Context-aware speech synthesis: A human-inspired model for monitoring and adapting synthetic speech

Mauro Nicolao

Department of Computer Science
University of Sheffield

January 2019

Dissertation submitted to the University of Sheffield
for the degree of Doctor of Philosophy in Computer Science

Supervisor: Prof. Roger K. Moore

Acknowledgements

The research leading to these results was funded by the EU-FP7 network SCALE (ITN-213850).

“La parole est moitié à celui qui parle, moitié à celui qui l’écoute” (The spoken word belongs half to him who speaks, and half to him who listens).

Michel De Montaigne, Essais, III, 13, De l’expérience.

Table of Contents

Abstract

The aim of this PhD thesis is to illustrate the development a computational model for speech synthesis, which mimics the behaviour of human speaker when they adapt their production to their communicative conditions.

The PhD project was motivated by the observed differences between state-of-the-art synthesiser's speech and human production. In particular, synthesiser outcome does not exhibit any adaptation to communicative context such as environmental disturbances, listener's needs, or speech content meanings, as the human speech does. No evaluation is performed by standard synthesisers to check whether their production is suitable for the communication requirements.

Inspired by Lindblom's Hyper and Hypo articulation theory (H&H) theory of speech production, the computational model of Hyper and Hypo articulation theory (C2H) is proposed. This novel computational model for automatic speech production is designed to *monitor its outcome* and to be able to *control the effort* involved in the synthetic speech generation.

Speech transformations are based on the hypothesis that low-effort attractors for a human speech production system can be identified. Such acoustic configurations are close to minimum possible effort that a speaker can make in speech production. The interpolation/extrapolation along the key dimension of hypo/hyper-articulation can be motivated by energetic considerations of phonetic contrast. The complete reactive speech synthesis is enabled by adding a negative perception feedback loop to the speech production chain in order to constantly assess the communicative effectiveness of the proposed adaptation. The distance to the original communicative intents is the control signal that drives the speech transformations.

A hidden Markov model (HMM)-based speech synthesiser along with the continuous adaptation of its statistical models is used to implement the C2H model.

A standard version of the synthesis software does not allow for transformations of speech during the parameter generation. Therefore, the generation algorithm of one the most well-known speech synthesis frameworks, HMM/DNN-based speech synthesis framework (HTS), is modified. The short-time implementation of speech intelligibility index (SII), named extended speech intelligibility index (eSII), is also chosen as the main perception measure in the feedback loop to control the transformation.

The effectiveness of the proposed model is tested by performing acoustic analysis, objective, and subjective evaluations. A key assessment is to measure the control of the speech clarity in noisy condition, and the similarities between the emerging modifications and human behaviour. Two objective scoring methods are used to assess the speech intelligibility of the implemented system: the speech intelligibility index (SII) and the index based upon the Dau measure (Dau).

Results indicate that the intelligibility of C2H-generated speech can be continuously controlled. The effectiveness of reactive speech synthesis and of the phonetic contrast motivated transforms is confirmed by the acoustic and objective results. More precisely, in the maximum-strength hyper-articulation transformations, the improvement with respect to non-adapted speech is above 10% for all intelligibility indices and tested noise conditions.

Glossary

AI articulation index

ASR automatic speech recognition

C2H computational model of Hyper and Hypo articulation theory

CMLLR constrained maximum likelihood linear regression

CoG Centre of Gravity

Dau Dau measure

DNN deep neural network

E-CPC English Consonant Production Control

E-VPC English Vowel Production Control

EM expectation maximisation

EMA Electromagnetic articulography

eSII extended speech intelligibility index

F0 fundamental frequency

FFT fast Fourier transform

GMM Gaussian mixture model

GMZ Gaussian mixture zone

GP Glimpse Proportion measure

H&H Hyper and Hypo articulation theory

HCI human-computer interface

HMM hidden Markov model

HTK HMM Toolkit

HTS HMM/DNN-based speech synthesis framework

HTS-C2H HTS-based C2H

HYO hypo-articutated

HYP hyper-articulated

I-CPC Italian Consonant Production Control

I-VPC Italian Vowel Production Control

LPC linear predictive coefficients

LSP line spectral pairs

LTAS long term average spectrum

MAP maximum a-posteriori

MFCC Mel-frequency cepstral coefficients

MGC Mel-generalized cepstral

MGLSA Mel-generalised log spectral approximation

ML maximum likelihood

MLLR maximum likelihood linear regression

NLP natural language processing

OOD out of domain

PCT Perceptual Control Theory

PESQ Perceptual Evaluation of Speech Quality

POS part-of-speech

SAMPA Speech Assessment Methods Phonetic Alphabet

SII speech intelligibility index

SNR signal-to-noise ratio

SPSS statistical parametric speech synthesis

SSNR segmental signal-to-noise ratio

STI speech transmission index

STRAIGHT Speech Transformation and Representation using Adaptive Interpolation of Weight Spectrum

SZ safe zone

TGSM trajectory generation simulation model

TTS text to speech synthesiser

VTLN vocal tract linear normalisation

WER word error rate

List of Figures

Preface

The content of this thesis has been previously presented in the following conferences, workshops, and journals:

2011 at the International Congress of Phonetic Sciences (ICPhS) 2011 the first reactive speech synthesis model was presented (Moore and Nicolao, 2011)

- Roger K Moore and **Mauro Nicolao**. *Reactive Speech Synthesis: Actively Managing Phonetic Contrast Along an H&H Continuum.* In ICPhS 2011, pages 1422–1425, Hong Kong, China, August 2011.

2012 at INTERSPEECH 2012, the C2H model adaptation was first presented and assessed (Nicolao and Moore, 2012a). At the SAPA-SCALE 2012 workshop, the TGSM simulation model was also proposed to simulate the navigation of the acoustic space in a reduced 2D space (Nicolao et al., 2012). At the workshop eNTERFACE 2012, the Matlab software suite XPLIC8 (Tang et al., 2012) was implemented, and it was used to provide an analysis of a real speech corpus (the P8-Harvard corpus) (Stylianou et al., 2012).

- **Mauro Nicolao** and Roger K Moore. *Establishing some principles of human speech production through two-dimensional computational models.* In SAPA-SCALE work- shop 2012, Portland, OR, August 2012.
- **Mauro Nicolao**, Javier Latorre, and Roger K Moore. *C2H: A Computational Model of H&H-based Phonetic Contrast in Synthetic Speech.* In INTERSPEECH 2012, Portland, OR, September 2012.
- Yannis Stylianou, Valerie Hazan, Vincent Aubanel, Elizabeth Godoy, Sonia Granlund, Mark Huckvale, Emma Jokinen, Maria Koutsogiannaki, Pejman Mowlaee, **Mauro Nicolao**, Tuomo Raitio, Anna Sfakianaki, and Yan Tang. *P8 - Active Speech Modifications. Technical report*, Metz, France, November 2012.

2013 at SPIN 2013, I gave an invited talk on the C2H speech synthesis model (Nicolao and Moore, 2013a). At SSW7, C2H was applied to Italian along with new acoustic analysis comparison with human production (Nicolao et al., 2013).

- **Mauro Nicolao**, Fabio Tesser, and Roger K Moore. *A phonetic-contrast motivated adaptation to control the degree-of-articulation on Italian HMM-based synthetic voices.* In SSW8, pages 107–112, Barcelona, Spain, September 2013.

2017 in an invited paper on Frontiers in Robotics and AI, the ideas behind context-aware speech C2H synthesiser are further extended to create a comprehensive explanatory model for all possible talking/listening agent interactions. HTS-C2H is proposed as the test case (Moore and Nicolao, 2017) to prove the feasibility of such approach to artificial talking agents.

- Roger K Moore and **Mauro Nicolao**. *Toward a Needs-Based Architecture for 'Intelligent' Communicative Agents: Speaking with Intention*. Frontiers in Robotics and AI, 4:66, December 2017.

Chapter 1

Introduction

The observation that human talkers adapt their speech according to their listening situation was established more than a century ago by Lombard (Lombard, 1911). According to theories such as Lindblom's Hyper and Hypo articulation theory (H&H) theory of speech production (Lindblom, 1990), such modifications are caused by the need to transfer information from the talker to the listener while minimising the effort and maximising the effectiveness of their communication.

The most common examples of adjustments are:

- speech rate modification,
- pitch shifting,
- spectral energy reallocation.

That is, humans make continuous adjustments, continuously assessing the effectiveness of their modifications. Similarly, in Levelt's Perceptual Loop theory (Levelt, 1989), this adaptation is described as a talker's inner process, driven by a perceptual loop which constantly monitors the spoken outcome to assure the success of the communication process.

1.1 Limitations of current speech synthesis systems

Standard automatic speech synthesis systems still exhibit a rather limited range of speaking styles as well as an inability to adapt to the listening conditions in which they operate (Moore, 2007b; Moore and Nicolao, 2017).

Whilst the raw technical performance of contemporary spoken language systems has improved significantly in recent years (as evidenced by corporate giants such as Microsoft and IBM continuing to issue claim and counter-claim as to whose system has the lowest word error rates (Xiong et al., 2016; Saon et al., 2017)), in reality, users' experiences with such systems are often less than satisfactory. Not only can real-world conditions (such as noisy environments, strong accents, older/younger users or non-native speakers) lead to very poor speech recognition accuracy, but the 'understanding' exhibited by contemporary systems is rather shallow. As a result, after some initial enthusiasm, users often lose interest in talking to *Siri* or *Alexa*, and they revert to more traditional interface technologies for completing their tasks (Moore, 2016).

One possible explanation for this state of affairs is that, while component technologies such as automatic speech recognition and text-to-speech synthesis are subject to continuous ongoing improvements, the overall architecture of a spoken language system has not changed for quite some time. Indeed, there is a W3C 'standard' architecture to which most systems conform (W3C-SIF, 2000), as shown in Figure 1.1. Standardisation is helpful because it promotes interoperability and expands markets, however it can also stifle innovation by prescribing sub-optimal solutions.

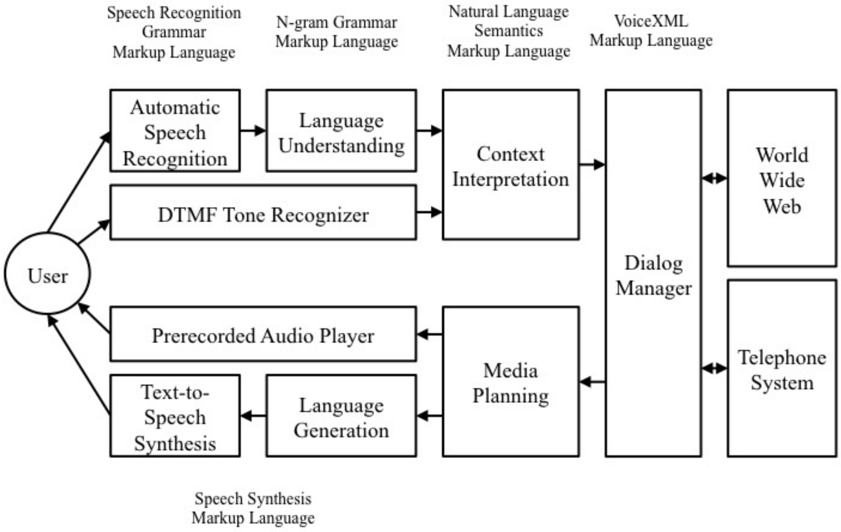


Figure 1.1: *Structure of the W3C 'standard'* Speech Interface Framework. *Figure adapted from (W3C-SIF, 2000).*

In the context of spoken language, there are a number of issues with the standard architecture depicted in Figure 1.1.

- The standard architecture reflects a traditional open-loop stimulus-response (‘behaviourist’) view of interaction; the user utters a request, the system replies. This is known as the ‘tennis match’ metaphor for language, where discrete messages are passed back and forth between interlocutors – a stance that is nowadays regarded as somewhat restrictive and old-fashioned (Bickhard, 2007; Fusaroli et al., 2014). Contemporary ‘enactive’ perspectives regard spoken language interaction as being analogous to the continuous coordinated synchronous behaviour exhibited by coupled dynamical systems: that is, more like a three-legged race than a tennis match (Cummins, 2011).

- The standard architecture suggests complete independence between the input and output components, whereas there is growing evidence of the importance of ‘sensorimotor overlap’ between perception and production in living systems (Wilson and Knoblich, 2005; Sebanz et al., 2006; Pickering and Garrod, 2007).

- The standard architecture fails to emphasise the importance of ‘user modelling’ in managing an interactive communication: that is, successful interaction is not only conditioned on knowledge about users’ directly observable characteristics and habits, but it also depends on inferring their internal beliefs, desires and *intentions* (Friston and Frith, 2015; Scott-Phillips, 2015).

- The standard architecture neglects the crucial teleological/compensatory nature of behaviour in living systems (Powers, 1973). In particular, it fails to acknowledge that speakers and listeners continuously balance the effectiveness of communication against the *effort* required to communicate effectively (Lombard, 1911) – behaviour that leads to a ‘contrastive’ form of communication (Lindblom, 1990).

As an example of context-dependant effort control, Hawkins provides an informative illustration of such *regulatory* behaviour in everyday conversational interaction (Hawkins, 2003). On hearing a verbal enquiry from a family member as to the whereabouts of some mislaid object, the listener might reply with any of the following utterances:

“I! ... DO! ... NOT! ... KNOW!”
“I do not know”

"I don't know"
"I dunno"
"dunno"
[ə̝̃ə̃ə̞̃]

where the last utterance is barely more than a series of nasal grunts. Which utterance is spoken would depend on the communicative context; the first might be necessary if the TV was playing loudly, whereas the last would be normal behaviour for familiar interlocutors in a quiet environment. Such responses would be both inappropriate and ineffective if the situations were reversed; shouting in a quiet environment is unnecessary (and would be regarded as socially unacceptable), and a soft grunt in a noisy environment would not be heard (and might be regarded as an indication of laziness).

Such *adaptive* behaviour is the basis of Lindblom's H&H theory of speech production (Lindblom, 1990), and it provides a key motivation for the model proposed in this thesis. It has been suggested that a new generation of talking agents should be developed that can adjust their *speech quality* and start to address behaviours exhibited by human talkers such as the H&H speech (Moore, 2007a, b; Moore and Nicolao, 2017).

The computational model of Hyper and Hypo articulation theory (C2H) (Moore and Nicolao, 2011; Nicolao et al., 2012, 2013), which is proposed in this thesis, represents the first *context-aware* model, that I know of, for *reactive speech synthesis*. Further, this synthesiser is embedded in a more general model of interaction between human or artificial agents (Moore and Nicolao, 2017). The general principle of reactive speech synthesis (or 'synthesis-by-analysis') exploits the ability of negative feedback control processes to monitor and adjust behaviour to achieve an intended perceptual effect (Powers, 1973).

The underlying idea of the C2H model is illustrated in Figure 1.2. Key features of this architecture are the *active control* on the text to speech synthesiser (TTS), and the comprehension model (or emulation of the *human speech recogniser, (HSR)* that is part of the negative feedback loop, which aims to minimise the . This emulation, which is effectively an *automatic speech recognition (ASR)*, is crucial for the artificial production model to adapt the speech production in order to maximise the receiver's reception.

Hence, a C2H synthesiser monitors the effect of its output and modifies its speech characteristics in order to maximise its communicative intentions.

In human speech, hyper/hypo-articulation behaviour is intrinsically related to the *effort* used by talkers during speech production. In synthesised speech, however, a valid effort measurement is less easy to achieve. In this thesis, phonetic-contrast

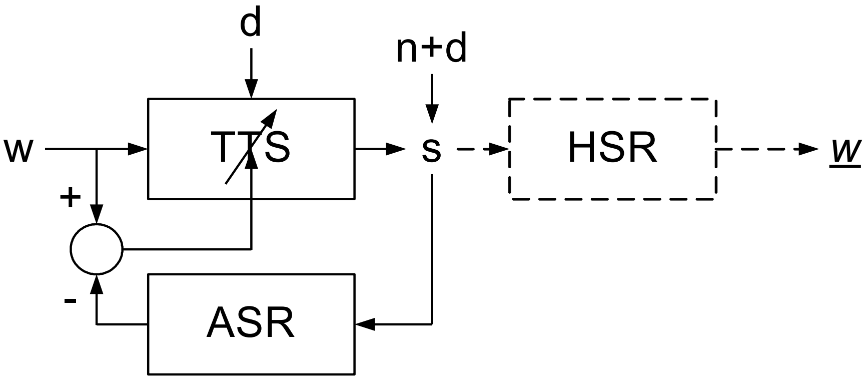


Figure 1.2: *Architecture for a reactive speech synthesizer, in which words (w) are converted to speech (s) which is subjected to noise (n) and disturbance (d). The synthesizer estimates the words perceived by the listener (w) using a feedback path involving Automatic Speech Recogniser ASR. A control loop compares w and w and the error signal drives the text-to-speech synthesis (TTS) to alter its output in such a way as to maximize recognition accuracy. The output of the reactive speech synthesiser is optimised to be received by a Human Speech Recogniser (HSR). Figure adapted from (Moore, 2007a).*

motivated transform is proposed to control the amount of energy (i.e., effort) that is used in the synthetic realisations. The hypothesis is that there are some speech configurations in the human speech production, named low-energy attractors, and that an interpolation/extrapolation along the key dimension of hypo/hyper-articulation can be obtained by controlling the distance to such attractors.

One of the key properties of the negative feedback loop is that it needs to provide a very efficient mechanism for assessing arbitrary disturbances. The implementation of the C2H model, presented here, measures the effectiveness of the communicative process in terms of speech intelligibility in noise conditions.

Recently, several studies have been proposed to tackle the adaptation of speech to environmental conditions (Tang and Cooke, 2010). A signal processing approach acting on the energy distribution and organisation and performed experiments with both natural and synthetic speech. Further studies concern automatic speech synthesis only, extending HMM-based synthesis to focus on optimisation of generated features in known noise conditions (Valentini-Botinhao et al., 2012), on data-driven adaptation of glottal source signal (Raitio et al., 2011a), and on interpolation/extrapolation of 'ad-hoc' trained models (Picart et al., 2011).

In order to test the degree of control on hypo/hyper-articulation speech, an implementation of the C2H model, with HMM-based parametric speech synthesiser (Zen et al., 2009) and STRAIGHT vocoder (Kawahara et al., 1999), is also developed in this thesis, and the intelligibility of the resulting speech utterances are evaluated.

Several hypotheses on the nature of the previously-introduced low-energy attractors are tested by modifying vowel and consonant production in an HMM-based synthesiser with scalable adaptation of its statistical models. The effectiveness of the proposed implementation is tested by performing acoustic analysis, objective, and subjective evaluations. A key assessment is to measure the control of the speech clarity in noisy conditions. Secondly, it is chosen to examine the similarities between the emerging system modifications and human behaviour itself. Different objective scoring methods are used in this thesis to assess the speech intelligibility: the speech intelligibility index (SII) and the index based upon the Dau measure (Dau). Among these metrics, the short-time implementation of SII, named eSII, is also chosen as the main perception measurement in the feedback loop to control the transformation.

1.2 Research questions

The context-aware speech synthesis model results from the search to answer some fundamental research questions on synthetic speech production.

1. The first research question regards the key human speech production characteristics that are missing in a standard speech synthesiser. *What are the characteristics that would allow a speech synthesiser to be aware of the communicative context?* An analysis of the principal theories of human speech production is reported.

2. The second fundamental question asks how the human production characteristics that are missing in synthesisers can be integrated in a speech synthesiser. Particular attention is given to Lindblom's H&H theory (Lindblom, 1990) and Powers' Perceptual Control Theory (PCT) (Powers, 1973). *How can standard theories of human speech production be integrated in a traditional synthesis system?* A computational model for speech synthesis is proposed to answer this question.

3. Another research topic concerns the transformations that can be applied to the synthesiser in order to balance the amount of energy and clarity of speech. *Can speech production energy be estimated in synthesisers? To what extent can phonetic contrast be a measure of the degree of effort in synthetic speech production?* Acoustic and intelligibility analyses are conducted to answer these questions.

4. The proposed energy-motivated transform predicts some expected modifications in the speech production. These have to be validated

against typical human behaviour. *Can the features of human speech in adverse conditions emerge from a phonetic-contrast transform?* Acoustic comparison between human and synthetic speech production is used to answer this question, comparing synthetic speech in adverse conditions with human speech in noise (Lombard, 1911).

5. The context-aware synthesiser should be able to track and follow the subsequent changes of the environment condition. The effects of adding a PCT-inspired perceptual feedback loop are investigated. *Can a perceptual feedback loop enable continuous adjustments of synthesiser production to follow the communicative context?*

Each question is addressed in a specific part of the thesis, as detailed next.

1.3 Thesis overview

Chapter 2 reports an overview of the salient aspects of human speech production. Some computational models of human speech production are also described.

Chapter 3 presents an overview of the state-of-the-art techniques for speech synthesis by machines.

Chapter 4 describes the main components of the C2H model. The fundamental principles behind its design are reported, its relevance to the research questions of this thesis are discussed.

In Chapter 5, an implementation of the C2H model is described using an HMM-based statistical parametric speech synthesiser. The implementation choices to create the voices, the transforms, and the control mechanisms are also detailed.

Chapter 6 presents the results of experiments assessing the effectiveness of the C2H model. Characteristics of human and C2H speech in adverse condition are considered and compared. Objective and subjective measures of intelligibility are also used to assess the quality of the proposed speech adjustments.

Chapter 7 summaries the main findings of this thesis, discussing the results, the effectiveness of the model, and some comments on future developments of the C2H approach.

Chapter 2

Speech Production in Humans

Contents

In order to model human speech production, some key characteristics of this behavioural phenomenon is described in this chapter. Some of the most affirmed theoretical explanations of the process that control human speech production are also reported. The main objective of this chapter is to highlight the behaviours that allow human speech to adapt to different contexts, addressing the research question n. 1 of this thesis.

2.1 Human speech production

The capability of conveying information using articulated sounds is a peculiar behaviour that characterises many living beings. Across many years of evolution, human beings have specialised in the use of these vocal sounds, and some complex speech mechanisms have been developed. Understanding the behaviours that control speech production and perception has always been of interest to research fields such as medicine, cognitive science, biology and computer science.

2.1.1 Speech communication

Speech communication identifies that set of behaviours that humans adopt to pass information among members of the group.

The main goal of speech communication is to transfer information from a speaker to a listening audience. Both talkers and listeners are active part of the process. The success of the communication might depend on the diverse motivation and communicative goal of the participants. Human speech communication has been so refined in centuries of evolution that people acquired the ability of transferring information with any level of abstraction, from concrete phenomena to ideas, concepts, and emotions. People are capable of compensating for most of the communicative conditions: noise, impairment, and disruptions.

Human speech relies on a combination of fundamental communicative units. These units may consist of articulatory gestures, phones, or morphemes. The range of gestures that can be produced by humans is quite wide and their extension can be scalable in a continuous space. Despite the continuous gesture space humans may use a relatively small number of configurations to render their speech vocalizations (Oudeyer, 2004). The set of gestures is also greatly language and cultural dependent.

Multi-modality

Human speech communication can use several parallel channels to exchange information (Levinson and Holler, 2014), rather than relying on the speech audio signal alone. Visual cues, such as hand or body *gestures*, *posture*, and other non-verbal modes, can communicate a lot of extra information to complete the speaker's or listener's communicative message. In particular, visual cues transferred by *lip reading* or in the listener's body language, can greatly increase the possibility of successful communication, when the environment channel is distorted by adverse conditions.

Finally, communication is also conducted via non-speech sound. For example, listener's *backchannel* responses are non-verbal sounds, a continuous assessment feedback to the speaker about the quality of a primarily one-way communication.

In principle, a context-aware computational model of speech production should also take these channel into account. In this thesis, however, only the main speech modality is modelled.

2.1.2 Variation in human speech

Variation is a crucial part of speech production in humans (Lindblom et al., 1992). When requested, speakers often fail to produce the same utterance twice, even in a controlled environment. Therefore, it is reasonable to assume that some innate component of the production system must generate chaotic variations in the speech signal, but without these variations affecting the general significance of the utterance. Moreover, even when speech is supposed to be neutral, it carries some additional information that listeners are able to decode, such as characteristics of the speaker, the listener, their relation, or the environmental conditions. Given that it is possible for humans to extract such information from the speech signal, it appears likely that speech variations thus constitute an additional fundamental dimension of communication.

Two main categories of variation can be identified in relation with the part of speech that is modified. The first type of variation includes the changes that speakers make on the information transferred. For example, a concept can be verbally expressed in several different ways, and variations in lexicon, language, grammar, intonation, gesture, and emotions, all contribute to the degree of complexity and abstraction in which the concept is uttered.

The second type of speech variation affects the acoustic characteristics of the sound itself, such as speech loudness, accent or inflection, spectral energy reallocation, production rate, and clarity. Exemplifying this, the Lombard speech reflex, described in § 2.1.2, explains a class of sound modifications that occur in presence of noise.

Lombard speech

The Lombard reflex (Lombard, 1911) is one of the earliest and more comprehensive description of variational behaviour in speech production. It describes the adjustments that a speaker performs in the presence of noise to compensate for the environment's influence and complete their communication with the listener. It is an unconscious reflex that occurs automatically, and is almost impossible to suppress, without specific training. It seems to be directly linked to the speaker's auditory feedback that analyses self-speech production, known as *self-monitoring* (Levelt, 1983). Evidence of the reflexive nature of this phenomenon has been gathered from experiments in which speakers (but not listeners) are exposed to noise over headphones. The talkers produce Lombard speech, despite the conscious awareness that potential listeners may not be experiencing the same noise (Junqua, 1996).

Several studies report the changes in speech that can be described as Lombard speech (van Summers et al., 2005; Garnier et al., 2006; Drugman and Dutoit, 2010). Common properties are:

- increased amplitude,
- mean word duration increment,
- changes in the fundamental frequency,
- changes in spectral tilt
- changes in format frequencies.

Most of these effects on speech can be explained by the observation that speech articulator movements that are more extensive than in quiet speech. This is known as hyper articulation.

When talkers try to enunciate clearly in noisy conditions, the Lombard reflex appears to be closely connected to the *speech production energetics*. In order to maintain the quality of the communication in adverse conditions, a certain quantity of energy has to be applied to speech realisations. This is reflected in more ample and precise articulatory movements, along with more intense airflow pressure, than in neutral speech realisations. This concept can be extended, and the most widely accepted approaches to speech communication modelling view production as only one side of the speech chain (Denes and Pinson, 2016). The additional energy that is observed in Lombard speech production is only a component of the total energy that both speaker and listeners have to balance to compensate the adverse effect of noise. The listener's energetic component is revealed in measurements of high level of cognitive load (i.e. neural activity level), when listening in noisy conditions.

Finally, an important characteristic of the Lombard reflex is that it is a source of speech variation that has specific causes and effects. Different degrees of noise perturbation produce proportional speech variation effects, and the predictability of these relationships allows repeatable experiments to be designed. Objective measures can therefore be used to quantify the degree of Lombard speech modifications subjected to multiple experimental conditions.

In this thesis, Lombard speech is considered the type of speech that a context-aware synthesiser should be able to produce in presence of a noise environment.

Sources of variation

Common sources of variation can be arranged in four categories (Hofe, 2011): *inter-personal*, *intra-personal*, associated with *clinical conditions*, and *listener-specific*. The first three variations arise from the speaker's physiology and education, social pragmatics, compensation of channel distortion, and any speech difficulties. Finally, the listener-specific category includes variation that a speaker produces to meet the needs of a specific audience, such as child-directed or acted speech.

Most speech variations are not fixed throughout the communication process, but evolve continuously due to ongoing changes of the context characteristics.

Some of the principal sources of variation are listed below (Hofe, 2011).

Emotions Emotional speech is one of the most prominent examples of speech variation in the literature concerned with speech technology. Some researchers try to investigate specific emotional states individually (Barra et al., 2006; Lee et al., 2005), other researchers build general frameworks within which specific emotions can be described in relation to each other (Scherer, 2003; Schröder, 2001). Several models place emotions in a 3-D emotional space, whose dimensions relate to specific voice characteristics (Grimm and Kroschel, 2007), such as, activation (active/passive), evaluation (positive/negative), and power (powerful/weak). Emotions in speech have also been investigated in conjunction with other modalities, such as facial expressions (de Gelder et al., 2013) and music (Juslin and Laukka, 2003).

Social pragmatics Speakers are known to vary their production depending on their social environment (Brown and Levinson, 1987). This results in dedicated speaking styles that individuals adopt in specific situations, for example with their family members or friends, in the presence of superiors, and at the workplace. Lindblom affirms the importance of modifying speech articulation for social acceptance (Lindblom et al., 1992), suggesting that *phonetic reductions presumably also serve to reduce the "social distance" between the two speakers*. That is, maximising speech clarity might not be the best communicative policy in every communicative context.

Errors and error compensation Speakers frequently make errors when they talk. This can be due to several cognitive and physiological causes, such as tiredness, low attention level, or articulatory movement obstruction. A number of typical speech behaviours that arise from errors, error correction, and error prevention are described in the literature: stop and restart due to speech disruptions (Levelt, 1983), or introduction of pauses and/or turn holders between words to reset the speech planning (Clark, 2002).

Coarticulation Coarticulation identifies the acoustic influence that neighbouring speech gestures within an utterance exert on each other (Hardcastle and Hewlett, 2006; Hura et al., 1992). This effect is not limited to direct neighbours but can also spread over longer segments. Generally, a speech sound influences preceding sounds more than succeeding ones, so coarticulation is therefore assumed to be connected to the planning of speech gestures. Coarticulation effects modify the speech signal consistently, when the production effort is reduced. Increment of speech rate and minimal amplitude of articulatory movements increase phone assimilation and approximation (van Bergem, 1993).

Environment compensation Environment conditions can influence speech production enormously. Several acoustic-phonetic consequences are observed in presence of communication barriers (Hazan and Baker, 2011) such as environmental noises (Lu and Cooke, 2007; Cooke, 2003) or the talker/listener's language proficiency (Lecumberri et al., 2016; Cooke et al., 2008). Such barriers define the linguistic and acoustic space from which speakers must sample their speech, if they want the communication to be successful.

Adaptation to the audience The speaker may also change their speaking style according to the audience that they are addressing. When adults – particularly parents – talk to children, they modify their speech (parentese or motherese) presumably to support infants in their language learning process. When teachers address a class, they may also change their linguistic and speech clarity style to help the students to minimise the learning effort.

Accommodation Communication accommodation (Giles, 2016) describes the tendency of two speakers involved in a speech interaction to converge towards a common acoustic and linguistic space. After a certain exposure to different accent, language proficiency, or semantic use of words, a speaker might adapt their production style to it, temporarily or permanently. These changes normally require prolonged two-way communication. As such, they are therefore not considered further in this thesis.

Effects on speech understanding

Listeners are normally aware of most of the effects of the previously described sources of variations in speech production. In presence of such variations, communication recipients adapt their understanding effort in order to successfully decode the speech message, as part of a joint communicative effort (Denes and Pinson, 2016). Noise and language barriers, production errors, etc. are normally classified as *adverse conditions* from the listener perspective, as understanding a spoken message in such disruptive conditions can be challenging task. If distortion is too severe, for example, no information can be transferred, regardless of any variation that is introduced by the speaker, A loud noise that completely masks the speech audio may force the listener to request a complete restart of the communication, or the switch to a different communicative modality. Moreover, decoding a message in a language with which the listener is unfamiliar may also require a too high cognitive effort, and this may impede any communication.

A complete description of the effect of speech variations on the listener's reception is beyond the scope of this thesis. However, the effects of adverse conditions on the listener that a talking agent (human or machine) must be aware of for successful speech comprehension to take place.

2.1.3 Context-awareness

Speakers are able to adapt their production continuously in reaction to the diverse sources of variations that are listed in the previous section. These circumstances influence an individual's speech dynamically, producing a communicative variation semantic that is typically shared by the members of the same language community. Effective communicators are normally fully aware of these context-derived variations, and can, where possible, exploit them proficiently to control the communication efficiency.

A crucial part of the human communication process is the ability of the speaker to *assess* the quality of their production in relation to the communicative context. One of the most well-accepted models of human speech production (Levelt, 1983) asserts that this is due to the mechanisms that detect (and correct) production errors (Postma, 2000; Hartsuiker and Kolk, 2001). This function is mainly devoted to the talker's auditory system that enables them to be aware of their own production – or *self-monitoring* (Levelt et al., 1999), – as well as sensing the environmental conditions. The auditory loop allows them to assess their outer speech quality against disturbances, and moreover to *predict* the quality of their production even before producing the actual sound. Humans seem thus to be able to use the internal representation of speech – or *inner speech* (MacKay, 1992) – and predict if it can correctly convey the desired information.

This prediction normally allows to adjust the linguistic and phonetic plan in order to increase the communication success.

Speech has evolved to be understood by humans and therefore speaker and listener share compatible tools to predict the success of communication. In direct communications, speakers are also able to detect the listener's requirements, that can be implicit (e.g., lack of attention), or explicit (e.g., spoken feedback). In a context-aware approach to speech production, the optimal speech outcome can therefore be modelled as the product of the *reaction* to an environmental context, rather the simple outcome of a set of behavioural rules. This will be discussed further in § 2.2.2.

2.1.4 Active control factors

Speech production can vary according to the *level of awareness* with which a speaker is willing to communicate.

This level of awareness can be compared to the *motivation* that drives the communicative intent. Each of the sources of variation that was previously reported (cf. § 2.1.2) can have larger or smaller effect on the talker's speech, depending on how much effort the speaker is willing to invest for optimising the variations. The talker's motivation influences the amount of effort that agents (talkers and listeners) are willing to invest in the communicative process. Depending on the degree of motivation, a talker, for instance, adopts more or less variations in their production to change the degree of clarity and the amount of errors.

2.2 Theoretical speech production models

The observations of speech behaviours that are described in the previous sections have led to several explanatory theories of speech production, and one of the most famous and controversial views on speech production, which is often also reflected in the standard speech synthesis systems, is the traditional open-loop stimulus–response ('*behaviourist*') view of interaction (Skinner, 1948). The verbal field is defined by Skinner as "*that part of behavior which is reinforced only through the mediation of another organism*". Therefore, another organism has to reinforce the speaker's behaviour, and direct their interaction modalities. That is, the speaker learns from experience and applies these techniques to the communication. In this approach, the speaker utters a sentence and waits for listener's response. Discrete messages are passed back and forth between interlocutors: a stance that is nowadays regarded as somewhat restrictive and old-fashioned (Bickhard, 2007; Fusaroli et al., 2014). Contemporary 'enactive' perspectives regard spoken language interaction instead as being analogous to the continuous coordinated synchronous behaviour exhibited by coupled dynamical systems: that is, more like a three-legged race than a tennis match (Cummins, 2011). Whereas the traditional speech production approach suggests complete independence between the input and output components, there is a growing importance of 'sensorimotor overlap' between perception and production in living systems (Wilson and Knoblich, 2005; Sebanz et al., 2006; Pickering and Garrod, 2007).

The computational model C2H that is proposed in this thesis is based on two main hypotheses. The first hypothesis states that the speaker's goal is to convey information to listeners. The second hypothesis is that there is a continuous trade-off between the effectiveness and the energetic cost of the communication. According to these hypotheses, speakers should try to maximise information throughput and minimise their energy expenditure.

The theories that contribute to modelling these behaviours are therefore discussed next in some details: the H&H and the PCT theories.

2.2.1 H&H theory

The most important inspiration for C2H derives from Lindblom's Hyper and Hypo articulation theory (H&H) theory of speech production (Lindblom, 1990). Lindblom proposes that speakers change their speech, according to different constraints, along a continuum that defines the degree of speech articulation: from hypo- to hyper-articulation. The constraints consist of energy consumption and information output. The mechanisms of control that are included in the H&H theory are defined as *output-oriented* and *system-oriented* control, as illustrated in Figure 2.1. The output-oriented control aims to maximise the clarity of the speech signal in order to support the listener's speech understanding. The system-oriented control aims, contrastively, to minimise the energy used in the speech signal

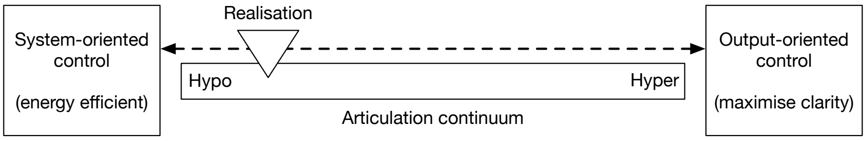


Figure 2.1: *According to H&H theory, both system- and output-oriented control influence the realisation of speech on the articulatory continuum. Adapted from (Lindblom, 1996).*

production. The two control goals interact in a perception loop, in which speakers monitor their speech outcome, its effects on the listener, and the effort used in the realisation. Corrections are made to adjust the error distance between the perceived communicative outcome and their prior intention. The resulting speech varies along a continuum from full-strength hypo-articulation to full-strength hyper-articulation.

When the production is pushed near the full-strength articulation extremes, it often results either extremely unintelligible or unreal-sounding speech.

Most of the types of variation, mentioned in the previous section (cf § 2.1.2), can be predicted by a theory such as H&H, in which energy and clarity are the driving forces. First, coarticulation and the Lombard reflex are motivated by energetics and clarity. Furthermore, characteristic accents that present language barriers may be linked to the specific articulatory strategies that allow speakers to produce sounds of their native language efficiently. It is hypothesised that the use of these energy-optimised strategies in another language produces the typical native-language dependent accents. Communication efficiency may cause talkers to cease to discriminate between specific sounds or to create new lexicon for specific tasks related to a specific community of people. Speech behaviour in different social contexts can also be related to energetics. For example, people may invest energy in the communication ans use a rather clear and well-formulated production when they communicate with their superiors to gain their approval.

Criticism of H&H theory

The main criticism against the H&H theory is motivated by the delay that such constant readjustments would require. Critics of speech production theories that rely on auditory feedback point out that there is a delay of around 60 ms between the generation of the motor command and the auditory feedback response in the brain (Guenther et al., 2005). This delay seems too long to allow for an immediate adaptation of articulatory gestures. This behaviour is more critical for consonants, that usually have much shorter duration, than for vowels, which typically are longer than 100ms (Junqua, 1996) and therefore can manage a similar delay. However, some forms of variation, such as the Lombard reflex, exhibit phone (particularly vowel) elongation as one of the most prominent features (Garnier et al., 2006). This lengthening may facilitate the accommodation of auditory feedback adjustment delays in adverse conditions.

Moreover, other feedback loops seem feature in the human communication process. For example, a short-latency feedback is proposed for the articulatory gestures from the sensory-motor system (Nasir and Ostry, 2006). As mentioned before, there is evidence of an internal loop that simulates the interaction between inner speech and the external environment (Borden, 1979; MacKay, 1992). This allows consequences of speech actions to be predicted, based on experience, even before the motor stimuli is generated (Levelt, 1983; Blackmer and Mitton, 1991). Auditory feedback measures of the clarity of past speech can be used to provide valuable predictions for the articulation of future speech. One of the most important criticisms originates from the work of Tabain. They studied particular languages with a high number of stop consonants (Tabain and Butcher, 1999). They found that the amount of variability of stop consonants is comparable to that observable in other languages with fewer stop consonants. These findings seem to confute the H&H theory, since a small articulatory distance between speech sounds should result in a reduced variability during production in order to maintain a sufficient perceptual distance between them. The clarity constraint of H&H theory relates to perceptual distances and the relationship to specific acoustic features may be non-linear. In a following study, Tabain analyses the influence of vowel sounds on preceding fricatives (Tabain, 2001), and reported small vowel-dependent variations in both articulatory and acoustic features. This seems to be contrary to the H&H theory, in which coarticulation should raise energy efficiency, taking advantage of allowable articulatory imprecisions. However, the H&H theory affirms that the overall energy expenditure must be minimised. No constraints about local energy behaviour are specified. It can be expected that certain gestures may inherently require more articulatory precision (and consequently energy) than others.

Evidence in favour of H&H theory

The presence of an output-oriented control system has been strongly evidenced (Lindblom, 2004). For example, the Lombard reflex seems to be originated by the need of maintaining the information flow. Moreover, it is an unconscious reflex that requires great effort to be suppressed (cf. § 2.1.2).

Speech acts, such as social pragmatics or acting, can represent another proof of output-oriented control. Speakers consciously modify their production to deliver a carefully crafted spoken performance that contains all the parameters (clarity, expressiveness, emotions, etc.) that match the audience expectations. Despite being a conscious effort, it can still be considered a further proof of the existence of an output-oriented control structure.

Though the Lombard reflex is listed as one of the main pieces of evidence in favour of the output-oriented control in the H&H theory, it remains true that a higher amount of energy is required to produce Lombard (rather than non-Lombard) speech. In a noisy environment, speech is loud and hyper-articulated (Garnier et al., 2006). For a given speech rate, related articulatory movements tend to produce both further and faster displacements of the articulators in Lombard speech. Since the articulators have a certain mass, movement amplitude is directly linked to energy consumption. Articulatory displacement is not the only energy expenditure, however, though speech production may initially seem to involve very few parts the human body, the cognitive load and brain resources that are needed to produce are quite remarkable. The air stream that is produced by lungs also requires energy. Both vocal effort and speech rate influence the amount of oxygen consumed by a speaker and hence require more calories to perform the related gestures (Moon and Lindblom, 2003).

The effects of output-oriented and system-oriented controls of the H&H theory may be quite difficult to quantify in some cases. A measure of the speech clarity can be inferred from analysing speech in noisy environment and assessing the Lombard reflex similarities. On the other hand, an energy consumption quantification may be more complicated to obtain as it is normally too intrusive to access the articulators and record their displacement.

The scientific objective of this thesis is to investigate the effectiveness of a computational model that can control the energy consumption and the clarity in speech production, based on pure acoustic motivations. This constitutes the proposed computational model of Hyper and Hypo articulation theory (C2H).

2.2.2 Perceptual Control Theory

H&H theory was introduced in the previous section (cf. § 2.2.1). It explains the variation in human speech behaviour as the result of a control loop architecture which monitors an internal and an external variable: energy and clarity. Powers' Perceptual Control Theory (PCT) formalises such an architecture by claiming that "*behaviour is the control of perception*" (Powers, 1973). That is, living organisms use their control structures to behave in such way that it would induce the desired – perceived – effects on the environment. This theory affirms that organisms do not aim to perform sequences of actions, but their goal is to achieve specific results.

It is commonly accepted that similar actions may generate different results, depending on external and internal factors, such as the environmental context or the individual's intent. Planning of an action for an individual could be a draining process, if it had to take into account of all those influences during the planning process. In PCT, managing results instead of actions is energetically convenient. Results are an internal representation of the desired state, that is simpler to visualise. Results are assessed by the perception loop that organisms use to perceive the environment.

The key PCT postulation is that the brain does not plan actions, but desired outcomes. This concept can be naturally transferred to speech production. When speakers produce speech, their brain does not explicitly plan any articulatory movement, but rather a representation of what the speech result should be. This representation allows for different degree of abstraction. It spans from the accurate assessment of the speech samples (i.e., signal similarity) to very high-level descriptions (e.g., expressiveness or clarity). The speech output is monitored by an auditory feedback loop. Translated into the H&H theory terminology, humans adjust their behaviour in an output-oriented way, controlling the effect of speech production (clarity). The feedback from the sensing organs is compared with the intentions, and if differences are detected, then appropriate actions are performed to reduce the discrepancy.

The PCT approach to speech production can explain the mechanisms of both the H&H theory and the Lombard reflex. It can also predict how clear speech can still be produced by test subjects in jaw perturbation experiments.

Application of PCT mechanisms to speech has caused some controversy. Some critics note that speakers with acquired deafness continue to speak, even though their auditory feedback path has been eliminated. It is emerging that the PCT feedback indicates a more general feedback loop, not only the auditory one. A wide range of possible feedback paths are available in humans. These include the tactile feedback from the speech articulators (Nasir and Ostry, 2006), the internal simulation of the speech process (Borden, 1979), and the observation of listener back-channelling reactions, which provide the perceived representation to compare with the intention. As a result, deaf speakers can actually provide supporting evidence of the self-monitoring auditory loop, as regular speech therapy to assist them to maintain the clarity of their speech (Brainard and Doupe, 2000).

The perception feedback loop of complex organisms monitors aspects of the environment as well as of the organism itself (e.g. external clarity and internal energy). In this theory, the brain reacts to all sources of feedback to produce targeted actions in response to some perturbation of the desired state. The existence of an ideal reference condition is a fundamental requirement to specify the desired state.

However, there are many cases in which the type of action and the intensity of its application are not unique. Actions can be then selected from a range of options. Experience or prediction of the action effect can help to select the optimal action.

2.2.3 Environment prediction

In the previous section, it is reported that the PCT can track the ongoing evolution of the environment (e.g., background noise changes) by trying to control the outcome of the interaction (e.g., maintaining speech clarity).

In this context, the uncertainty regarding the future evolution of the environment might be mitigated by the use of a predictive model of it. This is normally derived from experience, It should not be considered as set of rules for action planning, but instead as a set of constraints on the states in which the environment can evolve. In this way, the estimated reaction applied to maintain speech clarity can be interpreted as a form of mental simulation (or predictor) that emulates the consequences of possible actions prior to action selection (Hesslow, 2002; Grush, 2004). Another insight to emerge from this approach is that the depth of the search for a possible action can be the observed consequences regarded as analogous to effort, i.e., the amount of energy devoted to finding a solution (Moore and Nicolao, 2017).

In the H&H theory, a predictive model of the environmental effects on speech might help to reduce the latency of the reactions, and to exclude useless actions at an early inner-loop level. A speaker's previous experience can help them to estimate the environmental and contextual conditions and to change their speech output, accordingly. Listeners do not always need to extract every information detail from the speech signal. Prior knowledge of the listener is often combined with the signal information to decode the intention of the speaker (Lindblom, 1996). The role of knowledge becomes more important when the signal becomes more hypo-articulated. This mechanism is referred to as the *signal+knowledge* approach to speech decoding.

2.3 Computational models of speech production

This section describes some of the principles in the theoretical models in the previous sections can be translated into algorithms and function to define *computational models* of speech production. Computational modelling is the use of algorithms to simulate and study the behaviour of complex systems. It can be particularly effective when the internal mechanism of the systems is unknown, and only the inputs and outputs are measurable. In such cases, a computational model can be used in simulations in which the outcome has to be coherent with the system observations.

One of the earliest models of speech production can be identified in the source-filter representation (Fant, 1970) that is depicted in Figure 2.2. This model describes

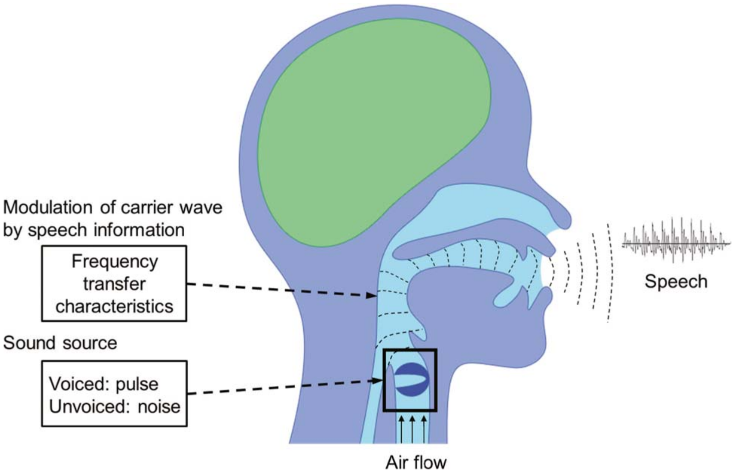


Figure 2.2: *Source-filter model of human speech production. Figure reproduced from (Tokuda et al., 2013)*

speech production as a two-stage process that originates from the generation of a sound excitation signal by the vocal cords, which is then shaped or filtered by the resonant properties of the vocal tract. The two parts of the model act independently. Most of the source spectrum shaping thus occurs in the oral cavity and optionally the nasal cavity. Although it has often been pointed out that, for various reasons, the linearity assumption is not strictly true, the model has been extremely influential in speech sciences for decades, forming the theoretical foundation for most of the parametric speech synthesis (see § 3.2) and for a wide range of applications in speech signal processing.

Other advanced computational models for speech production exploit the idea of the auditory feedback loop. Some of these further include somatosensory feedback, such as DIVA (Guenther and Perkell, 2004; Lane et al., 2007). The DIVA model of speech production assumes that lexical retrieval of strings of words leads to sequential activation of speech-sound map cells, each corresponding to a word, syllable, or phoneme. When one of these cells is activated, it sends signals to cells in the model's auditory, somatosensory, and primary motor cortical areas. These signals lead to production of the speech sound through a feed-forward system and a feedback system.

Some speech production models also produced physical or software implementations of the hypothesised processes. For example, Hofe's AnTon (animatronic tongue and vocal tract) (Hofe and Moore, 2008; Hofe, 2011) is a physical model of speech production that creates speech sounds that result solely from the anatomical structures that are implemented, rather than being artificially engineered. It applies the basic principles of the H&H theory by measuring the energy involved in the production as a function of the physical movements of the artificial articulators.

Atrianaki's MAGE system (Astrinaki et al., 2013) is one of the first implementations of continuous adjustments in a speech synthesiser framework, allowing reactive speech synthesis with short-latency control of the speech outcome characteristics. Although it allows modifications pf speech production that are compatible with Lombard speech, the outcome control is not driven by any human-related (auditory or somatosensory) feedback. Rather, the outcome is controlled by an interactive user interface, thus , this system cannot be regarded as a proper computational model in the sense required here.

Moore's PRESENCE (Moore, 2007a, b) is also a computational model for speech production which is designed to be applied to speech synthesis systems. His first challenge suggested that systems should talk 'clearly', and he noted that no contemporary text to speech synthesiser (TTS) had addressed the classic H&H behaviour exhibited by human talkers. Moore went on to develop this particular idea further and proposed a new approach to speech generation that

1. selects speech characteristics that are appropriate to the needs of the listener,
2. monitors the effect of its own output,
3. and modifies its behaviour according to its internal model of the listener.

The computational model C2H, which is proposed in this thesis, represents the first comprehensive model for *reactive speech synthesis* (Moore and Nicolao, 2011; Nicolao et al., 2012). This synthesiser is presented within a more general model of interaction between human or artificial agents (Moore and Nicolao, 2017). The general principle of reactive speech synthesis (or 'synthesis-by-analysis') exploits the ability of negative feedback control processes to monitor and adjust behaviour in order to achieve an intended perceptual effect (Powers, 1973).

Chapter 3

Speech Production in Machines

Contents

This chapter discusses various approaches to artificial speech generation, considering their usefulness for the specific focus of this work, which is to create a speech synthesiser that can *react* to external stimuli, such as environmental noises, and adjust its speech production.

Firstly, a historical overview of the different techniques that have been applied to speech generation is presented. Secondly, recent parametric speech synthesis methods are described, introducing the mechanisms and the terminology relevant to the creation of the context-aware speech synthesiser. The most crucial limitations of current speech synthesisers are highlighted, addressing the research question n. 1 of § 1.2. The final section overviews processes by which artificially-generated speech can be modified to reproduce some of the context adaptations that are observed in humans.

3.1 Speech synthesis

Speech synthesis is the process which enables machines (or *talking agents*) to generate speech-like audio output. Voice interaction is a common feature of many human-computer interfaces (HCIs). Speech synthesisers are used in a wide range of application areas, such as

- embodied virtual assistants (Amazon Echo, Google Home, Apple HomePod, etc.),
- speech enabled intelligent personal assistant (Siri, Alexa, etc.),
- automatic translation system output (Google Translate),
- language tutoring systems (correct pronunciation feedback),
- information access by telephone: interactive voice response (IVR),
- assistive technologies (aid to visually impaired users, voice reconstruction),
- interactive systems (e.g. games, simulators, toys).

Speech synthesisers have linguistic classes (a message or *concept*), normally expressed with formatted text, as input, and a speech *waveform* as output.

Speech synthesis can be regarded as the "inverse" process of speech recognition. In speech recognition, the redundant complexity of the audio signals is reduced to low-dimensional linguistic classes by clustering common speech characteristics, as depicted in Figure 3.1.

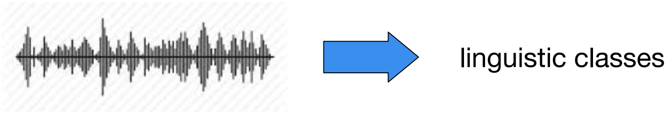


Figure 3.1: *Speech recognition*

On the other hand, in speech synthesis is a large-scale inverse problem: highly compressed linguistic classes (e.g., ideas, words, phones, etc.) are "decompressed" into audio, as depicted in Figure 3.2.

The objective of speech synthesis is to create a waveform which is:

- Meaningful: speech should be able to convey a message to the listener
- Intelligible: the message should be audible and clearly understandable

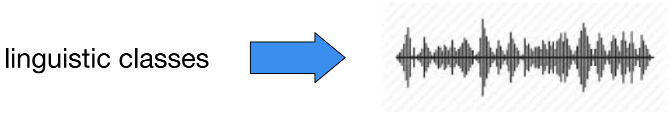


Figure 3.2: *Speech synthesis*

- Expressive: the intelligible message should be sound acceptable to listeners (in terms of naturalness, expressiveness, emotions, etc.)

The general diagram of an automatic speech synthesis pipeline is depicted in Figure 3.3. The descriptions of concept and natural language generations are outside the scope of this thesis, but are assumed to result in a formatted text string plus some attributes referring to the quality of the intended speech production. The text to speech synthesiser (TTS) box contains the elements commonly referred to by the term "speech synthesis". The *linguistic analysis* stage maps the input text string into a standard form; determines the structure of the input, and finally decides how to pronounce it. The *waveform generation* or *synthesis* of the speech signal converts the symbolic representation into an actual waveform.

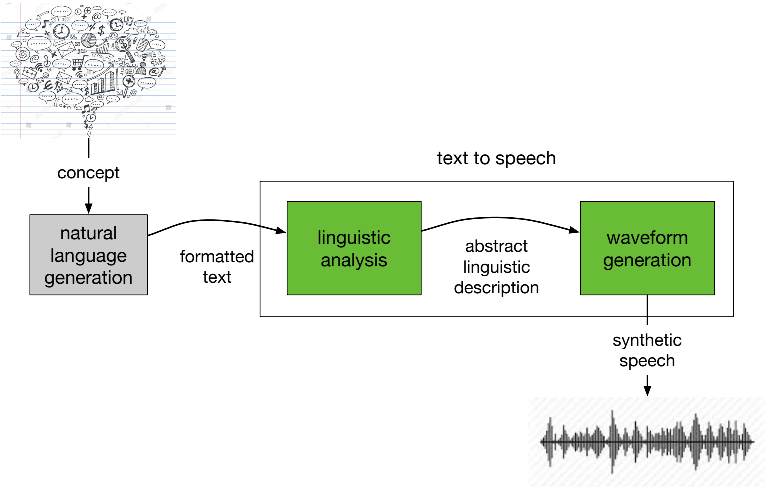


Figure 3.3: *Automatic speech synthesis*

3.1.1 History of speech synthesis

Human interest in speech production has led many researchers to attempt to produce artificial speech. The first production systems were mechanical devices that generated human-like sounds, normally no longer than a monosyllabic word.

The *apparatus* by von Kempelen (Dudley and Tarnoczy, 1950) is one of the earliest examples of these machines. However, only a limited range of speech sounds could be reproduced, such as the consonants [b], [p], [m], and [n].

Technological advances in the field of electronics allowed the development of electrical speech synthesis in the early 20th century. Electronic circuits, such as oscillators and filters that can manipulate a waveform, substituted the physical resonators. Vowels could be produced by combining periodic waveforms with different frequencies and amplitudes, producing the characteristic vowel formant frequencies. The first formant synthesisers were then created, including the Vocoder by (Dudley, 1939), in the 1930s.

The most successful way to produce artificial speech is based on the source-filter model described in the previous chapter (cf. § 2.3). In Figure 3.4, the functional block diagram of such a model is depicted. Recently, the term *parametric synthesis* has become widely used to describe synthesisers which are driven by a set of input parameters, typically those that define the excitation signal and the waveform spectral envelope. Formant frequencies, articulatory parameters, linear predictive coefficients (LPC), and Mel-generalized cepstral (MGC) coefficients are among the most popular parameters used to define the waveform spectral envelope.

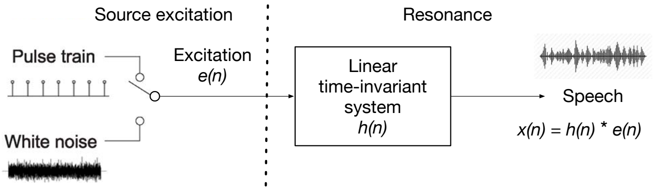


Figure 3.4: *A source-filter model that simulates the human speech production model reported in Figure 2.2. Adapted from (Tokuda et al., 2013).*

In the reminder of this section, brief descriptions of the principal synthesis methods are reported. These methods are typically based on the source-filter model, but the unit selection synthesis method discussed below is an exception to this rule.

Formant synthesis Formant synthesisers (Klatt, 1987; Holmes, 1983) were the first form of synthesiser that were based on the source-filter model of speech production (Fant, 1970), shown in Figure 3.4. The control parameters of a typical formant synthesiser are:

source parameters These describe the excitation signal, voiced/unvoiced characteristics, fundamental frequency, and intensity (loudness);

filter parameters These include the resonance frequencies (formants), bandwidth, and amplitude of the filters that shape the vocal tract.

A typical excitation signal is a pulse train, which is suitable for formant synthesisers due to its spectrum flatness (Oppenheim and Schafer, 2014). The pulses can also be shaped in such a way that they mimic the pulses produced by the vocal cords. However, the design of a more realistic excitation signal is not trivial as in humans it has highly non-linear characteristics (Drugman and Dutoit, 2010). The number of pulses per second controls the fundamental frequency of the speech signal, i.e., the voice pitch. In the case of unvoiced phones, an alternative excitation function is used, a random white noise signal, and filter parameters instead are responsible of the control of the synthesised speech spectrum. Rather good voice quality and intelligibility can be achieved using the formant synthesisers (Holmes, 1983, 1986), but adaptation to different speech styles or speaker identities is not easily achievable. Format synthesisers were quite popular until mid the 1990s, but were used mainly for research purposes.

Articulatory synthesis Articulatory synthesisers use the properties of human articulatory system models to generate vocal tract filters that shape the speech signal spectrum. Generation parameters are derived from physiology instead of acoustic properties of the vocal tract. The articulator positions, or the states of muscles controlling them are some of the most commonly used parameters. Two methods can be applied to synthesise speech from an articulatory model:

- the vocal tract filter can be computed from the articulatory configuration, and a standard source-filter model can be applied;
- the shape of the vocal tract can be modelled, and finite-state analysis of the air-stream behaviour in the vocal tract for that configuration is used to produces the speech waveform.

In the first approach, the challenge is to create a reliable mapping function between the articulatory movements and the vocal tract filter. The second approach requires a huge amount of knowledge and computational resource in order to derive solutions to the turbulence air-stream model.

These two synthesis approaches can also be regarded as parameter-to-sound systems, or *vocoders*. A vocoder is a system that takes a parametric representation of sound and generates a related speech waveform. Neither the formant nor the articulatory synthesisers provide methods to generate the sequences of parametric representations that will be needed to compose the speech message. Instead, this type of synthesisers are normally used for *copy-synthesis* or re-synthesis of speech.

In order to produce a more complete model of speech production, a speech synthesiser also has to provide the capability to generate the parametric representations from text (or concepts). For this reason, since the mid-1990s, several systems have been proposed to analyse the input text information, to convert it into linguistic and phonetic components, and to generate the sequence of parametric commands that produce the final waveform. These systems can be categorised into two main approaches: *waveform-based* or *model-based*.

Unit selection The *waveform-based* synthesisers do not use the source-filter approach to create speech audio. Instead, libraries of pre-recorded sound segments (units) that are concatenated to form the utterances (Taylor, 2009). In the mid 1950s, the first studies regarding the possibility of using sub-word segments of recorded speech to form new utterances were proposed (Harris, 1953). At the time, the recorded speech was segmented, cut, and recomposed mechanically from speech units that were stored on magnetic storage devices.

This type of synthesis is called also concatenative synthesis, as one of its fundamental components is the method that is used to concatenate individual units to form a new utterance.

Another challenge of unit selection synthesis is the algorithms that are used to choose the units to be concatenated. The objective function of the process aims to create a transition between two units that is minimally obtrusive. In Figure 3.5, the basic technique for selecting the best segments from a pre-recorded data set is displayed. It relies on the notions of *target cost*, which describes the similarity between the database sample and the required unit, and *concatenation cost*, which defines the degree of obstruction between the two units. Since this type of synthesis uses real speech samples, the output tends to sound very natural. For this reason, unit selection synthesisers are still the most widely used commercial synthesisers, e.g., Alexa and Siri. The main limitations of the unit selection synthesis are that they have large data storage requirements and proportionally large computational costs for the unit search algorithm.

Adaptation of unit selection speech to render specific expressiveness or new speaker identity, is normally difficult and often requires the creation of a new unit dataset for each speech style generated (Holmes and Holmes, 2001).

