 2013 (Fox et al., 2013). The three day full recordings are released in Interspeech 2016 along with the LM, the baseline speech recognition results and a Kaldi recipe to replicate the standalone training system (Liu et al., 2016). A website has been constructed for a wide access of relevant information about the SWC data: http://mini-vm20.dcs.shef.ac.uk/swc/SWC-home.html.

Chapter 5

Challenges in Real Natural Spontaneous Speech

Contents

5.1 Speech Recognition of Headset Recordings 74

5.1.1 Utterance duration, number of word and speaking rate 75

5.1.2 Emotional speech 80

5.1.3 Speaker and session difference 81

5.1.4 Competing speech 82

5.1.5 Conclusion 83

5.2 DSR: Factor Analysis with Simulated Data 84

5.2.1 Overlapping Speech 84

5.2.2 Reverberation 86

5.3 DSR: Factor Analysis with Real Data 92

5.3.1 Speaker attributes 94

5.3.2 Microphone attributes and speaker movement 96

5.3.3 Environment attributes and distributed microphone 99

5.4 Summary and Discussion 103

The previous Chapter 4 introduced the Sheffield Wargame Corpora (SWC), a multi-channel and multi-media database of recordings of the natural spontaneous multi-party speech conversations in native English. The recording configuration emphasises the

challenges for DSR in real applications such as the competing speech, the room reverberant, the background noise and the speaker movement. The experiments showed that the state-of-the-art acoustic models based on deep neural network (DNN) could not cope with all the challenges in SWC data, resulting with the WERs above 70% on distant recordings and above 40% on close-talking recordings. This chapter breaks down the influence factors on ASR performance in distant speech recordings and analyses the impact of each factor. Using the SWC data as a case study, Section 5.1 discusses the challenges for ASR with the close-talking recordings when the speech to be recognized is highly natural, spontaneous and emotional. Then Section 5.2 analyses the influence factors by examining the change in recognition performance when these factors are added one by one into the simulated distant speech recordings. Section 5.3 moves to the real distant recordings where it is more difficult to quantify the impact of each factor. Section 5.4 summarises this chapter by highlighting the major findings and conclusions.

This chapter involves a lot of experimental analysis. Most experiments follow the configuration used in previous chapter. For such experiments the configuration details are skipped and the reference will be provided to the corresponding sections where the details can be found. For the experiments with a different configuration, the difference will be highlighted and the configuration will be explained.

5.1 Speech Recognition of Headset Recordings

As mentioned in Section 4.4, on SWC data the WERs are above 40% even with headset recordings. In an earlier work published by the author (Liu et al., 2014), the acoustic model based on DNN-HMM-GMMs trained on 15.8 hours headset recordings in the AMI corpus (McCowan et al., 2005) leads to a WER of around 27%. In comparison with a DNN-HMM hybrid acoustic model trained with 13.5 hours headset recordings from SWC data based on the Kaldi recipe, the WER is more than 10% higher in absolute value (Section 4.4.3). The large performance difference is caused by the different properties of these two corpora. This comparison is of particular research interest because the AMI corpus is a typical representation for the large vocabulary meeting corpora with relatively controlled recording set-up in terms of the speaking style and the conversation topic. In comparison, the SWC is a typical representation for the real natural spontaneous multi-party conversational speech without a limit on the speaking style or the speech topic, even though the table-top game has naturally centralized the topic on Warhammer 40K. Since there is very limited data of the latter, most existing speech techniques have been developed on audio recordings of planned speech or read speech. As a consequence there is limited knowledge about their

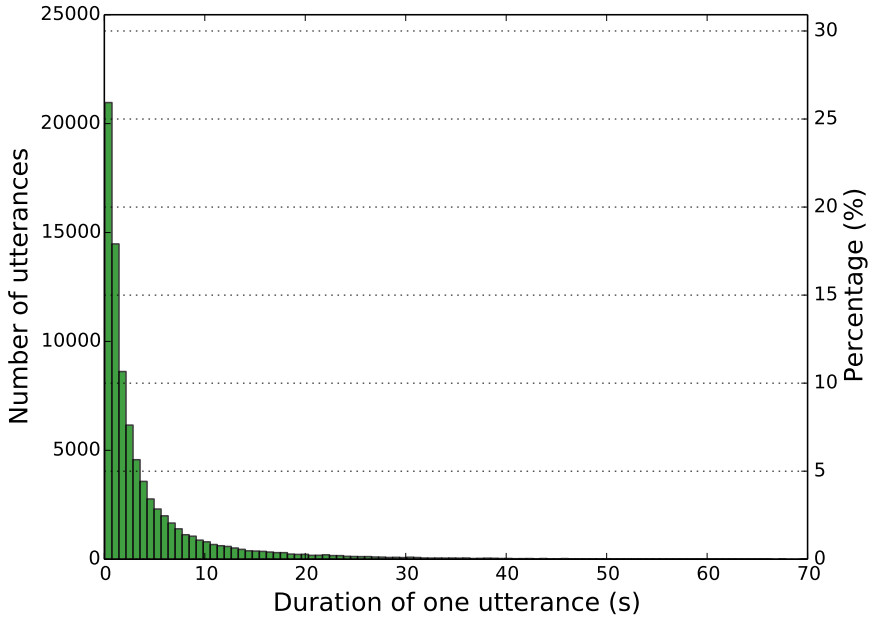


Fig. 5.1 *The average utterance duration histogram in the AMI corpus.*

weaknesses when these techniques are applied on real natural spontaneous multi-party conversational speech. In addition, the challenging factors that exist in headset recordings are very likely to exist in distant recordings as well. Therefore, this section uses the SWC as a study case to investigate the influence factors in the headset recordings of natural spontaneous multi-party conversations. The investigation is conducted from the following aspects: utterance duration, speaking rate, emotional speech, individual speaker difference and competing speech.

5.1.1 Utterance duration, number of word and speaking rate

In the general analysis provided in Section 4.2, the average utterance duration in SWC data is 2.2 seconds, while in AMI data it is above 4 seconds. In addition, Fig. 4.5a showed that in SWC the number of utterances decreases approximately exponentially as the utterance duration increases. Similar trend is observed with the AMI corpus, as illustrated in Fig. 5.1. The utterance duration is associated with the number of words spoken in one utterance and the talking speed, namely the number of words spoken per second. In this work the talking speed is also referred to as the speaking rate. These two factors are more interesting than the utterance duration, because the number of words is directly associated with the WER calculation, and the speaking rate could impact the recognition difficulty particularly in adverse conditions such as reverberant and noisy environment.

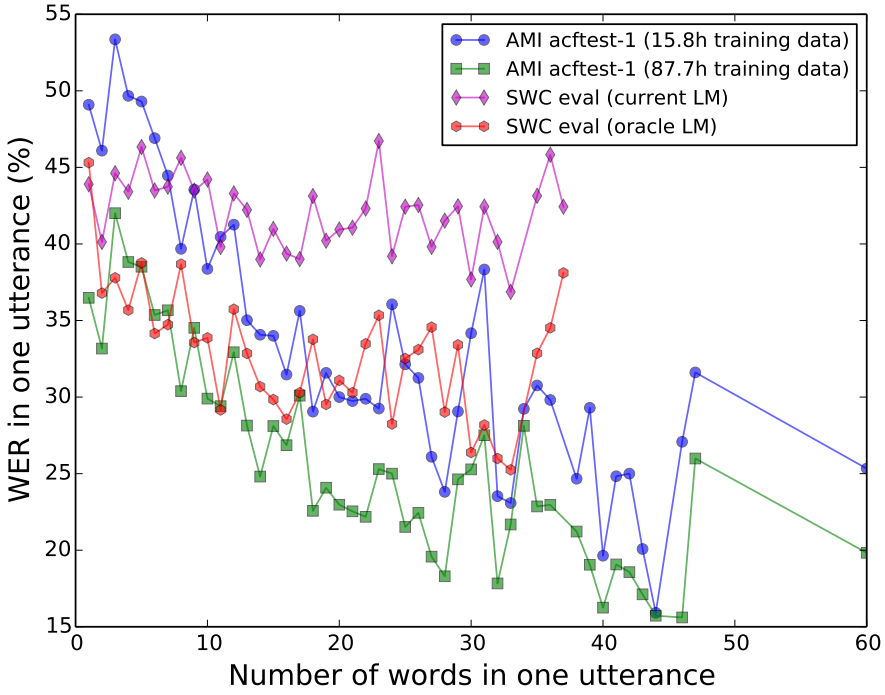


Fig. 5.2 *The average utterance level WER given different number of words in one utterance. "AMI acftest-1": the 6.1 hours evaluation data from AMI corpus with dataset defined by (Liu et al., 2014); "SWC eval": the 5.6 hours evaluation data from the "SA1" dataset definition of SWC data shown in Table. 4.7.*

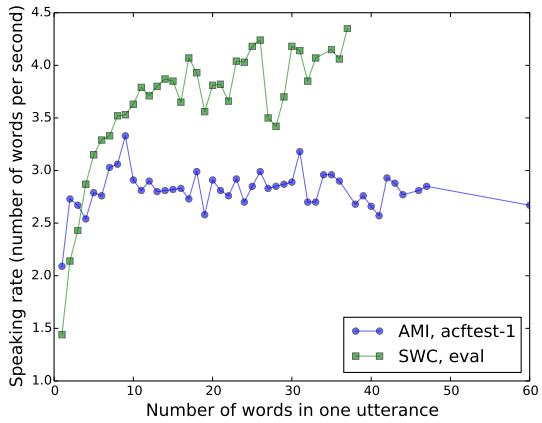
The relationship between the average number of words in one utterance and the WER of that utterance is illustrated in Fig. 5.2. The pink and red lines correspond to the WERs on the 5.6 hours SWC evaluation data, with the DNN-HMM hybrid acoustic model trained on the 13.5 hours SWC headset recordings in a standalone manner as described in Section 4.4.3. The WERs for the pink line are based on the 4-gram LM described in Section 4.4.3, and the WERs for the red line is based on a 4-gram LM trained with the SWC manual transcripts directly, *i.e.* the "oracle LM". The blue line corresponds to the WERs on 6.1 hours evaluation data in the AMI corpus, using DNN front-end and HMM-GMMs based acoustic model trained on 15.8 hours AMI headset recordings in a standalone manner (Liu et al., 2014). The green line corresponds to WERs on the same 6.1 hours AMI corpus evaluation data, using DNN front-end and HMM-GMMs acoustic model trained with 87.7 hours AMI headset recordings (Liu et al., 2014). Fig. 5.2 compares the statistics between the SWC data and the AMI corpus given similar amount of training data and evaluation data (blue line and pink line), the impact of adding more data for acoustic training on the AMI corpus (blue line and green line), and the impact of language model on SWC data

due to its unique topic (pink line and red line). Note that Fig. 5.2 only takes into account the number-of-word conditions with at least 10 utterances.

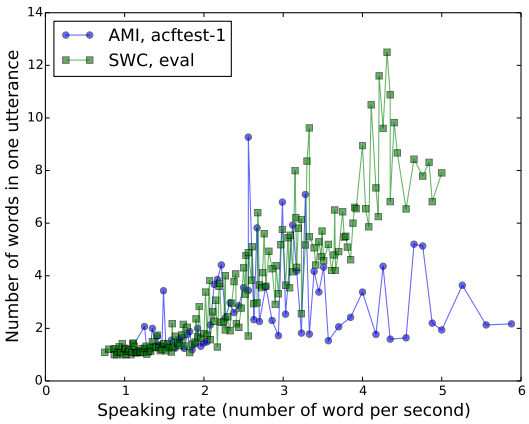
In all cases the average utterance level WER decreases as the number of words in a single utterance increases as shown in Fig. 5.2. This may be caused by two reasons. First, the utterances with small number of words tend to have a WER value that is more sensitive to any kind of recognition errors, due to a small denominator in the WER calculation. Second, short utterances with small number of words benefit less from the language model compared to long utterances. When comparing the AMI corpus (blue line) and the SWC data (pink line), the average WER decreasing gradient as the number of words in one utterance increases is larger in the AMI corpus than the SWC data. This may suggest that the LM for SWC needs further improvement. To quickly verify that, one 4-gram LM is trained with the manual transcripts of all SWC data. Such an "oracle LM" is combined with the same acoustic model used as before, and it helps to decrease the overall WER on the dev set from 42.0% to 33.0%, and the overall WER on the eval set from 42.0% to 33.2%. The red line in Fig. 5.2 shows an analysis of how the WER varies on the SWC evaluation data by the number of words in one utterance based on the decoding results with the oracle LM. The average WER per utterance is reduced significantly for utterances of any lengths except for those utterances with one word only. However even with the oracle LM, on long utterances iwth more than 20 words the overall gradient of the WER curves for the SWC data is not as steep as the AMI corpus. This difference implies that there is some other difference between the long utterances in the SWC data and the long utterances in the AMI corpus.

In the AMI corpus, increasing the amount of training data from 15.8 hours (blue line in Fig. 5.2) to 87.7 hours (blue line in Fig. 5.2) significantly improves the acoustic model performance, hence decreases the WERs on utterances of all lengths, particularly on short utterances with less than 10 words. While SWC has limited amount of data, it can be expected that a similar WER improvement can be observed with more training data. In addition, as shown on AMI data, increasing the audio data for the training of acoustic model and DNN front-end training decreases the WERs of utterances in all lengths. This suggests that the SWC recognition performance can be largely improved if more data is available for the training of acoustic model and DNN front-end.

In practice, speaker adaptation is often beneficial for WER reduction. This is the case with the AMI corpus as shown by Liu et al. (2015, 2014). In the baseline experiments explored in previous chapter, the speaker adaptation based on the fMLLR is unfortunately not beneficial at all (Table 4.9). The suspect is that the poor performance of the fMLLR based speaker adaptation is related to the vivid and temporally diverse speaking style in the SWC data. It is possible that dedicated speaker adaptive training may bring some



(a) *Average speaking rate per utterance at a given number of words per utterance.*



(b) *Average number of words per utterance at a given speaking rate per utterance.*

Fig. 5.3 *Relationship between speaking rate and the number of words in one utterance.*

WER improvement. Further research is still needed to thoroughly investigate the effective speaker adaptation methods for the SWC data.

The average WER of utterances with different number of words may reveal some weaknesses in the ASR system. In Fig. 5.2, SWC data has a smaller descending gradient in WER curves compared to the AMI corpus, and this suggests a weak LM. The current LM for SWC still needs further improvement because only limited text data is available for LM training. The blog data combined with conversational web data helps to alleviate this problem, but the analysis suggests that more in-domain data is needed.

The speaking rate varies from utterance to utterance. An analysis is carried out to investigate how much the speaking rate impacts the performance of speech recognition on headset recordings. Fig. 5.3 illustrates the relationship between the average speaking rate and the number of words per utterance, and the results are compared between the SWC data and the AMI corpus. Fig. 5.3a shows the average speaking rate (y axis) among the utterances with given number of words (x axis), and Fig. 5.3b shows the average number of words per utterance (y axis) among the utterances with given speaking rate (x axis). Fig. 5.3a suggests that the average speaking rate in the SWC data is higher than the AMI corpus due to the speech spontaneity and the natural speaking style in the SWC recordings. This potentially implies a bigger challenge to acoustic models from the SWC data than from the AMI corpus. Fig. 5.3a shows that the SWC data has a slower speaking rate in short utterances and a faster speaking rate in long utterances. The speaking rate in the SWC data increases from below 1.5 words per second to above 4.0 words per second as the utterance length increases. In comparison in the AMI corpus the speaking rate mostly stays between 2.5-3.0 words per second. Furthermore, Fig. 5.3b suggests that in SWC data high speaking

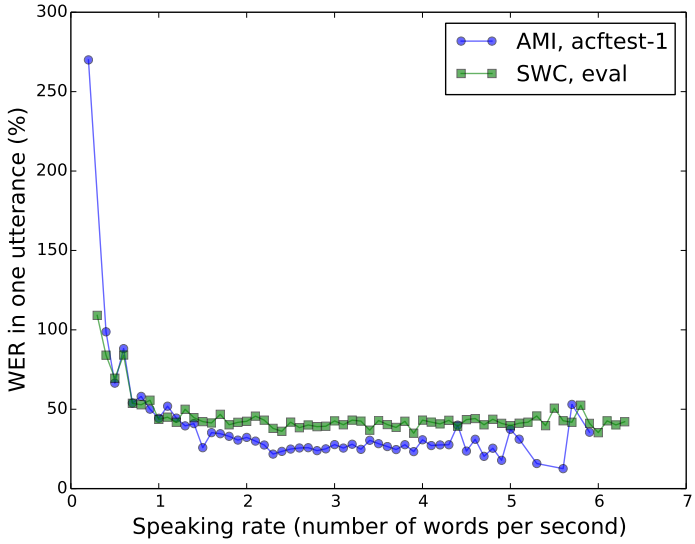


Fig. 5.4 *Average WER per utterance at a given speaking rate per utterance.*

rate has a larger chance to appear in long utterances while low speaking rate has a larger chance to appear in short utterances.

Fig. 5.4 further shows the average WER (y axis) among utterances of given speaking rate (x axis). It shows that counter-intuitively, WER decreases as speaking rate increases. This could be explained from the joint effect of the speaking rate and the number of words per utterance. From Fig. 5.3b, utterances with high average speaking rate tends to be long utterances, and according to the analysis on Fig. 5.2 long utterances tends to benefit more from LM thus having lower WERs. Therefore overall the utterances with high average speaking rate tend to have low WERs, as utterances with high speaking rate are very likely to be long utterances. Previously, Fig. 5.2 has shown that the AMI corpus has a steeper gradient in the descending WER curves than the SWC data, indicating a continuously increasing benefit from LM as the utterance length increases in the AMI corpus. That could explain the observation in Fig. 5.4 that the WER descending in the SWC data settles at a lower speaking rate with a higher WER value compared to the AMI corpus. The speaking rate where the average WER settles is around 2.5 words per second in the AMI corpus, and is slightly above 1.5 words per second in the SWC data. In both the AMI corpus and the SWC data, when the speaking rate increases to above 5 words per second, the WER increases slightly.

In summary, the recognition performance at utterance level is related to the number of words and the average speaking rate per utterance. Compared to short utterances, long utterances tend to have lower WERs due to a bigger benefit from the LM, even though long utterances also tend to have a faster speaking rate. A comparison between Fig. 5.2

Table 5.1 *Emotional speech analysis: number of laughs, average number of word per utterance and average WER per utterance.*

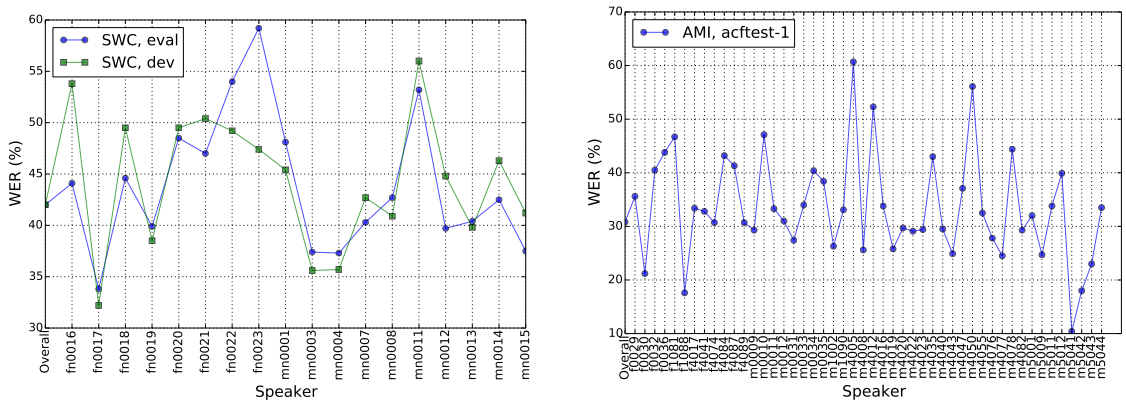
Dataset	#Laugh in one utterance	#Utterance	Average WER per utterance (%)	Average #word per utterance
dev	0	7793 (91.5%)	42.4	7.0
	1	609 (7.2%)	47.9	8.6
	2	111 (1.3%)	56.1	10.4
eval	0	7668 (91.3%)	42.3	6.9
	1	603 (7.2%)	50.7	7.9
	2	105 (1.3%)	53.3	9.2
	3	19 (0.2%)	52.0	15.9

Fig. 5.4 implies that the number of words per utterance seems to have a higher impact than the speaking rate on recognition performance in both the AMI corpus and the SWC data, particularly when the speaking rate is below 5 words per second. When the speaking rate is above 5 words per second, there is a slight increase in WER for both the AMI corpus and the SWC data.

5.1.2 Emotional speech

Another important difference between the SWC data and most existing speech corpora is that SWC recording involves natural emotional speech. In SWC2 and SWC3 transcripts, the emotional vocal sounds such as laugh is annotated along with the speech utterances. Therefore, it is possible to investigate the potential impact of emotional speech on the recognition performance. The investigation is performed by first grouping all speech utterances in the dev set and the eval set according to the number of laugh tags in that utterance, and then calculating the average WER in that speech utterance group. Table 5.1 shows the analysis results based on the recognition results from the standalone ASR system described in Section 4.4.3.

Table 5.1 shows that the average number of words per utterance increases as the number of laugh tags increases. This suggests that long utterances have a higher chance to include laughs. Among all transcribed speech utterances, around 9% of utterances have one or more laugh tags, and the average WER per utterance is 5-10% higher in absolute value compared to the remaining 91% speech utterances without laughs. Section 5.1.1 has shown that the average WER per utterance tends to decrease as the number of words per utterance increases. That is the case with the majority speech utterances which do not have emotional sounds. Table 5.1 suggests that 9% of utterances have emotional sounds such as laughs, and among such utterances the longer ones tend to have more laughs than the shorter ones.



(a) *WER per speaker in the SWC data.* (b) *WER per speaker in the AMI corpus.*

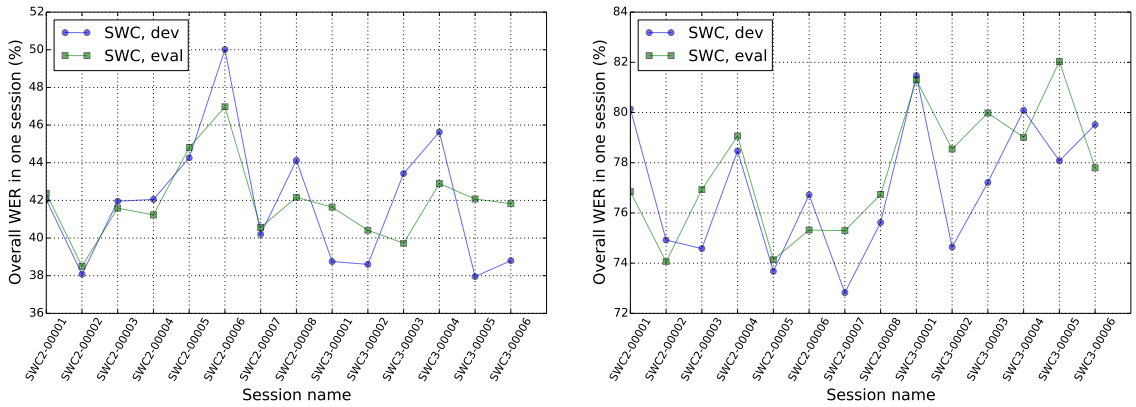
Fig. 5.5 *WER per speaker (male speaker IDs start with "m" and female speaker IDs start with "f").*

Compared to the speech utterances without laughs, the average WER of utterances with laughs is 8.4-11.0% higher in absolute value.

5.1.3 Speaker and session difference

Since there is a limited number of speakers in the SWC data, particularly the female speakers, it is worth investigating whether the recognition performance is imbalanced by the speaker individual difference or the gender difference. Therefore, the average WER per speaker is calculated based on the headset recording recognition results of the SWC headset recordings using the standalone system detailed in Section 4.4.3. In addition, the WER per speaker is compared with the WER per speaker in the AMI system published by Liu et al. (2014) based on a similar amount of training data and test data from the AMI corpus. As shown in Fig. 5.5a, in general the lack of female speech in the training data leads to a slightly higher WER for female speakers than male speakers. The variation of WER per speaker on SWC data is of similar range when compared on the AMI corpus.

Section 4.1 has mentioned that there is some slight difference in the recording configuration from session to session in the SWC data. To quantify the impact of such configuration difference, the average WER per session is calculated based on the recognition results of the SWC headset recordings using the standalone system introduced in Section 4.4.3. A similar analysis is performed on the recognition results based on single distant microphone recordings as a comparison. As shown in Fig. 5.6a, for headset recordings, most sessions have overall WERs within a normal variation range, except for Session "SWC2-00006", though there is no special recording arrangement for that session. For single distant microphones, as shown in Fig. 5.6b, the WERs on eval set in all sessions of SWC3 are higher



(a) *IHM.* (b) *SDM.*

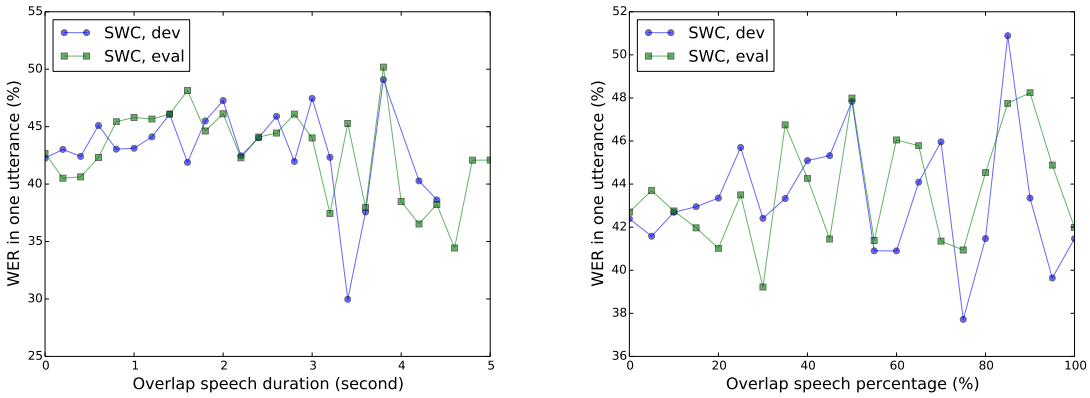
Fig. 5.6 *Overall WER per session in SWC using different microphones channels.*

than SWC2, potentially because of the two extra male speakers in the recording area whose voice is only recorded with distant microphones but not headset microphones.

5.1.4 Competing speech

It is common to have competing utterances with overlapped speech in spontaneous conversations. As indicated by the statistical analysis in Fig. 4.5d, around 50% of the transcribed utterances in SWC have at least one competing utterance. For headset recordings, the headset microphone is designed to reduce the volume of sound from sources beyond a certain distant to the microphone or a certain range of arrival angle, thus the effectiveness of suppressing competing speech is largely dependent on the headset microphone design and the distance between the target speaker and the loudest competing speaker. Theoretically the headset recordings are not completely free from the competing speech, though in practice this point is often neglected. In comparison, the distant recordings has more severe distortions from competing speech due to a lack of directivity and acoustic attenuation over distance in the microphone design.

To quantify the distortion of competing speech, WERs from the standalone system (Section 4.4.3) based on the headset recordings is analysed against the amount of speech overlap, *i.e.* the absolute amount of overlapped duration each utterance (Fig. 5.7a) and the percentage of the overlapped duration in each utterance (Fig. 5.7b). As shown in Fig. 5.7, the average utterance level WER on the headset recordings of the SWC evaluation dataset is not correlated with either the duration of overlapped speech or the percentage of overlapped speech in one utterance. This suggests that overlapped speech could be neglected in SWC headset recordings.



(a) *Average WER at a given overlapped speech duration of one utterance.*

(b) *Average WER at a given overlapped percentage of one utterance.*

Fig. 5.7 *Average WER per utterance of given level of speech overlapped.*

5.1.5 Conclusion

The SWC data is a representation of natural spontaneous multi-party conversational speech, and the AMI corpus is a representation of multi-party conversational speech in a business meeting style. To highlight the challenges in the real natural spontaneous multi-party conversational speech, this section has compared the recognition performance of two ASR systems trained and tested on the headset recordings from the SWC data and the AMI corpus respectively, using similar amount of data.

First, the analysis on average WER and utterance length in Section 5.1.1 suggests that in both SWC data and AMI corpus short utterances tend to have higher WERs compared to long utterances, and SWC has higher overall WERs than AMI corpus. One reason is that the performance on middle and long utterances are much poorer than AMI corpus. This potentially suggests that the current LM in SWC is yet to improve. In addition, as found in Section 5.1.1, the SWC utterances are much shorter than the AMI utterances on average, which leads to a big challenge to acoustic models. Second, among all annotated speech utterances in SWC recordings, around 9% utterances are of emotional speech with emotional vocal sound such as laughs. Such speech utterances have WERs on average 8-10% absolutely higher than the speech utterances without laughs (Section 5.1.2). It is also found that long utterances have a higher chance to contain emotional speech than short utterances. In terms of speaker and gender issues in SWC data, due to less female speech data compared to male speech data, the average WER on female speakers is higher than the average WER on male speakers in SWC. For SWC headset recordings, there is no particular impact spotted from recording session configuration or overlapped speech thanks

to the high quality of headset microphones. In comparison, such factors could introduce some challenges to DSR, which will be covered in details in the following sections.

5.2 DSR: Factor Analysis with Simulated Data

In Section 5.1, investigation is performed on factors that could contribute to the high WERs on SWC headset recordings compared to other speech corpora such as the AMI corpus. It was found that real natural spontaneous speech is more challenging due to the very short utterances and the emotional speech. Besides, the quality of LM and the limited amount of training data, particularly the female speech data, are also two important factors causing the high WERs in all SWC based recognition experiments. Some of the analysis conclusions in Section 5.1 also apply to the distant recordings. For example, the findings regarding the utterance duration, the number of words per utterance, the gender bias and the emotional speech. There are also some factors that have different impacts on speech recognition performance using distant microphone recordings compared to using headset recordings. This section investigates the influence factors of the SWC DSR performance based on the recordings from one or multiple distant microphones. The experiments in this section are mainly based on simulated data, so that when one factor is analysed, the other factors are fixed. The investigation will be conducted from different aspects: the overlapped speech (Section 5.2.1), the reverberation and a combination of both the overlapped speech and reverberation (Section 5.2.2). The analysis of background noise is omitted in this work though the background noise is also a very important factor in DSR.

5.2.1 Overlapping Speech

As pointed in Section 4.2 and illustrated in Fig. 4.5d, 50% of the transcribed utterances in SWC have at least one competing utterance that causes overlapped speech. In addition, it has been shown in Section 5.1.4 that there is no significant influence from overlapped speech on headset recordings in terms of speech recognition performance, thanks to the high quality headset microphones. In comparison, distant microphones cannot prevent overlapped speech in the distant recordings.

To quantify the impact of overlapped speech alone without reverberation or background noise, the recordings with overlapped speech is simulated by adding up the signals from four headset microphones. This is possible because they are synchronised at sample level during recording. Then acoustic models are trained and tested on the simulated data with overlapped speech in the same way with the standalone system described in Section 4.4.3. The scoring is performed with a NIST tool “asclite” which allows multiple reference

Table 5.2 *WERs with and without overlapped speech (%). "IHM": the original individual headset microphone recordings; "IHM.OL": simulated data with overlapped speech.*

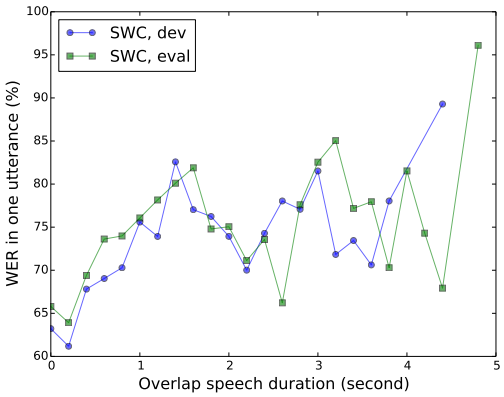
Training data	Test data	dev	eval	Overall			
				Sub.	Del.	Ins.	WER
IHM	IHM	41.3	41.2	29.9	6.9	4.4	**41.3**
	IHM.OL	68.4	70.7	43.4	13.7	12.4	**69.5**
IHM.OL	IHM.OL	65.1	66.4	44.5	14.4	6.8	**65.7**
	IHM	50.9	51.2	32.7	15.1	3.1	**51.0**

utterances for one hypothesis utterance. When there is overlapped speech, the "asclite" scoring tool identifies a hypothesis word as correct if it matches any candidate reference words among the overlapped utterances. Before such word level comparison the "asclite" tool performs the word alignment in a way so that the WER is minimised in each utterance. Another NIST scoring tool frequently used is "sclite" for ASR based on headset recordings. The difference between "asclite" and "sclite" is that the "sclite" only allows one reference utterance for one hypothesis utterance.

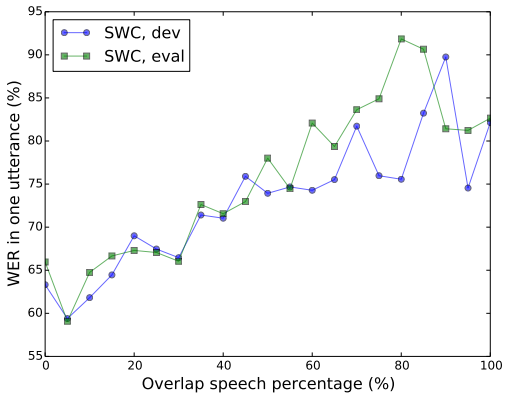
Table 5.2 shows the WERs on clean headset recordings and simulated data with overlapped speech. For a fair comparison, all the recognition results are scored with "asclite". As shown in Table 5.2, overlapped speech degrades the recognition performance at both training stage and testing stage. At the training stage, the existence of overlapped speech introduces some confusion to the acoustic model and degrades the overall WER on clean headset recordings, from 41.3% to 51.0%, namely by approximately 10% absolute. At testing stage, overlapped speech makes the recognition task more challenging. When the acoustic model is trained with clean headset recordings, the existence of overlapped speech in test data degrades the overall WER, from 41.3% to 69.5%, namely by more than 20% absolute. When the acoustic model is trained on simulated data with overlapped speech, the existence of overlapped speech in test data degrades the overall WER, from 51.0% to 65.7%.

When comparing the recognition performance on simulated data with overlapped speech, the acoustic model trained with simulated data outperforms the acoustic model trained with clean headset recordings, by 3.8% absolute in WER. However, such improvement in robustness against overlapped speech is at the cost of a 9.7% absolute WER increase on clean headset recordings. This suggests that multi-condition training based on the artificially overlapped speech data is a suboptimal strategy, particularly in the applications where the proportion of overlapped speech is relatively small.

When the acoustic model is trained and tested with matched data condition, the existence of overlapped speech degrades WER from 41.3% to 65.7%, *i.e.* more than



(a) *By absolute duration of overlapped speech in one utterance.*



(b) *By the percentage of overlapped speech in one utterance.*

Fig. 5.8 *Average utterance level WER with different amount of overlapped speech on SWC dev and eval dataset.*

20% absolute or 50% relative. Fig. 5.8 illustrates the average utterance level WER given different amounts of overlapped speech. It shows that the average utterance level WER increases as the percentage of overlapped speech in one utterance increases (Fig. 5.8b). This is different from an earlier observation in Fig. 5.7b on headset recordings where no significant correlation was found between the amount of overlapped speech and WER.

5.2.2 Reverberation

Reverberation is an important factor in DSR as it introduces convolutional distortion to the speech signal. As reviewed in Section 2.3.1, the convolutional distortion from room reverberation can be approximated with the room impulse response (RIRs), and reverberant speech can be simulated by convolving the RIRs with clean headset recordings. This section investigates the impact of reverberation on speech recognition performance with and without overlapped speech. The analysis is conducted in four stages with simulated reverberant speech, to better understand the impact of reverberation and its interaction with overlapped speech. The first stage focuses on the static reverberation effect without overlapped speech, *i.e.* the recordings from each headset microphone convolved with the same room impulse response. The second stage investigates the static reverberation effect in the presence of overlapped speech. The signals simulated in the first stage based on the headset recordings from multiple synchronised channels are mixed together additively. The third stage investigates the effect of changing reverberation. The RIRs measured in the same room but at different locations are used to simulate the speech signal with frequent speaker movement. The fourth stage compares the performance of dereverberation using

Table 5.3 *Analysis of the impact of overlapped speech and reverberation speech on WER with simulated data based on RIR from microphone "TBL1-01" at the center of table.*

Data	dev	eval	Overall			
			Sub.	Del.	Ins.	WER
IHM	41.3	41.2	29.9	6.9	4.4	41.3
static reverberation	51.5	51.7	36.0	10.1	5.5	51.6
static reverberation & overlap	72.1	73.7	49.1	18.5	5.3	72.9
SDM	76.4	77.3	39.1	35.5	2.2	76.8

the reverberant signals simulated with the RIRs corresponding to different microphone arrays.

As noted in Section 2.3.1, RIRs are measured with the swept sine wave signal to avoid the non-linear effects in the measurement system. Multiple RIRs are measured with a loudspeaker placed at different locations in the room. The loudspeaker is mounted on a portable support so that the height of loudspeaker can be adjust to the head height of an adult. The microphones for RIRs estimation are the same microphones used for distant speech recordings in SWC, and the microphone installation has been described previously in Section 4.1. Multiple recordings are taken and averaged for the same loudspeaker location to reduce the RIR measurement errors caused by background noise and occasional device artefacts. The recording of the swept sine signal for RIR estimation is designed and conducted by Dr. Charles Fox.

For the first stage, all the headset recordings are convolved with the same RIR independently. The same dataset and algorithms used for building the standalone system in Section 4.4.3 are employed here for training and decoding. For the second stage, the simulated reverberant signal for the first stage from multiple headset channels are added together to simulate a combination of overlapped speech and reverberation. The corresponding experiment results are shown in Table 5.3.

Table 5.3 shows that the reverberation alone ("static reverberation") increased the overall WER by 10.3% absolute or 24.9% relative compared to the clean headset recordings ("IHM"). Compared to the recognition performance with overlapped speech alone in both training and test as shown in Table 5.2 in previous section, reverberation on top of overlapped speech increased the overall WER from 65.7% to 72.9%, *i.e.* by 7.2% absolute or 11.1% relative. Therefore, reverberation leads to significant WER increase no matter whether there is overlapped speech or not. In particular, on SWC data the reverberation introduces more relative WER increase when there is no overlapped speech. In contrast, the overlapped speech introduces 19.3% absolute WER increase when it is added on top

Table 5.4 *WER improvement from multi-microphone based dereverberation and beamforming on simulated reverberant speech based on RIRs from "TBL1" array.*

Simulation configuration	Treating algorithm	dev	eval	Overall			
				Sub.	Del.	Ins.	WER
static reverb.	GWPE	48.9	48.8	34.0	9.9	4.9	48.9
	GWPE+wDSB	50.6	50.5	35.4	9.9	5.1	50.6
static reverb. & overlapped speech	GWPE	72.1	73.0	29.7	41.1	1.8	72.5
	GWPE+wDSB	72.1	73.2	48.9	18.4	5.3	72.6

of reverberation. This indicates that overlapped speech has a big contribution to the poor DSR performance on the SWC data.

In the baseline standalone system reported in Section 4.4.4, beamforming and dereverberation only brought very limited improvement to recognition performance. Thus it is worth investigating the performance of dereverberation and beamforming algorithms on simulated data. For that purpose, the RIRs from the 8 distant microphone in the circular array at the center of the table ("TBL1") given the same loudspeaker location are used to simulate the multi-channel reverberant speech recordings without background noise. The generalized weighted prediction error (GPWE) algorithm (Yoshioka and Nakatani, 2012) is applied on the simulated multi-channel reverberant speech signal. Since the simulated data does not have background noise and noise suppressing algorithms can potentially introduce speech distortion, it is anticipated that some beamforming algorithms might not improve the recognition performance on simulated reverberant speech without background noise.

Table 5.4 shows the results based on the multi-channel simulated reverberant speech. If one compares the results in Table 5.4 with Table 5.3 on simulated data without overlapped speech, dereverberation using GWPE decreased WER by 2.7% absolute or 5.2% relative. The effectiveness of GWPE became marginal when there is overlapped speech reducing overall WER from 72.9% to 72.5%, even though all speakers and all channels were simulated with the same single RIR. As expected, applying wDSB on top of GWPE degraded the recognition performance slightly no matter whether there is overlapped speech or not. This is caused by the speech distortion caused by the TDOA estimation error when applying wDSB. Therefore the comparison between Table 5.3 and Table 5.4 explains that overlapped speech is one key reason for the observation in Section 4.4.4 that the derverberation algorithm GWPE only introduced marginal improvement to the recognition performance on the SWC data.

