PERSONALISING SYNTHETIC VOICES FOR INDIVIDUALS WITH SEVERE SPEECH IMPAIRMENT

Sarah M. Creer

DOCTOR OF PHILOSOPHY
AT
DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF SHEFFIELD
SHEFFIELD, UK
AUGUST 2009

Table of Contents

List of Figures

Abstract

Speech technology can help individuals with speech disorders to interact more easily. Many individuals with severe speech impairment, due to conditions such as Parkinson's disease or motor neurone disease, use voice output communication aids (VOCAs), which have synthesised or pre-recorded voice output. This voice output effectively becomes the voice of the individual and should therefore represent the user accurately.

Currently available personalisation of speech synthesis techniques require a large amount of data input, which is difficult to produce for individuals with severe speech impairment. These techniques also do not provide a solution for those individuals whose voices have begun to show the effects of dysarthria.

The thesis shows that Hidden Markov Model (HMM)-based speech synthesis is a promising approach for 'voice banking' for individuals before their condition causes deterioration of the speech and once deterioration has begun. Data input requirements for building personalised voices with this technique using human listener judgement evaluation is investigated. It shows that 100 sentences is the minimum required to build a significantly different voice from an average voice model and show some resemblance to the target speaker. This amount depends on the speaker and the average model used.

A neural network analysis trained on extracted acoustic features revealed that spectral features had the most influence for predicting human listener judgements of similarity of synthesised speech to a target speaker. Accuracy of prediction significantly improves if other acoustic features are introduced and combined non-linearly.

These results were used to inform the reconstruction of personalised synthetic voices for speakers whose voices had begun to show the effects of their conditions. Using HMM-based synthesis, personalised synthetic voices were built using dysarthric speech showing similarity to target speakers without recreating the impairment in the synthesised speech output.

Acknowledgements

Many thanks must go to my supervisors, Phil Green and Stuart Cunningham, whose experience, knowledge and direction was always appreciated and gratefully received.

This research was sponsored by the Engineering and Physical Sciences Research Council.

This work would not have been possible without the contribution from Junichi Yamagishi whose willingness to share his expertise and ideas has been greatly appreciated. Thanks must also go to others at CSTR who made my visit there so useful and enjoyable.

Thanks to Margaret Freeman in the Department of Human Communication Sciences for the advice and assistance in finding participants. I am extremely grateful to all the participants in this study whose time and contributions have been essential and much appreciated.

Thanks to all those in the VIVOCA project and the CAST and RAT groups for their interest and involvement in this work and general support.

A thank you goes to SpandH, past and present for all the assistance, discussions, coffee-times and the willingness to tolerate the quizzes and sweepstakes. Thanks also to those in ICOSS floor 3 for the tea and the solidarity throughout all weather conditions.

It is good to have the opportunity to thank my friends and family for their great and continuing support. Finally thanks to Matt Gibson for the maths, the proof-reading but mostly his patience and kindness in all things.

Abbreviations

ALS Amyotrophic Lateral Sclerosis

ATR Advanced Telecommunications Research

CART Classification and Regression Tree

CASY Configurable Articulatory SYnthesiser

CP Cerebral Palsy

CVA Cerebrovascular Accident

DP Dynamic Programming

DRT Diagnostic Rhyme Test

EPG Electro-Palato-Graph

FDA Frenchay Dysarthia Assessment

FD-PSOLA Frequency Domain Pitch Synchronous Overlap and Add

F0 Fundamental Frequency

GMM Gaussian Mixture Model

HMM Hidden Markov Model

HSMM Hidden Semi-Markov Model

HTS H Triple S - HMM-based Speech Synthesis System

LPC Linear Predictive Coding

LP-PSOLA Linear Predictive Pitch Synchronous Overlap and Add

LSF Line Spectral Frequencies

MBE Multi-Band Excitation

MBROLA Multi-Band Resynthesis Overlap and Add

MCD Mel Cepstral Distortion

MDL Minimum Description Length

MFCC Mel Frequency Cepstral Coefficient

MND Motor Neurone Disease

MOS Mean Opinion Score

MRI Magnetic Resonance Imaging

MRT Modified Rhyme Test

MSD Multi-Space probability Distribution

MTVR ModelTalker Voice Recorder

PSOLA Pitch Synchronous Overlap and Add

RMSE Root Mean Squared Error

STRAIGHT Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum

SUS Semantically Unpredictable Sentences

ToBI Tone and Break Indices

VIVOCA Voice Input Voice Output Communication Aid

VOCA Voice Output Communication Aid

WER Word Error Rate

Chapter 1

Introduction

When individuals lose the ability to produce their own speech due to conditions such as Parkinson's disease (PD) or motor neurone disease (MND), they often look to other techniques to provide them with an alternative means to communicate. One such alternative is to use speech technology to provide a voice prosthesis or artificial replacement voice when that of the individual becomes unusable, particularly for interacting with listeners who are unfamiliar to them. *Voice Output Communication Aids* (VOCAs) can provide these individuals with an alternative method of communicating, taking an input and using a pre-recorded or synthesised voice to provide the output.

Currently available VOCAs provide a method of communication which attempts to recreate the natural oral communication that occurs between conversational partners. The use of VOCAs has been shown to provide a higher quality of life for individuals with speech impairment but this intervention can still lead to the abandonment of use of *augmentative and alternative communication* (AAC) devices [141]. An acceptability model for AAC devised in [122], describes a number of contributing factors to the acceptability of a communication aid, which includes aspects involving the conversational partner, the user and the technology used in the device. The aim of this thesis is to attempt to use technology to make communication aids more acceptable for the user. Specifically, this work argues that there should be a voice that more closely matches the vocal identity of the user themselves to use with a voice output communication aid. It goes on to investigate a potential technique for voice output personalisation to provide this choice for the user.

This chapter provides an introduction to the thesis with an overview of the target population for this work and discusses the implications of speech loss. Finally, an overview of the structure and content of the chapters is provided.

1.1 Population affected by speech loss and impairment

Speech impairment occurs as a result of injury to the brain, either acquired, through conditions such as motor neurone disease, or congenital, caused by conditions such as cerebral palsy (CP). These disorders are either progressive, such as Parkinson's disease or are sudden onset as a result of traumatic brain injury such as a stroke or cerebrovasular accident (CVA). Congenital dysarthria is usually stable in its presentation whereas acquired progressive disorders are usually preceded by having normal speech development and the diminishing neurological function leads to a progressive deterioration in the individual's ability to produce speech. Diminishing or reduced neurological function leads to a progressive deterioration or a sudden loss of motor control for an individual.

The severity of these conditions depends on the location and extent of the brain injury. The musculature involved in speech production is affected in the same way that motor control of other muscles are affected in these conditions. This means that the effect on speech is frequently coupled with physical disabilities. In progressive disorders, deterioration of speech is usually the first symptom to present [61, 86] and as motor control is lost, the severity of impairment increases and understanding the speech becomes more difficult. Disordered speech output resulting from conditions such as these is called *dysarthria*.

1.1.1 Dysarthria

Dysarthria is defined as:

> "a collective name for a group of neurologic speech disorders resulting from abnormalities in the strength, speed, range, steadiness, tone, or accuracy of movements required for control of the respiratory, phonatory, resonatory, articulatory and prosodic aspects of speech production. The responsible pathophysiologic disturbances are due to the central or peripheral nervous system abnormalities and most often reflect weakness; spasticity; incoordination; involuntary movements; or excessive, reduced, or variable muscle tone" [61].

It is a commonly acquired result of both progressive and sudden onset disorders, with statistics from 1995 reporting dysarthria affecting 170 people per 100000 in the UK [65].

Dysarthric speech differs in the severity of the symptoms and there is variability of the combinations of symptoms affecting the vocal apparatus depending on the type of dysarthria, which is related to the area of the brain where the injury is located. In general terms, these abnormalities in the speech mean that the vocal output has reduced intelligibility, affected voice quality and impaired prosody. In severe cases it can be completely unintelligible to naive listeners. The reduced ability to interact and communicate effectively leads to a lack of ability and ultimately motivation to interact socially [53, 65, 111, 148]. It also has implications for the self-identity of the individual [53], their relationships with others [53, 65, 111, 148] as well as implications for education and career prospects [65, 152].

1.2 Voice Output Communication Aids

When speech as a method of communication has been rendered unusable in some way, either progressively or as a result of a sudden onset disorder, an alternative must be used for this task that was previously natural and intuitive for the speaker. Maintenance of social interaction is vital for the avoidance of social withdrawal once the individual loses his or her own speech [128, 156, 168]. This help can come in the form of low-technology devices such as alphabet or picture boards to help communicate or high-technology solutions such as a VOCA. Using low-technology devices may be useful for communicating quickly and effectively with people who are well practised at that type of interaction. VOCAs offer the advantage that the output can be understood by most other people and interactions can take place with new conversational partners as well as enhance those occurring with the friends and family of the user. This can be useful in terms of providing more independence for the user in everyday situations [179]. One other advantage is that the use of the VOCA has been shown to increase the frequency of interactions that an individual has which allows the individual to build up social relationships and become more involved in the world around them [179]. This is in part due to the reduction of effort required to decode the message being communicated which can be very involved for the interactional partner and therefore restricted to those very familiar with the user. It is also partly due to the VOCA being able to more easily and explicitly gain the conversational partner's attention rather than using potentially ambiguous gestures that the individual would otherwise use to begin an interaction using a low-tech device.

A VOCA takes some kind of input and outputs a spoken message. The input is usually text which can be input manually using a keyboard or switches using a text or graphical display or it can be input using vocal control. The output is either a pre-recorded digitised voice, speech synthesis or a combination of these. *Speech synthesis* is the production of artificial speech which is used in situations where it is impractical or impossible to use actual speech. If a VOCA had a clearly delimited use or domain, with a highly restricted set of required utterances, it is possible and practical to record that set of utterances which could be played when required. For more general use as a replacement of speech function, where any novel utterance may be required, it is impossible for a speaker to record all possible utterances for communication and therefore speech synthesis is used in place of the user's or another's recorded voice.

How the communication aid is used is affected by the type of speech loss and the context in which an interaction is taking place. It can take the role of the individual's primary mode of communication, where the individual's condition has progressed to the point where their speech is unintelligible or entirely non-usable and the aid provides them with an alternative method of communication. It can also be used to augment some still functioning speech to clarify unintelligible utterances or when interacting with non-familiar listeners or in difficult listening conditions.

Factors which are important for the acceptability and therefore success of adoption of augmentative and alternative communication interventions have been indicated in [122]. This model groups factors into: milieu, person and technology. Expanding these component parts: *milieu* refers to external factors such as the conversation partner, including their attitudes towards the AAC device, funding available for the communication aid and the communication environment; *person* refers to factors relating directly to the user, including their condition, attitudes, personality, skills and personal needs; *technology* refers to features such as durability, ease of use, appearance, cost, possibility for customisation and voice output quality. There is some interaction here in that the technology influences the attitudes of both the communication aid user and the conversation partner.

Taking into account this model, problems with current VOCA use can be identified as contributing to potential acceptability issues. For example, the use of a VOCA over low-tech devices has advantages in that it can be used with unfamiliar interaction partners easily as the exchange attempts to replicate an interaction experience with which the conversation partners are familiar. However, allowances have to be made as the use of VOCAs cannot fully replicate the spoken language interaction that an unfamiliar conversational partner is expecting. There is a time delay between the input and output of the message to be conveyed, where the delay depends on the type of input employed and effort required by the VOCA user. The delay may also have implications for the content of the message and the type of interactional cues that are presented during interaction, for example turn taking [177].

Current VOCAs do not allow a full personalisation of the output speech, as an outlet for the conveyance of identity of the individual and of group membership. This personalisation of the output is also restricted to the content of the message, there is little or no provision of expression of personal characteristics or emotion through manipulation of the prosodic output. VOCAs lack an ability to use speech to convey features such as humour or sarcasm which the individual may be used to using and want to use in conversation with speaking partners. Expression of emotion or mood using prosodic information is also relied on for successful communication, showing understanding and resolving communication problems. These limitations specifically cause problems for social interaction, which can lead to social withdrawal and isolation for users of communication aids [128, 156, 168]. These limitations are discussed in more detail in chapter 2 which focusses on the impact of VOCA use on social interaction with a view to minimising social withdrawal and isolation.

1.3 Scope of the thesis

To improve the acceptability of VOCAs and maintain a high frequency of social interaction this thesis attempts to tackle one of the issues presented above: vocal personalisation, where personalisation is defined as producing a synthesised voice that sounds as similar as possible to the user before the individual was unable to use his or her own voice. It is acknowledged that vocal personalisation does not entirely solve the acceptability issue but offering a degree of personalisation to the vocal output contributes to addressing multiple factors as presented in the acceptability model such as customisation of the device, the attitudes of the conversation partner and user, service delivery and the output voice quality.

Preserving the identity of an individual through maintaining an individual's speaker characteristics in a VOCA overcomes social distance imposed by a device acting as an intermediary in a conversation, it preserves the voice as an identifier of the individual and it allows social bonds to form through associations with features realised in their voice along with the content. It has been suggested that the communication aid is an extension of the self where it represents the voice of that person:

> "If a voice communicates to the outside world everything that a person is, it should represent him/her accordingly" ([3] p139).

Having a personalised voice may increase motivation for the user to engage in social interaction and provides a means for individuals to have more control and choice over what represents them.

There are methods available for building personalised voices for people before they lose their speech. The ultimate personalisation is to record all required utterances before the voice deteriorates and store this on a communication aid. To produce any novel utterance, a new synthetic voice must be built. This requires building a database of recordings from which to build the synthetic voice, termed *voice banking*. There is currently no provision for personalising a voice for those individuals who have not banked speech recordings prior to their speech deteriorating. This is important for individuals whose emotional readiness inhibited their ability to address the potential loss of their own voice to make recordings of their voice pre-deterioration and also in cases where speech deterioration was sudden or not expected, such as in strokes. This personalisation using dysarthric speech data is the focus of the thesis.

The thesis does not directly address or measure the acceptability of personalised as opposed to non-personalised communication aids, it is hypothesised that personalised communication aids will have higher user and interaction partner acceptability than their non-personalised counterparts. In order to test this hypothesis, it is necessary to provide a method for the vocal personalisation of communication aids.

1.4 Thesis overview

The structure of the thesis is described below.

1.4.1 Chapter 2: VOCAs and social interaction

Chapter 2 firstly discusses in more detail the limitations of current VOCAs taking a social interaction perspective which can lead to a lack of acceptance of a device. The scope of the thesis is defined, with the focus presented as building personalised synthesised voices for individuals with dysarthria. To this end, the chapter presents the acoustic impairment caused by motor speech disorders in dysarthria. The chapter concludes that vocal personalisation can contribute to the acceptability of a VOCA following certain requirements for the voice output (it must be intelligible, natural-sounding, be similar to the user and be manipulable for prosody), technique (minimal data input, can use dysarthric data input, and there is a practical tool available) and user (having emotional readiness).

1.4.2 Chapter 3: Speech synthesis methods and evaluation

Chapter 3 provides a description of how speech synthesis is evaluated using both subjective measures using human listener judgements and objective measures based on acoustic feature comparisons. It provides information on the different techniques available for building voices for speech synthesis: articulatory synthesis, parametric synthesis, concatenative synthesis, model-based synthesis and voice conversion. It evaluates the appropriateness of these techniques for building personalised voices for speakers who have banked their voice pre-deterioration of their voice or once the deterioration has begun, based on the requirements detailed in chapter 2. The conclusion of this chapter is that model-based synthesis is the most appropriate technique for this task.

1.4.3 Chapter 4: HTS - HMM-based synthesis

Chapter 4 describes model-based synthesis in more detail, in particular the HTS toolkit ('H Triple S' - Hidden Markov Model-based Speech Synthesis System). HMM-based synthesis is a method which fulfils the requirements for building personalised synthetic voices as set out in chapter 2 using adaptation techniques originally used to deal with minimal data availability for training models for speech recognition. This technique allows the possibility of building personalised voices for speech data that has begun to deteriorate due to the individual's condition.

1.4.4 Chapter 5: Voice banking using HMM-based synthesis for data pre-deterioration

Chapter 5 describes experiments to build voices from non-disordered data, replicating the situation for voice building where data collection has been possible before deterioration of the voice has begun. It reports an evaluation of the voices using subjective measures of human listener responses and measures the amount of data needed to provide a voice resembling a target speaker. The comparison between these results and those published by other researchers in this area assesses whether the amount of speech data required applies to these speakers using the available set-up. It also investigates which acoustic features are used by the human listeners to make their judgements of similarity between the synthesised speech and the target speaker by training a neural network to replicate listener responses.

1.4.5 Chapter 6: Building voices using dysarthric data

Chapter 6 details experiments to build voices using speech-impaired data and describes case studies of building voices for three different individuals with dysarthric speech. It evaluates the different approach required for dealing with disordered data, as described in chapter 4.

1.4.6 Chapter 7: Conclusions and further work

Chapter 7 concludes the thesis, proposing further work in this area. The further work is directed towards developing a provision to build personalised synthetic voices for those individuals wanting to bank their voice pre- or post-speech deterioration.

Chapter 2

VOCAs and social interaction

2.1 Introduction

This chapter presents issues affecting the acceptability of AAC relating to the inhibition and facilitation of social interaction. Issues relating to personalisation of access, speed of interaction and personalisation and customisation of the output voice are considered. The scope of the thesis has been defined as building personalised synthesised voices for individuals with dysarthria. To this end, the chapter presents dysarthric speech in more detail, specifically the acoustic impairment caused by motor speech disorders. Finally a list of requirements for a VOCA is set out in terms of the specification of the voice output, the technique used and the emotional state of the person for such a personalised voice to be acceptable to a VOCA user.

2.2 Acceptability of AAC: social interaction perspective

Using VOCAs may provide opportunity for a higher quality of life for individuals with speech impairment over the use of low technology alternatives, but the use of VOCAs can still be abandoned even if they are functional and well-designed [141]. The acceptability model presented in chapter 1 presents an overall view of factors that should be taken into account when assessing the acceptability of an AAC device for a particular individual.