Statistical parametric speech synthesis In direct contrast to the selection of actual instances of speech from a database, in the *model-based* approach, a set of generative models such as HMMs, are used to map the linguistic analysis directly into the parametric space (Yoshimura et al., 1999; Ling et al., 2007; Black et al.,

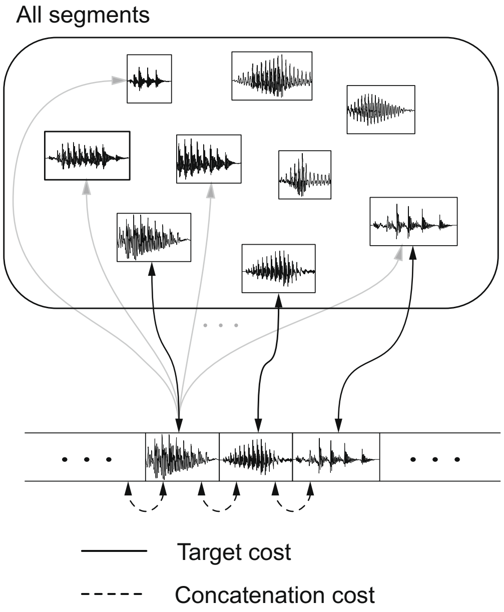


Figure 3.5: *Overview of the general unit-selection scheme. Solid lines represent target costs and dashed lines represent concatenation costs. Figure reproduced from (Zen et al., 2009).*

2007; Zen et al., 2007b). A detailed description of this statistical parametric speech synthesis (SPSS) is reported below in § 3.2.

End-to-end systems Following the recent increase in research applying neural networks to speech technologies, a new and effective type of synthesiser seems to be emerging: the end-to-end (E2E) synthesiser.

The best performing neural systems to date are WaveNet (van den Oord et al., 2016), a flexible model for audio generation which uses dilated, causal convolutions (with residual/skip connections) to form a conditional probability for the next time step value. For TTS tasks, WaveNet is conditioned on linguistic features from an existing TTS system and so is not fully end-to-end. In addition, its conditional model is auto-regressive and thus is prohibitively slow for many real-time applications. In return for these limitations, however, WaveNet produces very high-quality audio samples, surpassing strong concatenative and parametric baselines in naturalness.

Another attempt to move towards E2E system is DeepVoice (Arik et al., 2017) in which the entire TTS pipeline is implemented with neural networks. The computational time required is reduced from WaveNet, but the approach still requires separate training for the many different steps of the pipeline. The great complexity of training and deploying these models makes it harder to adapt the pipeline to new environmental contexts.

The field of the E2E synthesisers is an active research area, however the only fully E2E model so far seems to be Tacotron (Wang et al., 2017). Tacotron is able to produce samples of reasonable naturalness with much more efficiency than other systems as it has the significant advantage of operating at frame-level. The adaptation techniques and tuning of such complex system are still unexplored.

3.2 Statistical parametric speech synthesis

Statistical parametric speech synthesis (SPSS) (Zen et al., 2009) is one of the major approaches in the TTS systems. SPSS uses an acoustic model to represent the relationship between linguistic and acoustic features and a vocoder to render a speech waveform given a set of acoustic features. This approach offers various advantages over concatenative speech synthesis (Hunt and Black, 1996), such as a small generation model and the flexibility to change its voice characteristics (Yoshimura et al., 1997; Tamura et al., 2001; Miyanaga et al., 2007). However, the naturalness of the synthesized speech from SPSS is not as convincing as that of the best samples from concatenative speech synthesizers. Three major factors are reported that can degrade the naturalness: quality of vocoder, accuracy of acoustic model, and effect of over-smoothing (Zen et al., 2009).

Although there have been many attempts to develop a more accurate acoustic model for SPSS (Yoshimura et al., 1999; Zen et al., 2006; Shannon et al., 2013; Koriyama et al., 2014; Cai et al., 2015), the hidden Markov model (HMM) (Rabiner, 1989) remains the most popular approach.

SPSS still represents the high quality and flexible TTS method that offers full control over every aspect of the synthesised speech. The SPSS training and generation processes are described in Figure 3.6.

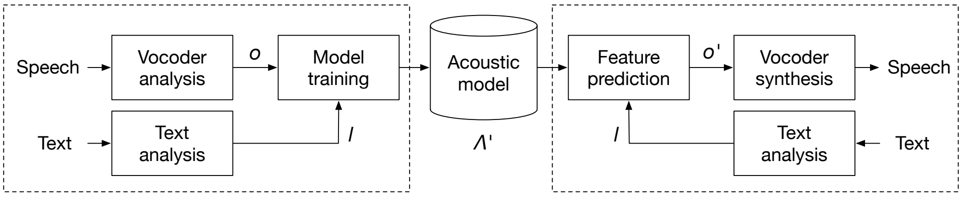


Figure 3.6: *Flowchart of a typical SPSS system. Adapted from (Zen, 2015).*

The *training* process aims to statistically model the relationship between linguistic and acoustic features. An acoustic model $\hat{\Lambda}$ is trained to model the conditional distribution of an acoustic feature sequence $\hat{\mathbf{o}}$, given a linguistic feature sequence $\hat{\mathbf{l}}$, as per:

$$\hat{\Lambda} = \arg\max_{\Lambda} p(\mathbf{o}|\mathbf{l}, \Lambda) \tag{3.1}$$

The model Λ in SPSS is normally described by HMMs (Rabiner, 1989).

At the *generation* or *synthesis* stage, a text to be synthesized is first converted to the corresponding linguistic feature sequence. Then the most probable acoustic feature sequence $\hat{\boldsymbol{o}}$ for the linguistic feature sequence $\boldsymbol{l}$ is predicted from the trained acoustic model $\hat{\Lambda}$ as:

$$\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{O}} p(\boldsymbol{o}|\boldsymbol{l}, \hat{\Lambda}) \tag{3.2}$$

Finally, a speech waveform is rendered from the predicted acoustic feature sequence using a *vocoder*.

3.2.1 HMM-based speech synthesis

HMMs are commonly used in speech synthesis to generate parameter streams for parametric synthesis (Tokuda et al., 2000, 2013). The main improvement in comparison with earlier rule-based formant synthesisers is the generation of the control parameters by statistical models. This allows the extensive statistical modelling tool-set developed for automatic speech recognition to be employed in speech synthesis. An example of such tool-sets is HMM/DNN-based speech synthesis framework (HTS) (HTS working group, 2012) which is model-based generative set of tools, based on HMM Toolkit (HTK) (Young et al., 2002), that has also been extended to generate articulatory parameters (Ling et al., 2008), as well as hybrid solutions using unit selection (Taylor, 2006; Black et al., 2007). However, the speech waveforms themselves are still generated by a source-filter structure akin to the one used in early formant synthesis methods. "Whilst most approaches aim to generate cepstral parameters, some generate formants and, in this sense, the HMM approach can be seen as a direct replacement for the provision of these rules by hand" (Taylor, 2009).

A significant benefit of HMM-based speech synthesis is that is has great flexibility in changing speaker identities, emotions, and speaking styles. The training and synthesis parts of a standard HMM-based speech synthesiser are depicted in Figure 3.7 and described next.

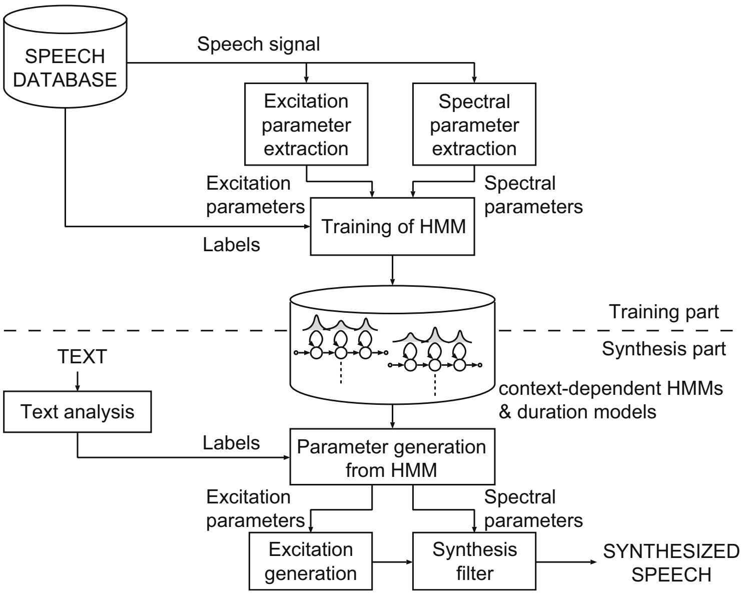


Figure 3.7: *Block-diagram of an HMM-based speech synthesis system. Figure reproduced from (Zen et al., 2009).*

Training

HMM-based synthesis with single Gaussian state-output distributions uses a statistical model described by (Zen, 2015)

$$
\begin{aligned}
p(\boldsymbol{o}|\boldsymbol{l},\Lambda) &= \sum_{\forall \boldsymbol{q}} p(\boldsymbol{o}|\boldsymbol{q},\Lambda)P(\boldsymbol{q}|\boldsymbol{l},\Lambda) && (3.3)\\
&\simeq \sum_{\forall \boldsymbol{q}} \prod_{t=1}^{T} p(\boldsymbol{o}_t|q_t,\Lambda)P(q_t|q_t-1,\boldsymbol{l},\Lambda) && (3.4)\\
&= \sum_{\forall \boldsymbol{q}} \prod_{t=1}^{T} \mathcal{N}(\boldsymbol{o}_t;\mu_{q_t},\Sigma_{q_t})a_{q_t q_{t-1}} && (3.5)
\end{aligned}
$$

where $\boldsymbol{o}_t$ is an acoustic feature vector at frame t, T is the number of frames, $\boldsymbol{q} = q_t, \dots, q_T$ is a sequence of hidden discrete states, q_t is a hidden state at frame t, μ_{q_t} and Σ_{q_t} correspond to the mean vector and covariance matrix associated with the state-output distribution at q_t, a_{ij} is the transition probability from state i to j, $a_{q_1 q_0}$ is the initial state probability of state q_1, $\boldsymbol{l} = l_1, \dots, l_P$ is a sequence of linguistic features associated with $\boldsymbol{o}$, l_p is a linguistic feature vector associated with p-th phoneme, and Λ denotes a set of context-dependent HMMs. Figure 3.8

shows an exemplar structure of an observation vector, $\boldsymbol{o}_t$. The static observations are often accompanied by their dynamic description to constraint the maximum variability permitted. Some computational simplifications are used to efficiently train different-order feature derivatives (reported in § 3.2.4 below).

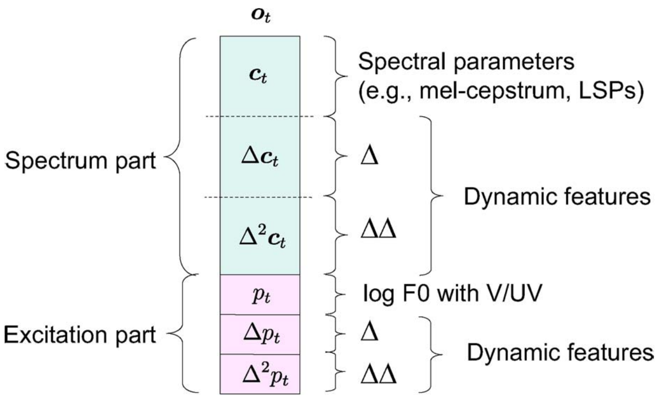


Figure 3.8: *Example of an observation vector. Figure reproduced from (Tokuda et al., 2013).*

The parameters of the HMMs can be estimated with the maximum likelihood (ML) criterion by the expectation-maximization (EM) algorithm (Rabiner, 1989). It can be seen from eq. 3.5 that $\boldsymbol{o}_t$ depends only on q_t; generative model statistics remain unchanged if the associated discrete states do not change. It is well known that the acoustic features of a particular phone in human speech are not only determined by the individual phonetic content but also affected by various background events associated with the phone. The background events which can affect the acoustic realization of a phone are referred to as its contexts. There are normally around fifty different types of contexts used in SPSS (Tokuda et al., 2002). The standard approach to handling contexts in HMM-based acoustic modelling is to use a distinct HMM for each individual combination of contexts, referred to as a context-dependent HMM. The amount of available training data is normally not sufficient for robustly estimating all context-dependent HMMs, however, since there is rarely sufficient data to cover all of the context combinations required. To address these problems, top-down decision- or regression-tree based context clustering (Odell, 1995) is widely used.

Synthesis

The generation stage of the HMM-based synthesis aims to find the most probable acoustic feature sequence $\hat{\boldsymbol{o}}$ given a linguistic feature sequence $\boldsymbol{l}$ and a set of trained context-dependent HMMs $\hat{\Lambda}$ (Zen, 2015). Eq. 3.2 can be approximated as

$$\begin{aligned}
\hat{\boldsymbol{o}} &= \arg\max_{\boldsymbol{o}} p(\boldsymbol{o}|\boldsymbol{l}, \hat{\Lambda}) \\
&= \arg\max_{\boldsymbol{o}} \sum_{\forall \boldsymbol{q}} p(\boldsymbol{o}, \boldsymbol{q}|\boldsymbol{l}, \hat{\Lambda}) && (3.6) \\
&\approx \arg\max_{\boldsymbol{o},\boldsymbol{q}} p(\boldsymbol{o}, \boldsymbol{q}|\boldsymbol{l}, \hat{\Lambda}) && (3.7) \\
&= \arg\max_{\boldsymbol{o},\boldsymbol{q}} p(\boldsymbol{o}|\boldsymbol{q}, \boldsymbol{l}, \hat{\Lambda}) P(\boldsymbol{q}|\boldsymbol{l}, \hat{\Lambda}) && (3.8) \\
&\approx \arg\max_{\boldsymbol{o}} p(\boldsymbol{o}|\hat{\boldsymbol{q}}, \hat{\Lambda}) && (3.9)
\end{aligned}$$

where $\hat{\boldsymbol{q}}$ is the predetermined state sequence derived from $P(\boldsymbol{q}|\boldsymbol{l}, \hat{\Lambda})$.

If the HMMs have left-to-right topologies and single Gaussian state-output distributions, the solution of eq. 3.9, when $\boldsymbol{o}_t$ has uniquely static features, becomes

$$\begin{aligned}
\hat{\boldsymbol{o}} &= \arg\max_{\boldsymbol{o}} \prod_{t=1}^{T} p(\boldsymbol{o}_t|\hat{\boldsymbol{q}}_t, \hat{\Lambda}) && (3.10) \\
&= \arg\max_{\boldsymbol{o}} \prod_{t=1}^{T} \mathcal{N}(\boldsymbol{o}_t; \mu_{\hat{q}_t}, \Sigma_{\hat{q}_t}) && (3.11) \\
&= \arg\max_{\boldsymbol{o}} \mathcal{N}(\boldsymbol{o}; \mu_{\hat{\boldsymbol{q}}}, \Sigma_{\hat{\boldsymbol{q}}}) && (3.12) \\
&= \mu_{\hat{\boldsymbol{q}}} && (3.13)
\end{aligned}$$

where $\mu_{\hat{q}_t}$ and $\Sigma_{\hat{q}_t}$ are the mean vector and covariance matrix associated with $\hat{q}_t$, and $\mu_{\hat{\boldsymbol{q}}} = [\mu_{q1}^{\mathsf{T}}, \ldots, \mu_{qT}^{\mathsf{T}}]^{\mathsf{T}}$ and $\Sigma_{\hat{\boldsymbol{q}}} = \text{diag}[\Sigma_{\hat{q}1}^{\mathsf{T}}, \ldots, \Sigma_{\hat{q}T}^{\mathsf{T}}]^{\mathsf{T}}$ are the mean vector and the covariance matrix over the entire utterance given $\boldsymbol{q}$.

Since this solution produces speech with clear discontinuities at the phone boundaries, dynamic feature smoothing (Tokuda et al., 1995a) is introduced. The derivatives of the vector in Figure 3.8 are addressed as functions of the static observations, i.e., $\boldsymbol{o}_t = \boldsymbol{W}\boldsymbol{c}$. With this assumption, eq. 3.12 becomes

$$\hat{\boldsymbol{c}} = \arg\max_{\boldsymbol{c}} \mathcal{N}(\boldsymbol{W}\boldsymbol{c}; \mu_{\hat{\boldsymbol{q}}}, \Sigma_{\hat{\boldsymbol{q}}}) \qquad (3.14)$$

A method to implement such vector generation is described in § 3.2.4 below.

The optimal generative model HMM is derived from the phonetic sequence and it is selected using a linguistic-based decision tree. A representation of such generation process is depicted in Figure 3.9.

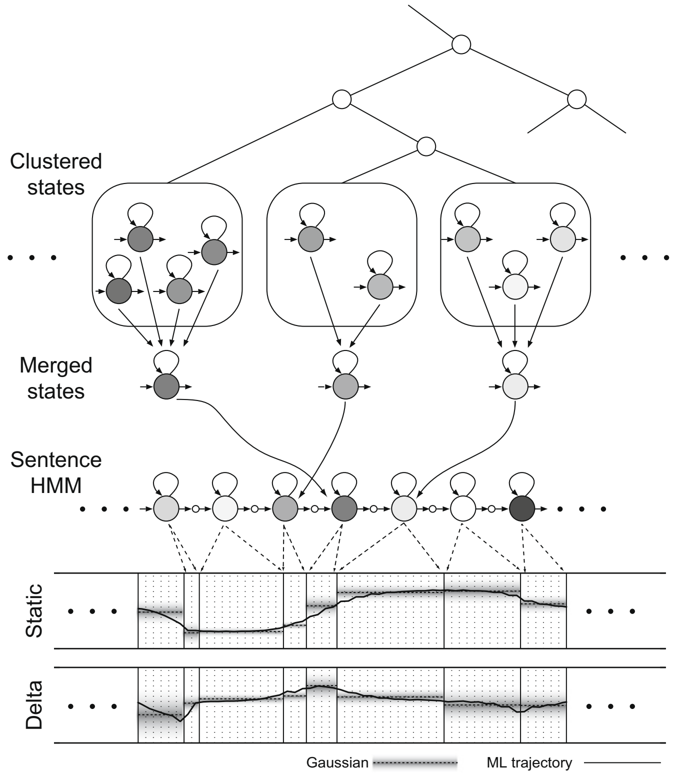


Figure 3.9: *Overview of HMM-based speech synthesis process based on decision tree clustering. Figure reproduced from (Zen et al., 2009).*

Vocoder

Various types of source-filter vocoder are typically used in HMM-based speech synthesis (Hu et al., 2013).

The simplest example of waveform generation from parameters is represented by the Mel-generalized cepstral (MGC) vocoder. A simple pulse/noise excitation is used for this vocoder. Although straightforward, this excitation model cannot fully represent natural excitation signals and often generates "buzzy" speech. Different types of coefficients may be used to represent the spectrum. Mel-cepstra are often used, providing a good approximation to the human auditory perception scale.

The Mel-generalised log spectral approximation (MGLSA) digital filter is also commonly used to filter the excitation signal to synthesise speech.

A more sophisticated method of analysis and re-synthesis of speech is represented by STRAIGHT (Kawahara et al., 1999). This method is successful in removing the periodicity effects of fundamental frequency (F0) on the vocal tract spectral shape. For spectral envelope extraction, both F0 adaptive spectral smoothing and compensatory time windows are used to transfer the time frequency-smoothing problem to the frequency domain. Aperiodicity of the signal is computed as the difference between the upper and lower envelope of the spectrum. For voiced frame, noise is calculated by modulating the randomness of the phase component according to aperiodicity. Finally, all parameters are sent to a minimum-phase filter with group delay phase manipulation to synthesise speech. Although STRAIGHT uses both aperiodicity and F0 adaptive spectral smoothing to solve the "buzzy" voice problem, the number of parameters required for both the spectrum and aperiodicity components is unsuitable for statistical modelling as it is the same size as the fast Fourier transform (FFT) length used. (Zen and Toda, 2005) proposed instead to use other lower dimensional parameters such as MGC or line spectral pairs (LSP) coefficients to represent the spectrum. In standard HMM-based SPSS, MGC is chosen for spectral parametrisation. Here, aperiodicity parameters are compressed by averaging the whole spectrum into sub-bands (e.g., 5 or 25 equally distributed sub-bands).

An alternative vocoder is WORLD (Morise et al., 2016), which allows generation of waveforms for real-time applications. Its sound analysis and manipulation are less accurate than STRAIGHT, but it is adopted very often due to its computational efficiency.

Other available vocoders that have been tested on HMM-based SPSS (Hu et al., 2013) are: Harmonic plus noise model (HNM) vocoder based on Mel-frequency cepstral coefficients (MFCC) and F0 (Erro et al., 2011); adaptive harmonic vocoder (Degottex and Stylianou, 2012); and harmonic vocoder with fixed parameters (Stylianou, 1996). For the source-filter vocoders, the deterministic plus stochastic model for residual (DSMR) vocoder (Drugman and Dutoit, 2012).

More recently, deep neural network (DNN)-based autoencoders such as WaveNet (van den Oord et al., 2016), allow substitution of the SPSS vocoder and the acoustic model at the end of the TTS pipeline. It still requires linguistic analysis and sample-level feature generation. This method can generate a high-quality voice, and it is controllable with input conditioning features, but it shows similar unpredictability as end-to-end systems due to its autoregressive nature. Moreover, it operates at speech sample level, which requires a great computational load.

Adaptation

When a speech synthesis system is trained, it is generally speaker-dependent, and exhibits a very limited speech style. These limitations derive from training the voice models on single-speaker data. Hence, voice timbre and prosody are learned for the average characteristics of the recorded data.

Different adaptive approaches have been developed to create new voice identities (Yamagishi and Kobayashi, 2007; Yamagishi et al., 2009), or unseen speech styles, including enhanced emotions (Schröder, 2001; Nicolao et al., 2006; Tesser et al., 2010).

A target speaker's voice can be created by adapting an average voice model, trained on multiple-speaker data. Speech style adaptation normally operates on single-speaker models and aims to change the expressiveness without affecting the speaker identity. All adaptation techniques normally require training with additional small amounts of target speech.

Several speaker adaptation techniques that were initially developed for HMM-based automatic speech recognition, can also be applied to synthesis. Traditional approaches, such as vocal tract linear normalisation (VTLN) (Saheer et al., 2009), maximum likelihood linear regression (MLLR) (Leggetter and Woodland, 1995), and maximum a-posteriori (MAP) (Gauvain and Lee, 1992; Lee and Gauvain, 1993) can be applied to the statistical models (or features) that generate the acoustic realisations.

Transforms can operate on $\hat{\boldsymbol{o}}$, or on $\hat{\Lambda}$ directly, and these aim to adapt the Gaussian mixture model (GMM) parameters to move the source acoustic space into the target one. The most common type of transform for HMM-based speech synthesisers, is MLLR that allows spectral, excitation, and duration parameters to be adapted.

3.2.2 DNN-based speech synthesis

As mentioned in the previous section, the clustered context-dependent acoustic models can be interpreted as large regression or decision trees that map linguistic features into statistics of acoustic features. Zen et al. proposed an alternative scheme that is based on a deep architecture (Zen et al., 2013), where the regression tree is replaced by a multi-layer artificial neural network.

Several differences can be observed between neural networks and traditional decision trees. While neural networks can compactly represent any relation learned from data (Bengio, 2009), decision trees cannot efficiently express extremely complex relations of input features. The partition of the input space operated by decision trees produces a poor generalization due to the reduction in local region data observations. Neural networks provide better generalization as weights are derived from all available training data (Hinton et al., 1984). Neural networks are therefore said ot provide a more analogous representation of the layered hierarchical structures of human speech production system.

The same adaptation techniques as in HMM-based synthesis can be applied to the output generative models that are produced by the DNN-based regression. More advanced techniques use conditioning vectors, such as *i-vector* (Wu et al., 2015), or *d-vector* (Doddipatla et al., 2017). These methods require an auxiliary conditioning vector to be added to the DNN input to obtain the appropriate adaptation. DNN synthesisers are trained with multi-speaker data, using a one-hot identity or style vector to condition the training.

Compared to an HMM-based speech synthesiser, the DNN-based synthesiser has some peculiar differences (Zen, 2015).

As previously mentioned, the mapping from a linguistic feature vector to an acoustic feature vector is provided by a network, rather than a decision tree. So, dynamic features are only used at the synthesis stages. The training of the DNN-based synthesiser is efficient as the phoneme- or state-level alignments are fixed during the process. The latency is $\mathcal{O}(T)$, which is similar to the HMM-based approach. The synthesis of an entire utterance is computationally much more expensive than the HMM. Visiting a decision tree is much faster than propagating though the DNN structure. The DNN-based synthesis seems to have a better degree of naturalness than the normal HMM (Zen et al., 2013). Finally, the weights in the DNN structures are of more difficult interpretation than the HMM decision tree coefficients.

In conclusion, DNN-based synthesis can generate better quality speech in comparison to HMM-based synthesis, but adaptation techniques, computational load, and structure interpretability are still more advantageous in the traditional HMM-based approach.

3.2.3 Software tools

The HMM/DNN-based speech synthesis framework (HTS) (HTS working group, 2012) is a statistical framework to train parametric synthetic voices and to use them to generate speech parameter sequences. This system has been used for decades as it is one of the most robust, flexible, and well-documented methods for synthesising speech. HTS has been mainly developed by the HTS group. The training part of HTS is implemented as a modified version of the HTK, which is a portable toolkit for building and manipulating HMMs. HTK is primarily used for speech recognition research, but has also been used for numerous other statistical modelling applications.

Merlin (Wu et al., 2016b, a) is a toolkit for building DNN models for SPSS. Merlin is written in Python and uses the Theano library for numerical computation. Detailed recipes are provided to build state-of-the art synthesis systems.

The HTS and Merlin tools must be used in combination with a *front-end text processor* such as Festival or MaryTTS as described below (cf. § 5.1.1), and a *vocoder* such as STRAIGHT or WORLD as already described (cf. § 3.2.1).

3.2.4 Standard HTS generation algorithm

The standard generation algorithm in the HTS software framework represents one of the most efficient implementations of the method for solving the parametric synthesis problem formulated in § 3.2.1.

As expressed in eq. 3.9, the generation process consists of finding the feature sequence $\hat{\boldsymbol{o}}$ that maximises $p[\boldsymbol{o}|\Lambda]$. Conditioning upon the total number of frames (i.e. the duration of the utterance) T in addition to HMM model $\hat{\Lambda}$ of eq. 3.2, the optimal phone label sequence $\hat{\boldsymbol{l}}$, is given by

$$\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{o}} p(\boldsymbol{o}|\hat{\boldsymbol{l}}, \hat{\Lambda}, T) \tag{3.15}$$

where $\hat{\boldsymbol{o}}$ and $\boldsymbol{o}$ are the $PT \times 1$ observation vector sequences and P is the dimension of each vector. T, which is the overall state duration derived from a dedicated statistical model, can be used to stretch the final spoken utterance to have the desired duration. Hence, $\boldsymbol{o}$ can be expanded into the sequence of vectors,

$$\boldsymbol{o} = \begin{bmatrix} \boldsymbol{o}_1 \\ \boldsymbol{o}_2 \\ \vdots \\ \boldsymbol{o}_t \\ \vdots \\ \boldsymbol{o}_T \end{bmatrix} \tag{3.16}$$

Along with the $M \times 1$ vector of static features, $\boldsymbol{c}_t = \{c_t(1), c_t(2), \ldots, c_t(M)\}^\top$, the first n-order derivatives of the features are also considered. These are dynamically computed using:

$$\Delta^{(n)}\boldsymbol{c}_t = \sum_{i=-L^{(n)}}^{L^{(n)}} \boldsymbol{\omega}^{(n)}(i)\boldsymbol{c}_{t+i} \tag{3.17}$$

In the standard HTS settings, $n = 2$. This is, the dimension of the vectors becomes $3MT \times 1$ and a single observation vector $\boldsymbol{o}_t$ can be written as:

$$\boldsymbol{o}_t = \begin{bmatrix} \boldsymbol{c}_t^\top \\ \Delta \boldsymbol{c}_t^\top \\ \Delta^2 \boldsymbol{c}_t^\top \end{bmatrix} = \begin{bmatrix} c_t(1) \\ c_t(2) \\ \vdots \\ c_t(M) \\ \Delta c_t(1) \\ \Delta c_t(2) \\ \vdots \\ \Delta c_t(M) \\ \Delta^2 c_t(1) \\ \Delta^2 c_t(2) \\ \vdots \\ \Delta^2 c_t(M) \end{bmatrix} \tag{3.18}$$

For the purpose of this discussion, it is assumed that the output feature vector is zero outside the utterance duration, e.g., $\boldsymbol{c}_t = \mathbf{0}_M$ for $t < 1$ and $t > T$. Further, **c** and $\Delta^{(n)}\boldsymbol{c}$ are assumed to be statistically independent.

The maximisation part of eq. 3.15 can be rewritten as

$$\max_{\boldsymbol{O}} P[\boldsymbol{o}|\hat{\Lambda}, T, \hat{\boldsymbol{l}}] = \max_{\boldsymbol{C}} P[\boldsymbol{o}|\hat{\Lambda}, T, \hat{\boldsymbol{l}}] = \max_{\boldsymbol{C}} \sum_{\text{all } \boldsymbol{q}} P[\boldsymbol{o}, \boldsymbol{q}|\hat{\boldsymbol{l}}, \hat{\Lambda}, T] \tag{3.19}$$

where **c** is the static feature sequence and **q** is the state sequence which can be:

$$\boldsymbol{q} = \begin{cases} \{q_1, q_2, \cdots, q_T\} & \text{for HMM with single pdf} \\ \{(q_1, i_1), (q_2, i_2), \cdots, (q_T, i_T)\} & \text{for HMM with pdf mixture} \end{cases} \tag{3.20}$$

Eq. 3.19 can be approximated as

$$\max_{\boldsymbol{O}} P[\boldsymbol{o}|\hat{\Lambda}, T, \hat{\boldsymbol{l}}] = \max_{\boldsymbol{C}} \sum_{\text{all } \boldsymbol{q}} P[\boldsymbol{o}, \boldsymbol{q}|\hat{\boldsymbol{l}}, \hat{\Lambda}, T] \approx \max_{\boldsymbol{C}} \max_{\boldsymbol{q}} P[\boldsymbol{o}, \boldsymbol{q}|\hat{\boldsymbol{l}}, \hat{\Lambda}, T] \tag{3.21}$$

The optimisation of eq. 3.21 should be done both for $\boldsymbol{q}$ and $\boldsymbol{o}$ simultaneously, however, this is impractical, since there are too many combinations of states $\boldsymbol{q}$ and mixtures $\boldsymbol{i}$ (Tokuda et al., 2000). Thus, some further simplifications must be used to reduce the complexity of the problem.

First, only the case of a single-mixture pdf HMM is considered.

Second, $P[\boldsymbol{o}, \boldsymbol{q}|\hat{\boldsymbol{l}}, \hat{\Lambda}, T]$ can be further simplified. According to the definition of conditional probability, it can be rewritten as

$$P[\boldsymbol{o}, \boldsymbol{q}|\hat{\Lambda}, T, \hat{\boldsymbol{l}}] = P[\boldsymbol{q}|\hat{\Lambda}, T, \hat{\boldsymbol{l}}]P[\boldsymbol{o}|\boldsymbol{q}, \hat{\Lambda}, T, \hat{\boldsymbol{l}}] = P[\boldsymbol{q}|\hat{\Lambda}, T, \hat{\boldsymbol{l}}]P[\boldsymbol{o}|\boldsymbol{q}, \hat{\Lambda}] \tag{3.22}$$

Now, the problem is separated into two parts: a) the maximisation of the *state sequence* probability, $P[\boldsymbol{q}|\hat{\Lambda}, T, \hat{\boldsymbol{l}}]$, given the text $\hat{\boldsymbol{l}}$ and the required overall duration T, and b) the maximisation of *acoustic parameter sequence* probability, $P[\boldsymbol{o}|\boldsymbol{q}, \hat{\Lambda}]$. The second maximization no longer depends on T, as it is only conditioned on the optimal state sequence $\boldsymbol{q}$. T is used to stretch the overall sentence duration, which impacts the total number of states in the utterance.

Consequently, the optimisation of eq. 3.21 can be computed as the maximisation of the two elements of eq. 3.22 separately:

$$\max_{\boldsymbol{o}} P[\boldsymbol{o}|\hat{\Lambda}, T, \hat{\boldsymbol{l}}] \approx \max_{\boldsymbol{c}} \max_{\boldsymbol{q}} P[\boldsymbol{o}, \boldsymbol{q}|\hat{\boldsymbol{l}}, \hat{\Lambda}, T] \approx \max_{\boldsymbol{q}} P[\boldsymbol{q}|\hat{\Lambda}, T, \hat{\boldsymbol{l}}] \max_{\boldsymbol{c}} P[\boldsymbol{o}|\boldsymbol{q}, \hat{\Lambda}] \tag{3.23}$$

In the standard HTS approach, the optimal feature vector sequence $\boldsymbol{c}$ is estimated after the computation of the most probable state sequence $\boldsymbol{q}$, which itself contains all the linguistic information about phone sequence and time duration.

Determining the state sequence q

Assuming that the HMM model Λ is a left-to-right model with no skip, then the probability of the state sequence $\boldsymbol{q}$ is characterised only by explicit state duration distributions. Once the phone sequence is known – from the linguistic analyser – and the duration of each state has been estimated, the sequence of possible states can be determined.

The computation of the state duration is done by independently training a statistical model. If the following expression for the logarithm of $P[\boldsymbol{q}|\hat{\boldsymbol{l}}, \hat{\Lambda}, T]$ is used:

$$\log P[\boldsymbol{q}|\hat{\boldsymbol{l}}, \hat{\Lambda}, T] = \sum_{k=1}^{K} \log p_{q_k}(d_{q_k}) \tag{3.24}$$

where the probability of d consecutive observation vectors for state q_k are described by a single Gaussian pdf

$$p_{q_k}(d) = \frac{1}{\sqrt{2\pi\sigma_{q_k}^2}} e^{-\frac{(d-m_{q_k})^2}{2\sigma_{q_k}^2}} \tag{3.25}$$

under the constraint of

$$\sum_{k=1}^{K} d_{q_k} = T \tag{3.26}$$

that the duration of each state in the optimal stare sequence, $\hat{\boldsymbol{q}}$, which maximises $\log P[\boldsymbol{q}|\hat{\boldsymbol{l}}, \hat{\Lambda}, T]$ under the above constraint can be obtained by using the Lagrange multipliers method (Yoshimura et al., 1998):

$$d_{q_k} = m_{q_k} + \rho \cdot \sigma^2_{q_k} \tag{3.27}$$

$$\rho = \frac{T - \sum_{k=1}^{K} m_{q_k}}{\sum_{k=1}^{K} \sigma^2_{q_k}} \tag{3.28}$$

where m_k and σ_k are the mean and variance of the duration distribution of state q_k. The parameter ρ can be used to control the *speaking rate*, in addition to the total frame length T. When ρ is set to zero, speaking rate becomes equal to the mean value. When ρ is set to a positive or negative value, speaking rate becomes faster or slower.

Equations for a single HMM Gaussian mixture

The probability of $\boldsymbol{o}$, eq. 3.16 given state sequence $\hat{\boldsymbol{q}} = \{q_1, q_2, \cdots, q_T\}$, with a single HMM Gaussian pdf and the derivative order $n = 2$, is described by:

$$P[\boldsymbol{o}|\hat{\boldsymbol{q}}, \hat{\Lambda}] = b_{q_1}(\boldsymbol{o}_1) b_{q_2}(\boldsymbol{o}_2) \cdots b_{q_T}(\boldsymbol{o}_T) \tag{3.29}$$

where the function $\{b_{q_t}(\boldsymbol{o}_t)\}$ is the product of the static and dynamic feature pdfs,

$$b_j(\boldsymbol{o}_t) = \mathcal{N}(\boldsymbol{c}_t; \mu_j, \Sigma_j) \cdot \mathcal{N}(\Delta \boldsymbol{c}_t; \Delta\mu_j, \Delta\Sigma_j) \cdot \mathcal{N}(\Delta^2 \boldsymbol{c}_t; \Delta^2\mu_j, \Delta^2\Sigma_j) \tag{3.30}$$

and $\mathcal{N}(\boldsymbol{x}; \mu_j, \Sigma_j)$, for the generic vector $\boldsymbol{x}$ at state j, is the Gaussian function:

$$\mathcal{N}(\boldsymbol{x}; \mu_j, \Sigma_j) = \frac{1}{\sqrt{(2\pi)^{MT} |\boldsymbol{\Sigma}_j|}} \cdot e^{(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)^\top \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_j))} \tag{3.31}$$

Computing the logarithm of $P[\boldsymbol{o}|\hat{\boldsymbol{q}}, \hat{\Lambda}]$ the following equation is obtained:

$$\log P[\boldsymbol{o}|\hat{\boldsymbol{q}}, \hat{\Lambda}] =$$

$$\frac{1}{2} \sum_{t=1}^{T} \log |\boldsymbol{\Sigma}_{q_t}| \quad -\tfrac{1}{2}(\boldsymbol{W}\boldsymbol{c} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{W}\boldsymbol{c} - \boldsymbol{\mu}) - \tfrac{3MT}{2} \log(2\pi) \tag{3.32}$$

where the vector $\boldsymbol{\mu}$ is the $3MT \times 1$ sequence of all the mean vectors,

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_{q_1} \\ \boldsymbol{\mu}_{q_2} \\ \vdots \\ \boldsymbol{\mu}_{q_t} \\ \vdots \\ \boldsymbol{\mu}_{q_T} \end{bmatrix} \tag{3.33}$$

$\boldsymbol{\Sigma}$ is the diagonal $3MT \times 3MT$ covariance matrix,

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{q_1} & \mathbf{0}_{3M\times3M} & \cdots & \mathbf{0}_{3M\times3M} & \cdots & \mathbf{0}_{3M\times3M} \\ \mathbf{0}_{3M\times3M} & \boldsymbol{\Sigma}_{q_2} & \cdots & \mathbf{0}_{3M\times3M} & \cdots & \mathbf{0}_{3M\times3M} \\ \vdots & \vdots & & \vdots & & \vdots \\ \mathbf{0}_{3M\times3M} & \mathbf{0}_{3M\times3M} & & \boldsymbol{\Sigma}_{q_t} & & \mathbf{0}_{3M\times3M} \\ \vdots & \vdots & & \vdots & & \vdots \\ \mathbf{0}_{3M\times3M} & \mathbf{0}_{3M\times3M} & \cdots & \mathbf{0}_{3M\times3M} & \cdots & \boldsymbol{\Sigma}_{q_T} \end{bmatrix} \tag{3.34}$$

and $\boldsymbol{o}$ is expressed as $\boldsymbol{W}\boldsymbol{c}$, where $\boldsymbol{W}$ is the $3MT \times MT$ matrix

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{w}_1^\top \\ \boldsymbol{w}_2^\top \\ \vdots \\ \boldsymbol{w}_t^\top \\ \vdots \\ \boldsymbol{w}_T^\top \end{bmatrix} \tag{3.35}$$

In more detail, $\boldsymbol{w}_t$ is a $3M \times MT$ matrix defined as

$$\boldsymbol{w}_t = \left[\ \boldsymbol{w}_t^{(0)},\ \ \boldsymbol{w}_t^{(1)},\ \ \boldsymbol{w}_t^{(2)}\ \right] = \tag{3.36}$$

$$= \left[\begin{array}{c|c|c}
\boldsymbol{o}_{M\times M} & \boldsymbol{o}_{M\times M} & \boldsymbol{o}_{M\times M} \\
\boldsymbol{o}_{M\times M} & \boldsymbol{o}_{M\times M} & \boldsymbol{o}_{M\times M} \\
\vdots & \vdots & \vdots \\
w^{(0)}(-L^{(0)}) \cdot \boldsymbol{I}_{M\times M} & w^{(1)}(-L^{(1)}) \cdot \boldsymbol{I}_{M\times M} & w^{(2)}(-L^{(2)}) \cdot \boldsymbol{I}_{M\times M} \\
\vdots & \vdots & \vdots \\
w^{(0)}(0) \cdot \boldsymbol{I}_{M\times M} & w^{(1)}(0) \cdot \boldsymbol{I}_{M\times M} & w^{(2)}(0) \cdot \boldsymbol{I}_{M\times M} \\
\vdots & \vdots & \vdots \\
w^{(0)}(+L^{(0)}) \cdot \boldsymbol{I}_{M\times M} & w^{(1)}(+L^{(1)}) \cdot \boldsymbol{I}_{M\times M} & w^{(2)}(+L^{(2)}) \cdot \boldsymbol{I}_{M\times M} \\
\vdots & \vdots & \vdots \\
\boldsymbol{o}_{M\times M} & \boldsymbol{o}_{M\times M} & \boldsymbol{o}_{M\times M}
\end{array}\right]
\begin{array}{c}
\text{1st} \\
\text{2nd} \\
\vdots \\
\text{(t-L)-th} \\
\vdots \\
\text{t-th} \\
\vdots \\
\text{(t+L)-th} \\
\vdots \\
\text{T-th}
\end{array}$$

where $w^{(n)}(i)$, with i in $[-L^{(n)} \dots L^{(n)}]$, are the window coefficients used to compute the n-th dynamic features from the static ones and $L^{(n)}$ is the relative length.

The scalar value ε is defined as follow,

$$\varepsilon(\boldsymbol{c}) = (\boldsymbol{W}\boldsymbol{c} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{W}\boldsymbol{c} - \boldsymbol{\mu}) \tag{3.37}$$

Maximising the function in eq. 3.32 requires finding the solution to the following equation system:

$$\frac{\partial \log P[\boldsymbol{o}|\hat{\boldsymbol{q}}, \hat{\Lambda}]}{\partial \boldsymbol{c}} = \mathbf{0}_{TM} \tag{3.38}$$

Since only ε, eq. 3.37, in eq. 3.32 depends on the observation, the sequence $\boldsymbol{c}$ that maximises eq. 3.38 is equivalent to the one that maximise eq. 3.37 and hence

$$\begin{gathered}
\frac{\partial \varepsilon(\boldsymbol{c})}{\partial \boldsymbol{c}} = \frac{\partial \left((\boldsymbol{W}\boldsymbol{c} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{W}\boldsymbol{c} - \boldsymbol{\mu})\right)}{\partial \boldsymbol{c}} = \mathbf{0}_{TM} \\
\frac{\partial \left(\boldsymbol{c}^\top \boldsymbol{W}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{W}\boldsymbol{c} - \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{W}\boldsymbol{c} - \boldsymbol{c}^\top \boldsymbol{W}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} - \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}\right)}{\partial \boldsymbol{c}} = \mathbf{0}_{TM}
\end{gathered} \tag{3.39}$$

and this happens when

$$\boldsymbol{W}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{W}\boldsymbol{c} - \boldsymbol{W}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} = \mathbf{0}_{TM} \tag{3.40}$$

from which comes

$$\boldsymbol{W}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{W}\boldsymbol{c} = \boldsymbol{W}^\top \boldsymbol{\Sigma}^{\mathbf{-1}} \boldsymbol{\mu} \tag{3.41}$$

$$\boldsymbol{R}\boldsymbol{c} = \boldsymbol{r} \tag{3.42}$$

where

$$\boldsymbol{R} = \boldsymbol{W}^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{W} \tag{3.43}$$

$$\boldsymbol{r} = \boldsymbol{W}^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \tag{3.44}$$

This set of linear equations can be solved efficiently with the Cholesky decomposition with $\mathcal{O}(T)$ operations.

HTS software implementation

In HTS, three modes are implemented to find the best sequences of states and observations formalised by eq. 3.19, (Tokuda et al., 2000).

mode 1 maximizing $P(\boldsymbol{o}|\hat{\boldsymbol{q}}, \hat{\Lambda})$ with respect to $\boldsymbol{o}$. The state sequence is a conditional observation. Pitch and spectral trajectories are computed by solving eq. 3.42 with the Cholesky matrix inversion. This method is fast and the whole utterance is generated at once.

mode 2 maximizing $P(\boldsymbol{o}, \boldsymbol{q}|\hat{\Lambda})$ with respect to $\boldsymbol{o}$ and $\boldsymbol{q}$. The best state sequence is determined by the duration models with eq. 3.25. If state output probabilities are assumed to be single-Gaussian, the solution is obtained by solving eq. 3.42 in the same way as mode 1.

mode 3 maximizing $P(\boldsymbol{o}|\hat{\Lambda})$ with respect to $\boldsymbol{o}$. This mode uses an algorithm based on expectation maximisation (EM), which finds a critical point of the likelihood function $P(\boldsymbol{o}|\hat{\Lambda})$. State sequences and mixture indices are both considered unobservable and are determined iteratively. Finally, the spectral characteristics are determined with the Cholesky decomposition.

Despite the different strategies available in the HTS software, these methods all eventually converge on using the Cholesky decomposition to solve the equation system in eq. 3.42. However, every mode implemented in the latest versions of HTS has some limitations when it comes to including it in a reactive framework which needs to modify its parameters during the generation process. In particular, one of these methods permits a fast modification in the generated features and nor a constant adjustment of the generative models. In HTS, speech features are generated utterance-by-utterance. Thus, modifications in the HMM statistical descriptions – e.g., adaptation – can only be applied at the beginning of the generation process.

This highlights the emerging need for a more flexible generation algorithm that allows frame-by-frame modifications of the generative models.

3.2.5 Recursive search generation algorithm

A recursive algorithm was proposed in the initial implementation of SPSS: the *recursive search generation algorithm*. This method yields the same performance as the latest Cholesky-based method, in term of quality of the generated feature sequence, but its computational complexity is higher, which results in a hundred-fold increase in generation time.

The recursive search algorithm was first introduced by Tokuda for a single GMM (Tokuda et al., 1995a), and later generalised to multiple GMMs (Tokuda et al., 1995b). This algorithm permits generative models to be selected in almost real time, allowing them to be manipulated with different transforms at each step of the synthesis process.

The recursive search generation algorithm was originally developed to compute the best acoustic feature sequence by comparing different sub-optimal state sequences derived from the concatenation of GMM choices, but some simplification can reduce the complexity of its implementation. As in the HTS standard generation algorithm described previously in § 3.2.4, it is assumed that the state sequence is already given (cf. eq. 3.25), and that the HMM has a single Gaussian per state.

First, the general algorithm for the optimal sub-sequence choice is described. Assuming that an initial feature sequence $\boldsymbol{c}$ is given, and that the state q_t is updated to $\hat{q}_t$, the system in eq. 3.42 can be written as

$$\hat{\boldsymbol{R}}\hat{\boldsymbol{c}} = \hat{\boldsymbol{r}} \tag{3.45}$$

in which the variable evolution is described by

$$\hat{\boldsymbol{R}} = \boldsymbol{R} + \boldsymbol{w}_t \boldsymbol{D} \boldsymbol{w}_t^\top \tag{3.46}$$

$$\hat{\boldsymbol{r}} = \boldsymbol{r} + \boldsymbol{w}_t \boldsymbol{d} \tag{3.47}$$

$$\boldsymbol{D} = \boldsymbol{\Sigma}_{\hat{q}_t}^{-1} - \boldsymbol{\Sigma}_{q_t}^{-1} \tag{3.48}$$

$$\boldsymbol{d} = \boldsymbol{\Sigma}_{\hat{q}_t}^{-1} \boldsymbol{\mu}_{\hat{q}_t} - \boldsymbol{\Sigma}_{q_t}^{-1} \boldsymbol{\mu}_{q_t} \tag{3.49}$$

These updating functions are similar to those of an adaptive filter such as the recursive least squares filter (Haykin, 2014). A recursive algorithm to obtain $\hat{\boldsymbol{c}}$ from $\boldsymbol{c}$ can be derived. The principal steps are listed in Table 3.1, with the substitution $\boldsymbol{P} = \boldsymbol{R}^{-1}$.

The sequence of steps needed to apply the algorithm of Table 3.1 are described in the following list:

1. determine an initial state sequence $\boldsymbol{q}$,

Table 3.1: *Algorithm to replace the sub-state q_t of a frame t with $\hat{q}_t$*

Substitute $\hat{\boldsymbol{c}}$, $\hat{\boldsymbol{P}}$ and $\hat{\varepsilon}$ obtained by the previous iteration to $\boldsymbol{c}$, $\boldsymbol{P}$ and ε respectively, and calculate:

$$\boldsymbol{\pi} = \boldsymbol{P}\boldsymbol{w}_t \tag{T.1}$$

$$\boldsymbol{\nu} = \boldsymbol{w}_t^\top \pi \tag{T.2}$$

$$\boldsymbol{\kappa} = \boldsymbol{\pi}\left\{\boldsymbol{I}_{3M} + \boldsymbol{D}\boldsymbol{\nu}\right\}^{-1} = \boldsymbol{\pi}\left\{\boldsymbol{I}_{3M} + \left(\boldsymbol{\Sigma}_{\hat{q}_t}^{-1} - \boldsymbol{\Sigma}_{q_t}^{-1}\right)\boldsymbol{\nu}\right\}^{-1} \tag{T.3}$$

$$\hat{\boldsymbol{c}} = \boldsymbol{c} + \boldsymbol{\kappa}\left\{\boldsymbol{\Sigma}_{\hat{q}_t}^{-1}(\boldsymbol{\mu}_{\hat{q}_t} - \boldsymbol{w}_t^\top \boldsymbol{c}) - \boldsymbol{\Sigma}_{q_t}^{-1}(\boldsymbol{\mu}_{q_t} - \boldsymbol{w}_t^\top \boldsymbol{c})\right\} \tag{T.4}$$

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{\varepsilon} + (\boldsymbol{\mu}_{\hat{q}_t} - \boldsymbol{w}_t^\top \hat{\boldsymbol{c}})^\top \boldsymbol{\Sigma}_{\hat{q}_t}^{-1}(\boldsymbol{\mu}_{\hat{q}_t} - \boldsymbol{w}_t^\top \boldsymbol{c}) - (\boldsymbol{\mu}_{q_t} - \boldsymbol{w}_t^\top \hat{\boldsymbol{c}})^\top \boldsymbol{\Sigma}_{q_t}^{-1}(\boldsymbol{\mu}_{q_t} - \boldsymbol{w}_t^\top \boldsymbol{c}) \tag{T.5}$$

$$\hat{\boldsymbol{P}} = \boldsymbol{P} - \boldsymbol{\kappa}\boldsymbol{D}\boldsymbol{\pi} = \boldsymbol{P} - \boldsymbol{\kappa}\left(\boldsymbol{\Sigma}_{\hat{q}_t}^{-1} - \boldsymbol{\Sigma}_{q_t}^{-1}\right)\boldsymbol{\pi} \tag{T.6}$$

2. given the initial state sequence, obtain a sequence **c**, **P** and ε,
3. for each frame $t = 1, 2, \ldots, T$:
    (a) calculate (T.1) and (T.2),
    (b) for each possible state of the frame t, calculate (T.3)-(T.5) and obtain $\log P[\boldsymbol{o}, \boldsymbol{q}|\hat{\boldsymbol{l}}, \hat{\Lambda}, T]$ using eq. 3.32,
    (c) choose the best state in the sense that $\log P[\boldsymbol{o}, \boldsymbol{q}|\hat{\boldsymbol{l}}, \hat{\Lambda}, T]$ is most increased by the state replacement,
4. choose the best frame in the sense that $\log P[\boldsymbol{o}, \boldsymbol{q}|\hat{\boldsymbol{l}}, \hat{\Lambda}, T]$ is most increased by the state replacement.
5. if $\log P[\boldsymbol{o}, \boldsymbol{q}|\hat{\boldsymbol{l}}, \hat{\Lambda}, T]$ cannot be increased by the state replacement at the best frame, stop iterating.
6. replace the state of the best frame by calculating (T.1)-(T.6) and obtain $\hat{\boldsymbol{c}}$, $\hat{\boldsymbol{P}}$ and $\hat{\varepsilon}$.
7. go to 2.

The most critical step is the choice of the initial $\boldsymbol{c}$, $\boldsymbol{P}$ and $\boldsymbol{\varepsilon}$ elements. These vectors can be chosen by assuming the existence of initial states $\{\bar{q}_t\}$ with parameters related to the static features only:

$$\bar{\boldsymbol{\mu}}_{\bar{q}_t} = \begin{bmatrix} \boldsymbol{\mu}_{q_t}^{(0)} \\ \mathbf{0}_{M\times 1} \\ \mathbf{0}_{M\times 1} \end{bmatrix} \tag{3.50}$$

and

$$\bar{\boldsymbol{\Sigma}}_{\bar{q}_t}^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}_{q_t}^{(0)} & \mathbf{0}_{M\times M} & \mathbf{0}_{M\times M} \\ \mathbf{0}_{M\times M} & \mathbf{0}_{M\times M} & \mathbf{0}_{M\times M} \\ \mathbf{0}_{M\times M} & \mathbf{0}_{M\times M} & \mathbf{0}_{M\times M} \end{bmatrix} \tag{3.51}$$

The $M \times 1$ vector $\boldsymbol{\mu}_{q_t}^{(0)}$ and the $M \times M$ matrix $\boldsymbol{\Sigma}_{q_t}^{(0)}$ are the mean vector of the static features $\boldsymbol{c}_t$, respectively. From eq. 3.43 and given that $\bar{\boldsymbol{\Sigma}}^{-1} = \text{diag}\{\bar{\boldsymbol{\Sigma}}_{\bar{q}_1}^{-1}, \ldots, \bar{\boldsymbol{\Sigma}}_{\bar{q}_t}^{-1}, \ldots, \bar{\boldsymbol{\Sigma}}_{\bar{q}_T}^{-1}\}$, it happens that $\varepsilon(\boldsymbol{c}) = 0$ and the following results are obtained:

$$\begin{aligned} \bar{\boldsymbol{P}} &= \bar{\boldsymbol{R}}^{-1} \\ &= (\boldsymbol{W}^\top \bar{\boldsymbol{\Sigma}}^{-1} \boldsymbol{W})^{-1} \\ &= \begin{bmatrix} \boldsymbol{\Sigma}_{q_1}^{(0)} & \mathbf{0}_{M\times M} & \cdots & \mathbf{0}_{M\times M} & \cdots & \mathbf{0}_{M\times M} \\ \mathbf{0}_{M\times M} & \boldsymbol{\Sigma}_{q_2}^{(0)} & \cdots & \mathbf{0}_{M\times M} & \cdots & \mathbf{0}_{M\times M} \\ \vdots & \vdots & & \vdots & & \vdots \\ \mathbf{0}_{M\times M} & \mathbf{0}_{M\times M} & & \boldsymbol{\Sigma}_{q_t}^{(0)} & & \mathbf{0}_{M\times M} \\ \vdots & \vdots & & \vdots & & \vdots \\ \mathbf{0}_{M\times M} & \mathbf{0}_{M\times M} & \cdots & \mathbf{0}_{M\times M} & \cdots & \boldsymbol{\Sigma}_{q_T}^{(0)} \end{bmatrix} \end{aligned} \tag{3.52}$$

and

$$\begin{aligned} \bar{\boldsymbol{c}} = \bar{\boldsymbol{R}}^{-1}\bar{\boldsymbol{r}} &= (\boldsymbol{W}^\top \bar{\boldsymbol{\Sigma}}^{-1} \boldsymbol{W})^{-1} \boldsymbol{W}^\top \bar{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\mu} \\ &= \begin{bmatrix} \boldsymbol{\Sigma}_{q_1}^{(0)} (\boldsymbol{\Sigma}_{q_1}^{(0)})^{-1} \boldsymbol{\mu}_{q_1}^{(0)} \\ \vdots \\ \boldsymbol{\Sigma}_{q_t}^{(0)} (\boldsymbol{\Sigma}_{q_t}^{(0)})^{-1} \boldsymbol{\mu}_{q_t}^{(0)} \\ \vdots \\ \boldsymbol{\Sigma}_{q_T}^{(0)} (\boldsymbol{\Sigma}_{q_T}^{(0)})^{-1} \boldsymbol{\mu}_{q_T}^{(0)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_{q_1}^{(0)} \\ \vdots \\ \boldsymbol{\mu}_{q_t}^{(0)} \\ \vdots \\ \boldsymbol{\mu}_{q_T}^{(0)} \end{bmatrix} \end{aligned} \tag{3.53}$$

Using eq. 3.52 and 3.53 in the algorithm, and the original definition of $\boldsymbol{\mu}_{q_t}$ and $\boldsymbol{\Sigma}_{q_t}$, for every t in $[1..T]$, the initial sequence of acoustic vectors from which to start the optimisation can be computed. It must be noted that this initialisation gives the optimised sequence $\boldsymbol{c}$ if the optimised $\boldsymbol{Q}$ sequence is given.

Since most of the elements in $\boldsymbol{w}_t$ are zeros, (T.6) has the higher computational complexity, which is $O(T^2M^3)$. When $\boldsymbol{\Sigma}_{q_t}$ is diagonal, it reduces to $O(T^2M)$ and finally, if only the S neighbouring frames ($S \ll T$) are assumed to influence the generation of the feature vectors, then it reduces further to $O(S^2M)$. Often S has an heuristically derived value around 30.

3.3 Artificial speech in adverse conditions

Since the beginning of electronic speech technology, a great deal of effort has been devoted to the enhancement of speech signals. One of the first examples (Blesser, 1969) used dynamic range compression to improve the quality of speech that was distorted by channel transmission. More recent approaches aim to reproduce the variations observed in humans (cf. § 2.1.2), and, in particular, to ultimately reproduce the effects of the Lombard reflex (cf. § 2.1.2). Indeed, the main objective of all these methods is to control the quality of produced speech in different levels and types of adverse conditions. Almost all models of speech production aim to *increase* the speech-in-noise intelligibility to its maximum available level. As pointed out already in § 1.1, however, very little attention has been given to balancing the overall amount of effort (energy), involved in the production.

An overview of some of the latest speech-in-noise techniques, and their performance is reported in (Cooke et al., 2013a). Three main categories can be identified among these approaches: a) 'near-end' enhancement of natural or synthesised speech, b) model-based synthesis with adaptation learned from human data, and c) model-based synthesis with adaptation emerging from human behaviour. The third of these us particularly critical for the work presented in this thesis; all are briefly introduced next.

Near-end speech enhancement This speech modification is the most commonly adopted approach and comprises all those techniques that can be applied to an already-available speech signal. Several degrees of complexity can be used to manipulate the signal at this stage: from a trivial loudness increment based on average signal-to-noise ratio (SNR) optimisation, to the analysis and modified re-synthesis of speech. A speech signal can be modified in order to reallocate its spectral energy in the most critical sub-bands (Tang and Cooke, 2010). The speech spectrum can be reshaped in combination with time stretching and dynamic range compression as in (Zorila et al., 2012). A linear time-invariant filter has also been proposed to redistribute speech energy across frequency in order to maximise the speech intelligibility index (SII) (Taal et al., 2013). Speech can be also parametrised into its fundamental components. Operating in the parametric domain allows efficient signal processing to redistribute spectral energy or modify the speech rate (Godoy et al., 2013; Koutsogiannaki and Stylianou, 2016). This type of modifications is normally very effective when it operates on clear speech recordings. However, when it is used to transform synthetic speech, which are more likely to contain distortions, the results might worsen the artefacts along with enhancing the speech signal.

Enhanced synthesis from human data Another approach that allows speech to be generated dynamically involves applying modification to a TTS system. Here, transforms are learned from observation of *human data* such as speech recordings. Previous section § 3.1.1 described that unit selection synthesisers can produce adequate clear speech, but require new voice inventories, i.e., pre-recorded data sets, for each specific style, statistical parametric synthesis allows more flexible adaptation. Transforms can be learned from 'ad-hoc' recorded audio (Picart et al., 2011, 2014) and datasets can be designed to contain different speech styles, including both hyper- and hypo-articulated speech. The transformation magnitude can be scaled such that interpolated versions can be generated to have different degrees of articulation. Transforms can be applied to different parts of the source-filter synthesis model. Thus, in addition to adapting spectral and phone-duration components, the glottal source signal can also be manipulated (Raitio et al., 2013). Different datasets can be recorded with breathy, normal, and Lombard speech and the specific glottal signal modifications can be learned for each speech style.

Enhanced synthesis from human behaviour The transforms that are applied to speech synthesisers can also be inspired by the observation of *human behaviour*. The idea is to embed computational human production models in the training and synthesis of synthetic speech. SPSS is the most typical method in which this approach can be applied. Speech modifications are the results that emerge from optimisation of the model parameters with respect to the selected modelled behaviour. In (Valentini-Botinhao et al., 2012), for example, an objective measure of speech intelligibility, Glimpse Proportion measure (GP) (Cooke, 2006), is introduced in the HTS objective function for parameter training. Created voices are then optimised to compensate for distortions measured with this index. The limitations of this method are that it only compensates for noises that are introduced during the training stage, and it does not allow for scaling of the degree of the compensation.

The computational model and its implementation, proposed in Chapter 4 and Chapter 5 of this thesis respectively, are developed following the third of these approaches (Moore and Nicolao, 2011; Nicolao et al., 2012, 2013). That is, C2H speech adjustments should emerge in reaction to the environment interferences. In principle, they should not be noise-dependant, and moreover they should enable both enhancement and reduction of speech quality.

Chapter 4

C2H, a Computational Model of H&H behaviour

Contents

This chapter introduces C2H, a computational model that incorporates the principles of Lindblom's H&H theory (cf. § 2.2.1) and Powers' PCT (cf. § 2.2.2). This model mainly represents the answer to the research question n. 2 of this thesis, and constitutes the basis on which context-aware automatic speech synthesis can be enabled.

The computational model is designed to account for the generation of diverse styles of speech (expressive, neutral, colloquial, clear, etc.) in reaction to the dynamic evolution of the communicative conditions. In this chapter, all these speaking styles are reduced to two main classes along the same energy-motivated dimension: high and low-effort (hyper/hypo-articulated) speech.

The following sections present all the principal components of the model are presented. In order to graphically illustrate some of the concepts of the C2H model, a simplified model, trajectory generation simulation model (TGSM), which is detailed in Appendix A, is used.

4.1 Context-aware speech synthesis

Lindblom's H&H theory models the human speech production process as the balance of two driving objectives, as discussed in § 2.2.1. The primary goal is for the talker to achieve their *communicative intent*, i.e., to deliver a meaningful concept to the listener. The second objective regards the amount of energy involved in the process. The talker aims to optimise the completion of the primary goal minimising the amount of energy involved. In order to assess the success of their communication, talkers must be aware of the context, in which speech occurs, and reacts to this with continuous adjustments to its speech production strategies.