In addition, a comparison between Table 5.3 and Table 5.4 indicates the best possible performance of the multi-channel dereverberation algorithm GWPE, when there is no

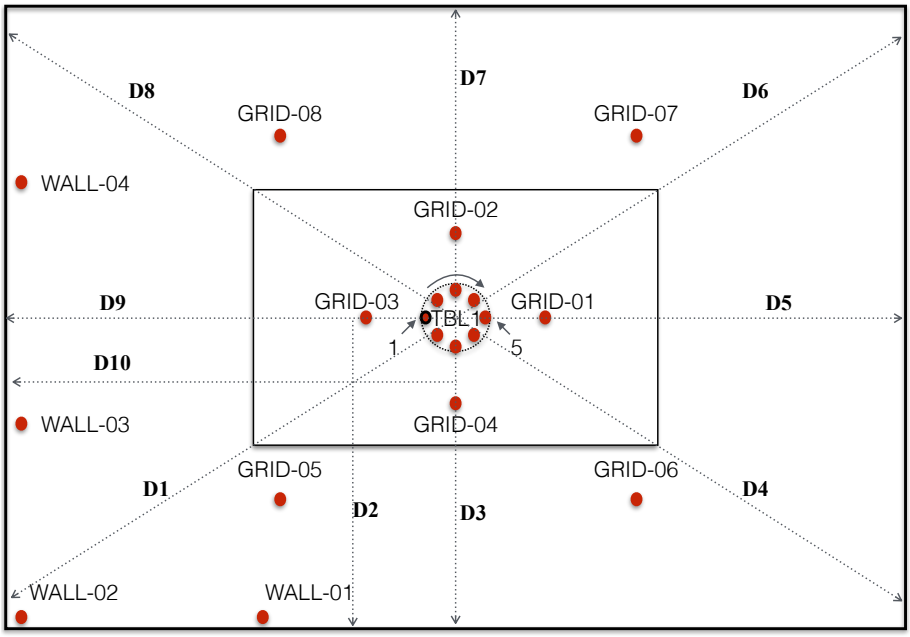


Fig. 5.9 *Loudspeaker location configuration to measure RIRs for simulating the speaker movement in the room.*

overlapped speech. For the simulated static reverberation without overlapped speech, the multi-channel dereverberation algorithm GWPE only recovered the WER degradation caused by static reverberation from 51.6% ("static reverberation" in Table 5.3) to 48.9% ("static reverberation" in Table 5.4), and there is still a big gap compared to the WER based on headset recordings, namely 41.3% ("IHM" in Table 5.3). GWPE reduced the WER by 2.7% absolute or 26.2% relative while there is still 7.6% absolute WER difference between using headset recording and using dereverberated simulated data. The situation is even more adverse to GWPE when there is overlapped speech, as previously discussed. This could explain the small improvement from GWPE on real multi-channel distant recordings as shown in Table 4.10 in Section 4.4.4.

For the third stage, the impact of changing reverberation due to speaker movement is investigated. For that purpose, the RIRs are measured with loudspeakers situated at different locations in the room. As shown in Fig. 5.9, the loudspeaker is placed at three different distances in 10 directions from the center of the table. The three different distances are approximately 0.15 m, 0.45 m and 0.75 m to the margin of the table horizontally. The 10 directions are labelled from "D1" to "D10" in Fig. 5.9. At all locations, the loudspeaker is mounted in a portable stand of adjustable height. Three heights, namely 1.4 m, 1.5 m and 1.6 m, are chosen so that the center of loudspeaker is approximately at the same height with the head of a standing adult human. Multiple measurements are performed with a swept sine wave signal at each location to reduce the measurement errors.

Table 5.5 *WER comparison between static reverberation and changing reverberation due to speaker movement.*

	With overlapped speech?	dev	eval	Overall			
				Sub.	Del.	Ins.	WER
static reverberation	No	82.3	81.8	51.4	27.7	2.9	82.1
reverberation changes every 10s	No	79.0	78.8	53.9	20.8	4.2	78.9
	Yes	86.1	86.3	55.6	28.3	2.3	86.2

Due to an unexpected technical problem, the RIRs measured for different sound location sources only cover up to 4 kHz in frequency. For speech recognition, an RIR with a bandwidth up to 8 kHz would be capable of describing the reverberation behaviour in the room. However an RIR covering only 4 kHz bandwidth will cut off the speech spectrum above 4 kHz and this will degrade the performance of speech recognition. Therefore, when examining the impact of changing reverberation due to speaker movement, the baseline for static reverberation is updated.

Table 5.5 shows the WERs for systems on the simulated reverberant speech with changing reverberation and the WERs for the new comparable static reverberation baseline. The RIR used for the "static reverberation" in Table 5.5 is measured when the loudspeaker is 0.75 m from the table in "D3" direction at a height of 1.5 m. To simulate the change in the reverberation caused by speaker movement, the RIRs are changed randomly every 10 seconds and convolved with the headset recordings. It is worth noting that the amount of training data and test data is the same as before in all cases with the same dataset definition. As shown in Table 5.5, the changing reverberation actually made the acoustic model slightly more robust to reverberation, improving the overall WER from 82.1% to 78.9%. Adding overlapped speech into simulated data with changing reverberation degrades the WER from 78.9% to 86.2% . Since the WER for the new static reverberation baseline in Table 5.5 is much higher than the correspondent in Table 5.2, it is not sufficient to conclude whether the overlapped speech is more harmful with changing reverberation or with static reverberation.

For the fourth stage, an investigation is conducted regarding the array difference in terms of the dereverberation effectiveness. Dereverberation experiments are performed on the simulated reverberant data using the RIRs estimated from the 8 microphones hanging from the ceiling grid ("GRID"). The performance is compared with the dereverberaiton experiments using the 8 microphones from the "TBL1" circular array. The microphones in the "TBL1" array and the "GRID" group are of the same hardware design. There are only two major differences between the "GRID" group and "TBL1" array that could lead to different RIRs. The first difference is the location of the microphones. As shown in

Table 5.6 *Dereverberation performance comparison using RIRs from different microphone arrays for simulated data.*

Microphone for RIR	Dereverberation	dev	eval	Overall			
				Sub.	Del.	Ins.	WER
TBL1-01	-	51.5	51.7	36.0	10.1	5.5	51.6
TBL1, 8 channels	GWPE	**48.9**	**48.8**	34.0	9.9	4.9	**48.9**
GRID-01	-	53.4	53.8	37.8	9.7	6.0	53.6
GRID-02	-	53.6	54.1	37.9	9.8	6.1	53.8
GRID-03	-	54.0	54.1	37.8	10.8	5.4	54.0
GRID-04	-	53.0	52.9	37.2	10.0	5.7	53.0
GRID-05	-	52.3	52.7	36.4	10.5	5.5	52.5
GRID-06	-	54.6	54.9	38.2	10.9	5.6	54.7
GRID-07	-	56.1	56.4	39.1	11.7	5.4	56.2
GRID-08	-	55.1	55.0	38.4	11.0	5.6	55.1
GRID, 8 channels	GWPE	**47.5**	**47.5**	33.0	9.6	4.9	**47.5**

Fig. 4.2, the microphones in "TBL1" array are placed in the center of the table to form a circular array with a diameter of 20 cm, while the microphones in "GRID" are hanging from the ceiling grid with a much larger average distance between any two neighbouring microphones. Since the microphones in "TBL1" array are mounted on a cylinder stand pointing up, the microphones record sound mostly from the upper sphere of the space above the table. In contrast, the microphones in the "GRID" group are fully exposed in space and they record sound from all directions.

Table 5.6 compares the recognition performance using the RIRs from different microphones for the simulated reverberant data, with and without multi-channel dereverberation based on GWPE. There is no overlapped speech in the simulated data, and beamforming is avoided as it has been found to cause performance degradation on the simulated reverberant data without background noise. Since the microphones in the circular array ("TBL1") are very close to each other, the recognition performance on the simulated reverberant data using the RIRs from these 8 are very similar, thus only the WER from microphone "TBL1-01" is reported in Table 5.6. Because the microphones in the ceiling grid group ("GRID") have larger distance to each other, the recognition performance on the simulated reverberant data using RIR from each microphone is detailed in Table 5.6.

When there is no dereverberation, the recognition performance on simulated data based on the RIRs from any microphone in "TBL1" array outperforms the microphones in the "GRID" microphone group by 0.9-4.6% WER absolute. However, the multi-channel dereverberation performance is better with the "GRID" group, with WER 1.5% lower in absolute value compared to the "TBL1" array. This suggests that the effectiveness

of dereverberation is dependent on the microphone array arrangement more than the recognition performance of each microphone in the array. Therefore, in practice with real recordings, there might be a compromise between the best performance from each microphone or the best performance from the whole microphone array when installing microphones. This case will be further investigated in Chapter 6.

5.3 DSR: Factor Analysis with Real Data

In the previous section analyses the factors that could potentially impact the DSR performance. The analysis was conducted on the simulated data as it allows a factor-by-factor investigation. However, in real recordings there are usually multiple factors interacting with each other, and they determine the recognition performance together in a more complex pattern than any single factor alone. The previous section has shown that overlapped speech makes it much more difficult to treat reverberation due to the interaction between the two factors. In the case with real data, such interaction will be more complicated among multiple factors. This will lead to the confusion in understanding the impact from each influence factor on the final WER, as well as the increased difficulty in treating any single factor. Therefore, this section uses SWC data as a case study and tries to list all possible factors in distant speech recordings. The impact of each factor on recognition performance as well as the interaction among multiple factors will be examined on real distant microphone recordings. In addition, the conclusions obtained in previous sections based on headset recordings and simulated data will be verified on real recordings in this section.

Fig. 5.10 lists all potential influence factors in the SWC distant microphone recordings. The many factors originate from the three categories of objects in the recording space: the speakers, the recording microphones and the environment. In this context, the environment refers to all the objects in the recording space which are not speaker or recording microphone. For example, the room and the furniture in the room are both parts of the environment. Each of the three "major factors" have some "attributes" that consistently impact the DSR performance. For example, the accent and gender of the speaker, the location and installation of the microphone and the average reverberation level of the room. In addition, the major factors interact with each other and trigger more complex attributes that also consistently impact the DSR performance. For example, one natural interaction between the speaker and the environment is the speaker movement in the room, which could have an impact on some speech enhancement algorithms such as beamforming. The

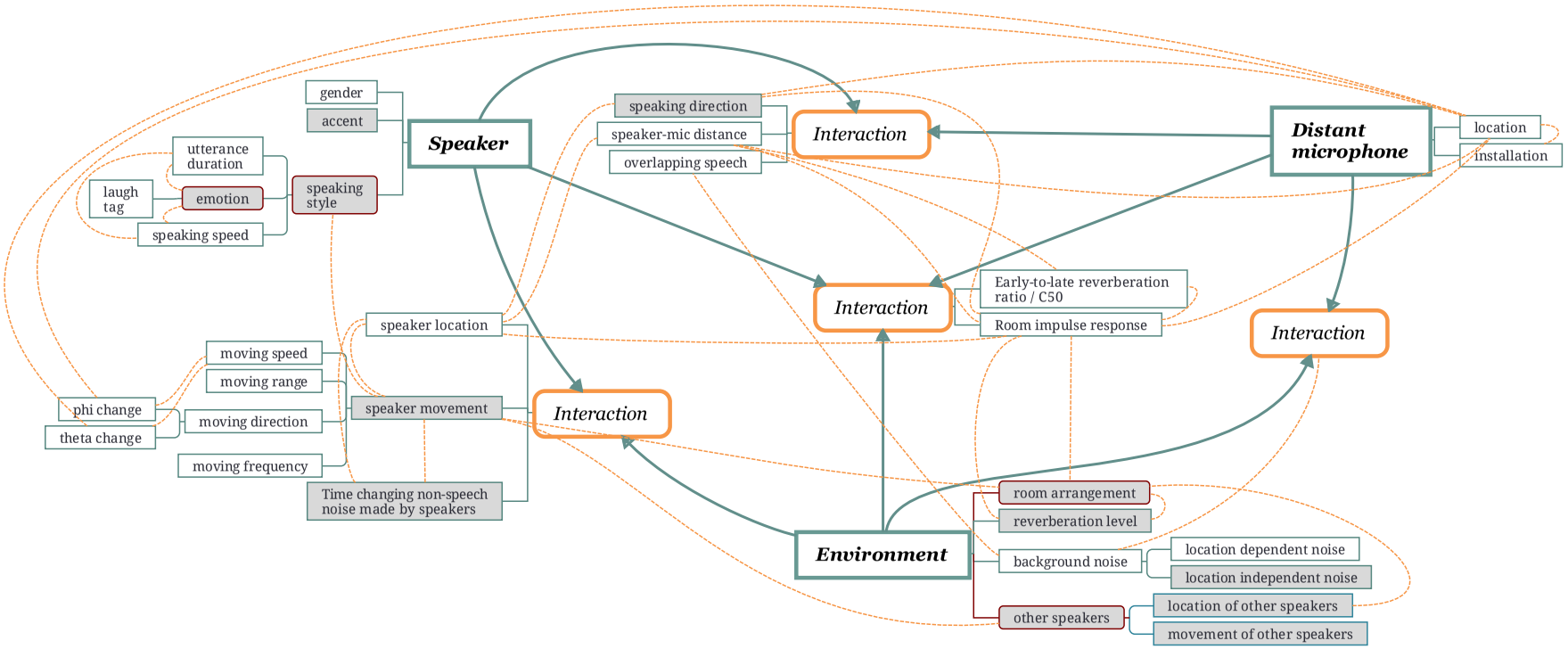


Fig. 5.10 *All physical factors in real distant speech recordings that impact speech recognition performance directly or indirectly (blue square box: the factor or attribute can be quantified; red box with round corner: the factor or attribute cannot be quantified; orange box round corner: an interaction, it is a special category as it is not a single concrete factor; box with grey background: the factor or attribute is omitted in this work; box with white background: the factor or attribute will be investigated in this work; orange dash line: the two factors or attributes are related; blue arrow: the factor is involved in an interaction).*

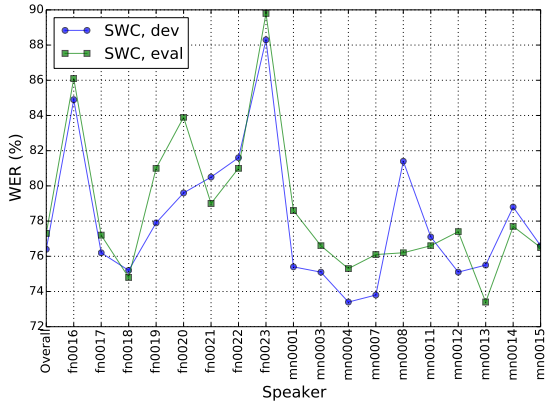
following sections will cover the attributes of each major factor regarding their impact on speech recognition performance.

5.3.1 Speaker attributes

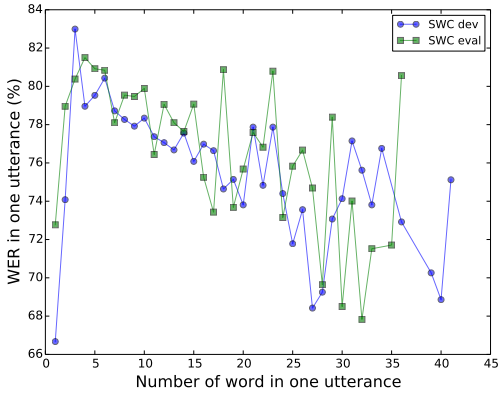
As discussed in Section 5.1, the recognition performance on SWC close-talking recordings is correlated with the speaker attributes such as gender, utterance duration and emotional speech. For example, the limited data female speech data contributes to a higher WER on female speech compared to male speech. In addition, long utterances with a large number of words tend to have lower WERs. The previously investigated speaker attributes on headset recordings also exist in disant recordings. In addition, there are other speaker related attributes that influence the recognition performance of distant recordings. As shown on simulated data in Section 5.2.1, overlapped speech from multiple speakers is an important factor that significantly degrades DSR performance. Besides, the interaction between speaker and distant microphone introduces attributes such as the speaking direction and speaker-microphone distance that potentially impact the reverberation level of the recordings as well as the recognition performance. Furthermore, the interaction between speaker and environment introduces factors such as speaker location, speaker movement in the room and changing background noise caused by speakers such as the foot step noise and the dice noise in the SWC data.

Fig. 5.11a shows the average WER per speaker based on the SDM recordings from one microphone in the circular array located at the center of the table ("TBL1-01"). Similar with headset recordings, the recognition performance is on average much higher with female speakers than male speakers mainly due to the insufficient female speech in the training data. Fig. 5.11b shows the average utterance level WER based on SDM recordings against the average number of words per utterance and average percentage of overlapped speech in one utterance. Similar to the case with headset recordings, short utterances with small number of words tend to have higher WER compared to long utterances with larger number of words, except for those utterances with only one or two words. As for overlapped speech, previous analysis based on simulated data has shown that the average WER per utterance increases significantly as the percentage of overlap in each utterance increases (Fig. 5.7b). In contrast, as shown in Fig. 5.11c, such correlation is not as strong for the real SDM recordings, potentially because the overall WERs are above 70% even without overlapped speech.

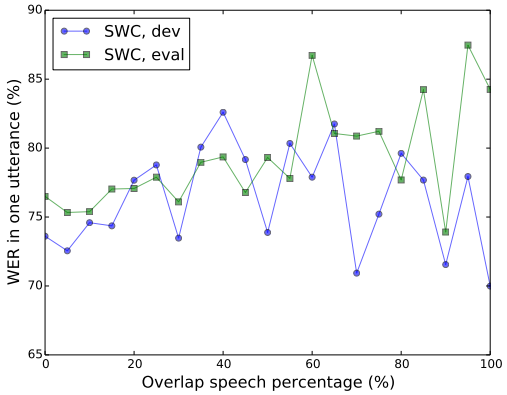
With regard to the speaking rate, in headset recordings based analysis in Section 5.1 the speaking rate is found to correlate with the number of words in one utterance and WER per utterance (Fig. 5.4). Similar correlation is observed in SDM recordings between



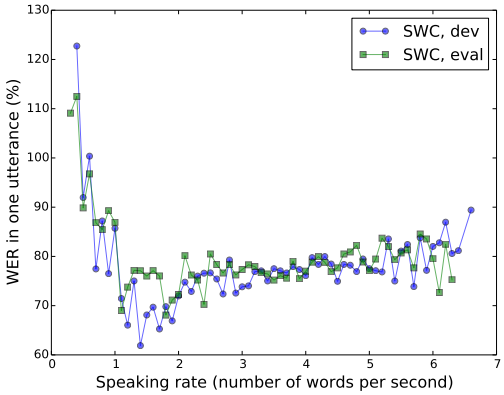
(a) *Speaker (female speaker IDs starting with "f" and male speaker IDs starting with "m").*



(b) *Number of words.*



(c) *Percentage of overlapped speech.*



(d) *Speaking rate.*

Fig. 5.11 *DSR WER and speaker dependent attributes analysis based on SDM recordings.*

speaking rate and WER per utterance. As shown in Fig. 5.11d, the highest WERs on the SDM recordings appear in utterances with very low speaking rate, and these utterances are often short with a small amount of words per utterance (Fig. 5.3). However, different from headset recordings, the SDM recordings are more sensitive to fast speaking rate. When the speaking rate is above 2 words per second in SDM recordings, the average WER per utterance also increases as speaking rate increases. In comparison, the average WER per utterance based on headset recordings does not increase much when the speaking rate is above 1.5 words per second (see Fig. 5.4).

For emotional speech, Table 5.7 shows how the average utterance level WER changes with the amount of laugh tags in that utterance. Similar with headset recording (see Table 5.1), the WER increases as the amount of laugh tags increases.

Table 5.7 *WER on SDM with emotional speech: laugh.*

Dataset	#Laugh tag in one utterance	#Utterance	Average WER per utterance (%)	Average #word per utterance
dev	0	7612 (91.5%)	74.3	7.0
	1	600 (7.2%)	84.0	8.6
	2	109 (1.3%)	88.4	10.4
eval	0	7440 (91.2%)	76.7	6.9
	1	596 (7.3%)	84.5	7.9
	2	102 (1.3%)	91.8	9.2
	3	19 (0.2%)	84.4	15.9

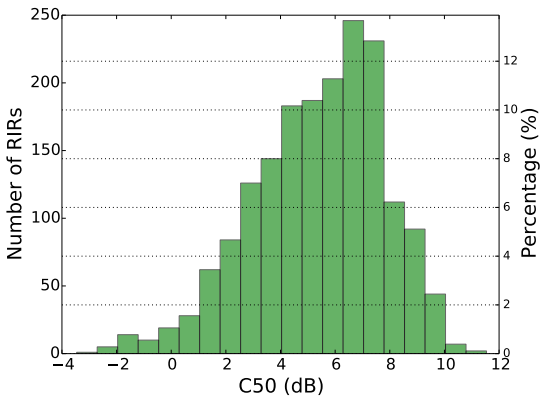
In SWC data, there is no annotation about speaker talking direction and there is very limited variation in speaker accent. Therefore, the impact of speaker accent and speaking direction are not investigated on SWC data.

5.3.2 Microphone attributes and speaker movement

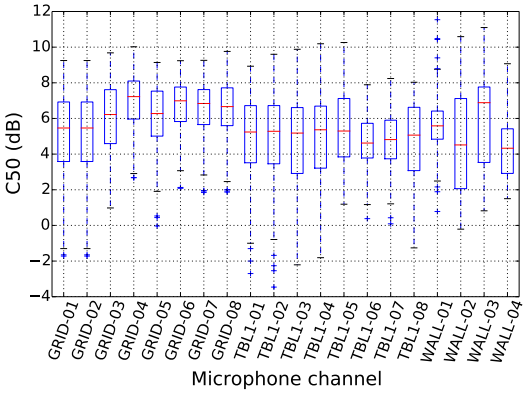
There are many distant microphones recording speech simultaneously in the SWC configuration. In particular, the 20 distant microphones shared among three recording days have exactly the same microphone hardware. Their only difference is the installation, namely the way of mounting and the installation location. Therefore, it is possible to investigate the advantageous and disadvantageous set-up in microphone installation for DSR with these 20 distant microphones. Such investigation is performed from two aspects.

The first aspect is about the difference in reverberation level of 20 microphone channels. As reviewed in Section 2.4, the reverberation level of microphone channel is compared with average C_{50} and T_{60} from the RIRs measured using corresponding microphone given different sound source locations in the room. Details about the RIR measurement have been presented in Section 5.2.2, and the geometry configuration for the microphones and loudspeaker can be found in Fig. 5.9. The reverberation time T_{60} is estimated based on energy decay curve of the octave band with its center at 1kHz. The overall energy decay curve is not used. Because of the technical problem mentioned in Section 5.2.2 that the spectrum of RIRs measured at different locations only covers 4 kHz, the reverberation time estimated from the full energy curve will be higher than the actual value.

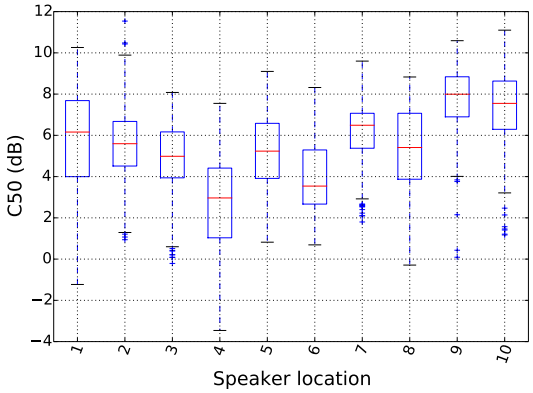
The histogram of C_{50} from all microphones given different sound source locations is shown in Fig. 5.12a. There is a large variation in C_{50}. Further analysis on the distribution of C_{50} per microphone channel and per sound source location is demonstrated in Fig. 5.12b and Fig. 5.12c respectively. Together with Fig. 5.12d, they suggest that the variation in C_{50} comes from the sound source location, the microphone location and the microphone



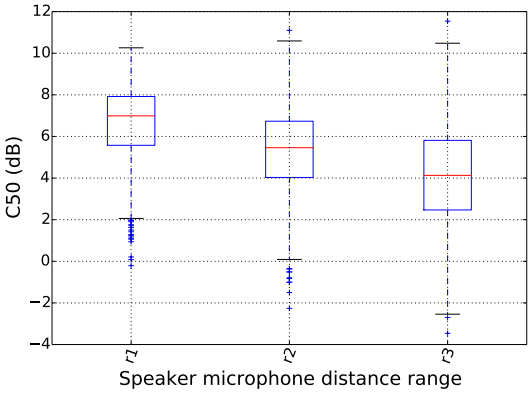
(a) C_{50} *histogram.*



(b) *Distribution of* C_{50} *by microphone channel.*



(c) *Distribution of* C_{50} *by speaker location.*

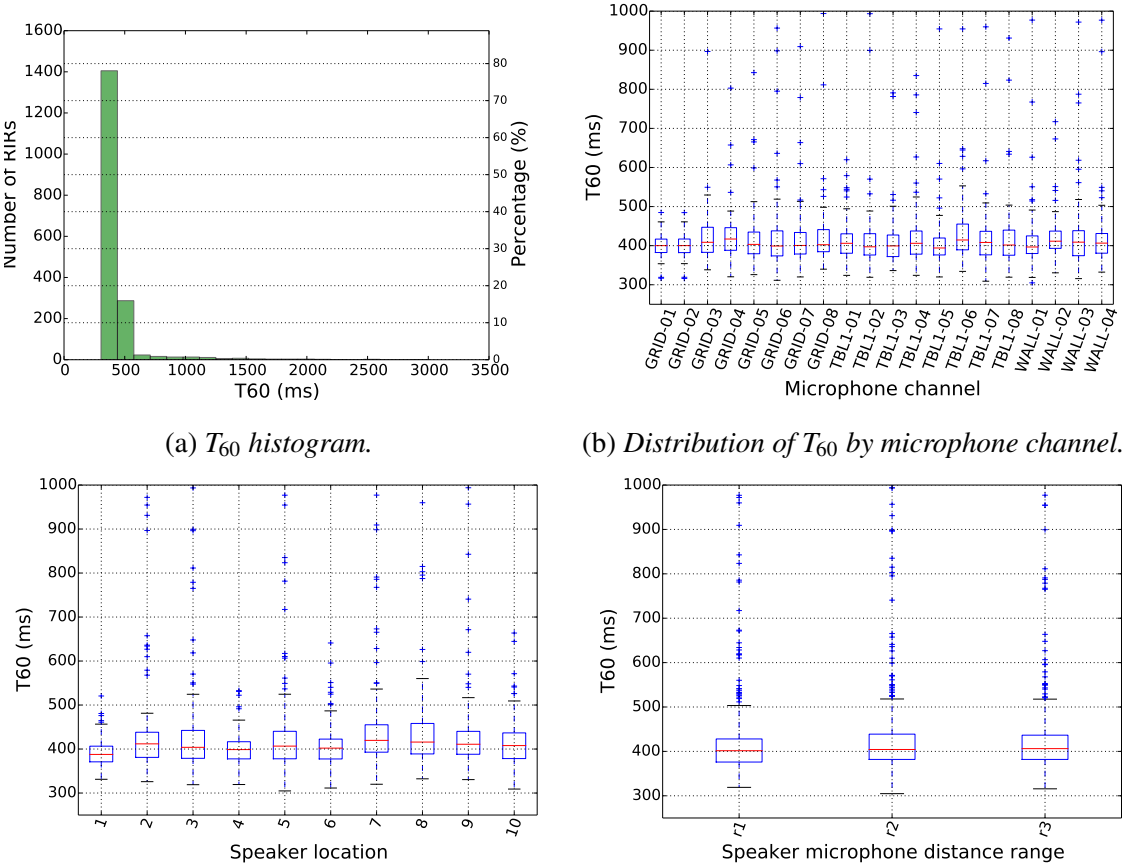


(d) *Distribution of* C_{50} *by speaker distance to the table (r1: 0.15m, r2: 0.45m and r3: 0.75m).*

Fig. 5.12 C_{50} *statistics: reverberation level variation with speaker location and microphone location.*

installation. In addition, the C_{50} variation caused by speaker movement (see Fig. 5.12c) is larger than the microphone difference (see Fig. 5.12b). In comparison, there is much smaller variation in reverberation time T_{60} caused by microphone configuration and speaker movement, as shown in Fig. 5.13. This is because reverberation is by definition a statistic metric of the overall sound energy decaying speed in a given room, and it is less sensitive to the location of microphone and sound source than the early-to-late reverberation ratio such as C_{50}.

The second aspect is about the average WERs on the recordings from each microphone. To avoid channel mismatch, the acoustic model is trained with the data from all microphone channels, *i.e.* multi-condition training with real recordings from all 20 distant microphones. Because multi-condition training increases training data by 20 times compared to SDM recording based acoustic model training, to avoid over-fit caused by 20 times replication of the same utterances, the adaptation system described in Section 4.4.2 is adopted. The



(a) T_{60} *histogram.*

(b) *Distribution of* T_{60} *by microphone channel.*

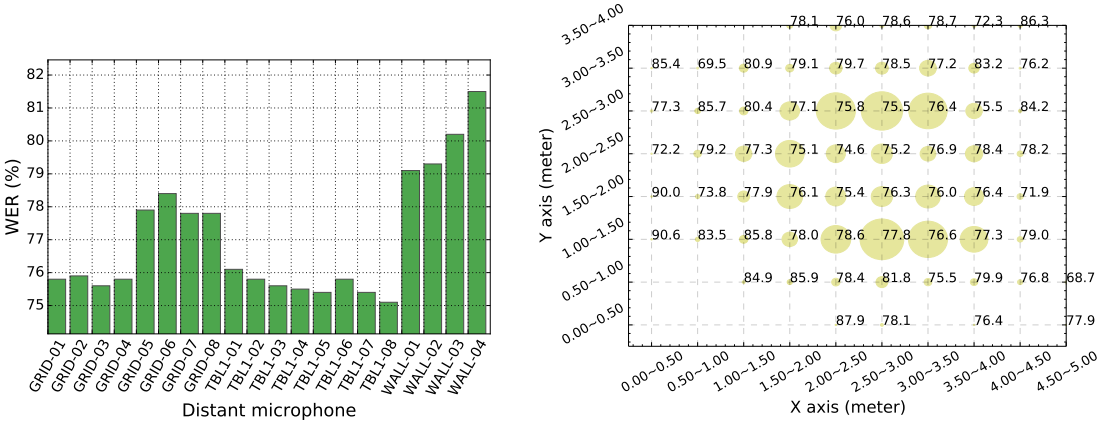
(c) *Distribution of* T_{60} *by speaker location.*

(d) *Distribution of* T_{60} *by speaker distance to the table (r1: 0.15m, r2: 0.45m and r3: 0.75m).*

Fig. 5.13 T_{60} *statistics: reverberation level variation with speaker location and microphone location.*

DNN front-end from the AMI corpus is adapted with the SWC1 recordings from all 20 microphones. The evaluation is conducted on the SWC2 recordings only because there is no gender mismatch between SWC1 and SWC2, while the gender mismatch between SWC1 and SWC3 could hide the impact of the microphone channel.

Fig. 5.14a illustrates the average WER on each microphone over all utterances in SWC2. The microphones in the circular array located at the center of the table ("TBL1-*") have the lowest average WERs, followed by the inner four microphones hanging from the grid ("GRID-01"∼"GRID-04"). There is a significant increase in WER from the inner 4 microphones hanging from ceiling grid to the outside 4 microphones hanging from the same ceiling grid ("GRID-05"∼"GRID-08"). The microphones located on the wall provides the highest average WER. Fig. 5.14b illustrates the average WER of utterances from all microphones given the speaker location. Since the utterance level WER gets very noisy in short utterances, only utterances with no less than 5 words are considered in the



(a) *Average WER per microphone using multi-condition acoustic model.*

(b) *Average WER by speaker location. The circular shadow size is proportional to the number of utterances at given speaker location.*

Fig. 5.14 *WER variation caused by microphone difference and speaker movement.*

calculation of the average WER per location area. Fig. 5.14b shows that the utterance level WER tends to increase when speakers move to the corner of the room.

In addition, as shown in Fig. 5.14b most utterances are spoken when the speaker is around the table. Therefore, the RIRs measurement configurations shown in Fig. 5.9 in Section 5.2.2 do not reflect the real movement of speaker. This explains the different ranks in the average microphone reverberation level by the average C_{50} as illustrated in Fig. 5.12b and average microphone quality by speech recognition performance in Fig. 5.14a. Fig. 5.12b indicates that on average the outer 4 microphones hanging from the ceiling grid ("GRID-05"∼"GRID-08") are the most advantageous. This is because the speaker location in RIR measurement covers a wide range of location far from the table. In comparison, the average WER per microphone shown in Fig. 5.14a suggests that the 8 microphones placed on the table are the most advantageous, closely followed by the 4 inner microphones hanging from ceiling grid. This is because the speakers stay mostly around the table and they often look downward at the table while talking.

Therefore, the optimal microphone installation for distant speech recognition is not only dependent on the room environment, but more importantly on the speaker behaviour, particularly the speaker location, the speaker movement and the speaking direction in the real recordings.

5.3.3 Environment attributes and distributed microphone

As illustrated in Section 5.3.2, DSR performance is largely affected by speaker movement and microphone installation, and the optimal microphone installation is highly dependent

on the speaker movement. There are a few potential implicit environment attributes behind this observation, such as the asymmetric room arrangement which causes variation in local background noise and reverberation level. There are also attributes from the interaction among speakers, microphones and environment, such as the early-to-late reverberation ratio which varies by microphone location and speaker location. Fig. 5.12c confirms that the reverberation variation at different locations of the room, by the variation in the early-to-late reverberation ratio C_{50} when speakers move around. Furthermore, Fig. 5.14b has confirmed the variation in speech recognition performance when speaker moves around in the room. Therefore, an obvious question is whether DSR could benefit from distributed microphones by always picking the most advantageous microphone for recognition task given the current speaker movement.

Ideally the optimal microphone is the microphone that provides recordings of the lowest WER for given speech utterance. This is a post-recognition selection that requires the reference transcript and the recognition scoring results, therefore it is not practical. But it provides a ceiling performance for any microphone selection strategies, thus it is further referred to as the "oracle selection". The second selection strategy is based on the average distance between the speaker and microphone in the given utterance, *i.e.* "minimum distance selection". This strategy employs the speaker location tracking results from the Ubisense system. The third selection strategy is the "maximum posterior selection". This strategy takes the idea of confidence score based on DNN posteriors in an early publication by the author in a joint work with Zhang et al. (2014). That published work suggests that a low posterior from the DNN front-end implies a high level of confusion in DNN against the hypothesis labels with provided acoustic features. Therefore, the microphone channel with lowest overall posterior in one utterance often indicates a high level of confusion. Based on this idea the microphone with highest overall posterior from the DNN front-end at given utterance is selected as the optimal microphone, thus it is referred to as the "maximum posterior selection". For a comparison purpose, the performance based on the recogniser output voting error reduction (ROVER) proposed by Fiscus (1997) is also explored, because it is a typical and frequently used system combination strategy for speech recognition. The confidence score used in ROVER is from the word level posterior calculated in a similar way with the utterance level posterior used in the "maximum posterior selection".

To avoid the training-test mismatch in terms of microphone channel, utterance level selection is conducted on 20 channel parallel decoding results from the adaptation system based on the multi-condition training which has been used in Section 5.3.2. The performance of different channel selection methods is compared on SWC2 in Table 5.8. Ideally with the oracle selection, the channel combination could bring down the WER by more than 10% absolute. However that can rarely be achieved. The minimum distance

Table 5.8 *WER based on distributed microphone selection using different strategies.*

	Microphone channels	Sub.	Del.	Ins.	WER
Best channel	TBL1-08	45.8	26.2	3.2	75.1
Worst channel	WALL-04	33.2	47.5	0.8	81.5
ROVER	TBL1	36.4	35.0	1.8	73.2
	TBL1+GRID	32.1	39.2	1.3	72.7
	TBL1+GRID+WALL	28.9	43.3	1.0	73.2
oracle selection	TBL1	40.7	24.7	2.3	67.6
	TBL1+GRID	38.3	24.3	2.1	64.7
	TBL1+GRID+WALL	37.5	24.6	2.0	64.1
minimum distance selection	TBL1	45.6	26.4	3.2	75.2
	TBL1+GRID	45.3	26.0	3.2	74.5
	TBL1+GRID+WALL	45.1	26.2	3.1	74.5
maximum posterior selection	TBL1	42.8	29.4	2.5	74.8
	TBL1+GRID	40.9	32.2	2.1	75.3
	TBL1+GRID+WALL	37.0	38.2	1.4	76.7

based channel selection outperforms the maximum posterior selection when there is a large number of microphones, for example 16 microphones (TBL1+GRID) or 20 microphones (TBL1+GRID+WALL). However when there is limited number of microphones, particularly when these microphones are located close to each other thus geometrically confusing (TBL1), the maximum posterior selection slightly outperforms the minimum distance based selection. ROVER gives the best performance among the blind channel combination methods tried. It is worth emphasising that ROVER is strictly word level channel selection. Thus it is not completely comparable with other strategies which select microphone channel at utterance level. The better performance in ROVER seems to imply that utterance level microphone selection in practice is suboptimal, and running multiple systems in parallel for channel combination might be better than channel selection if there is sufficient computation resource.

It is worth mentioning that when the results in Table. 5.8 are compared with the results from the adaptation baseline in Table. 4.8, the WERs on SWC2 did not improve with multi-condition training compared to the SDM recording based acoustic model training. In addition, the weighted delay and sum beamforming on 8 channel circular array (TBL1) achieved an overall WER of 71.6% on SWC2, and this is better than any channel selection methods tried above and the ROVER channel combination.

A further analysis is conducted on the impact of speaker-microphone distance in DSR performance on single microphone recordings. Fig. 5.15, Fig. 5.16 and Fig. 5.17 illustrate the utterance level WER of the utterances spoken at given ranges of speaker-microphone distance in SWC2 recordings. Each circle in Fig. 5.15, Fig. 5.16 and Fig. 5.17 represents

one speech utterance. The x-axis of the circle center corresponds to the distance between speaker and corresponding microphone averaged over the whole utterance. The y-axis of the circle center corresponds to the WER of that utterance. The size of the circle is proportional to the number of words in that utterance. For most microphones, as the distance between speaker and microphone increases, the minimum utterance level WER increases. This trend is particularly clear with the microphones from the circular array on the table ("TBL1") and the inner 4 microphones in the "GRID" group, all of which have relatively lower overall WERs than the other microphones (shown in Fig. 5.14a). However, there is a very large variation in the utterance level WER at any given distance for any microphone. This may be the reason why there is only marginal benefit from the utterance level microphone selection based on the minimum speaker-microphone distance (see Table 5.8).

Furthermore, Fig. 5.18, Fig. 5.19 and Fig. 5.20 illustrate the median of utterance level WERs among utterances in each speaker-microphone distance range on each microphone. The circle in each plot represents one distance range. The x-axis coordinate of the circle center corresponds to the distance range, the y-axis coordinate of the circle center corresponds to the median value of utterance level WERs, and the size of the circle is proportional to the number of utterances in corresponding distance range. The median WER increases as the speaker-microphone distance increases for all 8 microphones from "TBL1" array located at the center of the table and the inner 4 microphones from the "GRID" group hanging from the ceiling grid. In comparison, the outer 4 microphones in the "GRID" array and the 4 microphones distributed on the wall ("WALL") does not provide nice correlation between WER and speaker-microphone distance, and the average distance between speaker and microphone is larger than the case with microphones in "TBL1" array and the inner 4 microphones in "GRID" array.

Compared to the inner 4 microphones in the "GRID" array, the outer 4 microphones overall have slightly larger distance to speakers and usually they are not in the talking direction. The microphones distributed on the wall have the farthest average distance to any speaker and these microphones are rarely in the talking direction. Such differences suggest that the correlation between speaker-microphone distance and average speech recognition performance is only significant when the speaker-microphone distance is within 1.5 meters and when room reflections and background noise are not dominant in the recorded signal. This explains why the distance based channel selection could not benefit much from many microphones distributed in the room, when speakers often just move around the table and look downward to the center of the table when talking.