Addressing the problem of abandonment of technology which leads to the social isolation and withdrawal of potential AAC users [128, 156, 168], researchers have focussed on improving AAC specifically for social interaction [9, 83]. From a social interaction perspective, communication is not a static process of passing on information but it is a process that derives the message from a joint understanding between both interactional partners and the behaviour involved in transmitting that message [37]. A social interaction perspective therefore takes into account how AAC may inhibit or facilitate social interaction, how the perceptions of the device and the user contribute to the success of and motivation for interaction and how different communication environments influence the communication [9, 83].

As an example, it has been found that users of AAC can see a VOCA as a mode of communication that can discourage social closeness in some circumstances [156]. AAC users often employ multimodal strategies to communicate, using techniques to involve the conversational partner in trying to resolve a communication breakdown, such as gestures, vocalisations or eye contact. Using a VOCA, the conversational partner is less involved in the resolution, as the VOCA can take that role to some extent. The VOCA may be more suitable in interactions with unfamiliar listeners, as for family members or friends, this communication resolution partnership may be important to them in maintaining a close social bond. Therefore any AAC device which allows communication without inhibiting other interactional strategies and is relevant to that situation provides more successful social interaction [97].

2.3 Acceptability of VOCAs

Evidence suggests that positive attitudes toward non-speaking individuals are influenced by using voice output in their augmentative communication rather than using low-tech devices [75, 131]. These high-tech devices replicate more closely oral communication that interaction partners are more accustomed to using and provides a higher perception of the competence of the individual to communicate [75, 131]. This ease of communication in comparison to low-tech devices for both conversational participants also provides motivation for further interaction.

Crucially part of the acceptability model is the emotional state or *emotional readiness* [156] of the individual to accept the use of an AAC device along with its stigma of associations with impairment and illness [102] and the implications of accepting the future without complete vocal use. This is a psychological issue, which is highly specific to the user and can cause high risk of device abandonment if the individual is presented with a VOCA before they are ready to accept its use [156]. This highlights the need for VOCAs to be easy to use and adaptable to the user's needs at any point during a progressive deterioration. It is important that this emotional aspect does not eventually hinder the acceptability of a device which may otherwise have contributed to the quality of life in encouraging social interaction for that individual.

For VOCAs specifically, there are a number of communication issues which affect the successful acceptability of a VOCA in terms of the inhibition and facilitation of social interaction. A VOCA attempts to replace the voice of an individual but it does not sufficiently allow the replication of social interaction that communicating through the voice provides. Access to the speech output requires time to produce input. This violates interactional norms of timing causing barriers to communication [37, 43]. The output produced is also not personalised to the user therefore not fully representing him or her appropriately. This disassociates the output from the individual producing it, which is further enforced by the physical aspect of the device being a visible intermediary from which the sound is emitted, leading to the conversational partner frequently addressing the device rather than the person using it [3].

The following sections detail the approaches made to more closely match the expectations of conversation partners in terms of the speed of access for input and aspects of the synthesised speech voice quality output. Effects of these factors on the users' and listeners' attitudes are also outlined and hypotheses are made as to what further action is required to reduce inhibiting factors to social interaction.

2.3.1 Speed of access

In speaking person to speaking person conversations, speed of interaction between partners is a very precise and rule-governed process [177]. So much so that if messages are not conveyed with immediacy the expectations of the conversation partner are affected and some type of repair mechanism is expected [37, 43, 177]. When using communication aids, there is some accounting for the lack of adherence to these interactional rules in that the conversation partner can see the user providing input to the device. However, it has been shown that positive perceptions of communicative competence and positive attitudes towards the device by both the user and the conversation partner are related to the speed of the production of the interactional turn [203]. Having ease of access reduces the negative impact of passivity of contribution and therefore lack of control of the conversation that is sometimes found to contribute negatively to the conversation partner's perception of the user's communicative competence [34].

Improving social interaction must firstly take into account the personal requirements of that individual in terms of their particular physical, cognitive and linguistic capabilities. Once the physical and cognitive limitations have been established, appropriate techniques to allow quicker and more intuitive access to input can be discussed.

2.3.1.1 Physical aspects

Due to the underlying conditions causing speech impairment, many users of communication aids also have physical disabilities. Having a lack of motor control can make it difficult for individuals to use conventional input methods to a device such as a keyboard. Alternative input methods therefore must be used to fit the physical condition of the user more appropriately. For communication aid use this affects the positioning of the device, its visual display and the type of accessible input methods that are available to the individual [83, 97]. For users with physical disabilities, alternative access techniques can be used such as switches, which can be operated by a part of the user's body that retains motor control, or a head or mouth stick, which can operate a switch or keyboard. In combination with these input techniques, scanning and prediction techniques can make access easier, quicker and less fatiguing for the individual (see section 2.3.1.3 below).

Eye tracking is an emergent interface which could prove to be useful for users with physical disabilities. The technology, such as the MyTobii (Tobii) system is used to identify where an individual is looking on a screen which corresponds to a switch input for the communication aid. This technique may be affected by changes in lighting and by the fatiguing of the eye muscles and also inhibits the social interactional role played by eye contact between interactional partners [83].

Brain interface technology, currently at an experimental level, uses electrodes implanted into the brain or placed on the scalp to translate brain activity into control signals, for example [27]. The challenge for this type of interface, as with eye tracking, may also relate to the problems of simultaneously using this technology and fulfilling other social interaction strategies.

For those individuals with some speech function, an attempt has been made to provide a speech interface for VOCA users using automatic speech recognition (ASR). Mild to moderate dysarthric speakers have had some success with off-the-shelf speech recognition systems, although not achieving as low an error rate as speakers with no speech impairment [69, 199]. This is due to the increased variability found in dysarthric speech and the large difference between what the system is expecting and the speech produced by this population. An attempt has been made to extend the use of speech recognition to more severe dysarthric speakers, which found that the key to providing successful recognition is a smaller possible input set and the consistency of the speech produced [171]. The requirement of consistency of speech production for successful recognition provides a challenge for the use of ASR for dysarthric speakers with progressively deteriorating speech (see section 1.1).

Implementing ASR in a communication aid, the VIVOCA (Voice Input Voice Output Communication Aid) [79] project takes all degrees of severity of (currently, non-progressive) dysarthric speech as input and translates these consistently produced utterances into an intelligible synthesised speech output sequence. The VIVOCA also allows auditory feedback as the input is produced, which removes the need to visually focus on the device, facilitating eye contact with the conversation partner and reduces the impact of the device as an intermediary to the conversation.

2.3.1.2 Cognitive and linguistic capabilities

Techniques to aid speed and access of information have to be personalised to match the cognitive and linguistic capabilities and impairment of the individual. Using a communication aid in a social interaction setting involves multitasking in terms of having an awareness of the setting, listening and attending to the conversation partner and also accessing the correct input information to form and convey the appropriate message. Learning and remembering this process can be difficult for individuals with additional cognitive impairment. Using letter- or word-based input can be difficult for individuals with linguistic impairment who may struggle with spelling or syntactic structure. It has, for example, been shown that more successful communication outcomes have been reached using AAC interfaces with fewer cognitive and linguistic requirements for children learning to use communication aids [129].

2.3.1.3 Utterance access techniques

In addition to using appropriate physical, cognitive and linguistic input techniques for the individual, attempts have been made to increase the speed of access to the VOCA output utterance using phonotactic, lexical, semantic and syntactic constraints and information to predict following letters, words and phrases in the input or to fill in gaps in the output [212].

Input to a VOCA can be represented either as graphical symbols, which may assist in the speed of access for those individuals with cognitive or linguistic difficulties, or orthographic representations. The size and nature of the representations affects the speed of input. For example, input can be letter-based, which is a slow process. The speed of input can be increased using prediction techniques although this can increase the cognitive load for the user. It relies on the user having good linguistic ability to produce the words. As an advantage it can be very flexible and allows any input to be produced. A word-based interface is slightly faster and also allows prediction to increase the speed of input. This technique again can be reliant on both cognitive and linguistic ability. As the unit of input increases in size, the time taken from input to output decreases. Using phrase or sentence based units of input can allow an individual to produce an utterance using one button or switch press. The output is linguistically well-formed and prepared in advance but it may be problematic to alter or personalise it for a particular interaction. A limitation of these types of systems is the amount of utterances that can be stored and easily accessed.

Utterance-based systems can increase the speed of the output when combined with prediction knowledge based on semantic relations and the utterance's place in a structure of interaction (e.g. SchemaTalk [212] or Frametalker [82]). These systems work due to the predictability of many day to day conversation structures. VOCA access can therefore make use of these patterns as prior information to inform the likely path of a future conversation, providing quicker access to certain context-appropriate phrases or phrase structures. Contextually relevant information taken directly from the conversation partner has also been used to facilitate faster input speeds for VOCAs [221]. This system takes recognised noun phrases from the speech of the conversation partner using ASR and places them as directly accessible stored units in the VOCA interface.

Using pre-stored phrases to increase speed of message delivery, however, does not account for all the issues relevant to social interaction. Until the pragmatics of an interaction are accounted for, the interaction as a sequence of stored phrases will not be able to facilitate natural-sounding conversation. Pragmatics refers to meaning and language use within the context that it occurs [125]. The TALK system [204] and its extension Contact [138], take into account the processes involved in social interaction such as the variability of content and the flow of conversation which, if fulfilling expectations and social goals, should lead to an enjoyment of social exchange and the encouragement of positive attitudes towards the conversation partner's communicative competence [205]. Taking this point of view, responses do not need to be ideal or exact but should be appropriate and provided without delay. These systems provide rapid response buttons which within one switch press can provide appropriate backchannelling (utterances such as "mm-hmm" or "oh?", which encourage the continued flow of interaction) or general sentence responses, for example randomly selected expressions of sympathy, aphorisms or hedges, such as "that's a good question" [205].

In situations where high relevance of message content is important to convey wants and needs, such as interactions in shops or other service provision, there are trade offs between the speed of message delivery and message relevance [9, 203]. To maintain high relevance message content and keep to expectations of timing in conversation, using rapid response utterances in combination with more relevant message formulation was found to be a useful strategy to provide an overall more positive interaction experience for unfamiliar conversation partners [9]. A successful model of input for a VOCA should therefore be adaptable to the user's needs and particular conversation environment and context at that time, taking into account the expectations and requirements of both conversational partners.

2.3.2 Aspects related to the output voice

One of the features in the acceptability model as described in chapter 1 was the quality of the output voice. This is particularly relevant to the VOCA where attempts are made to provide a realistic replication of speaking person to speaking person communication. To sufficiently augment or replace a voice in these circumstances, an augmentative or alternative communication aid should be able to perform the same functions as the individual's own speech used to. These functions have been defined by Light [128], who states four different social purposes for interaction: a. to obtain wants and needs, b. for information transfer, c. for social closeness and d. for social conventions of politeness. Locke's view [135] encompasses that, broadening the categories of her social purposes of communication into: the transmission of impersonal facts (Light's a and b) and the construction, maintenance and enjoyment of social relationships (Light's c and d). For speech communication, an additional purpose is the portrayal of identity through the features contained in the voice. The following sections detail the acceptability of VOCAs in terms of these functions of speech and encouragement of social interaction, including the effect that the output has on the attitudes of the user and conversational partner.

2.3.2.1 Intelligibility

For successful social interaction using a communication aid and for acceptability of the device, the technology has to work well. Specifically for a VOCA, it must provide an output which is fully intelligible. *Intelligibility* is defined as the accuracy with which an acoustic signal is conveyed by a speaker and recovered by a listener [112]. *Comprehensibility* is an extension of intelligibility taking into account the understanding of the message being recovered [119].

Taking natural speech as a benchmark at which to aim in terms of intelligibility, natural speech is comprehended as accurately but more quickly than high quality synthesised speech for sentences and this difference becomes more marked in difficult listening conditions such as in reverberant noise or when attention is divided [119, 188]. No significant differences are reported in the comprehensibility of high quality synthesised speech and natural speech for discourse and more complex communication tasks. Comprehensibility of the different types of speech is dependent on the complexity of the situation and task in addition to the listening conditions. The increase in contextual information available during discourse results in an increase of comprehension [51, 59, 119].

There is evidence to suggest that listening to and understanding high quality synthesis uses a greater cognitive load than understanding natural speech, but this load decreases with exposure to that particular voice [119, 213]. These results hold for the DECtalk^{TM}(Fonix) speech synthesiser and suggest that speech synthesisers can be used to successfully communicate information to listeners if high intelligibility synthesisers are used.

In terms of preferences, people prefer listening to and assign more positive attitudes towards natural speech over synthetic speech. This is true unless the speech is disordered, when the preference shifts towards listening to synthesised speech due to the increased level of intelligibility [58]. There is evidence to suggest that there is a correlation between the preferences of listeners for a synthesis system and the intelligibility of that system's speech output [74, 136, 150].

High intelligibility for a synthesised voice is therefore required for acceptability of the use of a VOCA in terms of it providing the service for which is it designed and also in terms of positive attitudes towards the voice and social interaction with the user.

2.3.2.2 Naturalness

Providing an acceptable VOCA in terms of technological factors, relates to the quality of the voice. One criticism that is levelled at speech synthesis is that the output is still not as natural-sounding as natural speech itself [106]. This can cause barriers to natural interactions and decrease the acceptability of a VOCA. There is evidence to suggest that there is a correlation between the preferences of listeners for a synthesis system and its perceived naturalness [174, 188].

People are accustomed to using speech to interact and adapt their speech to that of others with whom they are communicating. Speech accommodation theory [190] states that when talking to another person it is natural to accommodate your own speech to the situation and with whom you are talking. The basis for this theory is that the adjustment of speech style is motivated by the expression of values, attitudes and intentions and that it is the individual's perception of the conversational partner's speech that determines their behavioural response. There will be a convergence to another's style of speech if they wish to show solidarity with them or they desire their social approval. The opposite is true if the individual shows a divergence away from the conversational partner or strictly maintains their own speech patterns. There is evidence that this type of behaviour extends to human-computer interaction, which has implications for naturalness of interaction with computerised speech in a communication aid.

Using a low naturalness voice, the robotic style of the output is a mismatch to the human that it represents. A person's interaction with machines is very different from their interaction with people and if the perception is that they are addressing the communication aid rather than the user, their style of interaction forms a further obstacle to the user's ability to make a social connection with their conversational partner. For example, in a study conducted of spoken enquiries to a telephone line giving travel directions [151], the agent on the telephone used either their natural voice or their natural voice vocoded, which altered the voice of the speaker to sound robotic but still intelligible. This attempted to fool the callers into thinking that they were interacting with a machine. This revealed that when talking to what they thought was a machine, the callers used much more concise language with a much smaller number of words and interactional turns than when the caller was talking to what they knew was a human.

A further study on interactions with a Wizard-of-Oz computer animated dialogue system, again fooling the user into thinking they were interacting with a machine, also found that people adapted their own speaking rate to that of the manipulated voice output of the system [10]. When interacting with computer animated characters with synthesised speech voices as part of an educational software tool, children were also found to adapt the amplitude and duration of their speech to that produced by the software [44, 50].

When interacting with a computer, the purpose is to find out or provide information and therefore the language used is more compact to convey that message so that the machine understands and extracts what is relevant. The prosody is also altered for the same intention. Interaction with a human is very different in that the communication is at a more abstract level. Ideas are being expressed and those ideas are able to be understood by the other human being [46]. An unnatural-sounding voice could then be providing an obstacle to having a usual human-to-human conversational interaction rather than a human-to-computer interaction.

Following this evidence, having a natural-sounding voice output associates the device more closely with the individual and promotes more natural social interaction, which is likely to increase acceptability of the VOCA. Having a high quality, more natural voice reduces listener fatigue and the cognitive load placed on the listener to understand the speech that is being presented to them. This also contributes to the positive attitudes towards the VOCA by both users and conversational partners.

2.3.2.3 Prosodic control

The intelligibility and naturalness of speech synthesis in VOCAs improves with the state of the art. The prosodic element of the synthetic speech contributes to both of these factors, in providing more natural speech which in turn contributes to the intelligibility of the output. There is, however, very little provision for manipulation of prosodic output which can be useful for constructing pragmatically appropriate phrases for social interaction purposes (e.g. the difference between "oh.", "oh?" and "oh!" [83]). The reduced ability to control or influence the prosody of the output also has implications for the expression of features such as humour or sarcasm along with different emotions or style of speech.

Attempts have been made to improve the output of VOCAs in terms of emotion, for example, the ModelTalker interface [28] provides a choice of emotions in which to produce the synthesised speech of a particular utterance, for example cheer and gloom.

If the acceptability of a VOCA is related to the ability of the device to be suitable for aspects of social interaction having natural-sounding, unambiguous expression for all interactions, then having more control available for the output prosody is required.

2.3.2.4 Representation of the user

From the point of view of a speaker, embarrassment and negative attitudes towards their own speech creates barriers to socialisation through taking part in interactions [148]. This can be extended to potentially having negative attitudes towards the voice in a VOCA that they are using to communicate with others. This offers the potential of being able to customise the voice to one that the speaker identifies with and is motivated to use.

The voice is an identifier of the person to whom it belongs and provides clues about the gender, age, size, ethnicity and geographical identity of that individual [36, 219]. In using a communication aid personal information as portrayed through the voice is immediately lost. That particular voice is an individual's identifier to family members, friends and acquaintances and, once interaction has begun, to new communication partners. For VOCA users, there is currently a restricted choice of the voices that are available to them to distinguish themselves from others and to represent themselves. For example the Lightwriter's^{TM}(Toby Churchill) English voices offer a choice of American English or British English of different sexes and includes a child's voice. The lack of choice of different voices for existing communication aids can cause practical problems in large groups of people who may all be using the same device. For example, in a classroom setting, it would be difficult to identify the person making a comment or asking a question if there are a number of students using the same voice in their synthesiser.

An individual has to like and identify with a voice on a VOCA for them to feel motivated to use it. If a voice does not contain appropriate information about an individual's identity, it restricts the individual's ability to form associations with others through their vocal features. It may also lead to disassociations when using an inappropriate voice, which has its own identity features which may not match those of the user and lead to a lack of motivation for the speaker to interact. This is detrimental to the individual where group membership is particularly important, for example, for cultural associations [6, 81, 170] and within age group, specifically adolescents [183], when group membership is key to well-being.