As discussed in section 1.1, very few speech synthesisers are *aware* of the environment in which they operate, and can *react* to compensate for its changes. The C2H context-aware computational speech production model aims to emulate the human H&H behaviour (Moore and Nicolao, 2011; Nicolao et al., 2012, 2013; Moore and Nicolao, 2017).

If it is to reproduce the speech production process according to the H&H theory, a computational model must include the capabilities to assess the success of the communication, as well as to measure the amount of energy required in the process. In machines, the communicative effort might no longer be a limited resource to be optimised. However, an unsuitable amount of energy in production often reflects in a higher-than-required effort on the listening side (Mattys et al., 2009), as discussed in § 2.1.2. Hence, in order to sound more human-like, machine-to-human communication must follow the same optimisation principles as human-to-human.

The general diagram of the C2H model is depicted in Figure 4.1. The model serves as a framework for *speech synthesis* and as such, its core component is the speech generation process. The objective of the model is to assess to what degree the

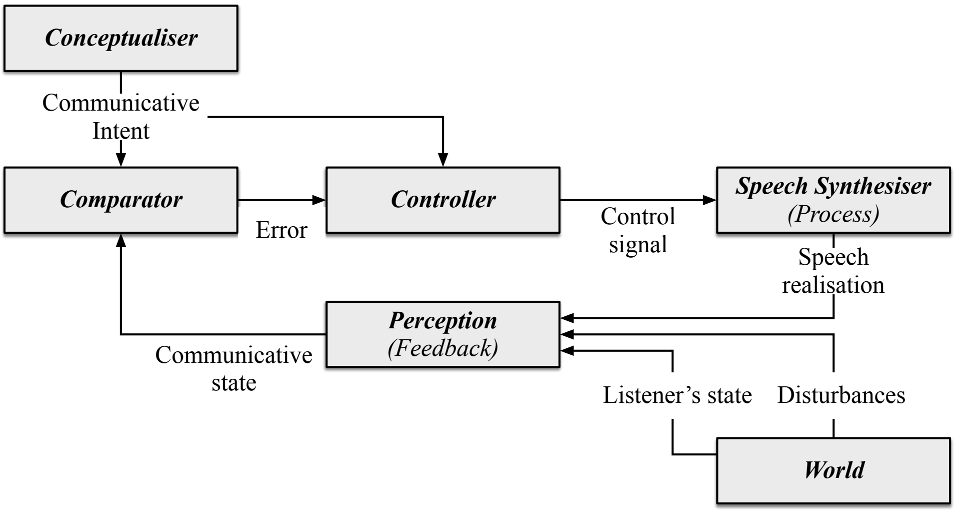


Figure 4.1: *High-level diagram of the C2H components. Square blocks represent the functional processes. Communication signals between blocks are also highlighted.*

communication process outcome matches the pre-defined intent, and it is based on the fundamental components of control theory: *comparator*, *controller*, *process*, and *feedback*, as shown in reactive speech synthesiser diagram of Figure 1.2.

The *communicative intent*, which originates from the *conceptualiser* consists of the motivations and goals that drive the whole communication process. These two inputs play a crucial role in speech generation as they can completely change the objective function to maximise/minimise. The range of possible intents can be very diverse. The goal of an automatic speech synthesis communication is often to transfer a text message to the listener. In order to deliver each single word of the text, the synthesiser objective is to produce extremely *clear* and *intelligible* speech. However, the goal could also be different, such as delivering the gist of the message. In this situation, only the intelligibility of a few meaningful key words of the message needs to be ensured. Therefore, different constituent parts of the utterance realisation can have different degrees of intelligibility. The message to deliver could also be an emotion rather than an informative content. In this case, it is voice expressiveness that needs to be optimised rather than speech clarity. A complete description of the range of possible intents is outside the scope of this thesis. However, it is noteworthy that some communicative goals may lead to reduced speech quality. If the system intent is to confuse the audience or hide some information from them, a muffled unintelligible speech signal might indeed be the optimal output. It is therefore necessary that the C2H model can control speech production in both directions of the hyper/hypo-articulation dimension.

The *comparator* assesses the distance between the communicative intent and the actual *communicative state* which results from the speech realisation. Such comparison is not trivial, and is heavily intent-dependent, potentially including an evaluation of the intended degree of intelligibility, and the number of elements of speech realisation that are successfully recognised by listeners among other viable dimensions. The C2H model assumes that each dimension of the communication that is measured by perception has a correspondent *reference* value from the conceptualiser intent.

The difference between the communicative intent and the communicative state (*error*) is transformed into *control signals* that drive the speech production via the *controller*. The control signals select the type of adjustment that optimise the generation process in order to achieve the goal. The degree of sensitivity of the controller can be varied by the motivation component of the intent. Motivation determines the level of attention that the model pays to the error. If the motivation is low, the controller neglects the error and no goal intent is pursued. That is, no message is delivered. Vice-versa, high motivation generates speech that aims to minimise every error component.

The awareness of the synthesiser is enabled by the *perception* loop (feedback). Linking the process and controller components of the model allows the speech production to be continuously assessed and adjusted even when the generation process is not yet complete. The perception loop is designed to probe the main dimensions of speech production. First, it can assess the level of communicative energy that results from the *speech realisation*. Second, it can observe or predict the effect of the realisation on the listener. Finally, since speech produced in a real environment is affected by *disturbances*, the model is also capable of analysing the communication channel conditions and measuring how these affect the quality of speech produced.

Each of the components of Figure 4.1 is fully described below.

4.2 Process: speech synthesis

The *speech synthesiser* is the core component of the C2H model. It generates the speech realisations on which the proposed model acts to improve the quality of the communication.

In principle, C2H is designed to be applied to any speech synthesis system. The model is sufficiently flexible to allow different TTS architectures, as long as they provide adequate techniques to allow control of their production. Some basic requirements are however necessary. Such speech synthesisers must:

- have *access* to the generation parameters during the entire synthesis process. C2H must be aware of its speech outcome during synthesis, even before the actual waveform is produced.
- allow *continuous* adjustment of the generation parameters. C2H must be able to modify the speech outcome at any step of the generation process.
- implement *robust* scalable modifications. C2H adjustments must be scaled depending on the effort/energy that needs to be applied to the production. Modified synthetic speech must always be consistent with a realistic production (without speech artefacts) and must be similar to the original speech characteristics in terms of same speaker identity, gender, language phone inventory, etc.

In summary, C2H requires a synthesis system which can react to the variation of the communication context at any stage of the speech production and can do so with variable intensity.

An overview of the several types of TTS and their characteristics is reported below in § 5.1. SPSS, emerges as the most suitable method in term of simplicity and flexibility. For this reason, in this chapter, SPSS is implied to be the underlying speech synthesis method of C2H.

4.2.1 Speech production adjustment

The TTS component of a context-aware speech synthesis model, such as C2H, must provide a way to act on its generation process and modify its outcome. The modifications must be controllable by a control signal and scalable in correspondence to the degree of energy required for the successful communication. The TTS must operate on reaction to the control signal originated in the controller (cf. § 4.5), which is dependent on the communicative error detected by the comparator (cf. § 4.4). The synthesis process is formalised as per eq. 4.1.

$$\text{speech} = \mathcal{F}_{\text{synthesis}}(\text{control signal}) \tag{4.1}$$

In order to function in C2H, the TTS adjustment components are required to have some important characteristics. They must be:

- *consistent* with the type of synthesiser. Speech modifications must be in the range that the generation algorithm is designed to achieve. For example, concatenative synthesisers might only allow signal post-processing modifications, whilst SPSS provides a larger set of controllable parameters.

- *effective* at changing the speech quality such that the communicative intent can be achieved. The operational points of the TTS modifications must be located on a dimension which affects the impact of speech on the listener. Magnitude, versus, and direction of the adjustments need to be controllable.
- *continuously* updated. The environmental conditions may vary during the TTS production and the perception feedback loop (cf. § 4.3), returns the communicative state with a relatively low latency. Speech modifications must therefore continuously follow the error signal and react to it.

Clearly, a large family of adjustments that comply with these requirements can be selected at diverse levels of communicative abstraction. Changes can be operated a) at the *conceptualiser* stage, by trying to express the goal in different modalities, b) at the *phonetic planning* level, by changing the driving commands to the synthesiser, or c) at the *signal level*, by keeping the linguistic information fixed and adapting the sound. Most synthesisers can only control their speech signal characteristics (intonation, speaker identity, expressiveness, etc.). Few of them can change their input phone sequence, and fewer are still able to reword the whole message.

Despite the possibility that all these modifications could, in principle, be implemented in C2H, the model currently focuses on the signal level actions alone. The speech transform proposed in this thesis is based on the adaptation capabilities that almost every standard speech synthesiser allows, using changes in speech duration, pitch, and spectral shape.

The idea behind the adjustment process in speech generation is also illustrated in Appendix A.3.3 with the exemplar TGSM.

4.2.2 Effort and degree of articulation in speech production

In human speech production, a speech sound is assumed to be clearly recognisable by a listener when its realisation is similar to an ideal production (i.e., phonemic target). The development of phonemic systems in a language are motivated by acoustic contrast. Speech sounds are categorised with diverse phonemic identities when their realisations are sufficiently separated, acoustically. The distinction of phonetic sounds (*clarity*) is correlated to the complete realisation of the related articulatory gestures. As reported in (Browman and Goldstein, 1992), “gestures can function as primitives of phonological contrast. That is, two lexical items will contrast if they differ in gestural composition”. Hence, if the gesture is not fully realised, it may overlap with other gestures, and the resulting phones produced can exhibit reduced phonological contrast. For clear speech, realisation requires complete articulatory gestures so larger articulator movements are required, and consequently a higher amount of effort is needed from the talker.

This well-established relationship between the effort of the communicative process, the degree of articulation, and the resulting phonetic contrast constitutes the basis of the C2H transformations, as was shown in Figure 4.2. This represents the answer to the research question n. 3 on this thesis (cf. § 1.2).

The most intuitive dimension along which the production energy can be controlled is the speech signal amplitude. The loudness level of a speech signal can be directly linked to the airflow pressure involved in its production. Therefore, the amplitude of a discrete-time signal $x[t]$ can be directly correlated to the amount of energy of the signal, E_x:

$$E_x = \sum_t |x[t]|^2 \tag{4.2}$$

In order to reproduce well-established behaviours (e.g., the Lombard reflex) other speech signal characteristic, such as formants, spectral shape, phone duration, etc., must be affected by the transformation. Hence, a transform that can modify these components is designed. As in the loudness control example, an important assumption in designing the required transform is that a metric can be defined to compute the *distance* between competing phonetic configurations, and the *direction*, along which the speech realisations are transformed.

Rather than focus on single characteristics of speech signal such as loudness, pitch or spectral energy, a more general motivating principle is proposed here whose effects on speech production imitates those that are observed in human speech behaviour.

4.2.3 Adjusting the phonetic contrast

Inspired by the H&H principles espoused by (**?**), adaptation of the C2H output is motivated by both articulatory and energetic manifestations of phonetic contrast. In particular, the notion of *low-energy attractors* is introduced.

Similar to the attractors in the theory of dynamical systems, these are hypothesised to be acoustic configurations toward which speech tends to be evolve if no extra effort is introduced in the communication. In the TGSM example of Appendix A, these configurations appear from the combination of two or more Gaussian mixture zone (GMZ). The interaction among equipotential Gaussian functions, where no outstanding effort (i.e., high weighting factor) is applied to any of these, generates virtual targets often in between the original GMZ targets (cf. Figure A.5).

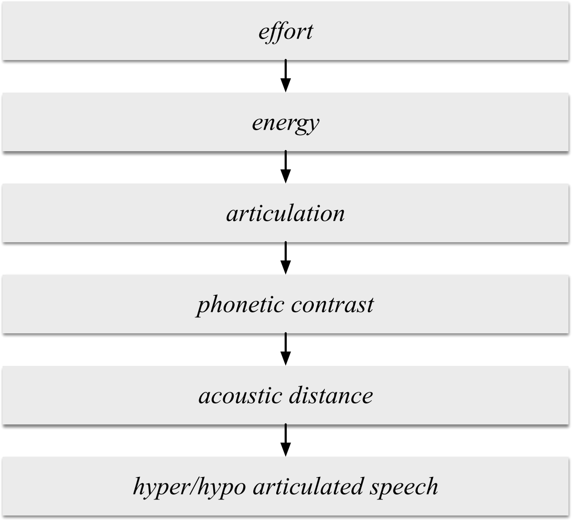


Figure 4.2: *Relationship chain between effort and synthetic speech production in order to create H&H speech.*

In C2H, the low-energy attractors in the acoustic configurations are the *low-contrastive (LC)* acoustic realisations towards which at least two competing phones tend to converge. The opposite *high-contrastive (HC)* configurations are realisations in which the acoustic differences of competing phones are enhanced. For example, in the utterances "*This is my pet*" versus "*This is my pot*", the ease with which a listener can distinguish between "*pet*" and "*pot*" depends on the effort put in to the pronunciation of the vowel by the speaker. A speaker is likely to produce very clear high-effort *hyper*-articulated output when there is poor contextual support and/or environmental noise: [pɛt] or [pɒt]. However, if the context is strong and/or the environment is quiet, a speaker is likely to produce a much less clear low-effort *hypo*-articulated output: close to [pət] (the neutral *schwa* vowel) for both "*pet*" and "*pot*".

The adaptation process of the synthesiser in C2H focuses on low-level signal modifications, but, due to the nature of the speech synthesis generation process, it must also be aware of the phonetic content of speech production.

Most languages have studies reporting lists of competing phones that constitute a source of confusion for human speech intelligibility. Language-dependent LC attractors are hence hypothesised for every phone to define the direction for the hyper/hypo-articulated speech transformation. These attractors are the most likely acoustic realisations towards which speech production converges when the effort reduces – HYO speech – and from which it moves when the intelligibility has to be increased – HYP speech. Once selected, an LC attractor in the acoustic space defines a specific direction along which each phone parametric representation should allow potential movements in order to decrease or increase the degree of articulation. It also defines a distance that specifies the maximum strength that can be used to scale the transform.

Interpolation/extrapolation along the key dimension of hypo/hyper-articulation in the domain of parametric speech synthesis can thus be obtained by controlling the distance from such attractors. The proposed adaptation is achieved with a linear transformation that allows for continuous adjustments of the speech output. The hypothesis is thus that by manipulating the acoustic distance between the realisation of different phones, it is possible to vary the output from HYO speech (i.e. by moving towards the LC attractor) to HYP speech (i.e. by moving in the opposite direction away from the LC attractor) with appropriate consequences for the clarity of the resulting output.

In this thesis, two sets of LC attractors are proposed for two languages: English and Italian, showing that the same energy-motivated transform can be applied to different languages. The success of the transformations is further examined in the experiments of Chapter 6.

Attractors for English

It is observed that both human (van Bergem, 1995) and synthetic (Picart et al., 2010) hyper-articulated (HYP) speech corresponds to an expansion of the vowel F1-F2 chart and, conversely, hypo-articutated (HYO) speech corresponds to a contraction of the vowel space. Figure 4.3 shows how formants of read vowels, which tend to be HYP, shift with respect to those of spontaneous ones, which tend to be HYO, in human speech. The clear trend movement towards the centre can be exploited to identify an LC position for English vowels. The hypothesis is that the mid-central schwa vowel [ə], which is phonetically distinct in the English language, defines the single LC attractor for *all* vowels. When communicative effort is diminished, vowel phonetic contrast is reduced, and they tend to converge to this point. Figure 4.4 illustrates an example for the phone [ɪ] of the transformation vectors along the hypo/hyper-articulation dimension.

The principle is that, whilst it is the case that a given vocalic speech sound can be changed in any direction in the high-dimensional space defined by its parametric representation, the particular location of the neutral *schwa* vowel [ə] defines a specific vector (Figure 4.4) along which it should be possible to produce output with either hypo-articulation – by moving towards [ə] – or hyper-articulation – by moving in the opposite direction away from [ə]. This idea is named *English Vowel*

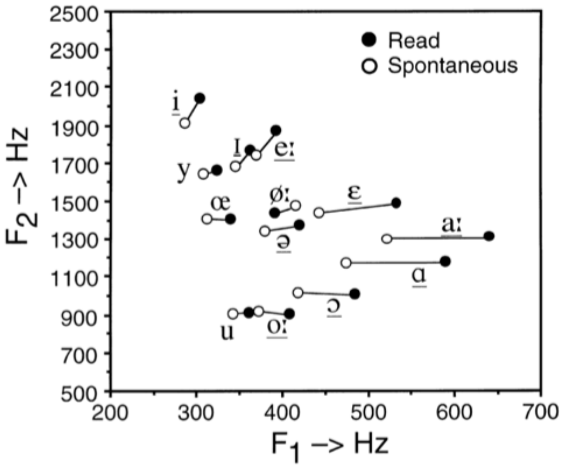


Figure 4.3: *Vowel chart in read (HYP) and spontaneous (HYO) human speech. Note that the vertical axis is inverted with respect to the normal vowel chart of Figure A.1. Figure reproduced from (van Son and Pols, 1999).*

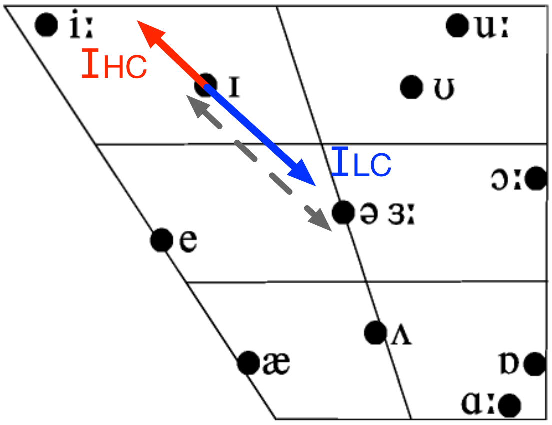


Figure 4.4: *Graphical representation of the transform vectors describing the English vowel HYO reduction (blue arrow, I_{LC}) and the HYP expansion (red arrow, I_{HC}). The dashed grey line shows the competitors used to define the transformation.*

Production Control (E-VPC), and was first introduced by (Moore and Nicolao, 2011) and (Nicolao et al., 2012).

The transform for consonants cannot be defined following the single low-contrastive configuration principle. The glottal plosive [ʔ] or glottal fricative [h] can be hypothesised as unique attractors, but their acoustic characteristics are quite different in term of spectral shape and pitch. Therefore, a simple vector transformation between any consonant and these phones cannot be defined easily.

Another approach to the consonant reduction cab be hypothesised by observing the confusion matrices of consonants in noise (van Son and Pols, 1999). Each consonant is considered to have a particular competitor that is acoustically very close, and hence potentially *confusable*. Once an acoustically similar competitor for every consonant is identified, it is assumed that every intermediate realisation generated along the dimension identified by each consonant-competitor pair is less contrastive than the original phones themselves. The low-contrastive point for each confusable pair of consonants is defined to be half-way between their realisations, as depicted for [d] and [i] in Figure 4.5, representing the LC – and hence low-energy – configurations. On the other hand, HC realisations are generated by moving to the opposite direction away from the half-way point.

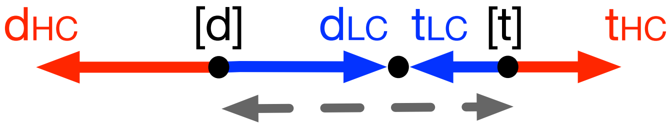


Figure 4.5: *Graphical representation of the English consonant transformation vectors describing the hypo-articulation (blue arrows, d_{LC} and t_{LC}) and the hyper-articulation (red arrows, d_{HC} and t_{HC}) phonetic productions. The dashed grey line shows the competitors used to define the transformation.*

To create this transformations, highly-confusable consonant pairs must be identified in English. The transforms that are trained to convert one phone into another are named *English Consonant Production Control (E-CPC)*. Pair choices can be motivated by different needs: the control of voiced-unvoiced contrast (e.g., [t] vs. [d]) or of confusion in noise (e.g., [t] vs. [p]) as per confusion matrices in (Miller and Nicely, 1955). A detailed map of the adopted contrastive phone pairs used in this thesis is reported in § 5.4.1, below.

Attractors for Italian

The vowel adaptation in the previous section takes advantage of some characteristics of the English language in which a vowel exists, [ə] that is widely recognised as the most common reduced phonetic configuration in hypo-articulated speech (Nicolao et al., 2012). The question therefore arises whether LC configurations can be also found in other languages, such as Italian, where low-energy phones cannot be explicitly labelled.

Italian is a seven-vowel language with some specific differences to English, such as

- the absence of low-energy phonemes such as /ə/ and /h/ in its phonemic inventory;
- vowel acoustic realisations, which typically stand close to the vocalic triangle outer border (F1-F2 chart);
- the variability of stress position in the word, along with the contrastive use of it;
- the contrastive use of consonant geminations.

Even though Italian language does not exhibit schwa in his vocalic system, it can be observed as allophones of some unstressed vowels in spontaneous speech, in some reduction phenomena or in some local dialects (Leoni et al., 1995). Thus, the Italian hypo-articulated speech is also assumed to contain one (or more) LC configurations towards which vowels are reduced. The contrastive use of stressed/unstressed vowels and the consonant gemination, which mostly affects the phone duration, can be also exploited to reduce or increase the acoustic distance between similar phones.

In order to model both Italian Vowel Production Control (I-VPC) and Italian Consonant Production Control (I-CPC) reduction, the same approach used with English consonants is proposed (Nicolao et al., 2013). Phonetically relevant competitors are identified for all phones, and the LC configuration is achieved by applying the half-strength transformation towards them. Ideally, this technique maps both competitors into the same acoustic realisation. As in the E-CPC, the HC configuration is achieved by moving the operational point along the same dimension but in the opposite direction. The main difference with English is the use of multiple target phones to train the speech transformation.

In the Italian transforms, the sole difference between vowel and consonant transforms consists of the criteria by which the competitors are selected. The Italian consonant competitor pairs are selected, analogously to English, by listing the most confusable pairs. The consonant competitors and transformations are similar to those of Figure 4.5. A comprehensive list of the consonant pairs can be found in the literature that studies the perceptual confusion/discrimination of Italian consonants in noise, see e.g., (Caldognetto et al., 1988). For vowels, the competitor is the opposite phone across the F1-F2 chart, as shown in Figure 4.6.

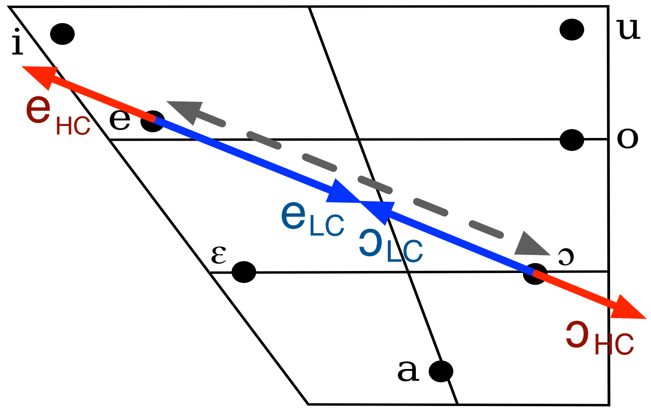


Figure 4.6: *An example of adaptation for Italian vowels, considering [e] and [ɔ]. The blue line refer to the transformations towards the LC point (HYO), the red line to the ones towards the HC configurations (HYP). The dashed grey line shows the competitors used to define the transformation.*

4.3 Perception: enabling context-aware speech synthesis

Speech is intended to be understood by human listeners. Therefore, its quality must be maximised with respect to the capability of a human audience to receive the message in it. Humans can assess the quality of the communication when they speak (Levelt, 1983), using a complex *perception* system to estimate how a listening audience will receive the spoken message. The quality of the *world* auditory scene, in which the communication takes place (e.g., the presence of noise, and/or language barriers) needs to be evaluated. If listeners are thought to face adverse conditions that prevent them from receiving the correct message, human speakers will adapt their production further (Hazan and Baker, 2011). Perception, which is continuously active during human communications, derives its information either from direct observations or from predictive models a) of the self, b) of the communication channel, and c) and of attitude of the listening audience. If direct observation is not possible, it is assumed that a prediction model may be used for the talking agent to estimate the communicative state of the audience. The degree of accuracy of the estimation depends on quality of the prediction models. Models are more accurate if the talker has prior experience about the same category of stimuli, as are commonly spoken and heard.

In control theory, monitoring the system state evolution is key to achieve stability of the desired behaviour. Continuous assessment of the auditory scene state provides the multi-dimensional *feedback* loop signal that allows the error computation. The feedback loop of C2H aims to estimate the *effort* involved in the synthetic speech realisation and the degree of success of such speech at delivering the intended message. It hence enables the model to be *context-aware* and *reactive* when the context in which it operates changes. This part of the C2H model addresses the research questions n. 2 and 5 of § 1.2.

The C2H monitoring loop represents the *perception* equivalent of the model. It encapsulates diverse layers of *sensors* that probe the multiple dimensions that contribute to estimating the *communicative state*. The measured state at a specific time constitutes the input to the comparator (discussed further in § 4.4). The comparator then computes the error against the communicative intent and produces the control signals that drive the speech synthesis transformations (as discussed further in § 4.5).

C2H sensors are designed to estimate the success of the synthetic speech at delivering the intended message. Analogously to humans, a model of the listening audience's perception is therefore used to assess the suitability of each specific TTS realisation.

The C2H model of the listeners requires a) the *self-monitoring* of synthetic speech production, b) the direct or indirect *probing* of environmental disturbances, and c) the assessment of the *listener's* state. The perception function is described by eq. 4.3.

$$\begin{aligned}\text{communicative state} &= \mathcal{F}_{\text{perception}}(\text{world}) \\ &= \mathcal{F}_{\text{perception}}(\text{speech}, \text{environment}, \text{listener})\end{aligned} \tag{4.3}$$

As highlighted by Moore's PRESENCE (Moore, 2007a) and MBDIAC (Moore, 2014) models, intentional speaking agents are currently lacking a model to predict the listener's needs (Moore and Nicolao, 2017).

An example of environment awareness in the exemplar TGSM space is reported in Appendix A.3.1.

4.3.1 Auditory system emulation

In humans, the auditory system is responsible for sensing environmental disturbances. Similarly, in C2H, information about the communication channel conditions can be gathered through sensors that assess the environment such as microphones. Continuous listening to the environmental acoustic scene allows the speech synthesis system to be aware of any adverse conditions that its speech needs to overcome. Alternatively, an inventory of previously-acquired disturbance descriptions can be used. Such models would be selected depending on prior knowledge such as the location description (e.g., train station, street, lecture theatre, silent room, etc.) or the communication medium (direct speech, radio, phone, etc.). In contrast to listening to the acoustic scene, using pre-defined noise models allows the agent to compensate only for average channel disturbances, since continuous reaction to environment condition changes is not available at the time of speech generation itself. A straightforward example of a disturbance estimation model assumes that channel noise energy is constant. Providing the expected noise energy level to the synthesiser allows the system to adjust the speech energy to maintain a constant signal-to-noise ratio (SNR). If the change in noise dynamics is large, part of the speech production may be either insufficiently under- or over-amplified.

Several features can be considered to directly measure disturbances that prevent a listener from receiving the synthesised message. The most commonly accessible quantity is the short-term noise energy level. Energy can be assessed across the whole frequency spectrum or the calculation could focus on specific critical bands that affect the speech frequency range. Perceptual importance functions are often used to emphasise parts of the spectrum that are the most important for human speech understanding, since different environmental disturbances can affect the speech spectrum differently.

Other quantities can be measured to assess the environmental state. The rhythm, evolution, and nature of the disturbances influence the ability of the synthesiser to deliver the complete message to listener.

4.3.2 Self-monitoring

The sensory-motor system, combined with the auditory system, allows humans to listen to their own speech production (Levelt, 1989; Postma, 2000). This ability is fundamental in allowing them to be aware of their own speech production and in assessing the effects of any adjustment made to their speech.

Moreover, humans predict the quality of their production in their mind before producing *overt speech*. Short latency correction can be applied at the *inner speech* – or *planned speech* – level to create realisations which are compatible with the intent (Levelt, 1983).

The C2H model must also be aware of its speech production during the communicative process. Analogously to the trajectory generation of the TGSM model (see Appendix A.3.4), in C2H the agent must be able to query the speech generation algorithm at each step of the communication process to assess weather the communicative intents are being fulfilled.

It is preferable to apply C2H to TTS systems that have accessible parameters. If available, synthetic inner speech can be used to estimate overt speech quality. Access to the internal state of the synthesiser allows the observation window to be reduced. This analogy with the human inner loop (Levelt, 1983) allows C2H to link once more to actual human behaviour. The latency of the system reaction, in terms of number of steps that need to be processed by the model before having an estimation, is determined by the speech generation frame rate. This latency is normally in the order of 50-100 ms, similar to the 100 ms or one-word latency observed in humans (Levelt, 1989).

As mentioned in § 4.2, not all TTS methods allow for direct access to the speech generation models. Standard SPSS system consists of three parts: *lexical analyser*, *parameter generation*, and *waveform synthesis* (cf. Chapter 3). This separation provides a parallel to human speech self-monitoring. The outcome of the parameter generation stage provides enough speech information, such as the energy spectrum and fundamental frequency, to build a meaningful representation of the speech quality, without having to synthesise the complete waveform. Having thus quality estimation available before the speech signal is generated enables C2H to start sending the perception signals to the comparator – and hence to the controller – with considerably reduced latency.

4.3.3 The listener's state

The listener's attitude on receiving a message further influences the success or otherwise of a communicative process. If listeners are motivated, they are inclined to put more effort into compensating for communication errors, compared to when they are not interested in the communication (Moore and Nicolao, 2017). If the speech synthesiser intent is to deliver a message to an unattentive listener, extra effort needs to be applied, e.g., increased loudness and clearer articulation. Human talkers continuously observe their speech recipients' cues (through audio and visual backchannel signals) and estimate their level of engagement in the communication.

A model is considered in C2H to describe the audience's reaction to the communicative process. The main purpose of this model is to establish the listener's motivation level. The presence and demographic metadata of listeners are information that can straightforwardly be acquired. For example, if a listener walks away from the communication scene, simple sensors can return a signal to prevent further speech production. If synthetic speech is deployed to deliver announcements in a school, the audience is likely to be young people whose main focus is unlikely to be listening out for school announcements. The first example represents an instantaneous (at phone- or word-level) signal. The second is longer term, as it might be valid for the entire use of the synthesiser. Short-term measures allow for faster feedback on the speech quality such that inadequacy can be detected – and corrected – quickly. The level of attention may gradually vary during speech production (e.g., a lecture-long speech), so continuous assessment (at the sentence or paragraph level) of the audience's attitude is then needed.

These measurements are quite complex to achieve with an automatic talking agent. They require either prior knowledge about the audience or direct access to their location. The direct method involves direct observation of signals such as listener's backchannel and gestures, or spoken confirmation of the message delivery. These signals are often adopted by dialogue systems to measure the level of engagement. They require several layers of automatic speech recognition, speech understanding, and, where possible, face expression analysis and emotion recognition. Gesture detection requires the speech synthesiser to be able to "see" the listener and interpret their actions. Listening to the communication feedback from the listener adds a great deal of complexity to the synthesiser, as it requires also understanding rather than just hearing capabilities. As such, these direct methods are accompanied by long (sentence-level) latency.

On the other hand, indirect methods mimic the prediction capability of humans to estimate the listener's state. In this case, a speech synthesis system needs to have access to a model that describes how the listener is behaving. This prediction model is stored in the perception block and is used to emulate how the listener is reacting to the synthesised speech. The prediction accuracy depends on the quality of the model. Synthetic speech and environmental conditions are input to this model. The measured level of attention is often computed by estimating the degree of cognitive load that is required by the listener to process the message. The main advantage of this indirect methods is that the listener's supposed behavioural state can be assessed during the speech production. Thus, a divergence in the reception of the delivered message is immediately detected, and countermeasures may be adopted before the speech generation is complete.

4.3.4 Automatically assessing the communicative state

The three types of environmental sensors described above can be analysed singularly and passed to the controller separately. In this way, a threshold on the energy of the noise and the speech, or a speech signal artefact detector can, for example, be used as control signals to drive the speech production.

The most effective way however to analyse the communicative state is to combine the information gathered through different sensors into a concise metric that summarises the degree of effectiveness at delivering the message. Communicative state quality can vary consistently depending on which operational point of the auditory scene at which the communication unfolds. This is depicted in Figure 4.7. The synthetic speech outcome may be processed in relation with the specific environment and listener's attitude, in order to model how effectively the listener is perceiving and parsing it.

Sensor data measures are closely entangled. For example, the speech outcome can be sufficiently effective in a quiet environment, but it can be completely unheard in a noisy context. If a language barrier is affecting the communication, even in a quiet environment, extra articulation effort is required for the listener to receive message.

Several methods exist to estimate the extent to which the intentional message has been received. In the following section, some of the methods that can be integrated into C2H are described.

Intelligibility indices

The easiest – yet effective – criterion to predict if a spoken message is successfully received is based upon the *intelligibility* of each element of the audio signal. The concept of speech intelligibility, however, involves several dimensions of the communication. First, the speech signal must be heard by the listener completely. Second, the message in the signal must have an appropriate code on which talker and listener agree (being grammatically correct, in a known language, etc.). Third, the speech signal must have a meaning that the listener can understand (for example, a comprehensible message).

Transferring this level of complexity to an automatic speech synthesiser can be challenging, and two types of approximations are illustrated that simplify the modelling of the listener's perception. The first type considers the intelligibility of the signal (a signal-processing approach). The second involves the capability of correctly recognising the content of the message (a model-based approach).

Speech signal clarity has proved to be a good prediction of the communicative error (Hazan and Baker, 2011). This technique is widely used in literature to evaluate speech audibility and to infer understanding properties (Tang et al., 2016). The intelligibility of speech can be regarded as a measure of the degradation of the audio signal at the listener's ear. Hence, signal processing techniques that measure the degree of distortion are often adopted. Though they neglect the semantic content of speech, these analyses allow speech intelligibility to be assessed with good correlation to human listener's subjective tests (Valentini-Botinhao et al., 2011). Moreover, the assessment can be done with relatively low latency. This is important because, to be as close as possible to the human auditory system, the

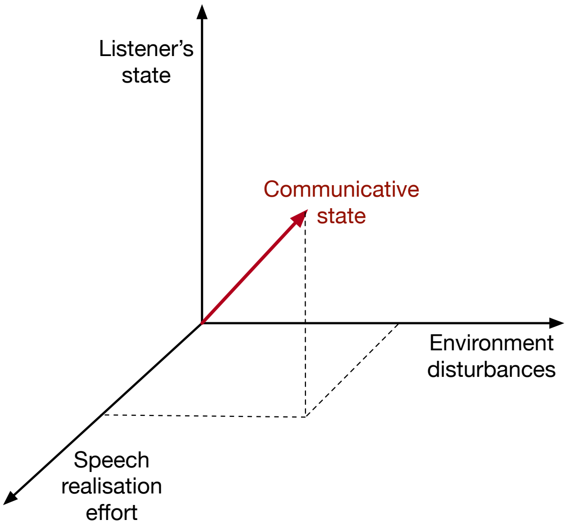


Figure 4.7: *Main components of the communicative states.*

C2H perception loop needs to capture communicative state variations and transfer them to the comparator instantaneously.

The most accurate intelligibility scores include several elaborate auditory processing stages (Tang et al., 2016). These measures compare an internal representation of the clean reference speech signal with an internal representation of the noisy signal in order to predict how intelligible the noisy signal is (Valentini-Botinhao et al., 2011).

Several procedures can be used for estimating speech intelligibility in normal-hearing listeners (Kates and Arehart, 2005). The *articulation index (AI)* (Kryter, 1962) was one of the first to be proposed. It has been further developed to produce some of the following indices,

The speech transmission index (STI) (Steeneken and Houtgast, 1980) is calculated by detecting the modulation amplitude for each sub band in the environment. The estimation of the SNR is formed by a weighted sum across frequency. STI uses speech-shaped noise or speech as the stimulus. The STI can be used for reverberating environment.

The speech intelligibility index (SII) The ANSI Standard (ANSI, 1997) calculates the signal-to-noise ratio (SNR) on a decibel (dB) scale for each frequency band in the frequency domain, considering the masking effects and specific auditory thresholds. The weighted SNRs are then summed to produce the intelligibility estimation. SII is effective with stationary, additive noise, and for bandwidth reducing filtering. The SII procedure contains a model of the auditory periphery, with known auditory thresholds for normal and impaired hearing. SII is an utterance-level average estimation. Since the C2H model requires real time measures, the eSII (Rhebergen and Versfeld, 2005) is proposed. Figure 4.8 shows the measurement of such an index when the environmental noise is fluctuating. The implementation of the time-varying intelligibility index will be reported in § 5.6.1.

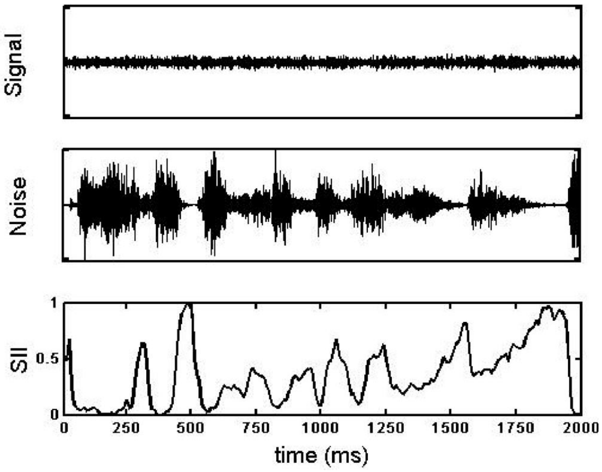


Figure 4.8: *Example of eSII analysis in relation to a speech signal in fluctuating noise. The upper plot shows a female speaker's speech signal. The middle plot represents a fluctuating speech-shaped masking noise. The lower plot displays the resulting index as a function of time. The SII averaged across time is equal to 0.35. Figure reported from (Rhebergen and Versfeld, 2005).*

Perceptual Evaluation of Speech Quality (PESQ) (ITU-T, 2001; Rix et al., 2001) is a measure designed for predicting the quality of speech signals transmitted over a telephone line. The measure includes an auditory transform and considers the masking phenomena for the comparison of this transformed representation, but cannot handle wideband speech signals because it was specially designed for narrowband signals.

Dau index (Dau et al., 1996a) is based on the human auditory model developed in (Dau et al., 1996b). The model is a time domain representation that incorporates aspects of temporal adaptation. The measure corresponds to the normalized correlation coefficient of the internal representation derived by the Dau model for reference and noisy signals. The correlation is evaluated over sliding window frames, computing the average of the values in high energy frames.

Glimpse Proportion measure (GP) (Cooke, 2006) derives from the Glimpse model for auditory processing. The assumption underlying this model is that listeners can reconstruct a speech signal in noise by listening out for short glimpses of speech that are relatively unmasked. The internal representation is derived using Gammatone filter banks, and measures the proportion of spectro-temporal regions in which speech is more energetic than noise. The GP value is also computed on a frame-wise basis.

Another model-based approach is to use *pronunciation assessment analysis* of the synthetic speech outcome. For example, the approach described in (Nicolao et al., 2015) can successfully pinpoint the phone and word level distance of a specific speech production to an ideal target example. In addition to the measure of intelligibility, this internal detection can also help to identify competing realisations that would generate word- and phone-level confusions. The competitor identity could also be used as a signal to the controller to determine the correct action to take in order to minimise that specific confusion (Cooke, 2009). The alternative is to use a statistical description of normal phonetic confusion.

Some of these methods are used in C2H to measure speech intelligibility.

The understanding models

The mere fact that a speech message is heard by the listener does not assure that the message is meaningful or that it can be understood. For example, it may contain nonsense words, or language, education, and cultural barriers can create misunderstanding. In order to design a speech synthesis system that checks the linguistic and semantic quality of its speech, an accurate model of the listener's speech recognition is required.

The recogniser model aims to replicate how the audio message is processed by the listening audience, assuming that the signal is audible at their ears. Several model-based sensors that predict the perception of the listener can be considered, each of them focusing on different aspects of the communicative intent. Some relevant speech understanding models are listed below.

The most immediate model that can be considered consists of *automatic speech recognition (ASR)*, which is tailored to the key listener's characteristics, considering native language, regional accent, and lexical vocabulary. A carefully trained automatic speech recognition (ASR) can emulate the understanding capability of a specific listener, and hence predict to what extent the message content is correctly received. Significantly low ASR performance might be an indicator that some synthetic speech production characteristics, such as speech rate, coarticulation effects, sound volume, or regionally accented voices, are not compatible with the recognition models. Hence, it is likely that the listener, modelled by the ASR, is not able understand the synthetic speech. Once the speech audio has been decoded and a written transcript is obtained, a *semantic analyser* can read the ASR output and extract the semantic content of the message to summarise its general meaning. Sometimes the communicative intent can be understandable even if the audio cannot be decoded entirely. Other aspects of the communicative intention can be detected with a *pitch/stress analyser* which can detect what type of intonation is perceived an *emotion detector* which can check what type of the expressiveness is conveyed, and a *speaker recogniser* which can predict whatever the listener can recognise the identity of the speaker.

The parametric speech synthesisers that are suitable for C2H, are based on statistical parametric acoustic models of speech. As a result, SPSS internal generative structure can be inspected to provide the basis of possible source error predictors. Confusion networks (Mangu et al., 2000) or confidence measures (Zha, 2014) extracted from the speech generative process can help to predict the degree of confusion that competing speech realisations might generate in listeners. A high number of active parallel hypotheses during the SPSS generation stage, is proportional to high uncertainty that an ASR would experience at recognising that speech outcome. The hypothesis here is that ASR and a human listener's uncertainties are proportional when the acoustic model is accurate.

An ASR-based perception loop can also return information about the identity of competing phones and words that generate recognition confusion. Such information can then be used to adjust the realisation effort in correspondence to those speech elements only.

The sensor implementation adds another layer of complexity. The models used in the prediction must be carefully tailored to the listener's characteristics. They need to be aware of what knowledge is available to listeners, what their cultural and linguistic backgrounds are, what their level of attention is, and their native language, etc. Model mismatch can become a source of measurement errors. This presents another analogy with the human communication process, since the human-to-human communication is also more effective when the speakers are aware of their audience in terms of their characteristics and needs (Hazan and Baker, 2011). Similarly, the C2H model produces a more successful speech outcome when it has access to accurate models of the listening audience (Moore and Nicolao, 2017). If direct access to the listener is unavailable, an approximate model, which incorporates some basic dimensions of speech, may be selected instead.

4.3.5 A synthesiser that evaluates the quality of its outcome

The perception part of C2H, in which the synthesiser can evaluate the quality of its outcome, represents the auditory feedback loop of the complete C2H system.

Perception consists of three main types of sensors: acoustic environment probing, self-monitoring of inner speech, and detection of listener's state. The data acquired through these sensors contributes to create a listening audience emulation. Such models normally aim to predict the level of intelligibility and/or understanding of the communicative process. The dimensions along which the prediction must operate are determined by the communicative intent (to maximise speech clarity, to deliver complete message, to transfer an emotional content, etc.) and by the quality of the assessing measures (reliability of intelligibility measure and awareness of the listener's needs).

4.4 Comparator: assessing realisations

A continuous connection between *perception* (§ 4.3) intent, and *controller* (§ 4.5) is essential for the C2H model to achieve a synthetic speech production system which *reacts* to the environmental disturbances with relatively short latency.

At every step of the speech synthesis process, C2H has the capability of checking to what degree the speech realisation is fulfilling the communicative intents.

The communicative state measured or predicted by the perception block needs to be contrasted with the goal of the communicative intent. This comparison happens in the *comparator* and is expressed by

$$\text{error} = \mathcal{F}_{\text{comparator}}(\text{communicative state, communicative intent}) \tag{4.4}$$

Here, the communicative state, communicative intent, and error are all multi-dimensional vectors, in which each component addresses a specific aspect of the intent: intelligibility, appropriateness, expressiveness, loudness, etc. The system distance to the intent is the *error* signal that drives the controller block, and that consequently causes the control signal which transforms the speech realisation.

In the TGSM example (cf. Appendix A.3.2) the comparator is essentially represented by the Euclidean distance between the trajectory and the target.

4.4.1 Multiple layer comparator

In standard control theory, a comparator only computes the difference between the reference and measured signals. In C2H, however, the conceptualiser intent can account for several dimensions. As a result, the comparison between communicative state and intent needs to allow multiple layers of control. For example, the perception block can return information related to self-monitored speech, to the outer environment, and to the listener's state. If a specific goal layer is prescribed in the communicative intent, the comparator must be tailored to compute the error prioritising that specific domain. Error layers belong to different domains: audibility, expressiveness, understanding. Perception layers that can be checked depend on the definition of intent and include a) the speech outcome semantic content that is planned in the intent (the text message), b) the production artefacts in the speech realisation (when clear speech is required), c) the audibility level of speech waveform in loud noise (when the goal is to be above an intelligibility threshold), d) the listener's word error rate (WER), that could not reflect audibility (when message transfer is the goal) and e) the speech expressiveness coherence with the goal.

The C2H multi-layer comparator error is input to the controller part of the model. This is theoretically also capable to handle multi-layer control signals, however, in the current design, the comparator is limited to the determining the distances of text message from completion, and of the speech waveform intelligibility from minimum goal threshold. The distance from complete text message is determined by how far in the synthesis the TTS is.

4.5 Controller: adjusting the communicative effort

As previously stated, the C2H model aims to produce Lindblom's H&H production model (see § 2.2.1). In the same way, the speech synthesiser in C2H constantly balances the effort applied to the transform (i.e., the transformation strength) against the effectiveness of the communication (i.e., matching the intent). The controller part of the model is thus designed to target the research question n. 5 of § 1.2.

Humans do not aim to maximise communicative success at all time. Lindblom's model and general linguistic observations state the minimal effort that achieves the desired degree of success is the optimal choice. Since listeners are accustomed to this speech behaviour, a speech synthesiser that is perceived as unnecessarily hyper-articulating would be received as unrealistic, if not irritating. The appropriateness of realisation loudness and articulation is more important than its maximisation. Depending on the communicative intent, a reduction of the production effort can counter-intuitively be the optimal strategy, for instance in cases where the goal is for the speech to sound friendly and colloquial rather than assertive and patronising. In the computer assisted language learning (CALL) context, for example, a friendly tone can be more appropriate than an impersonal reading voice.

Scaling the speech synthesiser transform is the role of the *controller* block of C2H, shown in Figure 4.1. The component behaviour is formalised in:

$$\text{control signal} = \mathcal{F}_{\text{controller}}(\text{error}, \text{motivation}) \tag{4.5}$$

The comparator error is converted into the *control signal* to appropriately scale the transform of § 4.2.1. As expressed in § 4.4, the error signal is multi-dimensional. The *error components* that are considered for the optimisation and the *amount of energy* that the system should invest in trying to minimise the error and increase the success rate of the communication are selected by the conceptualiser through its *motivation* component. The motivation component models the level of *attention* that talking agents use in the communicative process.

The conceptualiser can in principle modify the motivation level during production. However, in C2H, it is assumed to be selected before speech production starts, and it is constant throughout the entire communicative process.

Proportional–integral–derivative (PID) controller is a control mechanism widely used in classical control theory (Lipták, 2003). A PID controller continuously transforms the error $e(t)$ between intent and perception into the correction signal $u(t)$, based on *proportional*, *integral*, and *derivative* terms.

$$u(t) = K_P e(t) + K_I \int e(\tau) \mathrm{d}\tau + K_D \frac{\mathrm{d}e(t)}{\mathrm{d}t} \tag{4.6}$$

In the C2H controller, the proportional K_P and integration K_I coefficients are normally considered. For example, the error signal can directly act on the output signal amplitude and incrementally build the spectral adjustment variations.

As described in Appendix A.3.4, in the TGSM domain, the trajectory generation adaptation is normally a linear function of the error derived by the proximity functions.

4.5.1 Moving along the H&H dimension

The main purpose of the control in the C2H model is to compute a correction vector to adjust the transform and minimise the system error. The maximum correction that the system is allowed to produce is a function of the input motivation (*effort*) that is set in the communicative intent.

In the C2H model, the synthesiser intention and its communicative goal are assumed to be set with prior knowledge, and are defined externally. These input configurations are also presumed to be fixed for the entire duration of the communication, though the C2H model is sufficiently flexible to allow them to additionally evolve through time.

The error signals computed by the comparator determine the intensity of the correction signals. In C2H, these normally consist of loudness adjustments and speech parameter adaptations, as described in § 4.2.2.

As listed in § 4.3.4, many types of errors are expected to be measured by C2H. Consequently, the controller needs to address corrections on each of these error dimensions.

The motivation, which drives the controller, selects the aspects of the communication to maximised and the effort involved. Three foundation layers can be identified.

Audibility analyses the ratio between the produced speech and the environment disturbance energies. Inner speech needs to be accessed separately from the background noise. Transformations at this level are signal-based and they resemble the speech enhancement techniques. This layer neglects the semantic content message and the speech nature of the signal. E.g., SNR results are equivalent if applied on direct or time-inverted signals.

Intelligibility The most important frequency bands that are expected to carry the largest part of speech signal information are identified. Control along this dimension might result in spectral energy reallocation to enhance the contrast among critical bands of competing phones. This control again neglects the message content, and indeed extreme transformation which alter the phonetic identity as a result of such control. For example, “but, /b ʌ t/” might be adapted into “bit /b i t/”: the new phone sequence might be more intelligible, but the message changes.

Understanding controls the meaning mismatch between intentional and perceived messages. The semantic content of speech, its expressive valence, its appropriateness, and the listener’s attitude can all be evaluated. This approach requires sophisticated human understanding models. Transformations can act at a linguistic level, e.g., by changing the word sequence. The main limitation here is that the whole reaction time increases considerably, and it can at times even require a complete restart of the whole production.

If the goal is to optimise speech intelligibility, the measured index in the current environment (e.g., the eSII value) can be controlled by changing the adaptation strength proportionally to eq. 4.6. If speech understanding must instead be increased, the correction signal should act on the generative parameters (e.g., acoustic models in SPSS) such that the message is delivered in its entirety (e.g., maximising the accuracy of the ASR transcription).

One of the most important characteristics for creating a reactive C2H is that the control link between the perception and generation processes must have low latency. That is, any measurement received from the perception loop must be translated into a control signal with low latency. Here, the delay is proportional to the length of the observation needed to produce a control decision. High level perception signals, such as emotion, attitude, understanding require long observation windows. On the other hand, low-level perceptions, such as audibility and intelligibility, can generate an almost immediate reaction.

Not all errors can be corrected by a PID controller based on a gradient descent. Similarly to the human monitoring and self-repair mechanism (Levelt, 1983), if the error signal is too large, then volume change or speech adaptation are no longer adequate reactions. A very large error may need to trigger a different action such as the *interruption-and-restart* of the speech generation with the corrected version. Even if it is not listed in the set of available transforms, C2H can account for this type of major self-repair.

4.6 The complete C2H model

In this chapter, the complete C2H model structure is described.

The goal of C2H is to create a speech synthesiser that is aware of the context in which operates, and that is able to adapt to any changes of the communicative context. State-of-the-art synthesisers currently lack this capability. Therefore, the C2H model proposes adding a perception feedback loop to a standard TTS framework. Such negative control loop enables adjustment of the speech realisation according to the perceived communicative error.

The model design follows the principles of Levelt's perceptual loop theory (Levelt, 1989) and Powers' PCT (Powers, 1973). The perception module estimates the communicative state, a comparator computes the error with respect to the input intentions, and a motivation-driven controller modifies the speech production to minimise the error (cf. § 4.5).

The speech production adjustment mechanism is designed to model the principles of Lindblom's H&H theory, as discussed earlier in § 4.2.1. The communication process is the result of balancing the degree of success against the effort involved in the production.

The main idea of Lindblom's theory revolves around the observation that the *effort* that communicative agents adopt in their speech production is proportional to the degree of *articulation* of the phones (cf. Figure 4.2). An acoustic approach to transform the degree of articulation is proposed, where control of the phonetic distance in the acoustic space (*phonetic contrast*) is coupled with the control of the degree of articulation.

The overall process that describes the C2H synthetic speech generation, can thus be described by a recursive function in which all the principal components that influence the process are listed,

$$\text{speech}_{t+1} = \mathcal{F}_{\text{C2H}}(\text{speech}_t, \text{listener}_t, \text{environment}_t, \text{goal}_t, \text{motivation}_t) \qquad (4.7)$$

New speech at time $t + 1$ is created by descriptors of the communicative state at the t-th step (speech$_t$, listener$_t$, and environment$_t$), and by expressions of the communicative intent (goal$_t$ and motivation$_t$).

The high-level diagram of the C2H model was illustrated earlier in Figure 4.1. It can now be expanded in a more detailed description of the constitutive elements. The expansion shown in Figure 4.9 combines the arguments discussed in this chapter with some of the ideas proposed in the models by (Levelt, 1989) and (Hartsuiker and Kolk, 2001).

At the top of the Figure 4.9, the *conceptualiser* handles the communicative intent that has to be achieved. Communicative intent consists of a goal to be achieved (normally the synthesis of a text message), and a motivation that selects the communicative effort and dimensions to be assessed by C2H.

In the *speech synthesiser* part, the characteristic TTS components are displayed (cf. Figure 3.3). Three main stages are identified: lexical analysis, articulation, and waveform generation. The control signal acts both at lexical and articulation levels. The first produces the semantic content. The second contains the transform that changes the degree of articulation.

The *controller* stage produces the signal that adjusts the speech transform. It is driven by the error computation of the *comparator*. The error is generally the difference between the expected message and level of motivation, and the perception loop measurements.

The *perception* part of C2H mainly focuses on the signal intelligibility assessment. The main reason behind this design choice is the need to minimise the system feedback loop latency, and the fact that speech intelligibility measures can sample the environment faster than an speech understanding model. The perception sensors observe the inner speech generated by the synthesiser, and the environmental disturbances. The listener's feedback assessment is not directly

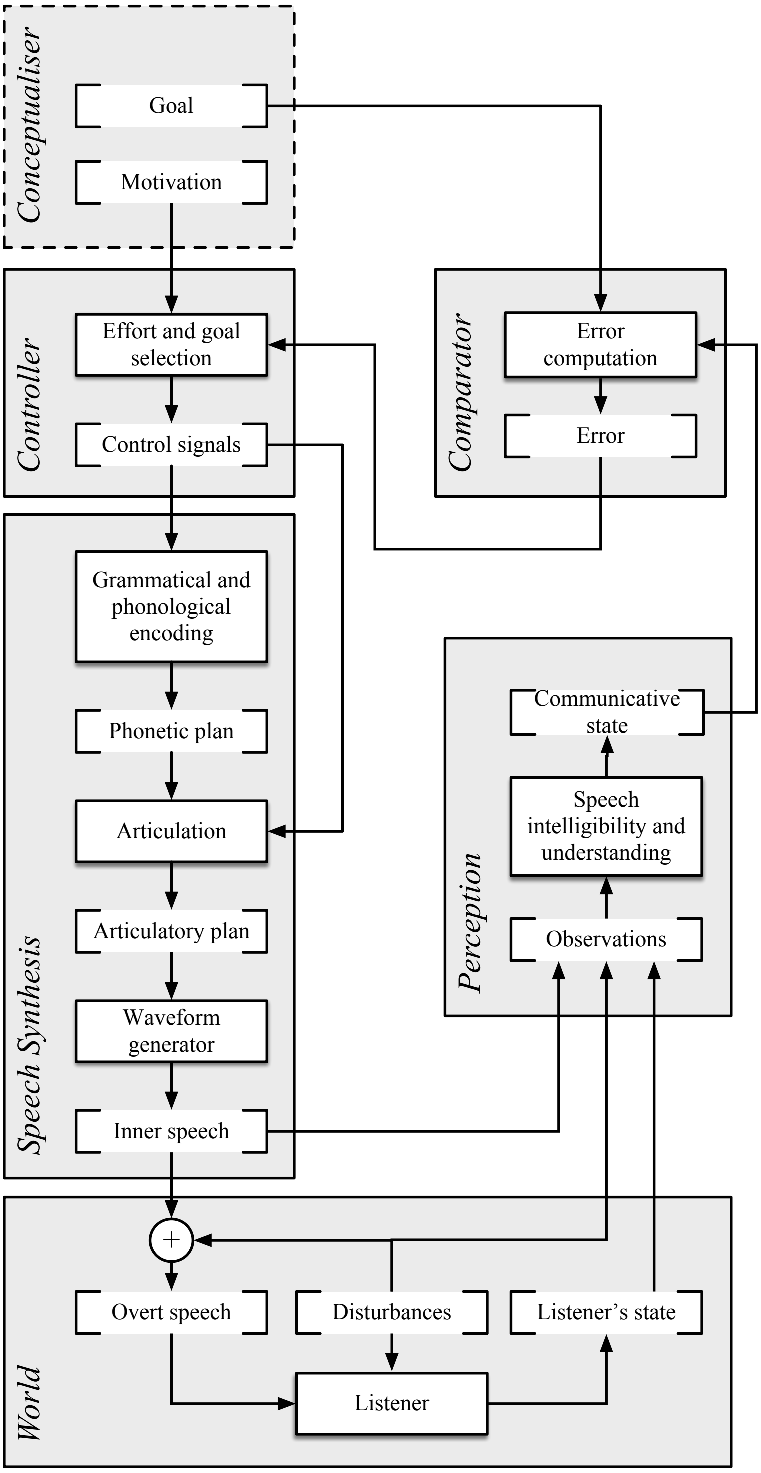


Figure 4.9: *The complete outline of the C2H model of speech production. The TTS constituents, on the left-hand side, the auditory perception loop, on the right-hand side, and the adaptive control, in the centre, blocks are highlighted. The framed boxes indicate the implementable functions. The bracket boxes identify the signals.*

measured in C2H at present, but are derived mostly from prior information. This helps to select the correct audience model (using age group, education, demographic, etc.), and is assumed to remain constant. The inner speech can be extracted either before or after the waveform is created. Articulation features, such as energy spectrum envelope and pitch contour, transport enough information to correctly estimate the speech qualities.

Communicative intent and *environmental context* characteristics are the sole external inputs of C2H. The former is the intentional scope for which the synthetic speech system is created. The latter contains all the elements that are external to C2H and cannot be controlled by the model.

Chapter 5

HTS-C2H: a C2H Implementation

Contents

The C2H model described in Chapter 4 is a flexible framework that enables a speech synthesiser to reproduce the Hyper and Hypo articulation theory (H&H) behaviour. This chapter introduces an implementation of the model so that experimental evidence can be gained to provide answers to the research questions proposed in § 1.2.

C2H comprises several blocks and functions, as were depicted in the diagram of Figure 4.9. However, not all of these functions are implemented. The main objective of the current implementation focuses on the *speech synthesis* process and its *adjustment* capabilities. Automatic assessment methods to *evaluate* the communication success are also investigated. The C2H *conceptualiser* and some parts of the *perception* feedback loop that would involve the direct observation of the listener (cf. § 4.3.3) are not implemented at the present, as they require further investigation that exceeds the scope of the research questions in this thesis.

This chapter addresses the answers to the research questions n. 2, 3, and 5, as it demonstrates the possibility to implement and test the proposed C2H model in a realistic scenario.

As suggested in § 4.2, statistical parametric speech synthesis (SPSS) is the most suitable synthesiser to apply in the model. Its implementation named HTS-C2H is therefore created to test to what extent the solutions described in the previous chapter can reproduce behaviours observed in human communication.

As introduced in § 4.2.1, C2H can operate in diverse languages. Language-dependent differences of the HTS-C2H implementation structure are therefore highlighted in addition.

5.1 Speech synthesiser

Figure 4.9 of Chapter 4 shows that the speech synthesiser to which C2H can be applied, must be in fact a TTS stage. Therefore, in order to provide an implementation to test the effectiveness of the computational model, a suitable TTS synthesiser must be selected. Such a synthesiser needs to be able to control the spectrum and the duration of speech effectively, and this control must be scalable and continuously applicable during synthesis (cf. § 4.2).

Table 5.1 summarises and compares the four most commonly available speech synthesis methods, and reveals their applicability for the C2H framework. The positive signs '+' indicate that the related characteristics exist, and they are reasonably accessible. The negative signs '-' state that those attributes may represent an obstacle against applying C2H to that synthesiser.

Most state-of-the-art TTS synthesisers generate their speech realisations in a single pass. In these TTS systems, the input goal (message, expressiveness, and motivation) is established at the beginning of the synthesis, along with the

Table 5.1: *Comparison of the four most common automatic TTS methods. The number of positive "+" and negative "-" signs indicates the degree of suitability of a TTS method for application in C2H.*

Characteristics	**Concatenative**	**SPSS**	**Physical model**	**End-to-end**
intelligibility	+++	++	-	+++
expressiveness	- - from inventory	++ scalable	++ scalable	++ scalable
training data	- single speaker high quality	++ multiple speakers high quality	- - single speaker, expensive EMA data	+ very large amount, any quality
computational cost	- memory	+ optimal search	- - dynamic equation solving	+++ several layers of neural networks
out of domain synthesis	- - require extra data	+ interpolation in the acoustic space	++ mapping articulation to unknown sounds	+ autoregressive generation
speech modifications	- - signal processing only	+ interpolation in the acoustic space	- interpolation in the articulatory space	+ using input conditioning features
analogies with human production	- -	+	++	-
real time reaction	- - sentence level, long	++ recursive generation, immediate	++ short latency	+ short latency
maturity	+++	+++	- -	-

generation models and the selected modifications. Once these synthesis parameters are chosen, the outcome of the entire utterance is determined. An example of such systems is the *concatenative* synthesiser. C2H cannot be applied to concatenative synthesis, as the modifications to speech with this method are not easily scalable. Partial realisations are not accessible before the entire speech signals are created.