5.4 Summary and Discussion

With SWC data as a study case, this chapter has performed a factor-by-factor experimental analysis regarding the challenges in DSR. The investigation is conducted step by step from the recognition performance on the headset recordings, followed by the recognition performance on the simulated distant speech recordings, to recognition performance on the real distant recordings from one or multiple microphones. The investigation uses the AMI corpus as a comparison to highlight the unique challenges in the real recordings of natural spontaneous conversational speech.

The analysis on the SWC headset recordings in Section 5.1 highlights a few challenges in the real recordings of natural spontaneous conversational speech: the very short utterance duration of 2.2 seconds on average and the emotional speech that have WER 5-10% higher in absolute value compared to normal speech. In addition, a comparison is conducted between the SWC data and the AMI corpus in terms of the average WER on utterances of different lengths. The results suggest that the current LM for SWC is suboptimal, and that the acoustic model of SWC would improve with more training data, particularly with more female speech data.

Following that, the analysis on simulated data in Section 5.2 investigates the impact of reverberation and overlapped speech, each factor alone and two factors in combination. While both reverberation and overlapped speech increases the WER, the impact of overlapped speech is larger than reverberation. It distorts both the acoustic model and the acoustic features. The WER in one utterance tends to increase as the percentage of overlapped speech in that utterance increases. When investigating the effectiveness of multi-channel based dereverberation algorithm GWPE, it is found that GWPE could only reduce 2.7% absolute WER while the reverberation distortion alone has increased the WER by 7.6% absolute. When there is overlapped speech, the improvement from GWPE shrinks further. A further experimental investigation on using different microphone arrays for dereverberation suggests that the optimal microphone combination for dereverberation might not be comprised of the microphones with the best performance as individual. In most applications the array with microphones close to each other are preferred over the group of distributed microphones. This can be attributed to the fact that the performance improvement from dereverberation tends to be less robust compared to the beamforming algorithms, and the beamforming algorithms prefer a small distance between microphones within an array.

In the further analysis in Section 5.3, the findings from analysis based on headset recordings and simulated data are verified on real recordings. Similar to the headset recordings, a WER descending trend as utterance length increases is observed in the SDM

recording based recognition results. In addition, the WER increases when the percentage of overlapped speech in one utterance increases, and when the speech is emotional with vocal sound such as laughs. Compared to the headset recordings, the distant recordings are more sensitive to the speaking rate, as the recognition results illustrate a more obvious growth in WER when the speaker rate increases above 2 words per second. With the speaker location tracking from the Ubisense system, it is found that most utterances are spoken when speakers are around the table and when they are likely to be facing downward. Such speaker behaviour makes the 8 microphones in the circular array located at the center of the table particularly advantageous, compared to any other microphones installed in the room. The pre-recording RIR measurement suggests that the most advantageous microphones are the four inner microphones hanging from the ceiling grid when speaker moves evenly in the room. However, such assumption on speaker location and movement is very different from the reality in the SWC recordings, suggesting that the optimal microphone installation is highly dependent on the targeted speaker activities. Because the speakers in the SWC recordings move around the table most of the time, there is no benefit found from combining recordings from microphones distributed around the room. Instead, the best overall performance on real distant recordings are achieved with multi-channel dereverberation and beamforming using the 8 channel circular microphone array located at the center of the table.

This chapter has covered a lot of factors in both headset recordings and distant recordings of real natural spontaneous conversational speech. However, one very important factor is skipped in the work of this chapter, *i.e.* the background noise. The background noise in SWC distant recordings includes the stationary noise such as the computer fan noise, as well as the non-stationary noise such as the cracking sound of the wood floor under the carpet caused by foot steps when speakers move around, dice noise in the game, occasional traffic noise from outside of the window and babble noise in a few recording sessions with invited viewers. The diversity in noise type and the complexity in noise statistics make a thorough analysis on background noise alone a significant amount of work. Therefore, the investigation on background noise is not covered by this work, and it can be considered for future work.

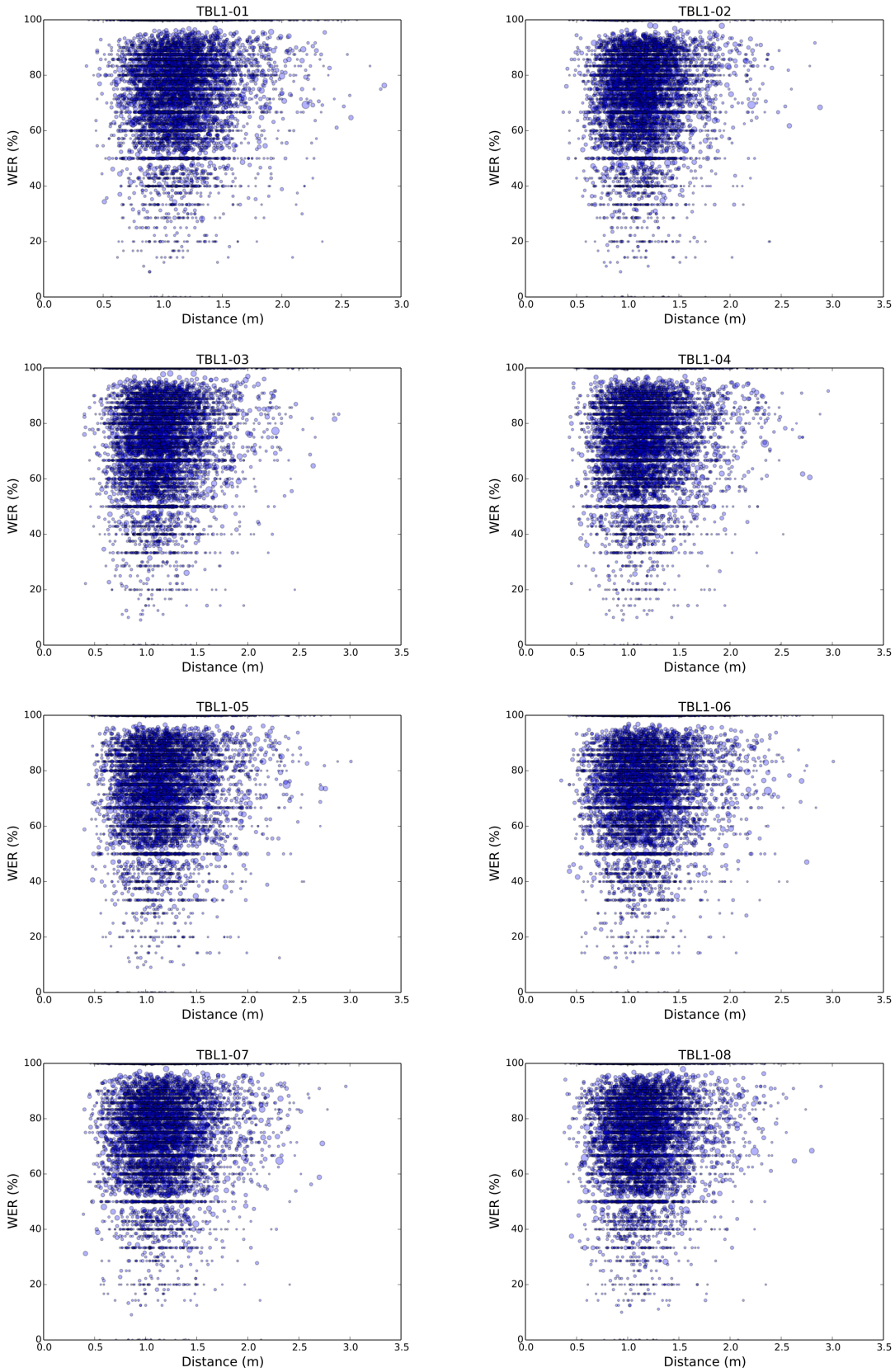


Fig. 5.15 *Utterance level WER as speaker-microphone distance changes - TBL1 microphones.*

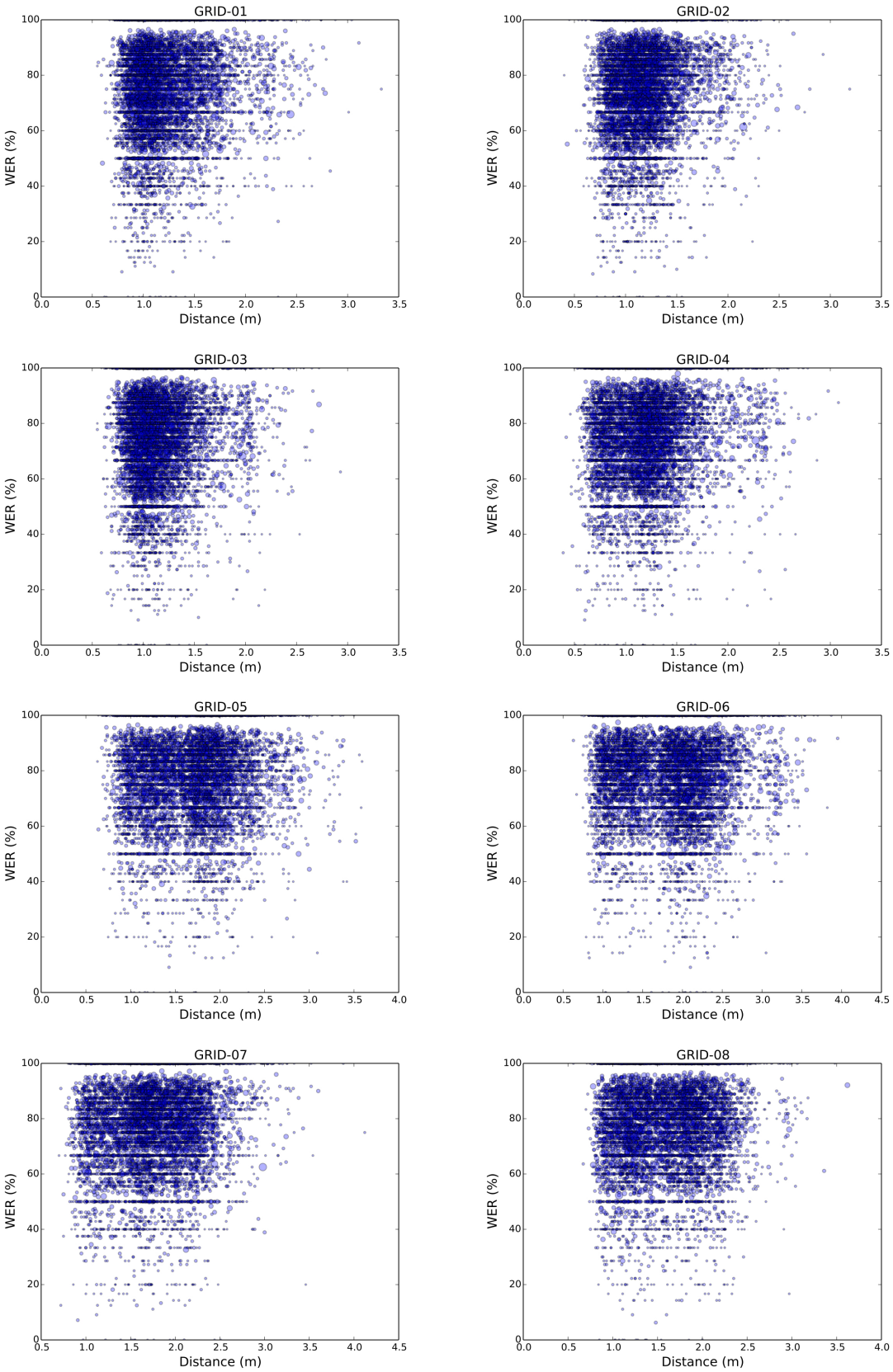


Fig. 5.16 *Utterance level WER as speaker-microphone distance changes - GRID microphones.*

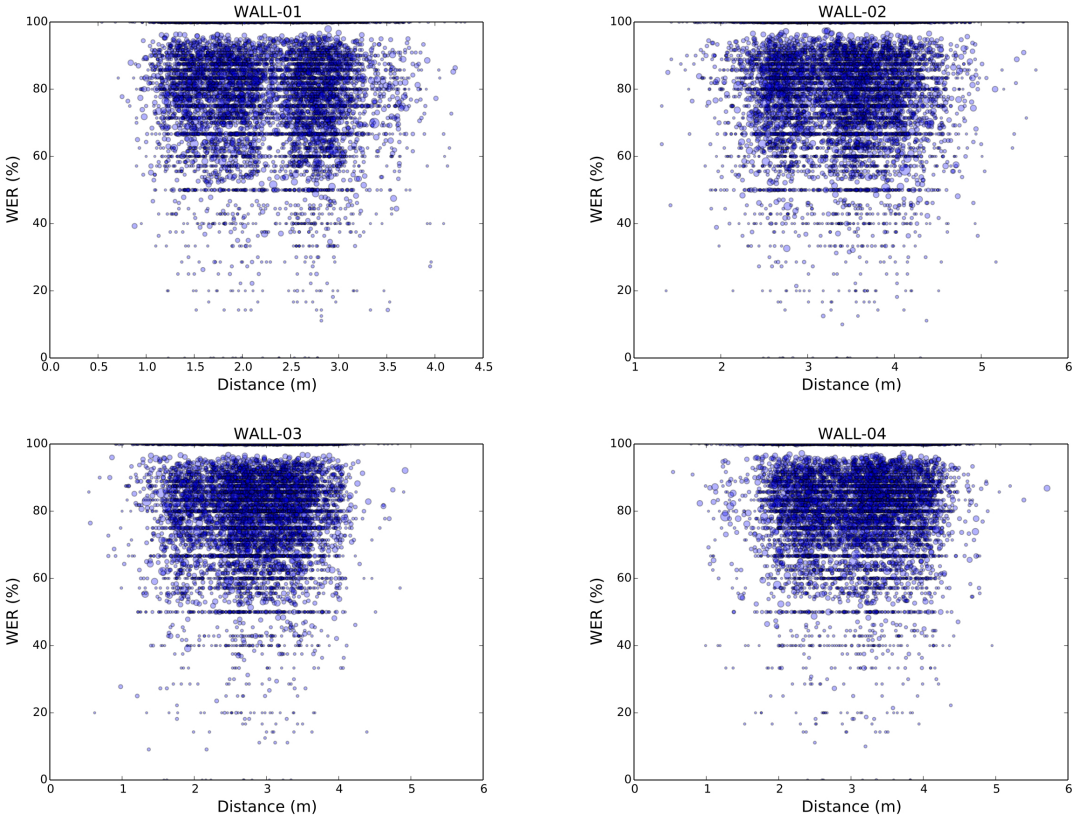


Fig. 5.17 *Utterance level WER as speaker-microphone distance changes - WALL microphones.*

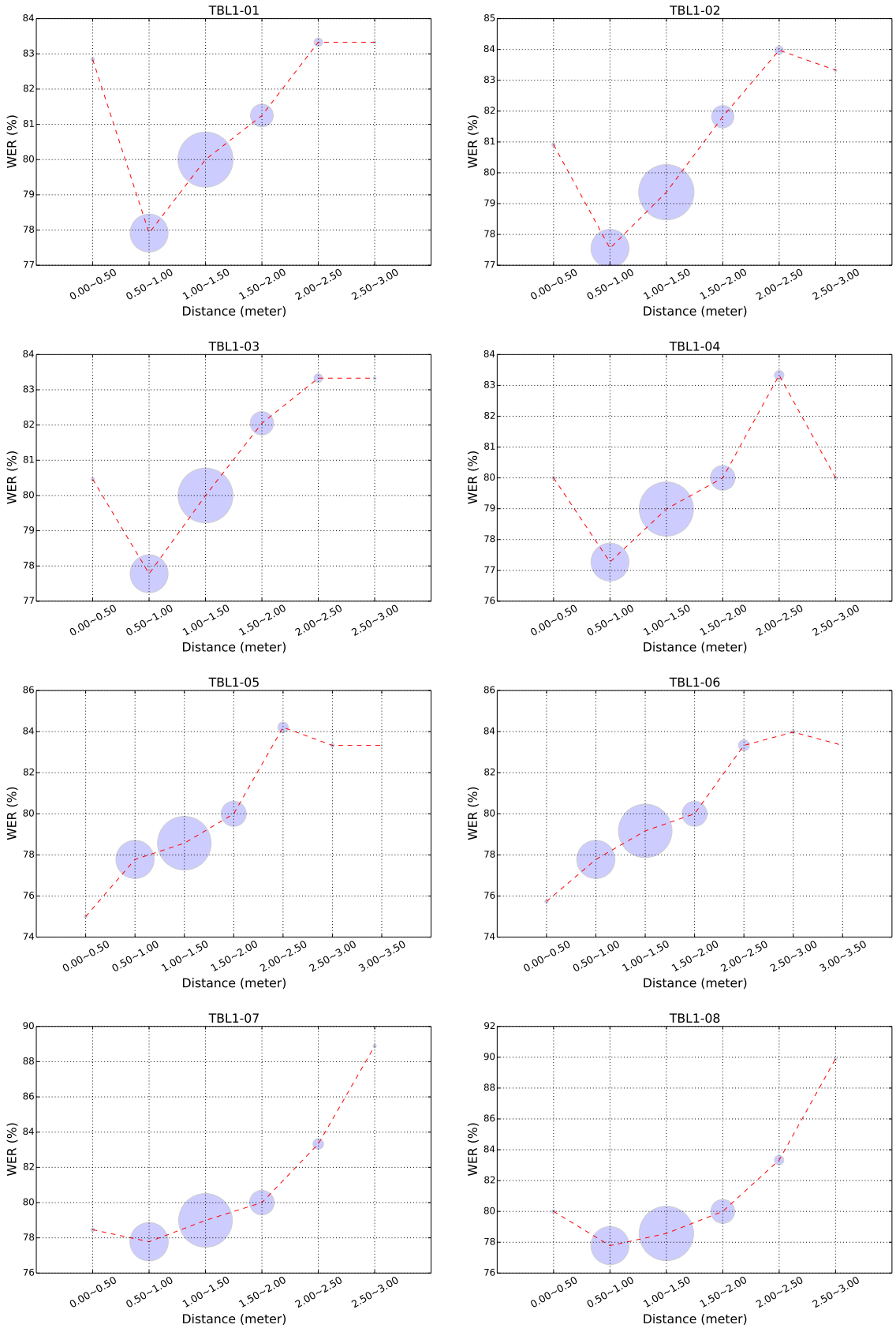


Fig. 5.18 *Average WER as speaker-microphone distance changes - TBL1 microphones.*

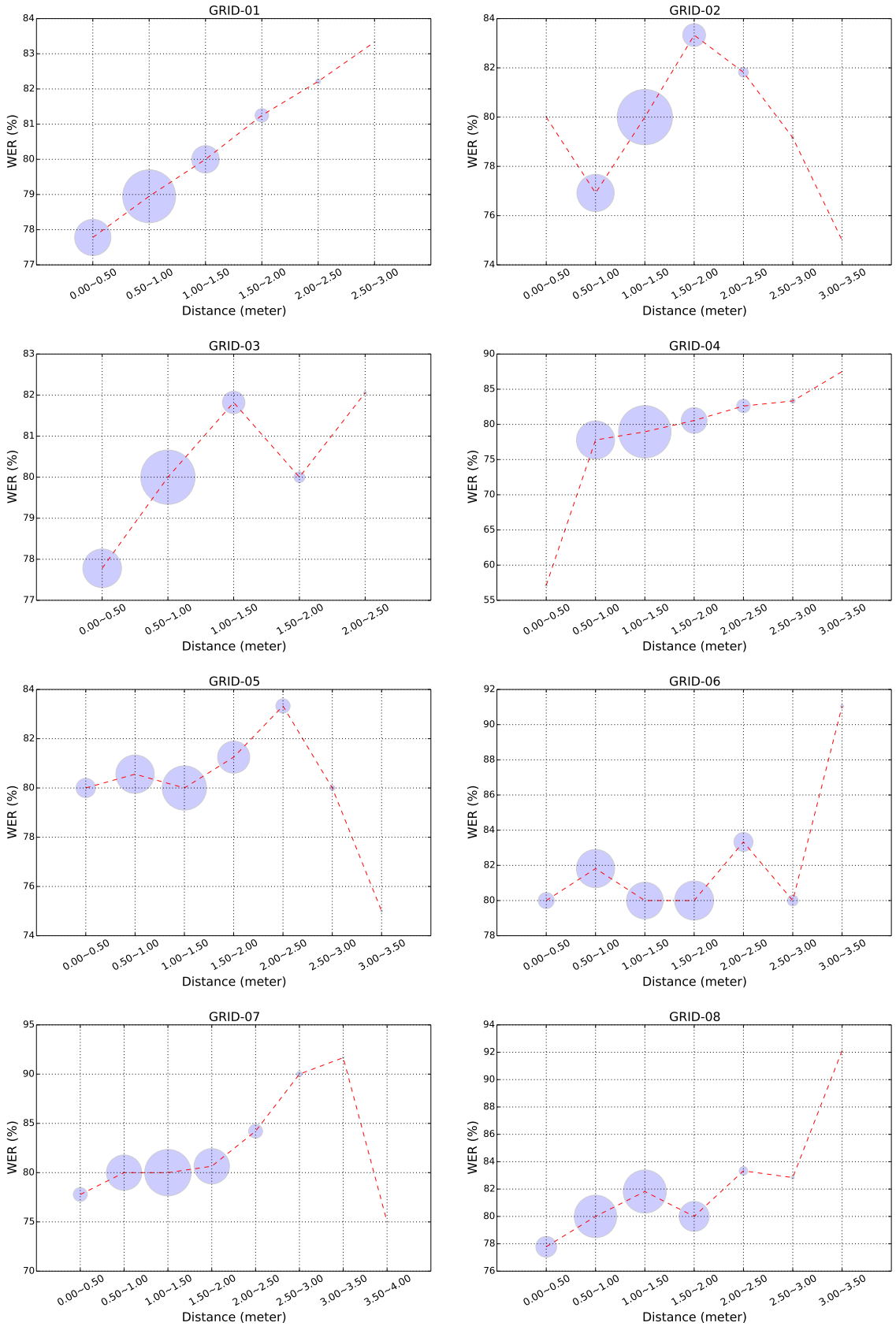


Fig. 5.19 *Average WER as speaker-microphone distance changes - GRID microphones.*

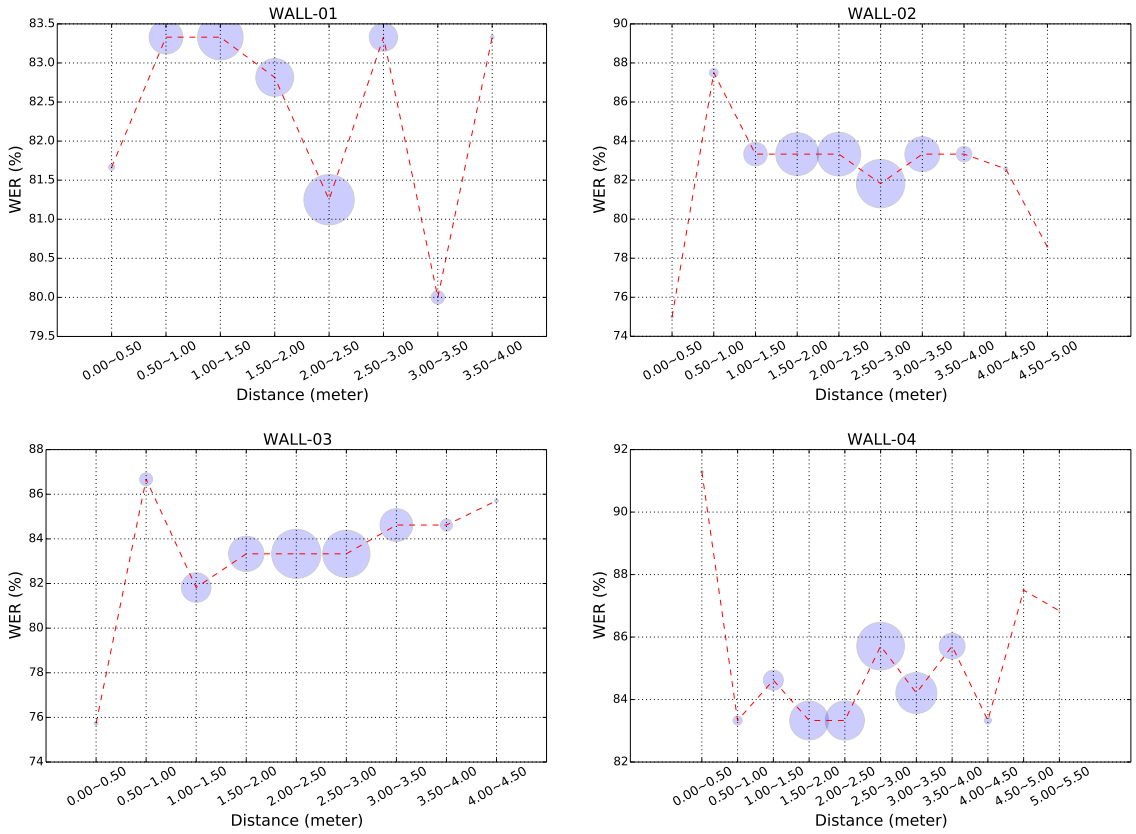


Fig. 5.20 *Average WER as speaker-microphone distance changes - WALL microphones.*

Chapter 6

Reverberation Modelling for Distant Speech Recognition

Contents

6.1 Complex Spectrogram Based Reverberation Modelling 112

6.2 The Local Phase and Magnitude Assumptions 117

6.3 Experimental Evaluation 119

6.3.1 The Local linear phase assumption 119

6.3.2 Local stationary magnitude assumption 120

6.3.3 Reverberation modelling for speech recognition 122

6.3.4 Reverberation variation analysis with the RIR spectrogram . . . 127

6.4 Summary and Discussion 131

Besides the background noise and overlapped speech, reverberation is one major factor in the distant recordings which limits the recognition performance. As discussed in Section 3.3, existing reverberation modelling at signal level via room impulse response (RIR) is very sensitive to any acoustic changes in the reverberant environment and the recording channel. The acoustic change that changes RIR however does not necessarily change the distant speech recognition (DSR) performance. For DSR it is the change in the feature pattern that directly impacts the recognition performance, as some variations at signal level can be normalised in the DSR front-end. Therefore the reverberation modelling based on the speech recognition feature is more important for understanding and estimating the impact of reverberation distortion to the DSR performance, as well as the treatment of the feature pattern distortion caused by reverberation.

Existing reverberation modelling based on speech features could not bring significant improvement to the reverberation robustness of DSR systems, because the feature pattern constructed with reverberation modelling tends to be oversimplified (Sehr and Kellermann, 2008, 2009). The work in this chapter investigates the impact of reverberation on the complex spectrogram of speech signal. The complex spectrogram based on the short time Fourier transform (STFT) is the building block of several popular features used in the state-of-the-art DSR systems, such as the Mel frequency cepstral coefficient (MFCC) and the logarithmic Mel filter bank coefficient. In addition recent research starts to use the spectrogram directly as the input to the deep network based front-end and acoustic models (Xiao et al., 2016). Therefore the reverberation modelling based on complex spectrogram will be of wide interest for many advanced DSR configurations. The goal of the investigation in this chapter is to provide an insight to the root of the feature distortion problem caused by reverberation, and to propose an accurate reverberation modelling based on the complex spectrogram.

This chapter is organized in the following order. Section 6.1 investigates the reverberation distortion on the complex spectrogram and further proposes the reverberation modelling for the frame level complex spectrogram. Section 6.2 discusses the two assumptions introduced by the proposed reverberation modelling and proposes the analytic formula to evaluate the errors introduced by the two assumptions. Section 6.3 reports the results of the experiments that evaluate the two assumptions and the reverberation modelling in speech recognition tasks. In the end, Section 6.4 summarizes the major findings in this chapter and discusses some influence factors not covered by this chapter.

6.1 Complex Spectrogram Based Reverberation Modelling

The STFT based spectrogram is used in the front-end during the feature generation by many state-of-the-art speech recognition systems. Two examples of the widely used speech recognition features are the MFCC and logarithmic Mel filter bank coefficient. This section performs an analytic investigation on how reverberation changes the speech spectrogram. The investigation does not further specialize on any specific features on top of spectrogram, so that the findings could generalize in many different feature configurations.

To minimise the confusion, the mathematical notations are introduced first. Denote the clean headset recording as $x(n)$, and the distant microphone recording of the same speech signal as $y(n)$, where n is the discrete time index or the sample index. The STFT of the

headset recording at time τ is thus

$$X(\tau,k) = \sum_{n=\tau-N+1}^{\tau} w(n-\tau)x(n)e^{-\frac{j2\pi k(n-(\tau-N+1))}{N}}, \tag{6.1}$$

where $w(n)$ is a window with N non-zero coefficients. The window slides over time thus truncates the recordings into frames on which the STFT for each frame is calculated independently. Thanks to the window function, the frequency leakage caused by the truncation is neglectable. N is also the number of frequency bins in the discrete Fourier transform (DFT) output with k as the frequency index. Similarly, the STFT of distant microphone recording is

$$Y(\tau,k) = \sum_{n=\tau-N+1}^{\tau} w(n-\tau)y(n)e^{-\frac{j2\pi k(n-(\tau-N+1))}{N}}. \tag{6.2}$$

It is worth emphasising that τ is a discrete time index similar to n. Its value reflects the updating rate of the STFT calculation when the recording is longer than the STFT analysis window. If the STFT is updated at every new sample in an extreme case, the time index τ increases by the same step size with n, *e.g.* 1. If the STFT is updated every l samples, the time index τ increases by a step size l times larger than n does. In most implementations of the ASR front-end, the frame size is 25-30 ms and the frame shift is 10 ms, *i.e.* the frame rate is 100 Hz and the STFT is updated every 10 ms.

Assume that there is no background noise or any other sound source, the only difference between the distant recording $y(n)$ and the headset recording $x(n)$ of the same signal is the presence of reverberation in the distant recording $y(n)$. Assume that reverberation could be approximated with a finite impulse response (FIR) filter in the time domain, namely the room impulse response (RIR):

$$\mathbf{h} = [h_0, h_1, ..., h_{M-1}]^T \tag{6.3}$$

where M is the effective length of the RIR that provides sufficient approximation accuracy. Therefore the distant recording $y(n)$ could be reconstructed with the clean headset recording $x(n)$ and the RIR $\mathbf{h}$ via

$$y(n) = \sum_{m=0}^{M-1} h_m x(n-m). \tag{6.4}$$

Substitute Eq. (6.4) into Eq. (6.2),

$$Y(\tau,k) = \sum_{n=\tau-N+1}^{\tau} w(n-\tau)\left(\sum_{m=0}^{M-1} h_m x(n-m)\right)e^{-\frac{j2\pi k(n-(\tau-N+1))}{N}}$$

$$
\begin{aligned}
&= \sum_{m=0}^{M-1} h_m \Bigg(\sum_{n=\tau-N+1}^{\tau} w(n-\tau)x(n-m)e^{-\frac{j2\pi k(n-(\tau-N+1))}{N}} \Bigg) \\
&= \sum_{m=0}^{M-1} h_m \Bigg(\sum_{n-m=(\tau-m)-N+1}^{\tau-m} w((n-m)-(\tau-m))x(n-m) \cdot \\
&\qquad\qquad e^{-\frac{j2\pi k((n-m)-(\tau-m-N+1))}{N}} \Bigg).
\end{aligned}
$$

Replacing index with $l = n-m$, with Eq. (6.1), it yields

$$
\begin{aligned}
Y(\tau,k) &= \sum_{m=0}^{M-1} h_m \Bigg(\sum_{l=(\tau-m)-N+1}^{\tau-m} w(l-(\tau-m))x(l)e^{-\frac{j2\pi k(l-(\tau-m-N+1))}{N}} \Bigg) \\
&= \sum_{m=0}^{M-1} h_m X(\tau-m,k). \qquad (6.5)
\end{aligned}
$$

This suggests that reverberant recording complex spectrogram $Y(\tau,k)$ is a convolution of the RIR $\mathbf{h}$ and the headset recording complex spectrogram $X(\tau,k)$. In addition the convolution is carried out independently in each frequency bin. It is worth emphasising that such frequency-independent convolution is based on the complex spectrograms with the STFT calculation updated at every recording sample.

In many speech recognition front-end implementations, the pre-emphasis is performed on a windowed piece of signal before performing STFT. Since the pre-emphasis does not affect the convolutional assumption with the RIR as shown in Eq. (6.4), the pre-emphasis can not change the conclusion in Eq. (6.5) either. This can be easily proved. Denote the pre-emphasis coefficient as a and the pre-emphasised signal with " $\hat{}$ ". For the headset recording and the distant recording respectively,

$$
\begin{aligned}
\hat{x}(n) &= x(n) - ax(n-1) \\
\hat{y}(n) &= y(n) - ay(n-1).
\end{aligned}
$$

Then

$$
\begin{aligned}
\hat{y}(n) &= \sum_{m=0}^{M-1} h_m x(n-m) - a \sum_{m=0}^{M-1} h_m x(n-1-m) \\
&= \sum_{m=0}^{M-1} h_m \Big(x(n-m) - ax(n-1-m) \Big) \\
&= \sum_{m=0}^{M-1} h_m \hat{x}(n-m). \qquad (6.6)
\end{aligned}
$$

Similarly, any operation or signal internal structure that could be approximated with autoregressive models will not change the conclusion in Eq. (6.5).

Eq. (6.5) provides a way to model reverberation accurately with complex spectrogram. However Eq. (6.5) also suggests that the high modelling accuracy requires the STFT to be calculated at the signal sampling rate, which is impractical considering the computation cost. In addition, the new way of modelling reverberation based on complex spectrogram will be completely equivalent with the RIR based reverberation modelling in the time domain, thus having the same level of sensitivity to any acoustic changes occurring in signal. To address that, two assumptions are introduced below, so that some summation items in Eq. (6.5) could be merged and the STFT updating rate could be reduced:

- *The locally linear phase assumption*: the STFT phase of the clean speech signal changes linearly within a very small temporal window, at a constant speed independent from speech content but dependent on frequency.

- *The locally stationary magnitude assumption*: the STFT magnitude of the clean speech signal does not change within a very small temporal window.

The justification of these two assumptions will be detailed in Section 6.2. Here it is first illustrated how the two assumptions simplify the reverberation modelling with the STFT updated at a reduced rate.

Assume that the locally linear phase assumption and the locally stationary magnitude assumption hold for any continuous N_{f} samples with sufficiently low errors. It is of particular interest when N_{f} equals the number of samples corresponding to 10 ms, *i.e.* the frequently adopted frame shift size in the speech recognition front-end. Denote the unwrapped phase of the clean speech STFT as:

$$\theta(\tau,k) = \angle X(\tau,k), \tag{6.7}$$

the linear phase assumption could be formulated as

$$\begin{aligned} \triangle\theta(\delta,k) &= \theta(\tau+\delta,k) - \theta(\tau,k) \\ &\approx 2\pi \cdot \frac{k}{N} \cdot \delta \\ &= \frac{2\pi k}{N} \cdot \delta \end{aligned} \tag{6.8}$$

or

$$\frac{\partial \angle X(\tau,k)}{\partial \tau} = \lim_{\delta\to 0}\left(\frac{\theta(\tau+\delta,k)-\theta(\tau,k)}{\delta}\right) = \lim_{\delta\to 0}\left(\frac{\triangle\theta(\delta,k)}{\delta}\right) \approx \frac{2\pi k}{N}. \tag{6.9}$$

Introduce $M_{\mathrm{f}} = \lceil M/N_{\mathrm{f}} \rceil$ to simplify the notation where $\lceil \cdot \rceil$ refers to the ceiling function. If N_{f} equals the number of samples in one frame shift, M_{f} equals the number of frames corresponding to the duration of RIR. Therefore Eq. (6.5) can be simplified:

$$\begin{aligned} Y(\tau,k) &\approx \sum_{m=0}^{M_{\mathrm{f}}-1} \left|X(\tau - mN_{\mathrm{f}},k)\right| \sum_{\delta=0}^{N_{\mathrm{f}}-1} h_{mN_{\mathrm{f}}+\delta} \cdot e^{j\theta(\tau-mN_{\mathrm{f}}-\delta,k)} \\ &= \sum_{m=0}^{M_{\mathrm{f}}-1} X(\tau - mN_{\mathrm{f}},k) \sum_{\delta=0}^{N_{\mathrm{f}}-1} h_{mN_{\mathrm{f}}+\delta} \cdot e^{j\triangle\theta(-\delta,k)} \\ &= \sum_{m=0}^{M_{\mathrm{f}}-1} X(\tau - mN_{\mathrm{f}},k) \sum_{\delta=0}^{N_{\mathrm{f}}-1} h_{mN_{\mathrm{f}}+\delta} \cdot e^{-\frac{j2\pi k\delta}{N}} \\ &= \sum_{m=0}^{M_{\mathrm{f}}-1} X(\tau - mN_{\mathrm{f}},k) H(m,N_{\mathrm{f}},k), \end{aligned} \tag{6.10}$$

where

$$H(m,N_{\mathrm{f}},k) = \sum_{\delta=0}^{N_{\mathrm{f}}-1} h_{mN_{\mathrm{f}}+\delta} \cdot e^{-\frac{j2\pi k\delta}{N}} \tag{6.11}$$

is actually the STFT of the RIR calculated every N_{f} samples, using a square window with an effective size of N_{f} without frame overlap between two sequential calculations. Eq. (6.10) indicates that the frame level complex spectrogram of the reverberant signal can be approximated with the convolution between the frame level complex spectrogram of the clean speech signal and the special complex spectrogram of the RIR. Therefore Eq. (6.10) is further referred to as the reverberation modelling based on frame-level complex spectrogram.

There are two important insights from this derivation. First, while existing literature suggests that reverberation causes distortion because the analysis window for STFT is shorter than the RIR (Raut et al., 2006; Sehr and Kellermann, 2008, 2009; Sehr et al., 2006), the above derivation shows that the reverberation distortion is mainly a consequence of the temporal change in the speech magnitude spectrum. Even though the analysis window is not shorter than the RIR, if the speech magnitude spectrum changes at such a fast rate that within the RIR duration the STFT for speech spectrogram has to be updated multiple times, the reverberation smearing will still exist. Here is one example: assume a scenario where the RIR is effectively 25 ms long and a 25 ms window is used for the STFT. Since the speech magnitude spectrum changes so fast, the STFT has to be updated every 10 ms, which means within 25 ms there will be multiple STFT calculations. In this example case there will still be reverberation distortion in the STFT spectrogram as well as in the features based on the STFT spectrogram. The actual degree of reverberation distortion is dependent

on the parameter value of RIR. More discussion about the reverberation distortion level will be covered in Chapter 7.

The second insight is about the time resolution in reverberation modelling controlled by N_{f}, which has been omitted by existing research (Raut et al., 2006; Sehr and Kellermann, 2008, 2009; Sehr et al., 2006). The N_{f} in Eq. (6.10) emphasises the reverberation modelling error introduced by the two assumptions made, *i.e.* the locally linear phase assumption and the locally stationary amplitude assumption. Details about the two assumptions will be covered in Section 6.2.

In a special case where the sound signal has its magnitude spectrum unchanged over a duration that equals the RIR length, the N_{f} can be increased to be of the same length with the RIR length. Then Eq. (6.10) reduces to a multiplication between the STFT complex spectrum of the clean speech and the complex spectrum of the RIR. In this case the distortion caused by reverberation could be easily factorized as an additive component in the logarithmic magnitude spectrum, and the reverberation distortion can be compensated with a mean normalisation or an additive bias.

6.2 The Local Phase and Magnitude Assumptions

The derivation in previous section has introduced two assumptions regarding the phase and magnitude properties of the clean speech complex spectrogram over a very short period of time. This section proposes the methods to evaluate the errors introduced by the two assumptions and the experimental results will be reported later in Section 6.3.1 and Section 6.3.2.

In the locally stationary magnitude assumption, it is assumed that the speech spectrum magnitude does not change within a small range of time corresponding to N_{f} samples. To quantify the error introduced by this assumption, a metric is proposed based on the local variance of the STFT magnitude in each frequency bin:

$$\mu_{\mathrm{mag}}(m,k,N_{\mathrm{f}}) = \frac{1}{N_{\mathrm{f}}} \sum_{n=mN_{\mathrm{f}}}^{(m+1)N_{\mathrm{f}}-1} \left|X(n,k)\right| \tag{6.12}$$

$$v_{\mathrm{mag}}(m,k,N_{\mathrm{f}}) = \frac{1}{N_{\mathrm{f}}} \sum_{n=mN_{\mathrm{f}}}^{(m+1)N_{\mathrm{f}}-1} \Big(\left|X(n,k)\right| - \mu_{\mathrm{mag}}(m,k,N_{\mathrm{f}})\Big)^2. \tag{6.13}$$

The variance is then normalised by the average STFT energy in corresponding frequency bin as a metric for the error introduced by the locally stationary magnitude assumption:

$$e_{\text{mag}}(k,N_{\text{f}}) = 10\log_{10}\left(\frac{\sum_{m=0}^{\lfloor T/N_{\text{f}}-1\rfloor} v_{\text{mag}}(m,k,N_{\text{f}})}{\sum_{m=0}^{\lfloor T/N_{\text{f}}-1\rfloor} \mu_{\text{mag}}(m,k,N_{\text{f}})^2}\right) \tag{6.14}$$

where T is the overall number of samples in the speech recording.