When asked which voice they would prefer if they had to use a VOCA [45] participants matched the most natural-sounding and gender-appropriate voice to themselves. This correlates with results from studies of ideas of how assistive technology should be designed, suggesting that individuals would prefer to have a voice on a communication aid that matched the characteristics of the person who was using it [130] and that any communication aid should be highly customisable for the wants and needs of users [3]. The evaluations in [45] and [130] consisted of participants who were themselves not speech-impaired. The results are therefore indicative of listener expectations and a theoretical idea of what might be acceptable to them if they were to use a communication aid.

Using a personalised VOCA where the output is that speaker's own voice pre-deterioration or an approximation to it, represents that individual's gender, size, age, geographical, social, ethnic and cultural identity as they were before they lost their voice. The individual identifies with all the features in the voice and it represented them accurately before their voice deteriorated.

The lack of available resources to personalise communication aids makes it difficult to state whether or not using a personalised voice is appropriate and positive for users of communication aids based on empirical evidence. There has been little provision for vocal personalisation of VOCAs until recently. Personalisation of VOCAs can occur to some extent by the recording of pre-stored phrases by a person pre-deterioration of their voice or by an individual whose voice could represent them. This opportunity is restricted to those particular utterances by the size of the communication aid memory and by the intelligibility of the individual's voice once they have acquired a VOCA.

For a VOCA to produce any novel utterance, not just those pre-stored by the individual, a synthesised voice must be built. The ModelTalker project [28] has provided a means to bank a voice and create a personalised synthetic voice, designed specifically for people with speech loss related conditions. It is designed to require a minimal amount of data input as individuals with these conditions find it difficult and fatiguing to produce large volumes of speech. However, this technique relies on the speaker recording his or her voice before deterioration has begun. There is currently no provision for building a synthesised voice based on dysarthric speech data.

Previously, Murray and Arnott [157] attempted to provide rapid personalisation of a voice for the DECtalk$^{\text{TM}}$(Fonix) synthesiser using two levels of editing: interpolation of the existing voices and other more detailed changes to the individual parameters. The EDVOX system permitted such interpolation to introduce a level of individualism into the voice but did not allow more detailed personalisation to reconstruct the voice qualities of the individual. Aimed at children and teenagers, the Tango$^{\text{TM}}$(BlinkTwice) communication aid currently provides opportunities to create a more appropriate voice for the user using voice morphing from recorded adult voices to that of a child.

There is evidence that people attach human-like attributes and associations to synthetic speech just as they do to natural speech and this can affect their attitude towards the person as well as the message being conveyed [39, 185]. Participants in the experiment in [185] perceived the synthetic voice 'speaker' as less truthful, knowledgeable and involved in the communication in comparison to the natural speech, preferring the natural speech output. The higher quality synthesis rating was closer to that of the natural speech for these factors than the lower quality synthetic output. This demonstrates that the quality of synthetic speech is correlated to the negativity of a listener's attitudes towards the individual using it. This negative attitude effect does seem to disappear and the listener is more tolerant of the synthesised speech, if the speaker is known to be speech-impaired and therefore has no other choice but to use a synthesised voice [184, 186, 187].

Further influences on the attitude of the conversation partner are the match between the synthetic voice and the user. Listener preferences of synthetic speech, matching the voices to potential users of communication aids, have revealed that there is a preference for gender-appropriate and age-appropriate voices [45] in addition to intelligence- and socially-appropriate voices [168].

Attitudes from a listener's point of view are therefore non-negative towards the individual who is using a communication aid via the perception of the voice. However, this is only the case if the speech is being understood by the listener and interaction is able to take place. These results only hold if the voice is of high quality, easily comprehensible and natural-sounding. Having a more appropriate voice which matches the characteristics of an individual is also likely to promote positive reactions to the speaker by conversational partners and increase the motivation for using the voice for the user.

Providing a means with which to capture the voice of an individual offers a level of customisation and personalisation to the output voice which provides some retention of identity, individualism and control over the output that the user may require.

2.4 Dysarthric speech

To investigate what issues need to be addressed in using dysarthric speech from which to reconstruct a synthetic version of an individual's voice, it is important to more thoroughly investigate the acoustic properties and impairments associated with dysarthric speech.

2.4.1 Types of dysarthria

The type of dysarthria classification is usually dependent on the location of the damage to the brain. These types were formalised by [49] as: spastic, flaccid, ataxic, hypokinetic, hyperkinetic and mixed.

2.5.1.1 Spastic dysarthria

Spastic dysarthria is associated with bilateral damage to the upper motor neuron area, such as is associated with amyotrophic lateral sclerosis (ALS). This type of dysarthria usually affects all components of speech production. Spasticity (or increased muscle tone) is usually found in the articulatory muscles causing a strained voice quality, hypernasality, a slow rate of speech, with slow imprecise articulation and invariant loudness and pitch output [61, 200].

2.5.1.2 Flaccid dysarthria

Flaccid dysarthria is commonly found as a result of stroke or cerebrovascular accident (CVA), caused by damage to the lower motor neurons. Common characteristics of flaccid dysarthria are breathy voice, hypernasality and imprecise consonant production, although other distortions occur which relate to the specific nerve damage affected by the condition [61, 155]

2.5.1.3 Ataxic dysarthria

Ataxic dysarthria is associated with damage to the cerebellum or its surrounding circuits. Its main characteristic is that of *ataxia*, or incoordination, of the articulators causing irregular articulatory problems, lack of control of prosodic features including loudness, duration and pitch [61, 154].

2.5.1.4 Hypokinetic dysarthria

Hypokinetic dysarthria is caused by damage caused to the substantia nigra component of the extrapyramidal tract in the brain. This type of dysarthria is usually associated with Parkinson's disease but it can be caused by other conditions. The main features associated with this type of dysarthria are related to rigidity of the articulators, having characteristics such as reduced and invariant loudness, monopitch output, breathy and strained voice with imprecise and perceived accelerated articulation [61, 198].

2.5.1.5 Hyperkinetic dysarthria

Hyperkinetic dysarthria is caused by damage to the basal ganglia component of the extrapyramidal tract. Features of this type of dysarthria are hoarse or strained voice, inappropriate interruptions of speech and voicing, hypernasality and a slowed speech rate [61, 197].

2.5.1.6 Mixed dysarthria

Mixed dysarthria combines two or more of the above types of dysarthria, most frequently the combinations occur where the areas of damage are in close proximity, e.g. flaccid-spastic dysarthria, when the damage is in the upper and lower neurons [61, 196].

2.5.1.7 Related contributions to disordered output

Dysarthria frequently occurs with other related disorders that can contribute to the speech output distortions found with these conditions. For example, *apraxia of speech* is a neurological impairment of both language and speech where the messages from the brain to the muscles for the production of speech are not properly received. Other disorders such as *dysphasia*, an impairment in communication where an individual has difficulty accessing words that they are trying to produce, leads to the output having more hesitations, pauses and requires increased effort in speaking [61].

The level of intelligibility of dysarthric speech varies not only due to the severity and type of disorder but also due to factors such as fatigue, health or context: the environment in which the interaction occurs and the conversation partner. Inter and intra-speaker variability makes dysarthric speech more difficult to understand, particularly when listeners are unfamiliar with disordered speech and also with the speaker.

2.4.2 Acoustic description of dysarthria

Dysarthric speech is by definition speech which contains abnormalities in all aspects of speech production: respiratory, phonatory, resonatory, articulatory and prosodic. The following sections detail the effects of the disorders on these areas of speech production, see [61, 66, 218, 238] for reference. A voice building technique must be able to deal with these effects on the acoustics to be able to reconstruct the voice of an individual to an acceptable level without recreating the impairment in the synthesis.

2.5.2.1 Respiratory problems

Respiratory abnormalities associated with dysarthric speech caused by loss of control of the musculature of the lungs are shallow inhalation and lack of control over exhalation [61]. This contributes to uncontrolled rushes of air during speech production causing the speaker to become quickly out of breath and leads to difficulties in controlling the intensity of the speech. There can be an overall effect of intensity decay or there can be marked variation where the control is lost and the speaker produces loud speech initially with a burst of air and the intensity decreases quickly during the utterance production. This lack of overall energy will be reproduced and where there is higher variability in the energy and intensity output, the variation will manifest itself into having an output with highly variable energy. These types of abnormalities are associated with ataxic and extrapyramidal hypokinetic dysarthria, the latter being typical of Parkinson's disease.

2.5.2.2 Phonatory problems

Phonatory problems associated with dysarthria are caused by lack of control or increased rigidity of the vocal folds, commonly found in conditions such as Parkinson's disease. This causes difficulties in setting the vocal folds into vibration to produce voiced sounds, leading to a period of unwanted vocal noise while building up the required amount of sub-laryngeal pressure to set the vocal folds in motion. Some of the non-speech sounds present in dysarthric output are produced by the vocal apparatus and have speech-like characteristics.

Phonatory substitution errors also occur in such conditions as apraxia of speech. This is a condition which affects an individual's ability to communicate through language as well as affecting the control of their articulators. One of the characteristics of this type of speech is the confusion of voiced and voiceless sounds, for example pronouncing the word 'dull' as 'tull', substituting the voiced alveolar plosive /d/ for its voiceless counterpart /t/ [238]. Abnormal voicing errors are also frequently found in dysarthric speech, usually caused by inappropriate timing and overlap of the articulator movements.

Other phonatory problems manifested in dysarthric speech produce a change in voice quality. Breathy or hoarse speech is common in flaccid and extrapyramidal hypokinetic dysarthria, where the vocal folds are weakened and complete adduction is not achieved so excess breath escaping through the glottis produces unwanted turbulent noise in the signal at high frequencies. A more strained or harsh quality of voice can occur in hypokinetic, spastic and ataxic dysarthria where the air is forced through a more tightly constricted larynx [61]. The respiratory problems as described above can also cause inadequate subglottal air pressure which leads to a strained voice quality where the individual attempts to produce longer utterances than is possible with their respiratory capabilities and produce speech using the residual air in their lungs.

2.5.2.3 Resonatory problems

Conditions associated with dysarthria usually affect the resonatory system. One particular resonatory problem, particularly associated with flaccid and spastic dysarthria, is the inability to raise the velum sufficiently to make a complete seal in articulating a non-nasal sound. This hypernasality not only makes the speech more nasal-sounding but the inability to form a complete seal means that air is wasted as it escapes through the nasal cavity, leading to the production of shorter phrases before inhalation is required. This decreased intra-oral pressure also means that stop consonants that require pressure build up behind the closures for release are not well-formed and often contain aperiodic components in the closure portion. This effect is called *spirantisation.*

2.5.2.4 Articulatory problems

Dysarthric speech has a high incidence of articulation control and timing problems, meaning that a target segment articulation is not always reached. This is particularly noticeable in the production of complex articulations or sequences, for example consonant clusters. Vowels are also affected, particularly in ataxic dysarthric speech. Conditions which cause dysarthria where nerve damage has occurred, for example in stroke patients, can result in certain sounds consistently being difficult to produce.

Apraxia of speech also shows instances of anticipatory errors in articulation, where a sound is articulated before its appearance in a word or phrase being produced [238].

2.5.2.5 Prosodic problems

The prosody of all speakers with dysarthria is affected by their condition. The above effects all have some interaction with the prosodic output of speech. Imprecision in articulation or timing control can alter the perceived rate of speech. Parkinson's speech, which is typically associated with extrapyramidal hypokinetic dysarthria, can have a perceived increased rate of speech, whereas spastic dysarthria, can result in a slowed rate of speech. This rate is usually highly variable across segments. The variability of the amount of control, speed and range of movements across articulators affects the rate, intonation and rhythm of the output synthesis.

The speaker's *F0* or *fundamental frequency*, the acoustic correlate of pitch, is also affected by motor speech disorders. Monopitch or much reduced variability of pitch in the output is commonly associated with spastic and hypokinetic dysarthria. This is particularly the case when the speech is affected enough for it to be telegraphic, where the speech is produced as a string of single words, often deleting some function words. These prosodic effects are realised in the synthesised speech output and produce over-smoothed pitch traces along with inappropriate rates of speech. These unexpected deviations from what is expected to be heard lead to problems relating to the intelligibility and perceived naturalness of the output.

The problems associated with phonatory and respiratory abilities of individuals with dysarthria interact with the prosodic output. A decrease in pressure throughout the vocal tract leads to difficulties in producing consistent loudness or a controlled variable loudness in situations such as producing stress. Monoloudness is a feature of all dysarthric types. The lack of control of exhalation also affects the output synthesis rhythm and intonation as phrases between inhalations are shortened or rushed to be completed.

2.4.3 Theory of dysarthric speech production

In looking to personalise a VOCA output using dysarthric speech, the synthesis technique could approach the issue from the underlying problem of the production of dysarthric speech. Appendix A details relevant theories of speech production to speech synthesis techniques, particularly with this approach, articulatory synthesis (see section 3.3.1). Alternatively, the problem can be tackled by addressing the consequences of dysarthric speech and being able to manipulate the acoustic output of dysarthria using models of unimpaired speech.

As has been discussed above, the problems of dysarthria do not occur solely at the segmental level, issues of timing and motor control imply that a method of synthesis that can take into account aspects of articulation at a hierarchical rather than a linear segmental approach would be better suited to this type of data. This issue is discussed in more detail in chapter 3.

2.5 Requirements of a VOCA

The above discussion has provided motivation for personalising a VOCA output voice to sound more like that of the VOCA user before their speech deteriorated. To find a suitable method for this task, the ideal requirements for the voice, the technique and the person involved have been extracted from the discussion above and restated as follows:

- Voice
  - Intelligible
  - Natural-sounding
  - Similar to the user
  - Manipulable for prosodic output
- Technique
  - Minimal data input
  - Can use dysarthric data input
  - Practical tool available
- Person
  - Have emotional readiness

These requirements can now form a specification for the selection of an appropriate voice building for speech synthesis technique, which provides the basis for chapter 3.

2.6 Conclusions

This chapter has discussed the problems associated with using voice output communication aids and the implications of these problems for social interaction. It has identified personalising the voice to sound like the VOCA user as a particular issue that is hypothesised as contributing to the encouragement of the user to accept and use the communication aid and increase their social interaction. It is acknowledged that providing a personalised voice will not solve the problem of VOCA acceptability but it may contribute to the overall issue. There is also currently no provision for building synthetic voices using dysarthric speech and therefore finding a technique which contributes to that application could be valuable. Detailing the effects of dysarthria on speech provides an insight into which speech parameters need to be tackled in building a synthetic voice using that type of input. A review of the acceptability issues has defined the requirements for a VOCA in terms of the output voice, the technique and the person. This list is used to structure the review of speech synthesis methods in the following chapter.

Chapter 3

Speech synthesis methods and evaluation

3.1 Introduction

This chapter details methods of synthesis that could be used for the task of personalising synthetic speech output. The requirements for an appropriate method were stated in chapter 2 as: having an intelligible, natural-sounding output which sounds like the target speaker and access to manipulate the prosody, the technique must use minimal training data input, have the potential to use speech data that has started to deteriorate due to a speech disorder and there must be a practical tool available for use.

To understand how intelligibility, naturalness and similarity to target speaker are determined, section 3.2 provides a description of how they are evaluated using both subjective and objective measures. Section 3.3 describes articulatory synthesis, parametric synthesis, concatenative synthesis, model-based synthesis and voice conversion techniques in terms of the appropriateness for the task. It concludes by summarising which techniques are available for use and identifies the most suitable method for personalisation of synthetic voices for people with severe speech impairment.

3.2 Evaluation

Evaluation of synthetic speech focusses on intelligibility and naturalness of the output. Techniques of *voice transformation* or *voice conversion*, changing a voice from one speaker to that of another, introduced a need for a further evaluation dimension: similarity of the speech output to a particular target speaker. The aim of speech synthesis is to produce a speech output that is acceptable along all these dimensions to human listeners. This means that most speech synthesis evaluations produce a subjective measure as conducted by human listeners. Any objective measure is expected to correlate with those listener judgements. Using human participants is time-consuming and costly and therefore not practically feasible to do frequently through the development process. Ideally, a naive listener response to any minor change in technique or parameter alteration would be performed but these practical restrictions mean that the evaluation is usually a non-formal judgement made by the researcher themselves. It is clear that an objective measure would be useful in this case to replicate listener judgements to relieve this inaccuracy in the developmental process. Human listeners are potentially unreliable in their judgements when their motivation for taking part is questionable or factors such as fatigue are introduced. A consistent objective measure provides additional reliability for the results when used in combination with the human listening experiment results.

Difficulties in finding such an objective measure for speech synthesis stem from an individual's capability to produce the same utterance with multiple 'correct' realisations, particularly in terms of prosody. This variation needs to be taken into account in using any objective measure of the speech. The following sections detail the types of subjective and objective evaluations that are conducted for intelligibility, naturalness and similarity to target speaker.

3.2.1 Intelligibility

Subjective measures for intelligibility usually consist of 'type-in' tests, where a listener is played a stimulus and asked to type or write down what they heard. How much of the stimulus they correctly identified provides the measure of intelligibility, usually word error rate (WER).

The type of stimulus can test different aspects of a speech synthesis system. For example, phonetically confusable sentences can test the intelligibility of individual segments and semantically unpredictable sentences can test the overall intelligibility independent from contextual information.

The modified rhyme test (MRT) [92] is a set of sentences where monosyllabic words are presented to a listener in a carrier phrase, such as 'now we'll say WORD again'. The items inserted into the phrase are taken from a list of phonetically confusable initial and final consonants in words, such as 'rig' and 'wig' or 'beat' and 'bead'. The MRT is an extension of the diagnostic rhyme test (DRT) [137], which tested for word initial consonants only. Type-in tests are performed using these structures and assigned an error metric. Confusion matrices can be built based on the results of the evaluations showing the confusability of individual segments which can diagnose which particular aspects of the synthesiser are causing problems for intelligibility.

Semantically unpredictable sentences (SUS) [13] are sentences which are syntactically well-formed but semantically meaningless. The use of semantically unpredictable sentences in speech synthesis evaluations means that the listener is restricted to using non-contextual cues to interpret the intelligibility of the stimulus sentence. Synthetic speech systems are designed to handle semantically predictable output and so using these stimuli sentences allows a test of completely unseen data and can standardise evaluations across different synthesis systems.