All SPSS architectures might in principle be used in C2H, since they normally allow for robust and scalable modifications of the generation parameters. However, some amendments to the standard generative algorithm are required in order to be able to control the adjustment strength of speech production during the generation process (cf. § 3.2.5).

Other types of synthesiser, such as the *physical-model* and *end-to-end* synthesisers, provide access to the state of the system at each production step. The outcome of the physical-model synthesisers is theoretically scalable as they have a direct correspondence to physical quantities such as articulatory movements, amplitude, and airflow pressure. However, this relationship, which is generally learned from experimental data, is heavily speaker dependent. Any modification can lead to unstable and unintelligible speech as the re-synthesis methods for these parameters are often inefficient.

End-to-end systems, on the other end, are innately autoregressive, therefore their output is constantly processed to generate the next samples. However, the level of complexity of such synthesisers requires a large amount of training audio. As a novel development, further research is needed to implement effective conditioning mechanisms to continuously scale the adjustment. In other words, such a system might be able to produce two different voices, but it is unlikely that it can generate unseen speech which is an interpolation between two models.

In conclusion, any TTS might be applied in the C2H model with differing degrees of change to their standard algorithms. However, from the foregoing analysis (summarised in Table 5.1), statistical parametric speech synthesis (SPSS) (Zen et al., 2009) emerges as the most suitable method in term of simplicity and flexibility of the implementation. In parametric synthesis, phone characteristics can be altered continuously in any direction of the high-dimensional space defined by their parametric representation, using any available adaptation techniques. As described in § 3.2, state-of-the-art SPSS allows two types of approaches: HMM-based (Tokuda et al., 2013) and DNN-based (Zen et al., 2013) parameter generation. At present, the quality of SPSS methods such as the DNN-based speech synthesis seems to outperform the HMM-based ones, but the computational load, adaptation techniques, and training process do not show the required stability and speed to be implemented in C2H.

TTS is composed of two main parts, depicted in § 3.1 and Figure 5.1: *linguistic analysis* and *waveform generation*.

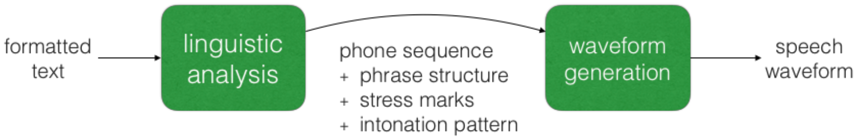


Figure 5.1: *Block diagram of standard TTS system.*

In particular, among the state-of-the-art SPSS waveform generators, HTS is chosen, described in § 5.1.2 below. HTS does not include any linguistic analysers. Valid text analyser options that can be combined with HTS are the Festival speech synthesis system (English, Spanish, etc.), the DFKI MaryTTS system (German, English, etc.), *Flite* combined with the *hts_engine* (English), *Open JTalk* (Japanese).

In the following paragraphs, the implementation choices for the linguistic analysis and waveform generation components of HTS-C2H are described.

5.1.1 Linguistic analysis

Linguistic analysis maps an input text into the standard representation that drives the parameter generation algorithm. It translates word-level information into phone sequences, as well as deciding how the synthesiser should pronounce it.

In HTS-C2H, the linguistic analyser must fulfil certain requirements. The input to this component is human readable text, perhaps with labels and tags to describe the type of expressiveness intended. The output is a phone-level description of the speech sound realisations: phone identities, durations, and positions in the word/sentence.

The initial stages of the process are designed to normalise the text and to disambiguate the sense of each word and symbol. The subsequent stages deal with the translation of words and any additional syntactically or semantically derived information, providing a compact description of the sounds to be generated.

In this thesis, two different frameworks for linguistic analysis are used, corresponding to the two different languages that the experimental part considers: English and Italian.

Festival

The linguistic analysis for the English language is provided by the standard Festival Speech Synthesis system. Festival is a general multi-lingual speech synthesis system developed at the Center for Speech Technology Research in Edinburgh (Clark et al., 2012; Taylor et al., 1998). It offers a full text-to-speech system with various software tools, as well an environment for the development and research of speech synthesis techniques. It is written in C++ with a Scheme-based command interpreter (Kelsey et al., 1998). Festival provides both lexical analyser and waveform generator for a general-purpose concatenative TTS architecture that uses the residual LPC synthesis technique.

In HTS-C2H, the lexical analyser for the English language is used. It is able to transcribe unrestricted text to speech.

Festival lexical analysis operates with text and linguistic/prosodic processing. The modules and the sequence of operations that need to be applied to perform a Festival lexical analysis are reported in Figure 5.2.

Festival is multi-lingual and voices in many languages have been developed, including English (UK and US), Spanish, Italian, and Welsh. The tools to build English voices are the most advanced.

MaryTTS

The linguistic analysis for the Italian language is provided by the modified version of the TTS software, MaryTTS (Multimodal Speech Processing Group, 2018; Schröder et al., 2011). MaryTTS is an open-source, multilingual TTS platform written in Java. It was originally developed as a collaborative project of DFKI's Language Technology Lab and the Institute of Phonetics at Saarland University. It is now maintained by the Multimodal Speech Processing Group in the Cluster of Excellence MMCI and DFKI. MaryTTS supports German, British and American English, French, Italian, Luxembourgish, Russian, Swedish, Telugu, and Turkish. It provides toolkits for quickly adding support for new languages and for building unit selection and HMM-based synthesis voices.

Natural language processing (NLP) is responsible for the prediction of speech-relevant data extracted from input text: phone symbols and intonation labels. Similar to Festival, MaryTTS analysis is organised in a modular way. The sequence of operation is reported in Figure 5.3. The output of the NLP component is a rich *MaryXML* structure.

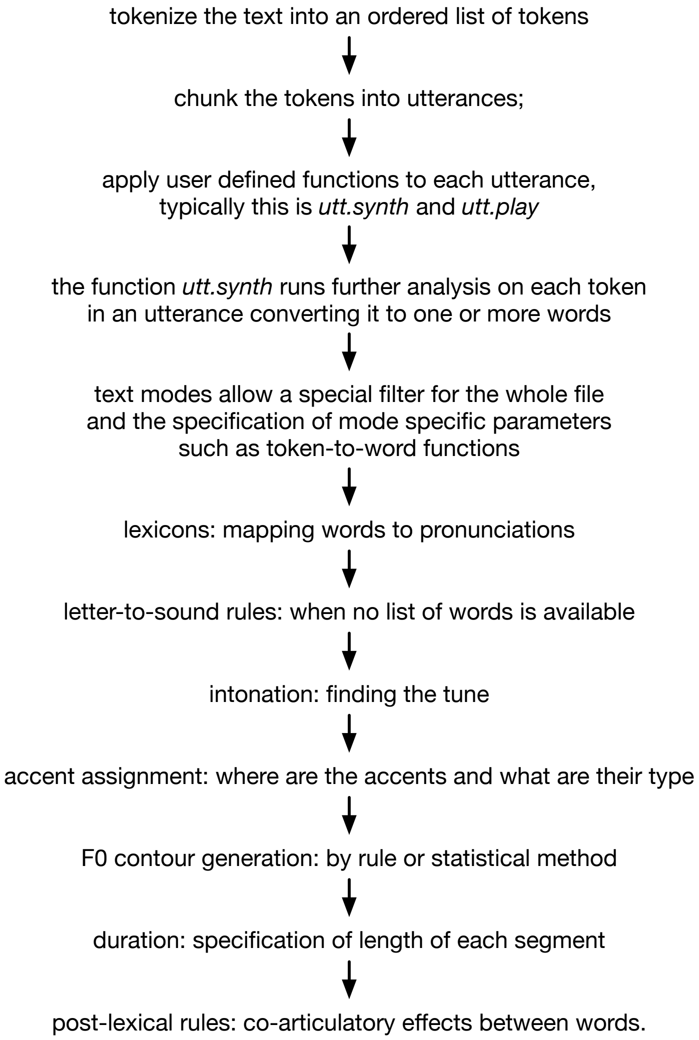

tokenize the text into an ordered list of tokens
chunk the tokens into utterances;
apply user defined functions to each utterance, typically this is *utt.synth* and *utt.play*
the function *utt.synth* runs further analysis on each token in an utterance converting it to one or more words
text modes allow a special filter for the whole file and the specification of mode specific parameters such as token-to-word functions
lexicons: mapping words to pronunciations
letter-to-sound rules: when no list of words is available
intonation: finding the tune
accent assignment: where are the accents and what are their type
F0 contour generation: by rule or statistical method
duration: specification of length of each segment
post-lexical rules: co-articulatory effects between words.

Figure 5.2: *Sequence of Festival modules to perform the lexical analysis.*

The Italian modules for MaryTTS (Tesser et al., 2013) have been developed by the ISTC-CNR research institute (ISTC, 2019) and they include: a) Italian lexicon and letter-to-sound rules, b) context dependent part-of-speech tagger, c) ToBI rules (Silverman et al., 1992) to predict symbolic prosody from text, d) a customised version of the Italian SAMPA phone set (UCL Phonetics and Linguistics, 1989).

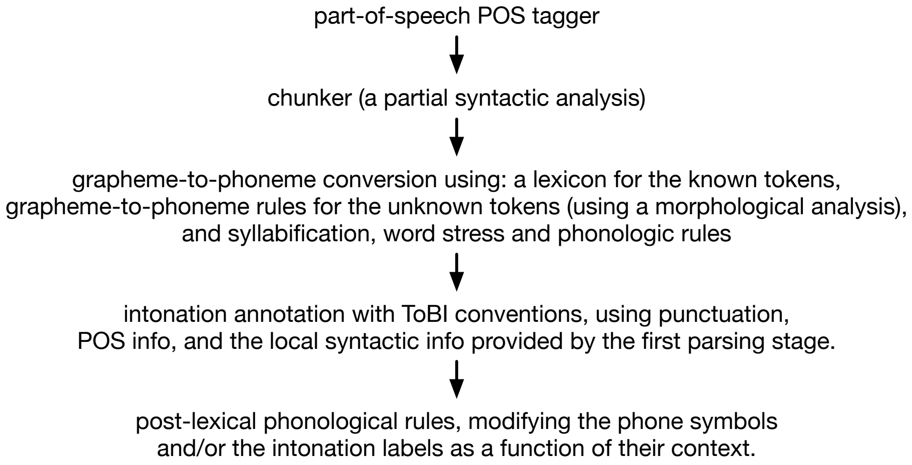

part-of-speech POS tagger
chunker (a partial syntactic analysis)
grapheme-to-phoneme conversion using: a lexicon for the known tokens, grapheme-to-phoneme rules for the unknown tokens (using a morphological analysis), and syllabification, word stress and phonologic rules
intonation annotation with ToBI conventions, using punctuation, POS info, and the local syntactic info provided by the first parsing stage.
post-lexical phonological rules, modifying the phone symbols and/or the intonation labels as a function of their context.

Figure 5.3: *Sequence of MaryTTS modules to perform the lexical analysis.*

5.1.2 Waveform generation

Waveform generation converts the articulate symbolic representation that result from the linguistic analysis into a speech waveform. The waveform generation stage in SPSS must convert the sequence of allophone symbols from linguistic analysis into a continuous speech waveform and apply the correct intonation to the waveform.

This stage's operations can be regarded as sequence-to-sequence regression task between phone-level and signal-level sequences. In the first domain, one linguistic specification vector is provided per each phonetic unit. In the second, the timeline expands each linguistic vector into one parameter vector for each frame. The regression task is followed by a vocoder stage in which the parameter sequence is converted into speech waveform.

HTS-C2H uses HTS as parameter generator (cf. § 3.2.3) and STRAIGHT as vocoder (cf. § 3.2.1).

Parameter generation

HTS 2.2 is chosen to provide the parameter generation functions of the HTS-C2H implementation. A *regression tree* followed by context-dependant HMMs represents the statistical models that drives HTS generation process. Since this HMM model adopts Gaussian functions to describe the output vector space, simple but effective adaptation techniques can be applied to the generation process to allow for an effective way to control the speech production.

Vocoder

The vocoder part of the synthesiser is not included in the standard HTS release. The MGLSA algorithm or STRAIGHT, see § 3.2.1, are compatible options to use in HTS.

In this thesis, the STRAIGHT analysis/synthesis tool is chosen to be the waveform parametrisation that is trained by the HTS-C2H statistical models. Compared to other vocoders, this vocoder has been proven to be effective in reducing the "buzzy" effect on the final speech realisation.

Since the C2H output speech parameters need to be transformed into the speech waveform during the generation, the HTS-C2H vocoder needs to be able to operate and synthesise limited portions of the signal. This can be achieved by implementing a buffer, in which speech parameters are accumulated, and which triggers the vocoding conversion when it contains sufficient speech material.

5.2 Synthetic voices

New English and Italian voices are developed for HTS-C2H, using the standard recipe provided with Festival.

The training of a single-speaker voice in a new language requires addressing the creation of linguistic analysis modules, such as: phone-set definition, token processing rules, prosodic phrasing method, word pronunciation (lexicon and/or letter to sound rules) and intonation (accents and F0 contour). On the other hand, duration and spectral waveform characteristics are learned from audio examples by training a set of context-dependent HMM using the HTK.

HTS-C2H uses the conventional HTS 2.2 (Zen et al., 2007a) training regime to derive the specific acoustic models (cf. § 3.2.1). The parameters of the HMMs can be estimated based on the maximum likelihood (ML) criterion by the expectation maximisation (EM) algorithm. Spectral parameters are modelled with Mel-generalized cepstral (MGC) features by context-dependent HMMs with continuous distributions. Neither conventional discrete nor continuous HMMs can be applied to F0 pattern modelling since F0 values are not defined in the unvoiced regions. The HMM-based SPSS system therefore uses multi-space probability distribution (MSD) distributions (Tokuda et al., 1999) for modelling F0. The voicing strengths for mixed excitation (Yoshimura et al., 2001) with continuous probability distribution.

All the steps of the voice training process in HTS are displayed in Figure 5.4. An exhaustive description of this process is outside the scope of this thesis. However,

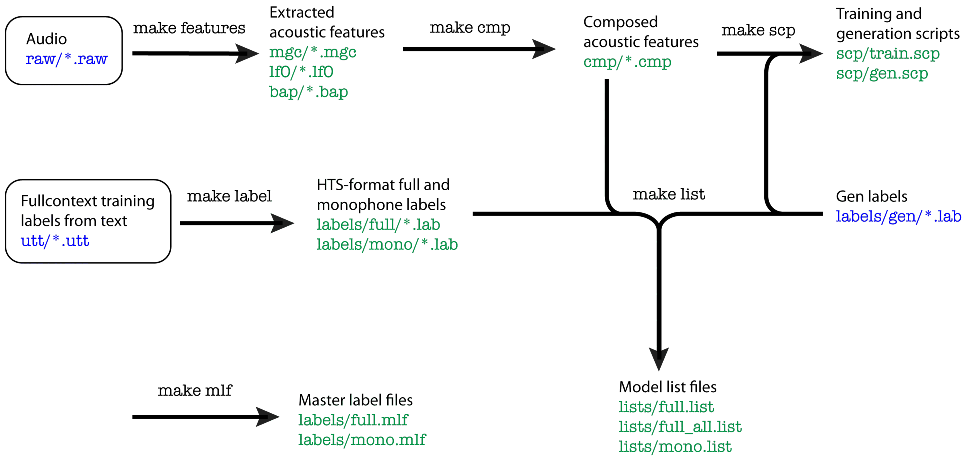


Figure 5.4: *Overview of all the individual analysis steps that are required to provide the parameters to train a new voice or synthesise a new utterance in HTS. Figure reproduced from (Cooper, 2019).*

it is important to emphasise that the process heavily relies on a good-quality text analysis (the *full-context training labels*), and alignment of speech and labels (the *master label files*).

All voices are trained with: a) speaker-dependent models, b) the default HTS parameter set, c) decision-tree based state clustering, and d) separate streams to model each of the static, delta and delta-delta features, e) single Gaussian models.

5.2.1 The English audio data

A female voice (*SLT*) and a male voice (*Nick*) are created for English (Nicolao et al., 2012; Nicolao and Moore, 2012b).

The *SLT* voice is trained on the 'CMU-ARCTIC SLT' corpus (Kominek and Black, 2003a), which consists of 1132 utterances spoken by an US English female. The corpus was recorded in a 16 bit 32 KHz format, in a sound proof room. The waveforms are stereo: one channel contains the actual acoustic waveform, the other has the electro-glotto-graph (EGG) signal (which is not used in the training). The database was automatically labelled using CMU Sphinx and the current FestVox labelling scripts. No hand correction has been made.

The English voice *Nick* is developed using speech data available from the LISTA Hurricane challenge (Cooke et al., 2012). The speech training corpus consists of about 3 hours of unmodified natural speech, spoken by a male British English speaker (*Nick*). Material consists of three different texts: 2023 newspaper style sentences, 300 sentences containing words from the modified rhyme test inserted in a carrier sentence (House et al., 1963), and 540 sentences from the Harvard corpus (Rothauser et al., 1969). The third corpus contains sentences that are arranged into phonemically-balanced subsets. Since the corpora are read speech by a highly-intelligible speaker, they can therefore be considered as intrinsically rather clear, i.e. hyper-articulated (Cooke et al., 2013a).

5.2.2 The Italian audio data

The two Italian voices used in the HTS-C2H experiments (Nicolao et al., 2013), are developed in collaboration with the Italian ISTC-CNR research institute (ISTC, 2019).

The female and male voices (*Lucia* and *Roberto*) are trained on two phonetically and prosodic balanced speech corpora, also recorded by ISTC-CNR.

For *Lucia* , the speech corpus consists of about 2 hours of material (approximately 1400 sentences), recorded by a non-professional female speaker with a Northern regional accent in a quasi-soundproof booth.

For *Roberto* , the voice is trained on a commercial corpus, available for research purposes. It contains about 3 hours (approximately 1900 sentences) of read speech, recorded by a professional male speaker in a quasi-soundproof booth.

Monophone and full context labels for both corpora are derived with the specific linguistic MaryTTS front-end for Italian text analysis, developed by ISTC-CNR (Tesser et al., 2013). HTK 3.4.1 (Young et al., 2002) toolkit is used to provide the phonetic alignment to the audio.

5.3 Parameter generation

The standard generation algorithm of HTS takes advantage of the Cholesky decomposition to increase the computational speed. As already expressed in § 3.2.4, this generation method allows fast computation, but eliminates the possibility of adapting the model parameters as the process unfolds. Therefore, an alternative generation method must be added to the HTS implementation of HTS-C2H.

The recursive search generation algorithm described in § 3.2.5 is therefore substituted for the standard HTS generation algorithm as it enables continuous manipulation of the generative models at each step of the generation process.

5.3.1 Speech parameter generation in HTS-C2H

The recursive search generation algorithm is adapted in the HTS-C2H implementation. In particular, the output feature distributions are modelled by single Gaussian functions, features are modelled with up to second order derivatives, and the effect of the delay is considered between processing the input linguistic features and the production of the output speech parameters.

The optimal sequence of states and mixtures $\{q_t, i_t\}$ with $1 \leqslant t \leqslant T$ should be chosen before parameter generation begins. To do this, the phone and the state-duration sequences must be computed independently from the acoustic features. For the phone sequence, an automatic text analyser such as the one in the Festival Speech Synthesis System (Clark et al., 2012) is normally used. For the state-duration sequence, the computation of the duration of each state is deduced with eq. 3.27 using an independently-trained statistical model. If the HMM comprehends multiple Gaussian mixtures, the mixture with the higher weight value is chosen.

The implemented algorithm assumes that the input state sequence is optimal. Thus, no state sequence optimisation is requested and the algorithm of Table 3.1 with the following substitutions,

$$\boldsymbol{\Sigma}_{q_t}^{-1} = \mathbf{0}_{3M\times 3M} \quad \text{and} \quad \boldsymbol{\mu}_{q_t} = \mathbf{0}_{3M} \tag{5.1}$$

$$\boldsymbol{\Sigma}_{\hat{q}_t}^{-1} = \boldsymbol{\Sigma}_t^{-1} \quad \text{and} \quad \boldsymbol{\mu}_{\hat{q}_t} = \boldsymbol{\mu}_t \tag{5.2}$$

can be changed into a simpler form. In this situation, eq. 3.48 and eq. 3.49 become

$$\boldsymbol{D} = \boldsymbol{\Sigma}_t^{-1} \tag{5.3}$$

$$\boldsymbol{d} = \boldsymbol{\Sigma}_t^{-1}\boldsymbol{\mu}_t \tag{5.4}$$

and the algorithm formula are those in Table 5.2.

Table 5.2: *Algorithm to compute the output feature* $\hat{\boldsymbol{c}}$*, given optimal state sequence.*

$\boldsymbol{\pi} = \boldsymbol{P}\boldsymbol{w}_t$	(ST.1)
$\boldsymbol{\nu} = \boldsymbol{w}_t^\top \pi$	(ST.2)
$\tilde{\boldsymbol{\kappa}} = \boldsymbol{\pi}\left\{\boldsymbol{I}_{3M} + \boldsymbol{D}\boldsymbol{\nu}\right\}^{-1}\boldsymbol{\Sigma}_t^{-1} = \boldsymbol{\pi}\left(\boldsymbol{I}_{3M} + \boldsymbol{\Sigma}_t^{-1}\boldsymbol{\nu}\right)^{-1}\boldsymbol{\Sigma}_t^{-1}$	(ST.3)
$\hat{\boldsymbol{c}} = \boldsymbol{c} + \tilde{\boldsymbol{\kappa}}(\boldsymbol{\mu}_t - \boldsymbol{w}_t^\top\boldsymbol{c})$	(ST.4)
$\hat{\varepsilon} = 0$ because no state optimisation is needed.	(ST.5)
$\hat{\boldsymbol{P}} = \boldsymbol{P} - \boldsymbol{\kappa}\boldsymbol{D}\boldsymbol{\pi} = \boldsymbol{P} - \tilde{\boldsymbol{\kappa}}\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\pi}$	(ST.6)

The computing is done for every n-order derivative with $0 \leq n \leq 2$, for each dimension m in $1 \leq m \leq M$ and for each t in $1 \leq t \leq T$. The number of elements used to compute the dynamic features with eq. 3.17 are $L^0 = 0$, $L^1 = L^2 = 1$ and the window coefficients:

$$\begin{aligned} \boldsymbol{\omega}^{(0)} &= 0 \\ \boldsymbol{\omega}^{(1)} &= [-1/2 \;\; 0 \;\; 1/2] \\ \boldsymbol{\omega}^{(2)} &= [1 \;\; -2 \;\; 1] \end{aligned} \tag{5.5}$$

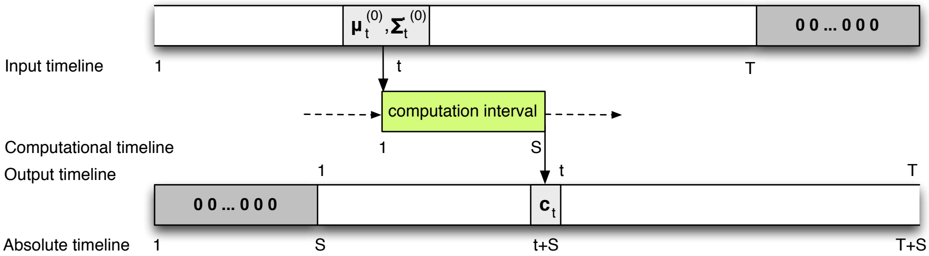


Figure 5.5: *Diagram of the relationship between the four different timelines (*input timeline*,* output timeline*,* computational timeline *and* absolute timeline*) in the implementation of the recursive algorithm.*

Ideally, T would be ∞ because the input stream could be endless, but, practically, the number of influenced neighbouring vectors is limited to S. So, inside the computation algorithm, the timeline, where range is $[1, T]$, is reduced to a small portion, $[1, S]$, as shown in Figure 5.5.

At input time t, the frame of interest is reinitialised, according to eq. 3.53 and eq. 3.52 to have:

$$\begin{aligned} c_{s=1,m} &= \mu_{t,m}^{(0)}, && \text{mean of the static features at state } q_t \\ P_{0,s=1,m}^{(0)} &= \Sigma_{t,m}^{(0)}, && \text{covariance of the static features at state } q_t \\ P_{0,s=1,m}^{(n)} &= 0, && \forall n \in [1,2] \end{aligned} \tag{5.6}$$

$\forall m \in [0, M-1]$. Thus, the matrices and the vectors in Table 5.2 can be written as:

$$\boldsymbol{\pi} = \{\pi^{(n)}_{u,m}\} \quad \text{and } \pi^{(n)}_{u,m} = \sum_{j=w_l}^{w_r} P^{(n)}_{u-j,s+j,m} w^{(n)}_j \tag{5.7}$$

$$\boldsymbol{\nu} = \{\nu^{(n)}_{m}\} \quad \text{and } \nu^{(n)}_{m} = \sum_{j=w_l}^{w_r} w^{(n)}_j \pi^{(n)}_{j,m} \tag{5.8}$$

$$\tilde{\boldsymbol{\kappa}} = \{\tilde{\kappa}^{(n)}_{u,m}\} \quad \text{and } \tilde{\kappa}^{(n)}_{u,m} = \pi^{(n)}_{u,m} \frac{\Sigma^{(n)}_{s,m}}{1 + \Sigma^{(n)}_{s,m} \nu^{(n)}_{m}} \tag{5.9}$$

$$\hat{\boldsymbol{P}} = \{P^{(n)}_{u,s,m}\} \quad \text{and } \hat{P}^{(n)}_{v-u,s,m} = \hat{P}^{(n)}_{u-v,s,m} = \tilde{\kappa}^{(n)}_{v,m} \pi^{(n)}_{u,m} \tag{5.10}$$

$$\hat{\boldsymbol{c}} = \{c_{s,m}\} \quad \text{and } \hat{c}_{s+u,m} = c_{s+u,m} + \tilde{\kappa}^{(n)}_{u,m} \left(\mu^{(n)}_{s,m} - \sum_{j=w_l}^{w_r} w^{(n)}_j c^{(n)}_{s+j,m} \right) \tag{5.11}$$

$\forall s \in [1, S]$, $\forall u \in [-R, w_r], \forall m \in [0, M-1]$ and with $w_r = L^{(n)}$ and $w_l = -L^{(n)}$. The first S steps of the algorithm are needed to fill the $\boldsymbol{P}$ matrix with the initialisation values. After these steps, the oldest vector $\boldsymbol{c}_{s=S}$ is ready to be sent to the output. All the elements in the vector $\boldsymbol{c}$ and the matrix $\boldsymbol{P}$ are shifted one step ahead ($c_{s+1,m} = c_{s,m}$ and $P_{u,s+1,m} = P_{u,s,m}$ $\forall s$ and $\forall m$) and the place left empty at the beginning of the vectors is filled with the parameters of next input model as in eq. 5.6. The optimised feature vector can thus be given to the output after S iterations from when the model parameters were first given to the input.

Though he computational complexity of the algorithm is initially $\mathcal{O}(T^2M^3)$, this reduces to $\mathcal{O}(T^2M)$, when Σ_{q_t} and $\Delta\Sigma_{q_t}$ are diagonal.

5.4 Phonetic contrast adaptation

The recursive generation algorithm of § 5.3.1 allows different acoustic model to be selected at each step. This enables each model to be manipulated before it is used in the algorithm. A strategy to train the phonetic-contrast motivated adjustments of § 4.2.3 is proposed next.

5.4.1 Adaptation of speech models

In C2H, a transform must be trained to map normal phone realisations onto low-contrast (LC) attractor acoustics, as per § 4.2.3. The resulting transformation is used to generate new unseen speech. This transform should hypothetically reduce the degree of perceived articulation of the synthetic speech. If the hypothesis is correct, the outcome of the adaptation is fully hypo-articulated (HYO) speech. Moreover, the inverse transformation should give rise to hyper-articulated (HYP) speech.

The required adjustments of acoustic and duration models are obtained using MLLR (cf. § 3.2.1). This technique is normally used to adapt HMM models to render new speaker identities (Yamagishi et al., 2009).

The use of linear transformation such as MLLR or constrained maximum likelihood linear regression (CMLLR) is crucial as their linearity allows to easily scale the intensity of the transformation. Moreover, MLLR computational cost is relatively low, thus the adaptation during the generation process can be done without a big effect on performance.

However, a procedure needs to be developed to identify the direction – i.e. MLLR parameters – along which the transform vector can be moved. In the following paragraphs, the training and the scaling of such transformations are described.

MLLR training with data augmentation

The MLLR transformation is estimated using a relatively small corpus of synthetic hypo-articulated speech. Two approaches are possible: collect *new specifically-recorded speech material* (Picart et al., 2014; Raitio et al., 2013) or *artificially generated new audio* (Nicolao et al., 2012; Nicolao and Moore, 2013b). The former approach is unsuitable for achieving the transformation of § 4.2.3, as the extremely unrealistic type of speech that is required from the speaker would be difficult to deliver. The latter approach is hence chosen to train the MLLR transform. This can be regarded as a data augmentation approach (Ragni et al., 2014). Data augmentation is often used to train speech recognition model for low-resource languages. It aims to increase the quantity of training data. This approach has an important theoretical advantage of being able to produce data when real examples are not available.

The artificially generated new audio consists of synthetic speech generated using HTS with input control sequences forced to have only LC attractors in them. The HTS synthesiser uses the speech models that are normally trained on the data set of § 5.2. Using decision-tree based clustering, HTS finds the most likely context-dependent acoustic model for all of the LC attractors, even those unseen in the original training corpora.

In detail, the speaker-dependent training procedure consists of the following steps:

1. Train speaker-dependent speech models using audio and transcription labels from standard training data set.

2. Substitute selected phone labels with related LC attractors in the input linguistic-analysis files.

3. Generate a hypo-articulated LC version of the original training corpus. No constraints on the duration of these realisations are set.

4. Train the MLLR adaptation of acoustic and duration models to match the new characteristics of the hypo-articulation corpus.

5. Repeat step 1-4 for each class of phones for which LC attractors need to be differentiated: e.g., vowels and consonants.

The MLLR transform that is obtained consists of a set of context-specific functions that modify the parameters of the context-dependent pdfs. The MLLR transform normally operates on the Gaussian model mean vectors only, but the covariance components can also be adapted. The resulting mean vectors and covariance matrices for each i-th HMM pdf $\{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}$ in the model can be written as, can be written as:

$$\boldsymbol{\mu}'_i = \boldsymbol{A}_i \boldsymbol{\mu}_i + \boldsymbol{b}_i \tag{5.12}$$

$$\boldsymbol{\Sigma}'_i = (\boldsymbol{H}_i^{-1})^{\intercal} \boldsymbol{g}_i \boldsymbol{H}_i^{-1} \tag{5.13}$$

where $\boldsymbol{H}_i \boldsymbol{H}_i^{\intercal} = \boldsymbol{\Sigma}_i^{-1}$ and the $P \times P$ matrix $\boldsymbol{A}_i$, and the $P \times 1$ vectors $\boldsymbol{b}_i$ and $\boldsymbol{g}_i$ are the transformation parameters. P is the size of the static feature vector plus, in this case, its n-order derivatives.

Low-contrastive reference generation

Figure 5.6 shows the functional diagram of the data augmentation procedure used to create the reference corpus to train the transformation parameters. Starting with the set of full-context labels (L0) used to build the standard HMM-based voice, a low-contrastive version of the labels (L1) is obtained through a phonetic transformation. L1 labels are used to generate the acoustic features (P1) representing the LC acoustic space. The most likely models for unseen-context phones are selected from the standard HMM models using decision-tree clustering. The time-aligned version of L1 (LA1) is mapped back into the standard phonetic domain (LA0). These labels along with the target generated parameters (P1), are used as reference to estimate the MLLR transform.

Phone substitutions, used to create the reference augmented data, are language dependent, as expressed in § 4.2.3. LC attractors in English are selected to be the

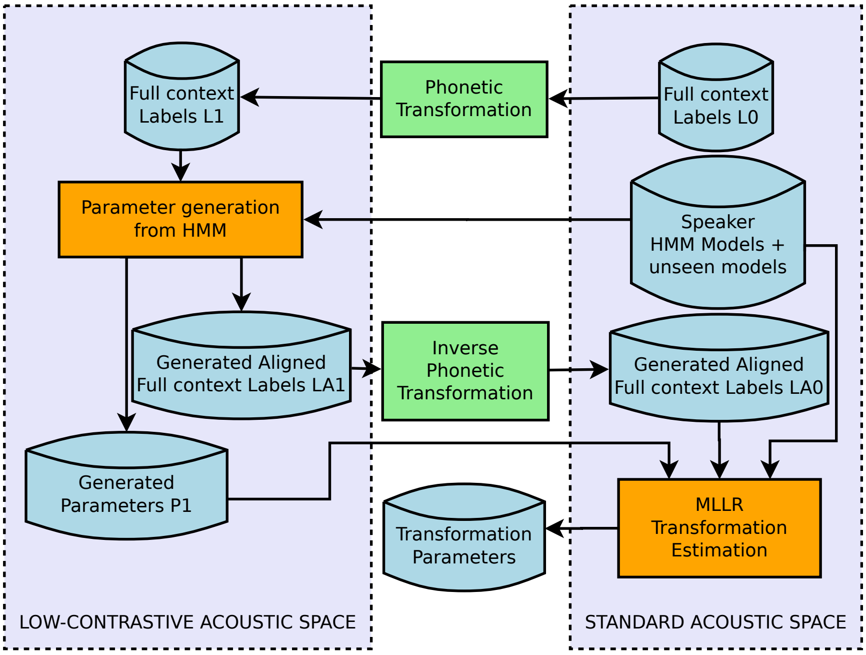

Full context Labels L1
Phonetic Transformation
Full context Labels L0
Parameter generation from HMM
Speaker HMM Models + unseen models
Generated Aligned Full context Labels LA1
Inverse Phonetic Transformation
Generated Aligned Full context Labels LA0
Generated Parameters P1
MLLR Transformation Estimation
Transformation Parameters
LOW-CONTRASTIVE ACOUSTIC SPACE
STANDARD ACOUSTIC SPACE

Figure 5.6: *Schematic diagram of the data augmentation preparation. Figure reproduced from (Nicolao et al., 2013).*

schwa phone [ə] for vowels, and the highest confusable competitor for consonants (Miller and Nicely, 1955). LC attractors in Italian are the competing phones that are more likely to be mistaken in adverse conditions. Vowel competitors are the opposite phones with respect to F1-F2 chart. Similarly to (Miller and Nicely, 1955), Caldognetto (Caldognetto et al., 1988) listed the Italian consonants that are undistinguishable in noise. Following these guidelines, confusable consonant pairs are chosen to be the contrastive pairs. Another set of consonant pairs is motivated by the contrastive use of gemination in Italian. Gemination is the consonant lengthening that differentiate two homophonic words. Geminated consonants are mapped into the corresponding non-geminated ones in order to increase the phonetic contrast.

A summary of the contrastive pairs used in the English and Italian MLLR training are displayed in Table 5.3 and Table 5.4 respectively.

Table 5.3: *Vowel and consonant mapping in English. STD column contains the original phones and CTR has the contrastive ones.*

STD		CTR	STD		CTR	STD		CTR
[a]	→	[ə]	[ʊ]	→	[ə]	[v]	→	[f]
[ə]	→	[ə]	[u]	→	[ə]	[θ]	→	[ð]
[ɑ]	→	[ə]	[ʏ]	→	[ə]	[ð]	→	[θ]
[ɔ]	→	[ə]	[w]	→	[ə]	[s]	→	[z]
[aʊ]	→	[ə]	[p]	→	[b]	[z]	→	[s]
[ə]	→	[ə]	[b]	→	[p]	[ʃ]	→	[ʒ]
[aɪ]	→	[ə]	[t]	→	[d]	[ʒ]	→	[ʃ]
[ɛ]	→	[ə]	[d]	→	[t]	[h]	→	[h]
[ə]	→	[ə]	[k]	→	[g]	[m]	→	[n]
[ɪ]	→	[ə]	[g]	→	[k]	[n]	→	[m]
[i]	→	[ə]	[tʃ]	→	[dʒ]	[ŋ]	→	[n]
[oʊ]	→	[ə]	[dʒ]	→	[tʃ]	[l]	→	[r]
[ɔɪ]	→	[ə]	[f]	→	[v]	[r]	→	[l]

Table 5.4: *Vowel and consonant mapping in Italian. STD column contains the original phones and CTR has the contrastive ones. Geminate consonants are not listed, but they are mapped to the corresponding non-geminate ones.*

STD		CTR	STD		CTR	STD		CTR
[a]	→	[u]	[f]	→	[p]	[dz]	→	[dʒ]
[e]	→	[o]	[t]	→	[k]	[dʒ]	→	[dz]
[i]	→	[ɔ]	[k]	→	[t]	[g]	→	[dʒ]
[o]	→	[e]	[ts]	→	[s]	[z]	→	[g]
[u]	→	[a]	[s]	→	[ts]	[l]	→	[ʎ]
[ɛ]	→	[o]	[tʃ]	→	[s]	[ʎ]	→	[l]
[ɔ]	→	[e]	[ʃ]	→	[tʃ]	[m]	→	[n]
[j]	→	[ɔ]	[b]	→	[d]	[n]	→	[m]
[w]	→	[a]	[d]	→	[b]	[ɟ]	→	[m]
[p]	→	[f]	[v]	→	[b]	[r]	→	[m]

5.5 Controlling the phonetic contrast

The generative speech model of HTS-C2H can be controlled at each step with the recursive generation algorithm of § 5.3.1. To achieve this, a method to control the phonetic-contrast motivated adjustments of § 5.4 must be implemented.

The control mechanism described here operates on the MLLR transform. It aims to scale the magnitude of the adaptation that is applied to the speech models. As hypothesised in § 4.2.3, the scaling of the transform magnitude is proportional to the degree of articulation, and hence to the speech production effort.

5.5.1 Scaling the MLLR transform

One of the important characteristic of MLLR transformation is its possibility to be scaled with different strength. The previously trained MLLR transform identifies the functions that change the acoustic space descriptors from the source $\{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}$ to the adapted space $\{\boldsymbol{\mu}'_i, \boldsymbol{\Sigma}'_i\}$ where i is the acoustic model index.

The transformation vectors $\{\boldsymbol{v}_i\}$ that move the mean values of the i-th GMM to the adapted space can be written as:

$$\boldsymbol{v}_i = \boldsymbol{\mu}'_i - \boldsymbol{\mu_i} \tag{5.14}$$

Combining eq. 5.14 and eq. 5.12, the $\boldsymbol{v}_i$ becomes:

$$\boldsymbol{v}_i = (\boldsymbol{A}_i - \boldsymbol{I})\boldsymbol{\mu}_i + \boldsymbol{b}_i \tag{5.15}$$

where $\boldsymbol{A}_i$ and $\boldsymbol{b}_i$ are the parameters of the MLLR for the i-th pdf, and $\mathbf{I}$ is the identity matrix.

Finally, the scaled mean vector, $\boldsymbol{\mu}_i^{\alpha^h}$, with the weighting factor $\alpha^h \geqslant 0$, can be expressed as the partial movement of $\boldsymbol{\mu}_i$ towards the target $\boldsymbol{\mu}'_i$:

$$\boldsymbol{\mu}_i^{\alpha^h} = \boldsymbol{\mu}_i + \alpha^h \cdot \boldsymbol{v}_i \tag{5.16}$$

Given the hypo-articulation transform parameters $\boldsymbol{\mu}'_i$ and $\boldsymbol{\Sigma}'_i$ as per eq. 5.12 and eq. 5.13, the scaled mean vector, $\boldsymbol{\mu}_i^{\alpha^h}$ and covariance matrix $\boldsymbol{\Sigma}_i^{\alpha^h}$ are computed as:

$$\boldsymbol{\mu}_i^{\alpha^h} = \boldsymbol{\mu}_i + \alpha^h(\boldsymbol{\mu}'_i - \boldsymbol{\mu_i}) = \alpha^h\boldsymbol{\mu}'_i + (1 - \alpha^h)\boldsymbol{\mu_i} \tag{5.17}$$

$$\boldsymbol{\Sigma}_i^{\alpha^h} = \alpha^h\boldsymbol{\Sigma}'_i + (1 - \alpha^h)\boldsymbol{\Sigma}_i = (\boldsymbol{H}_i^{-1})^\intercal(\alpha^h\boldsymbol{g}_i - \alpha^h + 1)\boldsymbol{H}_i^{-1} \tag{5.18}$$

When the transform is trained on contrastive pairs, the MLLR is applied with diminished strength (usually 50%) to the HMM models to reduce standard (STD) synthetic speech to the actual low-contrastive (LC) configuration, the mid-point between competitor acoustic realisations.

5.5.2 Linear transform inversion

In § 4.2.3, it is stated that the transformation towards hyper-articulated speech is hypothesised to be the inverse of the trained adaptation towards LC attractors: same domain, same orientation, but opposite sense. Therefore the transformation towards hyper-articulated speech could be regarded as the inverse of the MLLR transform vector defined by eq. 5.14. The vector $\boldsymbol{v}_i$, that defines the transformation orientation and sense towards hypo-articulated space, can then be inverted:

$$-\boldsymbol{v}_i = \boldsymbol{\mu}_i - \boldsymbol{\mu}'_i \tag{5.19}$$

Eq. 5.16 becomes

$$\boldsymbol{\mu}_i^{\alpha^h} = \boldsymbol{\mu}_i + \alpha^h \cdot (-\boldsymbol{v}_i), \qquad \text{with } \alpha^h \geq 0 \tag{5.20}$$

The inversion method transfers the negative sign of the vector to the weighting factor, eq. 5.16 can be applied by changing sign to α^h, $\alpha^H = -\alpha^h$.

$$\boldsymbol{\mu}_i^{\alpha^H} = \boldsymbol{\mu}_i + \alpha^H \cdot \boldsymbol{v}_i, \qquad \text{with } \alpha^H \leq 0 \tag{5.21}$$

For example, in the case of the same-magnitude inverse transform ($\alpha^H = -1$), eq. 5.21 becomes

$$\boldsymbol{\mu}_i^{-1} = -(\boldsymbol{A}_i + 2\boldsymbol{I})\boldsymbol{\mu}_i - \boldsymbol{b}_i \tag{5.22}$$

which means that the weighting factor should move towards negative values. Since eq. 5.17 and eq. 5.18 are still defined for those range of values, a continuum in which the formula exists is identified. From this point onwards, α^h and α^H will therefore be simply addressed as α and (5.17) and (5.18) can be written as

$$\boldsymbol{\mu}_i^{\alpha} = \boldsymbol{\mu}_i + \alpha(\boldsymbol{\mu}'_i - \boldsymbol{\mu_i}) = \alpha\boldsymbol{\mu}'_i + (1-\alpha)\boldsymbol{\mu_i} \tag{5.23}$$

$$\boldsymbol{\Sigma}_i^{\alpha} = \alpha\boldsymbol{\Sigma}'_i + (1-\alpha)\boldsymbol{\Sigma}_i = (\boldsymbol{H}_i^{-1})^{\mathsf{T}}(\alpha\boldsymbol{g}_i - \alpha + 1)\boldsymbol{H}_i^{-1} \tag{5.24}$$

and they are defined $\forall\alpha \in \mathbb{R}$. The magnitude of the inverted transformation towards hyper-articulation, in principle, has no phonetically-motivated constraints. However, if the transformed acoustic models are too far from the STD realisations, the feature domain becomes under-trained, and the synthetic speech sounds unrealistic. A discussion about the range of values for α and the effects when $\alpha \geqslant 0$ (hypo-) and $\alpha \leqslant 0$ (hyper-articulation) can be found in Chapter 6.

In conclusion, adaptations are trained such that both the forward and the inverse transformations can be applied with different magnitude, representing different operating points along the derived H&H axis. Proofs of such scaling effectiveness for the range of α values are shown in Chapter 6.

5.5.3 The control of α

The parameter α appears to be the critical element to control the strength of the hypo/hyper-articulation transform. Understanding how to handle this parameter is therefore of utmost importance. The transform scaling factor α is by definition time-varying, α_t, as it has to change frame-by-frame to follow the error signal.

The value of α_t is defined as a function of two main factors: the *motivation* m from the conceptualiser, and the *error* signal e from the comparator and perception feedback loop. This is,

$$\alpha_t = \mathcal{F}(m, e_t) \tag{5.25}$$

As previously expressed, motivation is assumed to be a fixed value, which describes the "attitude" of the synthesis system to overcome language barriers. It indicates the maximum amount of effort that can be involved in the process. Thresholds can be modelled to limit the range of α_t, i.e., the effort amplitude. The thresholds, α_{HYO} and α_{HYP}, are selected as the maximum HYO and the minimum HYP values respectively, with which the transform can operate. Hence, $\mathcal{F}(.)$ range can be limited either globally,

$$\alpha_{\text{HYP}} \leq \mathcal{F}(m, e_t) \leq \alpha_{\text{HYO}} \tag{5.26}$$

or locally,

$$\alpha_{\text{HYP}} \leq \frac{1}{B} \sum_{t-B}^{t} \mathcal{F}(m, e_t) \leq \alpha_{\text{HYO}} \tag{5.27}$$

where B is the length of the observation window that it is used to limit the $\mathcal{F}(.)$ values.

The α_t range is determined by the intrinsic qualities of the trained adaptation, but it is also a function of the motivation m. This factor can be included in the boundary constants,

$$\begin{aligned} \alpha_{\text{HYO}}^m &= m \cdot \alpha_{\text{HYO}} \\ \alpha_{\text{HYP}}^m &= m \cdot \alpha_{\text{HYP}} \end{aligned} \tag{5.28}$$

where $m \in [0, 1]$. The α_t range amplitude is then controlled by m, and it is $[\alpha_{\text{HYP}}^m, \alpha_{\text{HYO}}^m]$.

In the implementation of the HTS-C2H adaptation, the output of the perception loop is the principal component of the error signal. The error is defined as the difference between the measured and the intended intelligibility, eSII$_t$ and eSII$_{\text{intent}}$ respectively:

$$e_t = \text{eSII}_{\text{intent}} - \text{eSII}_{\text{perception}} = \text{eSII}_{\text{intent}} - \text{eSII}_t$$

where eSII_t is the *perceived intelligibility* and $\text{eSII}_{\text{intent}}$ indicates the *intended intelligibility*. The latter is assumed to be fixed for the whole duration of the synthesis, and it is derived from the conceptualiser intent.

Both the constant term, $\text{eSII}_{\text{intent}}$, and the observation term, eSII_t, vary between 0 and 1. Therefore, their difference e_t is in the range $[-1, 1]$. When e_t is positive, the intelligibility is not high enough. Hence, the degree of articulation must be increased to increase the communicative clarity (HYP, $\alpha_t \leq 0$). Negative e_t indicates that the speech is estimated to be clearer than is intended. The degree of articulation can therefore be reduced to reduce the system effort (HYO, $\alpha_t \geq 0$).

The parameters α_t^h and α_t^H of eq. 5.16 and eq. 5.21 are both reduced to α_t and are expressed as

$$\begin{aligned} \alpha_t &= \mathcal{F}(m, e_t) \\ &= \mathcal{F}(\alpha^m_{\text{HYO}}, \alpha^m_{\text{HYP}}, \text{eSII}_{\text{intent}}, \text{eSII}_t) \\ &= \mathcal{F}'(e_t) \end{aligned} \tag{5.29}$$

Since α^m_{HYO}, α^m_{HYP}, and $\text{eSII}_{\text{intent}}$ are constants defined by the conceptualiser, $\mathcal{F}(.)$ can be written $\mathcal{F}'(.)$ and the only variable is the time-varying measure, eSII_t, from the perception loop encapsulated in e_t .

Several mapping functions $\mathcal{F}'(.)$ can be adopted to map the error into the scalar control parameter, as long as these functions:

a) are monotonic,

b) allow some tolerance at the boundaries by mapping the error range $[-1, 1]$ into $[\alpha^m_{\text{HYP}} - \epsilon, \alpha^m_{\text{HYO}} + \epsilon]$ with $\epsilon \to 0$,

c) $\mathcal{F}'(e_t) < 0$, if $e_t > 0$ (HYP transform),

d) $\mathcal{F}'(e_t) > 0$, if $e_t < 0$ (HYO transform),

e) $\mathcal{F}'(e_t \approx 0.8) = \alpha^m_{\text{HYP}}$ (maximum HC configuration),

f) $\mathcal{F}'(e_t \approx -0.8) = \alpha^m_{\text{HYO}}$ (maximum LC configuration),

g) $\mathcal{F}'(e_t \approx 0) = 0$.

In HTS-C2H implementation and its relative experiments, a heuristic incremental function, inspired by PDI control principle, is chosen. It allows for small over-LC ($\alpha_t \geq \alpha^m_{\text{HYO}}$) and over-HC ($\alpha_t \leq \alpha^m_{\text{HYP}}$) configurations. It can be expressed as

$$\alpha_t = \mathcal{F}'(e_t) = \begin{cases} \alpha_{t-1} + \Delta\alpha_t & \alpha^m_{\text{HYP}} \leq \alpha_{t-1} \leq \alpha^m_{\text{HYO}} \\ \alpha_{t-1} & \text{elsewhere} \end{cases} \tag{5.30}$$

where $\alpha_0 = 0$ (no adaptation), and

$$\Delta\alpha_t = -0.1 \cdot \left(\frac{2}{1 + e^{-9 \cdot e_t^3}} - 1 \right) \tag{5.31}$$

The function of eq. 5.30 ensures that the α_t is always contained within the defined boundaries. The incremental component curve of eq. 5.31 is displayed in Figure 5.7 for the $[-1, 1]$ domain interval of e_t.

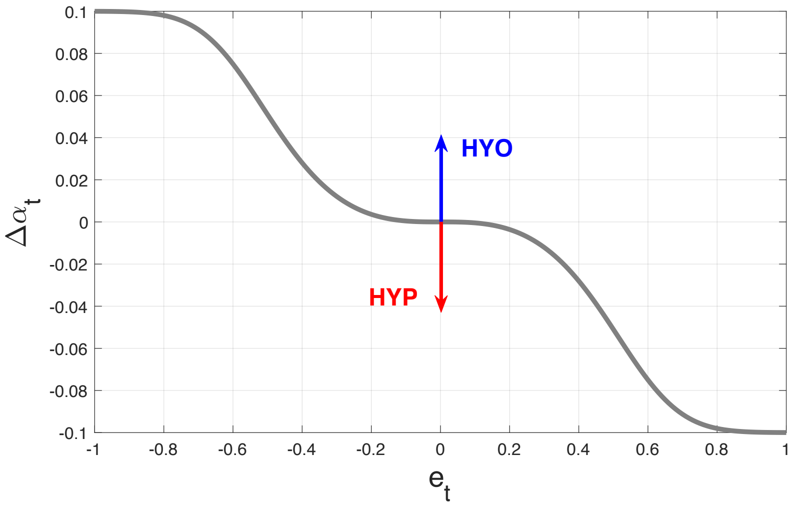


Figure 5.7: *Incremental component $\Delta\alpha_t$ of $\alpha_t = \mathcal{F}'(e_t)$. The HYO and HYP directions from the equilibrium position are also displayed.*

5.6 Monitoring speech quality

One of the key features in the C2H model is the perception feedback loop, that was discussed in § 4.3. The aim of this part of the model is to evaluate the synthetic speech quality.

Speech quality can refer to different dimensions of speech communication. The C2H comparator (cf. § 4.4.1), ideally requires a broad spectrum of speech quality measures, such as the audibility of the speech signal, clarity of the content, expressiveness, etc. However, some of these dimensions are not easily implementable and are themselves the topic of extensive research. In the experiments of this thesis, the assessment is limited to the *intelligibility* dimension, as it allows an automatic, low-resources, effective measure of how speech interacts with the environment.

Automatic methods for intelligibility assessment can here be effectively regarded as computational model of the listener.

5.6.1 Extended SII

The extended speech intelligibility index (eSII) (Rhebergen and Versfeld, 2005; Rhebergen et al., 2006) is an implementation of the SII, discussed in § 4.3.4, that allows intelligibility indices to be computed on a short-term window of the signal.

Standard SII requires that both clear speech and noise signals are known for the whole utterance duration to return an average intelligibility value. This model in its present form (ANSI, 1997) accurately describes intelligibility for speech in stationary noise but fails to capture the effects due to non-stationary disturbances. The extension to the model, eSII, has been proposed to predict speech intelligibility in both stationary and fluctuating noise. The basic principle of the extended approach is that both the speech and noise signals are partitioned into small time windows. Within each time window, the conventional SII is computed, yielding the speech information available to the synthesis framework at that instant. If the eSII values of these windows are averaged, the overall result remains coherent with the conventional SII value for that particular condition.

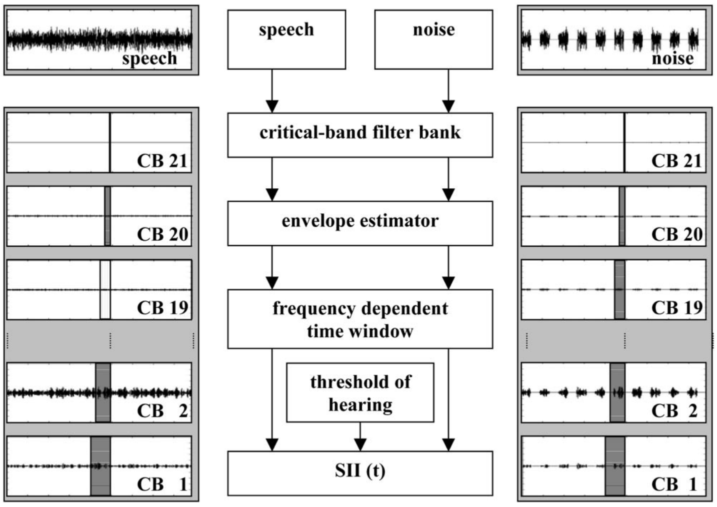


Figure 5.8: *Schematic overview of the calculation scheme for the eSII model. Figure reproduced from (Rhebergen and Versfeld, 2005).*

A block diagram of the calculation scheme of the eSII (Rhebergen and Versfeld, 2005) is presented in Figure 5.8 and can be described as follows:

1. the input speech signal and the input noise are separately filtered by a 21-critical-band (CB) filter bank;

2. the envelope of the input speech and noise are estimated in every CB; the instantaneous intensity is estimated in a frequency-dependent time window, as indicated by the shaded bars in Figure 5.8;

3. the window length is chosen to be relatively short in the higher bands and relatively long in the lower bands (CB1 = 35 ms to CB21 = 9.4 ms);

4. every 9.4 ms an eSII_t is calculated as described by (ANSI, 1997);

5. for each 9.4-ms steps, the instantaneous eSII_t is determined;

6. the ANSI SII for that speech-in-noise condition is determined by averaging across all instantaneous eSII_t values.

The eSII recipe that requires SII value for that particular speech-in-noise condition is calculated by averaging across all instantaneous eSII_t values, to determine the. However, in HTS-C2H, the instantaneous eSII values are used directly as the perception measurement that is returned to the comparator so that the distance from the communicative intent can be compared.

The eSII in HTS-C2H also slightly differs from the original because the filtering of step 1 and the envelope computation of step 2 are done in the frequency domain rather than in the time domain. These modifications do not affect the quality of the intelligibility estimation however, as there is a direct correspondence between the generated spectral features and the spectrum of the signal. Finally, the incremental step is also changed from 9.4 ms to 10 ms in the HTS-C2H implementation to synchronise the intelligibility assessment and the feature generation processes.

In Figure 5.9, an example of eSII estimation computed with the HTS-C2H perception loop is displayed for a speech audio in noise.

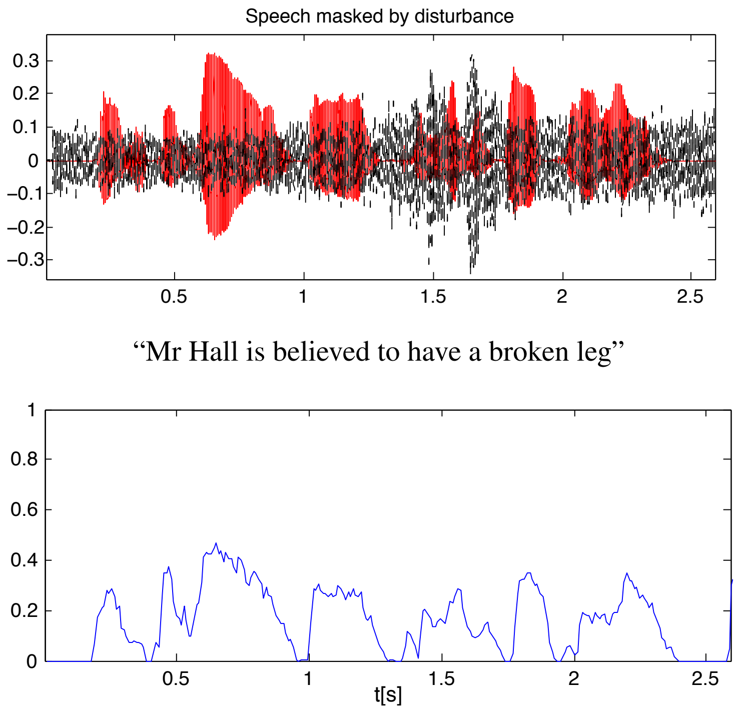


Figure 5.9: *Example of eSII computation for a synthetic speech signal (red waveform) in noise (black waveform). The corresponding eSII is displayed in the lower plot.*

5.7 The HTS-C2H complete framework

The proposed implementation of C2H, HTS-C2H, is summarised in this section, and connections between the functional blocks described in the above sections are highlighted.

HTS-C2H uses an HMM-based synthesiser, HTS, with regression tree model clustering. The standard generation algorithm is integrated with a specific recursive method that was illustrated in § 5.3.1.

Figure 5.10 shows the detailed functional diagram of the context-aware speech synthesiser implemented in this thesis with recursive short-latency system waveform generation and eSII perception loop.

Some set-up data is required to configure a synthesis process with HTS-C2H. In Figure 5.10, intent components such as the *text message*, the *effort level*, and the *required intelligibility* level are given to the system to produce the *overt speech* waveform. In this representation of the synthesis process, duration models are not

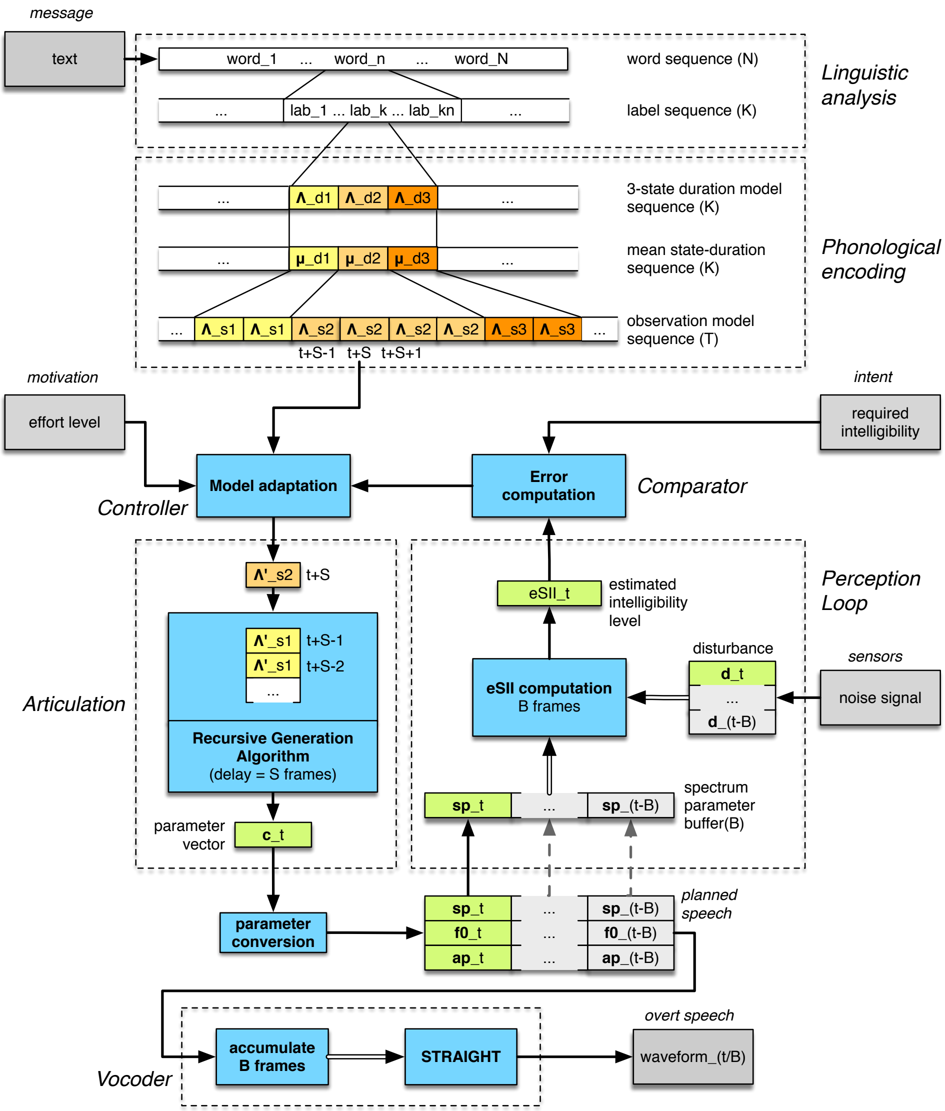


Figure 5.10: *The complete speech production process with HTS-C2H, using the recursive generation algorithm and eSII-based perception loop. Gray blocks represent input and output data and functional blocks are in blue. The delay S, deriving from the recursive generation, and the maximum waveform size B, determined by the output buffer size, are highlighted.*

adapted. The algorithm is synchronised with the generation output, as it is the speech realisation window that needs to contrast the disturbance at time t.

Each functional step of Figure 5.10 is magnified in the following figures.