In the locally linear phase assumption, it is assumed that the speech spectrum phase changes linearly within a small range of time corresponding to N_{f} samples, and the change speed is only dependent on frequency. The accuracy of the linear phase assumption is estimated indirectly with the Pearson product-moment correlation coefficient (Pearson, 1895). The Pearson linear correlation is estimated between the discrete time index τ and the unwrapped phase of STFT $\angle X(\tau,k)$ within a short time span of N_{f} samples, *i.e.* $\tau \in [mN_{\text{f}},(m+1)N_{\text{f}})$ for the m-th frame shift. Denoting such local Pearson correlation as $\rho_{\text{p}}(m,k,N_{\text{f}})$, it can be averaged over a large amount of data:

$$\bar{\rho}_{\text{p}}(k,N_{\text{f}}) = \frac{1}{\lfloor \frac{T}{N_{\text{f}}} \rfloor} \sum_{m=0}^{\lfloor T/N_{\text{f}}-1\rfloor} \rho_{\text{p}}(m,k,N_{\text{f}}). \tag{6.15}$$

In particular, to penalize the insignificant correlation, the linear correlation coefficient is set to 0 when the p-value is larger than 0.05, *i.e.*

$$\rho'_{\text{p}}(m,k,N_{\text{f}}) = \begin{cases} \rho_{\text{p}}(m,k,N_{\text{f}}), & \text{if } p \leq 0.05 \\ 0, & \text{if } p > 0.05 \end{cases} \tag{6.16}$$

$$\bar{\rho}'_{\text{p}}(k,N_{\text{f}}) = \frac{1}{\lfloor \frac{T}{N_{\text{f}}} \rfloor} \sum_{m=0}^{\lfloor T/N_{\text{f}}-1\rfloor} \rho'_{\text{p}}(m,k,N_{\text{f}}). \tag{6.17}$$

As shown in Eq. (6.14), Eq. (6.15) and Eq. (6.17), the error in the locally linear phase assumption and the locally stationary magnitude assumption is a function of the time resolution in reverberation modelling represented by N_{f}. When $N_{\text{f}} = 1$, there is no modelling error because Eq. (6.10) degenerates back to Eq. (6.5).

Within the context of speech recognition, a better method to indirectly evaluate the errors introduced by the two local assumptions is to compare the speech recognition performance using mismatched features for training and test. The features include those derived from reverberant signals directly, and those derived from constructed complex spectrogram via reverberation modelling using Eq. (6.10) given different temporal resolution N_{f}. Results and examples will be covered in Section 6.3.

6.3 Experimental Evaluation

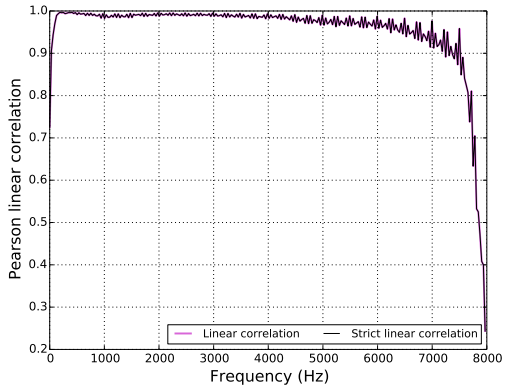
In this section, experiments are conducted to validate the locally linear phase assumption and the locally stationary magnitude assumption made in reverberation modelling. Section 6.3.1 and Section 6.3.2 employ the metrics proposed in Section 6.2 to evaluate the errors introduced by two assumptions, and Section 6.3.3 evaluates the reverberation modelling accuracy by speech recognition performance.

6.3.1 The Local linear phase assumption

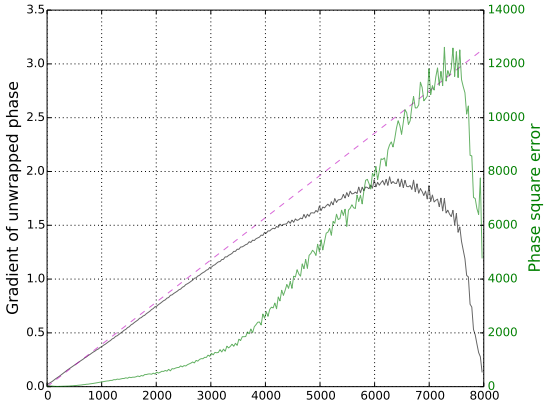
As discussed in Section 6.3.1, the local linearity in the phase spectrogram of the clean speech signal is evaluated with the Pearson product-moment correlation. The experiments are conducted on 200 speech utterances randomly selected from the evaluation dataset of SWC headset recordings in the "SA1" configuration (detailed in Section 4.4.1). The linear correlation coefficient is first averaged per utterance as shown in Eq. (6.15) and Eq. (6.17), and then averaged over all 200 utterances weighted by the utterance duration. The results are illustrated in Fig. 6.1.

As shown in Fig. 6.1a and Fig. 6.1c, the unwrapped phase of the clean speech STFT spectrum is approximately linear for most frequencies except the extremely low frequencies close to 0 Hz and the high frequencies close to the Nyquist frequency. The original Pearson coefficient (Eq. (6.15)) almost overlaps completely with the strict linear correlation coefficient (Eq. (6.17)). A comparison between Fig. 6.1a and Fig. 6.1c suggests that the local linear phase assumption holds better when N_{f} corresponds to 10 ms than 2.5 ms. This is because a larger value for N_{f} leads to more samples in each frame shift span where one linear correlation coefficient is calculated. Fig. 6.1b and Fig. 6.1d compares the gradient of the unwrapped phase estimated from linear regression (black line) with the gradient assumed in Eq. (6.9) (pink dash line), along with the linear regression error (green line). As shown, the assumption on phase gradient is sufficiently accurate for low and middle frequencies. The regression error (green line) increases as the frequency increases due to the fast phase change speed at high frequencies.

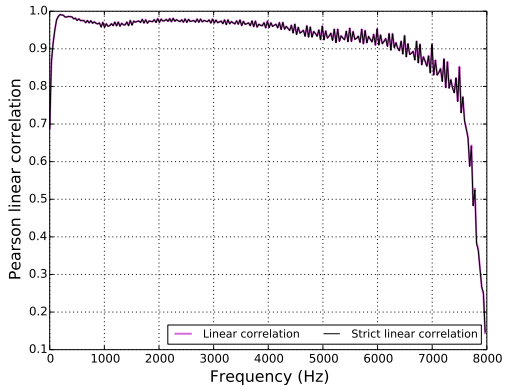
For a better understanding where the locally linear phase assumption breaks, Fig. 6.2 illustrates the magnitude and phase spectrogram from a small piece of headset recordings from the SWC data. As shown, the fluctuation in the gradient of the STFT unwrapped phase is associated with the fluctuation in the STFT magnitude. When there is an abrupt change in magnitude, there is also a fluctuation in the gradient of the unwrapped phase. In the complex spectrogram region with slow and smooth change in magnitude, the unwrapped STFT phase increases linearly with an approximately constant gradient.



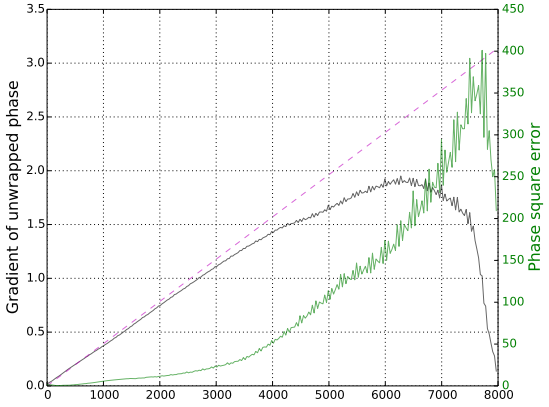
(a) *Pearson linear correlation between unwrapped STFT spectrum phase and the time index across different frequencies (N_f corresponds to 10 ms).*



(b) *Gradient from linear regression (black line) in comparison to assumed gradient $\frac{2\pi k}{N}$ (pink dash line), along with the regression error (N_f corresponds to 10 ms).*



(c) *Pearson linear correlation between unwrapped STFT spectrum phase and the time index across different frequencies (N_f corresponds to 2.5 ms).*



(d) *Gradient from linear regression (black line) in comparison to assumed gradient $\frac{2\pi k}{N}$ (pink dash line), along with the regression error (N_f corresponds to 2.5 ms).*

Fig. 6.1 *Validation of local linear phase assumption with 200 utterances randomly selected from "SA1" evaluation dataset of SWC headset recordings.*

6.3.2 Local stationary magnitude assumption

The locally stationary magnitude assumption is validated based on the variance-energy ratio in Eq. (6.14) using 200 utterances randomly selected from the evaluation dataset from the SWC headset recordings in the "SA1" configuration (detailed in Section 4.4.1). The variance-energy ratio is first calculated on each speech utterances and then averaged over all 200 utterances, weighted by the utterance duration.

Fig. 6.3 illustrates the average variance-energy ratio in each frequency bin for different N_f values. Overall for most frequencies the local temporal magnitude change in STFT

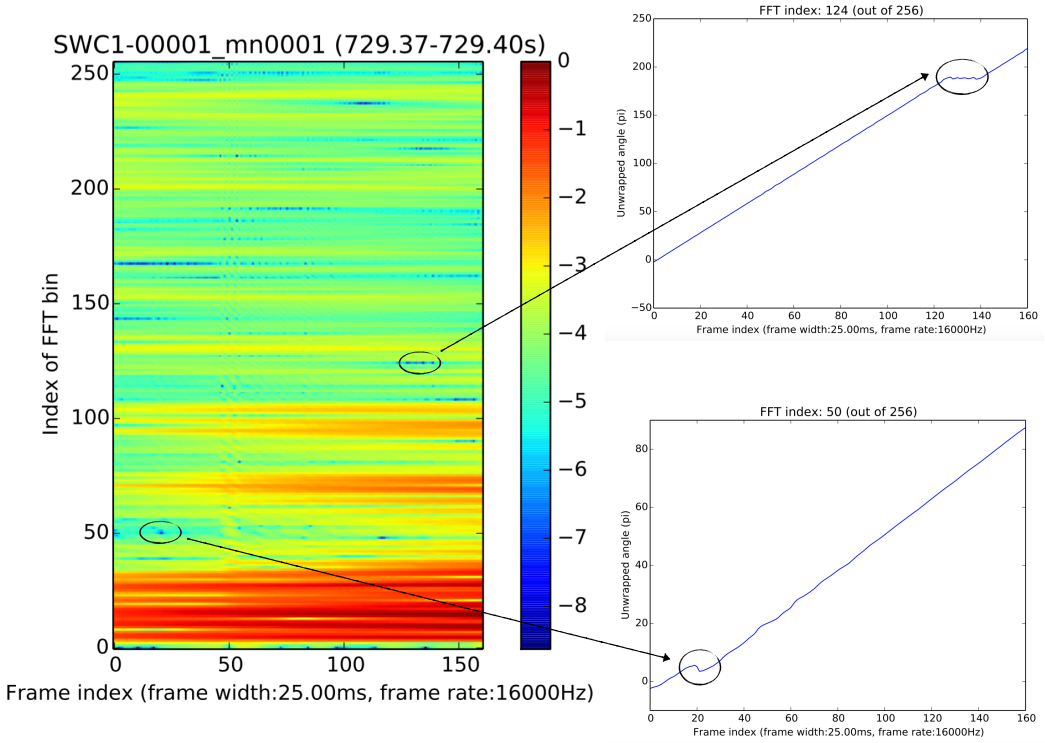
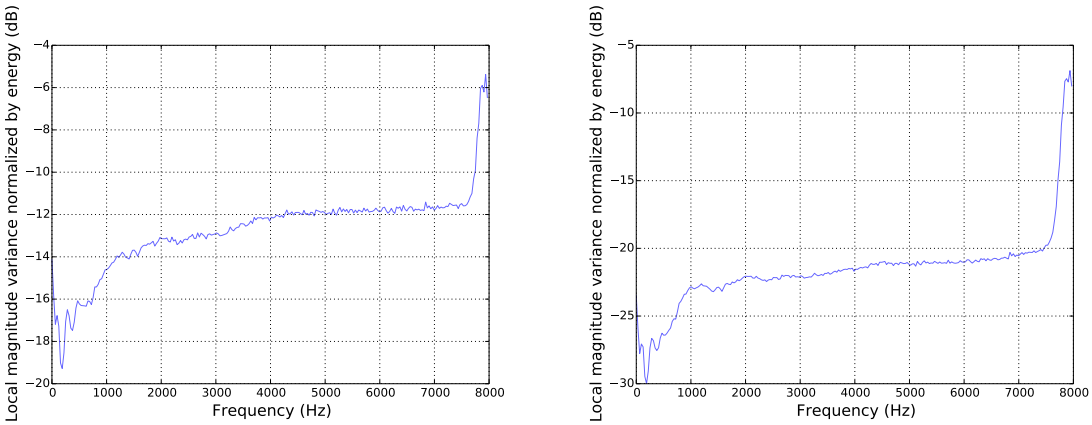


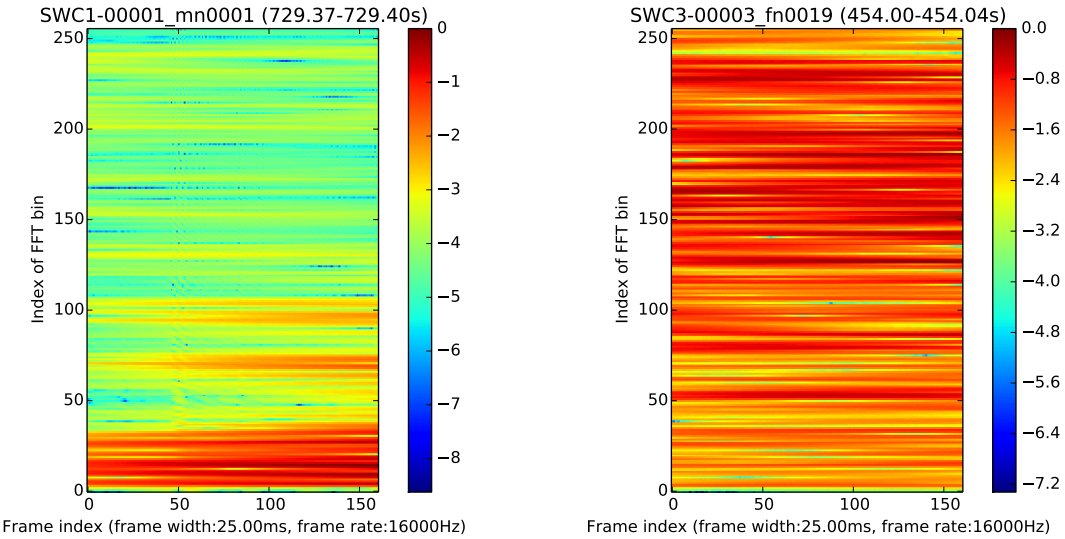
Fig. 6.2 *Magnitude (dB) and unwrapped phase of the complex spectrogram from a small proportion of speech sound (sampling rate: 16 kHz).*



(a) N_f *corresponds to 10 ms.* (b) N_f *corresponds to 2.5 ms.*

Fig. 6.3 *STFT magnitude change measurement via overall* e_{mag}.

complex spectrogram is very small compared to the spectrogram energy, thus the magnitude of clean speech could be assumed stationary within a small time span. In addition, the magnitude variation reduces as the frequency decreases and as the analysis time span shortens, namely N_f reduces. For a better understanding of the occasions where the assumption breaks badly, Fig. 6.4 illustrates the STFT magnitude spectrogram of two



(a) *Beginning 10 ms of a phoneme by a male speaker;*

(b) *Middle 10 ms of a phoneme by a female speaker.*

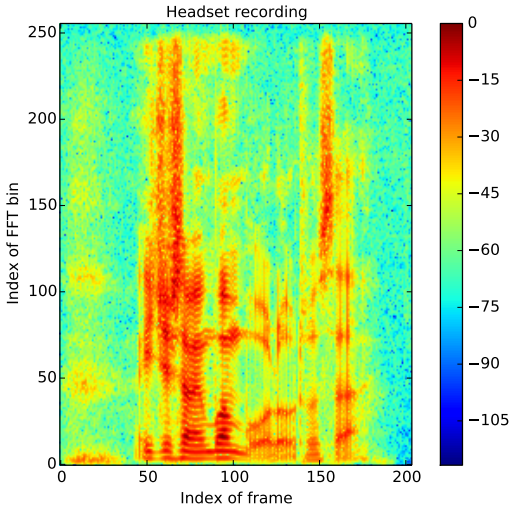
Fig. 6.4 *STFT amplitude change of speech signal (sampling rate: 16kHz) over 10 ms time span, with frame width being 25 ms (value in dB).*

pieces of audio. As shown, there is a larger magnitude change at the transition status from one phoneme to another phoneme (Fig. 6.4a), and at high frequency than at low frequency.

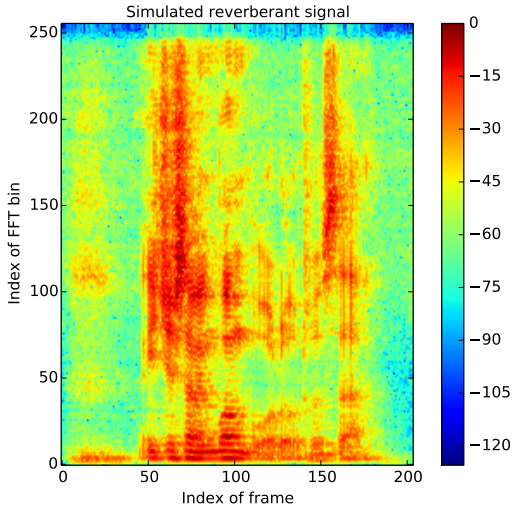
6.3.3 Reverberation modelling for speech recognition

The justification of the locally linear phase assumption (Section 6.3.1) and the locally stationary magnitude assumption (Section 6.3.2) has shown that the assumptions are reasonable for most regions in the complex spectrogram of the clean speech signal, except for the regions where the frequency is close to 0 Hz or to the Nyquist frequency and the regions at a transition stage from one phoneme to another.

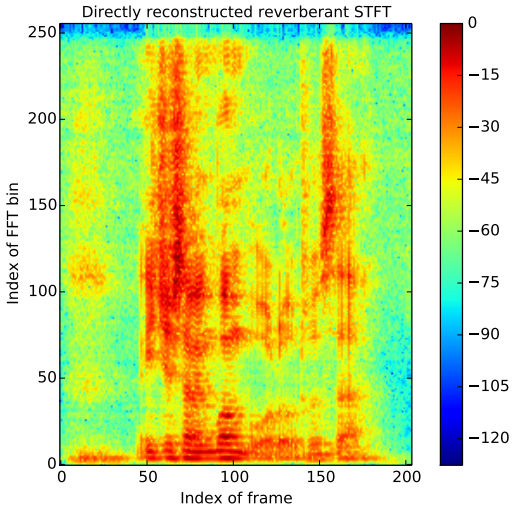
Fig. 6.5a shows the log magnitude spectrogram for one utterance taken from the SWC headset recordings. Fig. 6.5b illustrates the log magnitude spectrogram for the simulated reverberant signal of the same utterance. The reverberant signal is simulated by convolving the headset recording with the RIR measured in the SWC recording room. A comparison between Fig. 6.5a and Fig. 6.5b shows that reverberation decreases the temporal resolution of the pattern structure in the magnitude spectrogram in all frequencies, resulting in reverberation smearing. Fig. 6.5c illustrates the log magnitude of the complex spectrogram reconstructed with reverberation modelling (Eq. (6.5)), and Fig. 6.5d shows in dB the reconstruction error in the magnitude spectrogram when comparing Fig. 6.5c with Fig. 6.5b. In Fig. 6.5c, the x-axis corresponds to frame index τ in Eq. (6.5), the



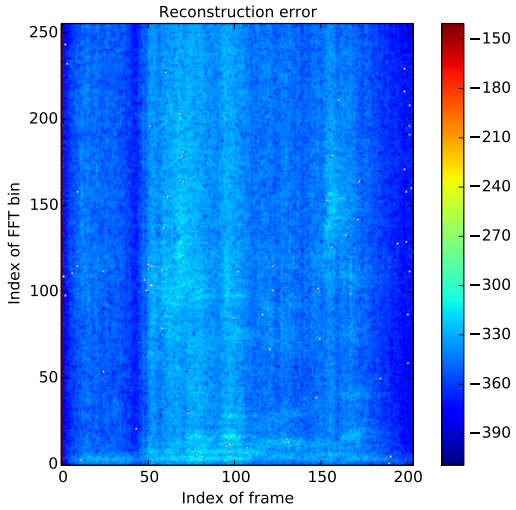
(a) *The magnitude spectrogram from headset recording speech signal;*



(b) *The magnitude spectrogram from simulated reverberant speech signal;*



(c) *The magnitude spectrogram reconstructed directly from the RIR and the complex spectrogram of the headset recording;*



(d) *The magnitude spectrogram of the reconstruction error in dB.*

Fig. 6.5 *The magnitude spectrogram with STFT updated at the signal sampling rate, corresponding to speech sentence: "Are you just having one warlord, what's that?" (dB)*

y-axis corresponds to frequency bin index k in Eq. (6.5), and the pixel value corresponds to $20\log_{10}|Y(\tau,k)|$.

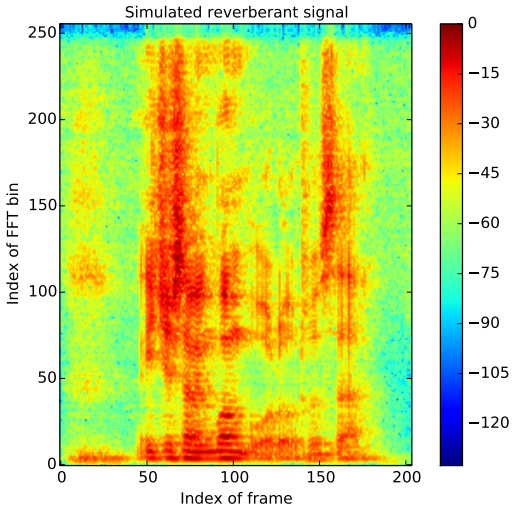
The very small errors in all regions of the magnitude spectrogram confirms Eq. (6.5) - the reverberant speech complex spectrogram could be accurately modelled with a convolution between the RIR and the clean speech complex spectrogram, and the convolution is conducted along the time index independently each frequency bin.

Fig. 6.6b, Fig. 6.6c and Fig. 6.6d illustrate the log magnitude spectrogram reconstructed with Eq. (6.10) with N_{f} corresponds to 10 ms, 5 ms and 1 ms. As shown, overall the reconstructed magnitude spectrogram preserves the patterns well in the magnitude spectrogram. The pattern approximation accuracy increases as N_{f} decreases because the temporal resolution of reverberation modelling increases.

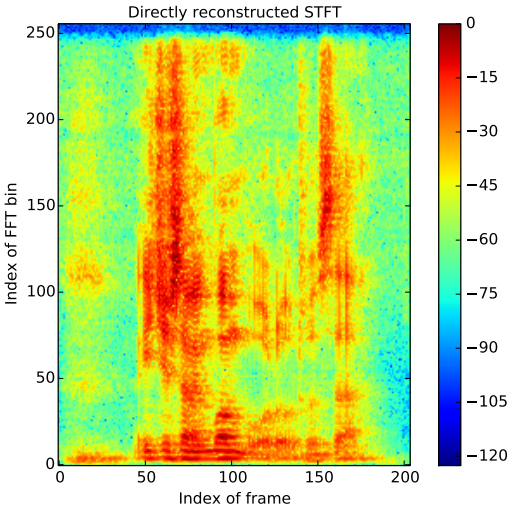
Fig. 6.7 further illustrates the residual error in reconstruction. It is worth noting that the energy distribution of the residual error magnitude spectrum shown in Fig. 6.7 has a clear correlation with the target magnitude spectrum shown in Fig. 6.6. During the calculation of the residual error, care has been taken to minimize potential alignment error and inconsistent magnitude normalization. Since the speech pattern in significant in the residual error, it can lead to misleading conclusions by using the plot in Fig. 6.7 to judge the reconstruction pattern accuracy. Unlike the accurate reconstruction shown in Fig. 6.6 where an accurate signal match sufficiently proves an accurate reconstruction of speech patten, an inaccurate signal match in Fig. 6.7 with informative speech pattern in the residual error is insufficient to judge whether the reconstruction is accurate or not.

To better evaluate the reverberation modelling accuracy in terms of speech recognition, a set of experiments is conducted on the SWC data using the DNN-HMM hybrid acoustic model structure described in Section 4.4.3. In the experiments, the acoustic model is trained with the features from the headset recordings. The acoustic model is then tested with the features generated from the spectrogram of the simulated reverberant signal, and the features from the spectrogram reconstructed using reverberation modelling by Eq. (6.10). The impact of the temporal resolution is investigated by comparing the WERs for different values of N_{f}. The language model and decoding configuration are the same as those used in the experiments in Section 4.4.3. The RIR used to simulate the reverberant signal is taken from a distant microphone in the circular table array ("TBL1-01"), and the details about the microphone geometry can be found in Section 4.1.

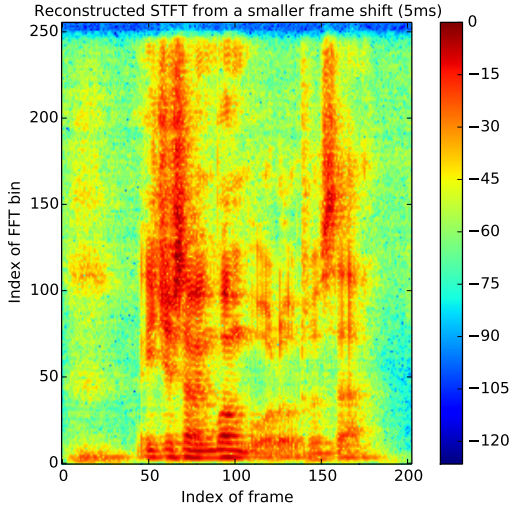
The reverberation modelling error could be observed in Table 6.1 by comparing the WERs based on the simulated reverberant speech and the WERs based on the reconstructed spectrogram using different N_{f} values. A high WER difference indicates a large reverberation modelling error in terms of the speech feature pattern. When N_{f} corresponds to 10 ms, namely the frequently adopted frame shift size, there is a 0.5% absolute WER



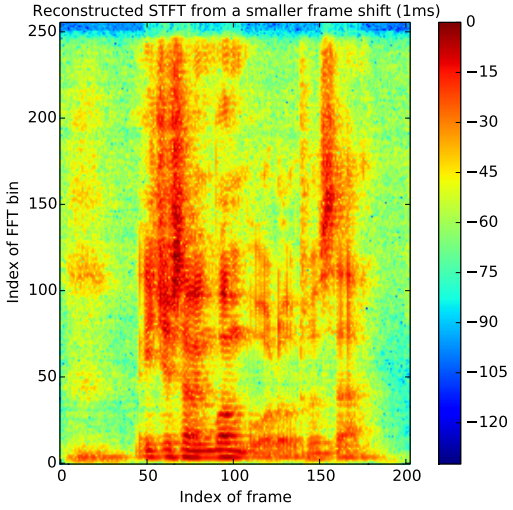
(a) *The magnitude spectrogram from simulated reverberant speech signal;*



(b) *The magnitude spectrogram reconstructed directly from the complex spectrogram of the headset recording and the special complex spectrogram of the RIR with STFT updated per 10 ms (N_f corresponds to 10 ms);*

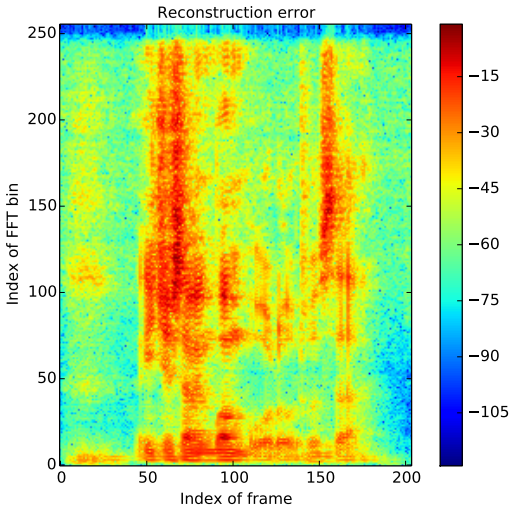


(c) *The magnitude spectrogram reconstructed directly from the complex spectrogram of the headset recording and the special complex spectrogram of the RIR with STFT updated per 5 ms (N_f corresponds to 5 ms);*

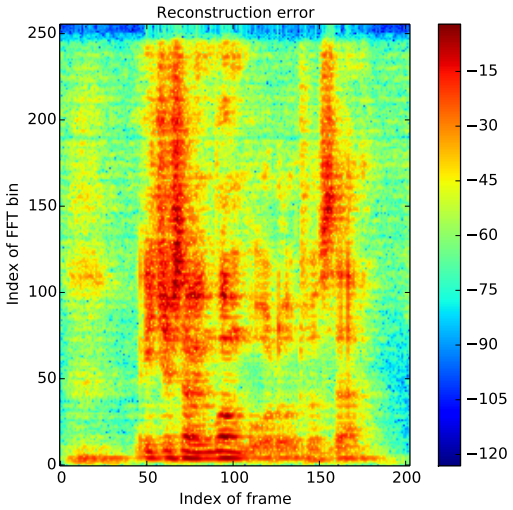


(d) *The magnitude spectrogram reconstructed directly from the complex spectrogram of the headset recording and the speical complex spectrogram of the RIR with STFT updated per 1 ms (N_f corresponds to 1 ms);*

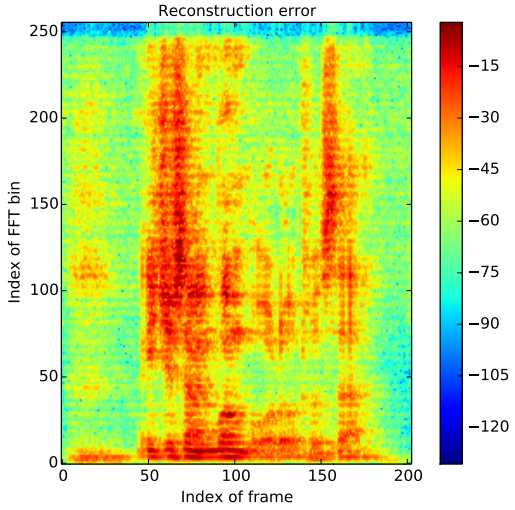
Fig. 6.6 *The magnitude spectrogram from reverberant signal and from complex spectrogram reconstructed based on the same speech utterance: "Are you just having one warlord, what's that?" (dB)*



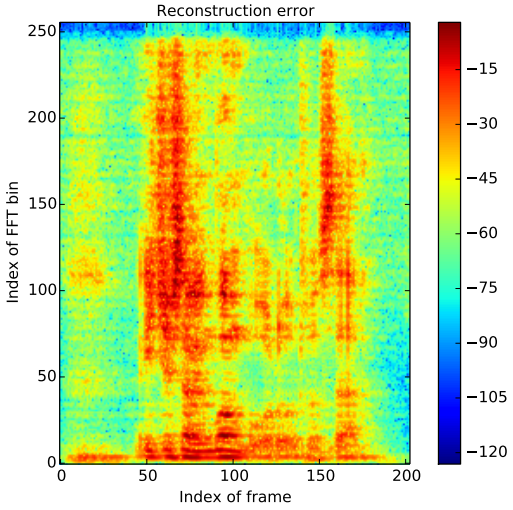
(a) *Residual error in magnitude spectrogram between Fig. 6.6a and Fig. 6.6b where N_f corresponds to 10 ms ;*



(b) *Residual error in magnitude spectrogram between Fig. 6.6a and Fig. 6.6c where N_f corresponds to 5 ms ;*



(c) *Residual error in magnitude spectrogram between Fig. 6.6a and when N_f corresponds to 2 ms ;*



(d) *Residual error in magnitude spectrogram between Fig. 6.6a and Fig. 6.6d where N_f corresponds to 1 ms ;*

Fig. 6.7 *Residual error in magnitude spectrogram using different time resolution when construction the complex spectrogram, based on the same speech utterance: "Are you just having one warlord, what's that?" (dB)*

Table 6.1 *WER and pattern approximation accuracy with different N_f value.*

The STFT for features of test data	dev set				eval set			
	Sub.	Del.	Ins.	WER	Sub.	Del.	Ins.	WER
headset baseline	*29.9*	*7.0*	*4.3*	*41.3*	*29.8*	*6.7*	*4.6*	*41.2*
reverberant speech signal (simulated)	35.9	10.6	5.2	**51.7**	36.3	10.2	5.3	**51.9**
reconstruct spectrogram with N_f∼10ms	35.5	10.5	5.1	51.2	36.1	10.0	5.3	51.4
reconstruct spectrogram with N_f∼5.0ms	35.8	10.3	5.4	51.5	36.2	10.0	5.5	51.8
reconstruct spectrogram with N_f∼2.5ms	35.9	10.3	5.5	51.6	36.3	9.9	5.6	51.8

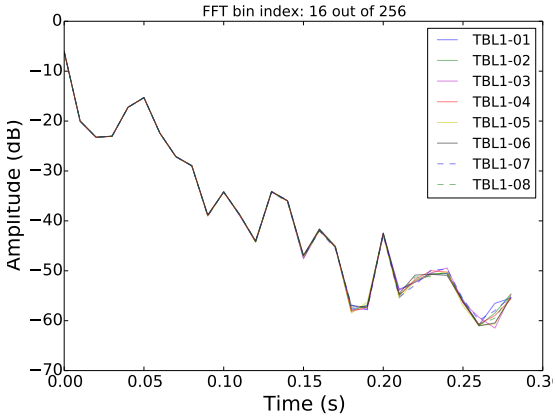
difference. The WER difference is further reduced when a smaller value is used for N_f. When N_f corresponds to 2.5 ms, the WER difference becomes marginal (0.1%). This indicates that the patterns in the reconstructed magnitude spectrogram are very close to the patterns in the magnitude spectrogram calculated directly from the simulated reverberant signal. Compared to the WER degradation caused by reverberation (>10% absolute), the WER difference caused by reverberation modelling error is so small that the error from the reverberation modelling based on Eq. (6.10) could be neglected in DSR applications.

A further experiment is conducted where the acoustic model is both trained and tested with features based on the reconstructed complex spectrogram using N_f corresponding to 10 ms. The WER is 50.6% for dev set and 51.0% for eval set, lower than the WERs from the acoustic model trained with the original reverberant spectrogram based features - 51.2% for dev set and 51.4% for eval set as shown in Table 6.1. The WERs based on the features generated on the reconstructed complex spectrogram are lower compared to the WERs based on the features generated from the reverberant speech signal directly, suggesting that the reverberation modelling with Eq. (6.10) simplifies the speech pattern from the complex spectrogram onwards. The pattern simplification is caused by the smoothing effects from the locally linear phase assumption and the locally stationary magnitude assumption, and the degree of simplification increases as the frame rate decreases. Similar pattern simplification has been observed by Sehr and Kellermann (2008, 2009) in the reverberation modelling based on Mel spectral features (Fig. 1 in their work). Compared to their work, the example plot in Fig. 6.6 and the WERs in Table 6.1 both suggest that the reverberation modelling based on complex spectrogram in this work (Eq. (6.10)) better maintains the speech pattern for speech recognition task.

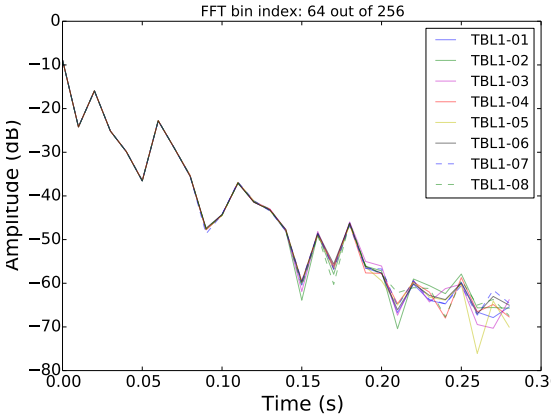
6.3.4 Reverberation variation analysis with the RIR spectrogram

Section 5.2.2 has conducted speech recognition experiments with the simulated reverberant data based on the RIRs measured in the SWC recording room using microphones installed at different locations. The experiment results illustrated in Fig. 5.12 have shown

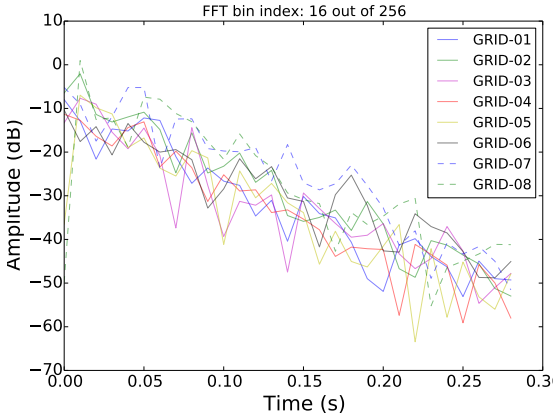
that both the speaker location and the microphone installation could impact the channel reverberation level estimated with C_{50}. With the reverberation modelling based on complex spectrogram (Eq. (6.10)), similar investigation could be conducted regarding the variation in reverberation modelling parameters due to speaker movement and microphone configuration difference. The investigation aims at providing some insights into how the variation in reverberation impacts the performance of dereverberation algorithms based on spectrogram.



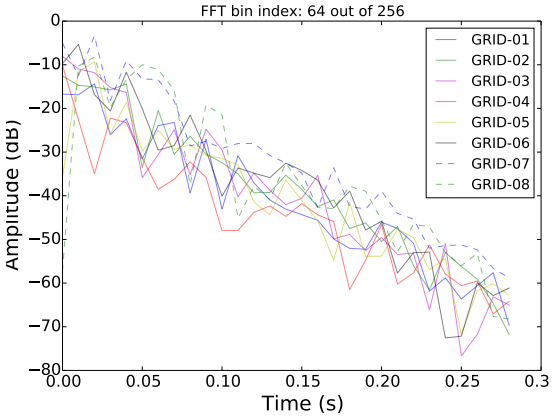
(a) *RIRs from 8 channel circular microphone array placed at the center of table ("TBL1") at 500 Hz.*



(b) *RIRs from 8 channel circular microphone array placed at the center of table ("TBL1") at 2000 Hz.*



(c) *RIRs from 8 channel microphones hanging from ceiling grid ("GRID") at 500 Hz.*

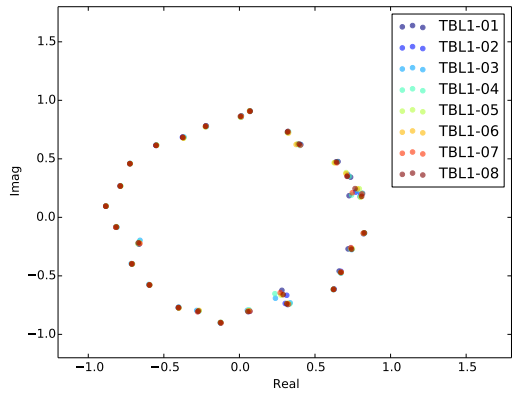


(d) *RIRs from 8 channel microphones hanging from ceiling grid ("GRID") at 2000 Hz.*

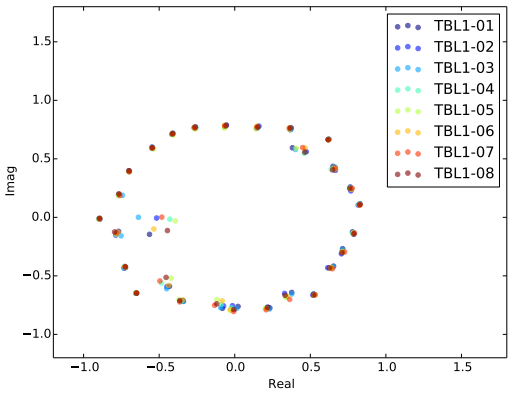
Fig. 6.8 *Magnitude spectrogram of RIRs measured with the same speaker location in SWC recording room.*

One particularly important question regarding the treatment of reverberation distortion is what kind of configuration in microphone and microphone array is advantageous to the dereverberation algorithms. Take the multi-channel dereverberation algorithm GWPE (Yoshioka and Nakatani, 2012) as an example, which conducts dereverberation on complex

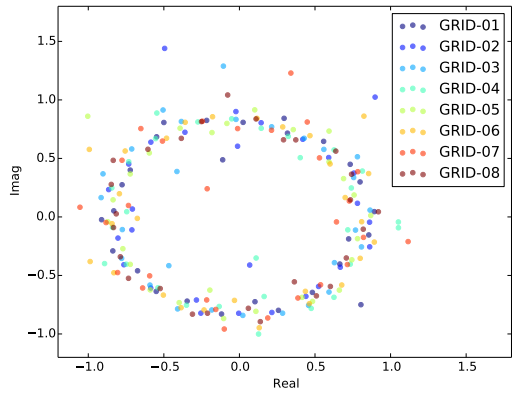
spectrogram. Since the reverberation modelling in Eq. (6.10) suggests that reverberation causes convolutional distortion on complex spectrogram, the RIR complex spectrogram is analysed for some insights into how the dereverberation algorithm is influenced by the microphone installation. The RIR magnitude spectrogram is calculated based on Eq. (6.11).