3.2.2 Naturalness

A measure of naturalness is usually elicited subjectively from listeners using a mean opinion score (MOS). This is based on a Likert-type five, seven or other point scale with each point representing a level of naturalness, for example 1=bad and 5=excellent. These perceptual categorical scales can vary on what is being tested, for example as applied to naturalness the scale would range from 'very unnatural' to 'very natural'. Synthesised speech samples are presented to a listener who assigns a rating appropriate to that output.

Alternatively a pairwise forced choice of stimuli takes place, where a listener is asked to choose one of two speech samples as being more natural than the other. Each experimental condition item is paired with all other versions of the same sentence. The results of these experiments result in an overall ranking of the naturalness of the different stimuli conditions.

Attempts have been made to find the underlying perceptual factors involved in making naturalness judgements using multidimensional scaling [144]. Listeners do not give equal weighting to the perception or attention of all the dimensions or features of a complex acoustic stimuli. In this study, multidimensional scaling was used to analyse and determine which psychoacoustic and physical components in synthetic speech, both sub- and supra-segmental, were most important to the listeners' judgements of naturalness. These types of analyses could eventually inform an objective measure for dimensions such as naturalness.

3.2.3 Similarity to target speaker

To evaluate similarity of speech to a particular target speaker, participants are presented with one or multiple samples of speech from the target speaker and a sample synthesised using the system under evaluation and asked to rate the similarity of the synthesis to the original speaker. This is usually done with a perceptual categorical five or seven point rating scale ranging from 'very dissimilar' to 'very similar', or 'sounds like a different speaker' to 'sounds like the same speaker'.

Alternatively, a direct comparison can be done using an 'ABX' method, where A and B represent stimuli produced using the different conditions under investigation and judges are asked which of the two, presented in a random order, is most like X, the corresponding target speaker stimuli. The perceptual categorical rating and the ABX methods can both be done where the word sequence of the synthesised conditions is different from that of the target speaker.

For objective measures of speaker similarity, direct comparisons can be performed by finding a distance metric between the synthesised speech and the original speech. This requires the two speech samples to have the same content. Such measures include mel cepstral distortion (MCD) [201], frequently used in voice conversion evaluation. MCD finds a distance between the values of extracted mel cepstral coefficients for the original speech and synthesised versions. This can be done using time alignment, for example using dynamic programming (DP), or relying on the two samples being similar in length and directly comparing frames of speech, implicitly penalising for any durational differences of segments within the synthesised speech. Other techniques can force the generation of features to have the same durational aspects as that of the target speech, which means that a direct frame comparison can be done [223]. This measure can either be performed on the whole utterance or on sections which reveal meaningful differences in terms of spectral characteristics, such as voiced sections or vowels.

The MCD comparison reveals the difference in spectral characteristics in isolation. Methods of comparison for similarity combine this measure with evaluations for other features such as root mean squared error (RMSE) or a correlation between original and synthesised output for fundamental frequency (F0), the acoustic correlate of pitch, and a comparison of vowel length [224]. Ideally, an objective measure should evaluate all of these features in combination to more closely replicate human perception of speaker similarity.

The objective measures described above are used to evaluate the effectiveness of a voice conversion technique where any overall effect of similarity is dealt with at frame level only. An objective measure may benefit from approaching this task from a higher supra-segmental level which takes into account overall pitch levels or smoothness of the output, for example, and may correlate better with human perceptual judgements.

3.3 Methods of synthesis

Chapter 2 detailed the requirements for the method of synthesis for building personalised voices for individuals using communication aids. The following sections detail articulatory synthesis, parametric synthesis, concatenative synthesis, model-based synthesis and voice conversion techniques in terms of the following requirements: the voice must be intelligible, natural-sounding, sound like the target speaker and provide access to manipulate the prosody; the technique must use minimal training data input and have the potential to use speech data that has started to deteriorate due to a speech disorder. In addition, there must be a practical tool available within which the research can be performed.

3.3.1 Articulatory synthesis

3.3.1.1 Introduction

To produce natural-sounding and intelligible synthesised speech, one approach is to model the movements of the vocal articulators. Articulatory synthesis provides a method of synthesis which can approach the problem of dysarthric speech modelling from a speech production perspective. This section describes articulatory synthesis and discusses its appropriateness for personalised speech synthesis.

3.3.1.2 Articulatory modelling

Articulatory synthesis attempts to model human articulators and the behaviour of the vocal folds to produce a synthetic output. This can be done through a description of the geometry and the dynamics of the vocal tract and articulators [84].

An articulatory synthesis system requires four components: accurate static configurations, accurate dynamic movement, ability to configure the system and a linguistic control of parameters [100]. The static configurations of articulators and dynamic motions must match observed configuration data for different speakers. The data used to inform this model is gathered using imaging such as X-ray, electropalatograph (EPG) and more recently magnetic resonance imaging (MRI), which is becoming more widely used to provide a three-dimensional picture of the speech process [162]. The transient nature of speech and the exact movements involved in the speech production of an individual can be captured using this technique. The aim is not to reconstruct all the possible movements in the vocal tract but to establish an optimally small number of parameters which can capture all the commonly occurring configurations and dynamic movements.

The set of articulatory parameters comprise those representing the vocal tract shape and those representing the excitation. This distinction is based on the source-filter model of production [68] which states that the source, the vibration of the vocal folds or turbulence created at a constriction in the vocal tract, and the filter, representing the shape of the vocal tract, are independent of each other and therefore separable. This theory is discussed in more detail in section 3.3.2.2. The parameters indicating the vocal tract configuration include features such as: lip aperture, lip protrusion, tongue tip height, tongue tip position, tongue height and velic aperture. The parameters for the excitation modelling include: glottal aperture, vocal fold tension and lung pressure. The acoustic model estimates the speech waveform from the sequence of geometrical functions with the corresponding sound source parameter functions as they change over time.

The configurability component refers to being able to configure the parameters sufficiently to capture the variation inherent between human vocal tracts. The CASY (Configurable Articulatory SYnthesiser) system, for example, attempted to account for speaker variation by altering parameters to more closely match the specification of the shape and size of various parts of an individual's vocal tract as viewed in magnetic resonant imagery [175].

The final component is that of using an appropriate method of input to the synthesiser which explicitly relates to the representation of the linguistic theory of speech production that is being used (see appendix A). For example, the CASY system [100, 175] uses a task dynamics (see appendix A.5) approach to speech production and relates the parameter descriptions to the lower level processes of physical articulatory movement.

Articulatory synthesis has the potential to simulate all aspects of human speech production, including dysarthria, from an articulatory level. Using an appropriate cognitive or linguistic parameterisation, the motor deficits which result in dysarthria could be modelled. Having knowledge of the underlying processes including the cognitive, motor control and articulatory processes involved in the production of disordered speech could provide more insight into how to best correct for these disorders in building a personalised articulatory synthesis system.

The gaps in knowledge of speech production currently prevent the development of a fully functional articulatory synthesis system [194]. It is currently used successfully for applications in speech therapy and fields where a simulation of the vocal tract is useful [15, 146, 210]. An example of the output of articulatory speech synthesis is available as example 3.1 on the attached sound file CD.

3.3.1.3 Personalisation requirements: summary

For synthesis, the articulatory information is usually input into a parametric synthesiser which can produce intelligible output. Where it uses parametric synthesis, it suffers from the same problems in producing natural-sounding speech (see section 3.3.2). Access to the prosody is available using, for example, a fundamental frequency parameter that can be altered accordingly.

For personalisation of an articulatory speech synthesis system, a comprehensive set of articulatory data is required and an appropriate model of speech production is not yet accounted for [194]. The data collection process is extremely time-consuming and expensive, particularly if MRI is used. An articulatory system is usually based on the movements of one individual and takes a lot of imaging data to provide a fully personalised system for that particular individual. It is however possible to alter already existing parameter sets to more closely match the speech of an individual [175]. For those individuals whose voices have started to show deterioration then this methodology could provide valuable diagnostic information and provide insight into the workings of a cognitive and articulatory system affected by dysarthria.

Using articulatory synthesis does not necessarily rely on a segmental approach to speech production, which may be a more appropriate technique to model speech with disordered inter-articulator motor control and timing. However, there may be a set of target configurations corresponding to particular segments, depending on the system.

3.3.2 Parametric synthesis

3.3.2.1 Introduction

Articulatory synthesis is an attempt to take knowledge about the human vocal tract and speech production and implement it to synthesise speech. *Parametric synthesis* represents the same articulations but in terms of the resultant acoustic output, attempting to model the resonant frequencies and their amplitudes of articulations in the vocal tract. Taking these parameters that represent the perceptually important characteristics, a waveform is generated by exciting a set of resonators, outputting the appropriate spectral features. This was first described by Dudley [60] who termed the system a *vocoder* (VOice CODER). It was a technique originally used in telecommunications applications to recreate speech from a transmitted coded representation. This type of synthesis is also based on the source-filter theory which is described in section 3.3.2.2.

This type of synthesis is dominant in voice output communication aids. The voices are highly intelligible [80, 149, 150], have a small memory footprint relative to other techniques and are robust to manipulation of prosodic parameters. However, these synthesised voices are comparatively not as natural-sounding and personalisation is difficult, as discussed below.

3.3.2.2 Source-filter theory

Parametric synthesis relies on the shape of the vocal tract and the excitation source being independent of each other and therefore separable. This is known as the source-filter theory [68].

The production of sound in the vocal tract is dependent on two things: the excitation and the articulation shaping the vocal tract. The excitation source is either the vibration of the vocal folds, the production of turbulence at a particular constriction in the vocal tract or a combination of the two. The configuration of the vocal tract modifies the excitation as it passes through by emphasising certain resonant frequencies (*formant frequencies*) and attenuating others.

The source-filter theory states that the vocal tract can be represented as a linear filter which varies over time as the shape of the vocal tract changes. This filter is modelled by a set of resonators which is then excited by a source. The assumption is that there is no other interaction between the source and the filter and so they can be modelled separately

[88, 121, 191]. The source-filter model is illustrated in figure 3.1.

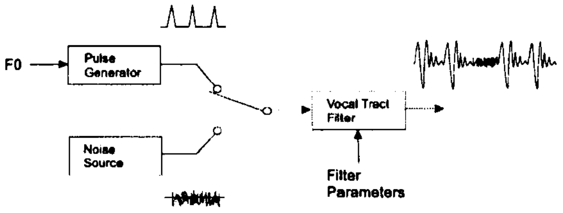


Figure 3.1: *Source-filter model (based on [76])*

3.3.2.3 Modelling the source

To model the source involves producing a model of both periodic and aperiodic waveforms. Voicing is produced by the periodic opening and closing of the vocal folds whereas voiceless sounds are produced by turbulent airflow at a constriction in the vocal tract. Modelling the glottal flow is highly complex so most source models make an approximation to it, usually using a pulse train. This is then modified by a filter to emulate the effect of the energy in voiced excitation occurring mostly at the lower frequencies. The fundamental frequency is then controlled by varying the frequency of the pulse train.

Voiceless sounds occur when the vocal folds are held open and the flow of air from the lungs passes freely through the glottis. Where there is a constriction in the vocal tract, the air pressure is altered and the flow of air takes on a turbulent quality. This can be crudely modelled by random noise, although it does not attempt to model the constriction itself.

These models also need to be able to produce mixed excitation for sounds that have both of these qualities, such as voiced fricatives. The more accurate the modelling of the source, the more natural-sounding the speech [2, 31, 116].

3.3.2.4 Modelling the filter

The filter attempts to reproduce the spectral characteristics of the vocal tract. Due to the movement of the articulators and changing shape of the vocal tract as a sound is produced, certain frequencies are attenuated and others are enhanced. The formants are the peaks in the frequency spectrum and can be modelled simply using a small set of poles. This has been shown to be a reasonable model of the vocal tract until nasal sounds need to be synthesised. Nasal sounds are created by the use of the nasal cavities in combination with the vocal cavities which introduces some anti-resonances where the nasal cavity dampens the sound. Therefore a model for a full range of speech sounds has to be more complex including both *poles* (peaks in the transfer function to model formants) and *zeroes* (troughs in the transfer function to model anti-formants).

3.3.2.5 Generating the waveform

The synthesis of the formants is achieved using a set of resonators. For example, a resonator takes the parameters of a resonance frequency and bandwidth to produce a transfer function. Anti-formant filters can also be used taking the same parameters. To model the entire vocal tract, each resonator represents a formant or anti-formant and they are connected together either in parallel or cascade configurations.

The cascade configuration, as shown in figure 3.2, has only one amplitude setting for all formants and the output of each resonator is the input to the next resonator. This configuration corresponds well to how the vocal tract works and is good for vowel synthesis. This configuration only requires the resonant formant frequencies and the amplitude gain for the whole system as parameters.

Connecting the resonators in parallel, as shown in figure 3.3, means that each resonator is controlled for its formant amplitude separately. The formants are produced simultaneously and then summed. This configuration allows for the introduction of zeroes as well as poles and is therefore better for modelling nasal sounds and other consonants than the cascade configuration. The parallel configuration requires more control parameters as input as each formant resonator has to have its own amplitude setting (A) and the resonators require both the formant frequencies (F) and their bandwidths (BW) as input. This means that parallel configurations demand more computational resources to deal with the increased control information and as such are used more in formant synthesisers for speech modelling rather than for practical speech synthesis. The JSRU/Holmes synthesiser used this configuration and produced a high quality output for male speech, although this was mainly due to lengthy manual optimisation of the parameters [89, 87]. An example of this synthesiser is available as example 3.2 on the CD.

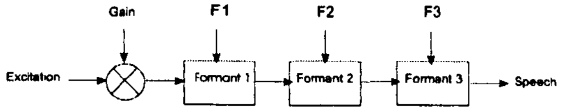


Figure 3.2: *Basic cascade formant synthesiser configuration (based on [124]. F=formant BW=bandwidth)*

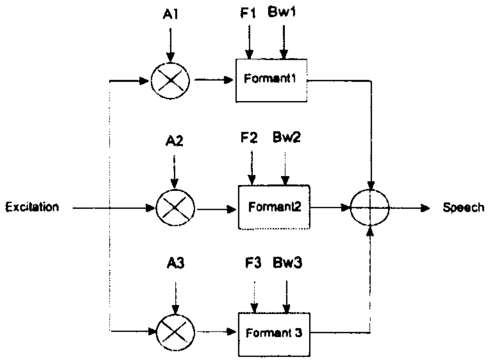


Figure 3.3: *Basic parallel formant synthesiser configuration (based on [124]. F=formant BW=bandwidth A=amplitude)*

The KlatTalk system [114] combined these two configurations into a hybrid formant synthesiser that used a parallel configuration for consonants, adding in separate resonators for nasal sounds and a cascade configuration for vowels. An example of the Klatt synthesiser is available as example 3.3 on the CD. The quality of this synthesiser led to the subsequent development of the MITalk [4] and DECtalk^{TM}systems (available as example 3.4 on the CD). This synthesiser used a total of 40 parameters, consisting of the frequency and bandwidth of the first six formants and their amplitudes. Other parameters are input time functions for frication, aspiration source amplitudes and other parameters that better model nasal sounds and the source.

Formant synthesis output has been described as having a slightly robotic or buzzy quality. This is partly attributable to updating the parameters at regular, non-phase-synchronous intervals. The regular parameter update results in an extra frequency component in the signal at resynthesis, introducing a 100 Hertz (Hz) buzz (depending at what frequency the parameters are updated). Another contributing factor to voice quality is the size of the analysis window. For windows of size 25 milliseconds (ms), it is likely that there is a varying number of pitch pulses in each window across the sample. This can result in a rougher quality of output as the F0 estimate is affected and rapidly changes in the resynthesis of the speech. If an analysis window is extended in size to capture more data to estimate the F0, it can overlap too much with unvoiced parts of the signal which also provides an unreliable F0 trace. Unvoiced sounds contain aperiodic high frequency components which when incorrectly assigned as being voiced, allow more of the high frequency components to be synthesised, which creates buzziness at boundaries between voiced and unvoiced sections. Attempts have been made to improve the voice quality by improving the quality of modelling of the excitation, for example, including ways of avoiding the definition of two different systems for the voiced/unvoiced excitation to reduce buzziness at voicing boundaries [2], using different source parameters for the glottal source model [116] and techniques based on manipulation of the phase [108].

3.3.2.6 Input to the synthesiser

Formant synthesisers require a way of setting the values for input to the resonators at synthesis time. One approach to defining the parameter values is using rules based on the acoustics and phonetics of natural speech. Rule-based synthesis is based on each segment having a target articulation set of parameter values. Parametric synthesis implements coarticulation theory (see appendix A.2) where the abstract representation of the segment that the speaker is aiming for is defined and variations from this target are due to the surrounding context. The values for these targets are derived from tables [31, 90] with target parameter values for every phone. These rules allow coarticulation theory to be implemented as a durational value providing a minimum period of time that the transition would need to take place. It takes the view that speech is generated on a hyper-hypo dimension, that articulations in natural speech are altered on a continuum depending on both the individual's control and contextual factors [132]. This allows undershoot of articulations in a shorter than usual segment to increase naturalness, to replicate the behaviour of the articulators in certain coarticulatory conditions.

The rules are generated by linguistic, phonetic and phonological knowledge to make an abstract representation from which to synthesise the utterance, based on the acoustic analysis of speech data [29, 90, 115, 127]. Klatt [114] details how these rules are applied to convert a string of phonemes into speech. The first step is to take the string of phonemes and using a set of rules, replace the phonemes with allophones (context-dependent realisations of the phonemes) which need to be articulated. Each segment is then assigned an inherent duration, which is taken from a list of rules in the program. Taking into account the wider context, rules applying to phrasal and sentential factors are then applied, further adapting the parameters. F0 contours are defined, determined by rules about the placement on intonational accents and smoothing between them to create a realistic output. Further rules are then implemented to characterise the voicing of the phones, their sound sources and the resonance properties. The final step is to take this parameterised string and input it into the formant synthesiser to output a speech waveform.

One of the problems of parametric synthesis is that there is one set of target parameter values from which a deviation occurs as detailed above. The lack of variability in the output due to this fixed set of target values contributes to a perceived unnaturalness of the synthesised speech [194].