Linguistic analysis translates the input text in a sequence of N words, and eventually into a sequence of K labels. Each label is described by a 3-state HMM. First, the standard (STD) duration model pdfs are used to generate the most likely label duration sequence. The label list is expanded according to each duration to create the sequence of T acoustic models that generate the speech parameters. Linguistic analysis is displayed in Figure 5.11.

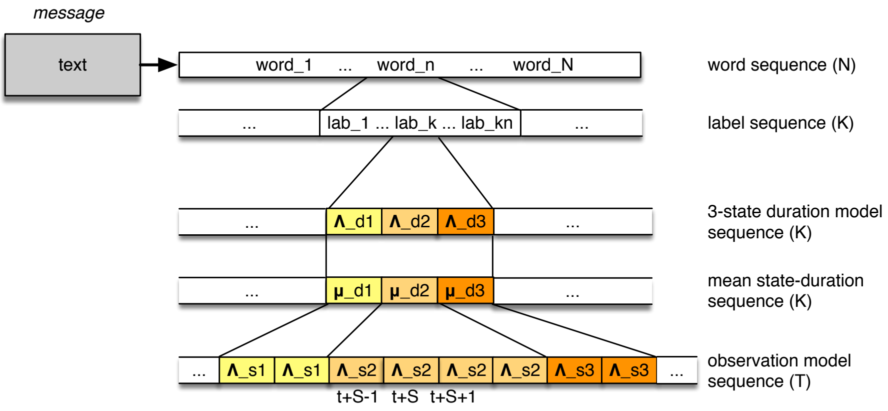

message
text
word_1 ... word_n ... word_N
word sequence (N)
... lab_1 ... lab_k ... lab_kn ...
label sequence (K)
... Λ_d1 Λ_d2 Λ_d3 ...
3-state duration model sequence (K)
... μ_d1 μ_d2 μ_d3 ...
mean state-duration sequence (K)
... Λ_s1 Λ_s1 Λ_s2 Λ_s2 Λ_s2 Λ_s2 Λ_s3 Λ_s3 ...
observation model sequence (T)
t+S-1 t+S t+S+1

Figure 5.11: *Block diagram of the HTS-C2H linguistic analysis and phonological encoding.*

At the same time, the perception feedback of HTS-C2H assesses the *intelligibility level* of the already-produced speech, see Figure 5.12. The spectrum parameter vectors of the latest B frames are contrasted against the measured environmental disturbance with the eSII algorithm, see § 5.6.1.

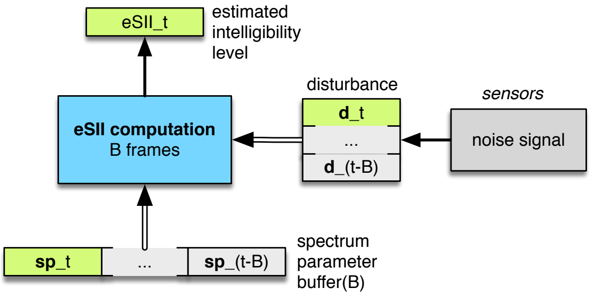


Figure 5.12: *Block diagram of the HTS-C2H perception feedback loop. B is the size of the window analysis that is used to compute the intelligibility index.*

The acoustic model sequence $\{\Lambda_k\}$ is adapted by the *controller* using the MLLR transform that controls the degree of speech articulation, as shown in Figure 5.13. The error, measured by the *comparator* between the estimated and the *required* intelligibility levels subsequentially determines the magnitude of the transform.

The *effort level* is also a factor that determines the sign and the value of α (cf. § 5.5.1).

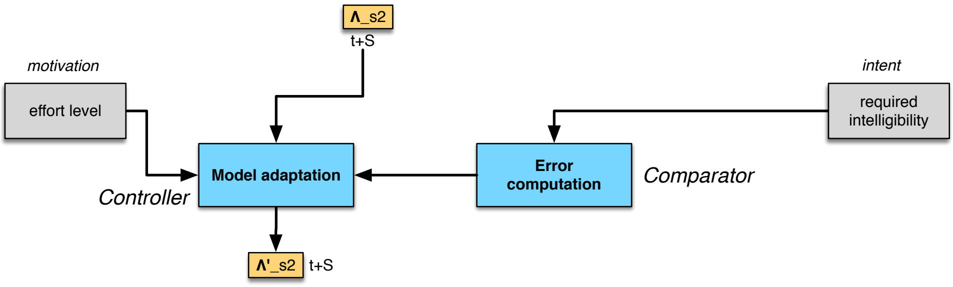


Figure 5.13: *Block diagram of the HTS-C2H controller and comparator.*

The adapted acoustic model sequence $\{\Lambda'_k\}$ is used by the recursive generation algorithm to produce the speech parameters c_t, as shown in Figure 5.14. A delay S is introduced by this process. The phonological encoding therefore needs to produce the acoustic models for the $[t, \ldots, t+S]$ frames *in advance* in order to enable the generation of the speech parameters at time t.

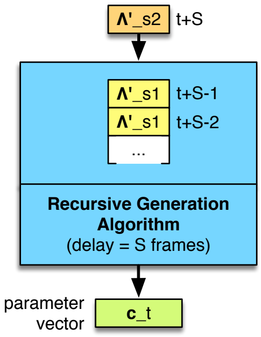


Figure 5.14: *Block diagram of the HTS-C2H waveform generation algorithm. The effect of the delay S, originating from the recursive generation algorithm, is considered.*

Speech parameters c_t are converted into the three components of the STRAIGHT analysis (spectrum, fundamental frequency, and a-periodic parameters) as discussed in § 5.1.2. These are used by the *vocoder* to produce the final waveform, as shown in Figure 5.15.

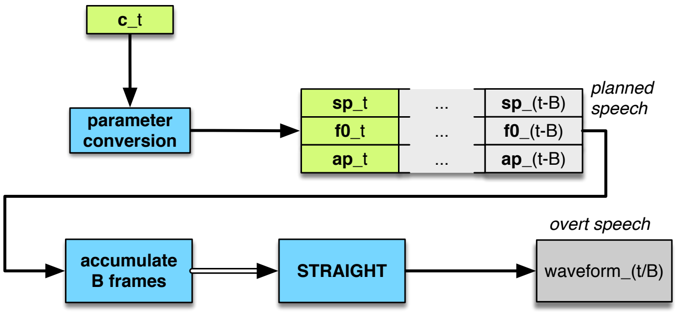


Figure 5.15: *Block diagram of the HTS-C2H vocoder stage with the STRAIGHT re-synthesis algorithm. sp, f0, and ap represent the spectrum, fundamental frequency, and a-periodic analysis parameters that are generated by HTS. B is the size of the output buffer.*

5.8 The HTS-C2H implementation

This chapter introduced an implementation of C2H, named HTS-C2H. It is designed upon the HMM-based SPSS TTS system, as it is considered to be the most suitable for actively controlling the synthetic speech production in terms of robustness and flexibility (cf. § 5.1). The HTS implementation of type of synthesiser is used.

Intelligibility of the synthetic speech output is chosen from the available perceptual domains to access the success of the communication (cf. § 4.3), as it provides a good correlation with the listener's capability of understanding the spoken message, and can be calculated with low latency (cf. § 5.6.1).

The HTS-C2H implementation therefore includes a speech synthesiser, and an automatic system using the eSII index to evaluate the synthesiser outcome.

An alternative recursive parameter generation algorithm is added to the standard HTS Cholesky decomposition to solve the optimisation in eq. 3.42 (cf. § 5.3.1). This alternative method is required to be able to adjust the generative model parameters during the synthesis process. This provides a method to react to sudden environmental changes (cf. § 5.7).

The next chapter considers similarities and differences between human behaviour and the HTS-C2H components discussed in this thesis

Chapter 6

Experimental Results

Contents

The HTS-C2H implementation, described in the previous chapter, is used in this chapter to test the proposed C2H model in term of its control of speech production, and evaluation of its outcome. The aim of this experimental part is to test the *difference* between a standard good-quality synthesiser and its modified version that is implemented according the C2H model. This chapter mainly answers the research question n. 4 (cf. 1.2).

Firstly, an acoustic analyses is performed to compare the changes that are observed with human speech in quiet and adverse conditions, to those observed on synthetic speech in similar conditions. As reference, an acoustic analysis on a recorded speech-in-noise corpus, P8-Harvard, is used (Stylianou et al., 2012). The complete overview of these results is reported in Appendix B.2.1. This also addresses the research question n. 1 (cf. 1.2), and provides an experimental validation of the answer, identified in § 2. Following this, the same analyses are carried out on the HTS-C2H English and Italian realisations synthesised with full-HYO (fHYO) and full-HYP (fHYP) strengths.

Secondly, the effectiveness of the energy-motivated transform is tested, specially in terms of its scaling of the synthetic speech quality as a function of the energy involved in the process (i.e., the parameter α of § 5.5.1). Synthesised speech samples with different magnitudes of the MLLR adaptation are produced, and both objective and subjective evaluations are then used to measure the degree of intelligibility of the adapted synthetic speech when immersed in different types of noise. This addresses the research questions n. 3 and 5 (cf. 1.2).

The HTS-C2H experiments are conducted on two languages: English and Italian. Testing the model on multiple languages in this way illustrates the generalisation capabilities of C2H, as well as highlighting the language-dependant performance differences.

6.1 Experimental parameters

Four voice models – two (female and male) per language (English and Italian) – are used in these experiments. These voices are trained on the speech data introduced in § 5.2, using the standard HTS training method described in § 3.2.1. These models are used to generate baseline speech references (STD) with use of the standard HTS generation procedure (cf. § 5.1.2) against which the experimental modifications can be compared. It should be noted that all these models are trained on clear read speech, therefore the phonetic space described by these statistical models can be considered to have a degree of articulation that is close to its physiological maximum.

The English voice models are characterised by: a) ∼38000 and ∼77000 context-dependent models for *SLT* and *Nick* , respectively; b) acoustic models with 5-state HMMs, 6 streams for state, 1 Gaussian mixture per state c) 231-dimensional parameter vectors (STRAIGHT spectrum + f0 + a-periodic components); d) separate duration models with 5 states per model, 1 stream, 1 Gaussian mixture per state, and 1-dimensional parameter vector.

Similar characteristics describe the STD Italian voices. They main difference is represented by the size of these voice models, as *Lucia* and *Roberto* have ∼74000 and ∼120000 context dependent models, respectively. This implies a generally higher complexity in modelling the available Italian training data compared to the English data. Both female and male Italian voices have been shown to have high-quality characteristics; *Lucia* was employed in robot-human interactions within the EU-funded project ALIZ-E (The ALIZ-E team, 2010), and the voice *Roberto* was selected for the commercial product MiVoq (MIVOQ s.r.l, 2013). Both voices have also received good scores in informal listening tests (Tesser et al., 2013).

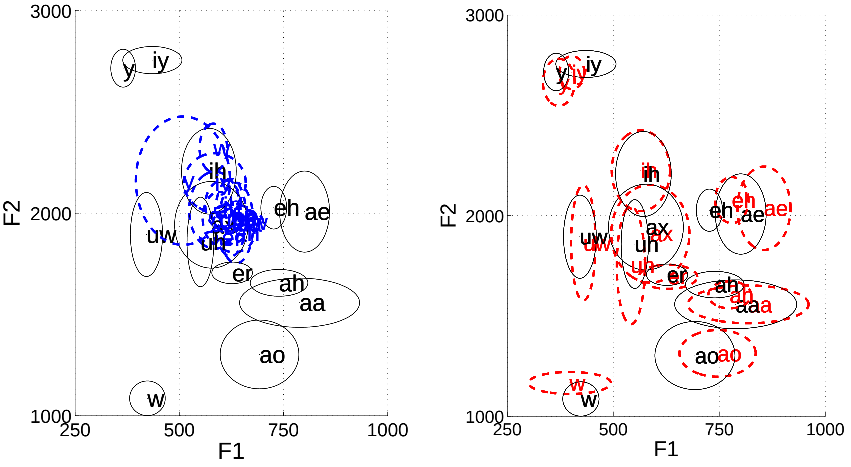
The speech synthesis samples that are used in the current experiments are generated from different sets of text material. The type of material is both language and task dependent. 200 text sentences from the Blizzard Challenge 2010 (SynSIG committee, 2010) are chosen to test HTS-C2H on the English *SLT* voice. The experiments with the English voice *Nick* are conducted on the test data provided by the Hurricane challenge § 5.2.1. These sentences serve as a common benchmark for the extensive evaluation in the challenge, as they allow comparison of HTS-C2H with other speech-in-noise synthesis methods. The Hurricane test set consists of the first 180 phonetically balanced sentences of the Harvard corpus (Rothauser et al., 1969). The test set for the Italian experiments consists of a set of 200 text sentences coherent with the phonetically balanced training data material (but not including it directly).

The analysis to generate the English linguistic labels from text is provided by the standard Festival tool set (cf. § 5.1.1). Italian linguistic labels are generated with MaryTTS (cf. § 5.1.1).

These test sentences are used as input to generate the full-strength direct transforms ($\alpha = \alpha_{HYO}$) and full-strength inverse transforms ($\alpha = \alpha_{HYP}$). Standard synthetic samples ($\alpha = 0$) of the same text set is provided as reference to compare the degree of modification. These samples are addressed respectively as *fully hypo-articulated* (fHYO), *fully hyper-articulated* (fHYP), and *standard* (STD) speech. The duration control of the synthetic speech is also computed by the adapted statistical model.

As discussed earlier in § 5.5.1, the MLLR transformation has to be scaled with the appropriate strength in order to reach the correct low-contrastive LC and high-contrastive HC operational points. The range of α thus needs to be assessed carefully. While the full-magnitude direct transform leads to a legitimate LC point in English Vowel Production Control (E-VPC), the LC configuration is obtained with half-strength transformation in English Consonant Production Control (E-CPC), Italian Vowel Production Control (I-VPC), and Italian Consonant Production Control (I-CPC). The α value to invert the MLLR transform has similar restrictions. The minimum α_{HYP} cannot be derived – as per α_{HYO} – from training motivation but must be assessed empirically for the different voice models. α_{HYP} is defined as the minimum value that does not generate unnatural artefacts in the speech realisation. The boundaries for α are defined in preliminary tests: a range of α values is applied, $\alpha \in [-2, 2]$, and the quality of the synthesis outcome is evaluated by extracting F1 and F2 mean values for every vowel. The highest and lowest values of α that still produce realistic vowels are chosen.

The resulting admissible ranges of α values are $[-0.8, 1]$ for E-VPC, and $[-0.7, 0.6]$, in E-CPC for the English voices. A brief analysis of the effect of E-VPC on the first two vowel formants for the selected α values is displayed in Figure 6.1.



(a) $\alpha = \alpha_{\text{HYO}} = 1$

(b) $\alpha = \alpha_{\text{HYP}} = -0.8$

Figure 6.1: *Modification of F1-F2 distribution for the SLT vowels with E-VPC adaptation. Ellipses indicate the F1-F2 variance-radius areas surrounding the average F1-F2 values. fHYO transformation effects are shown in 6.1(a), and fHYP transformation effects in 6.1(b).*

The fHYO F1-F2 distribution is displayed in 6.1(a) with blue-dashed ellipses, fHYP F1-F2 distribution is in 6.1(b) with red-dashed ellipses, and the reference STD F1-F2 distribution is shown in black colour. Formants are extracted with PRAAT software (Boersma and Weenink, 2018), and phones are displayed in the "CMU Pronouncing Phoneme Set" format (Carnegie Mellon University, 2015). In E-VPC reduction, the vowel space is effectively reduced, and the normal vowels tend to converge to the central part of the Figure 6.1(a). However, this is not a unique point, confirmed by the observation that schwa is not just a centralised vowel in human speech production, but is a sound that is assimilated with its phonetic and prosodic context (van Bergem, 1993). In the E-VPC expansion of Figure 6.1(b), on the other hand, the modifications are typically smaller. This is due to the almost-fully hyper-articulated characteristic of the STD vowel space. Nonetheless, this plot provides the first evidence of the effectiveness of the proposed MLLR transform. It modifies the vowel characteristics of synthetic speech in a comparable manner with that observed when human spontaneous and read speech are compared (cf. Figure 4.3).

English tests assess the effectiveness of the HTS-C2H model by undertaking acoustic analyses and objective evaluations of the intelligibility of the synthetic speech outcome of *SLT* and *Nick* voices. A subjective evaluation of the behaviour of *Nick* voice in adverse conditions is also reported. The Italian tests are similar to the English ones, comprising acoustic analyses and objective evaluations for both the *Lucia* and *Roberto* voices.

In order to test the synthetic speech signals in adverse conditions, all speech signals are normalised to have a constant RMS (RMS = -24 dBFS), and these are mixed with different noises. In the test with the English *SLT* and both *Lucia* and *Roberto* Italian voices, audio is mixed with three noise recordings, derived from an open source database of real sounds, such as

a) car engine noise recorded while driving (*CAR*),

b) babble noise recorded in a large hall (*BAB*),

c) competing speech from 2-3 English talkers (*ECS*).

The noise conditions for experiments using the *Nick* voice instead follow the Hurricane challenge guidelines (Cooke et al., 2013a). Nonetheless, the two types of disturbances are coherent with the previous experimental conditions. The masking noises are:

a) a fluctuating masker which is competing speech (*CS*) from a female talker producing read speech scaled to produce (utterance-wide) signal-to-noise ratio (SNR);

b) a stationary masker which is speech-shaped noise (*SSN*) whose long-term average spectrum matches that of the CS.

In order to normalise the speech energy with respect to the noise, the SSNR is computed (Hu and Loizou, 2008; Ma et al., 2009). SSNR is defined as the mean sound-to-noise ratio extracted only in the speech regions. It is computed using the VoiceBox toolbox for MATLAB (Brookes, 2012).

$$\mathrm{SSNR} = \frac{10}{K}\sum_{m=0}^{M-1}\left(\log_{10}\frac{\sum\limits_{n=mN}^{mN+M-1} s_R^2(n)}{\sum\limits_{n=mN}^{mN+M-1}\left(s_D(n)-s_R(n)\right)^2}\cdot \mathrm{VAD}_m\right) \tag{6.1}$$

where K is the number of speech segments in which there is speech activity, i.e., VAD$_m = 1$; M is the number of segments; N is the number of audio samples in each segment; $s_R(n)$ is the reference clean speech signal; and $s_D(n)$ is the speech corrupted by noise.

All noises are amplified to have fixed mean SSNR. Three SSNR levels are taken into consideration for these experiments: high, mid, and low SSNRs. For CAR, BAB, and SSN noises, these levels correspond to 1, -4, and -9 dB. For ECS and CS noises, these are -7, -14 and -21 dB.

6.2 Acoustic analysis

This section describes, a series of acoustic analyses that are performed on the various speech outcomes of the HTS-C2H implementation. The results are compared with observations of human speech production in adverse conditions.

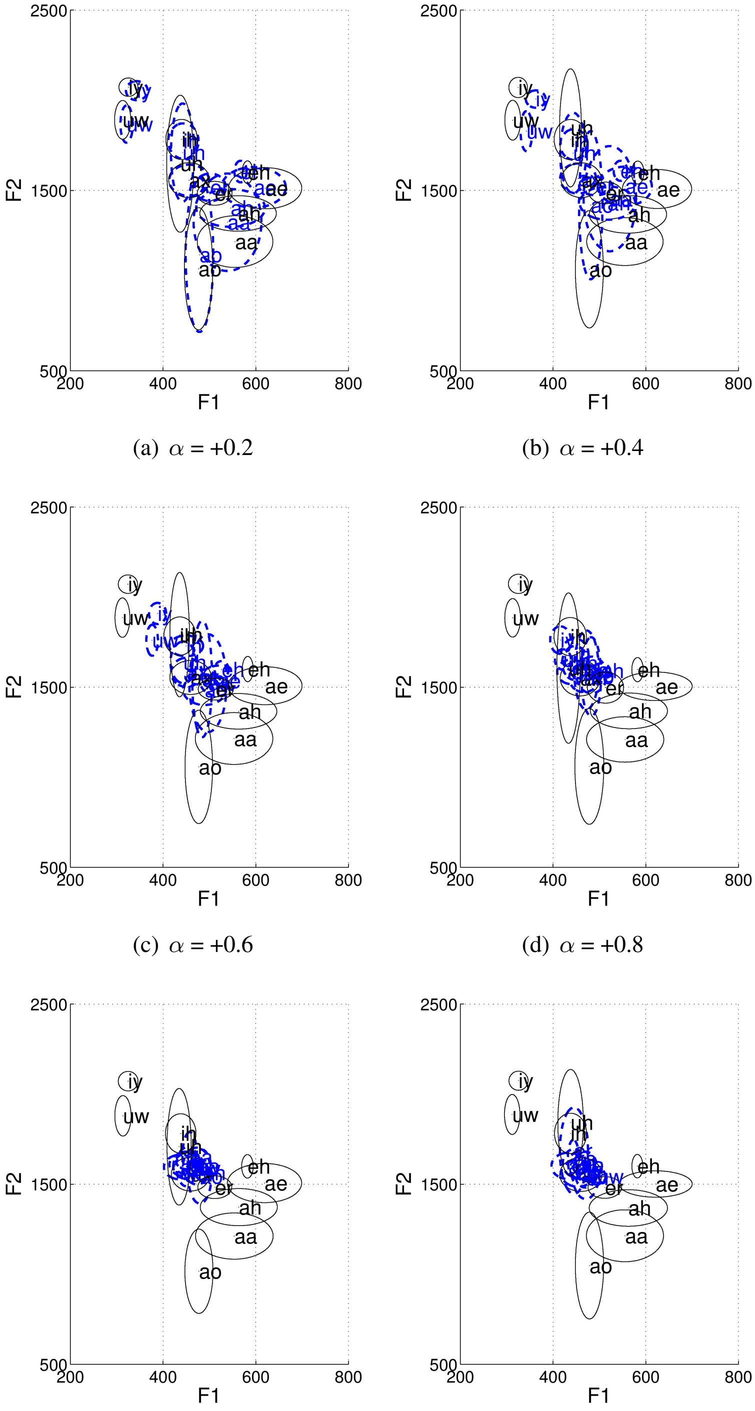
Much of the linguistic analysis in the literature uses the software PRAAT (Boersma and Van Heuven, 2001; Boersma and Weenink, 2018) as it provides state-of-the-art tools to measure the principal speech characteristics that are used in phonetic analysis. In order to extend such analyses, a set of Matlab functions is specifically implemented within the XPLIC8 software suite (Tang et al., 2012) in collaboration with other researchers (Stylianou et al., 2012). XPLIC8 is a Matlab graphic tool for performing a set of phonetic analyses on single or batch of signals. The details of this software and its analysis algorithms are reported in Appendix B.

The acoustic-phonetic analyses, implemented in XPLIC8 , are used in assessing the speech audio characteristics of human and HTS-C2H synthetic speech. Synthetic speech acoustic correlates are analysed in the rest of this section, whilst an extensive XPLIC8 analysis of a real human speech corpus in adverse conditions, the P8-Harvard corpus (Stylianou et al., 2012), is reported in Appendix B.2.

6.2.1 English HTS-C2H output

The synthetic speech produced by the English voice model *Nick* , and controlled by the HTS-C2H energy-motivated transform (cf. § 5.5.1) is acoustically analysed in this section.

The first assessment of the MLLR adaptation effects on the synthetic speech signal is done by measuring the vowel F1 and F2 formants on the outcome of HTS-C2H with various degrees of articulation from HYP to HYO, with $\alpha \in [-1.2, 1.2]$. Figure 6.2 and Figure 6.3 clearly depict the gradual control of the formant shift towards and away from the LC attractor (ə) respectively.



(a) $\alpha = +0.2$

(b) $\alpha = +0.4$

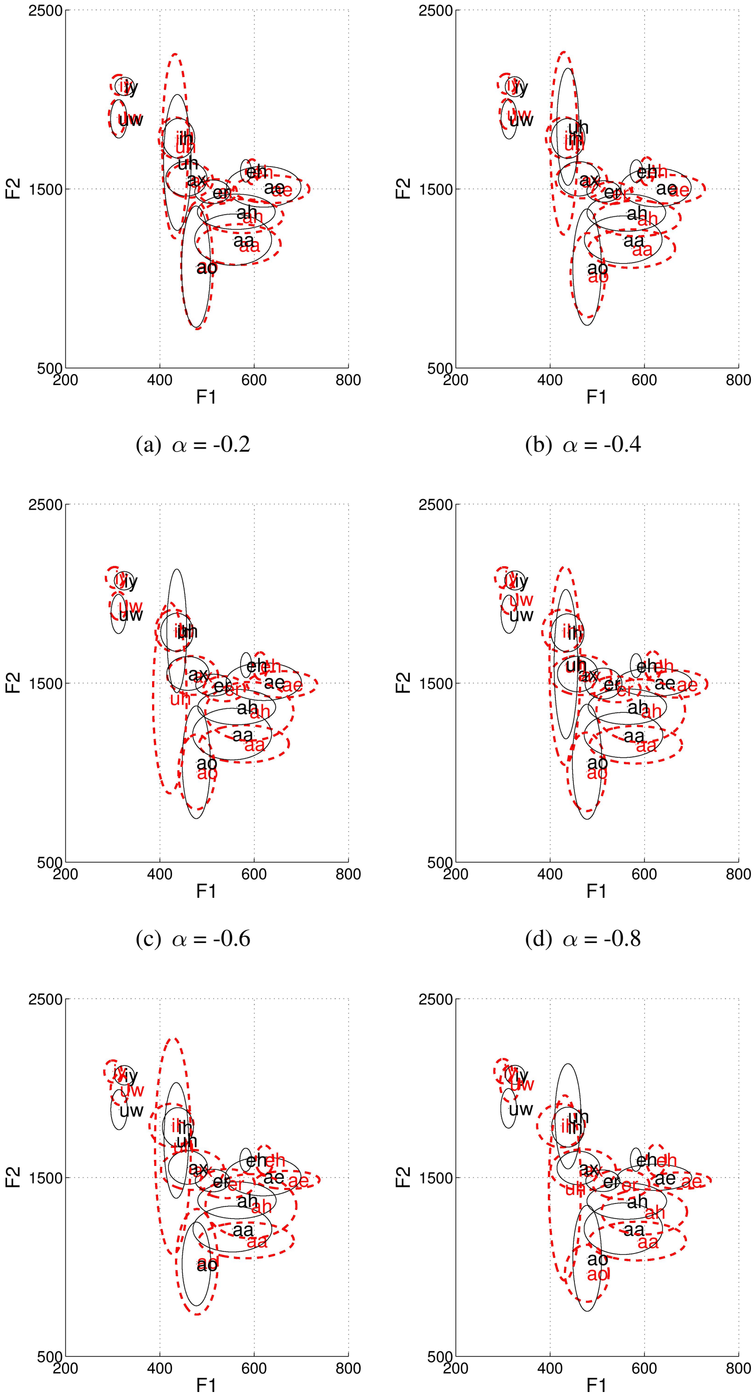
(c) $\alpha = +0.6$

(d) $\alpha = +0.8$

(e) $\alpha = +1.0$

(f) $\alpha = +1.2$

Figure 6.2: *Effect of the control of the E-VPC+E-CPC strength α toward hypo-articulation on vowel F1 and F2 formants, using the English male voice Nick .*



(a) α = -0.2

(b) α = -0.4

(c) α = -0.6

(d) α = -0.8

(e) α = -1.0

(f) α = -1.2

Figure 6.3: *Effect of the control of the E-VPC+E-CPC strength α toward hyper-articulation on vowel F1 and F2 formants, using the English male voice Nick .*

An analysis similar to that reported for the P8-Harvard corpus in Appendix B.2.1 is also performed on the *Nick* HTS-C2H outcome (Nicolao and Moore, 2013a). The acoustic phonetic analysis parameters are:

speech duration parameters the mean word duration (*MWD*), along with the mean sentence duration (*MSD*) and the mean pause duration (*MPD*);

spectral parameters the long term average spectrum (*LTAS13*), the spectral tilt (*Sp.Tilt*), the spectrum Centre of Gravity (*Sp.CoG*), and the vowel space area (*F1F2 area*);

pitch parameters the average fundamental frequency (*F0*), and its range (*F0 range*).

These values are extracted using the XPLIC8 software, as shown in Appendix B.2.1. Following the work of other researchers (van Son and Pols, 1999; Cooke et al., 2008; Hazan and Baker, 2011; Stylianou et al., 2012), these have been proven to be significantly correlated to the degree of clarity of speech.

The extracted values for the *Nick* voice are reported in Table 6.1 and Table 6.2.

Table 6.1: *Acoustic analysis of the three degrees of E-VPC adaptation for the Nick voice. The elongation/reduction w.r.t STD is given in parenthesis.*

Type of analysis	fHYO	STD	fHYP
MSD [s]	2.98 (-14.9%)	3.5	3.91 (+11.7%)
MWD [s]	0.27 (-15.6%)	0.32	0.36 (+12.5%)
MPD [s]	0.13 (-13.3%)	0.15	0.17 (+13.3%)
LTAS13 [dB SPL]	33.6 (-7.2%)	36.2	41.1 (+13.5%)
Sp.Tilt [dB/dec]	-6.2 (+6.9%)	-5.8	-4.7 (-19.0%)
Sp.CoG [Hz]	712 (-13.3%)	821	1024 (+24.7%)
F1F2 area [Hz^2]	1014 (-96.5%)	29021	70509 (+143.0%)
F0 [Hz]	172.6 (-0.9%)	174.1	174.7 (+0.3%)
F0 range [Hz]	146-185 (+21.9%)	151-183	145-190 (+40.6%)

The effects of the HTS-C2H control on the vowels (E-VPC, Table 6.1), and the control on the consonants (E-CPC, Table 6.2), is measured separately. The magnitude of the adaptation is chosen to be the maximum in both hyper- and hypo-articulation direction: $\alpha = \alpha_{HYP}$ and $\alpha = \alpha_{HYO}$ respectively. Raw analysis values along with the difference with respect to the appropriate STD reference are shown.

Table 6.2: *Acoustic analysis of the three degrees of E-CPC adaptation of the English male voice Nick . The elongation/reduction w.r.t STD is given in parenthesis.*

Type of analysis	fHYO	STD	fHYP
MSD [s]	3.43 (-2.0%)	3.5	3.6 (+2.9%)
MWD [s]	0.31 (-3.1%)	0.32	0.33 (+3.1%)
MPD [s]	0.14 (-6.7%)	0.15	0.16 (+6.7%)
LTAS13 [dB SPL]	35.4 (-2.2%)	36.2	38.4 (+6.1%)
Sp.Tilt [dB/dec]	-6.1 (+5.2%)	-5.8	-5.1 (-12.1%)
Sp.CoG [Hz]	547 (-33.4%)	821	1156 (+40.8%)
F1F2 area [Hz^2]	1014 (+44.1%)	29021	70509 (+93.3%)
F0 [Hz]	174.1 (+0.0%)	174.1	173.4 (-0.9%)
F0 range [Hz]	144-185 (+28.1%)	151-183	150-184 (+6.3%)

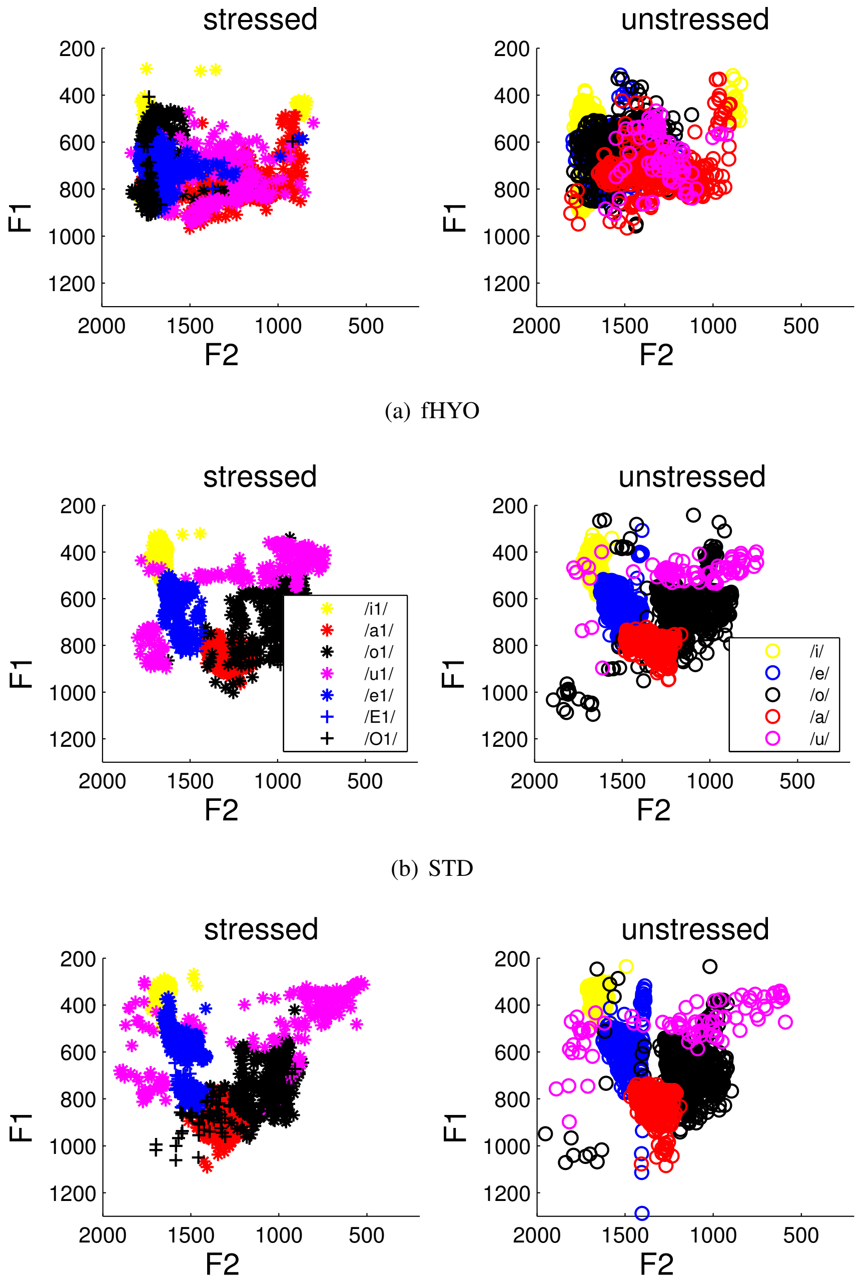
It can be seen here that the synthetic speech production using the HTS-C2H-controlled English male voice *Nick* follows the same trends as is seen with the human speech production. When the system controls the degree of articulation towards hypo-articulation ($\alpha > 0$), sentence and phones are shortened (see MSD and MPD in Table 6.1 and Table 6.2). The spectral energy distribution in the high frequencies (LTAS13) decreases. The spectral tilt increases, and its Sp.CoG moves towards the low frequency. The measured F1-F2 area also contracts, as expected. The median F0 is lower, whilst the F0 range remains constant. When the system instead controls the speech production towards hyper-articulation ($\alpha < 0$), the opposite behaviour is measured.

6.2.2 Italian HTS-C2H output

Similarly to the analysis reported in Appendix B.2.1 and in § 6.2.1, an acoustic phonetic evaluation is also performed on the Italian HTS-C2H speech outcome, as reported in (Nicolao et al., 2013).

For the Italian voices, the vowel F1 and F2 formants of the synthesiser outcome are displayed for only the two extreme degrees of articulations (fHYP, fHYO), and for the standard production (STD). Figure 6.4 and Figure 6.5 show these results for the two Italian voices, *Lucia* and *Roberto* , respectively.

Stressed and unstressed vowel behaviours are measured separately, as stress in Italian has a contrastive function. It is clear from these plots how both voices move from the confused and centralised positions of the HYO configuration

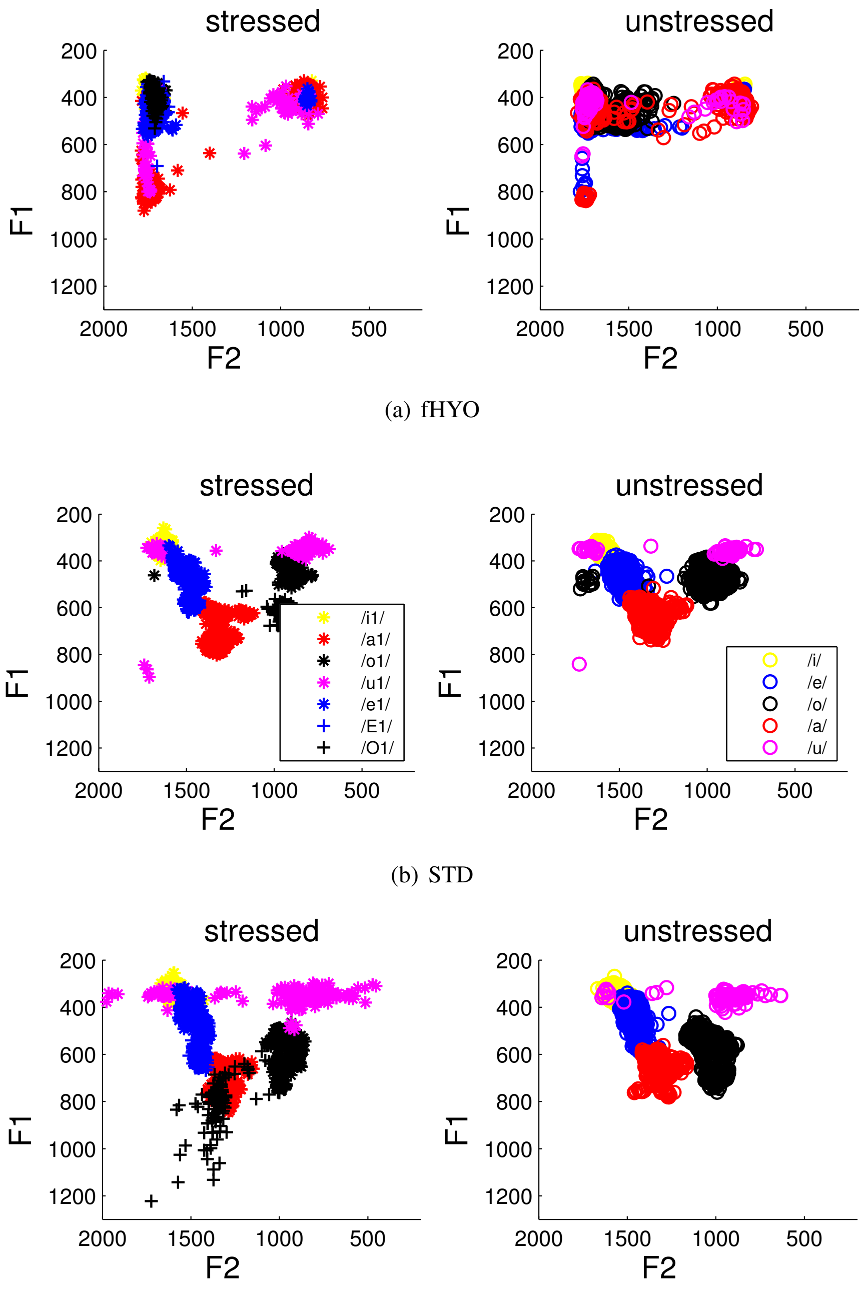


(a) fHYO

(b) STD

(c) fHYP

Figure 6.4: *Effect of the HYO (a) and the HYP (c) adaptation applied to the Lucia voice. The plot for STD voice vowels is also given for reference (b). SAMPA symbols are used in the legend. Units are Hz for both axes.*



(a) fHYO

(b) STD

(c) fHYP

Figure 6.5: *Effect of the HYO (a) and the HYP (c) adaptation applied to the Roberto voice. The plot for STD voice vowels is also given for reference (b). SAMPA symbols are used in the legend. Units are Hz for both axes.*

(Figure 6.4(a) and 6.5(a)) to the more separated and recognisable ones of HYP (Figure 6.4(c) and 6.5(c)). In the HYP vowels, the stressed [ɔ] (black cross + in figures) tends to migrate towards the more articulated [a] (red asterisk ⋆). The stressed HYO vowels in *Roberto* aggregate in three main positions rather than a unique central one. This confirms the idea that the low-contrastive configuration is not a unique position close to [ə], but takes an intermediate position depending on the surrounding phones.

Transformations seem to achieve a more effective reduction/expansion on the *Roberto* voice. It is noteworthy that the *Roberto* vowel variance (Figure 6.5) is quite limited with respect to the *Lucia* one (Figure 6.4). Indeed, the former is created using a professional speaker's voice whereas the latter denotes some regional accent influence. Moreover, the amount of recorded corpus is different: *Roberto* training corpus is one-third bigger the *Lucia* one. Even though the two adaptations, I-VPC and I-CPC, contribute to modify the signal simultaneously, the vowel charts behave similarly to what is observed for the schwa-based ones in (Nicolao et al., 2012).

The most common phenomena observed in Italian hyper/hypo-articulated speech are the same as what observed in English: formant shifting, spectral energy redistribution, speaking rate changes, and pitch modification. Therefore, an acoustic analysis with the same XPLIC8 tools is also performed. Average result values are shown in Table 6.3 for *Lucia* and in Table 6.4 for *Roberto* . In both tables, the clearest modifications are observed in the vowel space (*F1F2 area*) expansion/reduction. Even though this is imposed by design, nonetheless this observation, together with Figure 6.4 and Figure 6.5, it shows that the adaptation behaves correctly.

Table 6.3: *Acoustic analysis of the three degrees of adaptation of the Italian female voice Lucia . The elongation/reduction w.r.t STD is given in parenthesis.*

Type of analysis	**HYO**	**STD**	**HYP**
MSD [s]	5.75 (-5.1%)	6.06	6.38 (+5.3%)
MPD [s]	0.078 (-2.5%)	0.08	0.083 (+3.7%)
LTAS13 [dB SPL]	47.7 (-9.3%)	52.6	58.3 (+10.8%)
Sp.Tilt [dB/dec]	-5.6 (+7.7%)	-5.2	-4.7 (-9.6%)
Sp.CoG [Hz]	394.1 (-27.9%)	546.2	835.9 (+53.0%)
F1F2 area [Hz^2]	14115 (-90.1%)	142401	203959 (+43.2%)
F0 [Hz]	197.3 (-3.4%)	204.3	210.3 (+2.9%)
F0 range [Hz]	138-225 (-23.6%)	133-247	134-276 (+24.8%)

Other evident differences between the three sets of speech utterances appear in the spectrum energy shift (e.g. *Sp.CoG* and *Sp.Tilt*) and in the duration (*MSD* and *MPD*). The latter shows the tendency of the automatic system to elongate the speech production to increase phonetic contrast and vice-versa.

Table 6.4: *Acoustic analysis of the three degrees of adaptation of the Italian male voice Roberto . The elongation/reduction w.r.t STD is given in parenthesis.*

Type of analysis	HYO	STD	HYP
MSD [s]	4.72 (-14.2%)	5.50	6.28 (+14.2%)
MPD [s]	0.06 (-16.7.5%)	0.072	0.083 (+15.3%)
LTAS13 [dB SPL]	44.7 (-6.9%)	48.0	56.3 (+17.3%)
Sp.Tilt [dB/dec]	-6.3 (+8.6%)	-5.8	-4.9 (-15.5%)
Sp.CoG [Hz]	434.5 (-30.6%)	625.8	947.0 (+51.33%)
F1F2 area [Hz^2]	469 (-99.6%)	124518	143156 (+15.0%)
F0 [Hz]	119.7 (+2.9%)	116.3	112.7 (-3.1%)
F0 range [Hz]	73-143 (-6.7%)	68-143	67-162 (+26.6%)

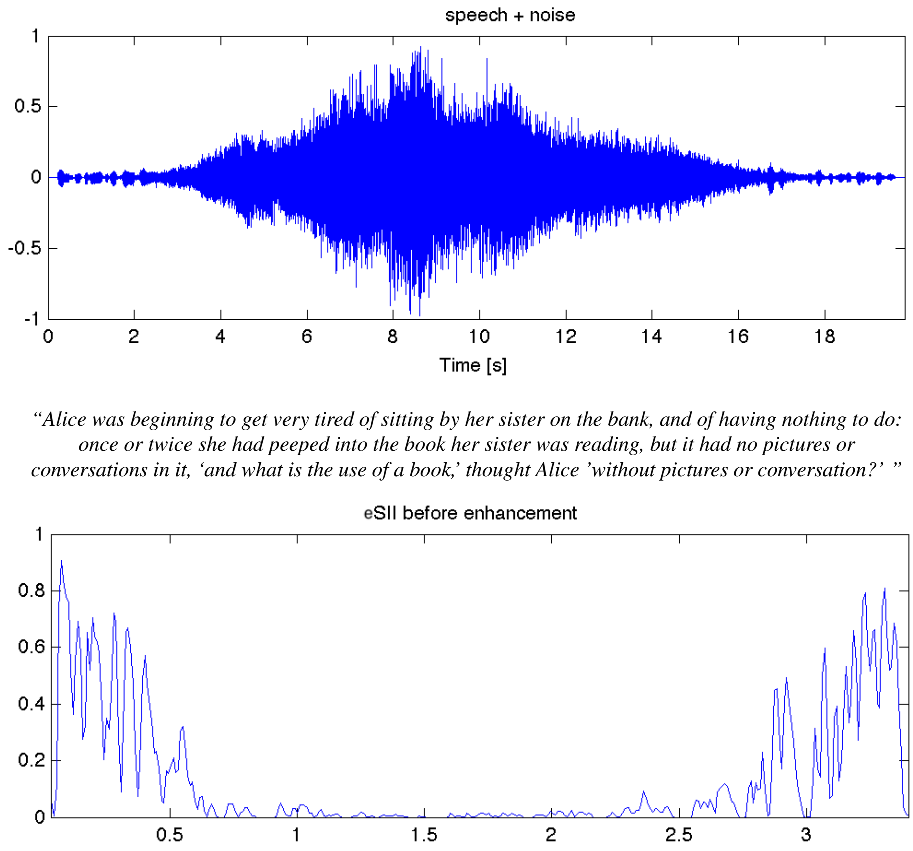
It is again noteworthy that all the observed synthetic speech characteristics emerge spontaneously from applying different degree of adaptation to the synthetic speech production model. Aside from the control of the phonetic contrast, no assumption or empirical rule in the adaptation training has been made to model these duration and spectral behaviours.

From the acoustic analysis, it can be concluded that the proposed transforms, acoustically-speaking, behave similarly to what is observed in human speech production for both English and Italian languages.

6.3 Perception feedback loop evaluation

One of the crucial elements of C2H is the ability to react to environmental changes. This ability is implemented via the perception loop that detects the evolution in disturbance, and by the controller that translates these changes in disturbance into a control signal.

If HTS-C2H is used with the conceptualiser motivation value m set to 0 in eq. 5.28, the α range collapses, and the synthesiser control mechanism is deactivated with the result that no speech adaptation is performed. In the Figures 6.6, plot of a speech signal generated in this way is displayed, mixed with a time-varying noise.

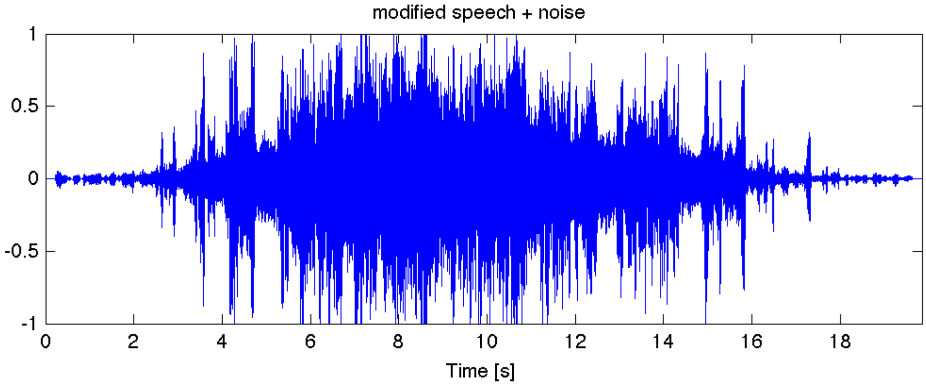


“Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, ‘and what is the use of a book,’ thought Alice ’without pictures or conversation?’ ”

Figure 6.6: *Example of speech synthesis signal in noise with inactive perception feedback. The corresponding eSII is also plotted.*

Since the overall energy of the speech realisation is fixed at the beginning of the synthesis and cannot be changed, the speech waveform is completely masked by the increasing loudness of the noise. This is clearly highlighted by the reduction in eSII, see § 5.6.1 (Figure 6.6, lower plot). The measured eSII values are reduced to almost zero (unintelligible) as the noise signal increases in amplitude.

If the HTS-C2H motivation m is greater than 0, the controller can transform the error signal into a control signal. The effect of the controlling mechanism of the speech production is shown in Figure 6.7. In this example, for clarity, only the speech signal amplitude (i.e., loudness) is adjusted. In Figure 6.7, it is evident how HTS-C2H operates to maintain the estimated intelligibility level (eSII) at the specific value determined by the conceptualiser intent, $\text{eSII}_{\text{intent}}$ of eq. 5.29. When the HTS-C2H context-aware system is active, the perception loop



"Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, 'and what is the use of a book,' thought Alice 'without pictures or conversation?' " (same of Figure 6.6)

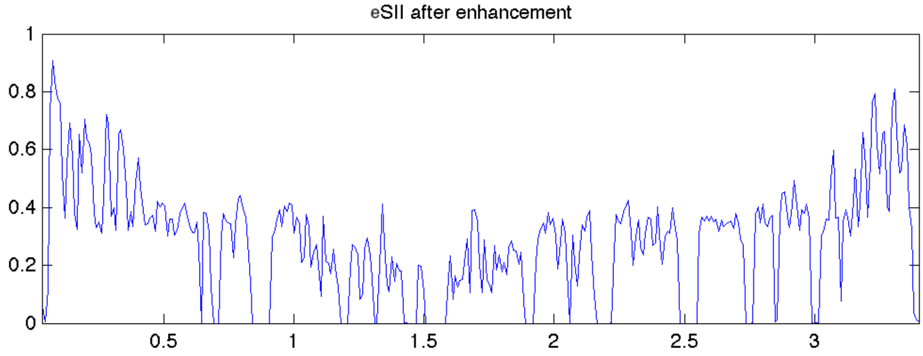


Figure 6.7: *Example of speech synthesis signal in noise with active perception feedback. Adjustments are applied to the speech signal loudness. The corresponding eSII is also plotted. The intent eSII threshold is set to 0.4.*

measures the intelligibility of the speech realisation, and if the value is below the intended threshold, the speech loudness is increased. If eSII is above the intelligibility required, the speech signal volume is reduced to allow the synthesiser to minimise the effort. Here, the α value that controls the transformation increases and decreases in reaction to the noise loudness. The peech signal and noise signal are the same of those as Figure 6.6, but the reactive control allow the speech signal to be audible and therefore more intelligible.

6.4 Objective evaluation of intelligibility control

Intelligibility is directly correlated to the effort involved in speech production. In the human speech analysis of Appendix B.2.1, it is observed that the degree of adverse condition severity – i.e., no barrier (NB) → babble channel (BAB) → vocoded channel (VOC) – is directly linked to the amount of required speech adjustment for an intelligible communication. The hypothesis is that the amount of adjustment observed in human speech emerges from an energy motivated transform, which controls the degree of phonetic contrast in synthetic speech production.

The following section illustrates the experimental results of an objective evaluation of intelligibility. In these experiments, automatic methods for speech intelligibility estimation (cf. § 4.3.4) are used to assess the degree to which the HTS-C2H adaptation changes the intelligibility in adverse conditions.

The experiments investigate both sides of the hyper/hypo-articulation spectrum, since, in some contexts (cf. § 4.5), fHYP speech might not be the optimal policy to achieve the communicative intent (cf. § 4.1). Reducing speech intelligibility often creates a friendlier and less assertive type of speech which might be more positively accepted by the listener.

Three different kinds of speech signals corresponding to three degrees of transformation magnitude (fHYP, fHYO, and STD) are mixed with the noises at different SSNRs, which is computed as described in § 6.1. Two automatic intelligibility estimation methods (SII and Dau) are used to analyse the HTS-C2H synthetic audio samples. Automatic intelligibility estimation methods are fairly reliable tools to score the speech synthesis clarity. Though most of them measure the audibility of a signal without taking phonetic content into account, some are proven to be highly correlated to human understanding performances. The Dau index in particular is reported to have a high accuracy at predicting speech clarity in noise (Valentini-Botinhao et al., 2011). Despite the lower correlation with human perception than the Dau index, an SII analysis is also reported. Along with the assessment of the speech control, the analogies with the more accurate Dau index provide evidence of the HTS-C2H perception loop adequacy at predicting the listener's understanding.

The intelligibility differences (improvement or degradation) between the hypo and hyper-articulated samples with respect to the standard ones are used as a metric to test the effectiveness of the C2H transform motivated by phonetic-contrast.

6.4.1 English HTS-C2H output

The English *SLT* and *Nick* voices are used to produce several speech samples with different degrees of articulation: HYO, HYP, and STD. The objective intelligibility measures evaluate the speech samples when they are mixed with the types of noises described in § 6.1.

The intelligibility differences, due to the noise type, on the English female speech samples *SLT* are summarised in Figure 6.8 and Figure 6.9, for the SII and Dau measures, respectively. Both indexes are averaged across the whole test set, and the results are illustrated in the two figures for the CAR, BAB, and ECS noises.

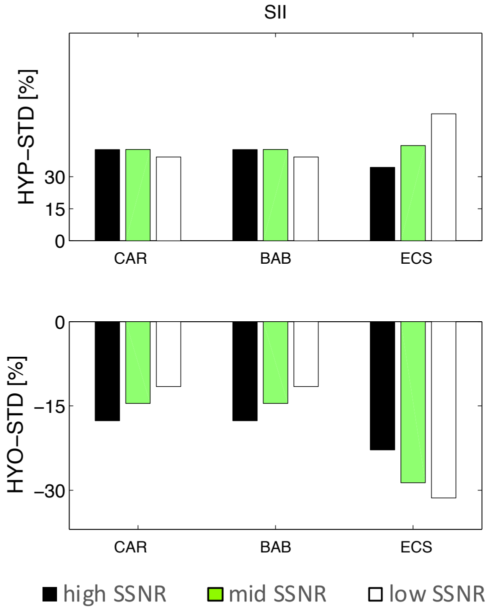


Figure 6.8: *Mean SII-differences (in percentage) between SLT STD speech and the HYO (below) and HYP (above) versions. CAR (left), BAB (centre), and ECS (right) noises and high (black), mid (green), and low (white) SSNR levels are considered.*

The detailed experiment results for *SLT* are listed in Appendix C.1.

In *SLT* experiments, both indexes measure consistent speech clarity improvement, when HTS-C2H is set to produce HYP speech: ∼ +30% and ∼ +15% with SII and Dau respectively. Speech clarity is reduced, on the other end, when the synthesiser produces HYO speech: ∼ -18% and ∼ -13% with SII and Dau respectively. No major differences are observed for different types or intensity of noise. It is also important to mention that SII and Dau indexes show a similar relative behaviour.

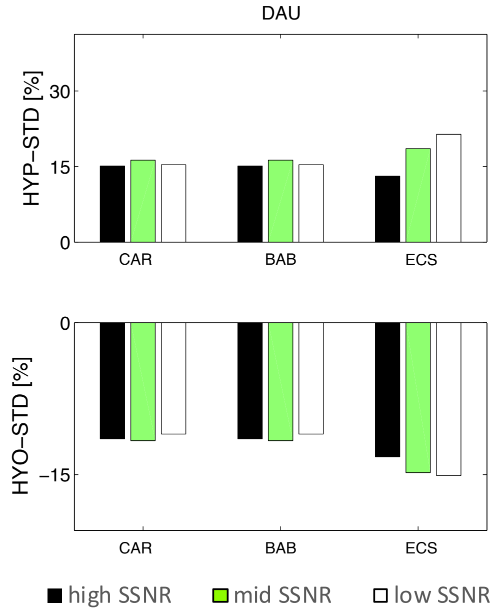


Figure 6.9: *Mean Dau-differences (in percentage) between SLT STD speech and the HYO (below) and HYP (above) versions. CAR (left), BAB (centre), and ECS (right) noises and high (black), mid (green), and low (white) SSNR levels are considered.*

The intelligibility differences on the English male speech samples *Nick* are summarised in Figure 6.10 and Figure 6.11, for the SII and Dau, respectively, for the SSN, and CS noises. The detailed experiment results for *Nick* are listed in Appendix C.2.

The observed behaviour with both indexes confirms the expected increment and reduction of the speech clarity for each noise. However, this time a dependency with SSNR levels is detected, such that the transform on this voice model increases its effectiveness at low SSNR. SII improves from around +5 to +10 % for both noises. Dau improves from $\sim$ +18 to +44 % for SSN, and from $\sim$ +6 to +17 % for CS. Clarity reduction is between -5 and -10% in SII, and between -10 and -20% in Dau. The SSNR dependency trend is similar in both SII and Dau.

6.4.2 Italian HTS-C2H output

Similarly to the English tests, the Italian *Lucia* and *Roberto* voices are used to produce several speech samples with different degrees of articulation: HYO, HYP, and STD. SII and Dau indexes are used to estimate the speech intelligibility of these test samples mixed with the types of noise as per specifications in § 6.1.

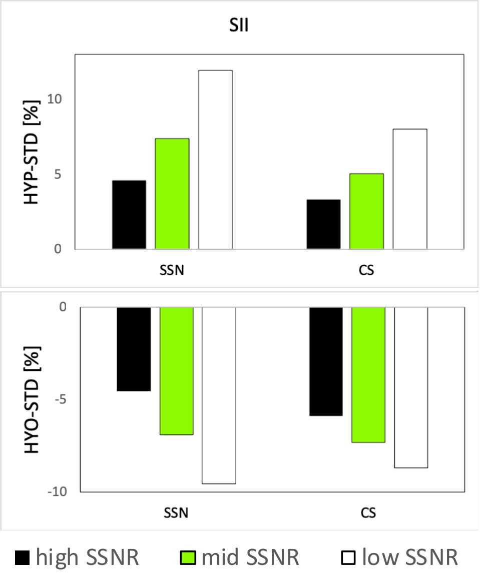


Figure 6.10: *Mean SII-differences (in percentage) between Nick STD speech and the HYO (below) and HYP (above) versions. CAR (left), BAB (centre), and ECS (right) noises and high (black), mid (green), and low (white) SSNR levels are considered.*

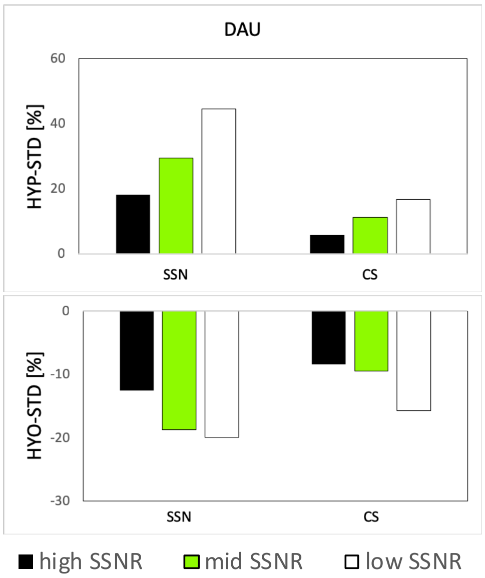


Figure 6.11: *Mean Dau-differences (in percentage) between Nick STD speech and the HYO (below) and HYP (above) versions. CAR (left), BAB (centre), and ECS (right) noises and high (black), mid (green), and low (white) SSNR levels are considered.*

The intelligibility differences on the Italian female speech samples *Lucia* are summarised in Figure 6.12 and Figure 6.13, for SII and Dau respectively. As before, both indexes are averaged across the whole test set, and the results are illustrated in the two figures for the CAR, BAB, and ECS noises. The detailed experiment results for *Lucia* are listed in Appendix C.3.

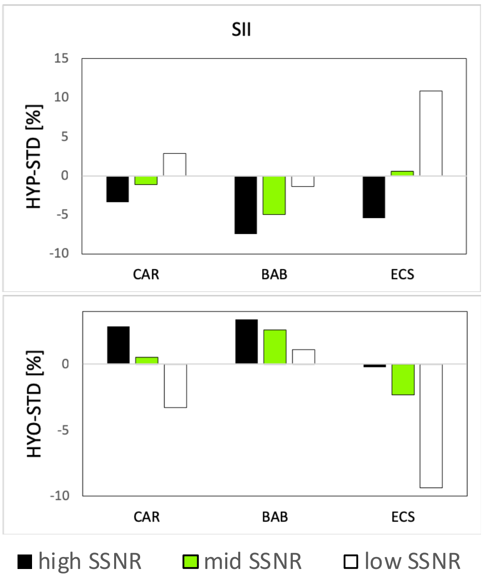


Figure 6.12: *Mean SII-differences (in percentage) between Lucia STD speech and the HYO (below) and HYP (above) versions. CAR (left), BAB (centre), and ECS (right) noises and high (black), mid (green), and low (white) SSNR levels are considered.*

The SII returns contrasting values for this voice model, and there is no evidence of a trend across SSNR. On the other hand, the Dau index shows a similar pattern of results to those measured in *Nick* , with analogous SSNR dependency also observed. Dau improvement spans from +3 to +70%, and the reduction is from -5 to -31%. In these experimental conditions, the SII does not seem to be coherent with the more reliable Dau. This may have happened due to some *Lucia* synthetic speech characteristics observed in Figures 6.4 and 6.5, where *Lucia* vowel realisations were less consistent than the ones for *Roberto* . This might be due to the nature of the training speech that is used to build the *Lucia* voice and the related transform. The recorded female speaker is not a professional voice talent, and her production had higher variability, as well as a mild regional accent.