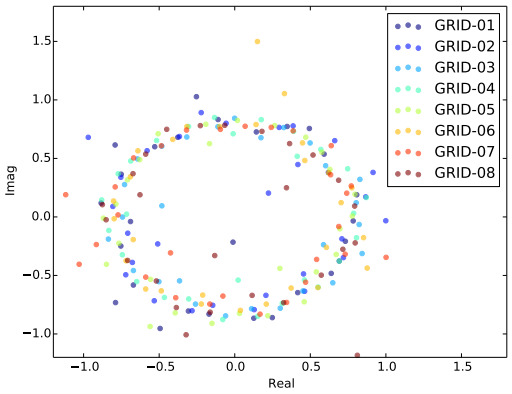
(a) 8 microphones from TBL1 (500 Hz).



(b) 8 microphones from TBL1 (2000 Hz).



(c) 8 microphones from GRID (500 Hz).



(d) 8 microphones from GRID (2000 Hz).

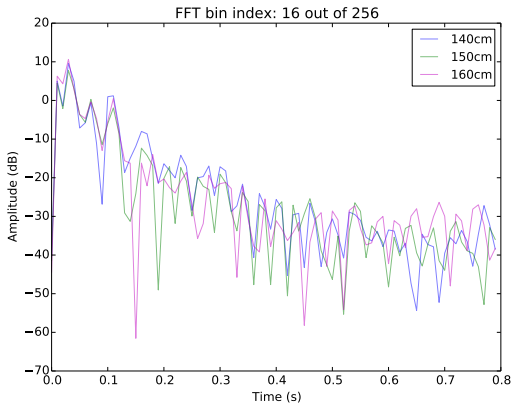
Fig. 6.9 *Zeros of equivalent complex valued filter based on reverberation modelling in Eq. (6.10) for the TBL1 microphones and the GRID microphones at different frequencies.*

Fig. 6.8 shows the RIR magnitude spectrogram at 500 Hz and 2000 Hz when the microphones are installed at different locations. As shown, the 8 microphones in the circular array placed on the table ("TBL1") have almost identical magnitude, indicating a very similar reverberation effect. In comparison, the 8 microphones hanging from the ceiling grid ("GRID") has a much larger variation in the magnitude spectrogram.

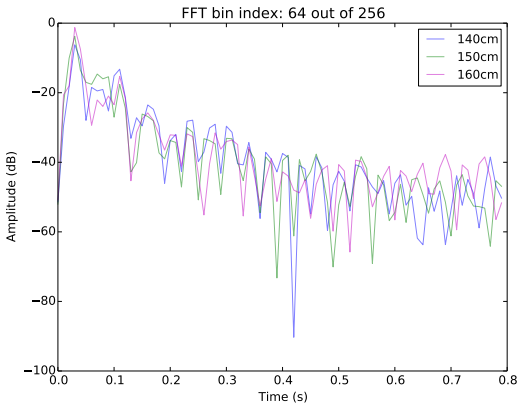
In an early work on the inverse filtering based multi-microphone dereverberation, Miyoshi and Kaneda (1988) pointed out that the mutli-microphone based dereverberation could outperform the single microphone based dereverberation if the multiple channels

do not share common zeros in RIRs. Fig. 6.9 illustrates the zeros of the RIR complex spectrogram corresponding to the microphone channels and frequencies shown in Fig. 6.8. The zeros are estimated by treating the RIR complex spectrogram as a complex valued FIR filter in each frequency bin. As shown, there are a lot of shared zeros in the RIR complex spectrogram from the microphones in the "TBL1" array, *i.e.* the no-shared-zero assumption does not hold for the microphones in the "TBL1" array. In comparison for the 8 microphones in the "GRID" group, the RIR complex spectrogram has very different zeros across microphone channels. This suggests that the 8 microphones hanging from ceiling grid ("GRID") are more suitable for multi-microphone based dereverberation than the 8 microphones from the circular table placed on the table ("TBL"). Actually the better performance of "GRID" microphones in GWPE based multi-microphone dereverberation was already observed in the experiment results in Table 5.4 in Section 5.2.2. Even though the microphones in the "TBL1" circular array are less reverberant than the microphones in the "GRID" group and the "TBL1" microphones have lower WERs, the "GRID" group outperforms the "TBL1" array with multi-microphone dereverberation.

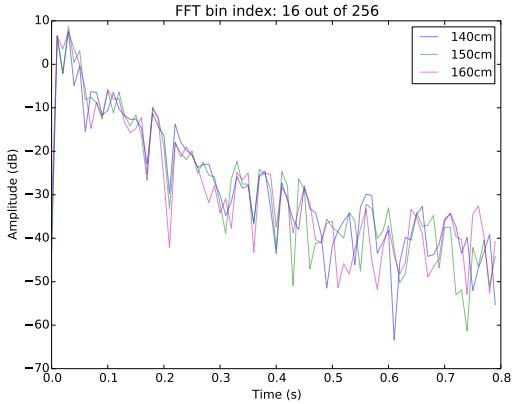
In applications where the microphone location is fixed once the recording starts, the RIR variation or reverberation variation comes mainly from speaker movements and the change in acoustic environment. In the SWC recording configuration, the acoustic environment change is much less compared to speaker movement. In addition, it is very difficult to quantify the acoustic environment change in a natural conversation scenario. Therefore the acoustic environment change is dropped out from the discussion here. In the SWC recordings, the RIRs have been measured by placing loudspeakers at different locations in the room (Fig. 5.9) and at three different heights (1.4 m, 1.5 m and 1.6 m). Details about the speaker location configuration can be found in Section 5.2.2. Fig. 6.10 shows a few examples of the RIR magnitude spectrogram with up to 20 cm difference in the speaker height, when the speaker location, the microphone configuration and the acoustic environment all stay the same. As shown, variations could be observed in the magnitude spectrogram at different frequencies even though there is only 10 cm change in the physical height of speaker. Similarly, Fig. 6.11 illustrates the RIR magnitude spectrogram variation when speaker is 0.15 m, 0.45 m and 0.75 m away from the table, and the RIRs are measured from the microphone "TBL1-01" located at the center of table. As shown, speaker movements also introduce variation in magnitude spectrogram.



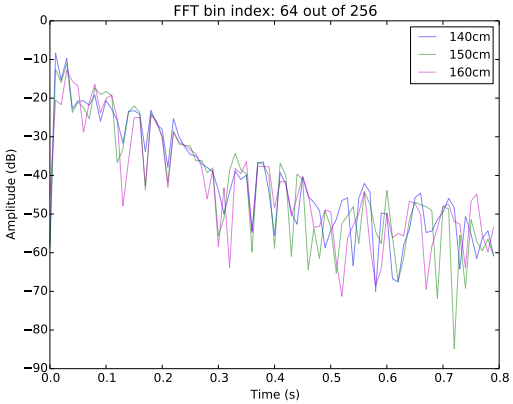
(a) *500 Hz, 0.15 m from the table in D1 direction as shown in Fig. 5.9.*



(b) *2000 Hz, 0.15 m from the table in D1 direction as shown in Fig. 5.9.*



(c) *500 Hz, 0.15 m from the table in D5 direction as shown in Fig. 5.9.*

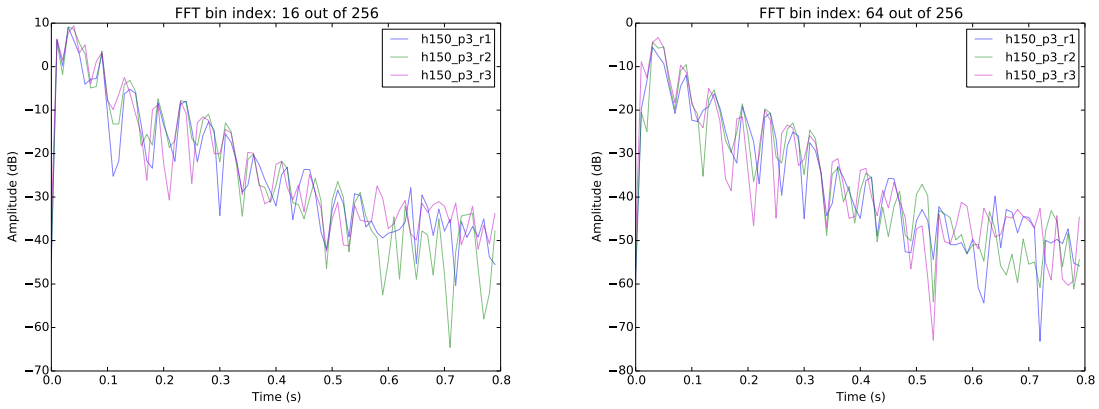


(d) *2000 Hz, 0.15 m from the table in D5 direction as shown in Fig. 5.9.*

Fig. 6.10 *Magnitude spectrogram of RIRs measured with slight variation in speaker height at the same location using the same microphone (TBL1-01).*

6.4 Summary and Discussion

This chapter has investigated the impact of reverberation on the speech complex spectrogram which is the building-block for the features used in many state-of-the-art ASR systems. The analytic analysis illustrates that the complex spectrogram of the reverberant speech is a convolution between the complex spectrogram of the clean speech signal and the RIR. This convolutional relationship suggests that the fast change in the speech magnitude spectrum is the fundamental reason for the pattern distortion in the reverberant speech spectrogram and the spectrogram based features. With two assumptions introduced, the reverberation modelling could be conducted with the frame level complex spectrogram. The two assumptions are the locally stationary magnitude assumption that the clean speech spectrogram magnitude does not change in a short time span, and the locally linear phase



(a) 500 Hz. (b) 2000 Hz.

Fig. 6.11 *Magnitude spectrogram of RIRs measured with the same microphone (TBL1-01) with speaker of different distances to the table in the same direction, i.e. D3 shown in Fig. 5.9. "r1": 0.15 m to the table; "r2": 0.45 m to the table; "r3": 0.75 m to the table.*

assumption that the clean speech spectrogram phase changes linearly in a short time span with a constant and frequency dependent gradient. Experimental evaluation based on the headset recordings from the SWC data suggests that the two assumptions are reasonable, and that the reverberation modelling based on frame-level complex spectrogram is sufficiently accurate for ASR applications. Compared to previous work, the reverberation modelling proposed in this work better preserves the patterns in speech features.

There are a few findings and contributions from the work in this chapter. First, the reverberation modelling based on a convolution between the RIR complex spectrogram and the clean speech complex spectrogram is sufficiently accurate for ASR tasks. This provides a solid support for the dereverberation algorithms based on frame level complex spectrogram. In particular, the recent progress in various deep network structures makes it possible to construct an acoustic model directly with low level features.

Second, the parameter analysis on the RIRs from different microphone configuration explains the performance limit in the multi-channel dereverberation based on microphones installed too close to each other - such microphone installation can result in many shared zeros in the RIR complex spectrogram factorization. This is consistent with the experimental observation in Table 5.4 in Section 5.2.2. Therefore the best microphone combination for multi-channel derevrberation is not necessarily the same with a combination of the microphones that provide the best recognition performance individually. The RIRs of the microphones to combine for multi-channel dereverberation should have different zeros in the complex spectrogram factorization. This could be achieved by increasing the distance among microphones.

Third, the variation in reverberation caused by speaker movements is briefly covered. With the reverberation modelling based on the frame-level complex spectrogram, it is shown that a speaker movement of 10 cm could cause a wide range of fluctuation in the RIR complex spectrogram across all frequencies. This suggests that the inverse filtering and the inverse processing based dereverberation will be generally challenged by the speaker movement in real applications.

The work in this chapter contributes to a better understanding of reverberation regarding how it distorts the features for speech recognition. To make this knowledge really useful in improving the DSR performance, the influence from background noise and overlapping speech should not be ignored. Chapter 5 has shown that the overlapping speech degrades the performance of dereverberation. The analysis on background noise is not covered by this work, but the background noise has also been widely observed to degrade the dereverberation performance. The background noise is additive to the speech component in the spectrogram while the reverberation distortion is convolutional. The treating filter for dereverberation could transform and amplify the noise component, resulting in the speech pattern distortion and as well as the WER increase. Therefore the dereverberation algorithms are usually combined with the de-noising algorithms in the DSR applications.

Chapter 7

Reverberation Measurement

Contents

7.1 Motivation 136

7.1.1 Reverberation distortion from the early reflections 136

7.1.2 Reverberation level and reverberation distortion level 140

7.2 Polynomial Reverberation Measurement 145

7.3 Phonetic Analysis Inspired Reverberation Measurement 147

7.3.1 Intra-phone smearing and inter-phone smearing 148

7.3.2 Combination of smearing indices 152

7.4 Fisher Ratio Based Discrimative Analysis 154

7.5 Experiment Results 156

7.5.1 Polynomial reverberation measurement 157

7.5.2 Phoneme duration and smearing ratio 160

7.5.3 Phonetic analysis inspired reverberation measurement 163

7.5.4 Fisher ratio based discriminative analysis 171

7.6 Summary and Discussion 172

The multi-condition training has gained much popularity recently due to its simplicity in implementation and its effectiveness in improving the overall robustness of DNN front-end and DNN acoustic model against reverberation. For multi-condition training to achieve a balanced selection of data that covers diverse reverberation conditions, the reverberation measurement becomes critical. Since the multi-condition training has the side-effect that the recognition performance degrades on relatively less reverberant data, recent research

has proposed to use model selection based on the reverberation measurement for a better balance between overall robustness against diverse reverberation conditions and the optimal performance in each reverberant condition. Therefore the reverberation measurement plays a key role in improving the reverberation robustness of current DNN based DSR systems. As discussed in Section 3.4, existing methods for reverberation measurement have been focusing on estimating the reverberation level of the acoustic environment and the recording channel rather than the reverberation distortion level in the speech feature pattern in a recognition task. This could cause a suboptimal decision in the data selection and the model selection when the reverberation level is evaluated on short recordings, because the difference between the reverberation level and the reverberation distortion level can be amplified by sound pattern sparsity in short recordings.

This chapter covers the novel research work aimed at improving existing methods to better estimate the reverberation distortion level. Section 7.1 illustrates the problems in existing methods, based on which Section 7.2 proposes a polynomial style reverberation measurement strategy for short recordings. The proposed method is the first research work according to the author's knowledge that takes the signal properties into consideration when evaluating the reverberation distortion level in a given environment. The following Section 7.3 explores the idea of phonetic pattern based reverberation analysis by Assmann and Summerfield (2004) in the context of measuring reverberation distortion level. The idea of self-masking and overlap-masking caused by reverberation as introduced by Kokkinakis and Loizou (2011) is borrowed to partition the reverberation distortion into the intra-phone smearing and the inter-phone smearing. Based on that partition, the overall distortion level is estimated by combining the reverberation distortion index of each part. Section 7.4 further discusses one implicit assumption made by the proposed strategy when estimating the feature pattern distortion level caused by inter-phone smearing. The experimental analysis of existing methods will be first illustrated in Section 7.1 regarding the issues in existing methods, and the experiment results of the proposed methods will be discussed in Section 7.5.

7.1 Motivation

7.1.1 Reverberation distortion from the early reflections

As reviewed in Section 2.4, existing research on reverberation measurement and dereverberation usually groups the reflected sound in the reverberant environment into early reflections and late reflections, or early reverberation and late reverberation (Valimaki et al., 2012). Such a partition originates from the research on human perception of reverbera-

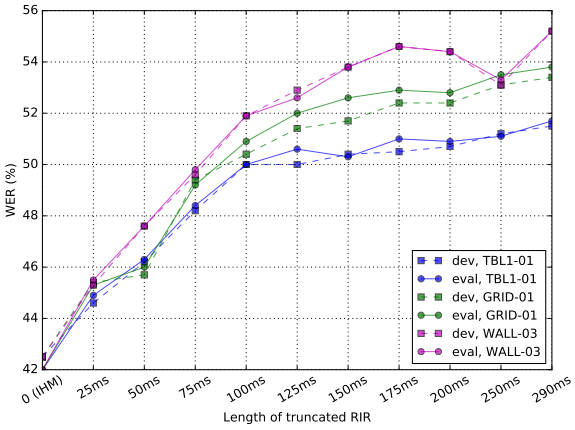
tion which suggests that early reflections and late reflections bring different perceptual experience in speech intelligence. The early reverberation contributes to sound colouration (Habets et al., 2006; Naylor and Gaubitch, 2005) which provides a positive impact on human intelligibility of speech, with an effect similar to increasing the strength of direct-path sound and the effective SNR (Bradley et al., 2003; Hu and Kokkinakis, 2014). In comparison, the late reflections cannot be integrated with direct sound and it causes a smearing effect of temporal blurring in the spectrogram, thus decreasing the speech intelligibility (Hu and Kokkinakis, 2014). Inspired by such findings on human perception of reverberation, existing reverberation metrics such as the early-to-late reverberation ratio (ELR) tend to integrate early reverberation with direct sound, and the dereverberation research tends to focus on the treatment of late reverberation.

However, some experiments on speech recognition using simulated reverberant recordings show that the early reflections can also degrade the quality of recorded speech, though the late reverberation might cause a higher level of speech feature distortion (Hu and Kokkinakis, 2014; Naylor and Gaubitch, 2005). The analytic investigation in Chapter 6 on how reverberation affects ASR features has shown that the fundamental reason for the negative impact from reverberation lies in the fast temporal change of the speech spectrogram, particularly the magnitude spectrogram. This suggests that theoretically the distortion can be caused by not only the late sound reflections but also the early sound reflections if the signal has a magnitude spectrogram changing fast enough. So far there has been very limited research conducted on the negative impact the early reverberation to speech recognition. Instead, recent research trend on reverberation measurement and dereverberation tends to ignore the impact of early reverberation. The structure of the widely adopted ELR based reverberation metric (Brutti and Matassoni, 2014, 2016; Parada, Sharma and Naylor, 2014) has an intrinsic blind spot regarding the distortion from the early reverberation, and the dereverberation algorithms also tend to relax the treatment on early reverberation (Yoshioka and Nakatani, 2012).

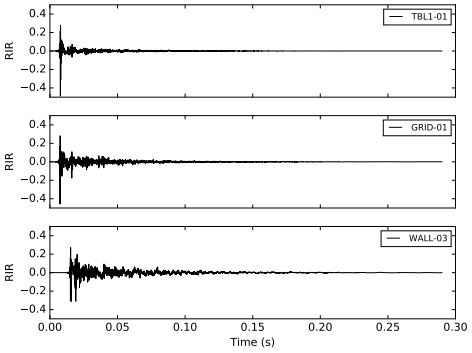
To validate the impact of early reverberation on the speech recognition performance, experiments are first conducted on simulated reverberant data by convolving the headset recordings in SWC data with RIRs truncated to different lengths. The acoustic modelling, the language modelling, the decoding configuration and the dataset definition all follow the setup in the DNN-HMM hybrid system in Section 4.4.3. To avoid the mismatch in channel, the system trained on the simulated data using one truncated RIR is tested on the simulated data using the same truncated RIR. The three RIRs involved in the experiments are estimated based on recordings of swept sine signal in the SWC recording room given the same speaker location using three different microphones: one located at the center of table ("TBL1-01"), one hanging from ceiling grid ("GRID-01") and one installed to the

Table 7.1 *Speech recognition performance comparison of simulated reverberant speech with full or truncated RIRs.*

		C_{50} (dB)	dev				eval			
			Sub.	Del.	Ins.	WER	Sub.	Del.	Ins.	WER
headset baseline		-	*29.9*	*7.0*	*4.3*	*41.3*	*29.8*	*6.7*	*4.6*	*41.2*
Full RIR (290 ms)	TBL1-01	17.46	35.7	10.4	5.4	51.5	36.3	9.7	5.7	51.7
	GRID-01	13.54	37.6	9.9	5.9	53.4	38.0	9.5	6.2	53.8
	WALL-03	9.93	38.8	10.8	5.6	55.2	38.9	10.3	6.0	55.2
Truncated RIR (50 ms)	TBL1-01	-	32.5	8.7	5.0	46.2	32.7	8.2	5.4	46.3
	GRID-01	-	31.9	8.7	5.1	45.7	32.3	8.2	5.4	46.0
	WALL-03	-	33.7	8.8	5.1	47.6	33.9	8.4	5.3	47.6



(a) *WERs and RIR truncation length.*



(b) *RIRs used in the truncation experiments.*

Fig. 7.1 *Early reverberation distortion via RIR truncation experiments.*

surface of the wall ("WALL-03"). The coefficients of the three RIRs are illustrated in Fig. 7.1b before any truncations. More details regarding RIR measurement could be found in Section 5.2.2. The speech recognition results are illustrated in Table 7.1 and Fig. 7.1a. To better interpret Fig. 7.1a, it is worth emphasising the meaning of x-axis in the plot. The increase in the length of RIRs after truncation equivalently unfolds the process of sound reflections adding up with the direct sound and the earlier arrival sound as time goes on.

Table 7.1 highlights the WERs when RIRs are truncated to 50 ms and when RIRs are not truncated at all (290 ms). As shown the WERs on reverberant data simulated with 50 ms long RIRs are 4.4-6.4% higher in absolute value compared to the WERs on headset recordings, and the reverberation after 50 ms increases WERs further by 5.3-7.7% absolute. Fig. 7.1a illustrates the incremental increase of WER as the length of truncated RIRs increases. The experiment corresponding to each dot in Fig. 7.1a is trained and tested with the same truncated RIR to avoid the RIR mismatch between training and test. As shown,

there is significant degradation in the recognition performance for all three microphone channels even when the RIR is truncated to 25 ms, *i.e.* RIR length being the same with signal processing window size in the ASR front-end. This suggests that the argument made by existing research (Yoshioka et al., 2012) that reverberation causes temporal feature smearing because the RIR is longer than the signal processing window in ASR front-end is problematic. Instead, the WER curves in Fig. 7.1a support the conclusions from the analytic investigation in Section 6.1 that the smearing effect caused by reverberation exists as long as there is a fast enough temporal change in the speech magnitude spectrogram. One more intuitive explanation is that even when the RIR is shorter than the frame width, the spectrogram in one frame of reverberant signal still contains the smearing components from the spectrogram in previous frames where the magnitude spectrogram might be very different.

There are four important messages that can be observed in Fig. 7.1a. First, the overall WER degrades accumulatively as the length of the RIR after truncation increases, and the increase of the RIR length refers to the arrival of more delayed sound reflections. Second, the early reverberation, typically defined as the sound reflections within 50 ms after the arrival of direct sound, can also cause significant WER degradation. Third, overall the WER increasing gradient decreases as the length of the RIR after truncation increases, and the WER increasing speed is generally larger when RIR is truncated to be shorter than 75 ms. This is reflected by the declining gradient of each WER curve in Fig. 7.1a, which may be caused by the declining RIR energy over time. Fourth, the multiple RIRs of different reverberation levels start to show significant channel difference in WER degradation when the RIR after truncation is larger than 75 ms, and overall the channel difference increases as the length of RIR after truncation increases. Since the increase in the length of the RIR after truncation equivalently unfolds the temporal process in which the late sound reflections add up to the earlier arrival sound in distant recordings, the four messages depict how early sound reflections and late sound reflections incrementally degrades WER and increases channel difference.

As mentioned in Section 3.4, one problem not fully addressed with the ELR based reverberation measurement is the optimal boundary between the early reverberation and the late reverberation. So far the partition boundary is determined empirically via experiments, and there have been different opinions on the optimal boundary between the early reverberation and the late reverberation in human speech perception, phoneme recognition and word recognition (Bradley, 2011; Brutti and Matassoni, 2014; Parada, Sharma and Naylor, 2014; Sehr et al., 2010). The ambiguity in the early and late reverberation boundary could be observed in the WER curves in Fig. 7.1a. The WER curves illustrate an incremental WER increase with all channels. In addition, the microphone channel difference is not

significant with the starting part of RIR, and it accumulates to a significant degree only with both the early and late reverberation. Therefore there is not a singular point with a dramatic change in the reverberation distortion behaviour.

As a consequence, the idea of partitioning reverberation into early reverberation and late reverberation for speech recognition is fundamentally arguable. In different tasks with different data, particularly with different RIRs, the experimental observations might vary regarding the optimal early-to-late reverberation partition boundary. For example according to Fig. 7.1a, 25 ms could be similarly effective to 50 ms as the early-late reverberation boundary in measuring the channel difference in reverberation. In fact, the C_{25} for "TBL1-01" is 12.82 dB, for "GRID-01" is 8.09 dB, and for "WALL-03" is 5.13 dB, providing the same rank order with C_{50} as shown in Table 7.1. In this case the only drawback of using a small value as boundary is an increase in the sensitivity to the direct sound alignment when RIRs are measured at different speaker-microphone distances, as shown in Fig. 7.1b.

7.1.2 Reverberation level and reverberation distortion level

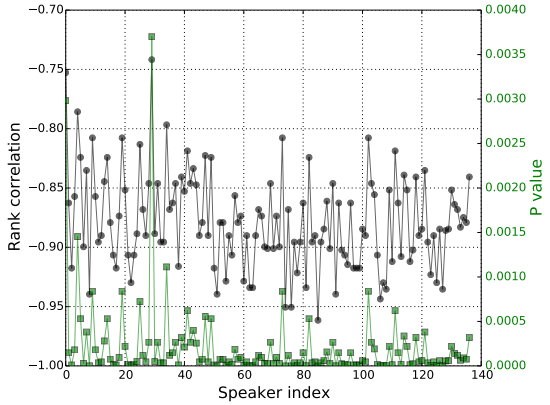
As reviewed in Section 2.4, reverberation metrics such as the ELR based C_{50} and reverberation time T_{60} have been adopted by DSR applications to improve the robustness of recognition performance against reverberation (Brutti and Matassoni, 2016; Parada, Sharma and Naylor, 2014; Parada et al., 2015; Sehr et al., 2010). As further pointed out in Section 3.4, such reverberation metrics are however designed to measure the reverberation level of the environment and channel rather than the reverberation distortion level of the speech feature patterns. Therefore they can be suboptimal due to the difference between the reverberation level and the reverberation distortion level. Inspired by Assmann and Summerfield (2004), Section 3.4 has briefly analysed such difference via the reverberation impact on different phonetic pattern structures. The analysis suggests that the difference between the reverberation level in channel and the reverberation distortion level in speech feature pattern could be amplified by the very limited types of speech feature patterns in short recordings. This argument can be experimentally verified by examining the correlation between the reverberation measurement results and the pattern recognition results such as the phoneme error rate (PER) and WER when the reverberation measurement is performed on recordings of different lengths.

The experiments verifying the correlation between C_{50} and PER are performed with the WSJCAM0 (Robinson et al., 1995) headset recordings and the 13 RIRs measured in the SWC recording room. The RIRs are measured given the same sound source location using 13 microphones installed at different locations in the room: eight microphones hanging

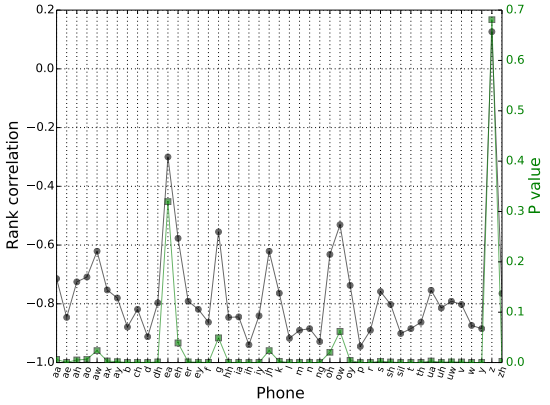
from a ceiling grid ("GRID-0*"), four microphones distributed on the wall ("WALL-0*") and one microphone placed at the center of the table ("TBL1-01"). The WSJCAM0 headset recordings are used because the WSJCAM0 database has provided high quality reference phonetic alignment. To avoid the confusion in the results caused by the mismatch in speaker and speaking style between training and test, new datasets are defined by using 60% of the speech utterances from each speaker for training, 20% speech utterances for development and validation, and the remaining 20% speech utterances for evaluation. A DNN is trained on headset recordings with TNet layer-by-layer using cross-entropy cost function. The trained DNN is used to perform phoneme classification on reverberant data simulated by convolving the WSJCAM0 headset recordings with the 13 RIRs from SWC recording room. For phoneme classification, the phonetic boundary is taken directly from the WSJCAM0 reference, so that the trained DNN only performs classification over given frames of features based on the maximum overall posterior. The reference alignment is used because in preliminary experiments it is found that the DNN based phoneme classification without provided phonetic boundary produces very noisy results, and the PER calculated in this way does not well correlate with WER.

Fig. 7.2 illustrates the Spearman rank correlation between PER and C_{50} regarding the microphone channel difference when PER is calculated with different amount of data. As shown in Fig. 7.2a, overall there is a very high negative rank correlation between C_{50} and PER per speaker regarding the microphone channel difference. When comparing across speakers, the Spearman rank correlation varies from around -0.75 to -0.95. When the PER is evaluated per phoneme, as shown in Fig. 7.2b, the variation in Spearman rank correlation is much larger, from -0.3 to above -0.9. In particular, phoneme /ea/ and phoneme /z/ do not have significant rank correlation between PER and C_{50} regarding the channel difference in reverberation. For other phonemes, the Spearman rank correlation varies from below -0.6 to above -0.9.

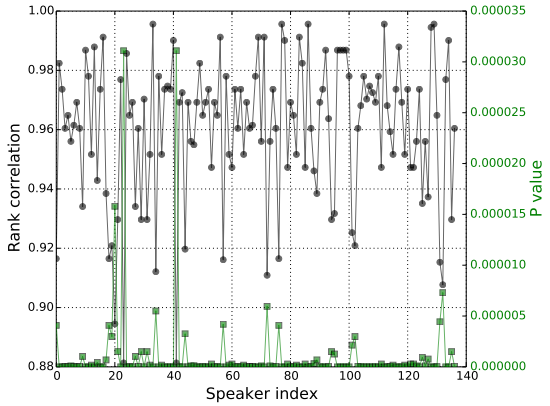
Furthermore, Fig. 7.2c illustrates the Spearman rank correlation between PER per speaker and the overall PER regarding the channel difference to provide an idea of the optimal performance of signal independent reverberation measurement. Fig. 7.2d illustrates similar analysis on the Spearman rank correlation between PER per phoneme and overall PER. The high rank correlation over all speakers in Fig. 7.2c suggests that there may exist a signal independent reverberation measuring method that could provide a high correlation between the reverberation score and the PER, when both the reverberation score and the PER are calculated per speaker. A comparison between Fig. 7.2a and Fig. 7.2c suggests that potentially C_{50} is not the optimal reverberation metric yet. Similarly, Fig. 7.2d suggests that there are some phonemes with the behaviour in phoneme level PER different from other phonemes when encountering diverse reverberation conditions, e.g. /ea/, /g/, /jh/ and



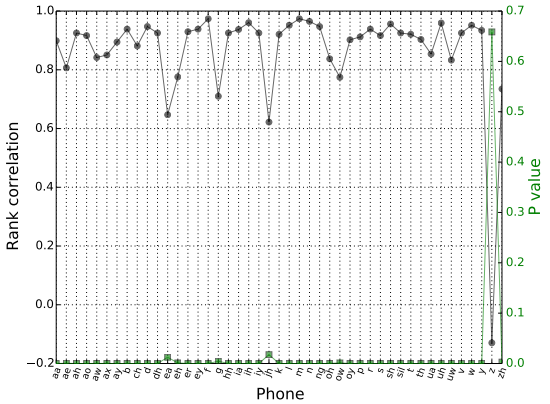
(a) *Spearman rank correlation between PER per speaker and C_{50}.*



(b) *Spearman rank correlation between PER per phoneme and C_{50}.*



(c) *Spearman rank correlation between PER per speaker and overall PER regarding the microphone difference on simulated reverberant data.*

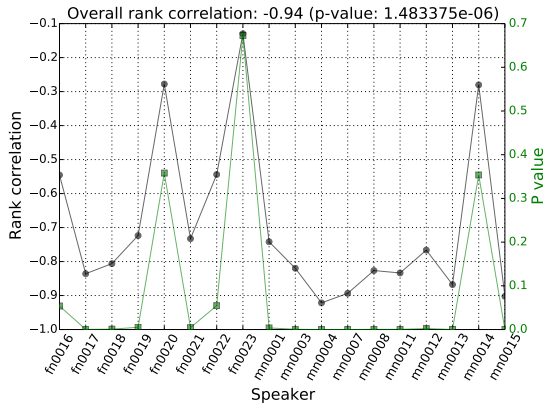


(d) *Spearman rank correlation between the PER per phoneme and the overall PER regarding the microphone difference on simulated reverberant data.*

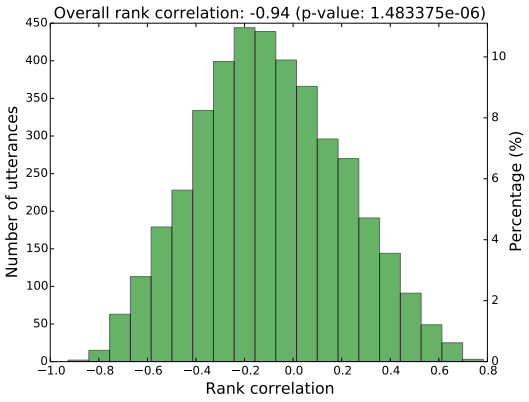
Fig. 7.2 *The Spearman rank correlation between the PER and the C_{50} at different data scales, and the Spearman rank correlation among PERs calculated at different scale, regarding the microphone difference on simulated reverberant data.*

/z/. A comparison between Fig. 7.2b and Fig. 7.2d suggests that C_{50} is not the optimal reverberation metric and it has some severe problem with phoneme /ea/.

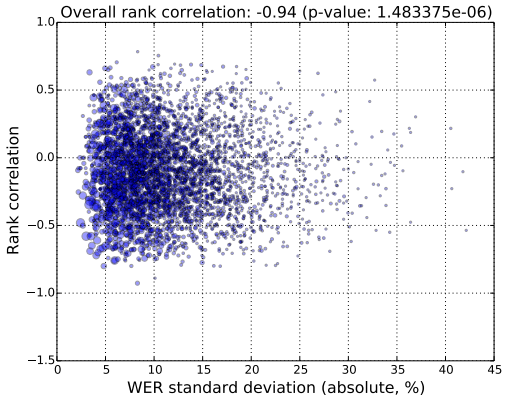
It is worth emphasising that all the PERs in Fig. 7.2 are based on the phoneme classification using reference alignment provided in WSJCAM0 corpus where the starting time and the ending time of each phoneme are labelled out in the audio recordings. Therefore the Spearman rank correlations shown in Fig. 7.2 are already of the best performance possible. In reality such high quality alignment is not available in a general set-up for phoneme classification, as a result the alignment errors could completely disrupt the correlation between PER and C_{50} and the correlation between PER and WER.



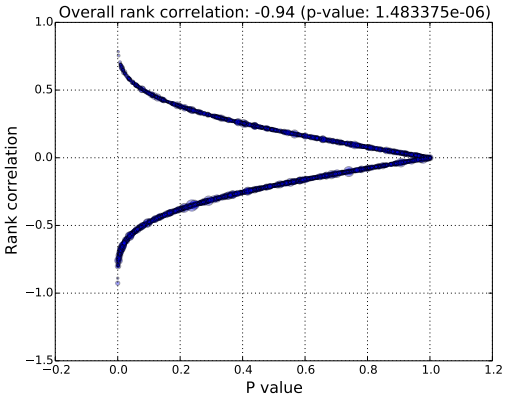
(a) *Spearman rank correlation between the overall WER per speaker and the C_{50} regarding the microphone difference with simulated reverberant speech (SWC hybrid system eval dataset).*



(b) *Histogram of the Spearman rank correlation between the utterance level WER and the C_{50} regarding the microphone difference with simulated reverberant speech (SWC hybrid system eval dataset).*



(c) *Spearman rank correlation between utterance level WER and C_{50} regarding the WER standard deviation (SWC hybrid system evaluation dataset).*



(d) *P-value scatter plot for the Spearman rank correlation shown in Fig. 7.3c.*

Fig. 7.3 *Spearman rank correlation between the WER and the reverberation score based on C_{50} at different scale: overall, per speaker and per utterance, based on simulated reverberant speech (SWC hybrid system evaluation dataset).*

Further experiments are conducted on SWC data regarding the correlation between the WER and the reverberation score from C_{50}. The same 13 RIRs are used to simulate reverberant data by convolving the RIRs with the headset recordings in SWC. One acoustic model is trained for each reverberant condition using data simulated with corresponding RIR, and tested on data of the same reverberant condition. Therefore unlike the experiments previously conducted for phoneme classification, there is no channel mismatch between training and test for word recognition here. The acoustic model training follows the DNN-

HMM hybrid system as detailed in Section 4.4.3. The dataset definition follows the "SA1" configuration shown in Table 4.7, so that there is no speaker mismatch between training and test. The same LM used in Section 4.4.3 is employed for decoding. The WER scoring is based on NIST tool "sclite", as there is no overlapped speech in the simulated data.

On evaluation dataset there is a high negative Spearman rank correlation of -0.94 between the C_{50} based reverberation score and the overall WER regarding the channel difference in diverse simulated reverberant conditions. The negative Pearson linear correlation is similarly high, being -0.94. On development dataset, the negative Spearman rank correlation between the C_{50} based reverberation score and the overall WER is also high, being -0.93, and the negative Pearson linear correlation is -0.94.

Fig. 7.3 shows the Spearman rank correlation between C_{50} and WER when WER is calculated at different scales, in a way similar to the PER analysis before. As shown in Fig. 7.3a, when the WER is calculated per speaker, some speakers have very low correlation between C_{50} and the WER over that speaker. This is particularly the case with female speaker "fn0016", "fn0020", "fn0023" and male speaker "mn0014".

Fig. 7.3b shows the histogram of the Spearman rank correlation between the utterance level WER and the C_{50} on evaluation dataset. Fig. 7.3c and Fig. 7.3d illustrate the corresponding scatter plot of the Spearman rank correlation and the p-value respectively. Only utterances with no less than 4 words are considered in the utterance level WER analysis. In Fig. 7.3c and Fig. 7.3d, each circle represents one utterance and the size of circle is proportional to the number of words in that utterance. As shown in Fig. 7.3b and Fig. 7.3c, when WER is calculated per utterance, there is a much larger variation in its correlation with C_{50} compared to the case where WER is calculated with more data, e.g. per speaker or over all data. This suggests that the utterance level model selection will be very noisy and will be far from the optimal if the selection is based on reverberation score from C_{50} or C_{50} based non-intrusive estimator.

In these experiments simulated reverberant data ensures that the channel difference in reverberation is the only factor causing the WER difference. Given the high rank correlation between the overall WER and the C_{50}, the poor correlation between utterance level WER and C_{50} (Fig. 7.3c and Fig. 7.3d) suggests that signal independent reverberation measurement such as the C_{50} based reverberation score fails to capture the reverberation effect on speech pattern in short recordings. As mentioned before, this is because existing methods estimate reverberation level independently from speech signal thus it could not accurately estimate the reverberation distortion level of the speech pattern in recognition tasks.