The difficulty in rule-based synthesis is in producing the rules at the initial stage. It is a time-consuming and laborious process to put together all the information needed to build a synthesiser. Attempts have been made to derive rules and target articulations from data automatically in combination with rules previously defined [85, 166]. Högberg's adaptation of the KTH GLOVE TTS system [31] added predictions of vowel formant frequencies to the phonetic rules which modified the default values for the parameter synthesis. These predictions were based on classification and regression tree (CART) models [25], which are described in more detail in section 3.3.3.8. An extension of this added in an improved automatic formant extraction algorithm and replaced voiceless fricatives with recorded versions and inserted them into the output. Using concatenative synthesis to add in recorded versions of those sounds which are unnatural-sounding when produced with a formant synthesiser is one way of improving the naturalness of this method. This increases the amount of storage space needed for the synthesiser and could produce some distortion in the signal at the concatenation points.

Other methods of input relate to underlying structures based on phonological theory, for example the Yorktalk system [40, 165] where phonological structures are assigned phonetic realisations which are then used as the input to the formant synthesiser. This method decouples the phoneme string from the phonetic realisation process allowing a non-segmental phonetic interpretation. This in turn removes any rapid changes in parameters and therefore distortions that occur in systems that place less emphasis on capturing the coarticulatory features of speech.

The SPRUCE system [126] also used phonological representations to remove discontinuity problems at the segmental level as input to a parametric synthesiser. The phonological representation was derived at the syllable level to which the phonetic realisations were assigned from an inventory of stored parameterised real speech syllables in combination with an appropriate prosodic parameter sequence at synthesis time. Using syllables as the base unit allowed a more natural capture of coarticulatory effects while the size of the stored inventory was manageable and an exhaustive closed set. By using the acoustic properties of actual stored speech, the output was intrinsically more natural by more closely matching the variability found in real speech, in comparison to the oversmoothed effect commonly found in parametric systems.

3.3.2.7 Advantages of parametric synthesis

The separation of source and filter means that parametric synthesis has a great deal of flexibility in modifications of prosody: F0, amplitude and duration. F0 modification is done by changing the parameter for the source excitation. Duration is altered by changing input values by a certain factor. Parametric synthesis has a relatively small memory footprint and this combined with its reliability and high quality intelligibility is the reason why it is only recently that concatenative synthesis has started to encroach on its previously monopolised area of assistive technology devices. The high quality DECtalk$^{\mathrm{TM}}$voices may also be slightly more robust in noise than other synthesis methods [137, 214] although the conditions used in these studies of speech synthesis in noise are not necessarily directly comparable. These studies do seem to agree that the more detail involved in modelling the speech and the more a system takes into account the natural variation in speech due to contextual factors, the more intelligible it is likely to be.

3.3.2.8 Personalisation

Optimising the parameters for personalisation of a parametric synthesiser is a time-consuming process. Murray and Arnott [157] attempted to provide rapid personalisation of a voice for the DECtalk$^{\mathrm{TM}}$synthesiser using two levels of editing: interpolation of the existing voices and other more detailed changes to the individual parameters. The EDVOX system permitted such interpolation to introduce a level of individualism into the voice but did not allow more detailed personalisation to reconstruct the voice qualities of the individual. Suggestions for personalisation have come from the KTH systems as set out above [85, 166], using input from an individual's speech in combination with the pre-defined rules for synthesis.

The main attraction of a formant synthesiser is its flexibility in manipulation of the parameters, although this optimisation may take time and is complex due to interactions between parameters. Tools which profit from establishing the interaction links, such as HLsyn [78, 189] are useful for this, in addition to results from experiments with correlation studies between personal characteristics and certain parameters, for example Schötz's [180] study of F0 contours and segment duration for age characteristics.

Using an individual's targets in a rule-based system is one option for personalisation although this is problematic as the acoustic correlates of phenomena such as the prosody and speaking styles of speech need more investigation [158]. This technique is a possibility to use for a person who is about to lose their voice although it may be more difficult for someone whose voice has already started to deteriorate as their targets for particular phones may be inaccurate. Parametric synthesis implements coarticulation theory and as such relies on a sequence of target parameter values. The nature of dysarthria is such that it is not just a question of not hitting targets but of the timing and control of the misarticulations. There would have to be a process to decide which targets to include and combine with pre-defined rules and targets of a donor speaker.

3.3.2.9 Personalisation requirements: summary

Parametric synthesised speech was the dominant type of synthesis found on VOCAS due to the small memory requirements and having a highly consistent intelligibility. The voice quality and the lack of variability in the output signal mean that it is lacking in naturalness but the voice quality of these types of synthesisers is improving with higher quality signal manipulation techniques and better modelling of the speech source. The flexibility of a parametric synthesiser lends itself to easy manipulation of the prosodic and other parameters. However, access to the appropriate parameters and the mapping between the parameters to particular characteristics of an individual is not always straightforward. There is potential to personalise a rule-based parametric synthesiser taking some data from an individual and combining it with pre-defined rules and target values.

3.3.3 Concatenative synthesis

3.3.3.1 Introduction

Concatenative synthesis output combines the naturalness of pre-recorded utterances with the ability to synthesise any novel utterance. Where it is impractical to pre-record every possible utterance, one alternative is to take recordings and segment them into smaller units of speech which are re-combined in different ways to make novel utterances. Concatenation does not simply join one chunk to another - the listener must be fooled into not perceiving the join. This is conventionally done with the pitch synchronous overlap and add (PSOLA) technique (see section 3.3.3.9). Concatenative synthesis can produce very high quality natural-sounding speech but requires a large amount of storage space and a lot of recording to provide a database from which to select and re-combine the speech. Once this voice is recorded it is not simple to personalise as this generally involves re-recording a new database.

3.3.3.2 Slot and filler systems

Where there is a highly constrained context such as the speaking clock or train announcements for railway stations, it is possible to record a closed set of units of information and re-combine them to produce all the possible output required. An example train announcement is:

"The next train to arrive on platform 5 is the 13.00 from Crewe."

The system could be broken down with the following structure leaving slots which are filled by other sections of recorded information:

"The next train to arrive on platform NUMBER is the TIME from PLACE."

The words in capital letters represent variable slots where all the possible platform numbers, times and places are recorded within that structured carrier phrase and then slotted into that framework as required. If enough care is taken to ensure that the intonation patterns remain constant in all the recordings, this slot and filler type synthesis is highly intelligible and natural-sounding. It requires a fraction of the amount of data needed to produce all possible combinations required for output and is therefore very suitable for such applications.

3.3.3.3 Units of language

For wider coverage of language and the ability to produce unrestricted output, more data must be provided from which to select the units for concatenation. Classical phonetic theory (see appendix A.1) makes the assumption that speech is composed of a sequence of discrete sounds (although overlapping or smoother together using coarticulatory processes) which make up a closed set of the underlying phonological segments of a language.

This closed set of units is the set of *phonemes* of the language. In classical phonetic theory, the phoneme is the smallest unit in a sound system that distinguishes meaning between words. This finite number of phonemes are used to produce any underlying representation of larger meaningful units such as syllables, words and sentences in a language. There are approximately 46 phonemes in British English depending on accent [103, 220]. They can be used as a base unit from which to derive a surface representation of *phones*, where a phone is defined as the acoustic realisation of a phoneme. A database of recordings that takes into account all of these phonemes and the contexts where they differ in pronunciation then has the coverage to create a synthesised voice capable of saying anything in the English language.

3.3.3.4 Contextual factors

Each phoneme has a range of variations in realisation: the *allophones*. The realised allophone depends on the articulatory context. As the articulators attempt to produce the phone sequence, the target articulations are not always reached, the movements overlap with each other and are influenced by the surrounding context. This overlapping with surrounding articulations is known as *coarticulation* [35, 120]. Prosodic factors such as stress or position in the syllable, word, or phrase influence the phonetic realisation of the phoneme. Any concatenation of units that does not take this into account makes the output sound disjointed, reducing the naturalness and intelligibility. Concatenative synthesis captures the natural prosodic effects that the speaker has used in the recording of the database although it is relatively difficult to change prosody that is not present in the recordings. Covering these prosodic contexts involves further recordings which is often not possible due to practical constraints such as the ability of the speaker, time restrictions and the increased storage space needed.

3.3.3.5 Unit type

Using different sized units for concatenation helps to overcome problems for this technique caused by coarticulation. Using phones as the unit for concatenation, where the join point is at the point of maximal coarticulation, results in a highly disjointed output. Moving the concatenation point to the period of maximal stability in the sound wave reduces that distortion. A *diphone* is therefore defined, which is the size of a phone but is two phones in sequence where each end of the diphone unit occurs in the middle section of each of the phones. This approach handles only coarticulation effects caused by immediate neighbour phonemes. It does not accommodate coarticulation or prosodic effects caused by a wider surrounding context.

By modelling the transition to one side of the target phoneme, the size of the minimum required inventory is N^2 units, where N is the number of phonemes in a language.

Smaller units than diphones have been used in concatenative synthesis with the aim of better modelling the contextual factors with less data. These smaller units can be used for multiple contexts in combination, reducing the amount of data for the same amount of coverage. Half phones have been used as the base unit for synthesis [14] which therefore allows diphone and phone synthesis in combination. Other systems have used sub-triphone HMM (Hidden Markov model) state level units for concatenation [55, 93, 172] or frame-sized segments [134]. These smaller units mean more points of concatenation and more potential for discontinuities at unit boundaries, but the reduced size of the unit means there is less within-unit variation.

Where the aim is to produce unlimited amounts of high quality speech, a large amount of data is needed to contain more coverage of phones, syllables, words and sentences that can be used in a unit selection system. Having a multi-unit length selection system allows a search algorithm to select the largest applicable unit to match the input, for example, [38]. The larger the unit selected the fewer join points which therefore minimises the potential for disruption for the listener. The more data that is available in the database provides more choice of units for concatenation. Full coverage of units is necessary to be able to produce all possible outputs but does not guarantee high quality in that output. Having more data means there is more chance of producing the required unit with a closer match to its surrounding units for concatenation.

3.3.3.6 Database design

The question can then be asked of how large a database would have to be to cover every potential utterance using this concatenative approach. Successful voices have been made for unit selection systems with databases for English consisting of approximately 80 minutes of speech using the 1132 phonetically balanced utterances of the Arctic database [11, 12, 117], diphone voices by recording around 1400 nonsense words in carrier phrases within the FestVox system [21] and 525 phonetically balanced ATR (Advanced Telecommunications Research) sentences for Japanese synthesis [19]. The concatenative synthesiser is dependent on full coverage of units and if a unit is not captured in the recorded database then that unit cannot be synthesised.

Having more data from which to select units is one way of increasing the likelihood of getting high quality output. Professional voice artists are usually employed to provide the recordings as they need to be of high quality and consistently recorded in quiet conditions. It is a difficult task to produce a large amount of accurate consistent speech with a natural-sounding prosody so the database is kept as small as possible while still maintaining full coverage of the units required. In the discussion of the Multisyn voice building algorithm, it was noted that larger databases have more coverage but in experiments, the increase in quality was not sufficient to justify the increased amounts of recordings needed and time taken to synthesise output utterances [38].

Restricting the domain, meaning topic area or application style, of a synthesiser improves the synthesis where the target utterance overlaps with content in the database [20, 176]. The unit selection process usually has a weighting to prefer the selection of a string of units for concatenation that occur adjacent to each other in the recorded database, for example the Multisyn algorithm for Festival [38]. Where the domain matches, the probability of getting a good match to that unit and its transition to the next unit in the database is much higher.

In construction of a diphone synthesiser database, all diphones are reproduced in carrier sentences to ensure prosodic consistency. Recording nonsense words such as "t-aa b-aa b-aa" to collect prosodic and context controlled units in the FestVox [21] system, has given way to phonetically balanced datasets such as the Arctic database [117] which compacts coverage of all US English diphones into 1132 sentences taken from out-of-copyright novels. The database consists of two sets: set A, which contains all diphones of US English in 593 sentences and set B, which has almost complete coverage in 539 sentences. They were designed to contain sentences between 5 and 15 words in length, with no out of date terminology, no confusable or difficult to pronounce words or names and simple grammatical structure to maximise readability. The Arctic database is a much simpler set to read for non-professional speakers and allows a range of unit sizes to be selected from it. Using real rather than nonsense words means there is an opportunity to select larger units than diphones therefore potentially minimising the number of join points in an utterance. Example 3.5 on the CD indicates the quality of limited domain synthesis and example 3.6 indicates Festival's general synthesis quality. Example 3.7 demonstrates an example of the author's own voice recording which contributed to building a Festvox voice, an example of this synthesis is available as example 3.8.

3.3.3.7 Measures

Selecting the most appropriate chunks to maximise intelligibility and naturalness of the novel utterance in concatenative synthesis involves evaluation metrics. These measures should evaluate how well that unit represents what is required at that point in the utterance and how well it concatenates to reduce any distortions in the output. Hunt and Black [96] discuss the notions of target cost and join cost for the selection of units in concatenative synthesis:

- The *target cost* is the cost of choosing the correct and appropriate unit in terms of what sounds make up the utterance to be synthesised and choosing the unit with, for example, the same surrounding context.
- The *join cost* is the cost between two units that are to be concatenated, as any mismatch at the join points introduces disruption into the signal.

These measures in combination are minimised to find the overall most appropriate string of units for concatenation. They are objective spectral measures and have been found to have reasonable correlations with subjective human perception of discontinuities in the synthesis output [215, 222].

3.3.3.8 Selection algorithms

In selecting units for concatenation, the selection process has to find the units that minimise the target and join costs. This selection is done using clustering techniques, collecting together similar acoustic units and classifying them using information provided by a text analyser. Units are indexed by their linguistic, phonetic and prosodic features such as: previous and following phonetic context, prosodic context, stress, position in the syllable and position in the phrase [22].

This information then provides a list of yes/no questions that can be asked about the units, for example, "is the following phone a nasal sound?". A classification and regression tree (CART) [25] method is then used to create a binary decision tree using acoustic measures of impurity to test how well these questions in particular orders split and eventually classify the actual acoustic units. This works by dividing the units up according to a splitting rule and a goodness of split criteria. The splitting rules are the yes/no questions, answers for which are provided by the text analysis of the units in the database. The goodness of split criteria compares the measures of impurity for each subgroup of the data for that question and finds the order of questions that best splits the data into the most homogenous subgroups. The available questions are the same for each phone type, depending on the information provided for each unit, but the algorithm only selects those questions that split the data significantly. This method deals with sparse data using a stopping criteria for splitting the data, such as a minimum number of units that are in a cluster before a question is asked or when the acoustic distance between all the units in the cluster becomes so small that there is no benefit to splitting the data any further. At this point the data is classified maximally into clusters.

This process results in a list of questions to reach the most homogenous subgroups. When unit selection is taking place, the questions are asked about the unit required and the tree is traversed. On reaching a cluster at the leaf node of a tree for each unit in the utterance, the best set of units is then found from this sequence of clusters. Unit selection can be thought of in terms of a first-order Hidden Markov model (HMM) state network where in this database of units, each unit is represented by a state in the network. The target cost corresponds to the state occupancy and the join cost corresponds to the transition probability. Using this representation allows the combined cost to be calculated using the Viterbi search algorithm [216] to find the best path through this network which minimises these costs [96].

The only units available for selection are those that occur in the database so any further minimisation of distortion at the join points relies on some processing of the signal.

3.3.3.9 Concatenation algorithms

Pitch synchronous overlap and add (PSOLA)

The process of concatenation is conventionally done using PSOLA [153]. Concatenation is not just simply joining the segments together in a sequence but uses signal manipulation to reduce the audible discontinuity at the join points. This technique manipulates the pitch and duration of the segments, both vital for the naturalness and intelligibility of the speech output. The aim is to persuade the listener that they are listening to a naturally produced utterance rather than a string of concatenated units. For vocoded speech, smoothing techniques were employed to deal with distortions at unit boundaries, usually taking a section around the join point and interpolating the parameters to smooth the joins [88].

The PSOLA technique directly manipulates the signal. This technique forces concatenation to take place at the point in the glottal cycle where the displacement is smallest, the glottal pulse has decayed and the following pulse is about to occur, so the discontinuity between segments is minimised. This point occurs at the rate of fundamental frequency and so the modifications take place pitch-synchronously.

The first stage in this process is to parameterise the speech by marking the pitch periods in the signal. In the time domain, the signal is segmented where the central point of each section is a labelled *pitchmark*: the point at the maximum amplitude of each glottal pulse. The section should be dominated by only one pulse and is usually the size of the distance to the next pitchmark either side of the dominating pitchmark. A window function is then applied to the section tapering the signal to the ends of the window. For unvoiced sections, the pitchmark is placed at a constant rate through the section. The join points of the segments are smoothed as the sequence of windowed segments is concatenated by overlapping and adding the signals together. For this technique to work it is therefore essential that the pitchmark labelling is accurate, which can involve checking by a human labeller.

For modification of pitch, the windowed segments can be overlapped and added with the pitchmarks closer together or placed further apart to increase and decrease the pitch respectively, see figure 3.4. Interpolation is performed between the waveforms when the pitchmarks are placed further apart. The range of the modification of the pitch is from half to twice that of the original signal. For modification of duration, the windows of speech are either replicated or removed from the signal. In the above ways, PSOLA manipulates the pitch and duration independently of the spectral envelope.

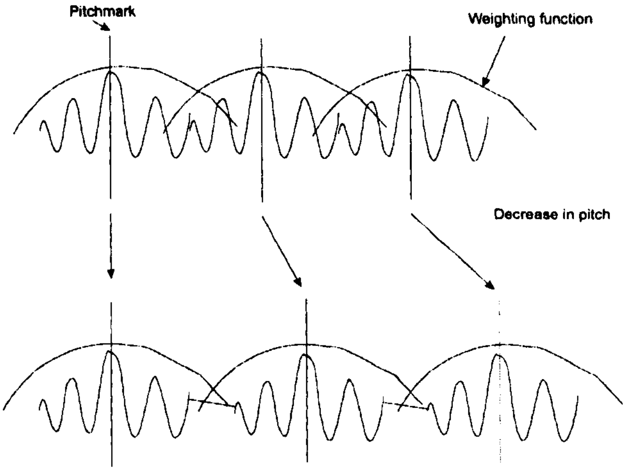


Figure 3.4: *Decreasing the pitch using PSOLA (based on [76])*

This technique can also be used in the frequency domain (FD-PSOLA). It uses a larger window size so the harmonics in the signal have better resolution. It allows modification of the spectral characteristics of the signal in the frequency domain to improve smoothing where concatenation is affected by spectral discontinuity before applying an inverse Fourier transform back to the time domain. It is, however, computationally expensive. Attempts to reduce memory size have been to combine coding and concatenation processes: linear predictive pitch synchronous overlap and add (LP-PSOLA). This parameterises the speech again separating out the source and filter, allowing for modifications of the prosodic features.