The intelligibility differences on the Italian male speech samples *Roberto* are summarised in Figure 6.14 and Figure 6.15, for SII and Dau respectively. Both indexes are averaged across the whole test set, and the results are illustrated in the

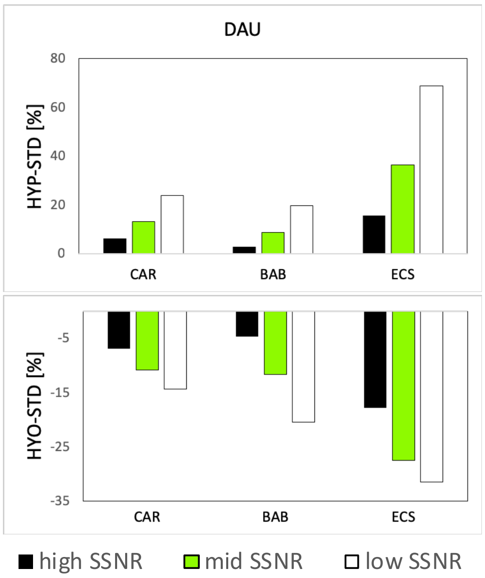


Figure 6.13: *Mean Dau-differences (in percentage) between Lucia STD speech and the HYO (below) and HYP (above) versions. CAR (left), BAB (centre), and ECS (right) noises and high (black), mid (green), and low (white) SSNR levels are considered.*

two figures for the CAR, BAB, and ECS noises. The detailed experiment results for *Roberto* are listed in Appendix C.4.

As for *Lucia* , the SII estimation again shows an opposite trend with respect to previous experiments. Here, the CAR noise seems to benefit of the largest improvement, whilst the BAB and ECS noises, mixed with high SSNR, exhibit an actual clarity reduction regardless of which direction the transform is moving to. On the other end, Dau index still shows a consistent trend which confirms the expected transform clarity modifications. Speech signal adaptations are more effective at mid/high SSNR levels. On average, the intelligibility deviation from the STD voice is around ±10% for both HYO and HYP.

Overall, the intelligibility analyses of *SLT* , *Nick* , *Lucia* , and *Roberto* speech transformed samples show a variation in intelligibility level with respect to the STD speech production. This intelligibility changes are coherent with the commanded degree of articulation in all types of noises for all four voices.

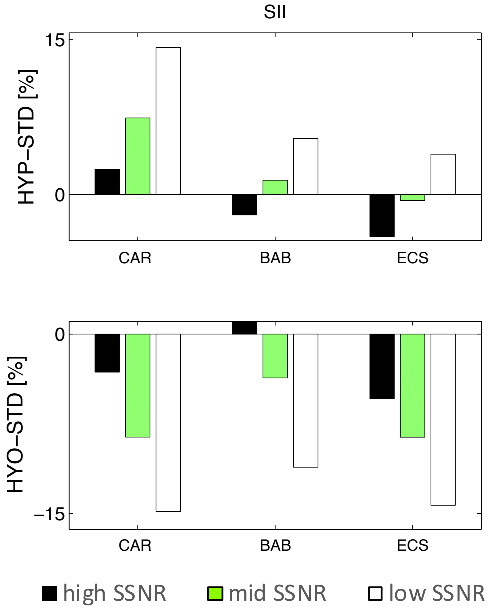


Figure 6.14: *Mean SII-differences (in percentage) between Roberto STD speech and the HYO (below) and HYP (above) versions. CAR (left), BAB (centre), and ECS (right) noises and high (black), mid (green), and low (white) SSNR levels are considered.*

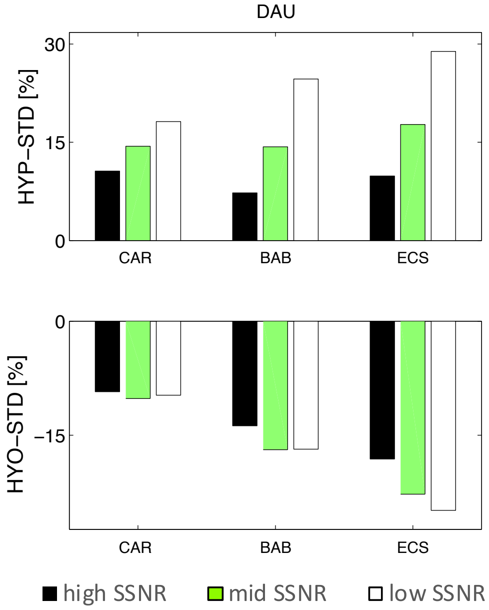
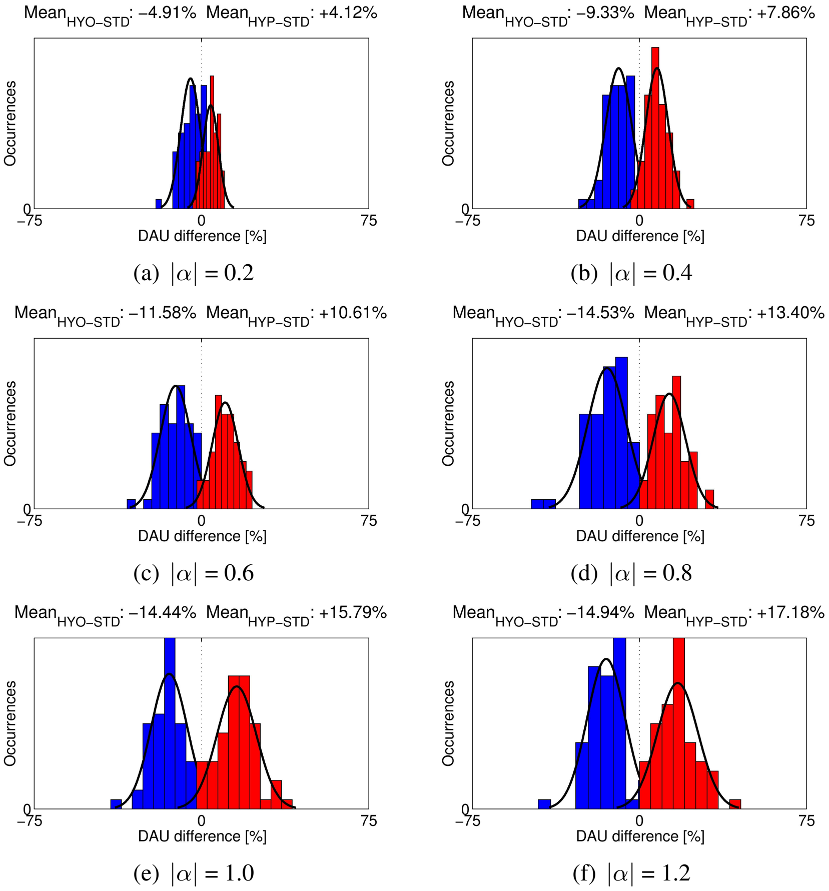


Figure 6.15: *Mean Dau-differences (in percentage) between Roberto STD speech and the HYO (below) and HYP (above) versions. CAR (left), BAB (centre), and ECS (right) noises and high (black), mid (green), and low (white) SSNR levels are considered.*

6.4.3 Scaling the control

The parameter α is used by the controller to adjust the magnitude of the transformation. Speech intelligibility is evaluated for different α to test how accurately the speech clarity can be controlled by such parameter.

Figure 6.16 reports the case of the Dau-estimated intelligibility results for the English *Nick* test samples, mixed with the SSN noise, and shows that the increment of $|\alpha|$ moves the speech clarity in noise towards the expected direction. Moreover, the transform strength can be scaled proportionally to the control values.



(a) $|\alpha| = 0.2$

(b) $|\alpha| = 0.4$

(c) $|\alpha| = 0.6$

(d) $|\alpha| = 0.8$

(e) $|\alpha| = 1.0$

(f) $|\alpha| = 1.2$

Figure 6.16: *Example of intelligibility control with the HTS-C2H transform. Intelligibility is estimated with the Dau index for SSN noise. The range of α is between -1.2 (fHYP) and +1.2 fHYO. The results of the transformation towards HYP (in red) and towards HYP (in blue) are displayed on the same plot.*

This experimental evidence demonstrates that the proposed HTS-C2H transform can effectively control the degree of clarity of a speech synthesiser, with the use a single parameter to reflect the desired degree of articulation.

6.5 Subjective evaluation results

Along with the objective evaluations of the HTS-C2H outcome reported in the previous part of the chapter, a subjective evaluation of the synthetic speech in noise is conducted to evaluate the effectiveness of the phonetic contrast-based transformation in adverse conditions.

The subjective evaluation results reported here are derived from the outcome of the Hurricane Challenge (Cooke et al., 2013a). The Hurricane Challenge was a large-scale open evaluation, organised by the members of the LISTA European project network, that aimed to assess 18 algorithms designed to enhance speech intelligibility. Two types of entries were evaluated within this challenge: algorithmically-modified and synthetically-generated speech waveforms. The HTS-C2H entry competed in the context of synthetic speech.

A HTS-C2H synthetic voice was trained on the English male speech corpus *Nick* that was provided to all the challenge entrants to train and adapt their systems, see § 5.2.1. A HTS-C2H fully hyper-articulated speech (fHYP) set was entered in the challenge, and the intelligibility in different noise types and levels was measured with extensive listening tests. Along with spectral adaptation, phone durations were controlled by the acoustic models (no copy synthesis). However, overall duration was kept within the imposed tolerance of ± 0.5s. The adapted speech entry was combined with two maskers at three SSNR and evaluated by listeners. Listeners could not change the output level at which the signals were played to them. Masker noises and SSNR levels are similar to those used in the objective evaluation, see § 6.1.

A total of 180 sentences x 2 maskers x 3 noise energy levels were submitted to the listening test. The experimental conditions are described in detail in (Cooke et al., 2013a). The evaluation used 175 participants who were native English adults of 19-27 years of age. Listeners had no speech and/or language disorders and passed an audiological screening. Stimuli were normalised to have the same RMS energy level and were presented to participants in dedicated sound-attenuated listening booths. Participants were instructed to transcribe what they could hear in the stimuli. The number of content words correctly identified represents the final score.

Table 6.5 presents the results in terms of the absolute percentage of correctly transcribed words for the unmodified natural speech (*plain*), the synthetic baselines (*TTS*), and the HTS-C2H stimuli.

Table 6.5: *Intelligibility subjective evaluation results of the Hurricane Challenge for the HTS-C2H entry. The numbers represent the keywords correct scores expressed in absolute percentages for plain, TTS, and HTS-C2H speech.*

		CS			**SSN**	
Type of speech	snrHi	snrMid	snrLo	snrHi	snrMid	snrLo
plain	85.1 ±1.5	57.0 ±2.4	24.8 ±1.9	88.3 ±1.3	63.0 ±2.2	17.3 ±1.8
TTS	59.7 ±2.3	31.3 ±1.9	11.7 ±1.3	63.7 ±2.2	32.8 ±2.1	6.8 ±1.2
HTS-C2H	45.6 ±2.1	24.7 ±1.7	10.8 ±1.4	46.3 ±2.0	22.4 ±1.6	7.6 ±1.1

Intelligibility differences of the HTS-C2H stimuli, relative to plain and TTS speech, are subsequently shown in Table 6.6. Equivalent intensity changes (EICs) are also computed (Cooke et al., 2013a). EIC indicates the amount in decibels by which plain speech would need to be changed to acquire the same intelligibility as a given synthetic type (Cooke et al., 2013b).

Table 6.6: *Comparison between the HTS-C2H entry and the plain and TTS baselines. The score is reported in terms of Equivalent intensity changes (EICs) [dB] and changes in keyword scores [percentage points] in parentheses.*

		CS			**SSN**	
HTS-C2H vs	snrHi	snrMid	snrLo	snrHi	snrMid	snrLo
plain	-10.42 (-39.5)	-7.58 (-32.3)	-5.48 (-14.1)	-6.39 (-42.0)	-5.23 (-40.6)	-2.7 (-9.7)
TTS	-3.09 (-14.1)	-1.77 (-6.6)	-0.52 (-1.0)	-2.09 (-17.4)	-1.56 (-10.5)	0.36 (0.8)

Table 6.6 shows that the intelligibility HTS-C2H fHYP stimuli appears to be reduced with respect to both plain and TTS speech. Only in the SSN noise condition with the lowest SSNR, the percentage of recognised word is higher than TTS, even though the increase is not statistically significant.

The intelligibility score of both TTS and HTS-C2H are lower than plain unmodified speech. This result is expected, due to the naturalness and intrinsic clarity of plain speech. The absence of artefacts and the natural prosody of plain speech increase the clarity of such speech with respect to any synthesised speech.

It is worth mentioning that the unmodified TTS voice, that represents the evaluation baseline was trained by the challenge organisers, and it is not the same baseline voice that constitutes the STD HTS-C2H voice *Nick* . The overall quality of the unmodified TTS models seemed higher than the models that were the base of the current set-up, and on which the hyper-articulation transform was applied.

A further explanation of the general reduction of HTS-C2H intelligibility is that the comparisons reported in the result tables are computed with respect to the unmodified TTS and not to the actual STD voice.

Finally, this challenge addressed only one aspect of HTS-C2H. According to the evaluation parameters, the disturbance RMS energy was imposed for the entire duration of each sentence. There were no significant loudness variations within the noise waveform, e.g., sudden increase or absence of disturbance, and no scaling of the speech transformation magnitude was applied. Hence, this listening test does not provide any information about the advantage of moving towards HYO speech. In the future, part-of-speech (POS) dependent control function could have been implemented to save energy – degree of articulation – in the function words and redistribute it in the content words that are crucial to be recognised.

Chapter 7

Discussion and Conclusions

In this chapter, the major findings of the work presented in this thesis are summarised and discussed. The final section presents an overview of the possible directions in which this research could be extended in the future.

7.1 Summary of the main contributions

The research that has led to this thesis has produced several contribution, listed below, which address research questions introduced in § 1.2.

These contributions are listed below.

A reactive speech synthesiser Inspired by the reasoning presented in (Moore, 2007a). This thesis presents the first computational framework that expands the traditional feed-forward speech synthesis system, to create a *reactive speech synthesiser* (Moore and Nicolao, 2011).

Active control of the degree of articulation The reactive speech synthesis framework is the basis for the design of the main contribution that this work, the computational model of Hyper and Hypo articulation theory (C2H), presented in Chapter 4. C2H proposes that speech synthesisers should follow the principle of balance between effort and communicative efficiency. This behaviour has been widely observed in human production, and it is defined by Lindblom's H&H theory (cf. § 2.2.1). The proposed model introduces a negative feedback to mimic the human perception loop. According to the Perceptual Control Theory (PCT) model of human behaviour (cf. § 2.2.2), the sensing feedback loop assesses whether the effects of the speaker's behaviour on the environment match their intentions. A complex computational model, such as C2H, therefore has two main requirements: the energy used in speech realisations must be *controllable*, and such adjustments must be permitted in a *continuous* manner.

Phonetic-contrast motivated transform The main obstacle for an effective implementation of such requirements is that standard speech synthesis does not take into account the *effort* that is involved in the generation process. Since it is not easy to directly quantify the effort/energy involved, this thesis postulates a link between acoustic phonetic contrast, the degree of articulation, and hence the total amount of effort (cf. Figure 4.2). The main idea that is proposed to create such phonetic-contrast motivated transform is to map the acoustic characteristics of normally articulated phones into *low-contrastive* configurations, see § 4.2.3. These configurations are specifically selected from literature to be the least contrastive among groups (vowels) or pairs (competitors) of phones.

TGSM In order to better understand the core principles to control the speech production, a simplified approach to the problem is proposed in Appendix A. The trajectory generation simulation model (TGSM) is a dimensionally-reduced model that compares the vowel realisation to trajectory generation in a 2-D space that simulates the F1-F2 chart. TGSM resulted an effective test environment to check the effect of design solutions – such as the controller function – before deploying them into C2H. This model allows, during the design stage, the importance of control of the distance from competing targets to be highlighted. It is observed that reaching the destinations of a set of targets is as important as avoiding passing through unselected points. Moreover, a link is created here between the length of the trajectory and the effort that is involved in the production.

HTS-C2H An implementation of the C2H model is created to test the behaviour of this new speech synthesis approach. This synthesiser is underpinned by an HMM-based statistical parametric speech synthesis (SPSS) implementation (discussed in Chapter 5). The perception loop consists of an intelligibility measure, such as the extended speech intelligibility index (eSII), which estimates the clarity of the produced speech in relation to the environmental noises. The speech generation algorithm is controlled by the value α that scales the magnitude of the transform in the hyper/hypo-articulation continuum. The phonetic-contrast motivated transform is trained on an *artificially augmented* dataset created by conditioning a standard synthesiser to generate only low-contrastive speech, see § 5.5.

English and Italian voices Two languages are chosen to test the independence of the C2H model from language-related features, such as the ə sound in English. Four different voices were trained along with their specific transforms (cf. § 5.2).

Recursive HTS The HTS 2.2 synthesis software uses the Cholesky decomposition to find the best speech parameter sequence. This method does not allow for model adaption during synthesis. A special implementation was therefore added to HTS in order to perform continuous adjustments in reaction to noise changes (cf. § 5.3).

Acoustic analysis tools As part of a related study at the eNTERFACE workshop in 2012, a set of specific tools for the analysis of speech-in-noise characteristics were needed. In collaboration with other researchers, a Matlab software suite was developed, named XPLIC8 , as shown in Appendix B. This software allows to perform some acoustic and phonetic analysis that are proven to be significant for understanding the modifications of speech in adverse conditions. The XPLIC8 tool, in addition to other analysis techniques, was used to process the P8-Harvard corpus (Stylianou et al., 2012) and the speech outputs of the HTS-C2H systems (cf. § 6.2). Correlation was observed between the modifications, operated by HTS-C2H in presence of different types of noises and different SSNR, and the analysis of natural speech in noise. The emerging modifications, measured on the fHYP HTS-C2H speech, are compatible with the characteristics of Lombard speech.

Effective control of the degree of articulation The main finding derived from the experiments of Chapter 6 can be summarised by the fact that the HTS-C2H is able to control the degree of synthetic speech clarity. The objective evaluation measures, SII and Dau indices, prove that the HTS-C2H full-strength transformations (fHYO and fHYP) modify the speech production intelligibility in the expected directions.

Finally, in § 6.3 an experiment showed how the HTS-C2H output adapts continuously when exposed to follows the non-stationary noise.

7.2 Discussion of the proposed model and research questions

The first objective of the research reported in this thesis was to identify the main behaviours that humans exhibit in speech communications that are not reproduced by state-of-the-art speech synthesisers (cf. *research question n. 1* in § 1.2). Particular attention has been paid to speech modifications required in adverse conditions. The identified type of speech that humans are most likely to reproduce is the result of the Lombard reflex. This behaviour is caused by the negative auditory and sensorimotor feedback loop, applied to the feed-forward direct production, and checks if the outcome has achieved the expected effect on the environment (as per the PCT model of § 2.2.2). The control force that adjusts the speech production aims to balance the correctness of communication and the amount of energy that is used by the process, as per the H&H theory of § 2.2.1. Here, the correct balance of the energy dictates the effort that the listener has to make to understand the message. Therefore, a complete model that can mimic the human speech production system must contain a mechanism to predict the listener's effort to understanding in addition.

The continuous speech production control and perception feedback loop is missing in standard speech synthesisers, which led the computational model of Hyper and Hypo articulation theory (C2H) proposed to answer to the *research question n. 2* of § 1.2. The model formalises the modalities and the computational blocks that need to be considered in order to create a context-aware speech synthesiser. One important feature of the model is that it allows the synthesiser to modify its production at multiple levels of abstraction. The complexity of the context analysis is determined by two joint factors: the communicative intent and the perceptual loop. The model, in principle, can operate on any component of the synthetic pipeline: to transform the message in the conceptualiser, to change the phonetic and prosodic prediction in the linguistic analyser, or to adapt the spectral and temporal characteristics in the waveform synthesiser. For tractability of the problem, the analysis of C2H mainly focuses on the control of spectral and temporal characteristics.

In order to model the H&H control mechanism, the C2H model proposes a link between the acoustic phonetic-contrast and the amount of energy that is used in production. From the literature discussed in § 2.2.1, it is known that the degree of articulation is directly linked to the amount of energy applied, and a higher degree of articulation increases the phonetic distance between pairs of phones. This link represents the answer to the *research question n. 3* of § 1.2. Using this approach, it is no longer necessary to train a model on a range of speaking styles and then select one style according to the prevailing communicative conditions, as had been proposed by other researchers. Rather, the challenge now is to determine the appropriate control strategy that would allow synthesised speech to be interpolated and/or extrapolated along the required dimension.

Assuming the existence of low-contrastive (LC) phone configurations, in which competing phones collapse into similarly perceived sounds, a transformation that can map all phones into their respective LC configuration can be learned. Provided that the transformation exhibits some specific properties such as linearity and inversion, then speech adjustments can also be adapted across the continuum range of styles from hypo- to hyper-articulation. The identification of the LC configurations and the creation of the mentioned mapping function are the crucial elements for the creation of such a transform.

In the HTS-C2H system, the LC configurations are language-dependent phone pairs that are known to be confusable in noisy conditions. The mapping function is an MLLR adaptation that is trained on artificially augmented data which is created by swapping each phone with its competitor. It is observed that it is best practice to train vowel and consonant adaptation functions separately.

An alternative method to define the mapping function, which can be applied only to parametric speech synthesisers, operates in the acoustic model space. A distance metric could be defined between the GMMs of a SPSS, and a set of linear functions could be computed that convert a model into its competitor. This approach would be almost equivalent to the one proposed in this thesis. An advantage of it may be that the set of transforms maps well-trained models, without artefacts that derive from the artificial data creation interfering with the learning process. On the other hand, the augmented data creation allows an interpolation between the available models to produce all required sounds, including unseen ones.

In principle, almost any speech synthesiser can be used in C2H. The only one that cannot readily fit into the framework is the unit selection synthesiser, as this does not allow for scalable style adaptation. If such synthesis had to be used, a work-around solution would be to apply a near-end signal processing to its output. DNN-based and end-to-end synthesisers can be used on this framework, but the complexity of the voice training, the computational load, and limited availability of scalable adaptation techniques has prevented being of use in this thesis. HMM-based synthesis is preferred in the proposed implementation of HTS-C2H, because, unlike the other synthesisers, it allows the speech production to be manipulated, progressively. HMM-based synthesisers with MLLR transformations can indeed perform the interpolation between two opposite degrees of articulation, as per the H&H requirements. Due to the linearity of the adaptation and the generative nature of the GMM, the speech output is also likely to maintain realistic speech-like characteristics even when it is transformed to its extreme hypo- and hyper-articulated boundaries. The standard version of the HTS cannot adapt its generation models continuously. Thus, in order to target the *research question n. 5* of § 1.2, the recursive generative method (Tokuda et al., 1994) was implemented in HTS. This allows the speech outcome to be manipulated at any stage of the generation process. The computational speed is indeed slower than the standard parameter estimation, however the generation process can produce a waveform and correct it in almost real-time on standard CPUs. The recursion algorithm also introduces a reaction delay due to the necessity of accumulate a certain number of past samples in order to generate new ones. This latency of 6 to 10 frames (60 to 100 ms) is comparable with the timescale of delays observed in human speech production.

Acoustic analysis in § 6.2 points out a correlation between the HTS-C2H speech outcome and human speech production recorded in adverse conditions (the P8-Harvard corpus of Appendix B.2.1). This confirms the answer to the research question n. 4 that a purely phonetic-contrast motivated speech adjustment may generate Lombard-like synthetic speech.

7.3 Limitation of the approach

It would be argued that the overall quality of the implemented synthesiser is limited. The generation of a high-quality standard voice model is in itself a quite challenging task. Datasets for speech synthesis training consist of a limited-amount of *read speech* recordings. The size of the data set and the limited variability of the recorded data might result in a voice with limited phonetic variation. Moreover, given that read speech is already quite clear (almost hyper-articulated), there is very limited acoustic space for the transformation to increase the degree of articulation, without ending up with speech that would sound unrealistic.

Another clear limitation of the implementation can be identified in the potential feedback analysis, which considers only the speech intelligibility analysis. Optimising the system with respect to intelligibility indexes generates speech with characteristics similar to humans, but no information can be extracted about the error between the perceived and the intentional messages. This limitation is mainly due to implementation issues that would introduce a long computational latency in the HTS-C2H adaptation control, if more advanced understanding models were used.

The objective evaluation of § 6.4 considers the effectiveness of the proposed transform. High SII and Dau values for low SSNR can be explained by the limitations of such intelligibility indexes. When the noise is much louder than the speech, the index ranges are very small (almost 0) for both transformed and standard speech. Hence, the ratio between these two values can easily diverge to large numbers.

Finally, the subjective evaluation of the implementation of the hyper-articulated output that was conducted in the Hurricane challenge has been reported, although at a first glance, the method would seem to reduce the degree of intelligibility further from the "standard" TTS. However, it must be kept in mind that a) natural speech production was proven to outperform any evaluated speech synthesis methods; b) the TTS baseline against which the intelligibility comparison is computed is not the standard model on which the constant transformation is applied, and unfortunately a direct comparison between the STD model and the Hurricane TTS baseline is not available; c) Table 6.6 actually shows that the intelligibility difference between baseline and HTS-C2H output decreases, when the adverse conditions become more severe. It can be argued that in very low SNR conditions, speech artefacts might interfere less with the intelligibility, and acoustic audibility itself becomes a more prominent limitation.

7.4 Future directions

The model that is presented in this thesis still has a great deal of room for potential improvement.

The C2H description of Chapter 4 comprises several layers of communicative success, and indicates the possible strategies to adjust the speech production. Several extensions to the model could include the implementation and testing of these layers. Four principal dimensions, along which further research could be done, are: the type of speech synthesiser, the voice and transform modelling, the perception feedback, and the control mechanism.

First of all, C2H is flexible in terms of which speech synthesisers is used, as long as the synthesiser complies with certain fundamental characteristics discussed in § 4.2. The speech synthesis field has been going through a huge transformation since the arrival of end-to-end models. The intrinsic autoregressive characteristics of these synthesisers, along with a potentially flexible conditioning mechanism to generate speech samples seems to constitute a suitable synthesiser candidate for C2H. The software and adaptation techniques available might still need some further refinement.

Secondly, it is likely that more recorded speech data might be needed in order to create more robust voice models and consequently more consistent transform. A multi-speaker model approach could also be attempted as rich phonetic variations in the voice model would allow a creation of a more realistic augmented dataset for the transform training stage.

Thirdly, the C2H perception loop describes several sensing levels. The current HTS-C2H implementation considers the intelligibility dimension of the communicative success. A further extension of the implementation could benefit from adding a better model of the listener's understanding of the speech, for instance by including an automatic speech recogniser.

Finally, HTS-C2H controls its production with a single parameter. Despite the simplicity and robustness of this control signal, it does not ensure the stability of the controlled system. A more comprehensive control-theory inspired could be researched to ensure the stability of the system.

HTS-C2H has recently been proven to be able to generate an impact outside the laboratory. Interest has been shown in applying the model to real scenarios, such as in manufacturing environments. Speech-enabled interfaces can be of potential benefit in such environments, where operators might need both hands free to follow prescribed instructions. A major challenge facing the deployment of speech interfaces in manufacturing plants is the level of background noise. A standard speech synthesiser often fails to be intelligible in adverse conditions, and thus an implementation of C2H is a viable solution to increase the clarity of speech in such challenging conditions.

References

Semi-supervised DNN training in meeting recognition, South Lake Tahoe, California and Nevada, Sept. 2014. IEEE.

ANSI. American National Standard Methods for Calculation of the Speech Intelligibility ANSI S3.5-1997, 1997.

S. Arik, G. Diamos, A. Gibiansky, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou. Deep Voice 2: Multi-Speaker Neural Text-to-Speech. *arXiv*, page arXiv:1705.08947, May 2017.

I. P. Association. International Phonetic Association website, 2019. URL `https://www.internationalphoneticassociation.org`.

M. Astrinaki, A. Moinet, J. Yamagishi, K. Richmond, Z.-H. Ling, S. King, and T. Dutoit. Mage - Reactive articulatory feature control of HMM-based parametric speech synthesis. In *SSW8*, pages 207–211, Barcelona, Spain, Aug. 2013.

R. Barra, J. M. Montero, J. Macias-Guarasa, L. F. D'Haro, R. San-Segundo, and R. Cordoba. Prosodic and Segmental Rubrics in Emotion Identification. In *2006 IEEE International Conference on Acoustics Speed and Signal Processing*, pages 1085–1088. IEEE, 2006.

Y. Bengio. Learning Deep Architectures for AI. *MAL*, 2(1):1–127, Nov. 2009.

M. H. Bickhard. Language as an interaction system. *New Ideas in Psychology*, 25 (2):171–187, aug 2007. ISSN 0732118X.

A. W. Black, H. Zen, and K. Tokuda. Statistical parametric speech synthesis. In *ICASSP 2007*, page 1229–1232, 2007.

E. R. Blackmer and J. L. Mitton. Theories of monitoring and the timing of repairs in spontaneous speech. *Cognition*, 39:173–194, May 1991.

B. Blesser. Audio dynamic range compression for minimum perceived distortion. *IEEE Trans. Audio Electroacoust.*, 17(1):22–32, Mar. 1969.

P. Boersma and V. Van Heuven. Speak and unSpeak with PRAAT. *Glot International*, 5(9/10):341–345, 2001.

P. Boersma and D. Weenink. PRAAT: doing phonetics by computer, 2018. URL `http://www.praat.org/`.

G. J. Borden. An interpretation of research of feedback interruption in speech. *Brain and Language*, 7(3):307–319, May 1979.

M. S. Brainard and A. J. Doupe. Auditory feedback in learning and maintenance of vocal behaviour. *Nature Reviews Neuroscience*, 1(1):31–40, Oct. 2000.

M. Brookes. VOICEBOX: Speech Processing Toolbox for MATLAB, 2012. URL `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`.

C. P. Browman and L. Goldstein. Articulatory Phonology: An Overview. Technical report, Haskins Laboratories, New Haven, CT 06511., 1992.

P. Brown and S. C. Levinson. *Politeness*. Some Universals in Language Usage. Cambridge Press, 2nd edition, 1987.

M.-Q. Cai, Z.-H. Ling, and L.-R. Dai. Statistical parametric speech synthesis using a hidden trajectory model. *Speech Communication*, 72:149–159, Sept. 2015.

E. M. Caldognetto, K. Vagges, and F. Ferrero. Intelligibilità e confusione consonantiche in Italiano. *Rivista Italiana di Acustica*, 1988.

Carnegie Mellon University. CMU Pronouncing Phoneme Set, 2015. URL `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`.

H. H. Clark. Speaking in time. *Speech Communication*, 36:5–13, Jan. 2002.

R. A. J. Clark, K. Richmond, and S. King. The Festival Speech Synthesis System, 2012. URL `http://www.cstr.ed.ac.uk/projects/festival/`.

M. Cooke. A glimpsing model of speech perception. In *ICPhS 2003*, pages 1425–1428, Barcelona, Spain, 2003. Department of Computer Science, University of Sheffield, UK.

M. Cooke. A glimpsing model of speech perception in noise. *JASA*, 119(3): 1562–1573–1573, Mar. 2006.

M. Cooke. Discovering consistent word confusions in noise. In *INTERSPEECH 2009*, Brighton, UK, 2009.

M. Cooke, M. L. G. Lecumberri, and J. Barker. The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception. *JASA*, 123(1):414, Jan. 2008.

M. Cooke, C. Mayo, B. Sauert, Y. Stylianou, C. Valentini-Botinhao, and Y. Tang. LISTA Hurricane challenge dataset, 2012. URL `http://www.listening-talker.org/the-hurricane-challenge/`.

M. Cooke, C. Mayo, and C. Valentini-Botinhao. Intelligibility-enhancing speech modifications: the Hurricane Challenge . In *INTERSPEECH 2013*, pages 1–6, Mar. 2013a.

M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, and Y. Tang. Evaluating the intelligibility benefit of speech modifications in known noise conditions. *Speech Communication*, 55(4):1–14, Feb. 2013b.

E. Cooper. TTS Voices , 2019. URL `http://www.cs.columbia.edu/~ecooper/tts/data.html`.

F. Cummins. Periodic and aperiodic synchronization in skilled action. *Frontiers in Human Neuroscience*, 5(170):1–9, 2011.

T. Dau, D. Puschel, and A. Kohlrausch. A quantitative model of the” effective” signal processing in the auditory system. II. Simulations and measurements. *JASA*, 99(6):3623–3631, June 1996a.

T. Dau, D. Puschel, and A. Kohlrausch. A quantitative model of the effective signal processing in the auditory system. I. Model structure. *JASA*, 99(6): 3615–3622, June 1996b.

B. de Gelder, B. M. C. Stienen, and J. Van den Stock. Emotions by Ear and by Eye. In *Integrating Face and Voice in Person Perception*, pages 253–268. Springer, New York, NY, New York, NY, 2013.

G. Degottex and Y. Stylianou. A Full-Band Adaptive Harmonic Representation of Speech . In *INTERSPEECH 2012*, pages 382–385, Portland, OR, USA, 2012.

P. B. Denes and E. N. Pinson. *The Speech Chain.* The Physics And Biology Of Spoken Language. Pickle Partners Publishing, Aug. 2016.

R. Doddipatla, N. Braunschweiler, and R. Maia. Speaker Adaptation in DNN-Based Speech Synthesis Using d-Vectors. In *INTERSPEECH 2017*, pages 3404–3408, ISCA, Aug. 2017. ISCA.

T. Drugman and T. Dutoit. Glottal-based Analysis of the Lombard Effect. In *INTERSPEECH 2010*, pages 2610–2613, Makuhari, Chiba, Japan, Sept. 2010. TCTS Lab, University of Mons, Belgium.

T. Drugman and T. Dutoit. The Deterministic Plus Stochastic Model of the Residual Signal and Its Applications. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(3):968–981, 2012.

H. Dudley. The Vocoder. *Bell Labs Rec*, 18:122–126, 1939.

H. Dudley and T. H. Tarnoczy. The Speaking Machine of Wolfgang von Kempelen. *JASA*, 22(2):151–166, 1950.

D. Erro, I. Sainz, E. Navas, and I. Hernaez. Improved HNM-based Vocoder for Statistical Synthesizers . In *INTERSPEECH 2011*, pages 1809–1812, Florence, Italy, 2011.

G. Fant. *Acoustic Theory of Speech Production*. With Calculations based on X-Ray Studies of Russian Articulations. Walter de Gruyter, Berlin, Boston, 1970.

D. Ferguson and A. Stentz. Anytime RRTs. In *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, pages 5369–5375, Beijing, China, Oct. 2006.

K. Friston and C. Frith. A Duet for one. *Consciousness and Cognition*, 36: 390–405, jan 2015. ISSN 10538100.

R. Fusaroli, J. Rączaszek-Leonardi, and K. Tylén. Dialog as interpersonal synergy. *New Ideas in Psychology*, 32:147–157, jan 2014.

M. Garnier, L. Bailly, M. Dohen, P. Welby, and H. Lœvenbruck. An acoustic and articulatory study of Lombard speech: Global effects on the utterance. In *INTERSPEECH 2006*, pages 2246–2249, Pittsburgh, PA, 2006.

J.-L. Gauvain and C.-H. Lee. MAP estimation of continuous density HMM. In *Proceedings of the workshop on Speech and Natural Language - HLT '91*, page 185, Morristown, NJ, USA, 1992. Association for Computational Linguistics.

H. Giles. *Communication Accommodation Theory*. Negotiating Personal Relationships and Social Identities Across Contexts. Cambridge University Press, 1 edition, Aug. 2016.

E. Godoy, M. Koutsogiannaki, and Y. Stylianou. Approaching speech intelligibility enhancement with inspiration from Lombard and Clear speaking styles. *Computer Speech & Language*, 28(2):629–647, Oct. 2013.

M. Grimm and K. Kroschel. Emotion Estimation in Speech Using a 3D Emotion Space Concept. In *Robust Speech Recognition and Understanding*. IntechOpen, June 2007.

R. Grush. The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and brain sciences*, 27(3):377–442, Jan. 2004.

F. H. Guenther and J. S. Perkell. A neural model of speech production and its application to studies of the role of auditory feedback in speech. *Speech motor control in normal and disordered speech*, page 29–49, 2004.

F. H. Guenther, S. S. Ghosh, and J. A. Tourville. Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language*, 96(2006):280–301, July 2005.

W. J. Hardcastle and N. Hewlett. *Coarticulation*. Theory, Data and Techniques. Cambridge University Press, Nov. 2006.

C. M. Harris. A Study of the Building Blocks in Speech. *JASA*, 25(5):962–969, 1953.

R. J. Hartsuiker and H. H. J. Kolk. Error monitoring in speech production: A computational test of the perceptual loop theory. *Cognitive Psychology*, 42(2): 113–157, 2001.

S. Hawkins. Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, 31:373–405, 2003.

S. Haykin. *Adaptive Filter Theory*. Prentice Hall, 2014.

V. Hazan and R. E. Baker. Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions. *JASA*, 130 (4):2139–2152, 2011.

G. Hesslow. Conscious thought as simulation of behaviour and perception. *Trends in Cognitive Sciences*, 6(6):242–247, 2002.

G. E. Hinton, J. L. McClelland, and D. E. Rumelhart. Distributed representations. In *Parallel distributed processing Explorations in the microstructure of cognition*, pages 77–109. 1984.

R. Hofe. *Biomimetic Vocal Tract Modelling*. PhD thesis, University of Sheffield, June 2011.

R. Hofe and R. Moore. Towards an investigation of speech energetics using 'AnTon': an animatronic model of a human tongue and vocal tract. *Connection Science*, 20(4):319–336, Dec. 2008.

J. N. Holmes. Formant synthesizers: Cascade or parallel? *Speech Communication*, 2(4):251–273, 1983.

J. N. Holmes. A parallel formant synthesizer for machine voice output. In *Computer speech processing*, pages 163–187. Prentice Hall International (UK) Ltd., Feb. 1986.

J. N. Holmes and W. J. Holmes. *Speech Synthesis and Recognition*. 2 edition, Jan. 2001.

A. S. House, C. Williams, M. H. L. Hecker, and K. D. Kryter. Psychoacoustic Speech Tests: A Modified Rhyme Test. *JASA*, 35(11):1899, 1963.

HTS working group. HMM-based Speech Synthesis System (HTS), 2012. URL `http://hts.sp.nitech.ac.jp`.

Q. Hu, K. Richmond, J. Yamagishi, and J. Latorre. An experimental comparison of multiple vocoder types. In *SSW8*, pages 135–140, Barcelona, Spain, Aug. 2013.

Y. Hu and P. C. Loizou. Evaluation of Objective Quality Measures for Speech Enhancement. *IEEE Trans. Speech, Audio & Language Processing*, 16(1): 229–238, Jan. 2008.

A. J. Hunt and A. W. Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *ICASSP 1996*, pages 373–376. IEEE, 1996.

S. L. Hura, B. Lindblom, and R. L. Diehl. On the role of perception in shaping phonological assimilation rules. *Language and Speech*, 35 (Pt 1-2)(1-2): 59–72, Jan. 1992.

ISTC. Istituto di Scienze e Tecnologie della Cognizione, 2019. URL `http://www.pd.istc.cnr.it`.

ITU Radiocommunication Bureau. ITU-R BS.1387-1: Method for Objective Measurements of Perceived Audio Quality , Mar. 2002.

ITU-T. ITU-T P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. ITU, Feb. 2001.

J.-C. Junqua. The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex. *Speech Communication*, 20 (1-2):13–22, Nov. 1996.

P. N. Juslin and P. Laukka. Emotional expression in speech and music: evidence of cross-modal similarities. *Ann. N. Y. Acad. Sci.*, 1000:279–282, Dec. 2003.

J. M. Kates and K. H. Arehart. Coherence and the speech intelligibility index. *JASA*, 117(4):2224, Apr. 2005.

H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication*, 27:187–207, 1999.

R. Kelsey, W. Clinger, and J. Rees. Scheme, 1998. URL `http://www-swiss.ai.mit.edu/~jaffer/Scheme.html`.

D. H. Klatt. Review of Text-to-Speech Conversion for English. *JASA*, 82(3): 737–793, Sept. 1987.

J. Kominek and A. W. Black. CMU-ARCTIC SLT, 2003a. URL `http://festvox.org/cmu_arctic`.

J. Kominek and A. W. Black. *CMU ARCTIC databases for speech synthesis*. Language Technologies Institute School of Computer Science Carnegie Mellon University, 2003b.

T. Koriyama, T. Nose, and T. Kobayashi. Statistical Parametric Speech Synthesis Based on Gaussian Process Regression. *IEEE Journal of Selected Topics in Signal Processing*, 8(2):173–183, Apr. 2014.

M. Koutsogiannaki and Y. Stylianou. Modulation Enhancement of Temporal Envelopes for Increasing Speech Intelligibility in Noise. In *INTERSPEECH 2016*, pages 2508–2512. ISCA, Sept. 2016.

K. D. Kryter. Methods for the Calculation and Use of the Articulation Index. *JASA*, 34(11):1689–, 1962.

H. Lane, M. L. Matthies, F. H. Guenther, M. Denny, J. S. Perkell, E. Stockmann, M. Tiede, J. Vick, and M. Zandipour. Effects of short-and long-term changes in auditory feedback on vowel and sibilant contrasts. *Journal of Speech, Language and Hearing Research*, 50(4):913–927, Aug. 2007.

M. L. G. Lecumberri, J. Barker, R. Marxer, and M. Cooke. Language Effects in Noise-Induced Word Misperceptions. In *INTERSPEECH 2016*, pages 640–644. ISCA, Sept. 2016.

C. H. Lee and J. L. Gauvain. Speaker adaptation based on MAP estimation of HMM parameters. In *Proceedings of ICASSP '93*, pages 558–561 vol.2. IEEE, 1993.

S. Lee, S. Yildirim, A. Kazemzadeh, and S. S. Narayanan. An Articulatory Study of Emotional Speech Production . In *EUROSPEECH 2005*, Lisbon, Portugal, 2005.

C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech & Language*, 9(2):171–185, 1995.

F. A. Leoni, F. Cutugno, and R. Savy. The vowel system of Italian connected speech. In B. P. Elenius K., editor, *ICPhS 1995*, volume 4, pages 396–399, Stockholm, 1995.

W. J. M. Levelt. Monitoring and self-repair in speech. *Cognition*, 14(1):41–104, 1983.

W. J. M. Levelt. *Speaking: From intention to articulation*. The MIT press, 1989.

W. J. M. Levelt, A. Roelofs, and A. S. Meyer. A theory of lexical access in speech production. *Behavioral and brain sciences*, 22(01):1–75, 1999.

S. C. Levinson and J. Holler. The origin of human multi-modal communication. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, 369(20130302):1–9, Sept. 2014.

B. Lindblom. Explaining phonetic variation: a sketch of the H&H theory. *Speech production and speech modelling*, 55:403–439, 1990.

B. Lindblom. Role of articulation in speech perception: Clues from production. *JASA*, 99(3):1683–1692, Mar. 1996.

B. Lindblom. The organization of speech movements: specification of units and modes of control. In *From Sound to Sense*, pages 86–97, Boston, MA, 2004.

B. Lindblom, S. Brownlee, B. Davis, and S.-J. Moon. Speech transforms. *Speech Communication*, 11(4-5):357–368, Oct. 1992.

Z.-H. Ling, L. Qin, H. Lu, Y. Gao, L.-R. Dai, R.-H. Wang, Y. Jiang, Z.-W. Zhao, J.-H. Yang, J. Chen, and G.-P. Hu. The USTC and iFlytek speech synthesis systems for Blizzard Challenge 2007. In *Blizzard 2007*, Dec. 2007.

Z.-H. Ling, K. Richmond, J. Yamagishi, and R.-H. Wang. Articulatory control of HMM-based parametric speech synthesis driven by phonetic knowledge. In *INTERSPEECH 2008*, pages 573–576, Brisbane, Australia, 2008.

B. G. Lipták. *INSTRUMENT ENGINEERS' HANDBOOK: Process Measurement and Analysis*, volume 1. 4th edition, 2003.

É. Lombard. Le Signe de l'Elevation de la Voix - The sign of the rise in the voice. *Ann. Maladiers Oreille, Larynx, Nez, Pharynx - Annals of diseases of the ear, larynx, nose and pharynx*, 37:101–119, 1911.

Y. Lu and M. Cooke. Speech production modifications produced by competing talkers, babble, and stationary noise. *JASA*, Dec. 2007.

J. Ma, Y. Hu, and P. C. Loizou. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. *JASA*, 125(5):3387–3405, May 2009.

D. G. MacKay. Constraints on theories of inner speech. *Auditory imagery*, pages 121–149, 1992.

L. Mangu, E. Brill, and A. Stolcke. Finding consensus in speech recognition: word error minimization and other applications of confusion networks . *Computer Speech & Language*, 14(4):373–400, 2000.

S. L. Mattys, J. Brooks, and M. Cooke. Recognizing speech under a processing load: Dissociating energetic from informational factors. *Cognitive Psychology*, 59(3):203–243, Nov. 2009.

G. A. Miller and P. E. Nicely. An analysis of perceptual confusions among some English consonants. *JASA*, Jan. 1955.

MIVOQ s.r.l. MIVOQ, 2013. URL `https://www.mivoq.it`.

K. Miyanaga, T. Masuko, and T. Kobayashi. A Style Control Technique for HMM-Based Speech Synthesis. In *INTERSPEECH 2004*, pages 1406–1413, Sept. 2007.

S.-J. Moon and B. Lindblom. Two experiments on oxygen consumption during speech production: vocal effort and speaking tempo. In *ICPhS 2003*, Barcelona, Spain, 2003.

R. K. Moore. PRESENCE: A Human-Inspired Architecture for Speech-Based Human-Machine Interaction. *IEEE Transactions on Computers*, 56(9): 1176–1188, Sept. 2007a.

R. K. Moore. Spoken language processing: Piecing together the puzzle. *Speech Communication*, 49(5):418–435, Jan. 2007b.

R. K. Moore. Spoken language processing: time to look outside? In L. Besacier, A.-H. Dediu, and C. Martín-Vide, editors, *2nd International Conference on Statistical Language and Speech Processing (SLSP 2014), Lecture Notes in Computer Science*, volume 8791, pages 21–36, Grenoble, 2014. Springer.

R. K. Moore. Introducing a Pictographic Language for Envisioning a Rich Variety of Enactive Systems with Different Degrees of Complexity. *Int. j. adv. robot. syst.*, pages 1–20, 2016.

R. K. Moore and M. Nicolao. Reactive Speech Synthesis: Actively Managing Phonetic Contrast Along an H&H Continuum. In *ICPhS 2011*, pages 1422–1425, Hong Kong, China, Aug. 2011.

R. K. Moore and M. Nicolao. Toward a Needs-Based Architecture for 'Intelligent' Communicative Agents: Speaking with Intention. *Frontiers in Robotics and AI*, 4:66, Dec. 2017.

M. Morise, F. YOKOMORI, and K. OZAWA. WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications. *IEICE Trans. Inf. Syst.*, E99-D(7):1877–1884, July 2016.

Multimodal Speech Processing Group. The MARY Text-to-Speech System (MaryTTS), 2018. URL `http://mary.dfki.de`.

S. M. Nasir and D. J. Ostry. Somatosensory Precision in Speech Production. *Current Biology*, 16(19):1918–1923, Oct. 2006.

M. Nicolao and R. K. Moore. Establishing some principles of human speech production through two-dimensional computational models. In *SAPA-SCALE workshop 2012*, pages 1–6, Portland, OR, Aug. 2012a.

M. Nicolao and R. K. Moore. Consonant production control in a computational model of hyper & hypo theory (C2H). In *LISTA workshop 2012*, Edinburgh, UK, May 2012b.

M. Nicolao and R. K. Moore. Actively Managing Phonetic Contrast Along an H&H Continuum in Automatic Speech Synthesis . In *SPIN workshop 2013*, Vitoria, Spain, June 2013a.

M. Nicolao and R. K. Moore. Analisi Qualitativa Del Modello C2H Per Il Controllo Del Contrasto Fonetico Nella Sintesi Del Parlato. In *AISV 2013*, pages 1–11, Venice, Italy, Nov. 2013b.

M. Nicolao, C. Drioli, and P. Cosi. Voice GMM modelling for FESTIVAL/MBROLA emotive TTS synthesis. In *INTERSPEECH 2006*, pages 1794–1797, Pittsburgh, Pennsylvania, 2006.

M. Nicolao, J. Latorre, and R. K. Moore. C2H: A Computational Model of H&H-based Phonetic Contrast in Synthetic Speech. In *INTERSPEECH 2012*, pages 1–4, Portland, OR, Sept. 2012.

M. Nicolao, F. Tesser, and R. K. Moore. A phonetic-contrast motivated adaptation to control the degree-of-articulation on Italian HMM-based synthetic voices. In *SSW8*, pages 107–112, Barcelona, Spain, Sept. 2013.

M. Nicolao, A. V. Beeston, and T. Hain. Automatic assessment of English learner pronunciation using discriminative classifiers. In *ICASSP 2015*, pages 5351–5355. IEEE, 2015.

J. J. Odell. *The Use of Context in Large Vocabulary Speech Recognition*. PhD thesis, Cambridge, UK, Mar. 1995.

A. V. Oppenheim and R. W. Schafer. *Discrete-Time Signal Processing*. Pearson Education , 2014.

P.-Y. Oudeyer. The self-organization of speech sounds. *Journal of theoretical biology*, 233(3):435–449, Dec. 2004.

B. Picart, T. Drugman, and T. Dutoit. Analysis and Synthesis of Hypo and Hyperarticulated Speech. In *SSW7*, pages 270–275. 7th ISCA Speech Synthesis Workshop 2010, Sept. 2010.

B. Picart, T. Drugman, and T. Dutoit. Continuous control of the degree of articulation in HMM-based speech synthesis. In *INTERSPEECH 2011*, pages 1797–1800, Florence, IT, 2011.

B. Picart, T. Drugman, and T. Dutoit. Automatic Variation of the Degree of Articulation in New HMM-Based Voices. *IEEE Journal of Selected Topics in Signal Processing*, 8:307–322, Apr. 2014.

M. J. Pickering and S. Garrod. Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, 11(3): 105–110, Mar. 2007.

H. Piéron. *Recherches sur les lois de variation des temps de latence sensorielle en fonction des intensités excitatrices*. *L'année psychologique*, 20(20):17–96, 1913.

A. Postma. Detection of errors during speech production: A review of speech monitoring models. *Cognition*, 77(2):97–132, 2000.

W. T. Powers. *Behavior: the Control of Perception*. Benchmark Publication Inc., 1973.

L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257–286, Feb. 1989.

A. Ragni, K. M. Knill, S. P. Rath, and M. J. Gales. Data augmentation for low resource languages . In *INTERSPEECH 2014*, Singapore, 2014.

T. Raitio, A. Suni, M. Vainio, and P. Alku. Analysis of HMM-Based Lombard Speech Synthesis. In *INTERSPEECH 2011*, pages 2781–2784, Florence, Italy, Aug. 2011a.

T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku. HMM-Based Speech Synthesis Utilizing Glottal Inverse Filtering. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(1):153–165, Jan. 2011b.

T. Raitio, A. Suni, M. Vainio, and P. Alku. Synthesis and perception of breathy, normal, and Lombard speech in the presence of noise. *Computer Speech & Language*, 28(2):648–664, Apr. 2013.

K. S. Rhebergen and N. J. Versfeld. A Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners. *JASA*, 117(4):2181–2192, 2005.

K. S. Rhebergen, N. J. Versfeld, and W. A. Dreschler. Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise. *JASA*, 120(6):3988–3997, 2006.

A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *ICASSP 2001*, pages 749–752. IEEE, 2001.

E. H. Rothauser, W. D. Chapman, N. Guttman, H. R. Silbiger, M. H. L. Hecker, G. E. Urbanek, K. S. Nordby, and M. Weinstock. IEEE Recommended Practice for Speech Quality Measurements. 17(3):225–246, 1969.

L. Saheer, P. N. Garner, J. Dines, and H. Liang. VTLN adaptation for statistical speech synthesis. In *ICASSP 2009*, pages 4838–4841. IEEE, Dec. 2009.

G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, D. Dimitriadis, X. Cui, B. Ramabhadran, M. Picheny, L.-L. Lim, B. Roomi, and P. Hall. English conversational telephone speech recognition by humans and machines. 2017. `https://arxiv.org/abs/1703.02136`.

K. R. Scherer. Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1-2):227–256, Apr. 2003.

M. Schröder. Emotional speech synthesis: A review. In *EUROSPEECH 2001*, 2001.

M. Schröder, M. Charfuelan, S. Pammi, and I. Steiner. Open source voice creation toolkit for the MARY TTS Platform. In *INTERSPEECH 2011*, Florence, Italy, 2011.

T. Scott-Phillips. *Speaking Our Minds: Why human communication is different, and how language evolved to make it special*. Palgrave MacMillan, London, New York, 2015.

N. Sebanz, H. Bekkering, and G. Knoblich. Joint action: bodies and minds moving together. *Trends in Cognitive Sciences*, 10(2):70–76, Feb. 2006.

M. Shannon, H. Zen, and W. Byrne. Autoregressive Models for Statistical Parametric Speech Synthesis. *IEEE Trans. Speech, Audio & Language Processing*, 21(3):587–597, Mar. 2013.

K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. ToBI: A standard for labeling English prosody. In *Second International Conference on Spoken Language Processing*, volume 2, pages 867–870, 1992.

B. F. Skinner. *Verbal Behavior*. B. F. Skinner Foundation, 1948.

H. J. M. Steeneken and T. Houtgast. A physical method for measuring speech-transmission quality. *JASA*, 67(1):318–326, Jan. 1980.

Y. Stylianou. *Harmonic Plus Noise Models for Speech, Combined with Statistical Methods for Speech and Speaker Modification*. PhD thesis, Paris, FR, 1996.

Y. Stylianou, V. Hazan, V. Aubanel, E. Godoy, S. Granlund, M. Huckvale, E. Jokinen, M. Koutsogiannaki, P. Mowlaee, M. Nicolao, T. Raitio, A. Sfakianaki, and Y. Tang. P8 - Active Speech Modifications. Technical report, Metz, France, Nov. 2012.

M. Svenstrup, T. Bak, and H. J. Andersen. Trajectory planning for robots in dynamic human environments. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, pages 4293–4298. IEEE, 2010.

SynSIG committee. Blizzard Challenge 2010, 2010. URL `http://www.synsig.org/index.php/Blizzard_Challenge_2010`.

C. H. Taal, J. Jensen, and A. Leijon. On Optimal Linear Filtering of Speech for Near-End Listening Enhancement. 20(3):225–228, Mar. 2013.

M. Tabain. Variability in fricative production and spectra: implications for the hyper- and hypo- and quantal theories of speech production. *Language and Speech*, 44(1):57–94, Mar. 2001.

M. Tabain and A. Butcher. Stop consonants in Yanyuwa and Yindjibarndi: locus equation data. *Journal of Phonetics*, 27(4):333–357, Oct. 1999.

M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi. Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR. In *ICASSP 2001*, pages 805–808, Salt Lake City, UT, USA, 2001. IEEE.

Y. Tang and M. Cooke. Energy reallocation strategies for speech enhancement in known noise conditions. In *INTERSPEECH 2010*, pages 1636–1639, Makuhari, Chiba, Japan, Sept. 2010.

Y. Tang, M. Nicolao, T. Raitio, E. Jokinen, M. Koutsogiannaki, E. Godoy, S. Granlund, V. Aubanel, A. Sfakianaki, P. Mowlaee, V. Hazan, and Y. Stylianou. XPlic8, 2012.

Y. Tang, M. Cooke, and C. Valentini-Botinhao. Evaluating the predictions of objective intelligibility metrics for modified and synthetic speech. *Computer Speech & Language*, 35:73–92, Jan. 2016.

P. Taylor. Unifying unit selection and hidden Markov model speech synthesis. In *INTERSPEECH 2006*, pages 1758–1761. Engineering Department, University of Cambridge, Trumpington Street, Cambridge, CB2 1PZ, UK, 2006.

P. Taylor. *Text-to-speech synthesis*. Cambridge Press, Cambridge, 2009.

P. Taylor, A. W. Black, and R. Caley. The Architecture of the Festival Speech Synthesis System. In *SSW3*, pages 147–151. International Speech Communication Association, Nov. 1998.

F. Tesser, E. Zovato, M. Nicolao, and P. Cosi. Two Vocoder Techniques for Neutral to Emotional Timbre Conversion. In *SSW7*, pages 1–6, Kyoto, Japan, Sept. 2010.

F. Tesser, G. Paci, G. Sommavilla, and P. Cosi. A new language and a new voice for MARY-TTS. In *9th national congress, AISV (Associazione Italiana di Scienze della Voce)*, Venice, Italy, 2013.

The ALIZ-E team. The ALIZ-E project, 2010. URL `http://www.aliz-e.org/`.

K. Tokuda, T. Kobayashi, and S. Imai. Recursive Calculation of Mel-Cepstrum from LP Coefficients. Technical report, Apr. 1994.

K. Tokuda, T. Kobayashi, and S. Imai. Speech parameter generation from HMM using dynamic features. In *ICASSP 1995*, pages 660–663. IEEE, 1995a.

K. Tokuda, T. Masuko, T. Yamada, T. Kobayashi, and S. Imai. An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features. In *EUROSPEECH 1995*, pages 757–760, Madrid, Spain, Sept. 1995b.

K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi. Hidden Markov models based on multi-space probability distributionfor pitch pattern modeling. In *ICASSP 1999*, 1999.

K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. Speech parameter generation algorithms for HMM-based speech synthesis. In *ICASSP 2000*, 2000.

K. Tokuda, H. Zen, and A. W. Black. An HMM-based speech synthesis system applied to English. In *SSW*, 2002.

K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura. Speech Synthesis Based on Hidden Markov Models. *Proceedings of the IEEE*, 101(5): 1234–1252, May 2013.

UCL Phonetics and Linguistics. SAMPA for Italian, 1989. URL `http://www.phon.ucl.ac.uk/home/sampa/italian.htm`.

C. Valentini-Botinhao, J. Yamagishi, and S. King. Evaluation of Objective Measures for Intelligibility Prediction of HMM-Based Synthetic Speech in Noise. In *ICASSP 2011*, Prague, May 2011.

C. Valentini-Botinhao, J. Yamagishi, and S. King. Mel Cepstral Coefficient Modification Based on the Glimpse Proportion Measure for Improving the Intelligibility of HMM-Generated Synthetic Speech in Noise. In *INTERSPEECH 2012*, pages 1–4, Portland, OR, June 2012.

D. R. van Bergem. Acoustic vowel reduction as a function of sentence accent, word stress, and word class. *Speech Communication*, 12(1):1–23, Mar. 1993.

D. R. van Bergem. Perceptual and acoustic aspects of lexical vowel reduction, a sound change in progress. *Speech Communication*, 16:329–358, Jan. 1995.

A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. WaveNet: A Generative Model for Raw Audio. *arXiv*, Sept. 2016.

L. van Maanen, R. P. P. P. Grasman, B. U. Forstmann, and E.-J. Wagenmakers. Piéron's law and optimal behavior in perceptual decision-making. *Frontiers in Neuroscience*, 5:1–15, Dec. 2011.

R. J. J. H. van Son and L. C. W. Pols. An acoustic description of consonant reduction. *Speech Communication*, 28(2):125–140, June 1999.

W. van Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow, and M. A. Stokes. Effects of noise on speech production: Acoustic and perceptual analyses. *JASA*, 84(3):917–928, Jan. 2005.

W3C-SIF. Introduction and overview of W3C speech interface framework. `http://www.w3.org/TR/voice-intro/`, 2000.

Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous. Tacotron: Towards End-to-End Speech Synthesis. *arXiv*, pages 1–10, Mar. 2017.

M. Wilson and G. Knoblich. The case for motor involvement in perceiving conspecifics. *Psychological Bulletin*, 131(3):460–473, 2005.

Z. Wu, P. Swietojanski, C. Veaux, S. Renals, and S. King. A study of speaker adaptation for DNN-based speech synthesis. In *INTERSPEECH 2015*, pages 879–883, Dresden, Germany, 2015.

Z. Wu, O. Watts, and S. King. Merlin: An Open Source Neural Network Speech Synthesis System. In *9th ISCA Speech Synthesis Workshop*, pages 202–207. ISCA, Sept. 2016a.

Z. Wu, O. Watts, and S. King. The Merlin toolkit, 2016b. URL `http://www.cstr.ed.ac.uk/projects/merlin/`.

W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig. Achieving human parity in conversational speech recognition. 2016. `https://arxiv.org/abs/1610.05256`.

J. Yamagishi and T. Kobayashi. Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training. *IEICE Trans. Inf. Syst.*, 90(2):533–543, 2007.

J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals. Robust speaker-adaptive HMM-based text-to-speech synthesis. *IEEE Trans. Speech, Audio & Language Processing*, 17(6):1208–1230, Aug. 2009.

T. Yoshimura, T. Masuko, K. Tokuda, T. Kobayashi, and T. Kitamura. Speaker Interpolation in HMM-Based Speech Synthesis System. In *EUROSPEECH 1997*, pages 2523–2526, Rhodes, Greece, 1997.

T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. Duration Modeling For HMM-Based Speech Synthesis. In *ICSLP 1998*, Sydney, Australia, Dec. 1998.

T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. Simultaneous Modeling of Spectrum, Pitch and Duration in HMM-Based Speech Synthesis. *IEICE Trans. Inf. Syst.*, 83(11):2099–2107, 1999.

T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. Mixed Excitation for HMM-based Speech Synthesis. In *Eurospeech*, 2001.

S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. The HTK book (for HTK version 3.2). Technical Report July 2000, Cambridge University, 2002.

H. Zen. Acoustic Modeling in Statistical Parametric Speech Synthesis – from HMM to LSTM-RNN . In *MLSLP*, pages 1–10, 2015.