7.2 Polynomial Reverberation Measurement

As discussed in Section 7.1.1, existing reverberation measurement based on ELR could not reflect the distortion effect from early reverberation, and the optimal boundary between early reverberation and late reverberation is also yet to be fully addressed. Based on the analytic investigation in Section 6.1, this section proposes a polynomial format reverberation distortion level measuring method based on RIR which better estimates the pattern distortion from early reverberation without the necessity of early-late reverberation partition. In addition, it also takes into account the speech magnitude spectrum change when estimating the distortion level.

With Eq. (6.10) it was shown that the complex spectrogram of reverberant speech could be approximated with sufficient accuracy by a convolution of the complex spectrogram of clean speech and the complex spectrogram of the RIR. As pointed in Section 6.1, with faster and frequenter change of speech magnitude spectrogram, there is more distortion caused by reverberation. Therefore the reverberation distortion level could be assumed to be proportional to the temporal change in the magnitude spectrogram of speech signal given that the speech signal magnitude is normalized globally. Such change in magnitude spectrogram can be quantified with the average speech magnitude difference given a time shift ($\triangle\tau$) in each frequency bin (k), which will be further referred to as "speech dynamic index" denoted as $D(\triangle\tau,k)$. Following the notation used in Section 6.1, the speech dynamic index could be approximated with

$$\begin{aligned} D(\triangle\tau,k) &= \frac{1}{\lfloor\frac{G}{2\triangle\tau}\rfloor}\sum_{i=0}^{\lfloor G/(2\triangle\tau)\rfloor-1}\frac{1}{\triangle\tau}\sum_{j=0}^{\triangle\tau-1}\Big|\big|X((i+1)\triangle\tau+j,k)\big|-\big|X(i\triangle\tau+j,k)\big|\Big| \\ &= \frac{1}{\triangle\tau\cdot\lfloor\frac{G}{2\triangle\tau}\rfloor}\sum_{i=0}^{\lfloor G/(2\triangle\tau)\rfloor-1}\Big(\sum_{j=0}^{\triangle\tau-1}\Big|\big|X((i+1)\triangle\tau+j,k)\big|-\big|X(i\triangle\tau+j,k)\big|\Big|\Big) \\ &\approx \frac{2}{G}\sum_{i=0}^{\lfloor G/(2\triangle\tau)\rfloor-1}\Big(\sum_{j=0}^{\triangle\tau-1}\Big|\big|X((i+1)\triangle\tau+j,k)\big|-\big|X(i\triangle\tau+j,k)\big|\Big|\Big) \end{aligned} \tag{7.1}$$

where $\triangle\tau$ represents the time shift and G represents the number of samples in the piece of clean speech signal to analysis. The dynamic index is calculated with two levels of averaging. The sum over j is for the first average over one instance of a phoneme, and the summer over i is for the second average across multiple instances of that phoneme. The dynamic index is calculated per frequency bin (k). As shown in Section 6.2 the magnitude of clean speech could be assumed to be stationary over a short period of time, the summation could be simplified with a reduced time resolution in magnitude spectrogram.

Since the complex spectrogram of the reverberant speech is approximately a convolution which involves the complex spectrogram of the RIR, the magnitude spectrogram of the RIR could be used to represents the contribution of the corresponding clean speech magnitude spectrogram change in reverberation distortion. Therefore the reverberation distortion in each frequency bin is assumed to be proportional to the magnitude spectrogram of the RIR as well even that the magnitude of all RIRs to compare have been normalized properly, *i.e.*

$$I_\beta(N_\mathrm{f},\triangle\tau,k) \propto \left|D(\triangle\tau,k)\right|\left|H_\beta(\triangle\tau,N_f,k)\right| \tag{7.2}$$

where β is the index of the RIR and the reverberation condition, and $I_\beta(N_\mathrm{f},\triangle\tau,k)$ refers to an estimation of the reverberation distortion level caused by the average magnitude spectrogram change in clean speech over a time shift of $\triangle\tau$ in the frequency bin k, given a time resolution of N_f when calculating the spectrogram. To estimate the overall reverberation distortion level, the energy of $I_\beta(N_\mathrm{f},\triangle\tau,k)$ is summed over all frequency bins except for $k=0$ and over all possible time shifts, leading to a polynomial formula for estimating overall reverberation distortion level:

$$I_\beta(N_\mathrm{f}) = \frac{\sum_{k=1}^{N-1}\sum_{\triangle\tau=0}^{M'-1}\left|D(\triangle\tau,k)H_\beta(\triangle\tau,N_\mathrm{f},k)\right|^2}{\sum_{k=0}^{N-1}E_X(k)E_\beta^{(\mathrm{d})}} \tag{7.3}$$

$$E_X(k) = \sum_{\tau=0}^{G-1}\left|X(\tau,k)\right|^2 \tag{7.4}$$

and

$$M' = \min\left\{\left\lfloor\frac{G}{2\triangle\tau}\right\rfloor,\left\lfloor\frac{M_\beta}{2\triangle\tau}\right\rfloor\right\} \tag{7.5}$$

where M_β is the length of the RIR and $I_\beta(N_\mathrm{f})$ is the polynomial reverberation score to estimate the reverberation distortion level. In Eq. (7.3) the normalisation over the signal energy $E_X(k)$ is performed to reduce the sensitivity of distortion level estimation against the duration and the volume of speech recordings. Similarly the normalisation over the direction sound energy in the RIR $E_\beta^{(\mathrm{d})}$ is to reduce the sensitivity against the RIR magnitude. The normalisation over direct sound energy could be realised in advance on RIRs, so that all RIRs have the same maximum coefficient magnitude or the same maximum magnitude spectrogram.

As shown in Eq. (7.3), the reverberation distortion level is estimated with a polynomial of RIR magnitude spectrogram, namely the polynomial reverberation score. Compared to the ELR based reverberation metrics, it avoids the partition of early and late reverberation,

and it could estimate the reverberation distortion level when the RIR is of any length. Since the polynomial reverberation score analytically relies on the signal spectrogram, it is more suitable as an estimation of the reverberation distortion level in given speech signal caused by given reverberant environment depicted by RIR. As discussed in Section 3.4, the proposed method can be extended to potentially replace the C_{50} based reverberation reference on simulated data to train non-intrusive reverberation estimator. In addition, the proposed method could potentially benefit the speech synthesis research in predicting the speech clarity degradation in diverse reverberant conditions.

7.3 Phonetic Analysis Inspired Reverberation Measurement

In the work by Parada, Sharma, Naylor and Waterschoot (2014), the degradation in phoneme recognition under reverberation is analyzed using different toolkits and ASR model structures. It is shown that phonemes vary in their robustness to reverberation, and that the confusion of phonemes due to reverberation lead to the increased deletions and substitutions for ASR. In that same work, a metric, the confusability factor, is presented to characterize the confusion of recognizing a phoneme in a Bayesian framework.

According to Kokkinakis and Loizou (2011), the reverberation distortion is a combination of self-masking and overlap-masking on phonetic structure for human perception. However in the work by Parada, Sharma, Naylor and Waterschoot (2014), the cross-phoneme effects due to "over-masking" is not investigated. Inspired by Kokkinakis and Loizou (2011), this section explores the idea of dividing reverberation distortion into two parts: the distortion from intra-phone smearing and the distortion from inter-phone smearing. Their distortion levels are respectively estimated with the intra-phone smearing index and the inter-phone smearing index calculated per phoneme, so that an overall reverberation distortion level per phoneme is estimated by combining the two indices.

The phonetic analysis inspired reverberation measurement will be detailed in two sections. Section 7.3.1 describes the intra-phone smearing index and inter-phone smearing index and Section 7.3.2 discusses the ratio between intra-phone smearing and inter-phone smearing when combining them together. The experiment verification will be covered in Section 7.5.

7.3.1 Intra-phone smearing and inter-phone smearing

Intra-phone smearing refers to the distortion in phonetic pattern caused by components from the same phoneme at an earlier time due to reverberation. In comparison, inter-phone smearing refers to the distortion in phonetic pattern caused by components from preceding phonemes due to reverberation. Therefore the difference between intra-phone smearing and inter-phone smearing lies in the distorting components. The intra-phone smearing is closely related to early reverberation and inter-phone smearing is closely related to late reverberation.

Since each phoneme has a characteristic pattern structure in the magnitude spectrogram, the distortion level of intra-phone smearing could be estimated with the polynomial reverberation score implemented at phoneme scale, *i.e.* based on the speech dynamic index averaged per phoneme using the observation examples of that phoneme:

$$D_\alpha(\triangle\tau,k) = \frac{1}{L_\alpha}\sum_{l=1}^{L_\alpha-1}\frac{2}{G_{\alpha,l}}\sum_{i=0}^{\lfloor\frac{G_{\alpha,l}}{2\triangle\tau}\rfloor-1}\sum_{j=0}^{\triangle\tau-1}\Big|\big|X_\alpha((i+1)\triangle\tau+j,k)\big|-\big|X_\alpha(i\triangle\tau+j,k)\big|\Big| \tag{7.6}$$

where L_α is the number of examples for phoneme α and $G_{\alpha,l}$ is the number of spectrogram frames for the l-th example of phoneme α. Therefore a similar smearing index could be

$$I_{\alpha,\beta}(N_\mathrm{f}) = \frac{\sum_{k=1}^{N-1}\sum_{\triangle\tau=0}^{\lfloor\frac{M'}{N_\mathrm{f}}\rfloor-1}\Big|D_\alpha(N_\mathrm{f}\triangle\tau,k)H_\beta(\triangle\tau,N_\mathrm{f},k)\Big|^2}{\sum_{k=0}^{N-1}E_\alpha(k)E_\beta^{(\mathrm{d})}} \tag{7.7}$$

$$E_\alpha(k) = \frac{1}{L_\alpha}\sum_{l=0}^{L_\alpha-1}\sum_{\tau=0}^{G_{\alpha,l}-1}\big|X_\alpha(\tau,k)\big|^2 \tag{7.8}$$

where

$$M' = \Big\lfloor\frac{\min\Big\{\max_{l\in[0,L_\alpha)}\{G_{\alpha,l}\},M_\beta\Big\}}{2\triangle\tau}\Big\rfloor \tag{7.9}$$

In addition, the variance of magnitude spectrogram in each phoneme could be used to emphasise the phonetic difference in temporal magnitude spectrogram change, and it is again normalised by phoneme energy, leading to an average variance-energy ratio per phoneme:

$$r_\alpha = \frac{1}{L_\alpha}\sum_{l=0}^{L_\alpha-1}\frac{\sum_{k=1}^{N-1}\sigma_{\alpha,l}^2(k)}{\sum_{k=1}^{N-1}e_{\alpha,l}(k)} \tag{7.10}$$

where

$$e_{\alpha,l}(k) = \frac{1}{G_l}\sum_{\tau=0}^{G_l}\left|X_\alpha(\tau,k)\right|^2 \tag{7.11}$$

$$\mu_{\alpha,l}(k) = \frac{1}{G_l}\sum_{\tau=0}^{G_l}\left|X_\alpha(\tau,k)\right| \tag{7.12}$$

$$\sigma^2_{\alpha,l}(k) = \frac{1}{G_l}\sum_{\tau=0}^{G_l}\left|\left|X_\alpha(\tau,k)\right| - \mu_{\alpha,l}(k)\right|^2 \tag{7.13}$$

By adding r_α to Eq. (7.7), the overall intra-phone smearing index for phoneme α is

$$I^{(\mathrm{l})}_{\alpha,\beta}(N_\mathrm{f}) = \lambda I_{\alpha,\beta}(N_\mathrm{f}) + r_\alpha \tag{7.14}$$

where λ is a tuning parameter applied on polynomial reverberation score to balance the value range of the polynomial reverberation score and the variance-energy ratio. The overall intra-phone smearing index for the β-th RIR over all phonemes could be approximated with

$$I^{(\mathrm{l})}_{\beta}(N_\mathrm{f}) = \sum_{\alpha} P(\alpha) I^{(\mathrm{l})}_{\alpha,\beta}(N_\mathrm{f}) \tag{7.15}$$

where $P(\alpha)$ is the probability of phoneme α, and it could be estimated by the ratio between the total duration of phoneme α and the overall speech duration given sufficient audio size.

The intra-phone smearing level is estimated based on the temporal change in the magnitude spectrogram of each phoneme. The same strategy can not be adopted for inter-phone smearing estimation, because there is a large variation in the magnitude spectrogram change as the smearing takes place in different phoneme contexts. In addition, due to the different energy distribution among phonemes, the inter-phone smearing could cause a change of energy distribution over frequencies. In the intra-phone smearing estimation, each frequency bin is treated independently and equally, thus the overall smearing index is an average of the smearing indices overall all frequencies. This strategy is not suitable for the inter-phone smearing level estimation, because the change in the energy distribution across frequencies caused by inter-phone smearing could not be well depicted. Therefore a different strategy is proposed to estimate the distortion level caused by inter-phone smearing.

In inter-phone smearing, the phonetic pattern in the frequency axis could be represented by a power spectrum vector which is realized by averaging the power spectrogram over time for each phoneme. In this way each phoneme could be represented with a vector in a high dimensional space. The inter-phone smearing introduces move, scaling and rotation

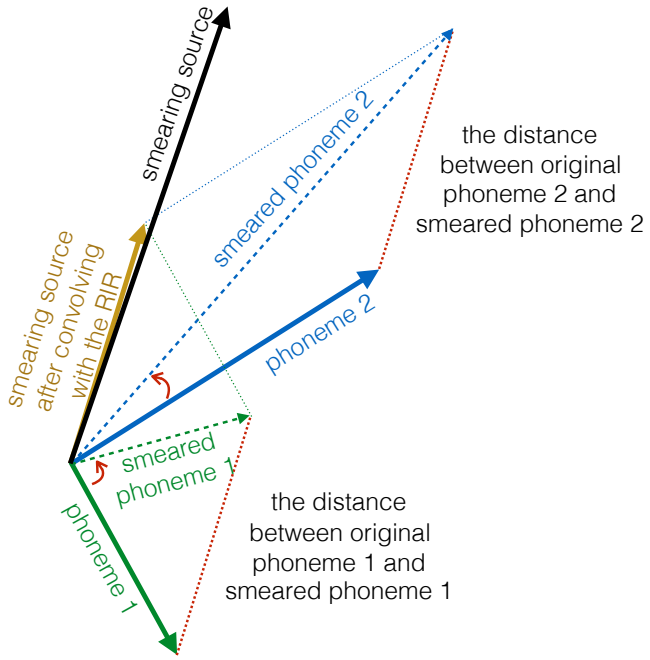


Fig. 7.4 *An illustration on how phonetic power spectrum pattern changes in a hyper space due to the linear effect of inter-phone smearing.*

to the power spectrum vector, as shown in Fig. 7.4. Since phonemes differ in energy, the cosine distance or the angle between the original power spectrum vector and the distorted power spectrum vector is adopted to describe the level of inter-phone smearing distortion. Thus the inter-phone smearing level is estimated by the amount of rotation in the power spectrum vector caused by the linear distortion from preceding phonemes as a consequence of the convolution with RIR spectrogram (Eq. (6.10)). Following this philosophy, the mathematical notation and calculation are covered below.

The power spectrum vector representing the phonetic pattern over frequency for phoneme α is denoted as $\mathbf{e}_\alpha$, and

$$\mathbf{e}_\alpha = [\bar{e}_\alpha(1), \bar{e}_\alpha(2), ..., \bar{e}_\alpha(N-1)]^T \tag{7.16}$$

$$\bar{e}_\alpha(k) = \frac{1}{L_\alpha} \sum_{l=0}^{L_\alpha - 1} e_{\alpha,l}(k) \tag{7.17}$$

where L_α is the number of examples to get sufficiently accurate average power spectrum for phoneme α and $e_{\alpha,l}(k)$ is the phonetic average energy in the k-th frequency bin of the l-th example for phoneme α. Similarly, a vector $\mathbf{e}_{\bar{\alpha}}$ is used to represent the statistic average power spectrum of phonemes preceding α that cause inter-phone smearing distortion:

$$\mathbf{e}_{\bar{\alpha}} = \sum_{\gamma,\, \gamma \neq \alpha} P(\gamma|\alpha)\mathbf{e}_\gamma \tag{7.18}$$

$$e_{\bar{\alpha}}(k) = \sum_{\gamma,\, \gamma \neq \alpha} P(\gamma|\alpha)\bar{e}_{\gamma}(k) \tag{7.19}$$

where $P(\gamma|\alpha)$ is the probability of phoneme α being distorted by phoneme γ. The convolution in Eq. (6.10) shows that in inter-phone smearing the complex spectrogram of preceding phonemes are weighted by the complex spectrogram of the RIR before adding onto the clean speech complex spectrogram. To approximate this effect, the power spectrum of the RIR corresponding to the inter-phone smearing ($\bar{e}^{(\mathrm{II})}_{\beta,\alpha}(N_\mathrm{f},k)$) is multiplied with the average speech phonetic power spectrum ($e_{\bar{\alpha}}(k)$) in each frequency, *i.e.*

$$\hat{e}_{\bar{\alpha},\beta}(k) \approx e_{\bar{\alpha}}(k)\bar{e}^{(\mathrm{II})}_{\beta,\alpha}(N_\mathrm{f},k) \tag{7.20}$$

$$\hat{\mathbf{e}}_{\bar{\alpha},\beta} = [\hat{e}_{\bar{\alpha},\beta}(1), \hat{e}_{\bar{\alpha},\beta}(2), ..., \hat{e}_{\bar{\alpha},\beta}(N-1)]^T \tag{7.21}$$

and for the power spectrum of the RIR

$$\bar{e}^{(\mathrm{II})}_{\beta,\alpha}(N_\mathrm{f},k) = \frac{1}{\lfloor \frac{M_\beta}{N_\mathrm{f}} \rfloor - \lfloor \frac{\bar{d}_\alpha}{N_\mathrm{f}} \rfloor} \sum_{\tau=\lfloor \frac{\bar{d}_\alpha}{N_\mathrm{f}} \rfloor - 1}^{\lfloor \frac{M_\beta}{N_\mathrm{f}} \rfloor - 1} \left| H_\beta(\tau, N_\mathrm{f}, k) \right|^2. \tag{7.22}$$

where $\bar{d}_\alpha$ corresponds to the average duration of phoneme α. The phoneme duration is involved in the average RIR power spectrum estimation above so that the RIR power spectrogram elements in the summation are only those contributing to the inter-phone smearing only. As the inter-phone smearing components are added to the original phoneme in spectrogram, the distorted power spectrum of phoneme α by the β-th RIR is approximated with

$$\mathbf{e}_{\alpha,\beta} \approx \mathbf{e}_\alpha + \hat{\mathbf{e}}_{\bar{\alpha},\beta} \tag{7.23}$$

and the distortion level by the inter-phone smearing could be estimated with the cosine distance

$$I^{(\mathrm{II,\,c})}_{\alpha,\beta} = 1 - \cos(\angle \mathbf{e}_\alpha - \angle \mathbf{e}_{\alpha,\beta}) \tag{7.24}$$

$$= 1 - \frac{\mathbf{e}_\alpha \cdot \mathbf{e}_{\alpha,\beta}}{\|\mathbf{e}_\alpha\| \|\mathbf{e}_{\alpha,\beta}\|} \tag{7.25}$$

or with the angle rotated

$$I^{(\mathrm{II,\,a})}_{\alpha,\beta} = \frac{1}{\pi} \left| \cos^{-1} \left(\frac{\mathbf{e}_\alpha \cdot \mathbf{e}_{\alpha,\beta}}{\|\mathbf{e}_\alpha\| \|\mathbf{e}_{\alpha,\beta}\|} \right) \right|. \tag{7.26}$$

Therefore the overall inter-phone smearing level by the β-th RIR over all phonemes is estimated with the overall cosine distance

$$I_{\beta}^{(\mathrm{II,\,c})} = \sum_{\alpha} P(\alpha) I_{\alpha,\beta}^{(\mathrm{II,\,c})} \tag{7.27}$$

or the overall rotated angle

$$I_{\beta}^{(\mathrm{II,\,a})} = \sum_{\alpha} P(\alpha) I_{\alpha,\beta}^{(\mathrm{II,\,a})}. \tag{7.28}$$

7.3.2 Combination of smearing indices

Section 7.3.1 has detailed the estimation of reverberation distortion level with the intra-phone smearing and the inter-phone smearing per phoneme. The overall reverberation distortion level could be approximated by combining the two parts additively. When the inter-phone smearing level is estimated with the angular distance,

$$I_{\beta}^{(\mathrm{a})}(N_{\mathrm{f}}) = r^{(\mathrm{I})} I_{\beta}^{(\mathrm{I})}(N_{\mathrm{f}}) + \lambda r^{(\mathrm{II})} I_{\beta}^{(\mathrm{II,\,a})} \tag{7.29}$$

and when the inter-phone smearing level is estimated with the cosine distance,

$$I_{\beta}^{(\mathrm{c})}(N_{\mathrm{f}}) = r^{(\mathrm{I})} I_{\beta}^{(\mathrm{I})}(N_{\mathrm{f}}) + \lambda r^{(\mathrm{II})} I_{\beta}^{(\mathrm{II,\,c})}. \tag{7.30}$$

The λ is a tuning parameter to balance the value range of the two parts. The $r^{(\mathrm{I})}$ and $r^{(\mathrm{II})}$ are the ratios of intra-phone smearing distortion and inter-phone smearing distortion in the overall reverberation distortion respectively. For each phoneme, similarly

$$I_{\alpha,\beta}^{(\mathrm{a})}(N_{\mathrm{f}}) = r_{\alpha}^{(\mathrm{I})} I_{\alpha,\beta}^{(\mathrm{I})}(N_{\mathrm{f}}) + \lambda r_{\alpha}^{(\mathrm{II})} I_{\alpha,\beta}^{(\mathrm{II,\,a})} \tag{7.31}$$

$$I_{\alpha,\beta}^{(\mathrm{c})}(N_{\mathrm{f}}) = r_{\alpha}^{(\mathrm{I})} I_{\alpha,\beta}^{(\mathrm{I})}(N_{\mathrm{f}}) + \lambda r_{\alpha}^{(\mathrm{II})} I_{\alpha,\beta}^{(\mathrm{II,\,c})} \tag{7.32}$$

where $r_{\alpha}^{(\mathrm{I})}$ and $r_{\alpha}^{(\mathrm{II})}$ are the ratio of intra-phone smearing and inter-phone smearing over phoneme α respectively.

The estimation of the intra-phone smearing ratio and the inter-phone smearing ratio is detailed below. Denote the average duration of phoneme α as $\bar{d}_{\alpha}$. For the τ-th STFT in the complex spectrogram of the reverberant recordings for phoneme α, namely $Y(\tau,k)$, according to Eq. (6.10) the chance of intra-phone smearing could be approximated by the ratio between the number of summation components in Eq. (6.10) from the same phoneme and the number of all summation components, *i.e.*

$$r_{\alpha}^{(\mathrm{I})}(\tau) = \frac{n_{\alpha}^{(\mathrm{I})}(\tau)}{M} \tag{7.33}$$

where M is the length of the RIR spectrogram and $n_\alpha^{(\mathrm{I})}$ the number of summation components in Eq. (6.10) from the same phoneme, thus

$$n_\alpha^{(\mathrm{I})}(\tau) = \min\{\tau, M\}. \tag{7.34}$$

For the whole phoneme α, the ratio of intra-phone smearing is thus

$$\begin{aligned} r_\alpha^{(\mathrm{I})} &= \frac{\sum_{\tau=0}^{\bar{d}_\alpha - 1} r_\alpha^{(\mathrm{I})}(\tau)}{\sum_{\tau=0}^{\bar{d}_\alpha - 1} 1} \\ &= \frac{1}{\bar{d}_\alpha} \sum_{\tau=0}^{\bar{d}_\alpha - 1} r_\alpha^{(\mathrm{I})}(\tau) \end{aligned} \tag{7.35}$$

when $\bar{d}_\alpha \leq M$,

$$\begin{aligned} r_\alpha^{(\mathrm{I})} &= \frac{1}{\bar{d}_\alpha} \sum_{\tau=0}^{\bar{d}_\alpha - 1} \frac{\tau}{M} \\ &= \frac{1}{\bar{d}_\alpha} \cdot \frac{\bar{d}_\alpha^2}{2M} \\ &= \frac{\bar{d}_\alpha}{2M} \end{aligned} \tag{7.36}$$

and when $\bar{d}_\alpha > M$,

$$\begin{aligned} r_\alpha^{(\mathrm{I})} &= \frac{1}{\bar{d}_\alpha} \Big(\sum_{\tau=0}^{M-1} \frac{\tau}{M} + \sum_{\tau=M}^{\bar{d}_\alpha - 1} \frac{M}{M} \Big) \\ &= \frac{1}{\bar{d}_\alpha} \Big(\frac{M^2}{2M} + (\bar{d}_\alpha - M) \Big) \\ &= 1 - \frac{M}{2\bar{d}_\alpha} \end{aligned} \tag{7.37}$$

therefore

$$r_\alpha^{(\mathrm{I})} = \begin{cases} \frac{\bar{d}_\alpha}{2M}, & \text{if } \bar{d}_\alpha \leq M \\ 1 - \frac{M}{2\bar{d}_\alpha}, & \text{if } \bar{d}_\alpha > M \end{cases} \tag{7.38}$$

thus the intra-phone smearing ratio for all phonemes is

$$r^{(\mathrm{I})} = \sum_\alpha P(\alpha) r_\alpha^{(\mathrm{I})}. \tag{7.39}$$

If the starting silence or pause before each phoneme is neglected, the ratio of inter-phone smearing can be approximated with

$$r_{\alpha}^{(\mathrm{II})} = 1 - r_{\alpha}^{(\mathrm{I})} \tag{7.40}$$

and the overall inter-phone smearing ratio is

$$r^{(\mathrm{II})} = 1 - r^{(\mathrm{I})}. \tag{7.41}$$

7.4 Fisher Ratio Based Discrimative Analysis

There is one implicit assumption made by the phonetic analysis inspired reverberation measurement in Section 7.3.1, *i.e.* the reverberation distortion level over one phoneme is independent from the reverberation distortion level over other phonemes. This implicit assumption can be problematic in the context of pattern recognition.

For example in a phoneme classification task, there are some phonemes easily confused with each other and they have small distances in feature space. For such phonemes even a small degree of feature distortion could cause a big degradation in the phoneme classification performance. In comparison there are also some phonemes being very different to other phonemes with a large distance in feature space. Such phonemes could be less sensitive to a small amount of feature distortion. The reverberation distortion measurement proposed in Section 7.2 and Section 7.3.1 does not take into account that such phonetic difference in classification difficulty is determined by the statistic properties of all phonemes jointly. Therefore this section suggests a Fisher Ratio based first order discriminative analysis on how the feature discriminability of different phonemes changes under reverberation distortion. The analysis is to reveal how much error the implicit assumption have introduced to the estimation of reverberation distortion level.

The Fisher Ratio based discriminative analysis is conducted on the phonetic average power spectrum which has been used in Section 7.3.1 to represent each phoneme in a hyper space for inter-phone smearing estimation. The discriminative analysis is performed among all phoneme classes in each frequency bin first and then the Fisher Ratio is averaged across frequencies. The results calculated this way will be referred to as the "discriminative score". Therefore the overall discriminative score for all phonemes are calculated with

$$J_{\mathrm{F}} = \frac{1}{N-1} \sum_{k=1}^{N-1} J_{\mathrm{F}}(k) \tag{7.42}$$

$$J_{\mathrm{F}}(k) = \frac{v_{\mathrm{b}}(k)}{v_{\mathrm{w}}(k)} \tag{7.43}$$

where $v_{\mathrm{b}}(k)$ is the between-class variance and $v_{\mathrm{w}}(k)$ is the within-class variance.

$$v_{\mathrm{b}}(k) = \sum_{\alpha} P(\alpha)(\bar{e}_{\alpha}(k) - \bar{e}(k))^2 \tag{7.44}$$

where the average power spectrum overall phonemes is

$$\bar{e}(k) = \sum_{\alpha} P(\alpha)\bar{e}_{\alpha}(k). \tag{7.45}$$

Similarly, the within-class variance is

$$v_{\mathrm{w}}(k) = \sum_{\alpha} P(\alpha)v_{\alpha}(k) \tag{7.46}$$

and the variance for phoneme α is

$$v_{\alpha}(k) = \frac{1}{L_{\alpha}} \sum_{l=0}^{L_{\alpha}-1} (e_{\alpha,l}(k) - \bar{e}_{\alpha}(k))^2. \tag{7.47}$$

For each phoneme, the discriminative score is calculated by averaging the discriminative scores from the Fisher discriminative analysis conducted in pair with all other phonemes one by one.

$$J_{\alpha,\mathrm{F}}(k) = \frac{1}{Q-1} \sum_{\beta} J_{\alpha,\beta,\mathrm{F}}(k) \tag{7.48}$$

$$J_{\alpha,\beta,\mathrm{F}}(k) = \frac{v_{\alpha,\beta,\mathrm{b}}(k)}{v_{\alpha,\beta,\mathrm{w}}(k)} \tag{7.49}$$

where Q is the number of phonemes in total. The between-class variance and within-class variance are respectively

$$v_{\alpha,\beta,\mathrm{b}}(k) = \Big(\bar{e}_{\alpha}(k) - \frac{P(\alpha)\bar{e}_{\alpha}(k) + P(\beta)\bar{e}_{\beta}(k)}{P(\alpha)+P(\beta)}\Big)^2 \tag{7.50}$$

$$v_{\alpha,\beta,\mathrm{w}}(k) = \frac{P(\alpha)v_{\alpha}(k) + P(\beta)v_{\beta}(k)}{P(\alpha)+P(\beta)}. \tag{7.51}$$

Thus

$$J_{\alpha,\mathrm{F}} = \frac{1}{N-1} \sum_{k=1}^{N-1} J_{\alpha,\mathrm{F}}(k) \tag{7.52}$$

Therefore, the discriminative score overall all phonemes J_{F} and the overall discriminative score per phoneme $J_{\alpha,\mathrm{F}}$ could be calculated on the power spectrum of headset recordings and and on the power spectrum of reverberant recordings. If the change in

discriminative score caused by reverberation distortion is correlated with the change in phoneme classification or word recognition performance, the implicit assumption mentioned before could have introduced significant errors in reverberation measurement. If the change in discriminative score is not correlated with the change in phoneme classification or word recognition performance, the errors introduced by the implicit assumption might be neglected.

7.5 Experiment Results

The reverberation scores are evaluated by their correlation with the PER in the phoneme classification task or its correlation with the WER in the speech recognition task. The Spearman rank correlation is preferred over the Pearson linear correlation in most scenarios because the work in this chapter has targeted the application of reverberation measurement on the model selection for DSR implemented at utterance level. When the rank correlation is very low or the rank correlations are similar when comparing multiple scenarios, the linear correlation will be highlighted.

The work in this section evaluates the performance of reverberation score in a more extensive way than existing literature. A good reverberation score should be capable of depicting both the reverberation level in different environment and microphone channels and the signal dependent reverberation sensitivity. Therefore the evaluation will involve the correlation based on the statistics estimated at multiple different scales. To avoid the confusion caused by multiple factors in the correlation calculation, the following phrase patterns will be used when referring to a correlation:

The correlation between A and B regarding the C difference.
The correlation between A and B regarding C.

They both mean that "A"s and "B"s are first grouped and averaged by the value of the influence factor "C", before calculating the correlation between the averaged "A"s and the averaged "B"s. For example,

The correlation between the reverberation score and the PER regarding the speaker difference.

actually means that the reverberation score and PER in each environment condition are first averaged per speaker before calculating their correlation, given multiple environment conditions.

The estimation of signal dependent reverberation sensitivity is conducted at three scales: speaker, utterance and phoneme. When the reverberation score is evaluated regarding its

capability of depicting the speaker individual difference in the reverberation sensitivity, the reverberation score is averaged per speaker and the WER is averaged per speaker as well, *i.e.*

The correlation between reverberation score and WER regarding the speaker difference.

When the reverberation score is evaluated regarding its capability of depicting the signal dependent difference in the reverberation sensitivity at utterance level, the reverberation score is averaged per utterance and the WER is calculated per utterance as well, *i.e.*

The correlation between reverberation score and WER regarding the utterance difference.

When the reverberation score is evaluated regarding its capability of depicting the phoneme difference in the reverberation sensitivity, the reverberation score is averaged per phoneme and the PER is averaged per phoneme as well, *i.e.*

The correlation between reverberation score and PER regarding the phoneme difference.

When the reverberation score is evaluated regarding its capability of depicting the environment and channel reverberation level, the reverberation score is averaged per channel and the WER is averaged per channel as well, *i.e.*

The correlation between reverberation score and WER regarding the channel difference.

7.5.1 Polynomial reverberation measurement

Section 7.2 proposed a polynomial method to estimate the reverberation distortion level based on the signal spectrogram and the RIR spectrogram, and the produced polynomial reverberation score is an alternative to the C_{50} based reverberation score as the reference for training non-intrusive reverberation estimators. As mentioned in Section 7.1.1 and Section 7.1.2, there are two major motivations for the work on polynomial reverberation score, *i.e.* increasing the distortion estimation accuracy of early reverberation and improving the correlation between the estimated reverberation score and the performance of pattern recognition task on short recordings. Therefore the experiments in this section evaluate the polynomial reverberation score in comparison with the popular C_{50} based reverberation score from these two aspects.

The reverberation measurement experiments are conducted on reverberant data simulated by convolving SWC headset recordings with RIRs measured in SWC recording room. The RIRs are measured with the same speaker location from 13 different microphones, *i.e.* 8 microphones hanging from the ceiling grid ("GRID-0*"), 4 microphones installed on the

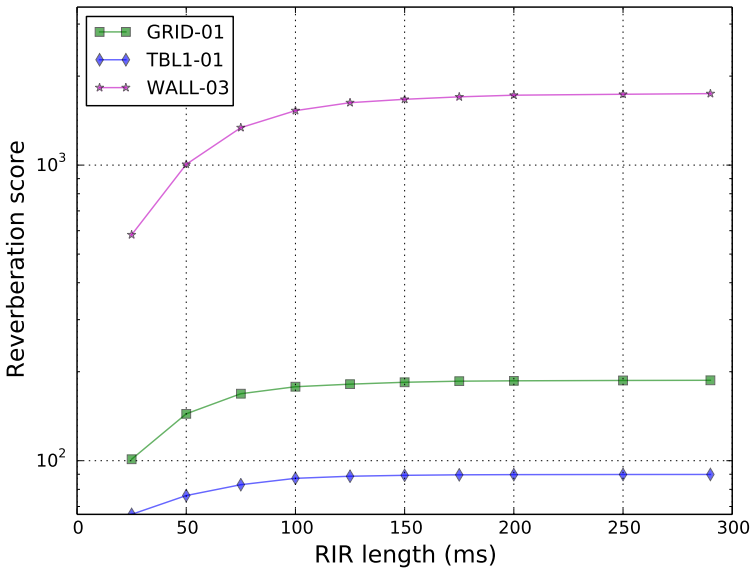
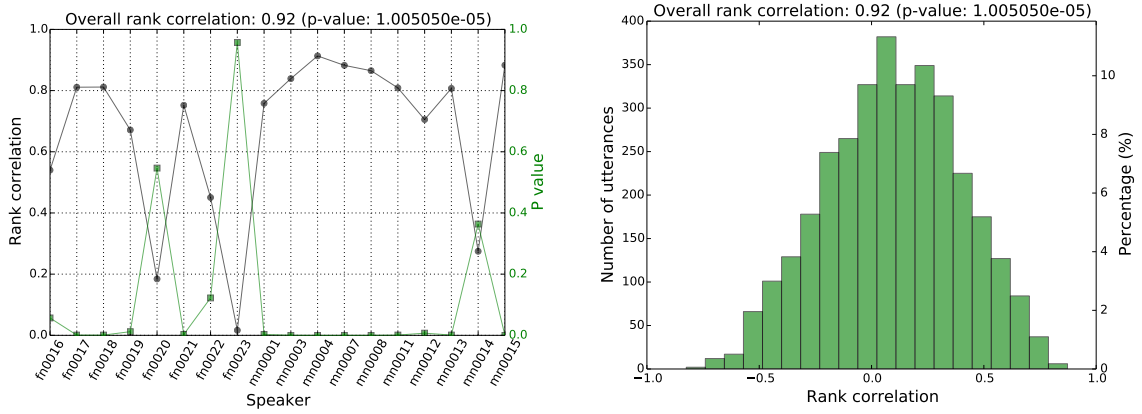


Fig. 7.5 *Overall polynomial reverberation score as the length of truncated RIRs increases.*

wall ("WALL-0*") and one microphone located at the center of the table ("TBL1-01"). One reverberation score is calculated per utterance. The reverberation scores for the same speech utterance in different reverberant conditions are compared with the WERs on the same utterance based on simulated data of corresponding reverberant conditions. To avoid the mismatch in reverberant condition between training and test, WERs are calculated from the decoding results of DNN-HMM hybrid systems which are trained and tested independently over the simulated data of each reverberation condition. The remaining setup in ASR system follows the standalone training system detailed in Section 4.4.3.

In the experiments evaluating the performance of polynomial reverberation score on early reverberation distortion estimation, the RIRs are truncated in a similar way with Section 7.1.1 to create the reverberant data with different levels of early reverberation distortion. Fig. 7.5 shows the overall polynomial reverberation score as the length of truncated RIRs increases. The overall reverberation score is an average of reverberation score on all utterances. The plot should be compared with the line plot in Fig. 7.1a regarding WERs on eval dataset.

As shown in Fig. 7.5, the overall polynomial reverberation score increases incrementally as the RIR length increases. The increase in reverberation score is consistent with the WER increase shown in Fig. 7.1a as the length of the truncated RIR increases. Therefore the different levels of distortion by early reverberation is reflected by the proposed reverberation score. In terms of microphone channel, there is a very clear difference among the reverberation scores corresponding to different microphone channels regardless of



(a) *Spearman rank correlation between the reverberation score and the WER regarding the speaker difference.*

(b) *Histogram of the Spearman rank correlation between the utterance level reverberation score and the utterance level WER regarding the channel difference.*

Fig. 7.6 *Correlation between the polynomial reverberation score and the WER regarding the channel difference.*

the RIR truncation length. In comparison, as shown in Fig. 7.1a the difference among WERs corresponding to different channels is very small when the RIRs are truncated to be less than 50 ms. In addition, the vertical distance among the reverberation curves is not proportional to the vertical distance among the WER curves in Fig. 7.1a. These two observations imply some potential normalisation difficulty in the polynomial reverberation score, which will be discussed further in Section 7.6.

To evaluate the performance of the polynomial reverberation score on limited recording data at speaker level and on short recordings at utterance level, an experiment is conducted with the previously mentioned 13 RIRs. The experiment configuration is the same as before except that the RIRs are not truncated. Fig. 7.6a illustrates the Spearman rank correlation between the polynomial reverberation score and WER regarding the speaker difference. Fig. 7.6b shows the histogram of the Spearman rank correlation between the polynomial reverberation score on each utterance and the WER of corresponding utterance regarding reverberation condition.

As shown, the Spearman rank correlation between the overall WER and the polynomial reverberation score over all utterances in eval dataset is 0.92, which is slightly worse compared to C_{50} which is -0.94 (Fig. 7.3). Comparing Fig. 7.6a with Fig. 7.3a which evaluates the Spearman rank correlation speaker by speaker, the polynomial reverberation measurement did not improve the correlation on the speakers who had low rank correlation between C_{50} and speaker level WER. Comparing Fig. 7.6b with Fig. 7.3b, both C_{50} and

polynomial reverberation score have a wide distribution of the Spearman rank correlation between the reverberation score and the WER regarding the utterance difference.

In summary, similar to C_{50} the polynomial reverberation score provides a high rank correlation with WER over a large amount of data, and it well reflects the different reverberation distortion levels in different environments and channels. When the amount of data is reduced and the rank correlation between the reverberation score and the WER is calculated regarding the speaker difference or the utterance difference, the polynomial reverberation score does not improve over C_{50}.

7.5.2 Phoneme duration and smearing ratio

As shown in Eq. (7.38) in Section 7.5.2, the ratio of intra-phone smearing is dependent on the phoneme duration and the effective length of the RIR. Therefore a survey of average phoneme duration is first conducted using TIMIT data (Garofolo et al., 1993) which has accurate manual annotation of phonemes and phonetic boundary. As shown in Fig. 7.7, there is a large variation in the average phoneme duration across different phonemes, from below 20 ms (/b/) to above 150 ms (/aw/, /oy/). In terms of phoneme categories, as shown in Table 7.2, 51.1% of speech duration is occupied by vowels whose average duration per phoneme is 96 ms, followed by fricatives (18.1%) whose average duration per phoneme is 91 ms.