Another modification which attempts to reduce the memory size required and also provide a smoother output was the use of multi-band excitation (MBE) as the coding technique, which represents voiced speech as a sum of harmonically related sinusoids. In the multi-band resynthesis overlap and add (MBROLA) technique [62], explicit pitchmarking is not required as the fundamental frequency is modified once the units have been concatenated, removing the possibility of discontinuities due to the fundamental frequency of the segment.

Further join cost minimisation

To further reduce the join cost, other techniques are used separately or in conjunction with PSOLA. Festival uses an optimal coupling technique [42] where units for concatenation have movable join points to find the point at which the join cost is minimised.

3.3.3.10 Personalisation

Concatenative synthesisers can produce high quality speech which sounds very natural and intelligible. However, this technique is not suited to easy personalisation of a voice. To build a concatenative synthesiser, a speaker has to provide the synthesiser with a database of units. Producing a database of an appropriate size is difficult for a speaker with unimpaired speech as the recordings have to be high quality, consistent and accurately produced. This is especially true if the database is designed for maximal coverage of units using minimal data where there may be fewer instances of each unit in the database. For people with a speech disorder, how much data can be produced is dependent on that individual and the stage of their progressive disorder. The output of the concatenative synthesiser is a combination of the recordings made so if they contain disordered productions then the output also contains these segments. In some cases, the individual may not be able to produce particular sounds at all. If an individual's voice has not begun to deteriorate and they are able to produce a substantial amount of data, this technique is open to them to bank their voice.

FestVox [21] provides the tools and environment to build a synthetic voice for an individual. The amount of data an individual is required to provide depends on the type of synthesiser they build. The Arctic database (see section 3.3.3.6) can be used to make a successful voice [11, 12] comprising 1132 phonetically balanced sentences. Recording this database takes longer than the total of approximately 80 minutes of speech as the recording time must include breaks to maintain a consistent voice quality, also taking into account errors in production that have to be re-recorded. High quality recordings are required to ensure an intelligible and natural-sounding output. Using an appropriate domain, having accurate labelling with consistently spoken data, this can be successful way to build a new synthetic voice.

An approach developed specifically for people with progressive speech disorders to bank their voices is ModelTalker [28]. ModelTalker Voice Recorder (MTVR) provides the individual with a user-friendly tool to collect the database of recordings which can be done in their own homes. It requires no detailed knowledge by the individual of computers or phonetics. The individual is provided with a prompt and then repeats it to the recorder. MTVR then illustrates whether the speech is at a consistent pitch and amplitude and whether the pronunciation is phonetically accurate. If the data produced is acceptable then the utterance is accepted for entry into the database and if not then the person is prompted to repeat the same utterance. Once the recordings are completed, the database can be uploaded and the user has a synthesised voice built for them. It attempts to deal with potential disruptions in the synthesis by screening the data that it is collecting for consistency to minimise reliance on signal processing techniques at the join points. The complete database to be recorded consists of approximately 1800 utterances or 40 to 50 minutes of actual speech. The data consists of a set of utterances to provide broad *biphone* (phone plus following phone in this terminology) coverage and high frequency words. In addition there are a set of utterances that are likely to be of direct benefit for communication aid users, such as requests and personalised items that the individual is likely to use frequently. In effect, this creates a limited domain synthesis database within a larger set. An example of a ModelTalker voice built using the author's data is available as example 3.9 on the CD.

A subset of personalised data for daily interaction was also used in an attempt to build a personalised system for a patient with amyotrophic lateral sclerosis (ALS) using ATR's speech synthesis system, CHATR [30, 98]. This individual's voice was still functioning but he was breathing with the aid of a ventilator, emphasising the need to minimise the amount of data to be recorded. A phonetically balanced subset of the ATR database comprising 129 sentences was recorded, combined with texts familiar to the individual and a set of sentences which the individual prepared consisting of utterances for daily interaction. This again, in effect, creates a limited domain synthesiser within a larger set. This ensures that at least utterances required for that domain are produced at a high quality.

For those people whose voices have begun to deteriorate, attempts have been made to add their units into a database built from recordings of other voices [41, 99]. Combining these databases of units allows those units that are well articulated to be used in conjunction with some 'donor units' which the individual may have had difficulty in producing. These techniques achieved good results when the units to be added were those carrying minimal speaker identity information, such as fricative units. The donor units could be from the speaker of a similar age, accent, size and same gender, to maximise similarity of output. Consistency of recording conditions is essential for producing good quality concatenative synthesis. Replicating these conditions for a donor speaker would therefore have to be taken into account. Creating natural-sounding output with minimal distortion at join points is a difficult enough task using a database built using professional speaker data, so combining units from a donor speaker introduces more inconsistencies into the system and therefore reduces the likelihood of a high quality output.

This donor unit method relies on the individual's speech deterioration being only at the segmental level. For example, deterioration in phonatory function affecting the output voice quality would be retained in the synthesised voice. This problem would also have to be addressed, potentially through voice conversion techniques (see section 3.3.5).

3.3.3.11 Personalisation requirements: summary

This section discussed concatenative synthesis and its advantages and disadvantages for building a personalised synthetic voice. Concatenative synthesis produces intelligible, natural-sounding speech which at its best is very high quality although for this level of quality, there is a high data volume requirement. The prosody is not easy to manipulate at a suprasegmental level and if out of domain utterances are required, the output quality is inconsistent, sometimes to the point where the synthesis is completely unacceptable to the listener. See examples 3.11 and 3.12 on the CD which demonstrate the inconsistency of concatenative synthesis. Example 3.10 is an original recording taken from speaker 1 with which to compare examples 3.11 and 3.12. Concatenative synthesis has the potential to provide high quality personalised voices for those wishing to bank their voices particularly when personalised utterances are included in the database for recording. It is less suitable for those whose voices have begun to deteriorate.

3.3.4 Model-based synthesis

3.3.4.1 Introduction

Model-based speech synthesis uses *Hidden Markov models* (HMM) [173] to probabilistically model and generate sequences of *feature vectors*, discrete representations of the speech signal at a segment of time. HMMs have been successfully used in speech recognition to characterise sequences of feature vectors and these properties have recently been exploited in the speech synthesis field. This section looks at the use of HMMs and other statistical parametric systems in speech synthesis and applicability for use in personalising synthetic voices.

3.3.4.2 HMM-based synthesis

HMMs have been used in speech synthesis as part of concatenative systems for aligning and segmenting corpora [57], using states of an HMM as units for concatenation [54, 55, 93, 142, 172] and also for generating speech from HMMs themselves [56, 140, 206, 209, 232, 235].

HMMs are trained on a corpus of speech data to produce statistical models of the acoustics. Novel speech utterances are then formed by concatenating the appropriate models, generating a sequence of feature vectors (a discretisation of the signal) from the model sequence from which a speech waveform is synthesised. This technique has some of the disadvantages of parametric synthesis in its slightly robotic voice quality (see section 3.3.2.5) but has the advantages of producing highly intelligible, consistent output and is more robust to inconsistent recording conditions than concatenative systems [225]. Unlike parametric synthesis, these data-driven techniques do not demand human intervention for tuning any synthesis parameters; the variation is captured in the corpus of data on which the models are trained. Using HMMs also creates the opportunity to use speaker adaptation techniques developed for speech recognition to personalise the system with minimal data.

The HTS toolkit (H Triple S - HMM-based speech synthesis system) [228, 234, 235], an extension to the HTK speech recognition toolkit [233], provides a research tool for HMM-based synthesis. This toolkit is described in more detail in chapter 4.

The original speaker-dependent HTS system, where models are trained on one speaker's speech only, proved to be successful in the Blizzard evaluation when trained with large amounts of speech. Using full Arctic datasets (see section 3.3.3.6) of approximately 80 minutes of speech, speaker-dependent HTS achieved the highest rating in mean opinion score evaluation in 2005 for naturalness and had the highest intelligibility [11, 236]. This system uses a high quality vocoding technique for feature extraction and resynthesis - STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) [110], which reduces the buzzy quality of the output associated with parametric techniques.

Speaker adaptation techniques, originally developed for speech recognition, have been adapted for HMM-based synthesis. A *speaker-independent* model, or *average voice* is firstly estimated using speech data from multiple speakers. The average voice starting point shares some characteristics with the target speaker. It provides a strong prior for the *adaptation data*, data taken from one speaker used to adapt the models, and enables robust estimation of the target speaker model (see section 4.2.2). Given these speaker-independent models, adaptation of the model parameters towards a particular speaker is possible using minimal input data from that speaker. In the Blizzard challenge 2008, the average-voice-based system achieved equal highest naturalness and equal highest intelligibility for the voice built using one hour of UK speech data [229]. In comparison experiments, using 100 sentences or approximately 6-7 minutes of speech data, this *supervised* (where the transcription of the adaptation data is known and labelled) adaptation procedure surpasses the speaker-dependent technique using voices trained on between 30 and 60 minutes of speech [223, 227]. Example 3.13 on the CD is an example of a voice built using Arctic set A of the author's data to build a speaker-adapted HTS voice.

Experiments have been reported where average voice models are adapted towards a target speaker using *unsupervised* adaptation [113]. Unsupervised adaptation is where the transcription of the adaptation data is unknown and estimated by a speech recognition process. The recognised transcription is then used to label the data in a supervised adaptation procedure. This process showed that overall adaptation to the prosodic characteristics was reasonably successful but using unsupervised techniques decreased the intelligibility slightly. However, the impact of using a possibly incorrect transcription for the adaptation was not as severe as expected, which could be due to recognition transcription errors being acoustically similar to the true transcription and therefore using similar transforms for adaptation.

When using speech-impaired data, the individual's speaker characteristics should be captured without replicating the impairment in the output synthesis. HTS stores the adapted duration, fundamental frequency, aperiodic components of a periodic signal and spectral information as separate models. It allows access to the voicing decisions and energy components of the models. This structure allows information from the average voice model to be selectively substituted for that of the speaker where necessary to compensate for the effects of the impairment. Using this type of synthesis, rather than deal with the underlying problems associated with speech production with dysarthria and building a model of that type of speech behaviour, the approach is to deal with the consequences of dysarthria in the signal. The distortions can be identified in the output signal and a robust model of unimpaired speech is used to reconstruct the disordered signal while still attempting to retain the individual's speaker characteristics. This is described in more detail in section 4.7. The prosodic parameters are easily manipulable, for example, there is access to the rate of speech production and manipulation of the variation of the pitch output.

Recent work using HMM-based synthesis provides promising results for altering voice characteristics and individual phone identity by integrating articulatory information into the acoustic model framework of HTS [133]. This allows a manipulation of the synthesised speech output by altering articulatory parameters that contribute to speaker characteristic information such as vocal tract length. The procedure could provide more options for choices of voice output as manipulated from the original speaker although direct personalisation still requires the initial data input. The research used a speaker-dependent HMM-based synthesis technique and therefore requires more data from which to build the models. If this technique was implemented for adaptation, it could provide opportunities to identify and reconstruct consistent articulatory problems captured in dysarthric speech with a smaller amount of data.

3.3.4.3 Other related methods

Extending HTS, the CLUSTERGEN system [17] trains models and also uses these models to synthesise speech. This system does not use HMMs to generate the speech, but uses the same parameterisation techniques as HTS. It parameterises the speech so that there is a 25 dimensional feature vector comprising 24 Mel Frequency Correlation Coefficients (MFCCs) and the F0 trace. Once the speech has been parameterised, each feature vector then has prosodic and phonetic features assigned to it, rather than assigning that information to the phone unit. These features contain information about the unit itself, its surrounding context and its position at higher levels in the utterance. The vector is also labelled with which state it is assigned to and its position in the state. The feature vectors are clustered using CART models for each state assignment. A separate CART model is built which models the state durations.

At synthesis time, the text is converted to a list of context-dependent phone labels which correspond to a state sequence. The duration tree is traversed to find the state duration. To find the feature vector values for each state, questions are asked to traverse the associated state tree and the mean of the feature vectors present in the corresponding leaf node is used. This set of feature vectors is smoothed and combined with voicing decisions determined from the labels and used to resynthesise the speech using a filter.

Experiments using the CLUSTERGEN system report that using 200 utterances is a sufficient amount to produce acceptable synthesis for this system, although this amount was derived using the mel cepstral distortion measure only [17]. Experiments using CLUSTERGEN for the Blizzard challenge 2008 showed that it had a smooth and intelligible output, but it was not preferred in listening tests over a unit selection voice due to its lack of naturalness [18].

3.3.4.4 Personalisation

For the HTS average voice based system to build a voice, 100 sentences or 6-7 minutes of non-disordered speech data is required to produce a highly intelligible, with acceptable naturalness output which sounds like the person who recorded the adaptation data. This is much less data than is required for any other synthesis technique. From 200 sentences of data is needed to produce a CLUSTERGEN voice which is intelligible, although experimental reports suggest that HTS has a higher naturalness quality [106].

For those people whose voices have begun to deteriorate, there are possibilities to use model-based synthesis. Where a minimal amount of data is needed for personalisation using adaptation, it may be possible to select intelligible portions of speech from a larger database. This relies on a certain level of speech intelligibility and that within a set of recordings there will be some examples of intelligible speech that can be modelled. In combination with a data selection technique, adaptation techniques allow a more selective adaptation procedure, adapting only those features which contain speaker characteristics or using information from the data-rich average voice model where the disordered speech has been affected.

3.3.4.5 Personalisation requirements: summary

This section has looked at statistical parametric synthesis techniques and the advantages and disadvantages of these systems. These data-driven techniques take advantage of the flexibility of not relying on particular physical instances of pre-recorded speech, as with concatenative synthesis, but being able to accurately estimate segments that do not appear in the recorded data. These techniques are therefore useful for building voices requiring minimal data input. An advantage over parametric synthesis is that there is little involvement required to manipulate the individual parameters to produce a reasonable output. Improvements to the voice quality made using STRAIGHT vocoding for HTS led to that system having success in the 2005 Blizzard Challenge [11] and performing well in subsequent evaluations with its average-voice-based adaptation techniques, particularly achieving equal highest scores for both intelligibility and naturalness in the 2008 challenge using one hour of UK English data [229]. There is access for manipulation of the prosodic parameters of the output speech. The structure of the system leads to the potential for the selective substitution of information from the average voice model to reconstruct voices that show the effects of dysarthria.

3.3.5 Voice conversion

3.3.5.1 Introduction

Voice conversion is the process of converting an utterance of an original speaker (the *source speaker*) so that it sounds like that which would be produced by another speaker (the *target speaker*). This is not usually considered to be a method for speech synthesis however it has potential applications for personalising a system. It could be used to convert a whole database of speech to a target speaker before use in a synthesis system as detailed above or embedded within systems. The conventional conversion processes involve using parallel sets of data, i.e. the same utterances spoken by both source and target speakers, and estimating a mapping between them. The advantage of this system is that to perform a conversion requires a fraction of the amount of data required to build a voice than build a new synthesis system from scratch. This section will look at the methods used for voice conversion and its possible applications in personalised speech synthesis.

3.3.5.2 Methods of conversion

Voice conversion usually requires the alignment of parallel corpora between a source and a target speaker. Features are extracted from both target and source utterances and the frames are time-aligned. This forms the basis for estimating a mapping of the acoustic space of the source to that of the target speaker. The mapping can then be applied to new, unseen data. The usual method of voice conversion involves a manipulation of the spectral features and some simple modification of the excitation and prosody.

The acoustic space of the speakers for voice conversion has been represented by various features, such as linear predictive coding (LPC) coefficients [192], formant frequencies [1, 163] or line spectral frequencies (LSF) [8, 104, 164, 169], which are closely related to formant frequencies and contain individual speaker characteristics [8, 101].

The mapping between the source and target acoustic space is done using discrete or continuous methods. Discrete methods use codebooks to represent the speech of both source and target speakers and estimate a mapping between the two [1, 8]. Having an inflexible one-to-one mapping between the source and target codebooks leads to discontinuities in the speech signal, leading to inconsistent output when the converted speech is used as a database for speech synthesis.

Continuous functions have also been used in voice conversion, including neural network models [163], Gaussian mixture model (GMM) [201] or Hidden Markov model (HMM) based conversions [164], reverse vocal tract length normalisation [63] and linear transforms [192, 104, 230].

There has been some research into ways of converting voices without using an explicit parallel database which allow voice transformation functions to be estimated based on minimal data from a source speaker. Ye and Young put forward an approach where the source and target's training data are indexed in a database and labelled at state level using an HMM-based recogniser [231]. The target data equivalent to the source data can then be compared based on this labelling, estimating the linear transform between them. A global transformation over the utterance can then be performed over all the data. In contrast, Kain and Macon's [104] approach takes the view that only transforming the segments where appropriate data is available was preferable for a speech synthesis system, not altering the source data feature vectors that do not have a corresponding target. This approach relies on finding a source speaker who is perceptually very similar to the target speaker. These methods require significantly less data and suggest that it is possible to make a successful transformation without having a full inventory of sounds produced by the source or target speakers. They do both conclude however that the more data that is available, the more successful the conversion.

Voice conversion can be done post-synthesis, taking concatenative synthesis voices and used a smaller database towards which they converted that voice [77]. This approach took advantage of using a large database to produce high quality concatenative synthesis voices and then personalising it to the required extent using a smaller database gathered from the target speaker. They concentrated on altering characteristics such as the prosody used rather than spectral manipulation. This approach relies on using a source speaker with similar speech to the target speaker. Such an approach is not a practical solution for a real time synthesiser as it requires a large amount of storage space for the concatenative synthesis unit database, extra computation for the conversion stage and the data from the target speaker to estimate the conversion needed to personalise the system.