H. Zen and T. Toda. An overview of Nitech HMM-based speech synthesis system for Blizzard Challenge 2005. In *EUROSPEECH 1995*, 2005.

H. Zen, K. Tokuda, and T. Kitamura. Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences. *Computer Speech & Language*, 21(1):153–173, Jan. 2006.

H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda. The HMM-based Speech Synthesis System (HTS) Version 2.0. In *SSW6*, pages 294–299, Bonn, Germany, Aug. 2007a.

H. Zen, T. Toda, and K. Tokuda. The Nitech-NAIST HMM-based speech synthesis system for the Blizzard Challenge 2006. *IEICE Trans. Inf. Syst.*, Dec. 2007b.

H. Zen, K. Tokuda, and A. W. Black. Statistical parametric speech synthesis. *Speech Communication*, 51(11):1039–1064, Nov. 2009.

H. Zen, A. Senior, and M. Schuster. Statistical parametric speech synthesis using deep neural networks. In *ICASSP 2013*, pages 7962–7966. IEEE, 2013.

T.-C. Zorila, V. Kandia, and Y. Stylianou. Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression . In *INTERSPEECH 2012*, Portland, OR, USA, Sept. 2012.

Appendix A

TGSM: a Trajectory Generation Simulation Model

Speech synthesis is a very complex process, which involves the generation and control of high-dimensional parameter vectors. In SPSS, for example, the feature vectors that drive the speech waveform generation can contain more than 200 elements. In such high-dimensional acoustic space, the effects of any modification to the generation process can be very difficult to visualise. Therefore, a trajectory generation simulation model TGSM is proposed, which allows for generating trajectories in a low-dimensional space that can be intuitively displayed. Visual representation is useful to understand different optimisation strategies to design a context-aware synthesiser and to test their effectiveness.

A.1 A phonetically-inspired trajectory space

Speech generation can be regarded as the evolution in time of a feature vector which is input to the waveform generator. Understanding the effect of diverse optimisation criteria on the trajectories is crucial to assess the quality and to tune the parameters of the modifications that are the constituent parts of the complete C2H model.

The simulation model, TGSM, is chosen to operate in a two-dimensional (2D) space, and it is designed to visualise the effect of the different optimisation criteria and control functions. In order to create an intuitive but yet effective model, which maintains the link to the original speech synthesis problem, some principles have to be ensured in its design.

- Since feature vectors in speech synthesis describe connected high-dimension surfaces in the acoustic space, *continuous* trajectories evolving in time are produced when reducing the number of dimensions to 2.

- Speech synthesis aims to realise a sequence of phone targets. The goal of TGSM is to visit a series of target positions in a given order and with predetermined accuracy.

- The 2D space is defined by a set of point coordinates that identify the targets, and by other points that identify the competitors that the trajectory needs to avoid.

- Proximity functions are defined to measure the precision with which the trajectory visits the targets.

- Trajectory-evolution is defined by two components: velocity and direction. Both are *continuously updated* as a reaction to the current trajectory position and to the proximity function stimulus. An example is inspired by observations on human behaviour (van Maanen et al., 2011). That is, the speed at which the trajectory moves towards the target is directly proportional to the distance from it.

- The optimal trajectory can evolve as function of the *motivation* with which the trajectory reacts to the stimulus.

- The system evolves with a discrete time step. Some physiologically-motivated constrains are applied to the maximum distance that can be covered by a trajectory in a single time step.

Even if a strict relationship with the original acoustic space is not emphasised in the above principles, the 2D resulting space can be regarded to have strong similarities with the most common representation of the vowel space, the F1-F2 chart, see Figure A.1 In such space, a vowel is identified by two coordinates: the values of its first two formants. The variation of those values produces a modification of the vowel identity. Therefore, the trajectory generation process can be also regarded as a simple vowel sound generation.

Different strategies to navigate this reduced space are tested. Keeping in mind the analogies to the vowel chart, visualisations in this space can be often extended to the speech synthesis acoustic space.

The TGSM space is defined by a set of points, $\{\mathbf{p}_n\}$, $n = 1, \ldots, N$. Those points can be chosen randomly or to resemble the vowel positions in the F1-F2 chart. At

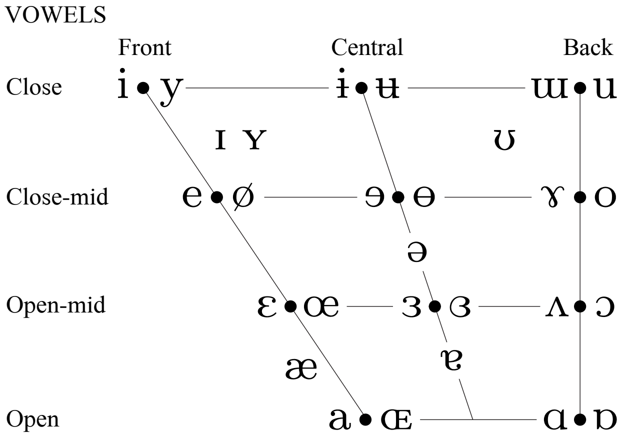


Figure A.1: *The IPA cardinal vowel chart representation. Adapted from (Association, 2019).*

each generation, a subset of points is marked as *targets*, $\{\mathbf{ph}_l\}$, $l \in [1, L]$, $L \leq N$. The $N - L$ inactive points represent the obstacles to avoid – the *competitors* – in the trajectory generation, $\{\mathbf{c}_k\}$, $k \in [1, T]$. The goal is for the trajectory $\boldsymbol{x}_k$ to *visit* every target in the *correct order*, avoiding to pass close to the competitors.

An external observer recognises that a point $\mathbf{p}_n$ is visited by the trajectory, if the trajectory distance from $\mathbf{p}_n$ is smaller that the distances from all the other competing points and it is below a fixed threshold ϵ:

$$\begin{array}{lll} \mathbf{p}_n \text{ is visited} & \text{if} & ||\boldsymbol{p}_n - \boldsymbol{x}_k|| < \epsilon \\ \mathbf{p}_n \text{ is not visited} & \text{if} & ||\boldsymbol{p}_n - \boldsymbol{x}_k|| \geq \epsilon \end{array} \tag{A.1}$$

where $||.||$ is the Euclidean distance. Ideally, ϵ should be close to 0 to minimise recognition errors.

The function that controls the trajectory generation must follow the same principle of the recognition process of eq. A.1 to maximise the trajectory effectiveness. A proximity function which emulates eq. A.1 and determines if the trajectory visits a point must be available. It is also crucial that the function assesses the generation process continuously.

At each k-th step, only one target, $\mathbf{ph}_{\hat{l}}$, is active. Consequently, all other points are considered competitors. The target changes as soon as it is marked as visited. The visiting sequence is fixed – similarly to the C2H intent –, and the following expression is chosen to decide when to switch the active target:

$$\mathbf{ph}_k = \begin{cases} \mathbf{ph}_0 & k \leq 0 \\ \mathbf{ph}_{\hat{l}} & \text{if } \mathbf{ph}_{k-1} = \mathbf{ph}_{\hat{l}} \text{ and } \mathbf{ph}_{\hat{l}} \text{ is not visited} \\ \mathbf{ph}_{\hat{l}+1} & \text{if } \mathbf{ph}_{k-1} = \mathbf{ph}_{\hat{l}} \text{ and } \mathbf{ph}_{\hat{l}} \text{ is visited} \\ \mathbf{ph}_0 & k \geq T \end{cases} \tag{A.2}$$

where $\mathbf{ph}_0$ represents an initial neutral position, $\hat{l}$ is the index of the current target, and T is the maximum time to visit all targets.

The details of the trajectory generation process are discussed later in this chapter as an introduction to the C2H components. Here, two naive examples of trajectory generation, that have no perception feedback, are introduced.

In the first example, the physiological constraint that limits the maximal speed of the trajectory is removed. Therefore, the optimal trajectory consists of a sequence of disconnected points whose coordinates correspond to the targets. The trajectory visits the target $\{\mathbf{ph}_l\}$ for a fixed amount of time and has no transition paths. The trajectory can be expressed as:

$$\boldsymbol{x}_k = \mathbf{ph}_l \quad \forall k, \quad D_{l-1} < k \leq D_{l-1} + d_l \tag{A.3}$$

where d_l is the number of steps – time – for which the trajectory is allow to visit the target, and D_l is the sum of previous durations $D_l = \sum_{i=0}^{l-1} d_i$.

In the second example, the trajectory generation is constrained to use the same number of steps to transition between consecutive targets. Therefore, the trajectory consists of fixed-length segments along the shortest path (w.r.t the Euclidean distance) to the next target. The resulting trajectory is represented by:

$$\boldsymbol{x}_k = \frac{k-D_{l-1}}{d_l}(\mathbf{ph}_l - \mathbf{ph}_{l-1}) \quad \forall k, \quad D_{l-1} < k \leq D_{l-1} + d_l \tag{A.4}$$

in which the time d_l defines the duration of the transition l. In this trajectory, the time d_l for the transition is pre-defined and no deviation from the shortest-path direction is allowed.

Both above examples assume that a target $\mathbf{ph}_{\hat{l}}$ is visited when $\boldsymbol{x}_k = \mathbf{ph}_{\hat{l}}$, i.e., when $\epsilon = 0$ in eq. A.1. In the following paragraphs, two criteria are described that aim to emulate eq. A.1 in order to predict whether the points can be considered as visited by an external observer.

A.2 The proximity function

In order to visit an active target, the trajectory $\boldsymbol{x}_k$ has to get closer to it than to any other surrounding points. This proximity constraint which is specified by a metric or *proximity function* defines an area around each point $\mathbf{p}_n$ (targets and competitors), in which the trajectory is considered as visiting that point. Two types of proximity functions are considered in TGSM, that respectively define two types of areas: the *safe zones* and the *Gaussian mixture zones*.

A.2.1 The safe zones

In the TGSM first approximation the proximity function, a set of safe zones SZ_n are defined and they are assumed to be circular. The SZ_n radius is different for each point and it is defined as half of the distance between $\mathbf{p}_n$ and its closest competitor. An example of such space is displayed in Figure A.2.

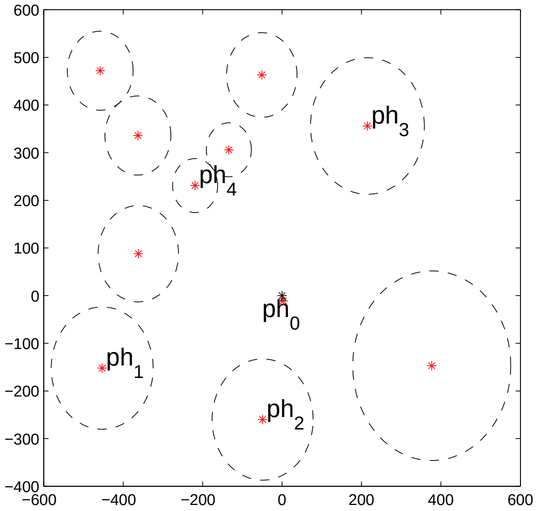


Figure A.2: *Example of the 2-D TGSM space with 11 random points. SZ are displayed with dashed-line circles. Four targets ($\{\boldsymbol{ph}_{1,\dots,4}\}$) and the neutral position, $\{\boldsymbol{ph}_0\}$, are also marked.*

Given the sequence of L target vowels, $\mathbf{SQ} = \{\mathbf{ph}_l\}$ with $l \in [1, L]$, the proximity of the k-th point in the trajectory, $\boldsymbol{x}_k$, to the target $\mathbf{ph}_{\hat{l}}$ is computed by

$$\Delta \boldsymbol{x}_k^{\hat{l}} = ||\mathbf{ph}_{\hat{l}} - \boldsymbol{x}_k|| \tag{A.5}$$

which reproduces the hard decision of eq. A.1.

When the trajectory is inside a SZ, the related point is considered visited. In details,

$$\begin{array}{lll} \mathbf{p}_{\hat{l}} \text{ is visited} & \text{if } \exists \hat{k} \;\; | & \boldsymbol{x}_k \in \mathrm{SZ}_{\hat{l}} \\ \mathbf{p}_{\hat{l}} \text{ is not visited} & \text{if } \forall k & \boldsymbol{x}_k \notin \mathrm{SZ}_{\hat{l}} \end{array} \tag{A.6}$$

Moreover, combining eq. A.5 and eq. A.6, the criterion becomes

$$\boldsymbol{x}_k \in \mathrm{SZ}_{\hat{l}} \Longleftrightarrow \Delta \boldsymbol{x}_k^{\hat{l}} < R_{\hat{l}} \tag{A.7}$$

where $R_{\hat{l}}$ is the $\mathrm{SZ}_{\hat{l}}$ radius.

The link with the correspondent vowel space is that all the points in a safe zone SZ_n can be thought as a set of phonetic realisations of the phoneme $\mathbf{ph}_n$ and they cannot hence be mistaken for any other phone.

Moreover, the neutral position represents a low-energy configuration toward which trajectories tend to converge when the goal is fulfilled or the motivation is not a sufficient stimulus.

A.2.2 The Gaussian mixture zones

The previous criterion represents a quite drastic simplification which has only a weak link to the acoustic representation of speech. A further step towards the real problem is then introduced, which is inspired by the affinity between the 2D space and the F1-F2 chart, Figure A.1.

The positions of the points $\{\mathbf{p}_n\}$ are explicitly selected to be vowels in the F1-F2 chart. The mean formant values are extracted from the audio of the CMU-arctic SLT corpus (American English female voice)(Kominek and Black, 2003b). The mean values, $\boldsymbol{\mu}_{\mathbf{p}_n}$, along with their variances, $\boldsymbol{\sigma}_{\mathbf{p}_n}$, define the GMZ which represents the statistical proximity to the vowel most likely realisation, see Figure A.3. The likelihood of each GMZ target – amplitude of each peak – varies according the evolution of goal sequence.

In Figure A.3, the overlapping GMZ of four targets and the neutral position are displayed. All competitors are assumed to have the same likelihood values. Colour gradient areas describe the perceived proximity of the trajectory to the target. If the trajectory enters a red area, normally the target is marked as visited. The red areas that surround the targets are analogue to the safe zones of the previous paragraph. The hard boundaries of the safe zones are replaced with soft likelihood-based ones. Another important difference that can be observed in Figure A.3 is that the target Gaussian functions overlap with other the competitor ones when their positions are close. Also, subsequent targets interfere in the trajectory realisation detection. In this scenario, visiting a target requires extra energy in order to avoid zones in which the likelihood of two or more points is similar.

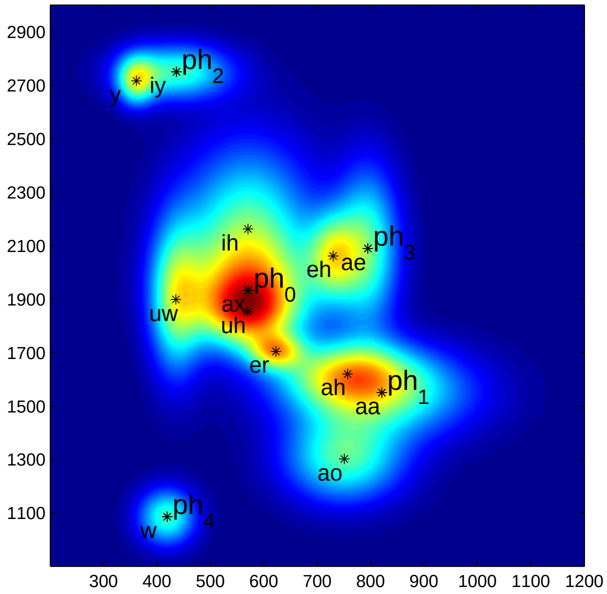


Figure A.3: *Example of the 2-D TGSM space representing 11 English vowels with the proximity likelihood function. Four targets* $\boldsymbol{ph}_{1,\ldots,4}$ *(/aa/, /iy/, /ae/, and /w/), along with the neutral position* $\boldsymbol{ph}_0$*, /ax/, are also shown.*

Inspired by the strategies adopted for the trajectory planning in physical environments (Svenstrup et al., 2010), the proximity function to control the optimal trajectory generation is designed as follows. In contrast to the hard decision of eq. A.5, the GMZ proximity function of the k-th point in the trajectory, $\boldsymbol{x}_k$, to the target $\mathbf{ph}_{\hat{l}}$ returns a soft decision as per,

$$\Delta \boldsymbol{x}_k^{\hat{l}} = G_{\text{SQ}}(\boldsymbol{x}_k, \hat{l}) \tag{A.8}$$

where $G_{\text{SQ}}(\boldsymbol{x}, \hat{l})$ is function of the current position and the current active target, $\mathbf{ph}_{\hat{l}}$, and the target sequence, **SQ**, is fixed.

$G(.)$ can be expanded in a sum of weighted Gaussian functions:

$$G_{\text{SQ}}(\boldsymbol{x}, l) = \sum_{n=0}^{N} a_n(\mathbf{SQ}, l) \cdot G_n(\boldsymbol{x}) \tag{A.9}$$

where $n \in [1, N]$ is the index of all the points in 2-D space, and the functions $G_n()$ are $\mathbf{p}_n$-specific Gaussians that are learned from data examples. The coefficients, $\{a_n(\mathbf{SQ}, l)\}$, are the weighting parameters that depend on the target sequence and on the current active target l. These weighting factors allow the proximity function to change when the targets in **SQ** are visited. Normally, $a_n(\mathbf{SQ}, l) \geq 0$ for the target, and $a_n(\mathbf{SQ}, l) \leq 0$ for competitors. The parameters $\{a_n(\mathbf{SQ}, l)\}$ can also be used to modify the overall motivation associated to the system. The enhancement of a Gaussian function would increase the effort to keep the trajectory moving towards the associated target for a longer time.

Although the control parameters in (A.9) are merely proportional, they are effective in reproducing some formant generation behaviours that can be observed in human speech. An example can be identified in the approximation of the vowel realisations. When motivation (a_n) is low, Gaussian mixture is flat. That results in a wide red areas that generate less precise trajectory.

A target $\mathbf{ph}_{\hat{l}}$ is assumed to be visited when current trajectory $\boldsymbol{x}_k$ reaches a position which maximises the mixture $G_{\text{SQ}}(\boldsymbol{x}, \hat{l})$. Analogously to eq A.6, this relation can be described as follow,

$$\begin{array}{lll} \mathbf{p}_{\hat{l}} \text{ is visited} & \text{if } \exists \hat{k} \;\; | & \Delta G_{\text{SQ}}(\hat{k}, \hat{l}) \leq \epsilon \\ \mathbf{p}_{\hat{l}} \text{ is not visited} & \text{if } \forall k & \Delta G_{\text{SQ}}(k, \hat{l}) > \epsilon \end{array} \tag{A.10}$$

where $\Delta G_{\text{SQ}}(k, \hat{l}) = |G_{\text{SQ}}(\boldsymbol{x}_k, \hat{l}) - G_{\text{SQ}}(\boldsymbol{x}_{k-1}, \hat{l})|$ and ϵ is a minimum-increment threshold. Hence, if the $\Delta G_{\text{SQ}}(k, l)$ moves closer with increment less than ϵ, the target is considered as visited.

Once the sequence SQ and the motivation $\{a_n(\text{SQ}, l)\}$ are defined, the GMZ space is determined. The trajectory is hence generated to navigate such a space to reach the closest local maxima starting from the low-energy configuration. Interesting effects emerge when the GMZ space is left to evolve with different configuration parameters. Since the trajectory direction is motivated by the local maximum, it might not be just a straight line, but bend around the Gaussian surface. If two close targets have comparable amplitude, the trajectory might end in an intermediate point, similarly to the effect of phone approximation. If the motivation is low, the trajectory might move and visit another competitor, resulting in what can be regarded as a mispronunciation.

A.3 Similarities between TGSM and C2H

The simulation model TGSM was introduced to provide a simplified graphical representation of the complex C2H model of § 4. In the following sections, a detailed explanation is reported of the links between the principal components of TGSM and C2H.

A.3.1 Proximity detection

In the simulation model TGSM, the complex communicative state estimation translates into a much simpler task than in C2H (see § 4.3). The state of the system is fully described by the position of the trajectory in the 2D space.

Since the goal of this system is to visit all the targets with a certain level of accuracy, the distance between the trajectory and the active target can be regarded as the effective description of the system state. In particular, TGSM sensors need to extrapolate two pieces of information: the trajectory current *position* and the *distance* to the target. The first measure can be directly derived from the information that are accessible to the process. The current position estimation, or self-monitoring, is direct query to the trajectory generation process, which returns the coordinates of the current position. The second measure requires the non-trivial assumption that the active target and all competitor positions are known by the system. This estimation can be regarded as equivalent to the environment and listener sensors in the C2H space. The distance to the target can be estimated by the proximity functions described in Appendix A.2.1 and Appendix A.2.2. Both conditional rules in eq. A.6 and eq. A.10 return measures of vicinity of the trajectory to the targets. The two rules depend on the positions of targets and competitors. These points can be perturbed by external causes (environmental disturbances). Target SZ shapes can be resized due to external factors. The SZ radius and the strength of the competitor influence are controlled by a scaling factor which expands or reduces the 'target-visited' area and hence the sensitivity of the sensor. Therefore, the trajectory has to move closer than in *clean* conditions to visit it.

In the GMZ space description, the target Gaussian function can be masked by a Gaussian mixture function that introduces a distance measure uncertainty. Trajectory distance to the target needs to be reduced in order for proximity function to detect the target as visited.

This estimation can be regarded as a simulated intelligibility model of the listener, which returns the degree of confusion of an observer (the listener) at measuring the distance of a trajectory realisation to the target (speech intelligibility).

In TGSM, the trajectory length (duration) is the main measure of the energy involved in the generation. In the acoustic space, articulation loci and phone durations can be also regarded as similarly linked to speech production effort.

A.3.2 Trajectory position error

Analogously to the C2H comparator of § 4.4, a function that assesses the distance to the target is also used in the TGSM domain. The TGSM comparator function measures the error between the input target sequence and the trajectory position. The error is expressed either as a binary value, which expresses whether the proximity function labels the target as visited, or a real value that is used as control signal to drive the controller. The error is computed from the measurement about the trajectory state (equivalent to the perception output) and the prior knowledge on the position of the targets (intents). The SZ and the GMZ criteria of Appendix A.2.1 and Appendix A.2.2 respectively are the proximity functions that generate the error. The related eq. A.6 and eq. A.10 express the binary decision that a target is successfully visited.

A.3.3 Trajectory adaptation

In the TGSM 2D space, the trajectory generation needs to be modified according to the proximity error signal. Such adaptation consists of a set of linear transforms that can adjust *direction* and *velocity* of the trajectory path. In eq. A.3 and A.4, two intuitive generation-control examples have been introduced. The output trajectory either visits the target positions with no transitions (infinite speed and no direction), or it moves to the targets in a fixed number of steps until the target coordinates are reached (uniform duration and shortest-path direction), respectively. Both methods cannot control the amount of energy involved in the process and the degree of accuracy of the realisations.

In order to achieve that, the direction and velocity, at which the path $\boldsymbol{x}_k$ moves, must be controlled more effectively. Modifying these two trajectory parameters results in a change of shape and length of the path in the 2D space. The points that the trajectory visits, the durations of the transitions, and the proximity to the targets are affected.

Two examples of trajectory transform are proposed to control the energy in the generation. In the first one, expressed by eq. A.11, the velocity is uniform, but the direction changes according to target positions and the proximity function values, eq. A.6 and A.10.

$$\boldsymbol{x}_k = \quad \boldsymbol{x}_{k-1} + \tfrac{k-D_{l-1}}{d_l} \cdot \boldsymbol{A}(\theta_k) \cdot \Delta\boldsymbol{x}^l_{k-1} \qquad D_{l-1} < k \leq D_{l-1} + d_l \qquad \text{(A.11)}$$

where d_l is the duration of each transition and D_{l-1} is the overall number of time steps to the previous target. $\Delta\boldsymbol{x}^l_{k-1} = (\mathbf{ph}_l - \boldsymbol{x}_{k-1})$ is the vector identifying the shortest path to the next target, and $\boldsymbol{A}(\theta_k)$ is the matrix representing the direction rotation of θ_k-degree, $\theta_k \in [0, 2\pi]$. θ_k is controlled by the error signal measured by the proximity functions. The error signal consists of the logic decision on whether the next step visits a competing target. The rotation in eq. A.11 implies that the position, $\boldsymbol{x}_{k-1}$, can potentially evolves towards any direction of the 2D space at uniform speed. The rotation parameter θ_k is determined by a controlling mechanism reacting to the trajectory position perception error. If $\boldsymbol{x}_k$ visits the active target, target is switched to the next.

The second type of transform is expressed by eq. A.12. The direction is fixed to be the shortest path between targets, whilst the velocity is adapted reacting to the trajectory position.

$$\boldsymbol{x}_k = \quad \boldsymbol{x}_{k-1} + v(e_{k-1}) \cdot \Delta\boldsymbol{x}^l_{k-1} \qquad D_{l-1} < k \leq D_{l-1} + d_l \tag{A.12}$$

The velocity $v(.)$ is defined as a function of the trajectory error to target, e_{k-1}, which consists of the distance between target and trajectory. The Euclidean distance is used to control $\boldsymbol{x}_k$ speed, $e_{k-1} = ||\Delta\boldsymbol{x}^l_{k-1}|| = ||\mathbf{ph}_l - \boldsymbol{x}_{k-1}||$. The function $v(.)$ is inspired by the Pieron's law (Piéron, 1913), which models the relationship between stimulus and response in human sensorimotor system. According to this model, the average human response-time is quicker when the stimulus is stronger. In TGSM, Pieron's law is designed as an exponential function, in which the position error is the stimulus and the velocity is the response. The function can be expressed by

$$v(d_k) = a_1 \cdot e^{\left(d_k + \frac{b_1 \cdot b_2}{b_3}\right)^2} \tag{A.13}$$

in which a_1, b_1, b_2, and b_3 are control constants.

The transforms in the TGSM domain can be linked to those in the real acoustic space. Both adaptations control the degree of accuracy of the input target realisations. Modifications are driven by a control signal which is a function of the realisation error. The different accuracy of both trajectory and speech generation is controlled by coefficients applied to the generation parameters. The amount of energy involved in a realisation emerges from *accuracy* optimisation motivations. If a high degree of accuracy is required, this is proportional to an increment of the standard *effort* involved in the trajectory realisation.

A.3.4 Dynamic trajectory generation

In the TGSM domain, trajectory generation and adaptation are controlled by the error derived by the proximity functions. The sensitivity of the TGSM controller can be interpreted as the correction parameter of the trajectory generation *effort*: i.e., the *motivation* of the system to avoid external observer's trajectory recognition errors. Controlling the trajectory $\boldsymbol{x}$ means to select the appropriate generation parameters that minimise the distance to target.

As expressed in the previous section, Appendix A.3.3, the TGSM transforms operate both on trajectory direction and velocity. The correction parameters are chosen according to the error signal computed by the proximity functions.

Motivation in the TGSM controller modifies the parameters of the proximity functions in order to change the course of $\boldsymbol{x}$. When motivation is standard (STD), the target ph$_l$ safe zone SZ$_n$ is a circle and its radius is half the distance to the closest competitor. The SZ radius can be adjusted by the motivation to be larger or smaller than this distance. These sensitivity changes influence how trajectory $\boldsymbol{x}_k$ moves towards ph$_l$. Generated trajectories can be very different as shown in later examples.

If motivation is low (*HYO*), SZ is enlarged, and consequently proximity functions consider a target as visited at a larger distance than with STD motivation. On the other hand, if motivation is high (HYP), SZ size is reduced, and the trajectory is forced to move for a longer time in order to visit a target. The recognition by an external observer of the visited target sequence, eq. A.1, is clearly influenced by this sensitivity change. In HYP mode, many points in $\boldsymbol{x}_k$ are in the SZ and therefore satisfy the recognition criterion of eq. A.1. The observer has a larger amount of information (points) than with the standard SZ to determine the target sequence. Even if observations are affected by measurement errors or perturbation, they are more likely to produce a correct recognition. Vice-versa, in HYO mode, the accuracy of external observations is decreased.

Motivation affects GMZ sensitivity similarly by changing the weights associated to the Gaussian functions. If a target weight is increased (HYP), the surface gradient of the surrounding area becomes steeper. Since the target-visited criterion for GMZ is that $\Delta G_{\mathrm{SQ}}(k, l) \leq \epsilon$ in eq. A.10, more steps towards the target are needed to obtain an increment smaller than ϵ. If the target weight is reduced, surface is flatter, and the trajectory stops earlier than in STD mode.

The length of the trajectory (number of steps or duration) is hence directly proportional to motivation.

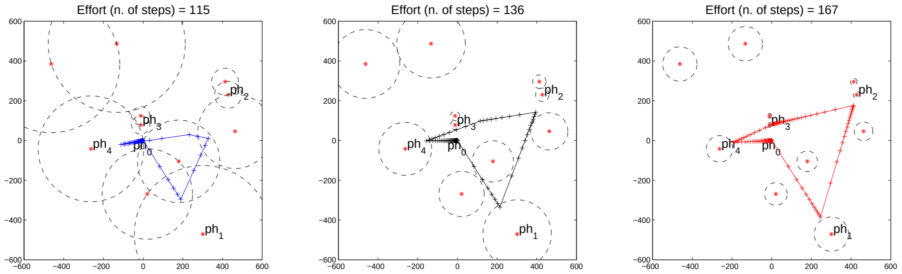
Two transforms, eq. A.12 and eq. A.11, are tested to correct the generation process in TGSM. Their application in relation to the SZ proximity function definition is detailed below.

The first control criterion scales the trajectory speed and aims to find the shortest path which visits each target. The second criterion controls the direction and makes the trajectory to avoid the competitor SZ.

Minimising the distance to target

Minimising the distance between the trajectory $\boldsymbol{x}_k$ and targets ph_n is the first criterion used to compute the desired path. The trajectory speed is adjusted by the transform in eq. A.12. The parameters that control speed are proportional to the trajectory position error. The error derives from the proximity function and it depends from the distance to the target. The trajectory point moves along a straight line to the next target ph_l (the shortest path), until this is visited, then moves to the ph_{l+1} direction. The SZ size is determined by the motivation.

Examples of three trajectories, $\{\boldsymbol{x}_k^{\mathrm{HYO}}\}$, $\{\boldsymbol{x}_k^{\mathrm{STD}}\}$, and $\{\boldsymbol{x}_k^{\mathrm{HYP}}\}$, resulting from applying low, standard, and high motivation respectively, are plotted in Figure A.4. Different trajectory lengths and distances to final targets of the three paths are clearly visible. The diverse evolutions of these trajectories are solely motivated by energy criteria.


Effort (n. of steps) = 115

Effort (n. of steps) = 136

Effort (n. of steps) = 167

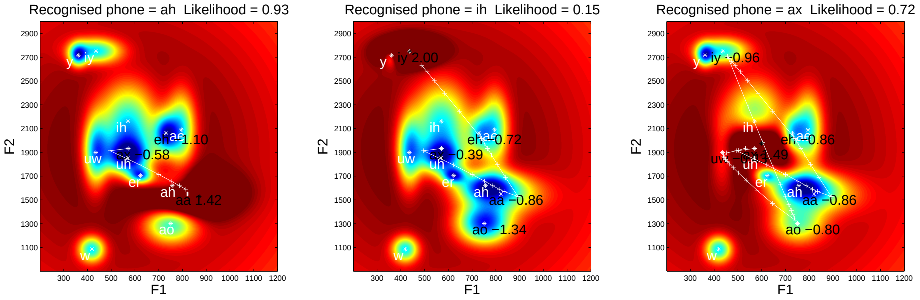
$\boldsymbol{x}_k^{\mathrm{HYO}}$ $\boldsymbol{x}_k^{\mathrm{STD}}$ $\boldsymbol{x}_k^{\mathrm{HYP}}$

Figure A.4: *Trajectory examples with different degrees of motivation to reach the same targets, $\boldsymbol{ph}_{1..4}$, of Figure A.2, through the shortest path.*

This strategy represents a generation algorithm that aims to maximise one aspect of the accuracy (the target distance) and it also controls the effort with which the goal can be achieved. Accuracy is influenced by a *motivation* component, which is defined as the maximum effort that the system is configured to invest in the generation. The motivation component is added to TGSM to represent the influence that such control parameter potentially has in the parallel domain of C2H speech production.

This first control criterion is intuitive and effective, but it clearly has issues arising from its naivety. For example, nothing prevents $\boldsymbol{x}_k$ from crossing the other targets, while it is moving towards its destination. Hence, some competitors might be recognised as visited even though they are not real targets and the observer detects a different path from the desired one.

Example of STD trajectory generation that minimises the distance to the targets in the 2D space defined by the GMZ is shown in Figure A.5. Trajectory velocity at time k is proportional to the steepness of the Gaussian surface surrounding the trajectory point $\boldsymbol{x}_k$.



Towards [aa] Towards [iy] Complete trajectory

Figure A.5: *STD path generated in the GMZ space using linear trajectory planning. Analogies with the F1-F2 vowel space for 11 English vowels are highlighted. The target sequence is [aa],[ah],[iy],[ao],[uw].*

Maximising the distance from competitors

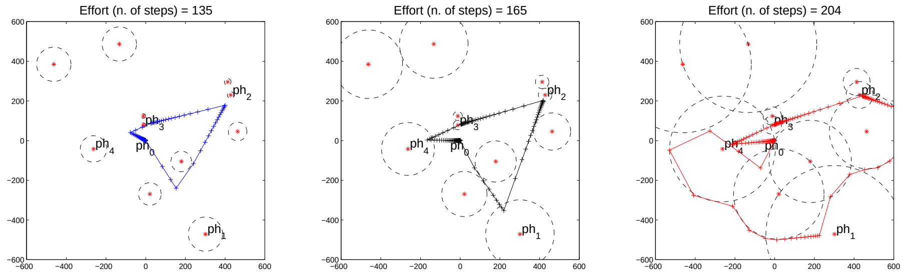
The previous strategy often creates trajectories which intersect some competitor, before reaching the correct target. In order to avoid it, eq. A.11 is used, which adds a rotation coefficient, $\boldsymbol{A}(\theta_k)$, to change the trajectory direction.

The trajectory goal is still to move towards the next target with the minimum length path, however, the control parameter, θ_k, is also function of the position error. This parameter activates an extra dimension in the error signal. If a newly generated point of the trajectory, $\boldsymbol{x}_{k+1}$, moving towards SZ$_l$, falls into a competitor zone, SZ$_m$, an error is signalled.

Examples of the three trajectories, $\{\boldsymbol{x}_k^{\mathrm{HY0}}\}$, $\{\boldsymbol{x}_k^{\mathrm{STD}}\}$, and $\{\boldsymbol{x}_k^{\mathrm{HYP}}\}$, resulting from applying low, standard, and high-effort respectively, are plotted in Figure A.6.

In this case, besides the target SZ$_l$ size, motivation also inversely controls the competitor SZ$_m$ sizes. A high effort reduces the SZ size of the target and increases the SZ area of the competitors to avoid. The trajectory always moves outside all competing SZ. Generated path lengths are longer than those by the previous method.

This strategy emphasises the importance of controlling the distance from competitors as well as from the target, in order to minimise false target recognitions. The main limitations of this method are represented by the binary



x_k^{HY0} x_k^{STD} x_k^{HYP}

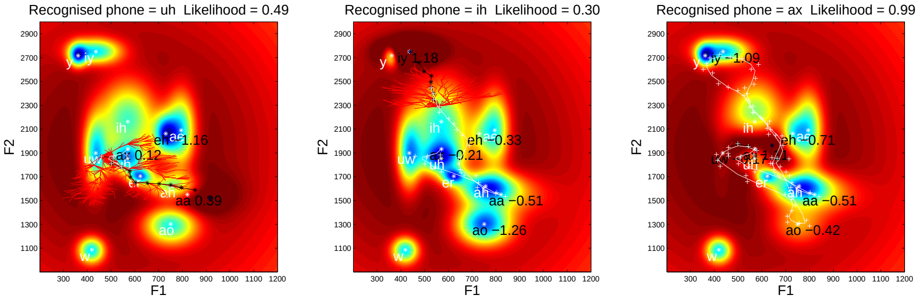
Figure A.6: *Trajectory with different degrees of motivation to reach the same targets, $\boldsymbol{ph}_{1..4}$, of Figure A.2 by maximising the distance from competitors.*

decision function that decides if the target $\mathbf{ph}_l$ is visited, and by the rotation matrix that often selects the sub-optimal direction.

The clear analogy with speech production is that synthesisers need to achieve the correct realisation of each element (phone) in the speech audio, but also ensure that the closest competitors are avoided. The simulation model proves the importance of such control as well as its feasibility.

Other interesting characteristics are that the same phone sequence can generate different trajectories and that the same parametric model is used to generate the vowels and to recognise them (speaker's and listener's model are shared). A suitable recogniser's model in the perception loop is therefore crucial.

A further example of optimal STD path generation that aims to maximise the distance to competitors is shown in Figure A.7. Optimal direction is selected using



Towards [aa] Towards [iy] Complete trajectory

Figure A.7: *Trajectory planning using in the GMZ space. Analogies with the F1-F2 vowel space for 11 English vowels are highlighted. The target sequence is [aa], [ah], [iy], [ao], [uw].*

the rapidly-exploring random tree (RRT) method (Ferguson and Stentz, 2006). The search random tree is displayed in red colour. Once the tree reaches the next target, the first branch of that tree defines the direction of the next movement.

Appendix B

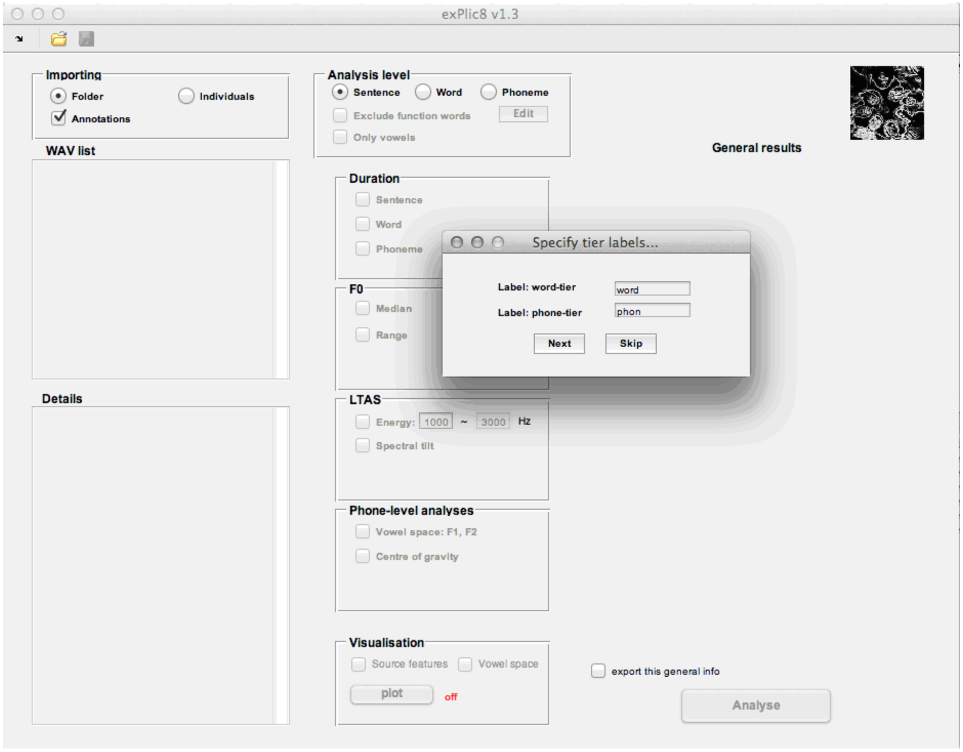
The XPLIC8 analysis software

XPLIC8 is one of the output of the eNTERFACE 2012 project (Stylianou et al., 2012). It consists of a MATLAB-based graphic tool for carrying out a series of analyses on single or batch of signals. It comprises of a set of functions for acoustic-phonetic measurement of speech, as typically used in speech science and phonetics research.

XPLIC8 is able to perform the seven acoustic-phonetic analyses and the two visualization methods on sentence, word and phoneme levels that are listed below:

- Analysis
  - duration [s],
  - F0 median [Hz],
  - F0 range [Hz],
  - long term average spectrum (LTAS) energy between a specified frequency range [dB SIL],
  - spectral tilt [dB/oct],
  - vowel space: F1 [Hz] and F2 [Hz] values,
  - centre of gravity Centre of Gravity (CoG) [Hz].
- Visualization
  - Source features analysis (Raitio et al., 2011b)
    - LPC Spectrum
    - Harmonic-to-noise (HNR) ratio plot

        - ∗ Average glottal flow waveform
    - Vowel space plots
        - ∗ Plot of F1/F2 of tense vs. lax vowels
        - ∗ Plot of mean F1/F2 for all vowels
        - ∗ Plot of centre of gravity for /i/-/ɒ/-/ɔ/



● Folder ○ Individuals
☑ Annotations

● Sentence ○ Word ○ Phoneme
□ Exclude function words
□ Only vowels

□ Sentence
□ Word
□ Phoneme

□ Median
□ Range

□ Energy: 1000 ~ 3000 Hz
□ Spectral tilt

□ Vowel space: F1, F2
□ Centre of gravity

□ Source features □ Vowel space

□ export this general info

Label: word-tier: word
Label: phone-tier: phon

Figure B.1: *The* XPLIC8 *GUI*

Analyses can only be performed on certain levels of accuracy that relies on the existence of corresponding annotation files for the signals. The detailed results from the analyses can be exported in plain text format that can be used as direct input for statistical applications such as the *SPSS* or *R* software for further analysis.

B.1 Analysis algorithms

The analysis algorithms, developed during this project and incorporated in the GUI XPLIC8 , are described below.

B.1.1 F0 estimation

A rough F0 trajectory prediction is performed prior to actual pitch detection. This is done in two stages: The first stage is to high-pass filter the speech signal in order to remove possible low frequency noise, followed by defining the rough F0 range. This is performed by using simple inverse filtering of the speech signal in order to remove most of the formants and then integrating the signal in order to get a signal close to glottal flow. This is done frame-wise with a 40-ms window. The rough fundamental period is estimated by evaluating the autocorrelation sequence of the signal and then finding the maximum peak that corresponds F0 between 50 and 500 Hz. Those frames with low energy or high zero-crossing rate (ZCR) are classified as unvoiced. F0 range is defined as:

$$F0_{\text{min}} = \text{median}(f_0)^{\frac{1.2}{5}} \tag{B.1}$$

$$F0_{\text{max}} = 2.2\text{median}(f_0) \tag{B.2}$$

The actual pitch detection takes place after the initial estimation of the F0 range. The analysis window size is adjusted to the estimated F0 range so that it is twice the lowest fundamental period (2/$F0_{min}$). The glottal inverse filtering method used in F0 estimation is iterative adaptive inverse filtering (IAIF) which estimates the glottal flow signal of the frame using linear prediction such that the fundamental period from the vibratory glottal flow waveform can be estimated. The fundamental period is estimated again finding the maximum peak of the autocorrelation sequence.

For post-processing, two highest peaks are saved: First, the post-processing involves forming a continuous trajectory from the two trajectories. This is based on the relative jump of the trajectories compared to a local F0 median. Second, 5-point median filtering is applied to smooth out outliers. Third, the unvoiced parts are set to zero based on the energy, ZCR, autocorrelation peak value, and gradient index. Fourth, the F0 trajectory is filtered with a 3-point medial filter. Finally, the median F0 is defined as the median of the non-zero values of the trajectory. The $F0_{min}$ and $F0_{max}$ are defined as the minimum and maximum non-zero F0 values of the trajectory.

B.1.2 LTAS energy in specified frequency ranges

The energy is computed as the intensity in sound intensity level (SIL) dB on the specified frequency range. The input sample is windowed with a 5-ms rectangular window without overlap and a 1024-length Fourier transform (using the *fft()* function) is computed for each frame. To obtain the normalized intensity for each frame, the energy in the specified frequency range is normalized by the length of the FFT, the length of the window (in samples) and the sampling frequency. Finally, the normalized intensities of all the frames are summed and the corresponding decibel value is computed by using the reference value $I_0 = 10e^{12}$.

B.1.3 Spectral tilt

The average spectral tilt is computed by fitting a regression line to 1/3-octave band energies of the LTAS in logarithmic scale. The LTAS is computed in 5-ms frames without overlap. For each frame, a 2048-length Fourier transform (with the *fft()* function) is computed and the LTAS is obtained as the mean of the absolute values of the Fourier transforms over all frames. The average energy in the LTAS for each third-octave band is computed and normalized with the width of the band. These values are then transformed to logarithmic scale and a first-degree polynomial fit is estimated (using function *polyfit()*). The average spectral tilt (in dB/octave) is three times the value of the first coefficient of the polynomial.

B.1.4 Vowel space (F1, F2)

The formant extraction tool returns the formant values in the middle point of the selected segment. It uses PRAAT (Boersma and Weenink, 2018) to extract the formant values for each consecutive frame in the selected speech segment and the cheapest paths through those values. Then, the values related to the centre of the time interval are chosen. This function returns formant info for every selected phone and this data is also used to plot the vowel space. Most of the analysis options are already optimised and cannot be changed: Time step = 0.01 s, Maximum formant number $=$ 7, Number of paths to tracks = 5, Formant search range ceiling = 6500 Hz, Pre-emphasis filter lower limit = 50 Hz, Duration of the analysis window (0.025 s). For a detailed description of these parameters, please refer to the online PRAAT manual (Sound to Formant (Burg) and Formant Track)

Formant extraction The sound is re-sampled (Sound: Resample) to a frequency of twice the value of maximum formant and a pre-emphasis filter is also applied (Sound: Pre-emphasize (in-line)). For each analysis window, a Gaussian-like window is applied and the LPC coefficients are as per the algorithm by Burg, as (Childers, D.G., 1978) and (Press, W.H. et al., 1992). The number of ”poles” in this algorithm is set as twice the maximum number of formants. The algorithm finds the best peaks in the selected range of frequency (between 0 Hz and the maximum formant value). Then, all formants below 50 Hz and above the ceiling minus 50 Hz are removed because very low frequency (near 0 Hz) and very high frequency (near the maximum) peaks cannot usually be associated with the vocal tract resonances and they are likely to be artifacts of the LPC algorithm.

Formant tracking After the formant candidate extraction, a tracking on these values is performed in order to rearrange the peaks to obtain the best formant tracks. This command uses a Viterbi algorithm with multiple planes and chooses the cheapest path through all the previously selected peaks (Formant Track). The cost function for one track (e.g. 2) with proposed values $F_{2,i}$ $(i = 1...N$, where N is the number of frames) is:

$$\begin{aligned} CostFunction = \sum_{i=1}^{N} frequencyCost \frac{|F_{2,i} - referenceF_2|}{1000} \\ + \sum_{i=1}^{N} bandWidthCost \frac{B_{2,i}}{F_{2,i}} + \\ + \sum_{i=1}^{N-1} transitionCost |log_2 \frac{F_{2,i}}{F_{2,i+1}}| \end{aligned} \tag{B.3}$$

where *frequencyCost*, *bandWidthCost*, *transitionCost*, and *referenceF2* values are fixed and all set to 1. Analogous formulas compute the cost of other tracks. The procedure will assign those candidates that minimize the sum of all-track costs.

B.1.5 Centre of gravity (CoG)

The Centre of Gravity is a measure of the spectrum energy distribution. The average spectrum on the speech segment is computed. It uses the PRAAT software. Given the complex spectrum, S(f), f is the frequency, the CoG is computed by

$$\int_0^\infty f|S(f)|^p df \tag{B.4}$$

divided by the“energy”

$$\int_0^\infty |S(f)|^p df \tag{B.5}$$

The value of p is chosen to be 2. For further details please refer to the online PRAAT manual (Spectrum: Get the centre of gravity).

B.1.6 Source features

For details of F0 prediction refer to F0 estimation. The polarity is estimated by comparing the positive and negative energy of the glottal flow derivative signal. If the negative energy is greater, the speech signal most likely has positive polarity (and vice versa). After F0 and polarity detection, a suitable window size is selected for estimating the parameters ($3/F0_min$). Iterative adaptive inverse filtering (IAIF) is applied to the speech signal to separate the vocal tract transfer function and the voice source signal. Then, various parameters are extracted, such as:

- F0 and voiced/unvoiced decision [1]
- LPC and FFT spectra of voiced speech
- LPC and FFT spectra of unvoiced speech
- LPC and FFT spectra of vocal tract
- LPC and FFT spectral of voice source
- Speech energy
- Harmonic-to-noise ratio (HNR)
- H1-H2 value of the glottal flow signal
- Normalized amplitude quotient (NAQ)
- Individual glottal flow pulses and their average

The harmonic-to-noise ratio is evaluated by peak picking of the harmonics and then comparing the magnitude difference between the harmonics and the inter-harmonic valleys. These values are averaged to five equivalent rectangular bandwidth (ERB) bands. Normalized amplitude quotient is evaluated for each glottal flow pulse and thus averaged to one value for each frame. Finally, all the estimated unique glottal flow pulses are interpolated to constant length and averaged to estimate the average glottal flow waveform. Parameters are post-processed with median filtering. Statistics of the parameters are evaluated with 95% confidence intervals.

[1] Only available when single WAV file is selected and the analyses are performed on sentence level.

B.2 The P8-Harvard corpus acoustic analysis

The acoustic-phonetic analyses, implemented in XPLIC8 , are used to assess the audio characteristics of the human speech-in-adverse-condition corpus, named *P8-Harvard*.

The P8-Harvard (Stylianou et al., 2012) corpus is a speech dataset that contains audio recordings of speaker pairs (speaker A and speaker B) communicating in adverse conditions. Three different communication barriers are presented to speaker B, during the recordings: *no barrier* (NB), *babble additive noise* (BAB), and *vocoder speech filter* (VOC). Diverse speaking styles are observed as a reaction to the diverse communication conditions. Only speaker A's output is considered in this analysis, as they are not aware of the nature of speaker B's barrier. Therefore, their speech compensations to overcome the communication difficulties are entirely extrapolated from experience and listener's long-term feedback about successfully understanding the message. A more detailed description of the P8-Harvard corpus can be found in (Stylianou et al., 2012)

A summary of the extracted indices is reported below.

F0 extraction the F0 detection is based on glottal inverse filtering and autocorrelation peak detection. The algorithm implemented to extract the F0 and the F0 range from the speech signals is described in Appendix B.

LTAS XPLIC8 estimates the LTAS. This measure returns the energy level in specified frequency ranges.

Formant extraction The XPLIC8 formant extractor algorithm computes the best candidates in the selected segments and finds the cost-efficient paths through those values. The function returns the formant value at the centre of the selected speech segment.

Centre of gravity extraction The XPLIC8 function for measuring the CoG of the average spectrum of the selected speech segment is also implemented in XPLIC8 . CoG is an important measure of the spectrum energy distribution and it also relies on the XPLIC8 tool.

F1-F2 chart In order to isolate the vowel instances in the corpora, all of the speech is segmented using an HTK-based audio-to-text aligner. No manual corrections are performed. For each vowel instance, formant analysis is performed using the PRAAT algorithm. The representative pair of F1 and F2 values for each vowel instance is then taken as the values at the centre of the speech segment. For each vowel, the mean over all of the vowel instances is trimmed, with 95% of the data kept, in order to limit the influence of potential outliers. Then, the convex polygon fit that encompasses all of the data points is calculated in order to represent the maximum area that the points in the vowel space span.

No-Speech detector A Matlab voice activity detection (VAD) function is also implemented to detect parts of speech signal with no proper speech content (NS, no-speech). They can be silence, or pauses between words, or stop-consonant closures, etc. The NS detector relies on a *low-loudness* detection function based on the perceptual speech quality (PSQ) measure (ITU Radiocommunication Bureau, 2002). The function identifies NS segments by selecting the parts in which the loudness is below a certain threshold (15% of the normalised signal loudness). The phonetic and linguistic annotations are used to label some of these NS chunks more specifically. According to the linguistic context in which the NS part is located, the function labels the following type of NS:

- S (speech): part of signal with loudness above threshold.
- NS (no-speech): generic low-loudness part of signal
- NS[SIL] (silence): low-loudness part of signal at the beginning/end of the sentence
- NS[SC] (stop consonant pause): low-loudness part of signal, which is part of a stop consonant inside a word
- NS[IW] (inter-word pause): low-loudness part of signal between two separate words

Mean duration analysis The NS detector is also used to produce reliable information about the duration of the S/NS parts of speech. This allows for a more accurate measurement of different levels of mean duration analysis for each type of condition, since the inter-word durations within utterances can be subtracted to the annotated word durations. The mean word duration (MWD) is the most common duration analysis.

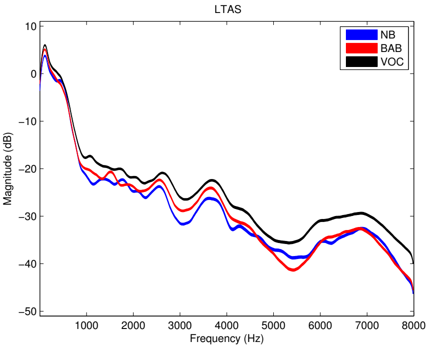
Analyses as NS detection, and phone, word, and sentence duration computations rely on the existence of corresponding annotations for the audio signals. The level of accuracy is hence directly linked to the level of accuracy of the annotation. In recorded speech, manual or automatic force-alignment between audio and text must be performed.

B.2.1 Results

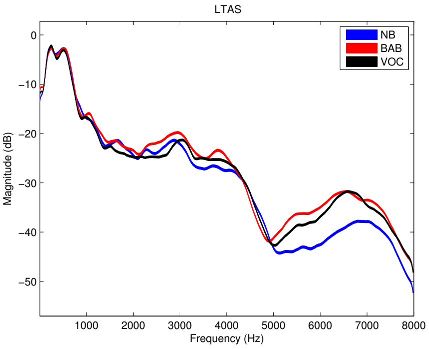
The joint-effort acoustic and phonetic analysis conducted at the eNTERFACE 2012 workshop (Stylianou et al., 2012) is reported here below.

F0 These estimated values for the P8-Harvard corpus are evaluated with the analysis of variance (ANOVA) technique. As expected, the F0 *median* is significantly higher for the female speaker ($p < 0.001$) than for the male one. It is also significantly higher in the VOC condition than in the NB ($p < 0.001$) and BAB conditions ($p < 0.001$). In the BAB, F0 is also higher than NB conditions ($p < 0.001$). F0 range also vary across conditions: it is broader in BAB than in both NB ($p = 0.018$) and VOC ($p < 0.001$). However, F0 range do not seem to differ between the NB and VOC conditions ($p = 0.067$).

LTAS Previous studies correlate the increase of clear speech intelligibility with respect to casual speech to an increment of the LTAS value in the high frequency band 1-3kHz (LTAS13). Figure B.2 depicts the LTAS for speakers A2 (left) and A1 (right) of the P8-Harvard corpus for the three conditions of the P8-Harvard corpus. The male speaker increases his energy above 1000Hz especially for the VOC condition and less on the BAB. The female speaker slightly increases the energy between 2000-4000Hz for the BAB condition and a significant increase above 5000Hz.



(a) male

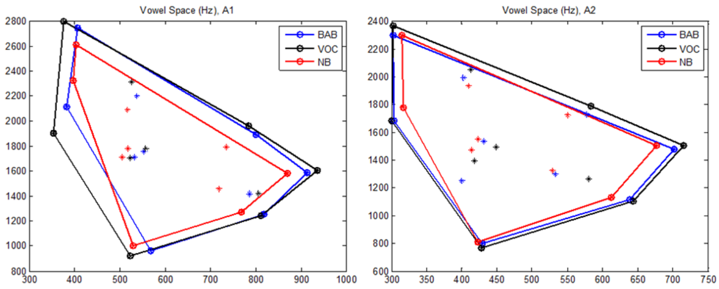


(b) female

Figure B.2: *LTAS of the male (left) and female speaker (right) for three different conditions NB, BAB and VOC. Figure reproduced from (Stylianou et al., 2012).*

LTAS13 for the corpus is evaluated with ANOVA. This measure depends on both speakers and conditions. Post-hoc paired t-tests show that the BAB condition is greater in intensity (mean= -3.1 dB) than the VOC (mean= -3.6 dB), and NB conditions (mean= -6.9 dB) ($p < 0.001$). There is also a significant interaction of speaker and condition ($p < 0.001$). Post-hoc analyses show that there are significant speaker-specific strategies in terms of intensity ($p < 0.001$): for A1, the BAB condition has a greater intensity than the VOC condition (mean difference between VOC and BAB = -2.3 dB), while for A2, the VOC condition has a greater intensity than the BAB condition (mean difference between VOC and BAB = 1.2 dB).

F1-F2 chart Figure B.3 depicts the largest-area polygon fit that identifies the vowel space in the three noise conditions for speakers A1 and A2. The picture shows the 4 tense and 6 lax vowels (95% trimmed means). Per-vowel analysis is



Speaker A1 (female) Speaker A2 (male)

Figure B.3: *F1-F2 chart of the speaker A1 and A2 recordings in the three NB, VOC, and BAB conditions. The largest-area polygon fit for the 4 tense and 6 lax vowels (95% trimmed means). Figure reproduced from (Stylianou et al., 2012).*

run on the values using a mixed-model ANOVA, with vowel as a between-subjects factor, and condition (NB, BAB, VOC) as a within-subjects factor. The analysis shows a significant condition effect on all three vowels /i/, /ɒ/ and /ɔ/ for speaker A1 ($p = 0.0398$) but no effect for speaker B. So, for speaker A, vowel space expands as follows: NB < BAB < VOC.

To explain these results, it must be kept in mind that speakers A have no knowledge on the type of disturbance to which speakers B are exposed. Therefore, they try to adopt the best communicative strategy from their experience repertoire that allows them to successfully transfer information. The success of their communication attempts is determined by the spoken feedback from speakers B.

No-speech parts The No-Speech (NS) detector of XPLIC8 is applied to the P8-Harvard database. Table B.1 contains the number of NS segments for each category, speaker and condition and Figure B.4 has the average number of the total number of inter-word pauses per each utterance.

Table B.1: *Instances of different types of NS returned by* XPLIC8 *No-Speech analysis. SIL identifies start-/end-utterance silences, SC are the stop-consonant silences, and IW are inter-word pauses. NC reports the number of not classified silences. Adapted from (Stylianou et al., 2012).*

	A1			A2		
NS type	NB	BAB	VOC	NB	BAB	VOC
NS[SIL]	295	293	298	277	276	276
NS[SC]	437	492	529	433	470	520
NS[IW]	208	274	375	147	186	257
NS[NC]	1161	1386	1615	947	1059	1247

The results show an increasing number of NS parts in the speech along with the difficulties in the communication. The instances of each type of NS follow the same trend: $\#\text{NS}[.]_{\text{VOC}} > \#\text{NS}[.]_{\text{BAB}} > \#\text{NS}[.]_{\text{NB}}$ for both speakers, even though the male speaker tends to compensate less for the adverse conditions, as reported by other analysis. A significant increase of NS[IW] is observed between the VOC barrier and the other two conditions in both speakers, as Figure B.4 explicitly shows. This confirms that when the communication channel is really destructive and the speaker has no direct access to assess the channel, the most likely speaker's strategy is to greatly decrease the speaking rate.

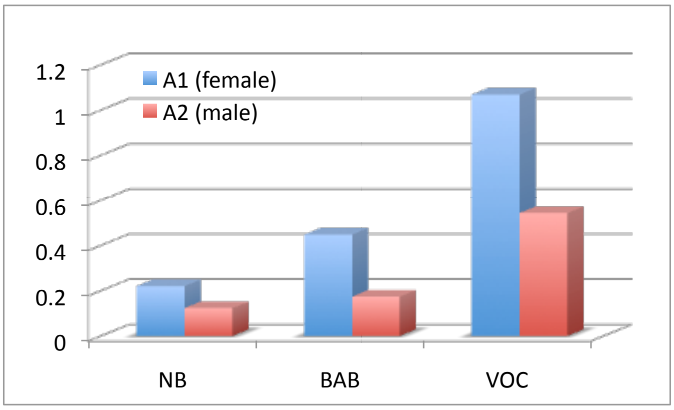


Figure B.4: *Average number of inter-word pauses (NS[IW]) for each utterance in different conditions. Figure reproduced from (Stylianou et al., 2012).*

Further insight can be gained by looking at the durations of the different silence categories. Figure B.5 shows that, apart from expected silences (NS[SIL]), all types of silences – NS[IW] particularly – undergo duration increase from NB to BAB to VOC. In contrast, speech durations remain stable, highlighting a possible speaker strategy of reducing speech rate by detaching words.

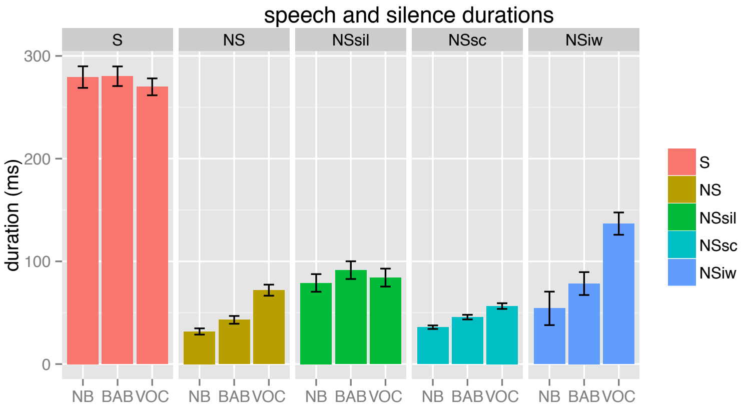


Figure B.5: *Mean speech and silence durations for speakers A1 and A2 across NB, BAB and VOC. Error bars are 95% confidence interval. Figure reproduced from (Stylianou et al., 2012).*

MWD The mean word duration (MWD) for each type of condition is measured accurately using the silent detector, since the inter-word durations within utterances can be identified and subtracted to the word durations.