Table 7.2 *Phoneme category definition in the TIMIT corpus ("DPP": average duration per phoneme; "Per.": percentage in overall duration without silence).*

Phoneme category	Phonemes in TIMIT corpus	Per. (%)	DPP (ms)
stop	b, d, g, p, t, k, dx, q, pcl	9.8	41
affricate	jh, ch	1.3	70
fricative	s, sh, z, zh, f, th, v, dh	**18.1**	**91**
nasals	m, n, ng, em, en, eng, nx	7.4	57
semivowel and glide	l, r, w, y, hh, hv, el	12.3	64
vowel	iy, ih, eh, ey, ae, aa, aw, ay, ah, oy, ow, uh, uw, ux, er, ax, ix, axr, ax-h	**51.1**	**96**
overall	all above without silence	100.0	82.5

Since TIMIT is a corpus of read speech recordings with clear pronounciation, more surveys are conducted on conversational speech using the headset recordings from the WSJCAM0 corpus (Robinson et al., 1995), the AMI corpus (McCowan et al., 2005) and the SWC Day One (SWC1) (Fox et al., 2013). As AMI and SWC1 do not have the reference annotation for phonetic boundary, the force alignment is performed with manual transcription and with the in-domain state-of-the-art speech recognition systems based

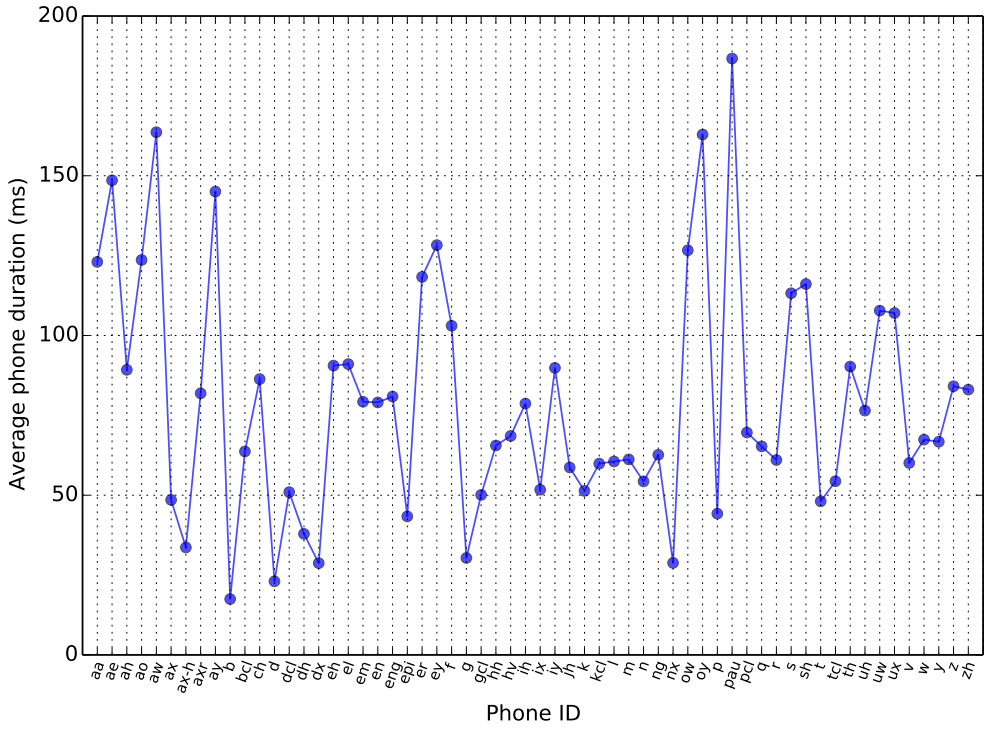


Fig. 7.7 *Average phoneme duration in the TIMIT corpus (/pau/ refers to silence and pause).*

Table 7.3 *Duration and duration percentage of different sounds in conversational English speech. "Dur.": overall duration (hours); "DPP": the average duration per phoneme (ms); "Pct.": percentage in overall duration without silence (%).*

Phoneme category	Phonemes	AMI Corpus			SWC1			WSJCAM0		
		Dur.	Pct.	DPP	Dur.	Pct.	DPP	Dur.	Pct.	DPP
stop	b, d, g, k, p, t	10.04	16.1	67	0.91	16.5	70	6.27	21.4	80
affricate	ch, jh	0.71	1.1	91	0.06	1.0	85	0.50	1.7	118
fricative	dh, f, hh, s, sh, th, v, z, zh	10.27	**16.4**	**84**	0.92	**16.7**	**81**	6.76	**23.0**	**98**
nasal	em, en, m, n, ng	6.83	10.9	86	0.61	11.0	83	3.13	10.7	69
liquid	el, l, r	4.10	6.6	75	0.36	6.6	77	1.99	6.8	65
semivowel	w, y	2.84	4.5	75	0.26	4.7	82	0.82	2.8	84
vowel	aa, ae, ah, ao, aw, ax, axr, ay, eh, er, ey, ih, iy, ow, oy, uh, uw	27.69	**44.3**	**90**	2.39	**43.4**	**90**	9.90	**33.7**	**72**
overall	all above without silence	62.5	100.0	83	5.5	100.0	82	30.3	100.0	80

on headset recordings. The acoustic models are based on the bottleneck features from the DNN front-end whose configuration follows the setup published by Liu et al. (2016,

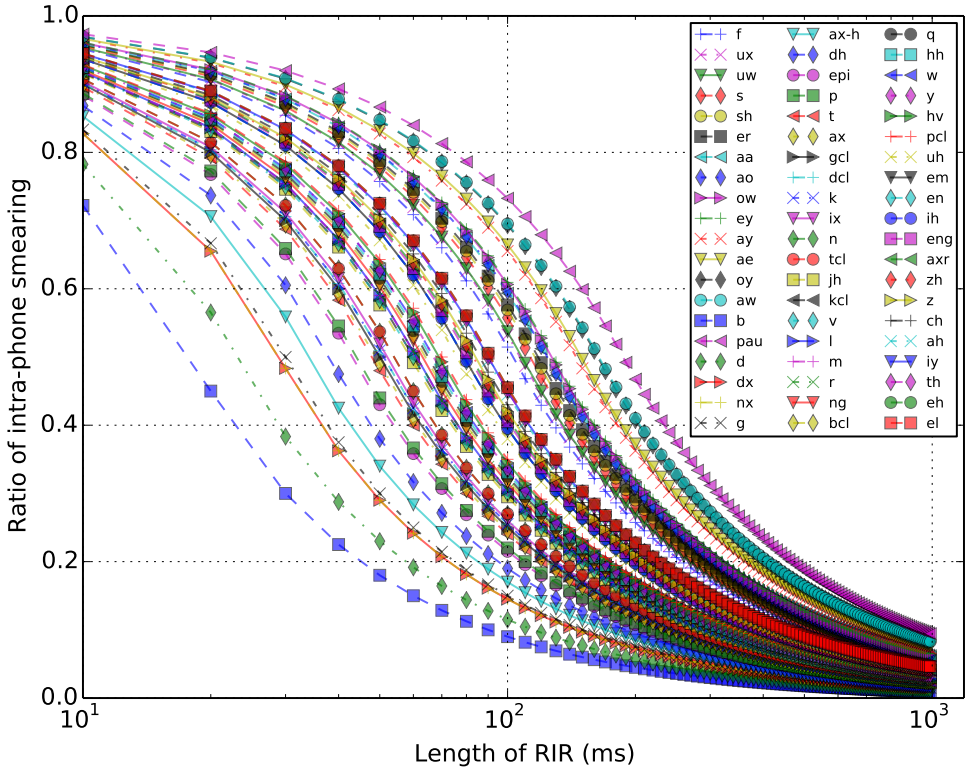


Fig. 7.8 *Based on the average phoneme duration for each phoneme in TIMIT.*

2014). The average phoneme duration in these three corpora are shown in Table 7.3. As shown, the duration proportion and average duration of each phoneme category are very similar between AMI corpus and SWC recordings, considering potential alignment errors. As shown in Table 7.3, more than half of the duration in the conversational English speech are phonemes with an average duration from 80 ms to 90 ms, which is similar to the read speech in the TIMIT corpus (Table 7.2). Compared to TIMIT, AMI and SWC1, the WSJCAM0 corpus is slightly different in the duration percentage of each phoneme category and the average duration per phoneme category. This could be potentially caused by a different speaking style or the alignment errors.

Taking the average phoneme durations in TIMIT as examples because of the accurate manual phonetic annotation, Fig. 7.8 shows the ratio of intra-phone smearing calculated with Eq. (7.38) as the length of RIR increases. As shown when the RIR is shorter than 100 ms, there is a large variation in the ratio of intra-phone smearing. As the length of RIR increases, there is a higher ratio of inter-phone smearing than intra-phone smearing, and a smaller variation among phonemes in terms of intra-phone smearing ratio. In particular, when the RIR length is 50 ms, the intra-phone smearing probability is 67.4% in TIMIT.

Therefore, when using 50 ms as the early-late reverberation boundary, there is up to 30% of inter-phone smearing, *i.e.* reverberation smearing from preceding phonemes, and the ratio of inter-phone smearing can be particularly large for short phonemes such as /b/, /n/, /dx/, /g/, /ax-h/, /epi/ and /p/.

Since the phoneme duration statistics is similar across multiple corpora of different speaking styles and different speech topics (Table 7.2 and Table 7.3), it can be expected that the conclusions about the ratio of intra-phone smearing based on the phoneme duration of the TIMIT corpus are also transferable to other corpora.

7.5.3 Phonetic analysis inspired reverberation measurement

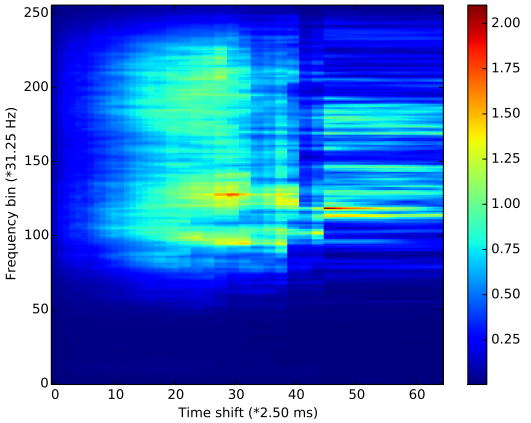
For the phonetic analysis inspired reverberation measurement, besides the polynomial reverberation score which will be applied at a phonetic level to estimate the intra-phone smearing index, there are two other components introduced. First, the normalised magnitude-spectrum variance as shown in Eq. (7.10) was introduced to emphasise the phonetic difference in energy distribution over frequency. Second, the inter-phone smearing index as shown in Eq. (7.27) or Eq. (7.28) is introduced to emphasise the impact of power spectrum difference between neighbouring phonemes in reverberation distortion. Since the phonetic analysis inspired reverberation measurement employs three components, for a better understanding of each component the experiments are organized in the following way. First each of the three components is investigated independently, in terms of its capability in depicting the reverberation distortion level in different channels, as well as its capability in depicting the signal dependent sensitivity to the reverberation distortion from the same channel. Then the components are combined and any complementary benefit from the combination will be investigated. The three components are combined in two stages. First the normalised magnitude-spectrum variance is combined with the polynomial reverberation score as intra-phone smearing index, with the combination weighting parameter to be tuned to explore the maximal complementary benefit in the combination. Then the intra-phone index is combined with inter-phone smearing index, and the combination weighting parameter is again tuned.

As shown in Section 7.3.1, phonetic analysis inspired reverberation measurement combining the intra-phone smearing index and the inter-phone smearing index requires phonetic spectrogram statistics. Therefore a reliable phonetic annotation is needed for calculating spectrogram statistics for each phoneme. For this reason the WSJCAM0 corpus is chosen for the experiments in this section since it has a reference annotation for the phonetic boundaries and its data size is sufficient to construct advanced phoneme classifier based on DNNs. The performance of the reverberation measurement are evaluated by the

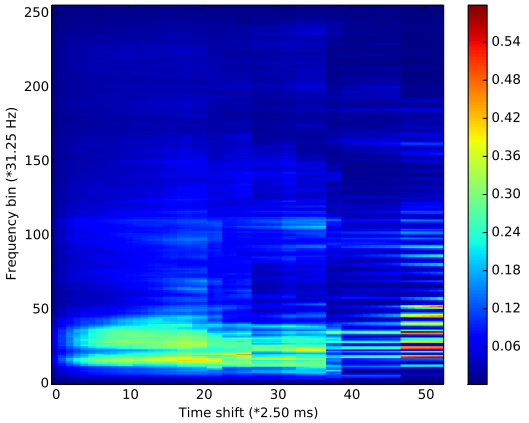
correlation between the reverberation score and the PER in a phoneme classification task. There are two reasons why the PER is preferred to the WER at this stage. First, it allows a direct comparison between the phonetic analysis inspired reverberation score and PER on the same phoneme. Second, the PERs are the results from the system described in Section 7.1.2 which has already implied a high correlation between PERs and WERs when PERs are calculated with the reference phonetic boundary. In the experiments in Section 7.1.2, C_{50} has shown high Spearman rank correlations with both the overall WER and the overall PER regarding the channel difference, indicating that overall there is a high correlation between PER and WER and a good reverberation score should provide high correlations with both PER and WER at the same time. Therefore this section evaluates the performance of reverberation measurement based on the PER first, using reverberant data simulated by convolving WSJCAM0 headset recordings with 13 RIRs from the SWC recording room (same with Section 7.1.2). If good performance is observed, further experiments could be conducted to evaluate the reverberation measurement performance against WER.

Experiments are first conducted to evaluate the performance of the intra-phone smearing index based on polynomial reverberation score applied at phonetic level alone. Fig. 7.9 illustrates the speech dynamic indices for phoneme /ch/ and phoneme /oy/, acquired with 100 examples of each phoneme. The pattern in speech dynamic index along the frequency axis (Fig. 7.9a and Fig. 7.9b) matches the average power spectrum of the corresponding phoneme (Fig. 7.9c and Fig. 7.9d). The analytic formula in Eq. (7.7) suggests that the polynomial reverberation score is based on the element-wise multiplication between the speech dynamic index and the RIR magnitude spectrogram. Therefore the speech dynamic index examples shown in Fig. 7.9 suggest a fundamental difference between the phonetic polynomial reverberation score and the existing reverberation score: the final polynomial reverberation score is mainly determined by the spectrogram elements of the RIR in the frequency bins where the phoneme has high energy.

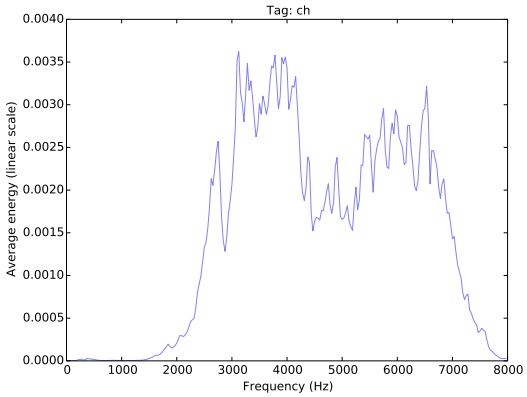
Fig. 7.10a shows the overall intra-phone smearing index based on the polynomial reverberation score in various reverberation conditions (black line) and the corresponding overall WER (green line). The Spearman rank correlation between the two lines is 0.88, indicating that the intra-phone smearing index based on polynomial reverberation score applied on phonetic level alone can depict the channel difference in reverberation to a large degree. This correlation is slightly lower compared to the Spearman rank correlation of -0.94 between C_{50} and PER (Section 7.1.2), as well as the previously reported Spearman rank correlation of 0.92 between the polynomial reverberation score implemented at utterance level and the overall WER on SWC data regarding the channel difference (Section 7.5.1). Such degradation is expected because by applying the polynomial reverberation score on a smaller time scale, *i.e.* from utterance level to phoneme level, the information



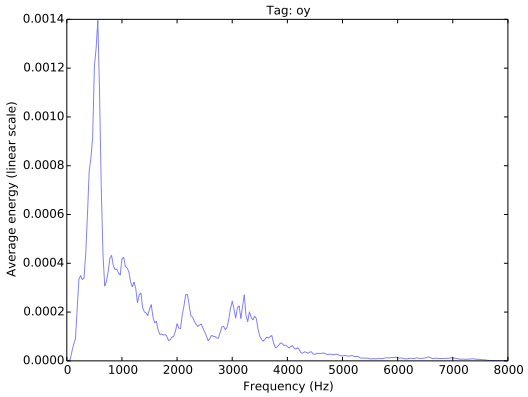
(a) *Speech dynamic index for phoneme /ch/.*



(b) *Speech dynamic index for phoneme /oy/.*



(c) *Average power spectrum for phoneme /ch/.*

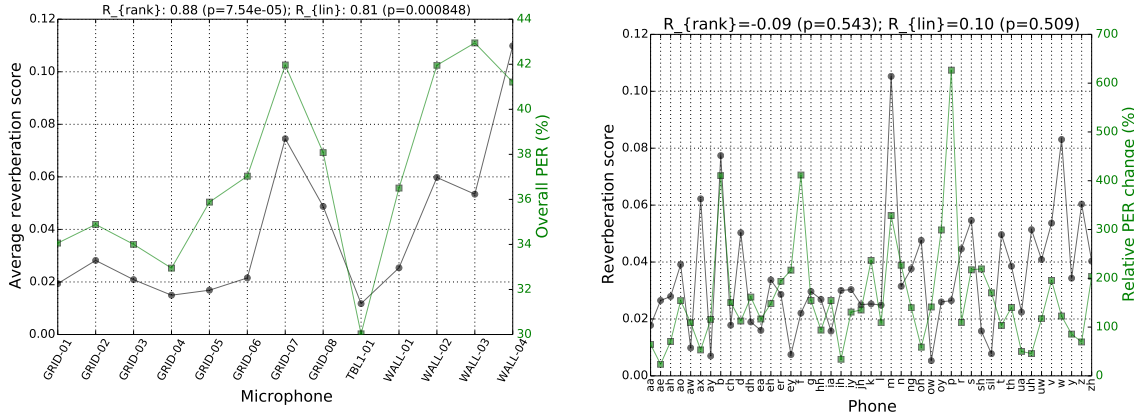


(d) *Average power spectrum for phoneme /oy/.*

Fig. 7.9 *Examples of speech dynamic index and average power spectrum.*

about reverberation smearing across phonemes is cut off, leading to inevitable performance degradation.

Fig. 7.10b examines the capability of the polynomial reverberation score based intra-phone smearing index in depicting the phonetic reverberation sensitivity across multiple reverberant conditions. The black line shows the reverberation score and the green line shows the relative change in the PER from headset recordings to simulated reverberant data, both averaged over data of all 13 reverberant conditions on the corresponding phoneme. Fig. 7.10b suggests that the intra-phone smearing index based on the polynomial reverberation score could not reflect the phonetic difference in reverberation sensitivity, as the Spearman rank correlation between the reverberation score and the PER regarding the phoneme difference is -0.09. This is expected because in Section 7.5.1 it has been shown that the polynomial reverberation score alone could not track the signal difference in reverberation sensitivity in short recordings well (Fig. 7.6b).



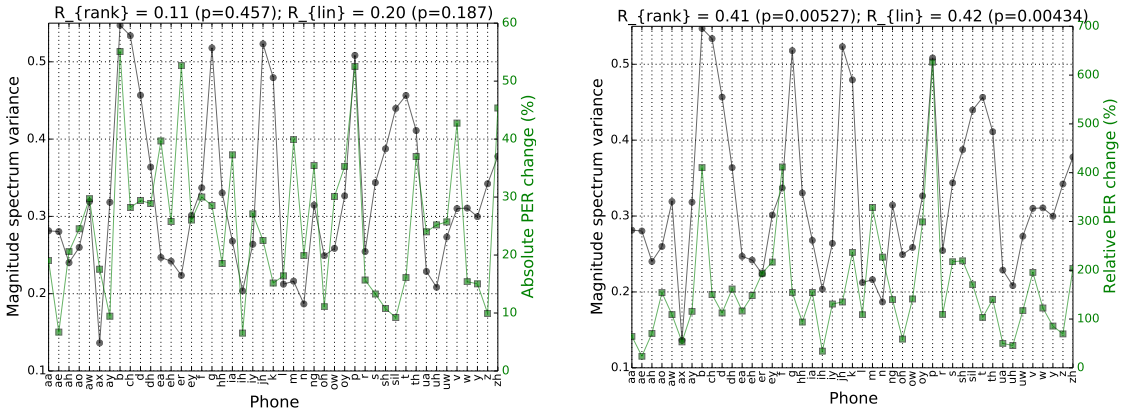
(a) *Polynomial reverberation score per channel and overall PER per channel.*

(b) *Polynomial reverberation score per phoneme and average PER increase per phoneme in 13 reverberant conditions.*

Fig. 7.10 *Polynomial reverberation score for intra-phone smearing and PER change due to channel difference and phonetic difference, on the WSJCAM0 evaluation set of simulated reverberant data.*

Further experiments are conducted to examine the properties of the first component introduced in phonetic analysis inspired reverberation measurement, *i.e.* the normalised magnitude spectrum variance. Since the normalised magnitude spectrum variance is independent from RIRs, it is only evaluated regarding its capability of depicting the phonetic difference in reverberation distortion sensitivity given the same reverberant environment and channel. Fig. 7.11 compares the normalised magnitude spectrum variance (black lines) with the absolute PER increase from headset recordings to simulated reverberant data (green line in Fig. 7.11a) and the relative PER increase from headset recordings to simulated reverberant data (green line in Fig. 7.11b), when both the absolute PER increase and absolute PER increase are the average over all 13 reverberant conditions over corresponding phoneme. The average magnitude spectrum variance per phoneme and the average power spectrum per phoneme are both calculated with 100 examples of corresponding phoneme in WSJCAM0 headset recordings. As shown, overall the normalised magnitude variance is not well correlated with the phonetic difference in terms of average PER degradation in reverberant environments, though it does highlight a few phonemes that are particularly sensitive, such as /b/ and /p/. Further experiments will be conducted later to investigate whether it introduces any complementary effect when combined with the polynomial reverberation score for intra-phone smearing index.

The proposed inter-phone smearing index, *i.e.* the third component in the phonetic analysis inspired reverberation measurement, is designed to be dependent on both signal and RIRs to depict both the phonetic difference in sensitivity to the same reverberant



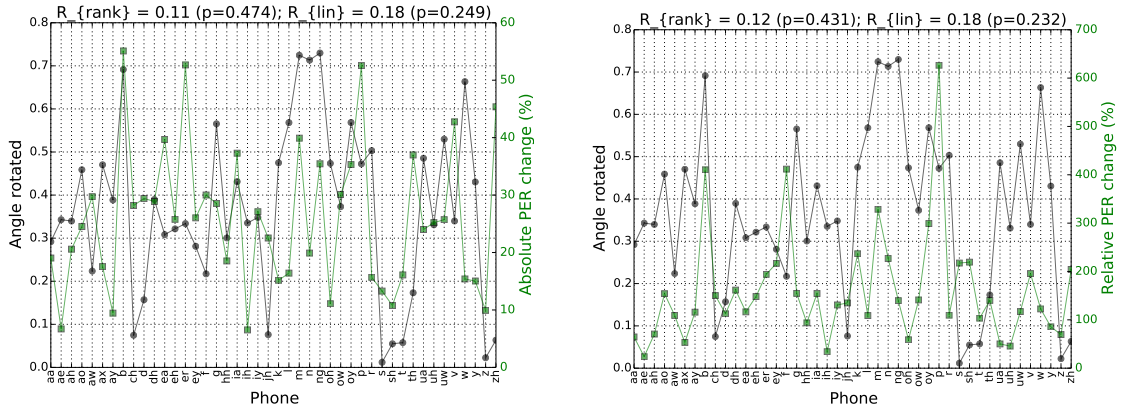
(a) *Average absolute increase in PER.* (b) *Average relative increase in PER.*

Fig. 7.11 *Average PER increase in comparison with magnitude spectrum variance in terms of phoneme difference.*

channel and the channel difference in reverberation distortion level over all phonemes. Therefore its performance is evaluated in both aspects. In the experiments below the inter-phone smearing is implemented with the angle rotated in the high dimensional power spectrum space due to reverberation. Three preceding phones are considered for inter-phone smearing in the experiments. Similar to Fig. 7.11, Fig. 7.12 compares the inter-phone smearing index (black line) with the absolute PER increase from headset recordings to simulated reverberant data (green line in Fig. 7.12a) and the relative PER increase from headset recordings to simulated reverberant data (green line in Fig. 7.12b), all averaged over 13 reverberant conditions per phoneme. As shown, inter-phone smearing index has very limited correlation with the PER degradation regarding the phoneme difference.

Furthermore, Fig. 7.13 compares the overall inter-phone smearing in various reverberant channels (black line) and the overall PER in corresponding channel (green line). As shown, the Spearman rank correlation between the overall inter-phone smearing index and overall PER regarding the channel difference in reverberation is -0.42, indicating very limited capability of the implemented inter-phone smearing index in depicting the reverberation level in different channels. In addition, while the analytic derivation in Section 7.3.1 expects a positive correlation, the results shown in Fig. 7.13 suggest the opposite trend. This point will be discussed further in Section 7.6.

Therefore, experiments conducted so far investigating the three components independently suggest that only the polynomial reverberation score applied at phoneme level for intra-phone smearing could well depict the channel difference in reverberation, while none of the three components alone could well reflect the phonetic properties in terms of



(a) *Average absolute increase in PER.* (b) *Average relative increase in PER.*

Fig. 7.12 *Average PER increase in comparison with average power spectrum vector rotation caused by reverberation (3 preceding phones taken into consideration).*

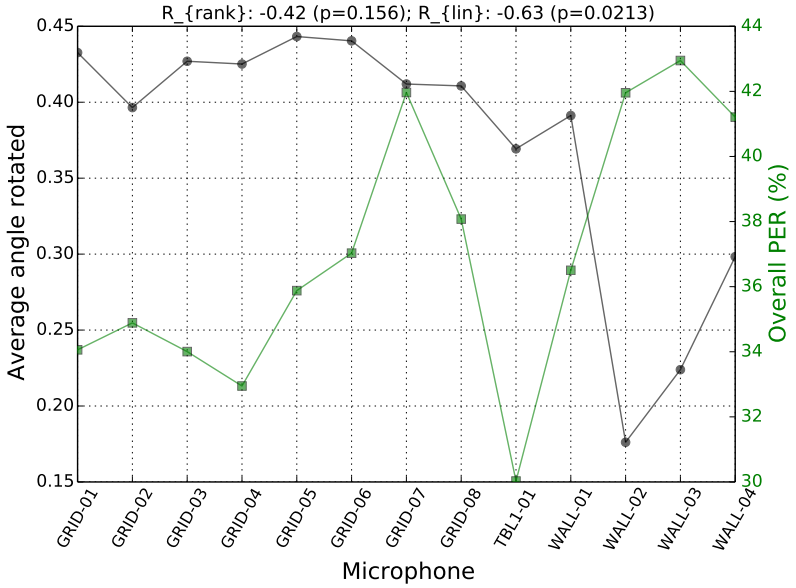
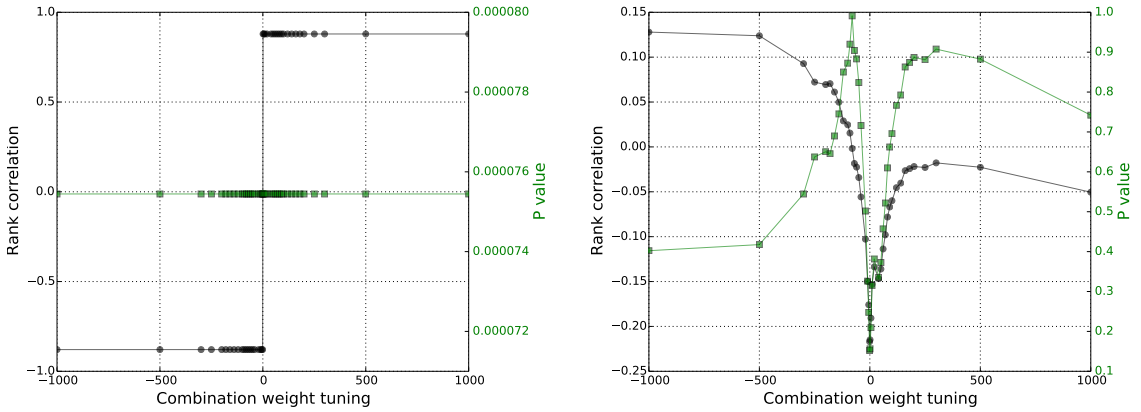


Fig. 7.13 *Average PER increase in comparison with average power spectrum vector rotation caused by reverberation (3 preceding phonemes taken into consideration).*

different reverberation sensitivity. The following experiments in this section investigate the performance of the reverberation score when multiple components are combined.

First the polynomial reverberation score and the phonetic normalised magnitude spectrum variance are combined as the intra-phone smearing index. Since the normalised magnitude spectrum variance is independent from the environment and microphone channel, tuning the combination weighting parameter will not change the performance on measuring the reverberation level of different reverberant channels (black line in Fig.



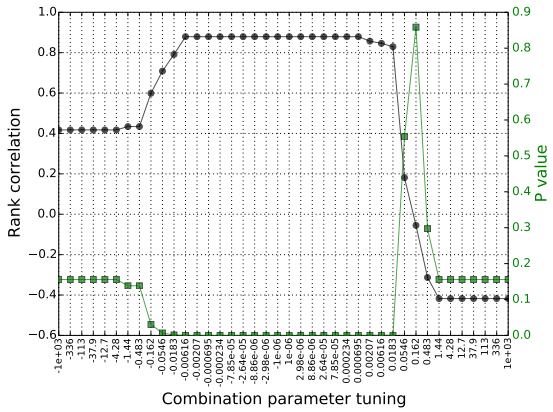
(a) *Spearman rank correlation with overall PER regarding the channel difference.*

(b) *Spearman rank correlation with relative PER increase per phoneme regarding the phoneme difference.*

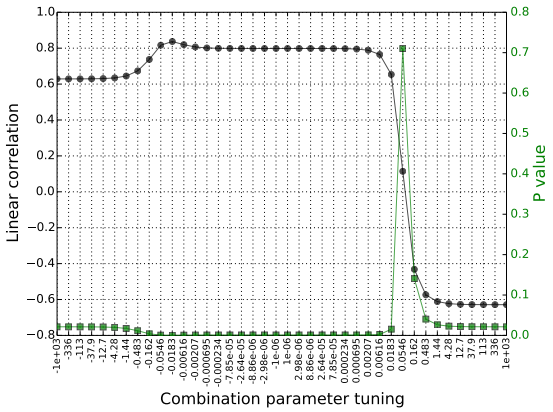
Fig. 7.14 *Correlation change when tuning the combination weight in intra-phone smearing index (black line: Spearman rank correlation; green line: p-value).*

7.14a). The tuning does change the Spearman rank correlation between the intra-phone smearing index and the relative PER increase regarding the phoneme difference, however no significant rank correlation is observed no matter how the combination weighting parameter is tuned, as illustrated by the black line in Fig. 7.14b. This is possibly caused by the results previously observed that neither the polynomial reverberation score nor the normalised magnitude spectrum variance could well depict the phonetic difference in reverberation sensitivity. In addition, there seems no complementary benefit when combining the two components additively together. Therefore in further experiments the normalised magnitude spectrum variance component is dropped from intra-phone smearing index, and the polynomial reverberation score applied at phoneme level is used as intra-phone smearing index to directly combine with inter-phone smearing index.

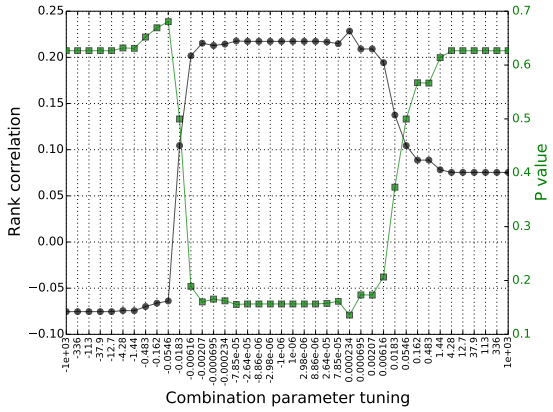
The reverberation score combining the intra-phone smearing index and the inter-phone smearing index is first evaluated by its correlation with the overall PER on simulated data regarding the channel difference. Fig. 7.15a and Fig. 7.15b respectively show the Spearman rank correlation and Pearson linear correlation between the overall reverberation score and PER regarding the microphone channel difference when tuning the combining weight parameter λ (Section 7.3.2). Overall the highest correlation between PER and reverberation score regarding the channel difference is achieved when the combination weight λ has a value close to 0. That suggests adding the inter-phone smearing index to the intra-phone smearing index based on polynomial reverberation score is not beneficial. This may be related to the earlier observation that the rotation angle based inter-phone smearing index does not provide a good indication of the reverberation level in environment and



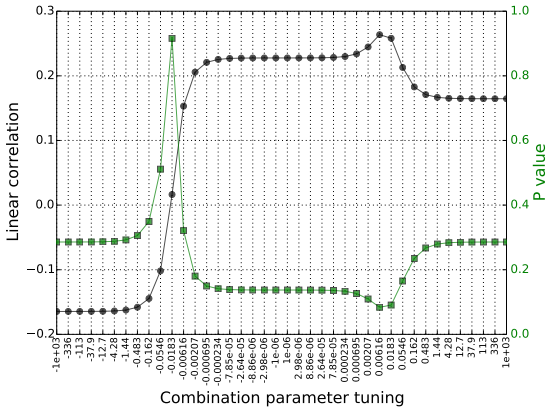
(a) *Spearman rank correlation between the reverberation score and the overall PER regarding the channel difference.*



(b) *Pearson linear correlation between the reverberation score and the overall PER regarding the channel difference.*



(c) *Spearman rank correlation between the reverberation score and the average relative PER increase in 13 simulated reverberant environments, regarding the phoneme difference.*

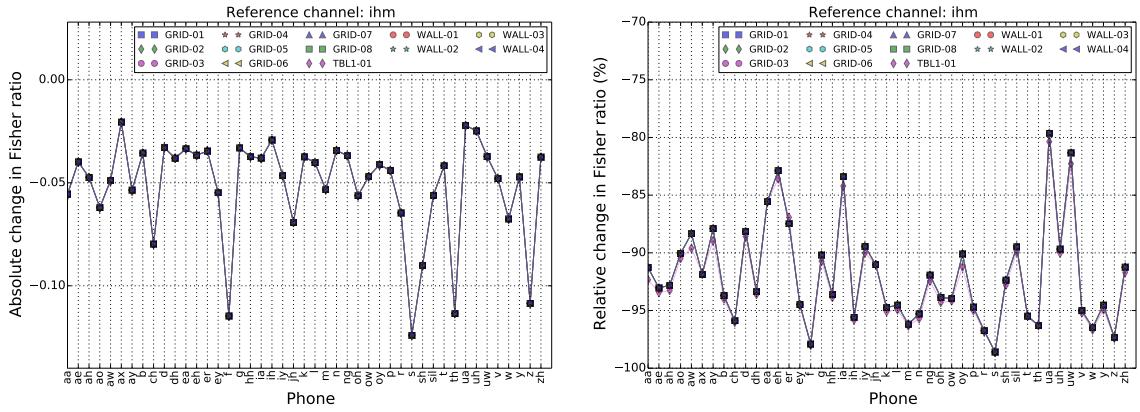


(d) *Pearson linear correlation between the reverberation score and the average relative PER increase in 13 simulated reverberant environments regarding the phoneme difference.*

Fig. 7.15 *Experimental examination of the correlation between reverberation score and the overall PER on WSJCAM0 based simulated data.*

microphone channel. In addition the inter-phone smearing index does not introduce any complementary benefit when combined with the polynomial reverberation score based intra-phone smearing index.

Furthermore, the combined reverberation score is evaluated by its correlation with the relative PER increase from headset recordings to simulated reverberant data regarding the phoneme difference in reverberation sensitivity. Fig. 7.15c and Fig. 7.15d show that inter-phone smearing does not help to improve the correlation with the average relative PER increase caused by reverberation, *i.e.* it does not help to improve the accuracy of reverberation score in depicting the phonetic difference in reverberation sensitivity. This is



(a) *Absolute change in Fisher ratio.* (b) *Relative change in Fisher ratio.*

Fig. 7.16 *The change of Fisher ratio due to reverberation.*

potentially because there is no complementary benefit when combining the intra-phone smearing index based on polynomial reverberation score and the inter-phone smearing index based on the rotation angle, and the earlier analysis in Fig. 7.10b and Fig. 7.12b has shown that each factor alone is good at depicting the phonetic difference in reverberation sensitivity.

7.5.4 Fisher ratio based discriminative analysis

As discussed in Section 7.4, the inter-phone smearing index based on the phonetic power spectrum vector has one implicit assumption: the reverberation distortion level of one phoneme is independent from the reverberation distortion level of other phonemes in the same classification task. The experiments on phoneme classification in Section 7.5.3 showed that the proposed inter-phone smearing index could not well depict either the phonetic difference or the microphone channel effect in reverberation distortion. Such experiment results can be caused by the errors from the implicit assumption. Therefore this section investigates the impact of the implicit assumption on the inter-phone smearing index. The investigation is based on the Fisher linear discriminative analysis of the phonetic power spectrum vector, and it is achieved in two steps. The Fisher ratio is first calculated following the algorithm detailed in Section 7.4, separately on WSJCAM0 headset recordings and on the reverberant data simulated with WSJCAM0 headset recordings and the 13 RIRs from SWC recording room. Then the correlation between the change in Fisher ratio and the change in PER regarding the channel difference and the phoneme difference in reverberation distortion is analysed.

It is found that the overall Fisher ratio J_{F} has a high Pearson linear correlation with PER regarding the microphone channel difference in reverberation distortion, being -0.82 with p-value of 3.7e-4. In comparison the Spearman rank correlation is as low as -0.047, *i.e.* there is no rank correlation. Fig. 7.16 shows the change in the Fisher ratio per phoneme $J_{\alpha,\mathrm{F}}$ due to reverberation. When comparing the average absolute change in Fisher ratio per phone due to reverberation with the average absolute change in PER per phoneme on the same eval dataset due to reverberation, the Spearman rank correlation is 0.22 and the Pearson linear correlation is 0.19. Similarly, when comparing the average relative change in Fisher ratio per phoneme with the average relative change in PER per phoneme on eval dataset due to reverberation, the Spearman rank correlation is -0.17 and the Pearson linear correlation is -0.24. Therefore, the Fisher ratio could only reflect the channel difference but not the phonetic difference regarding the sensitivity to pattern distortion caused by the reverberation.

As the Fisher ratio represents the first order discriminative analysis on the distance among multiple classes in the same classification task, the results implies that the idea of adding components that better depicts the between-class difference is unlikely to improve the capability of reverberation score in estimating signal dependent distortion sensitivity to the same environment reverberation. Or the power spectrum is not a good feature space for that purpose.

7.6 Summary and Discussion

This chapter has detailed the research work on improving the existing reverberation measurement regarding the capability of depicting early reverberation distortion and signal dependent reverberation distortion level in short recordings. Based on the analytic study of reverberation in previous chapter, a polynomial format reverberation score is first proposed in Section 7.2. In the experiments with simulated reverberant data by convolving the SWC headset recordings with RIRs truncated to different length, it is shown that such polynomial reverberation score could well represents the accumulated increase in WER with distortion accumulated from early reverberation and late reverberation. In addition, the polynomial reverberation score shows a high rank correlation with WER regarding the channel difference in reverberation. Therefore the polynomial reverberation score improves the estimation of the early reverberation distortion level compared to the currently popular reverberation score based on ELR. A further investigation of the correlation between the reverberation score using the polynomial method and the WER regarding the utterance difference suggests no improvement over C_{50} based reverberation

score. Thus the polynomial method is yet to improve in its capability of depicting signal dependent reverberation sensitivity in short utterances.

Taking the idea of "self-masking" and "overlap-masking" in the phonetic analysis on the human perception of reverberation by Kokkinakis and Loizou (2011), a new reverberation measurement structure is proposed by partitioning reverberation distortion into intra-phone smearing and inter-phone smearing and by combining the reverberation indices from each part. The polynomial reverberation score is applied at phoneme level to measure intra-phone smearing. In the experiments on simulated data based on the WSJCAM0 headset recordings and SWC RIRs, the polynomial reverberation score used as the intra-phone smearing index showed high correlation with PER regarding the microphone channel difference. However it does not well correlate to the phonetic difference in the sensitivity to the same reverberant environment. Therefore a phonetic magnitude spectrum variance normalised by corresponding power spectrum is explored to improve the phonetic sensitivity of intra-phone smearing index. However the experiments suggests that the added component does not introduce any benefit.