3.3.5.3 Modification of dysarthric speech

For speakers whose voices have begun to deteriorate, another direction pursued is to make modifications to the dysarthric speech to result in an improved intelligibility output. The Speech Enhancer$^{\mathrm{TM}}$(Voicewave Technology Inc.), for example, filters and amplifies certain parts of the speech signal to enhance the speech of a dysarthric speaker. This has reduced impact when the speaker's impairment has very complex and variable articulation problems. Hosom et al. [91] set out a method to improve the understanding of the contribution of modifying various factors of dysarthric speech to the intelligibility of the output. Using the Nemours database of dysarthric speech [145] the combinations of modifications revealed that the results are highly speaker-dependent and that there is a lack of ability to generalise the intelligibility modifications over all speakers. This research was extended to take the parameters that were found to affect the intelligibility of dysarthric speech and modify them accordingly, outputting a resynthesised speech signal [105]. The main finding of this research was that although the removal of breathiness in dysarthric speech improved the perceived intelligibility, the signal processing added extra artefacts in the output therefore counteracting the attempted enhancement.

Voice conversion has been used to personalise speech output from a speaking aid for individuals who have had a laryngectomy [159, 160]. The aim was to produce a more natural-sounding output for the electrolarynx device, using voice conversion techniques between a source speaker and the speech of the post-laryngectomy patient with an electrolarynx as the target speaker. The results of this study showed that with this technique it was possible to increase naturalness although this personalisation reduced the intelligibility slightly. There were no explicit speaker similarity experiments reported although mel cepstral distortion measures showed that the produced speech was moving closer to the target.

3.3.5.4 Personalisation

Voice conversion offers an opportunity to provide a personalised target voice for a synthesiser. The difficulties in this approach are that there is generally a need for parallel databases of speech to accurately estimate a transformation between the source and target voices. This may be possible for voice banking although it will rely on a certain amount of data to be produced and also the more similar the voice is to the source speaker, the more successful the conversion will be. This is particularly true for those techniques where part of the source speaker data is used in the synthesis. Where the synthesis technique relies on unit selection concatenative synthesis, the problems of high storage requirements are again encountered which suggests that this method may be inappropriate for this task.

For those whose speech has begun to deteriorate, the target speech for conversion is not complete. The task becomes more difficult as the speech produced contains only a partial representation of the target. Being selective about the target data and then applying global transformations may be one way to deal with this problem although the quality of the conversion increases with the amount of data available.

For those individuals who have lost their voice completely, there are promising results for using voice conversion techniques to increase naturalness of output with an electrolarynx, although this is a very new research area and this type of data is unavailable for this thesis.

3.3.5.5 Summary

This section has reviewed voice conversion techniques. Voice conversion is a method of converting a source voice into a target voice. The usual method involves parallel corpora of data and transformations to be estimated between the two sets of data. This technique requires a fairly large amount of storage space and data. Attempts to provide conversion methods which require much smaller amounts of data and data which is non-parallel have relied on the source speaker having similar features to the target. The problem with applying conversion techniques to dysarthric speech is the lack of generalisability across speakers as the modifications required are highly speaker-dependent. For those modifications that have been applied, the impact of the conversion is reduced by signal degradation introduced by these manipulations.

3.4 Conclusions

The requirements for the task of personalising voices for speech synthesis can be split into two types of requirements: quality of output and practicalities of the task. Quality of output requires intelligibility, naturalness, similarity to the target speaker and manipulation of prosodic factors. Practicalities required for the task are using minimal data, providing possibilities for using deteriorated data and having an available resource with which to work. A summary of the suitability of these techniques in terms of the practical requirements is listed in table 3.1. The quality of output requirements are summarised in table 3.2. Articulatory synthesis and voice conversion are not listed in this table as they are generally used as input to parametric, concatenative or model-based systems for synthesis. Comparisons in the output table are made for concatenative and model-based synthesis taken from results from UK English voice B in the Blizzard challenge 2008 evaluation [106]. This voice is built with one hour of data, most closely matching the voice building requirements of minimal data that is attempted in this thesis.

From the review of synthesis methods, the most comprehensive coverage of the requirements for the problem of vocal personalisation of synthesised speech is to use model-based synthesis due to its ability to produce intelligible, natural-sounding voices that sound like the speaker. It allows easy access to the prosodic factors and requires comparatively small datasets of recordings particularly when using adaptation techniques. In addition to using this technique with unimpaired speech data, it has the potential to deal with dysarthric speech data. This is explored in more detail in chapter 4.

Method	Data required	Available tools?	Possibilities for dealing with deteriorating speech?
Articulatory	Detailed articulatory knowledge (X-ray/MRI etc.) for all sounds of the language.	No	Donor articulatory information
Parametric	Detailed articulatory, aerodynamic and temporal information for all sounds of the language.	Fonix, Klattools	Donor speaker parameters
Concatenative	80 mins phonetically balanced recorded speech.	FestVox	Donor units
Model-based	6-7 mins recorded speech.	HTS	Adaptation, data selection and using model information from average-voice.
Voice conversion	Parallel data sets of source and target speakers to create personalised dataset. Amount dependent on synthesis technique.	HTS + others	Global transformation.

Table 3.1: *Summary of the practical considerations for selecting a synthesis method for building a personalised voice*

Method	Intelligibility	Naturalness	Similarity to speaker	Prosodic manipulation
Parametric	High - no significant difference between high quality synthesiser and natural speech [119]	Low - robotic quality	Not personalised	Yes
Concatenative	Festvox baseline in Blizzard challenge: approx 40% WER for UK voice B (built with 1 hour of data - decreases with more data used)	High - Festvox baseline in Blizzard challenge: 3.1 MOS for UK voice B	High - Festvox baseline in Blizzard challenge: 3 (median) MOS for UK voice B	No access
Model-based	approx 28% WER for Blizzard UK voice B	High - 3.6 MOS for Blizzard UK voice B	High - 3 (median) MOS for UK voice B	Yes

Table 3.2: *Summary of the output requirements for building a personalised voice with each synthesis method*

Chapter 4

HTS - HMM-based synthesis

4.1 Introduction

The previous chapter set out an argument for the appropriateness of using hidden Markov model (HMM)-based synthesis for the task of both voice banking with non-deteriorated speech and building personalised synthetic voices for speech which is progressively deteriorating due to a motor speech disorder. This chapter describes HMM-based synthesis in more detail, specifically providing an overview of the HTS toolkit [234], software that provides a research tool for HMM-based synthesis. Information is provided on the discretisation of the waveform and the features used and the prosodic and phonetic elements taken into account for the models. The different methods used to construct a speaker-specific voice are discussed and an overview is provided of the speech synthesis procedure. The applicability and usefulness of HMM-based synthesis for dysarthric speakers is also discussed.

Note that, for the purposes of this chapter, it is assumed that the reader is familiar with the HMM statistical model and the related hidden semi-Markov model (HSMM) used in HMM-based speech synthesis. Appendix B.1 provides an informal introduction to this model for the less familiar reader or the reader is directed to the tutorial in [173].

4.2 HTS overview

Two strategies exist to construct models corresponding to a particular speaker. The first technique available in HTS is to create a speaker-dependent model using speech solely from that speaker, as described in section 4.2.1. The second method is to adapt an existing model using the speech of a particular speaker, as described in section 4.2.2.

4.2.1 Speaker dependent system

Figure 4.1 shows how speaker-dependent models are constructed using HTS. Initially, the input speech waveforms are segmented into utterance level segments. The continuous waveforms are then discretised into *feature vectors*: a compact, discrete representation of speech characterising the acoustics of the signal. Context-dependent labels are derived from the orthography of the utterances using the text analysis component of Festival [23] and used in combination with the extracted feature vectors to build the models.

Synthesis is performed by providing the labels corresponding to the text to be synthesised, which identifies the appropriate models which are then used to produce the speech waveform for that particular utterance.

Building a speaker-dependent model requires a large amount of data to fully capture the characteristics of an individual's speech. The Blizzard challenge annual evaluation of building synthesised voices using standardised Arctic data sets, approximately 80 minutes of speech, rated the speaker-dependent 2005 HTS system highest in a mean opinion score evaluation for naturalness and had the lowest human word error rate score representing high intelligibility [11, 236].

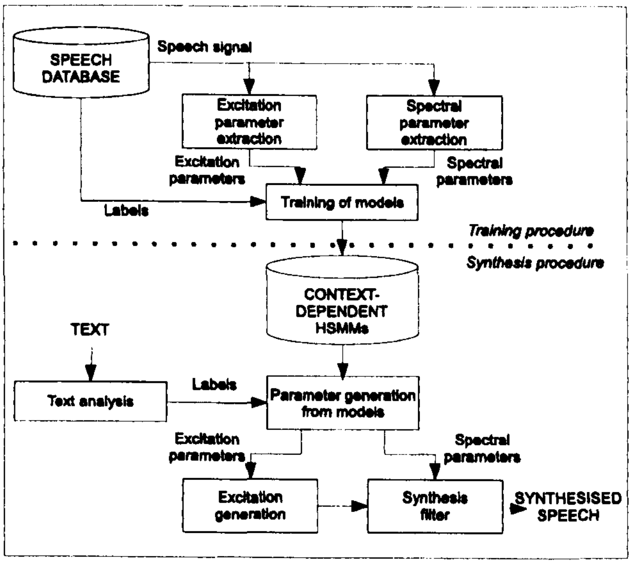


Figure 4.1: *An overview of the basic HMM-based speech synthesis speaker-dependent system (figure based on [235])*

4.2.2 Speaker adaptation system

Figure 4.2 shows the adapted average voice HTS system. The data is input in the same way as the speaker-dependent system but in this system, a speaker-independent model or *average voice* is built and the adaptation data, taken from one speaker, is used to adapt this model towards that speaker's characteristics, defined here as the *target speaker.*

Using this technique allows adapted models to produce synthesis based on the target speaker using much less data than is required for speaker-dependent models. In a comparative evaluation between the two techniques, using 100 sentences or approximately 6-7 minutes of speech data for the adaptation procedure was judged to be better than the speaker-dependent technique using models trained on between 30 and 60 minutes of speech taken from one speaker [227, 228].

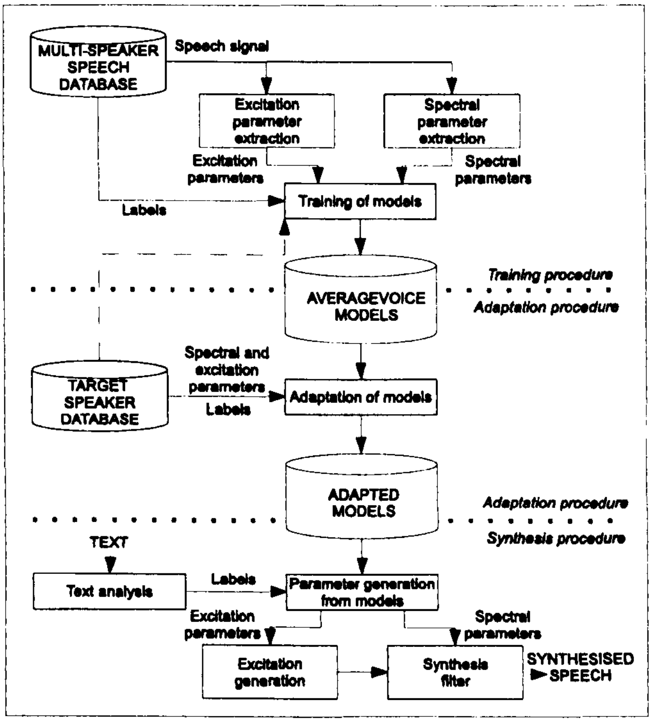


Figure 4.2: *An overview of HTS HMM-based speech synthesis speaker-independent average-voice adaptation system (figure based on [228])*

4.3 Feature vectors

To build the models, the data has to be discretised into feature vectors, defined as a compact, discrete representation of speech characterising the acoustics of the signal. In speech recognition, feature vectors designed for the task of sound discrimination are used in order to accurately transcribe or classify speech. This is usually a representation of the spectral acoustics without F0 information. In contrast to recognition, speech synthesis is not a classification task; the aim is to reconstruct the speech signal as accurately as possible to produce a natural-sounding output. For this reason, a more appropriate feature vector should be used for speech synthesis that captures more information about the speech signal.

The feature vectors comprise separate streams of spectral features including energy, log F0 and band aperiodicity.

To capture the spectral component of the individual's speech, 40 mel cepstra are used. This relatively high number of features in comparison to the usual 12 used in speech recognition means that more fine detail of the signal is captured which contributes to the natural-sounding percept of the reconstructed signal. The dynamic elements of the speech must also be captured, therefore the spectral stream of the feature vector is 120-dimensional, consisting of 40 mel cepstra (including energy), their deltas and delta-deltas [72].

The log F0 of the frame is also captured in the feature vector, with its deltas and delta-deltas which contributes to modelling the overall pitch of the signal.

The band aperiodicity component represents the relative energy of aperiodic components in the periodic signal in 5 different frequency bands. The deltas and delta-deltas of these values are also present in the feature vector. A more detailed discussion of the features used for HMM-based synthesis is deferred to appendix B.2.

For speech recognition, feature vector extraction usually occurs for frames of speech every 10 ms, which provides sufficient detail to characterise the speech for discrimination of sounds. For speech synthesis modelling, a higher temporal resolution is typically used and feature vectors are extracted from the speech every 5 ms with a window size of 25 ms. HTS simultaneously models these features to ensure that the alignment between the spectral features and the prosodic features remains consistent.

4.4 Model details: contexts and parameter sharing

The acoustic structure of a sound varies depending on its surrounding context due to the continuous movement of the articulators in the production of speech. For speech recognition, the unit modelled by the HMM is usually a *triphone*: a phone-sized unit which takes into account the previous and following phone. Speech recognition aims to discriminate between sounds to classify them correctly using the minimal information required to do so. Speech synthesis aims to reproduce the speech waveform as accurately as possible, retaining information which contributes to the naturalness of speech. For the speech synthesis task, richer contextual information is therefore used as it contributes to the generation of phonetic and prosodic elements of the output synthesised speech. In HTS contextual phonetic and prosodic information is provided at the phoneme, syllable, word, phrase and utterance levels. A full list of contextual information incorporated into these context-dependent phoneme models is given in appendix B.3. The text analysis method used to extract this information from an orthographic transcription is also detailed in appendix B.3.

Due to the number of contexts taken into account for each model, training a model for every possible context-dependent phoneme requires an impractically large amount of data. This data sparsity problem is addressed by sharing the parameters of the state output distribution between acoustically similar states. This sharing is performed using decision trees (see section 3.3.3.8) which define clusters of acoustically similar states.

The tree is constructed using questions based on the detailed phonetic and prosodic contextual labels. For example, these questions include "is the previous phoneme a vowel?" (L-vowel?) or "is the following syllable stressed?" (R-stressed?). The minimum description length (MDL) criterion [181] is used to determine both the structure and complexity of the decision tree. Once the tree is constructed the parameters are tied between the states which correspond to the same leaf node of the tree.

Different contextual factors (i.e. the questions) affect the acoustic distance between vectors for duration, spectral information, log F0 and aperiodicity and so these feature streams are clustered independently of each other. This means that there are separate models for each of these features, which are combined at synthesis time.

4.5 Model construction

4.5.1 Speaker dependent models

Figure 4.1 shows how speaker-dependent models are constructed using HTS. Monophone HSMMs are initialised using a uniform segmentation of the feature vectors to the model labels. Further training is performed using the embedded Expectation-Maximisation algorithm [52]. The monophone HSMMs are then extended to clustered context-dependent HSMMs and further re-estimated.

4.5.2 Speaker adaptation models

Using HSMMs for synthesis allows adaptation techniques to be used as an alternative to the speaker-dependent approach, see figure 4.2. These techniques are used in speech recognition to adapt the models to better represent the characteristics of an individual's speech with minimal data. In the same way, for synthesis it is possible to adapt *speaker-independent* models, trained on large amounts of data from multiple speakers, to more closely match a speaker's individual voice characteristics using minimal input data.

Further details on the construction of the average voice model and the subsequent adaptation techniques are given in appendix B.5.

4.6 Synthesis

The first stage of synthesis is to convert the orthographic text to be synthesised into a sequence of context-sensitive labels as described in appendix B.3. A composite HSMM is then created by concatenating the context-dependent models corresponding to this label sequence. A duration is then assigned to each state in the composite HSMM which maximises the likelihood of the state duration probability density.

The feature sequence of maximum conditional probability, given the input state sequence and models, including the probability distributions over the deltas and delta-deltas as well as those for the static features, is found using the feature generation algorithm [208]. This feature sequence is subsequently converted into a waveform using the STRAIGHT vocoder [110].

4.6.1 Global variance

The statistical nature of this technique results in spectral details being averaged out with high priority placed on producing a smooth output trace for each feature. In an attempt to improve the speech output and reduce this oversmoothing, refinements to the feature generation algorithm were introduced which model the utterance level variance (also called the *global variance*) of each stream. For each utterance in the adaptation data, for each set of features: mel cepstra, log F0 and aperiodicity, a variance is calculated. The mean of these variances and a variance of these variances is calculated across all the utterances in the data set. The global variance is integrated into the feature generation algorithm and ensures that the features generated more accurately reflect the utterance level variance of the data rather than oversmoothing the cepstral coefficients, log F0 and aperiodicity output. Introducing this technique has improved the quality of the synthesis [202, 226].

4.7 HMM-based synthesis and dysarthric speech

As mentioned in section 4.2.1, approximately 80 minutes of speech is sufficient to train a speaker-dependent model capable of generating natural, highly intelligible speech. For individuals who have difficulties with their speech production, it may be inappropriate and impractical to collect this amount of data. The speaker adaptation approach described in section 4.2.2 is a potential solution to this problem, as it avoids the need to collect such large volumes of speaker-specific data, while maintaining natural, intelligible speech.

For those speakers whose speech has begun to deteriorate, the approach may be suboptimal for the following reason. Where errors in production or misalignments between models and acoustics occur, HTS models the disfluencies in the data and therefore reconstructs them in the output. The following sections further detail the issues involved for using HTS with dysarthric data and attempt to provide a solution to reconstruct the voices of individuals, compensating for any impairment captured in adapting the models.