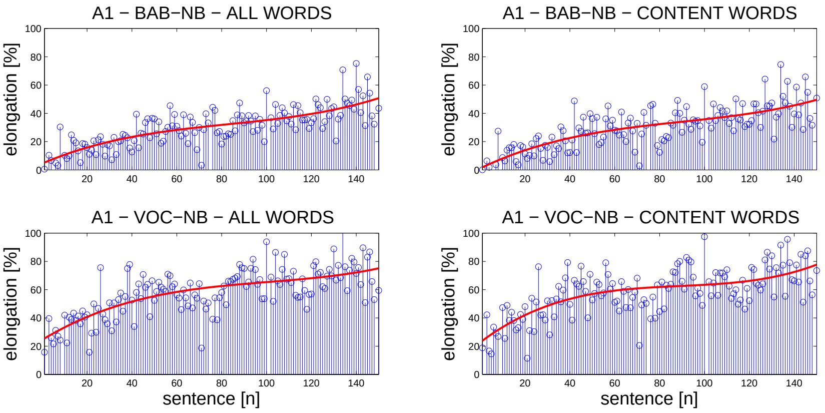


Figure B.6: *Longitudinal evolution of the MWD elongation for speaker A1. The MWD for all the words is shown on the left, whereas there is the content-word MWD only. The 3rd-order polynomial curve that fits the data is also displayed. Figure reproduced from (Stylianou et al., 2012).*

Figure B.6 and Figure B.7 display the change of MWD (word elongation) in the VOC and BAB conditions with respect to the NB condition, during the

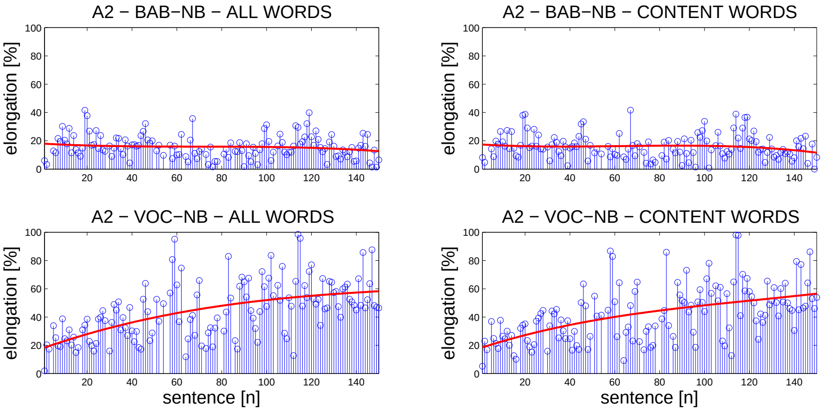


Figure B.7: *Longitudinal evolution of the MWD elongation for speaker A2. The MWD for all the words is shown on the left, whereas there is the content-word MWD only. The 3rd-order polynomial curve that fits the data is also displayed. Figure reproduced from (Stylianou et al., 2012).*

experimental sessions. First observation is that all speakers elongate their speech production, especially in the worst condition (VOC). This evolves along the sessions. However, this is not consistent between the two speakers for the BAB condition. Speaker A2 maintains the all MWD and the content MWD stable. Speaker A2 is found to be generally less effective in the compensation, he slightly elongated the speech ($\sim 20\%$), only in the VOC barrier case but he does not adjust his speech any further. This lack of efficiency is confirmed by the amount of the errors the listener made which are much more compared to the errors he made during the session of speaker A1.

In Figure B.6 and Figure B.7 the red line is a 3rd-order polynomial fitting curve that shows the data trend. Three different stages emerge in all sessions, particularly for speaker A1 and the most stressful (VOC) condition. At the beginning, the speakers start with their normal speech style (i.e. almost the same MWD as the NB condition), but as soon as they receive intelligibility feedbacks from the listener, they adapt their speech realisation by increasing the effort (i.e. word duration). Hence, an elongation increment is seen at the beginning of the session. In the central part, the elongation w.r.t NB is constant. The hypothesis is that speakers and listeners agree that the current elongation is effective for the communicative conditions, and no further adaptation is needed. In the final part of the experiment, an increasing MWD elongation is measured, especially for speaker A1. It is hypothesised that she is trying to overcome listener's new difficulties. These might be due to the listener's fatigue after extensive exposure to such difficult communicative conditions. Understanding a speech message in a severe adverse condition requires considerable cognitive load. In the same conditions, speaker A2 seems to cease making the effort to elongate, maybe due to a lack of motivation towards the end of the session.

Appendix C

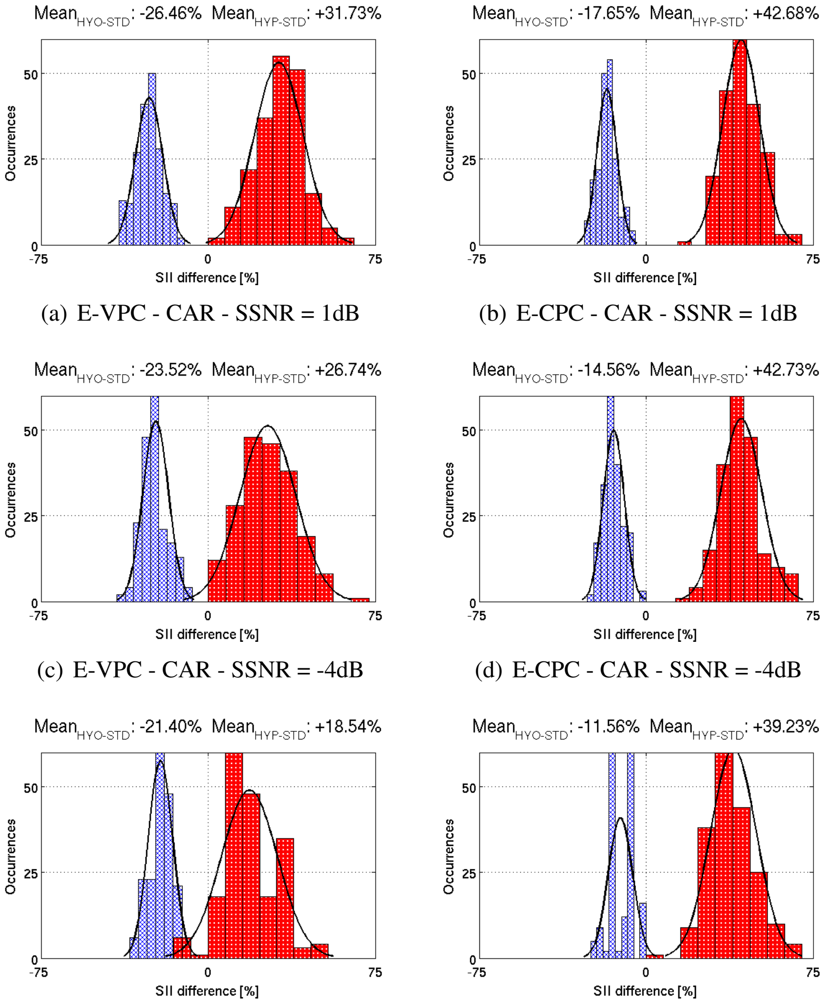
Detailed objective results

This part reports the complete list of the objective evaluation results computed with two different intelligibility-in-noise estimation methods: the SII (ANSI, 1997) and **Dau** (Dau et al., 1996b) indexes.

Implementations of the SII and Dau indexes are used to estimate the intelligibility of the HTS-C2H speech samples in high, mid, and low SNR noise conditions as per the experimental settings in § 6.1.

C.1 *SLT*

The audio samples produced with the English female voice *SLT* results are evaluated with SII in Figure C.1, Figure C.2, and Figure C.3, and with Dau in Figure C.4, Figure C.5, and Figure C.6.



(a) E-VPC - CAR - SSNR = 1dB

(b) E-CPC - CAR - SSNR = 1dB

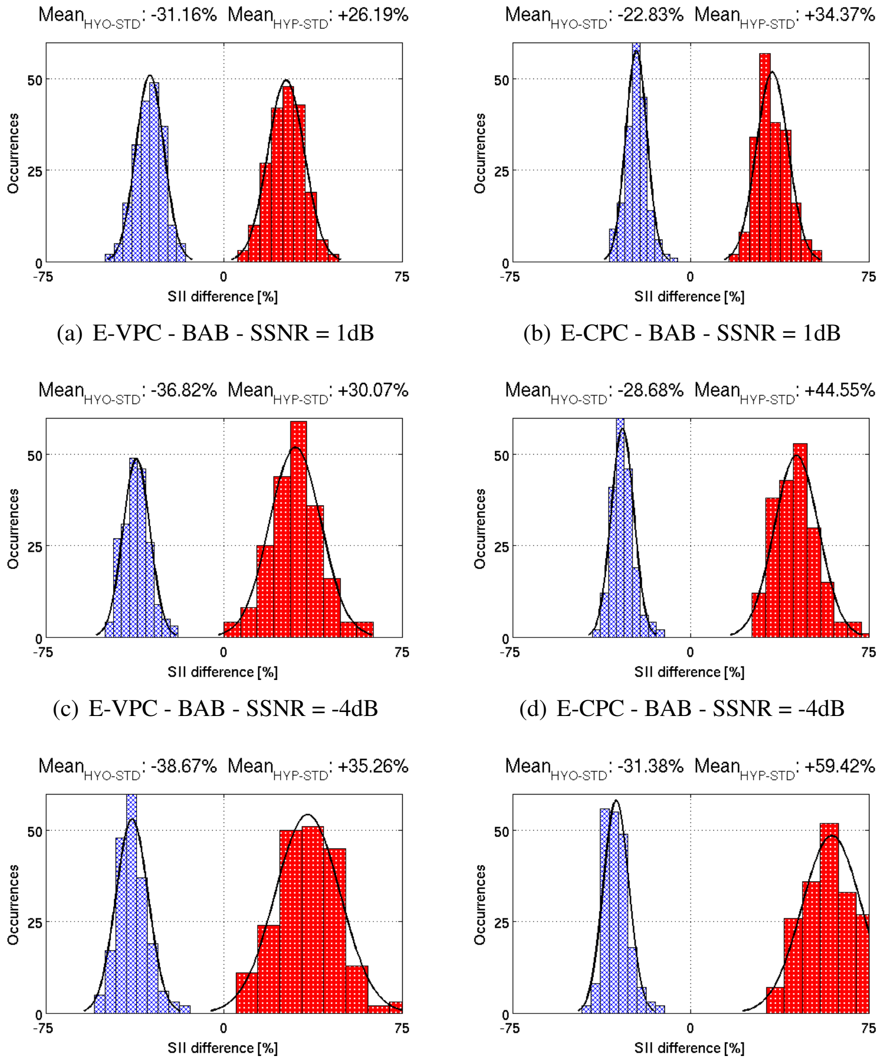
(c) E-VPC - CAR - SSNR = -4dB

(d) E-CPC - CAR - SSNR = -4dB

(e) E-VPC - CAR - SSNR = -9dB

(f) E-CPC - CAR - SSNR = -9dB

Figure C.1: *Distribution of the **SII** value differences (in percentage) between HYO and STD speech (blue-crossed histograms), and between HYP and STD speech (red-dotted histograms) controlled with E-VPC and E-CPC separately. Speech signals are mixed with **CAR noise** with high, mid, and low SSNR.*



Mean$_{HYO\text{-}STD}$: -31.16% Mean$_{HYP\text{-}STD}$: +26.19%

(a) E-VPC - BAB - SSNR = 1dB

Mean$_{HYO\text{-}STD}$: -22.83% Mean$_{HYP\text{-}STD}$: +34.37%

(b) E-CPC - BAB - SSNR = 1dB

Mean$_{HYO\text{-}STD}$: -36.82% Mean$_{HYP\text{-}STD}$: +30.07%

(c) E-VPC - BAB - SSNR = -4dB

Mean$_{HYO\text{-}STD}$: -28.68% Mean$_{HYP\text{-}STD}$: +44.55%

(d) E-CPC - BAB - SSNR = -4dB

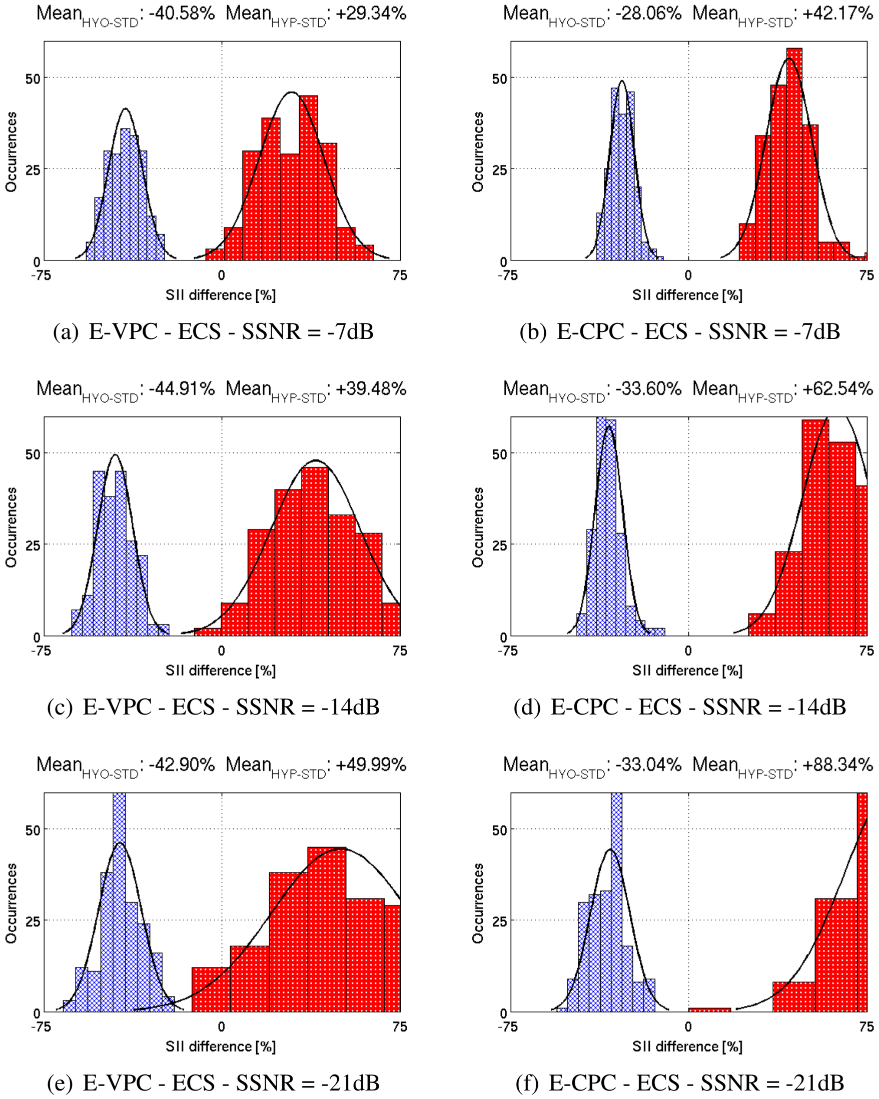
Mean$_{HYO\text{-}STD}$: -38.67% Mean$_{HYP\text{-}STD}$: +35.26%

(e) E-VPC - BAB - SSNR = -9dB

Mean$_{HYO\text{-}STD}$: -31.38% Mean$_{HYP\text{-}STD}$: +59.42%

(f) E-CPC - BAB - SSNR = -9dB

Figure C.2: *Distribution of the* ***SII*** *value differences (in percentage) between HYO and STD speech (blue-crossed histograms), and between HYP and STD speech (red-dotted histograms) controlled with E-VPC and E-CPC separately. Speech signals are mixed with* ***BAB noise*** *with high, mid, and low SSNR.*



(a) E-VPC - ECS - SSNR = -7dB

(b) E-CPC - ECS - SSNR = -7dB

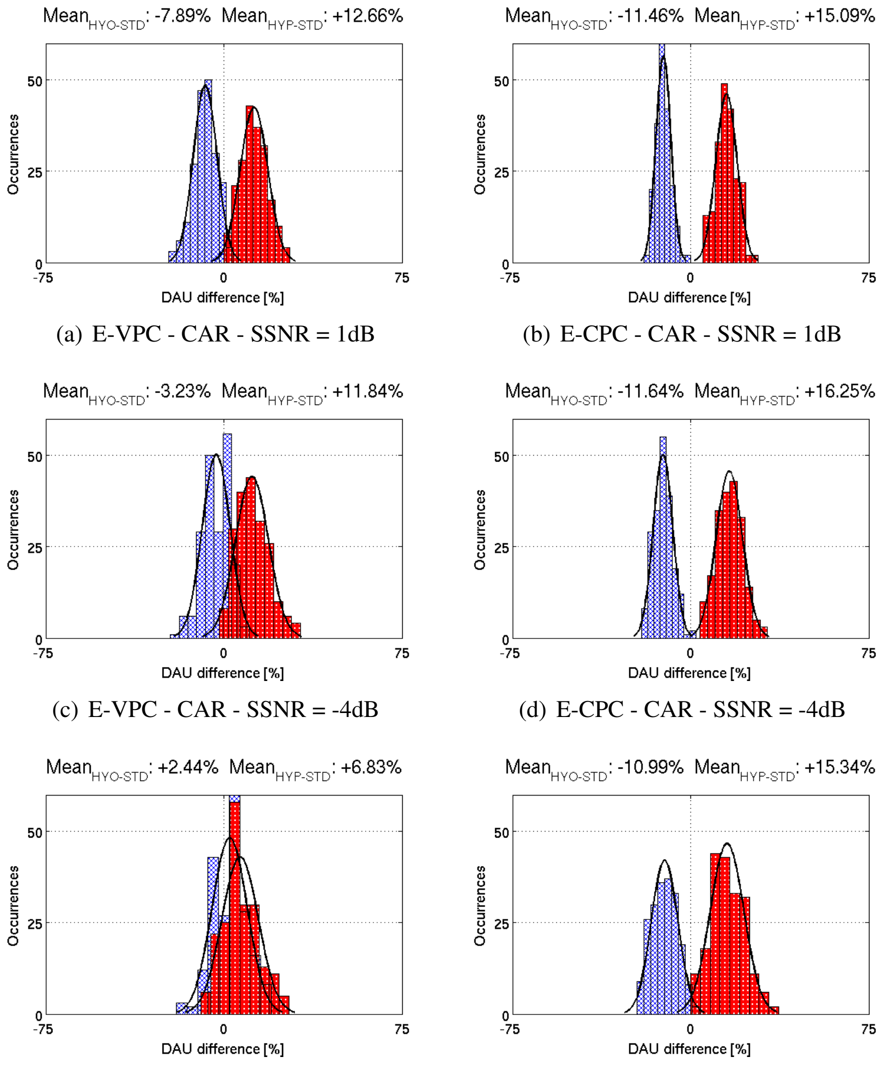
(c) E-VPC - ECS - SSNR = -14dB

(d) E-CPC - ECS - SSNR = -14dB

(e) E-VPC - ECS - SSNR = -21dB

(f) E-CPC - ECS - SSNR = -21dB

Figure C.3: *Distribution of the **SII** value differences (in percentage) between HYO and STD speech (blue-crossed histograms), and between HYP and STD speech (red-dotted histograms) controlled with E-VPC and E-CPC separately. Speech signals are mixed with **ECS noise** with high, mid, and low SSNR.*



(a) E-VPC - CAR - SSNR = 1dB

(b) E-CPC - CAR - SSNR = 1dB

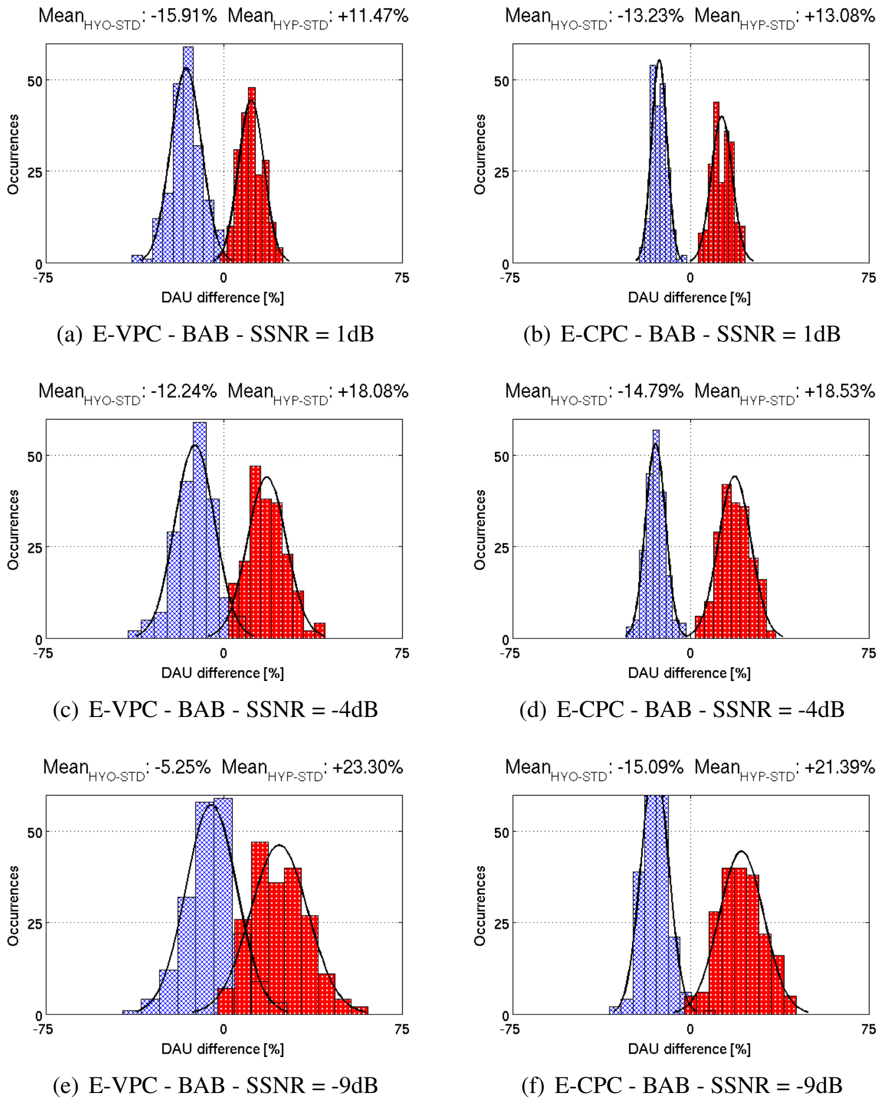
(c) E-VPC - CAR - SSNR = -4dB

(d) E-CPC - CAR - SSNR = -4dB

(e) E-VPC - CAR - SSNR = -9dB

(f) E-CPC - CAR - SSNR = -9dB

Figure C.4: *Distribution of the* ***Dau*** *value differences (in percentage) between HYO and STD speech (blue-crossed histograms), and between HYP and STD speech (red-dotted histograms) controlled with E-VPC and E-CPC separately. Speech signals are mixed with* ***CAR noise*** *with high, mid, and low SSNR.*



(a) E-VPC - BAB - SSNR = 1dB

(b) E-CPC - BAB - SSNR = 1dB

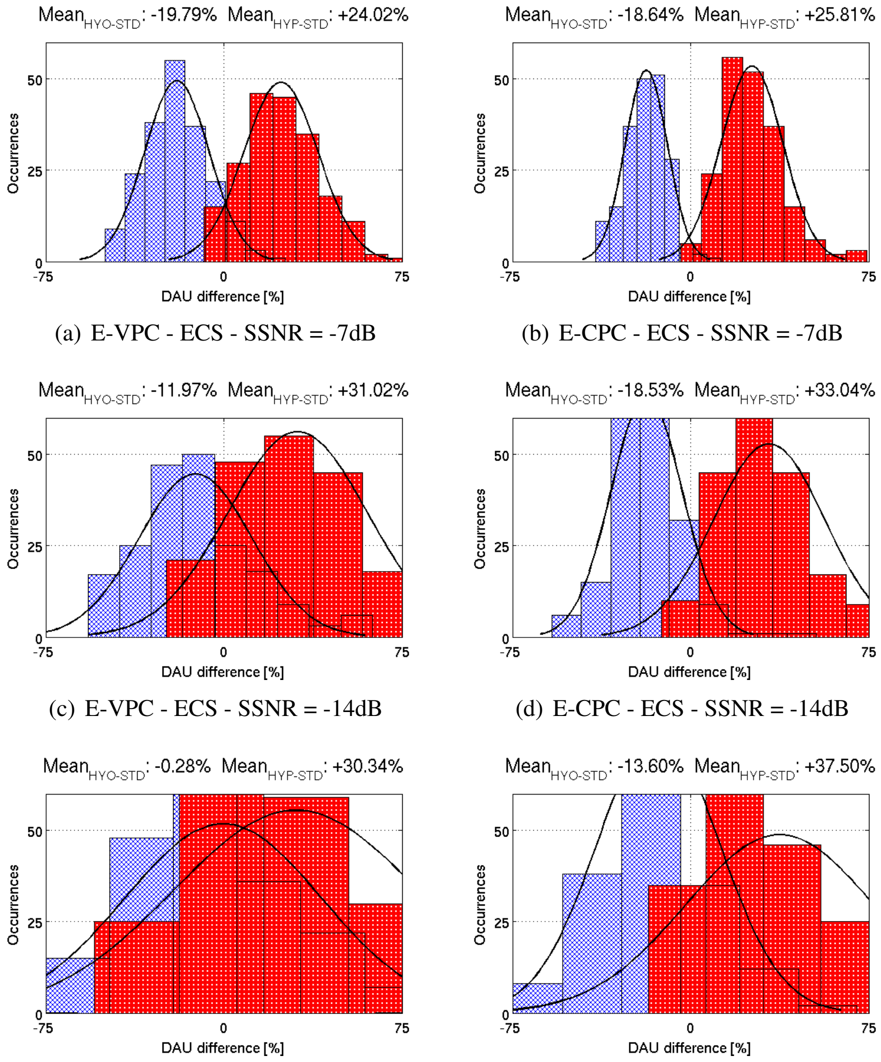
(c) E-VPC - BAB - SSNR = -4dB

(d) E-CPC - BAB - SSNR = -4dB

(e) E-VPC - BAB - SSNR = -9dB

(f) E-CPC - BAB - SSNR = -9dB

Figure C.5: *Distribution of the* ***Dau*** *value differences (in percentage) between HYO and STD speech (blue-crossed histograms), and between HYP and STD speech (red-dotted histograms) controlled with E-VPC and E-CPC separately. Speech signals are mixed with* ***BAB noise*** *with high, mid, and low SSNR.*



(a) E-VPC - ECS - SSNR = -7dB

(b) E-CPC - ECS - SSNR = -7dB

(c) E-VPC - ECS - SSNR = -14dB

(d) E-CPC - ECS - SSNR = -14dB

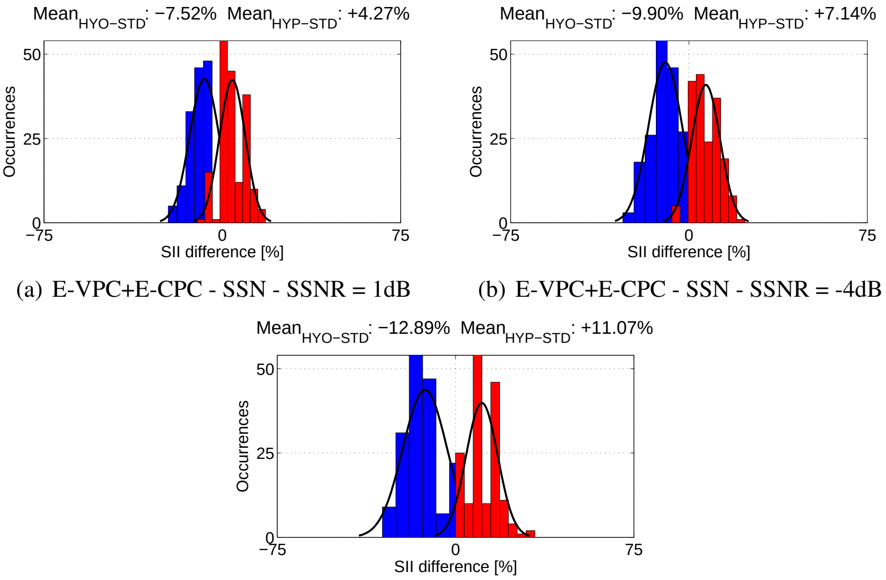
(e) E-VPC - ECS - SSNR = -21dB

(f) E-CPC - ECS - SSNR = -21dB

Figure C.6: *Distribution of the* ***Dau*** *value differences (in percentage) between HYO and STD speech (blue-crossed histograms), and between HYP and STD speech (red-dotted histograms) controlled with E-VPC and E-CPC separately. Speech signals are mixed with* ***ECS noise*** *with high, mid, and low SSNR.*

C.2 *Nick*

The audio samples produced with the English male voice *Nick* results are evaluated with SII in Figure C.7 and Figure C.8, and with Dau in Figure C.9, and Figure C.10

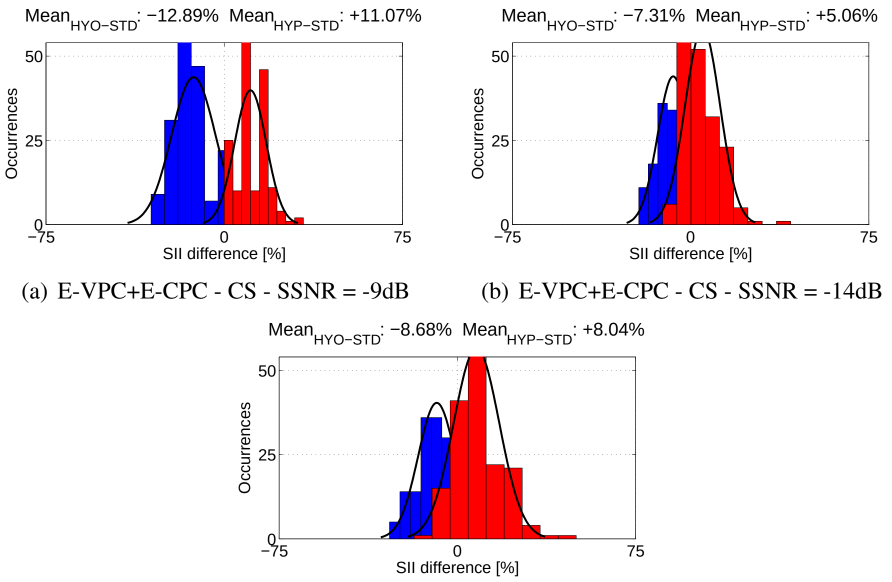


(a) E-VPC+E-CPC - SSN - SSNR = 1dB

(b) E-VPC+E-CPC - SSN - SSNR = -4dB

(c) E-VPC+E-CPC - SSN - SSNR = -9dB

Figure C.7: *Distribution of the* ***SII*** *value differences (in percentage) between HYO and STD speech (blue histograms), and between HYP and STD speech (red histograms) controlled with E-VPC and E-CPC simultaneously. Speech signals are mixed with* ***SSN noise*** *with high, mid, and low SSNR.*

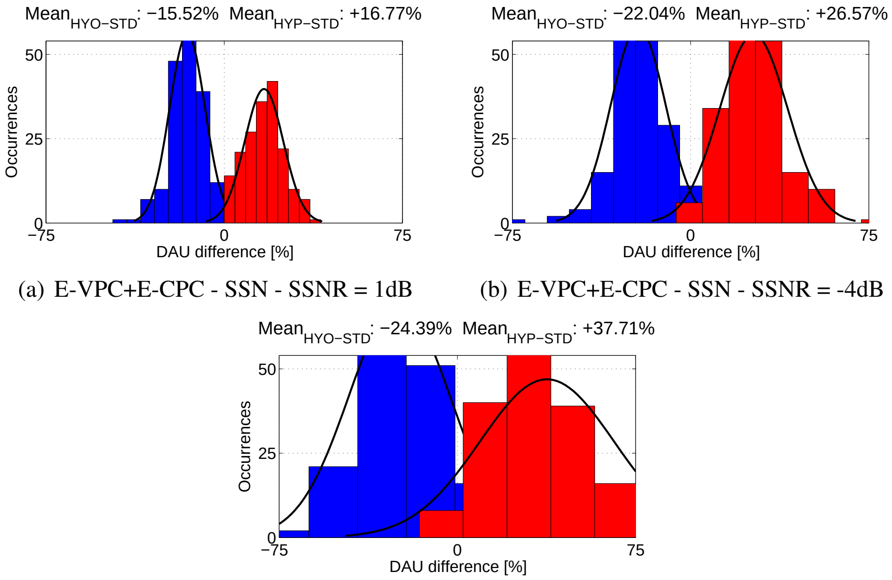


(a) E-VPC+E-CPC - CS - SSNR = -9dB

(b) E-VPC+E-CPC - CS - SSNR = -14dB

(c) E-VPC+E-CPC - CS - SSNR = -21dB

Figure C.8: *Distribution of the* ***SII*** *value differences (in percentage) between HYO and STD speech (blue histograms), and between HYP and STD speech (red histograms) controlled with E-VPC and E-CPC simultaneously. Speech signals are mixed with* ***CS noise*** *with high, mid, and low SSNR.*

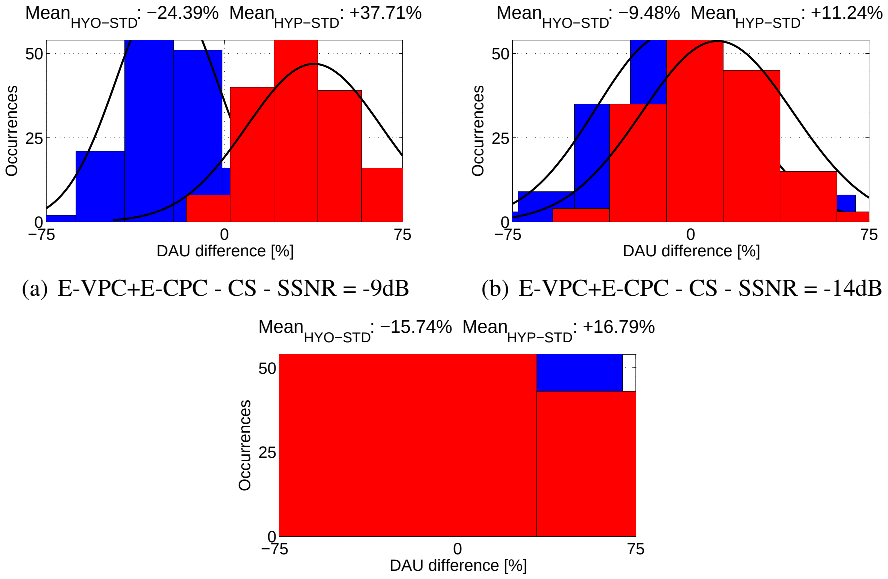


(a) E-VPC+E-CPC - SSN - SSNR = 1dB

(b) E-VPC+E-CPC - SSN - SSNR = -4dB

(c) E-VPC+E-CPC - SSN - SSNR = -9dB

Figure C.9: *Distribution of the* ***Dau*** *value differences (in percentage) between HYO and STD speech (blue histograms), and between HYP and STD speech (red histograms) controlled with E-VPC and E-CPC simultaneously. Speech signals are mixed with* ***SSN noise*** *with high, mid, and low SSNR.*



(a) E-VPC+E-CPC - CS - SSNR = -9dB

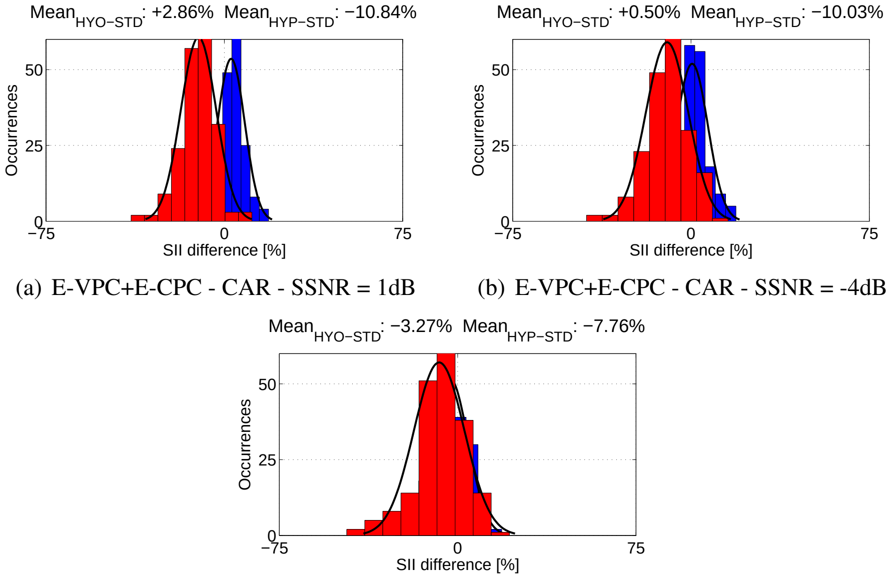
(b) E-VPC+E-CPC - CS - SSNR = -14dB

(c) E-VPC+E-CPC - CS - SSNR = -21dB

Figure C.10: *Distribution of the* ***Dau*** *value differences (in percentage) between HYO and STD speech (blue histograms), and between HYP and STD speech (red histograms) controlled with E-VPC and E-CPC simultaneously. Speech signals are mixed with* ***CS noise*** *with high, mid, and low SSNR.*

C.3 *Lucia*

The audio samples produced with the English male voice *Lucia* results are evaluated with SII in Figure C.11, Figure C.12, and Figure C.13, and with Dau in Figure C.14, Figure C.15 and Figure C.16.

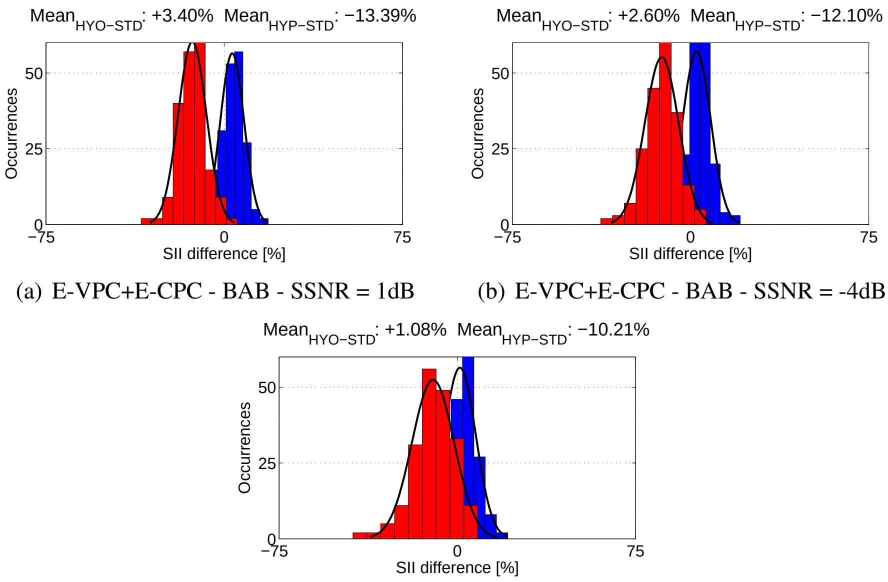


(a) E-VPC+E-CPC - CAR - SSNR = 1dB

(b) E-VPC+E-CPC - CAR - SSNR = -4dB

(c) E-VPC+E-CPC - CAR - SSNR = -9dB

Figure C.11: *Distribution of the **SII** value differences (in percentage) between HYO and STD speech (blue histograms), and between HYP and STD speech (red histograms) controlled with E-VPC and E-CPC simultaneously. Speech signals are mixed with **CAR noise** with high, mid, and low SSNR.*

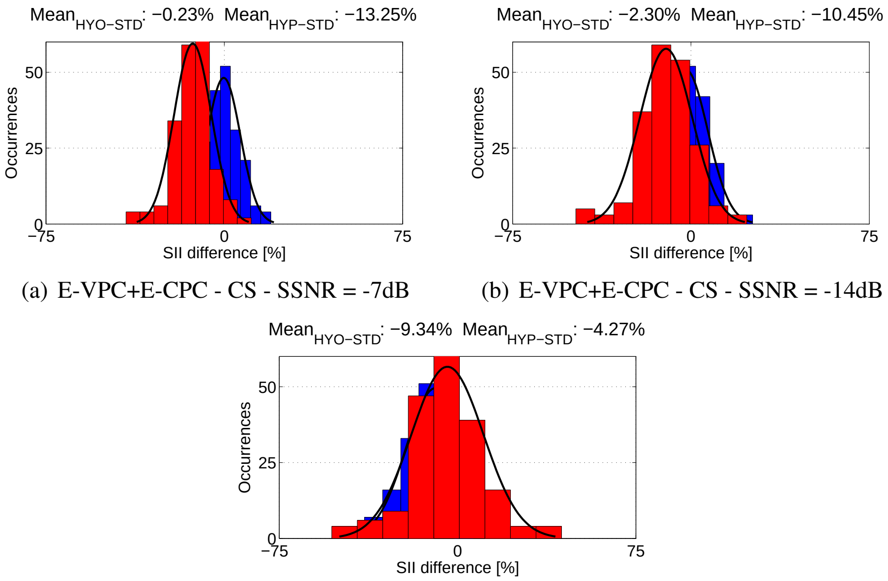


(a) E-VPC+E-CPC - BAB - SSNR = 1dB

(b) E-VPC+E-CPC - BAB - SSNR = -4dB

(c) E-VPC+E-CPC - BAB - SSNR = -9dB

Figure C.12: *Distribution of the **SII** value differences (in percentage) between HYO and STD speech (blue histograms), and between HYP and STD speech (red histograms) controlled with E-VPC and E-CPC simultaneously. Speech signals are mixed with **BAB noise** with high, mid, and low SSNR.*

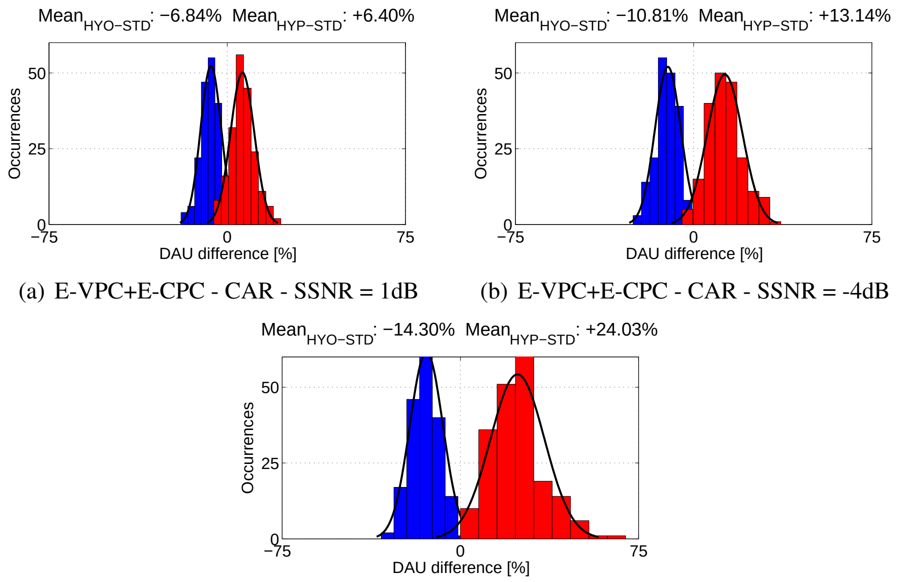


(a) E-VPC+E-CPC - CS - SSNR = -7dB

(b) E-VPC+E-CPC - CS - SSNR = -14dB

(c) E-VPC+E-CPC - CS - SSNR = -21dB

Figure C.13: *Distribution of the **SII** value differences (in percentage) between HYO and STD speech (blue histograms), and between HYP and STD speech (red histograms) controlled with E-VPC and E-CPC simultaneously. Speech signals are mixed with **ECS noise** with high, mid, and low SSNR.*

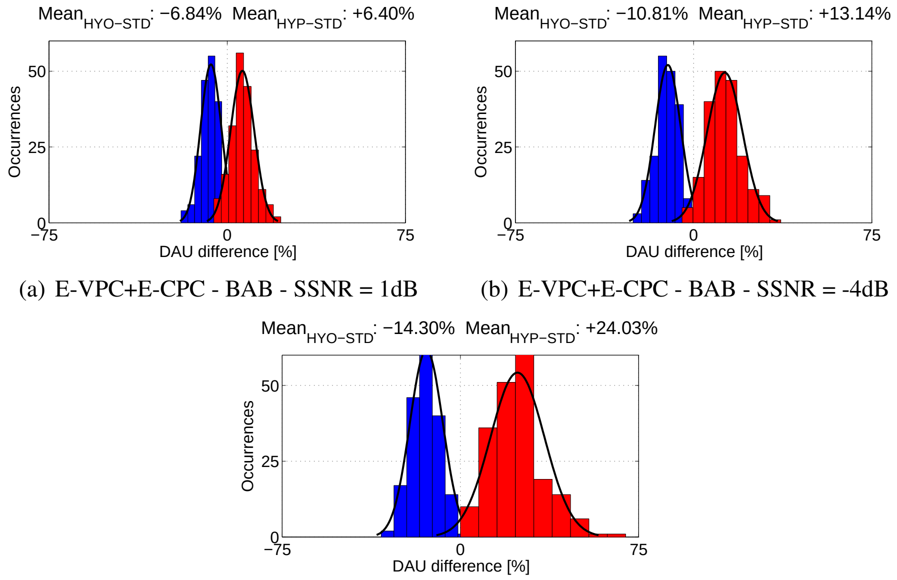


(a) E-VPC+E-CPC - CAR - SSNR = 1dB

(b) E-VPC+E-CPC - CAR - SSNR = -4dB

(c) E-VPC+E-CPC - CAR - SSNR = -9dB

Figure C.14: *Distribution of the **Dau** value differences (in percentage) between HYO and STD speech (blue histograms), and between HYP and STD speech (red histograms) controlled with E-VPC and E-CPC simultaneously. Speech signals are mixed with **CAR noise** with high, mid, and low SSNR.*

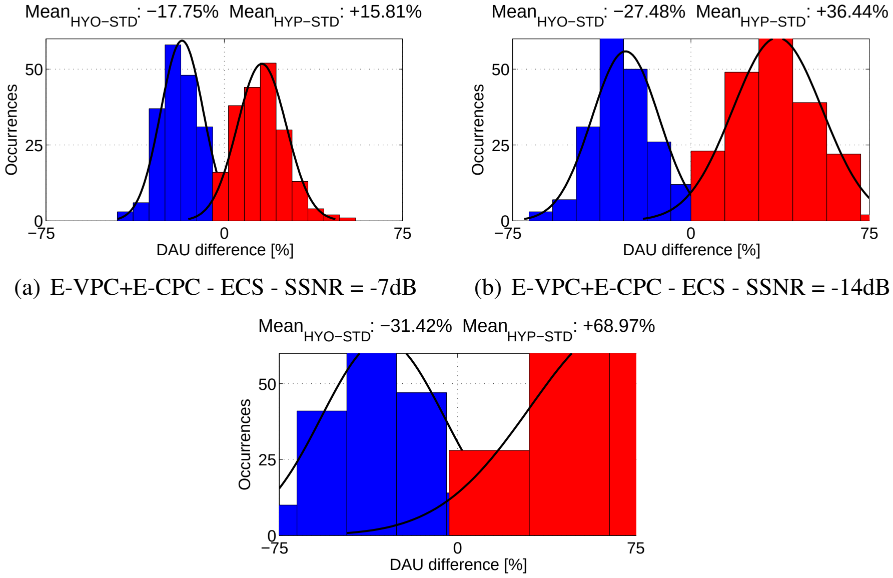


(a) E-VPC+E-CPC - BAB - SSNR = 1dB

(b) E-VPC+E-CPC - BAB - SSNR = -4dB

(c) E-VPC+E-CPC - BAB - SSNR = -9dB

Figure C.15: *Distribution of the **Dau** value differences (in percentage) between HYO and STD speech (blue histograms), and between HYP and STD speech (red histograms) controlled with E-VPC and E-CPC simultaneously. Speech signals are mixed with **BAB noise** with high, mid, and low SSNR.*



(a) E-VPC+E-CPC - ECS - SSNR = -7dB

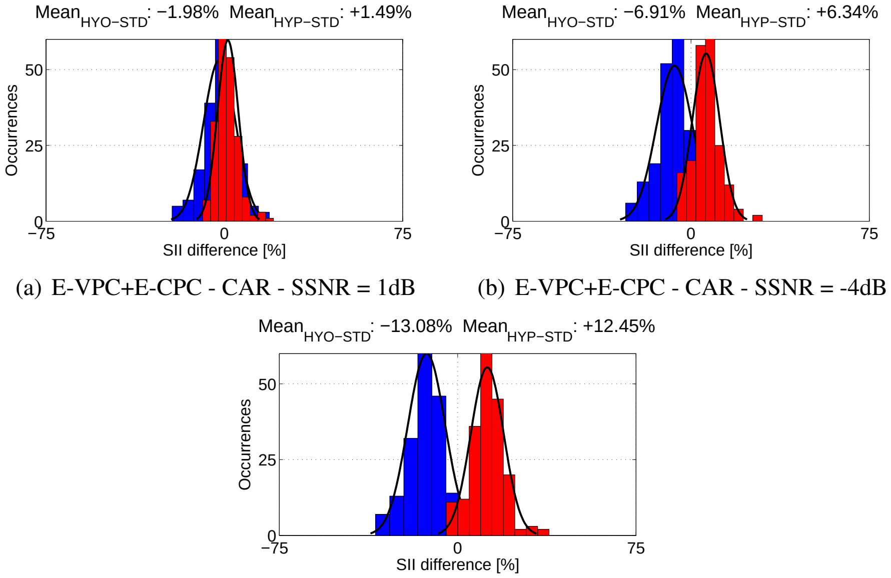
(b) E-VPC+E-CPC - ECS - SSNR = -14dB

(c) E-VPC+E-CPC - ECS - SSNR = -21dB

Figure C.16: *Distribution of the* ***Dau*** *value differences (in percentage) between HYO and STD speech (blue histograms), and between HYP and STD speech (red histograms) controlled with E-VPC and E-CPC simultaneously. Speech signals are mixed with* ***ECS noise*** *with high, mid, and low SSNR.*

C.4 *Roberto*

The audio samples produced with the English male voice *Roberto* results are evaluated with SII in Figure C.17, Figure C.18, and Figure C.19, and with Dau in Figure C.20, Figure C.21 and Figure C.22.

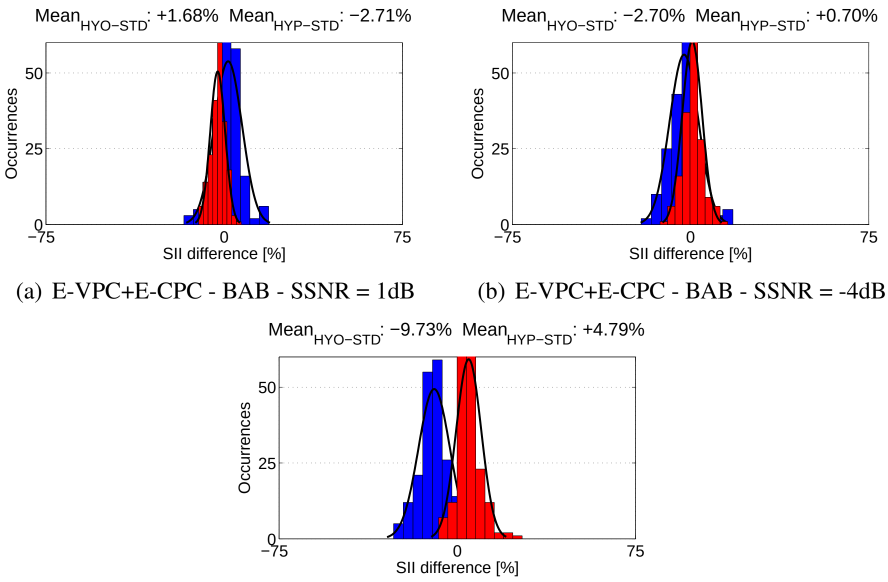


(a) E-VPC+E-CPC - CAR - SSNR = 1dB

(b) E-VPC+E-CPC - CAR - SSNR = -4dB

(c) E-VPC+E-CPC - CAR - SSNR = -9dB

Figure C.17: *Distribution of the **SII** value differences (in percentage) between HYO and STD speech (blue histograms), and between HYP and STD speech (red histograms) controlled with E-VPC and E-CPC simultaneously. Speech signals are mixed with **CAR noise** with high, mid, and low SSNR.*

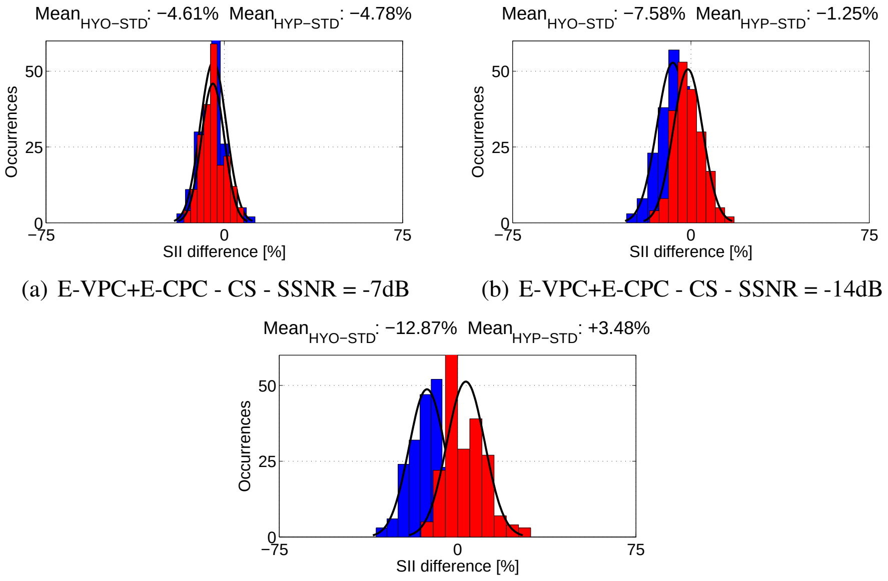


(a) E-VPC+E-CPC - BAB - SSNR = 1dB

(b) E-VPC+E-CPC - BAB - SSNR = -4dB

(c) E-VPC+E-CPC - BAB - SSNR = -9dB

Figure C.18: *Distribution of the* ***SII*** *value differences (in percentage) between HYO and STD speech (blue histograms), and between HYP and STD speech (red histograms) controlled with E-VPC and E-CPC simultaneously. Speech signals are mixed with* ***BAB noise*** *with high, mid, and low SSNR.*

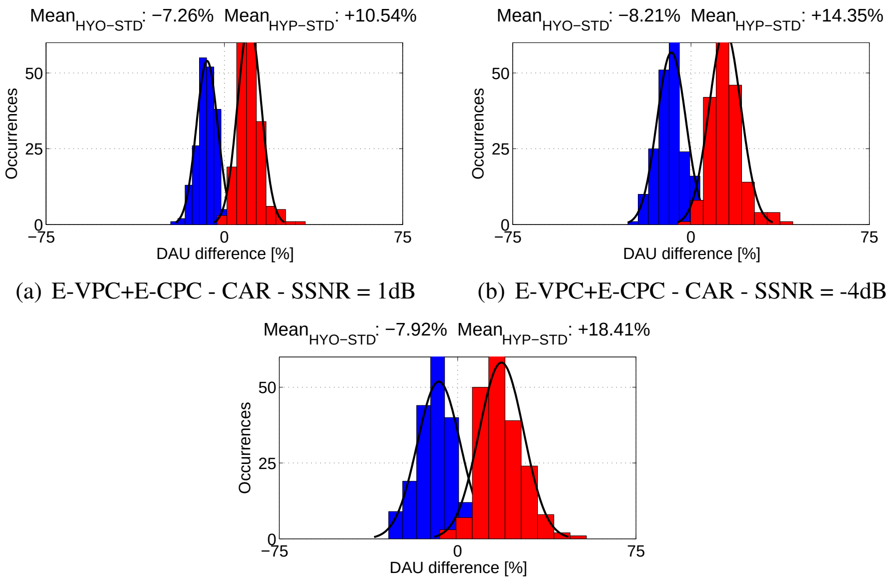


(a) E-VPC+E-CPC - CS - SSNR = -7dB

(b) E-VPC+E-CPC - CS - SSNR = -14dB

(c) E-VPC+E-CPC - CS - SSNR = -21dB

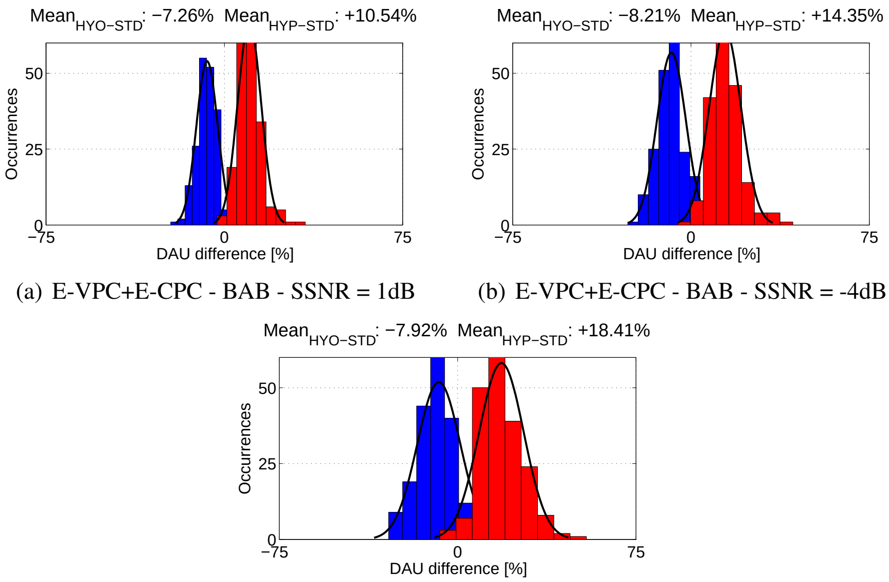
Figure C.19: *Distribution of the* ***SII*** *value differences (in percentage) between HYO and STD speech (blue histograms), and between HYP and STD speech (red histograms) controlled with E-VPC and E-CPC simultaneously. Speech signals are mixed with* ***ECS noise*** *with high, mid, and low SSNR.*



(a) E-VPC+E-CPC - CAR - SSNR = 1dB (b) E-VPC+E-CPC - CAR - SSNR = -4dB

(c) E-VPC+E-CPC - CAR - SSNR = -9dB

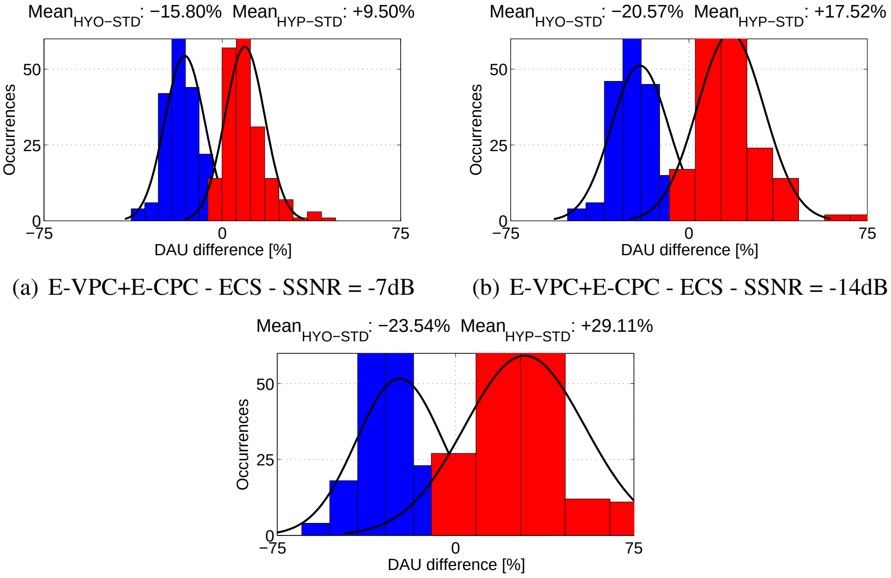
Figure C.20: *Distribution of the* ***Dau*** *value differences (in percentage) between HYO and STD speech (blue histograms), and between HYP and STD speech (red histograms) controlled with E-VPC and E-CPC simultaneously. Speech signals are mixed with* ***CAR noise*** *with high, mid, and low SSNR.*



(a) E-VPC+E-CPC - BAB - SSNR = 1dB (b) E-VPC+E-CPC - BAB - SSNR = -4dB

(c) E-VPC+E-CPC - BAB - SSNR = -9dB

Figure C.21: *Distribution of the* ***Dau*** *value differences (in percentage) between HYO and STD speech (blue histograms), and between HYP and STD speech (red histograms) controlled with E-VPC and E-CPC simultaneously. Speech signals are mixed with* ***BAB noise*** *with high, mid, and low SSNR.*



(a) E-VPC+E-CPC - ECS - SSNR = -7dB

(b) E-VPC+E-CPC - ECS - SSNR = -14dB

(c) E-VPC+E-CPC - ECS - SSNR = -21dB

Figure C.22: *Distribution of the* ***Dau*** *value differences (in percentage) between HYO and STD speech (blue histograms), and between HYP and STD speech (red histograms) controlled with E-VPC and E-CPC simultaneously. Speech signals are mixed with* ***ECS noise*** *with high, mid, and low SSNR.*