To measure the distortion level by inter-phone smearing, the average power spectrum per phoneme is employed as a phonetic vector representation in a hyper space. The inter-phone reverberation distortion could change the phonetic vector in both its magnitude and angle. Since in speech recognition the feature magnitude is usually normalised, the rotation angle by reverberation distortion is adopted to represent the level of inter-phone smearing distortion. Experiments on WSJCAM0 data shows that such a inter-phone smearing index does not provide a high correlation with PER regarding the channel difference or the utterance difference. As a result combining it with the intra-phone smearing index based on the polynomial reverberation score does not introduce any benefit.

Further analysis is performed regarding one crucial implicit assumption made by the inter-phone smearing index, *i.e.* the reverberation distortion level of one phoneme is independent from the reverberation distortion level of other phonemes. The Fisher ratio is employed to explore the importance of global information in accurately estimating the distortion level of each individual classification class. The Fisher ratio was shown to have a high linear correlation with overall PER regarding the channel difference in reverberation. However it does not show a high correlation with PER regarding the phoneme difference in the sensitivity to reverberation distortion. This implies that other methods based on discriminative analysis on power spectrum will have similar problems in improving the correlation with PER regarding the signal dependent difference to the reverberation distortion.

Throughout the exploration research work in this chapter, it was noticed that there is some hidden problems regarding normalisation. The first problem is the RIR energy

normalisation. In all experiments the RIRs have been normalised by the maximum absolute value in the RIR coefficients beforehand. This might not be optimal while the energy based normalisation could introduce negative disruption to polynomial reverberation score when RIRs are not of the same length. Later when multiple components are combined for an overall reverberation score, a few weighting parameters are introduced to balance the different value ranges caused by completely different strategies adopted in calculating each component. In addition, when phonetic information is introduced to reverberation measurement, the normalisation over different phonetic energy is another issue.

The normalisation problem can be also observed from the experiment results. The polynomial format reverberation score provides a high rank correlation with PER and WER regarding the channel difference, but the linear correlation is much lower. This is particularly obvious when examining the early reverberation distortion where the RIRs are truncated to different lengths. The change in the reverberation scores based on the polynomial method is not proportional to the change in WER regarding different levels of reverberation. Another example is the performance evaluation of the rotation angle based inter-phone smearing index in Fig. 7.13 where each RIR is normalised by its maximum value in the spectrum. If the RIRs are not normalised, both the rank correlation and linear correlation will drop to below 0.2, and if the RIRs are normalised by overall energy, the linear correlation will increase to -0.66. In addition, the correlation value is negative, which opposes to the positive correlation expected by the analytic work in Section 7.3.1. This is potentially caused by the magnitude normalisation of the average phonetic power spectrum.

The work presented in this chapter achieved one primary goal out of two. The polynomial based reverberation score achieved the primary goal of measuring both early reverberation distortion and late reverberation distortion without the necessity of an optimal boundary for early reverberation and late reverberation. This chapter has explored multiple strategies and their combination to improve the capability of reverberation measurement in tracking the signal dependent sensitivity to the reverberation distortion. The partition of reverberation distortion into intra-phone smearing and inter-phone smearing is for the first time explored in estimating the reverberation distortion level. The experiment results suggest that future work is needed to improve the reverberation distortion estimation in short recordings which suffers the most from the signal variation.

The work in this chapter has avoided directly training a DNN with the utterance level WER change as reference to non-intrusively predict the recognition performance given different reverberation conditions. It is because this blind DNN strategy will inevitably lead to a reverberation measurement system highly dependent on the recognition systems involved. In addition such a strategy has gone beyond the category of reverberation

measurement because the DNN estimator will also react to any mismatch caused by speaker, gender, accent, background noise, *etc.* Instead the work in this chapter set its foundation on the analytic study about how reverberation distorts feature pattern in the front-end. Based on the analytic study a series of methods are hand-crafted to estimate the reverberation distortion level. Future work could consider a combination of machine learning based method and reverberation modelling based method. The intra-phone smearing and inter-phone smearing based reverberation measurement explored in this work highly relies on phonetic annotation of high quality. Future work should try to avoid such high dependence.

Chapter 8

Summary, Discussion and Future Work

Contents

8.1 Distant Speech Recognition of Real Natural Spontaneous Conversations 177

8.2 Reverberation Modelling and Measurement 179

8.3 Future Work 180

8.1 Distant Speech Recognition of Real Natural Spontaneous Conversations

The work presented in this thesis has aimed to improve the DSR performance on the real natural spontaneous multi-party conversations. For this purpose the Sheffield Wargame Corpora (SWC) was collected with simultaneous multi-microphone audio recording, multi-camera video recording and Ubisense based speaker location tracking. The SWC is manually annotated and transcribed, leading to 24.6 hours speech from 22 native English speakers, 14 being male and 8 being female. SWC is a unique database for three reasons. First, it is the first free native English speech corpus for research that is based on the real natural spontaneous multi-party conversations. Second, it is the first speech corpus with free and natural speaker movement accompanied with speaker location tracking. Third, it is a speech corpus including both the headset recordings and the multi-microphone distant recordings, released with a Kaldi recipe for the other researchers to replicate or to improve the work using the state-of-the-art ASR systems.

Recognition systems in two state-of-the-art structures involving DNNs are evaluated on SWC data. The DNN-HMM-GMM structure employs DNNs as the front-end to generate

bottleneck features for HMM-GMM training, and the parameters in the DNN-HMM-GMM are adapted from another system of the same structure but trained on a much larger corpus, *i.e.* the AMI corpus. The other DNN-HMM hybrid structure employs DNNs in acoustic modelling, and the parameters in DNN-HMM are trained in a standalone fashion using the recordings from SWC only. The training and evaluation of two structures are performed on the headset recordings, the single distant microphone recordings and the signal enhanced from multiple distant microphone recordings using beamforming. Overall the WERs are above 40% on the headset recordings and above 70% on the distant recordings, suggesting a high level of challenge for speech recognition on SWC data. The implementation of beamforming and dereverberation algorithms only brings small performance improvement. The lowest overall WER on distant recordings is 71.3%, and it is achieved with a combination of the multi-channel dereverberation algorithm GWPE and the MVDR beamforming employing TDOAs estimated from the speaker location tracked by the Ubisense system.

Using the SWC data as a study case, further analysis is conducted to understand the influence factors and the main challenges for the speech recognition of real natural spontaneous multi-party conversations. It is found that the real natural spontaneous multi-party conversational speech has a few unique properties that differentiate the SWC data from existing corpora. There are the very short utterances with an average duration of 2.2 seconds, the high proportion (around 50%) of utterances partly or completely overlapped with their competing speech utterances, the emotional speech with a big gender difference and speaker movement while talking. It is found that the short utterances and the emotional speech are two important reasons for the high WERs on both the headset recordings and the distant recordings. Further investigation employs the simulated data for a factor-by-factor analysis to quantify the impact of reverberation and overlapped speech on DSR performance. This is followed by a comparable investigation on real recordings regarding the interaction among multiple factors in application. The overlapped speech is found to be a key influence factor for the high WERs in DSR. Besides, reverberation and background noise are also two major factors that contribute to high WERs in DSR. Multi-channel based dereverberation and beamforming algorithms have been applied on both the simulated data and the real recordings, and it is found that the interaction among multiple factors can dramatically decrease the effectiveness of the enhancement algorithms. One typical example is that overlapped speech decreases the dereverberation performance significantly.

8.2 Reverberation Modelling and Measurement

An investigation has been performed both analytically and experimentally on how reverberation distorts the spectrogram for speech recognition features. It is found that the reverberation distortion in the speech complex spectrogram could be accurately approximated with a convolution between the clean speech complex spectrogram and the RIR complex spectrogram, *i.e.* the reverberation modelling. Based on the reverberation modelling, a polynomial reverberation score is proposed to estimate the reverberation distortion level on short speech utterances. The proposed polynomial reverberation score is found to provide a high rank correlation with the overall WER in terms of the microphone channel difference in reverberation. In addition, the polynomial reverberation score avoids a strict partition between early reverberation and late reverberation, as well as the disputed selection of the optimal boundary for the early-late-reverberation partition. When the rank correlation between the WER and the reverberation score is evaluated on a very small amount of data, e.g. one speech utterance, the polynomial reverberation score performs similar to the existing C_{50}, and both fail to well depict the different reverberation sensitivity due to the signal properties, particularly the phonetic properties.

Therefore further effort has been devoted to improving the polynomial reverberation score so that a higher rank correlation with WER could be achieved when both the reverberation score and WER are calculated on short recordings. Inspired by the phonetic analysis by Kokkinakis and Loizou (2011) on how reverberation causes the self-masking and the overlap-masking across phonemes, the reverberation distortion level is estimated from two aspects: the intra-phone smearing and the inter-phone smearing. Since the temporal pattern in the spectrogram for the same phoneme is relatively stable, the polynomial reverberation score is used for the intra-phone smearing index. In addition, the temporal variance in magnitude spectrogram normalised by energy per phoneme is used to emphasise the overall level of phonetic magnitude spectrogram change over time. For the inter-phone smearing index, the level of reverberation distortion is estimated with the change in the average power spectrum in each phoneme caused by the inter-phone smearing, particularly the rotation of the average power spectrum in the hyper space. The intra-phone smearing index and inter-phone smearing index are further combined additively. The experiments suggest that the phonetic difference in reverberation sensitivity is too challenging to measure with any of the proposed strategies in the intra-phone smearing index or inter-phone smearing index, alone or in combination.

To investigate the reasons for the poor performance of the phonetic analysis based reverberation measurement, one implicit assumption made in the inter-phone smearing level estimation is validated. The implicit assumption is that the smearing level of one

phoneme is independent from the smearing level of other phonemes. This assumption could be problematic because both the phoneme recognition and the word recognition are multi-class classification tasks where the discriminability of one class is dependent on the other classes. Therefore Fisher discriminative analysis is conducted to examine the change in the discriminability of the phonetic average power spectrum due to the reverberation distortion. The experiment results suggest that there is not a significant correlation between the change in the Fisher score per phoneme and the average change in the PER on corresponding phoneme in terms of phonetic difference, though there is indeed a high negative linear correlation between the overall Fisher score change and the overall PER change regarding the environment and microphone channel difference. Therefore the concern on the overall discriminability omitted by the inter-phone smearing seems not to be the reason for the poor performance of the reverberation measusrement based on the intra-phone smearing and inter-phone smearing.

Further discussion points out other potential reasons for the poor performance of the reverberation measurement in depicting the phonetic difference of reverberation sensitivity. The value normalisation has been highlighted as one likely issue, including the normalisation of different phonetic energies in different frequencies and the normalisation of different RIRs.

8.3 Future Work

The analysis on DSR using the SWC data for a case study covered a variety of influence factors in the real distant recordings except for the background noise. The background noise in SWC data is very diverse, including both the stationary background noise such as the computer fan noise and the changing background noise such as the wood floor cracking sound when the players walk around in the room. The background noise does play an important role in the high WERs in the DSR of SWC data, and it represents the large range of background noise in the real domestic applications. Therefore for a complete understanding of all the environment factors in DSR, future analysis should be conducted to quantify the impact of the real background noise on DSR performance. The effectiveness of the de-noising algorithms could be conducted on both the simulated data and the real distant recordings in a similar way with the analysis work conducted in this thesis regarding the effectiveness of dereverberation algorithms.

In the work on reverberation measurement, the polynomial reverberation score is found to provide a high rank correlation with the WER and the PER, while further work is still needed to address the value normalisation issue so that a high linear correlation

with the speech recognition performance could be achieved as well. In addition, more research is needed to improve the reverberation distortion level estimation accuracy on short recordings where the signal dependent reverberation sensitivity needs to be taken into consideration. The progress in the reverberation measurement on short recordings will be critical for both the data selection and the model selection used by the state-of-the-art ASRs system to achieve a balance between the overall robustness against diverse levels of reverberation and a good performance in each reverberant condition.

References

Abdel-Hamid, O. and Jiang, H. (2013), Fast speaker adaptation of hybrid NN/HMM model for speech recognition based on discriminative learning of speaker code, *in* 'IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', pp. 7942–7946.

Abdel-Hamid, O., Mohamed, A. R., Jiang, H. and Penn, G. (2012), Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition, *in* 'IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', pp. 4277–4280.

Anguera, X., Wooters, C. and Hernando, J. (2007), 'Acoustic beamforming for speaker diarization of meetings', *IEEE Transactions on Audio, Speech, and Language Processing* **15**(7), 2011–2022.

Assmann, P. and Summerfield, Q. (2004), *The Perception of Speech Under Adverse Conditions*, Springer New York, New York, NY, pp. 231–308.

Barker, J., Marxer, R., Vincent, E. and Watanabe, S. (2015), The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines, *in* 'IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)', Scottsdale, AZ, United States.

Barker, J., Marxer, R., Vincent, E. and Watanabe, S. (2016), 'The third 'CHIME' speech separation and recognition challenge: Analysis and outcomes', *Computer Speech and Language* .

Bell, P., Swietojanski, P. and Renals, S. (2013), Multi-level adaptive networks in tandem and hybrid ASR systems, *in* 'IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', pp. 6975–6979.

Bengio, Y., De Mori, R., Flammia, G. and Kompe, R. (1992), 'Global optimization of a neural network-hidden Markov model hybrid', *Trans. Neural Network* **3**(2), 252–259.

Berouti, M., Schwartz, R. and Makhoul, J. (1979), Enhancement of speech corrupted by acoustic noise, *in* 'Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '79.', Vol. 4, pp. 208–211.

Bishop, C. M. (1995), *Neural Networks for Pattern Recognition*, Oxford University Press, Inc., New York, NY, USA.

Boll, S. (1979), 'Suppression of acoustic noise in speech using spectral subtraction', *IEEE Transactions on Acoustics, Speech and Signal Processing* **27**(2), 113–120.

Bradley, J. S. (2011), 'Review of objective room acoustics measures and future needs', *Applied Acoustics* **72**(10), 713–720.

Bradley, J. S., Sato, H. and Picard, M. (2003), 'On the importance of early reflections for speech in rooms', *The Journal of the Acoustical Society of America* **113**(6), 3233–3244.

Brutti, A. and Matassoni, M. (2014), On the use of early-to-late reverberation ratio for ASR in reverberant environments, *in* 'IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', pp. 4638–4642.

Brutti, A. and Matassoni, M. (2016), 'On the relationship between early-to-late ratio of room impulse responses and ASR performance in reverberant environments', *Speech Communication* **76**, 170–185.

Checka, N., Wilson, K. W., Siracusa, M. R. and Darrell, T. (2004), Multiple person and speaker activity tracking with a particle filter, *in* 'IEEE International Conference on Acoustics, Speech, and Signal Processing', Vol. 5, p. 881.

Chen, J., Benesty, J., Huang, Y. and Doclo, S. (2006), 'New insights into the noise reduction wiener filter', *IEEE Transactions on Audio, Speech, and Language Processing* **14**(4), 1218–1234.

Cooke, M., Morris, A. and Green, P. (1997), Missing data techniques for robust speech recognition, *in* 'IEEE International Conference on Acoustics, Speech, and Signal Processing', Vol. 2, pp. 863–866.

Cui, X., Goel, V. and Kingsbury, B. (2015), 'Data augmentation for deep neural network acoustic modeling', *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* **23**(9), 1469–1477.

Dahl, G., Ranzato, M., rahman Mohamed, A. and Hinton, G. E. (2010), Phone recognition with the mean-covariance restricted Boltzmann machine, *in* J. Lafferty, C. Williams, J. Shawe-taylor, R. Zemel and A. Culotta, eds, 'Advances in Neural Information Processing Systems 23', pp. 469–477.

Delcroix, M., Yoshioka, T., Ogawa, A., Kubo, Y., Fujimoto, M., Ito, N., Kinoshita, K., Espi, M., Hori, T., Nakatani, T. et al. (2014), Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the REVERB challenge, *in* 'Proc. REVERB Workshop'.

Deng, L., Acero, A., Jiang, L., Droppo, J. and Huang, X. (2001), High-performance robust speech recognition using stereo training data, *in* 'IEEE International Conference on Acoustics, Speech, and Signal Processing', Vol. 1, pp. 301–304.

Deng, L., Droppo, J. and Acero, A. (2004), 'Enhancement of log mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise', *IEEE Transactions on Speech and Audio Processing* **12**(2), 133–143.

Deng, L., Droppo, J. and Acero, A. (2005), 'Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion', *IEEE Transactions on Speech and Audio Processing* **13**(3), 412–421.

Deng, L., Droppo, J. and Acero, A. (May 2002), A bayesian approach to speech feature enhancement using the dynamic cepstral prior, *in* 'IEEE International Conference on Acoustics, Speech, and Signal Processing', Vol. 1, pp. 829–832.

Doire, C., Brookes, M., Naylor, P. A. and Jensen, S. (2015), 'Data-driven statistical modelling of room impulse responses in the power domain', *Proceedings of the European Signal Processing Conference (EUSIPCO)* pp. 2466–2470.

Eaton, J., Gaubitch, N. D., Moore, A. H. and Naylor, P. A. (2015), The ACE challenge - corpus description and performance evaluation, *in* '2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)', pp. 1–5.

Eneman, K. and Moonen, M. (2007), 'Multimicrophone speech dereverberation: Experimental validation', *EURASIP Journal on Audio, Speech, and Music Processing* **2007**(1), 051831.

Fiscus, J. G. (1997), A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER), *in* 'IEEE Workshop on Automatic Speech Recognition and Understanding', pp. 347–354.

Fox, C., Christensen, H. and Hain, T. (2012), Studio report: Linux audio for multi-speaker natural speech technology., *in* 'Proc. Linux Audio Conference'.

Fox, C., Liu, Y., Zwyssig, E. and Hain, T. (2013), The sheffield wargames corpus., *in* 'The 14th Annual Conference of the International Speech Communication Association (Interspeech 2013)', Lyon, France.

Furui, S. (1981), 'Cepstral analysis technique for automatic speaker verification', *IEEE Transactions on Acoustics, Speech and Signal Processing* **29**(2), 254–272.

Furuya, K., Sakauchi, S. and Kataoka, A. (2006), Speech dereverberation by combining MINT-based blind deconvolution and modified spectral subtraction, *in* 'IEEE International Conference on Acoustics Speech and Signal Processing Proceedings', Vol. 1.

Gales, M. and Young, S. (2008), 'The application of hidden Markov models in speech recognition', *Foundations and Trends in Signal Processing* **1**(3), 195–304.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S. and Dahlgren, N. L. (1993), 'DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM', *NASA STI/Recon Technical Report N* **93**.

Gatica-Perez, D., Lathoud, G., Odobez, J. M. and McCowan, I. (2007), 'Audiovisual probabilistic tracking of multiple speakers in meetings', *IEEE Transactions on Audio, Speech, and Language Processing* **15**(2), 601–616.

Gaubitch, N. D., Habets, E. A. P. and Naylor, P. A. (2008), Multimicrophone speech dereverberation using spatiotemporal and spectral processing, *in* 'IEEE International Symposium on Circuits and Systems', pp. 3222–3225.

Giri, R., Seltzer, M. L., Droppo, J. and Yu, D. (2015), Improving speech recognition in reverberation using a room-aware deep neural network and multi-task learning, IEEE - Institute of Electrical and Electronics Engineers.

Graves, A., rahman Mohamed, A. and Hinton, G. E. (2013), 'Speech recognition with deep recurrent neural networks', *CoRR* .

Grezl, F., Karafiat, M., Kontar, S. and Cernocky, J. (2007), Probabilistic and bottle-neck features for LVCSR of meetings, *in* 'IEEE International Conference on Acoustics, Speech and Signal Processing', Vol. 4, pp. 757–760.

Guidorzi, P., Barbaresi, L., D'Orazio, D. and Garai, M. (2015), 'Impulse responses measured with MLS or swept-sine signals applied to architectural acoustics: An in-depth analysis of the two methods and some case studies of measurements inside theaters', *Energy Procedia* **78**, 1611–1616. The 6th International Building Physics Conference (IBPC).

Haas, H. (1972), 'The influence of a single echo on the audibility of speech', *J. Audio Eng. Soc* **20**(2), 146–159.

Habets, E. A. P. (2005), Multi-channel speech dereverberation based on a statistical model of late reverberation, *in* 'IEEE International Conference on Acoustics, Speech, and Signal Processing', Vol. 4, pp. 173–176.

Habets, E. A. P., Gannot, S. and Cohen, I. (2006), Dual-microphone speech dereverberation in a noisy environment, *in* 'IEEE International Symposium on Signal Processing and Information Technology', pp. 651–655.

Hain, T., Wan, V., Burget, L., Karafiat, M., Dines, J., Vepa, J., Garau, G. and Lincoln, M. (2007), The AMI system for the transcription of speech in meetings, *in* 'IEEE International Conference on Acoustics, Speech and Signal Processing', Vol. 4, IEEE, p. 357.

Harvie-Clark, J. and Dobinson, N. (2013), The practical application of G and C50 in classrooms, *in* 'INTER-NOISE and NOISE-CON Congress and Conference Proceedings', Vol. 247, Institute of Noise Control Engineering, pp. 1510–1520.

Hennebert, J., Ris, C., Bourlard, H., Renals, S. and Morgan, N. (1997), Estimation of global posteriors and forward-backward training of hybrid HMM/ANN systems, *in* 'European Conference on Speech Communication and Technology (EUROSPEECH 97), Rhodes, Greece', pp. 1951–1954. Some of the files below are copyrighted. They are provided for your convenience, yet you may download them only if you are entitled to do so by your arrangements with the various publishers.

Hermansky, H., Ellis, D. P. W. and Sharma, S. (2000), Tandem connectionist feature extraction for conventional HMM systems, *in* 'IEEE International Conference on Acoustics, Speech, and Signal Processing', Vol. 3, pp. 1635–1638.

Hikichi, T., Delcroix, M. and Miyoshi, M. (2007), 'Inverse filtering for speech dereverberation less sensitive to noise and room transfer function fluctuations', *EURASIP Journal on Advances in Signal Processing* **2007**(1), 1–12.

Hochreiter, S. and Schmidhuber, J. (1997), 'Long short-term memory', *Neural Comput.* **9**(8), 1735–1780.

Hornik, K., Stinchcombe, M. and White, H. (1989), 'Multilayer feedforward networks are universal approximators', *Neural Networks* **2**(5), 359–366.

Hu, Y. and Kokkinakis, K. (2014), 'Effects of early and late reflections on intelligibility of reverberated speech by cochlear implant listeners', *The Journal of the Acoustical Society of America* **135**(1), 22–28.

Jacobsen, F. (1979), *The Diffuse Sound Field: Statistical Considerations Concerning the Reverberant Field in the Steady State*, Report: Laboratoriet for Akustik, Acoustics Laboratory, Technical University of Denmark.

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A. and Wooters, C. (2003), The ICSI meeting corpus, pp. 364–367.

Jeub, M., Schafer, M. and Vary, P. (2009), A binaural room impulse response database for the evaluation of dereverberation algorithms, *in* 'Digital Signal Processing, 2009 16th International Conference on', pp. 1–5.

Kingsbury, B. (2009), Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling, *in* 'IEEE International Conference on Acoustics, Speech and Signal Processing', pp. 3761–3764.

Kinoshita, K., Delcroix, M., Gannot, S., P. Habets, E. A., Haeb-Umbach, R., Kellermann, W., Leutnant, V., Maas, R., Nakatani, T., Raj, B., Sehr, A. and Yoshioka, T. (2016), 'A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research', *EURASIP Journal on Advances in Signal Processing* **2016**(1), 7.

Ko, T., Peddinti, V., Povey, D. and Khudanpur, S. (2015), Audio augmentation for speech recognition, *in* 'The 16th Annual Conference of the International Speech Communication Association (Interspeech)'.

Kodrasi, I. and Doclo, S. (2012), Robust partial multichannel equalization techniques for speech dereverberation, *in* 'IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', pp. 537–540.

Kodrasi, I. and Doclo, S. (2014), Joint dereverberation and noise reduction based on acoustic multichannel equalization, *in* 'The 14th International Workshop on Acoustic Signal Enhancement (IWAENC)', pp. 139–143.

Kokkinakis, K. and Loizou, P. C. (2011), 'The impact of reverberant self-masking and overlap-masking effects on speech intelligibility by cochlear implant listeners (l)', *The Journal of the Acoustical Society of America* **130**(3), 1099–1102.

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012), ImageNet classification with deep convolutional neural networks, *in* F. Pereira, C. Burges, L. Bottou and K. Weinberger, eds, 'Advances in Neural Information Processing Systems', Curran Associates, Inc., pp. 1097–1105.

Kuttruff, H. (2000), *Room Acoustics, Fourth Edition*, E-Libro, Taylor & Francis.

Kuttruff, H. (2009), *Room Acoustics, Fifth Edition*, 5 edn, CRC Press.

Lee, H., Pham, P., Largman, Y. and Ng, A. Y. (2009), Unsupervised feature learning for audio classification using convolutional deep belief networks, *in* Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams and A. Culotta, eds, 'Advances in Neural Information Processing Systems 22', Curran Associates, Inc., pp. 1096–1104.

Liao, H. and Gales, M. J. F. (2008), 'Issues with uncertainty decoding for noise robust automatic speech recognition', *Speech Communication* **50**(4), 265–277.

Lim, J. S. and Oppenheim, A. V. (1979), 'Enhancement and bandwidth compression of noisy speech', *Proceedings of the IEEE* **67**(12), 1586–1604.

Lincoln, M., McCowan, I., Vepa, J. and Maganti, H. K. (2005), The multi-channel wall street journal audio visual corpus (MC-WSJ-AV): specification and initial experiments, *in* 'IEEE Workshop on Automatic Speech Recognition and Understanding', pp. 357–362.

Liu, J. and Yang, G.-Z. (2015), 'Robust speech recognition in reverberant environments by using an optimal synthetic room impulse response model', *Speech Communication* **67**, 65–77.

Liu, Y., Fox, C., Hasan, M. and Hain, T. (2016), The sheffield wargame corpus - day two and day three, *in* 'The 17th Annual Conference of the International Speech Communication Association (Interspeech)', San Francisco, USA.

Liu, Y., Karanasou, P. and Hain, T. (2015), An investigation into speaker informed DNN front-end for LVCSR, *in* 'IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', pp. 4300–4304.

Liu, Y., Zhang, P. and Hain, T. (2014), Using neural network front-ends on far field multiple microphones based speech recognition, *in* 'IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', pp. 5542–5546.

Lu, L., Ghoshal, A. and Renals, S. (2013), Acoustic data-driven pronunciation lexicon for large vocabulary speech recognition, *in* 'Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)'.

Ma, W.-K., Vo, B.-N., Singh, S. S. and Baddeley, A. (2006), 'Tracking an unknown time-varying number of speakers using tdoa measurements: a random finite set approach', *IEEE Transactions on Signal Processing* **54**(9), 3291–3304.

Maganti, H. K., Gatica-Perez, D. and McCowan, I. (2007), 'Speech enhancement and recognition in meetings with an audio-visual sensor array', *IEEE Transactions on Audio, Speech, and Language Processing* **15**(8), 2257–2269.

Marković, I. and Petrović, I. (2010), 'Speaker localization and tracking with a microphone array on a mobile robot using von mises distribution and particle filtering', *Robotics and Autonomous Systems* **58**(11), 1185–1196.

Matassoni, M., Astudillo, R. F., Katsamanis, A. and Ravanelli, M. (2014), The DIRHA-GRID corpus: baseline and tools for multi-room distant speech recognition using distributed microphones, *in* 'The 15th Annual Conference of the International Speech Communication Association (Interspeech)', pp. 1613–1617.

McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V. et al. (2005), The AMI meeting corpus, *in* 'The 5th International Conference on Methods and Techniques in Behavioral Research', Vol. 88.

Miyoshi, M. and Kaneda, Y. (1988), 'Inverse filtering of room acoustics', *IEEE Transactions on Acoustics, Speech, and Signal Processing* **36**(2), 145–152.

Moreno, P. J., Raj, B. and Stern, R. M. (1996), A vector taylor series approach for environment-independent speech recognition, *in* 'IEEE International Conference on Acoustics, Speech, and Signal Processing', Vol. 2, pp. 733–736.

Nagata, Y., Tatekura, Y., Saruwatari, H. and Shikano, K. (2004), 'Iterative inverse filter relaxation algorithm for adaptation to acoustic fluctuation in sound reproduction system', *Electronics and Communications in Japan (Part III: Fundamental Electronic Science)* **87**(7), 15–26.

Nakadai, K., Okuno, H. G., Kitano, H., Okuno, H. G. and Kitano, H. (2002), Real-time sound source localization and separation for robot audition, *in* 'IEEE International Conference on Spoken Language Processing', pp. 193–196.

Naylor, P. A. and Gaubitch, N. D. (2005), Speech dereverberation.

Otsuka, T., Ishiguro, K., Yoshioka, T., Sawada, H. and Okuno, H. G. (2014), 'Multichannel sound source dereverberation and separation for arbitrary number of sources based on bayesian nonparametrics', *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **22**(12), 2218–2232.

Parada, P. P., Sharma, D., Lainez, J., Barreda, D., Naylor, P. A. and Waterschoot, T. v. (2014), A quantitative comparison of blind C50 estimators, *in* 'The 14th International Workshop on Acoustic Signal Enhancement (IWAENC)', pp. 298–302.

Parada, P. P., Sharma, D., Lainez, J., Barreda, D., van Waterschoot, T. and Naylor, P. (2016), 'A single-channel non-intrusive C50 estimator correlated with speech recognition performance', *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (99).

Parada, P. P., Sharma, D. and Naylor, P. A. (2014), Non-intrusive estimation of the level of reverberation in speech, *in* 'IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', pp. 4718–4722.

Parada, P. P., Sharma, D., Naylor, P. A. and Waterschoot, T. v. (2014), Reverberant speech recognition: A phoneme analysis, *in* 'IEEE Global Conference on Signal and Information Processing (GlobalSIP)', pp. 567–571.

Parada, P. P., Sharma, D., Naylor, P. A. and Waterschoot, T. v. (2015), 'Reverberant speech recognition exploiting clarity index estimation', *EURASIP Journal on Advances in Signal Processing* **2015**(1), 1–12.

Parihar, N., Picone, J., Pearce, D. and Hirsch, H. G. (2004), Performance analysis of the aurora large vocabulary baseline system, *in* '12th European Signal Processing Conference', pp. 553–556.

Pavlidi, D., Griffin, A., Puigt, M. and Mouchtaris, A. (2013), 'Real-time multiple sound source localization and counting using a circular microphone array', *IEEE Transactions on Audio, Speech, and Language Processing* **21**(10), 2193–2206.

Pearson, K. (1895), 'Note on regression and inheritance in the case of two parents', *Proceedings of the Royal Society of London* **58**, 240–242.

Polack, J.-D. (1988), La transmission de l'é nergie dans les salles, PhD thesis, Université du Maine, Le Mans, France.

Potamitis, I., Chen, H. and Tremoulis, G. (2004), 'Tracking of multiple moving speakers with multiple microphone arrays', *IEEE Transactions on Speech and Audio Processing* **12**(5), 520–529.

Povey, D. and Woodland, P. C. (2002), Minimum phone error and I-smoothing for improved discriminative training, *in* 'IEEE International Conference on Acoustics, Speech, and Signal Processing'.

Qian, Y., Bi, M., Tan, T. and Yu, K. (2016), 'Very deep convolutional neural networks for noise robust speech recognition', *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **24**(12), 2263–2276.

Rath, S. P., Knill, K., Ragni, A. and Gales, M. J. F. (2014), Combining tandem and hybrid systems for improved speech recognition and keyword spotting on low resource languages, *in* 'The 15th Annual Conference of the International Speech Communication Association (Interspeech)'.

Raut, C. K., Nishimoto, T. and Sagayama, S. (2006), Model adaptation for long convolutional distortion by maximum likelihood based state filtering approach, *in* 'IEEE International Conference on Acoustics Speech and Signal Processing Proceedings', Vol. 1.

Ravanelli, M., Cristoforetti, L., Gretter, R., Pellin, M., Sosi, A. and Omologo, M. (2015), The DIRHA-ENGLISH corpus and related tasks for distant-speech recognition in domestic environments, *in* 'IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)', pp. 275–282.

Renals, S. and Hain, T. (2010), *Speech Recognition*, Wiley-Blackwell, pp. 297–332.

Robinson, T., Fransen, J., Pye, D., Foote, J. and Renals, S. (1995), WSJCAM0: A british english speech corpus for large vocabulary continuous speech recognition, *in* 'In Proc. ICASSP 95', IEEE, pp. 81–84.

Sagayama, S., Yamaguchi, Y., Takahashi, S. and Takahashi, J. (1997), Jacobian approach to fast acoustic model adaptation, *in* 'IEEE International Conference on Acoustics, Speech, and Signal Processing', Vol. 2, pp. 835–838.

Sainath, T. N., Weiss, R. J., Wilson, K. W., Narayanan, A., Bacchiani, M. and Senior, A. (2015), Speaker location and microphone spacing invariant acoustic modeling from raw multichannel waveforms, *in* 'IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)', IEEE.

Sak, H., Senior, A. W. and Beaufays, F. (2014), 'Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition', *CoRR* .

Saon, G., Sercu, T., Rennie, S. J. and Kuo, H.-K. J. (2016), 'The IBM 2016 english conversational telephone speech recognition system', *CoRR* .

Saon, G., Soltau, H., Nahamoo, D. and Picheny, M. (2013), Speaker adaptation of neural network acoustic models using i-vectors, *in* 'IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)', pp. 55–59.

Sehr, A., Habets, E. A. P., Maas, R. and Kellermann, W. (2010), Towards a better understanding of the effect of reverberation on speech recognition performance, *in* 'Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC)', Tel Aviv, Israel.

Sehr, A. and Kellermann, W. (2008), New results for feature-domain reverberation modeling, *in* 'Hands-Free Speech Communication and Microphone Arrays (HSCMA)', pp. 168–171.

Sehr, A. and Kellermann, W. (2009), Strategies for modeling reverberant speech in the feature domain, *in* 'IEEE International Conference on Acoustics, Speech and Signal Processing', pp. 3725–3728.

Sehr, A., Zeller, M. and Kellermann, W. (2006), Distant-talking continuous speech recognition based on a novel reverberation model in the feature domain, *in* 'The 7th Annual Conference of the International Speech Communication Association (Interspeech 2006)', pp. 769–772.

Seide, F., Li, G., Chen, X. and Yu, D. (2011), Feature engineering in context-dependent deep neural networks for conversational speech transcription, *in* 'IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)', IEEE.

Seltzer, M., Yu, D. and Wang, Y. (2013), An investigation of deep neural networks for noise robust speech recognition, *in* 'IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

Sercu, T. and Goel, V. (2016), 'Advances in very deep convolutional neural networks for LVCSR', *CoRR* .

Shankland, R. S. (1977), 'Architectural acoustics in america to 1930', *The Journal of the Acoustical Society of America* **61**(2).

Sharma, D., Hilkhuysen, G., Gaubitch, N. D., Naylor, P. A., Brookes, M. and Huckvale, M. (2010), Data driven method for non-intrusive speech intelligibility estimation, *in* 'The 18th European Signal Processing Conference', pp. 1899–1903.

Stan, G.-B., Embrechts, J.-J. and Archambeau, D. (2002), 'Comparison of different impulse response measurement techniques', *Journal of the Audio Engineering Society* **50**(4), 249–262.

Stewart, R. and Sandler, M. (2010), Database of omnidirectional and B-format room impulse responses, *in* 'IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)', pp. 165–168.

Strobel, N., Spors, S. and Rabenstein, R. (2001), 'Joint audio-video object localization and tracking', *IEEE Signal Processing Magazine* **18**(1), 22–31.

Sturim, D. E., Brandstein, M. S. and Silverman, H. F. (1997), Tracking multiple talkers using microphone-array measurements, *in* 'IEEE International Conference on Acoustics, Speech, and Signal Processing', Vol. 1, pp. 371–374.

Swietojanski, P., Ghoshal, A. and Renals, S. (2013), Hybrid acoustic models for distant and multichannel large vocabulary speech recognition, *in* 'IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)', pp. 285–290.

Swietojanski, P., Ghoshal, A. and Renals, S. (2014), 'Convolutional neural networks for distant speech recognition', *IEEE Signal Processing Letters* **21**(9), 1120–1124.

Thiele, R. (1953), 'Richtungsverteilung und zeitfolge der schallrückwürfe in räumen', *Acta Acustica United with Acustica* **3**(Supplement 2), 291–302.

Tüske, Z., Golik, P., Nolden, D., Schlüter, R. and Ney, H. (2014), Data augmentation, feature combination, and multilingual neural networks to improve ASR and KWS performance for low-resource languages, *in* 'The 15th Annual Conference of the International Speech Communication Association (Interspeech)'.

Valimaki, V., Parker, J. D., Savioja, L., Smith, J. O. and Abel, J. S. (2012), 'Fifty years of artificial reverberation', *IEEE Transactions on Audio, Speech, and Language Processing* **20**(5), 1421–1448.

Valin, J. M., Michaud, F., Hadjou, B. and Rouat, J. (2004), Localization of simultaneous moving sound sources for mobile robot using a frequency- domain steered beamformer approach, *in* 'IEEE International Conference on Robotics and Automation', Vol. 1, pp. 1033–1038.

Valin, J. M., Michaud, F. and Rouat, J. (2006), Robust 3D localization and tracking of sound sources using beamforming and particle filtering, *in* 'IEEE International Conference on Acoustics Speech and Signal Processing Proceedings', Vol. 4.

Vermaak, J. and Blake, A. (2001), Nonlinear filtering for speaker tracking in noisy and reverberant environments, *in* 'IEEE International Conference on Acoustics, Speech, and Signal Processing', Vol. 5, pp. 3021–3024.

Veselý, K., Ghoshal, A., Lukáš, B. and Povey, D. (2013), Sequence-discriminative training of deep neural networks, *in* 'The 14th Annual Conference of the International Speech Communication Association (Interspeech)', number 8, International Speech Communication Association, pp. 2345–2349.

Viikki, O. and Laurila, K. (1998), 'Cepstral domain segmental feature vector normalization for noise robust speech recognition', *Speech Communication* **25**(1-3), 133–147.

Wen, J., Gaubitch, N. D., Habets, E., Myatt, T. and Naylor, P. A. (2006), Evaluation of speech dereverberation algorithms using the MARDY database, *in* 'Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)', Paris, France.

Wölfel, M. and McDonough, J. (2009), *Distant Speech Recognition*, Wiley.

Xiao, X., Watanabe, S., Erdogan, H., Lu, L., Hershey, J., Seltzer, M. L., Chen, G., Zhang, Y., Mandel, M. and Yu, D. (2016), Deep beamforming networks for multi-channel speech recognition, *in* 'IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)'.

Xiong, F., Goetze, S. and Meyer, B. T. (2014), Estimating room acoustic parameters for speech recognizer adaptation and combination in reverberant environments, *in* 'IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', pp. 5522–5526.

Yoshioka, T., Karita, S. and Nakatani, T. (2015), Far-field speech recognition using CNN-DNN-HMM with convolution in time, *in* 'IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)'.

Yoshioka, T. and Nakatani, T. (2012), 'Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening', *IEEE Transactions on Audio, Speech, and Language Processing* **20**(10), 2707–2720.

Yoshioka, T., Sehr, A., Delcroix, M., Kinoshita, K., Maas, R., Nakatani, T. and Kellermann, W. (2012), 'Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition', *IEEE Signal Processing Magazine* **29**(6), 114–126.

Young, S. (1996), 'A review of large-vocabulary continuous-speech', *IEEE Signal Processing Magazine* **13**(5), 45.

Zhang, P., Liu, Y. and Hain, T. (2014), Semi-supervised DNN training in meeting recognition, *in* 'IEEE Spoken Language Technology Workshop (SLT)', South Lake Tahoe, USA.

Zhang, W., Habets, E. and Naylor, P. A. (2010), On the use of channel shortening in multichannel acoustic system equalization.

Zhang, W., Khong, A. W. H. and Naylor, P. A. (2009), Acoustic system equalization using channel shortening techniques for speech dereverberation, *in* 'The 17th European Signal Processing Conference', pp. 1427–1431.

Zuo, G., Liu, W. and Ruan, X. (2003), Telephone speech recognition using simulated data from clean database, *in* 'IEEE International Conference on Robotics, Intelligent Systems and Signal Processing', Vol. 1, pp. 49–53.