4.7.1 Misalignment between data and labels

For non-disordered data collection, it is assumed that an individual is able to be prompted with an utterance and produce the contents and the articulations accurately. This is not a trivial task for an individual with dysarthria. The speech may take longer to produce, it takes more effort and there will be frequent insertions, repetitions and misarticulations (see section 2.4). HTS labels the data based on an orthographic transcription of the original prompt from which the context-dependent phonetic and prosodic information is assigned. The labelled section of data is then assigned to its corresponding model. If the utterance produced does not accurately match what is expected then the resulting labelling will be inaccurate and introduce inaccuracies into the HSMM modelling.

In the adaptation stage, at each iteration of the process, an alignment between the data and current models is performed. If the alignment path score is too low, the utterance is rejected from the adaptation data. This is a problem with disordered data in that where insertions or deletions occur in the data, the whole utterance is rejected. Within that utterance there is likely to be some intelligible and therefore usable speech that is rejected from the adaptation process unnecessarily. Speech production is a difficult and effortful task for individuals with impaired speech and therefore a way of maximising the use of this data is sought.

A potential problem for HTS using dysarthric adaptation data is that there can be inappropriate or unpredictable pauses inserted into an utterance, between words or syllables within words. Where the pauses are not explicitly labelled as such in the orthographic transcription, they are assigned to a non-pause model, thus causing a mismatch between the models and the data.

Relabelling the speech depending on what speech is actually produced would solve this problem. This is not a trivial task for dysarthric speech, however. The insertions in dysarthric speech consist of both speech and non-speech sounds. Unintelligible speech or non-speech sounds do not coincide with a phonemic label. The labelling requires the context-dependent phoneme to be part of a syllable, word and utterance. If any of these levels of description are unavailable, decisions have to be made about how best to incorporate this into the label sequence. An advantage of HTS's approach to labelling is the lack of a requirement for human intervention for the labelling process. Any relabelling introduces a large amount of time-consuming expensive man-power into the procedure.

4.7.2 Proposed approach

A proposed approach to dealing with dysarthric speech with HTS is detailed below. The approach deals with maximising the use of the data and reducing the effects that dysarthria has on the voice and speech, as detailed in section 2.4. The two approaches to reducing the effects of dysarthria on the output synthesis are: data selection and feature selection.

4.7.2.1 Data selection

To solve the misalignment and labelling problem and maximise the amount of data that is available to be used for adaptation, intelligible sections of speech can be extracted from the recordings and associated with the corresponding sections of the full phonetic and prosodic context transcription. This removes the need for relabelling or decision-making on what inserted information to incorporate in the labelling process. Although not necessarily an accurate representation of what was actually produced by the speaker and the surrounding context, this method links the speech produced with the cognitive planning of what was intended to be said, as shown through the presence of anticipatory coarticulation in the data [107]. This approach reduces the problem of data rejection and allows a much higher percentage of data to be used for speakers with more severe dysarthria.

Selecting the data to match the labelling deals with the problem of anticipatory errors that can occur in some types of dysarthria. Matching the data to the labels maintains the correct sequence of labels to adapt the associated models with the appropriate data.

This data selection technique produces a fluent synthesised voice output, rather than one containing the disfluencies found in dysarthric speech, by using only intelligible segments of speech data. Removal of incorrect or unintelligible sections from the data disposes of 'bad' data that skews the modelling towards an inappropriate target. This process separates out those sections that are not well-articulated or where the timings of the articulators are not coordinated or controlled enough to produce an intelligible output. By selecting those sections of speech for adaptation that are intelligible to a listener, speaker characteristics are retained and intelligibility of the output is ensured. This procedure also removes unwanted unlabelled silences which can prevent accurate modelling of the data.

If the individual makes consistent errors in the articulation of specific segments and every example of a particular segment is removed from the adaptation data, the average voice model can provide an appropriate model for that segment based on the multiple speakers' data from which it is estimated in combination with the global adaptation algorithm.

Data selection can be done manually using a human listener to make a judgement on whether a section is intelligible. Manual data selection is an extremely time-consuming process. It is also inconsistent in that as a human becomes more exposed to the speech of an individual, they become more attuned to understanding it and therefore the acceptable intelligibility level changes [32]. Ideally, an automatic process would be used to replicate a naive human listener's selection for intelligibility, reducing the time taken to complete this process and select the data consistently.

Figure 4.3 demonstrates how the data can be selected for intelligibility. This figure shows a short section of an Arctic database sentence as spoken by a dysarthric speaker (speaker 5 in chapter 6): arctic a0251 "I may manage to freight a cargo back as well". The transcription panels show three levels of labelling. The panel nearest the waveform shows what different fragments are present in the speech files. They consist of pauses (labelled 'pau'), words or syllables (shown in forward slash delimiters '//'), non-vocal sounds (labelled 'noise') and vocal insertions (labelled 'vocal'). The central panel shows where the words occur in the phrase and the topmost panel shows which sections of the phrase are selected as being intelligible and therefore usable as adaptation data. This phrase is available on the sound file CD as example 4.1.

Due to the number of insertions in this phrase, without data selection, the whole utterance would be rejected by the first iteration of the adaptation procedure. Using data selection, at least two intelligible words could be used that would otherwise be discarded and are guaranteed by the selection process to provide reliable estimates of those corresponding context-dependent phonemes.

This example shows the complexity of trying to automate the data selection procedure. Detecting non-speech noise including silences and extraneous noise may be possible but there is a high occurrence of vocalised noise. The vocalised insertions in the dysarthric output have speech-like characteristics which makes it more difficult to automatically discriminate them from the speech that is to be retained in the adaptation data. Automation of data selection is not a trivial task and therefore for the experiments in chapter 6, the data was manually selected.

4.7.2.2 Feature selection

If there are errors in a dysarthric individual's speech, it would be useful to only use for adaptation those features that are not affected by the disorder. The remaining affected features would not be used as target speech for adaptation but the corresponding features in the starting point average voice model would be retained. The structure of HTS allows an approximation to this behaviour. The feature vectors are extracted and used to adapt the HSMMs simultaneously, but the spectral, log F0, aperiodicity and duration features are represented in separate streams and re-combined only when generating the synthesised speech. Therefore post-adaptation, certain features of the speech can be substituted with

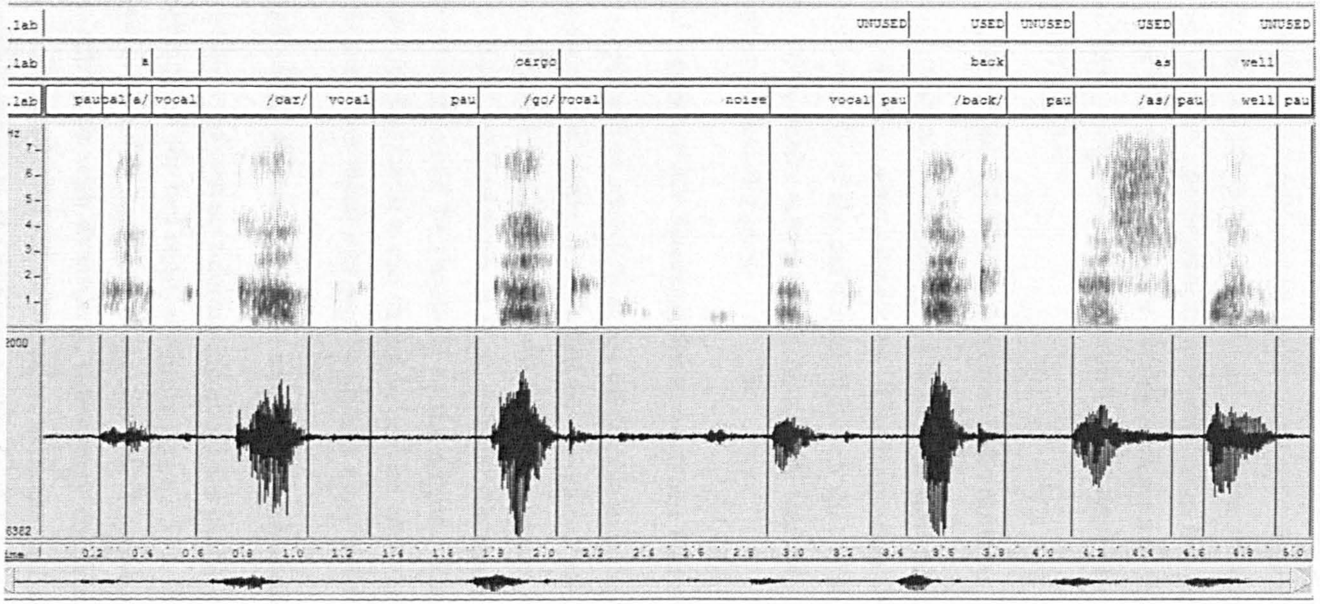


Figure 4.3: *The phrase "a cargo back as well" as spoken by a dysarthric speaker. It is labelled to show which sections of the phrase are usable as adaptation data (USED/UNUSED). The central pane indicates the word boundaries. The third pane shows the type of segment: pauses (pau), words or syllables (shown between //), non-vocal sounds (noise) and vocal insertions (vocal).*

those of the original average voice and used to reconstruct those features in the dysarthric speaker model which have been affected by the individual's condition. Figure 4.4 shows the type of substitutions that can be made using information from the average voice and the dysarthric participant speaker model to create an output speaker model. Features that capture the speaker characteristics are taken from the participant speaker model and information from the average speaker model reconstructs those features affected by the speaker's condition. Different feature substitutions depend on the individual's condition and stage of deterioration.

Use of the participant speaker's spectral information

As noted above, most of the inaccurate articulation that occurs in the adaptation data is removed during the data selection process. The adaptation data that remains is therefore a robust characterisation of the participant's speech that is not affected by the dysarthria. Retaining the participant's spectral information in this way allows the retention of the individual's speaker characteristics.

Use of global variance for spectral features from the average voice model

The actual articulations of sounds that are intelligible may be highly variable in dysarthric data. This spectral variability is modelled by the global variance parameter (see section 4.6.1), which influences the utterance level spectral variance during the parameter generation process. Where this variance is high, as could be the case for dysarthric data, constraining this measure could be beneficial to the output. The average voice model global variance for the spectral features can therefore be used with the participant speaker spectral features to constrain the variability and produce a more well-defined spectral output.

Use of energy information from the average voice model

The feature vectors used contain information about the overall energy of the speech frame as the zeroth coefficient of the mel cepstral feature vector. This feature may be highly variable due to the speech disorder, as stated in section 2.4. To solve the problems in reproducing energy of the individual with dysarthria, this energy coefficient can be selected from the average voice model and used in combination with the mel cepstral coefficients from the participant speaker model in the output speaker model. This smooths the output if there

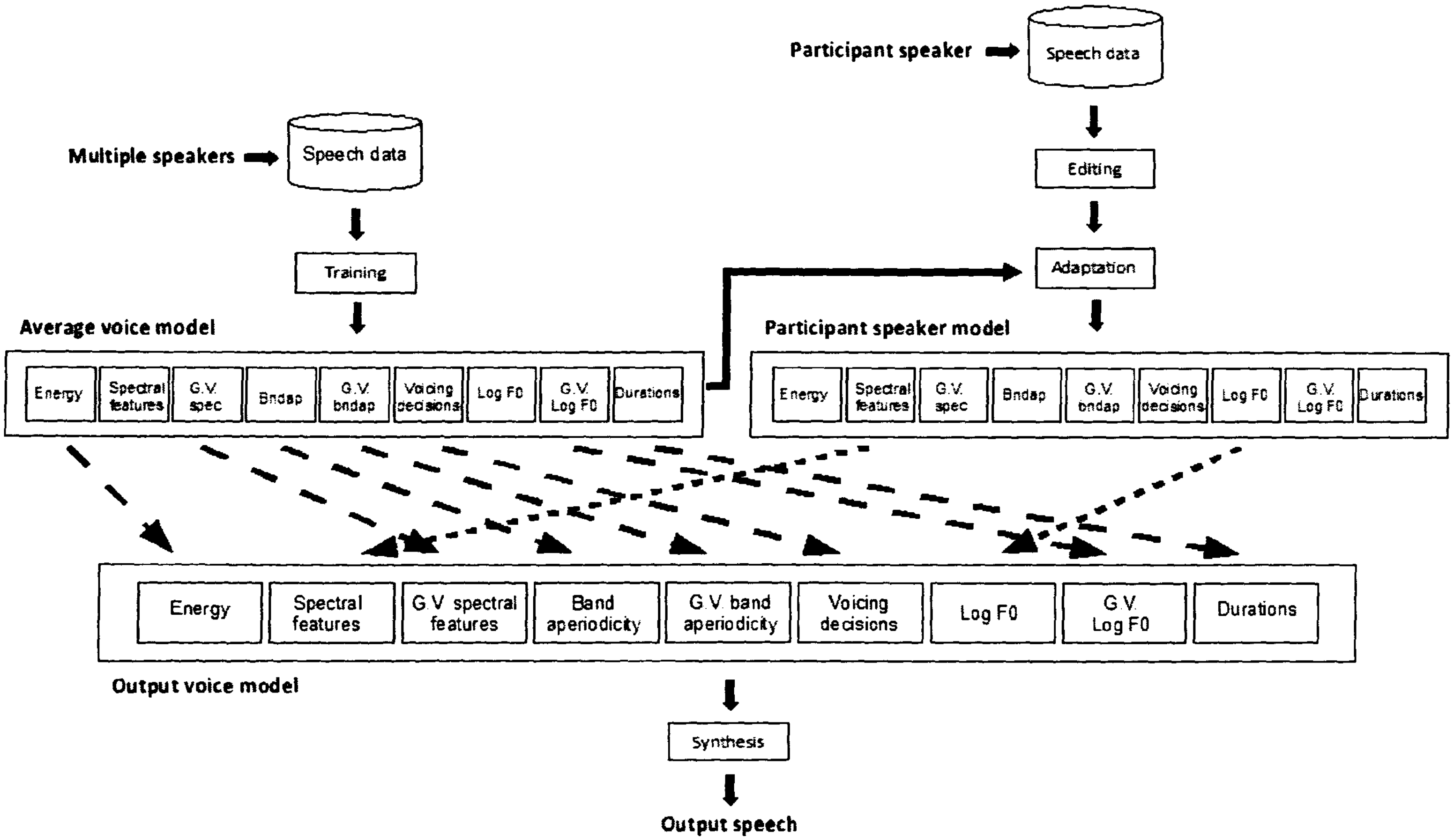


Figure 4.4: *Possible component feature selection to produce an output speaker model with speaker characteristics taken from the speaker model and components taken from the average voice model which compensate for the effects of the individual's condition. Bndap is band aperiodicity and G.V. is global variance.*

is much variation in the energy in the original speech and produces a more appropriate speaker energy if the speaker's voice has either reduced or elevated energy levels.

Use of the participant speaker's F0 information

The F0 of the speaker should be used in the output speaker models, if it has not been adversely affected by the condition, as it contains information specific to the speaker and contributes to the recognition of the voice as belonging to that particular individual. This will be shown in chapter 5 where F0 contributes significantly to listener responses of similarity between synthesised and target speech.

Use of voicing decisions from the average voice model

Where there are phonatory irregularity problems such as abnormal production of voicing, voicing initiation and reduced control of the vocal folds, voicing decisions can be isolated from the average voice log F0 model and used in the output speaker log F0 model. A voice would need to be consistently producing an incorrect voice decision to change the voiced/voiceless weighting sufficiently, but this possibility may occur with dysarthric speech.

Use of global variance for log F0 from the average voice model or alteration of values to match preferences of the speaker

Where the speaker has either a monopitch or highly variable prosodic quality due to the condition, the global variance of the log F0 can be altered to make the pitch range more appropriate. This can be done either by changing the mean of the global variance to that of the average voice or altering it to an amount which seems appropriate for that speaker. This parameter can be customised to suit the preference of the speaker and how this alteration affects the intelligibility and naturalness of the synthesised speech.

Use of aperiodicity information from the average voice model

The individual with dysarthria may have altered voice quality caused by reduced control of the larynx and weakened or tightened vocal folds. This causes an abnormal setting of the vocal folds, either causing excessive breath through the glottis or having to force the air through the constricted glottal area, in either case producing unwanted turbulent noise in the signal. Substitution of the aperiodicity models from the average voice alters the voice quality effect to match that of the average voice. It could also contribute to the minimisation of spirantisation effects found in the closure period of stop consonants.

Use of global variance for aperiodicity from the average voice model

Using the average voice global variance of the aperiodicity may also help to constrain the potential inflated variability of the aperiodicity in the speaker models caused by the individual's condition.

Use of duration information from the average voice model

For dysarthric speakers, the duration of segments is highly variable and often disordered, causing distortions in the rhythm and intonation of the output, therefore contributing to the lack of intelligibility in the speech. This problem is partly dealt with in the data selection process for adaptation but this selection process will not remove the variability that occurs when the speech is of varying speeds but well-articulated. By using the average voice model duration probability distributions, a consistent and reliable estimate of the duration of the segments will be produced.

Alteration of speech rate to match the preferences of the speaker

To make the output synthesis more appropriate and preferable for the user, the speech rate can be altered during synthesis, using the average voice model relative durations as a starting point.

Summary

Table 4.1 summarises which aspects found in dysarthric speech can be solved by data selection and substitution of average voice model information into the participant speaker model to produce an acceptable output speaker model.

4.8 Conclusion

HMM-based synthesis is a promising technique to use to build personalised voices for individuals. It can produce intelligible and comparatively, to other synthesis systems, natural-sounding output and its requirements for data input are significantly smaller than for any

Problem	Solution
Maximising use of data available for adaptation	Data selection
Articulation problems	Data selection
Highly variable articulation accuracy	Data selection and use average voice global variance for spectral features
Highly variable or intensity decay	Use average voice energy
Laryngeal voice onset problems	Use average voice voicing decisions
Incorrect voicing in segments	Use average voice voicing decisions
Reduced F0 range	Use average voice or altered global variance for log F0
Altered voice quality	Use average voice aperiodicity
Highly variable or inappropriate segment duration	Use average voice durations
Highly variable or inappropriate speech rate	Use average voice durations and alter output rate

Table 4.1: *Proposed solutions for reconstructing voices showing dysarthric features.*

other synthesis technique. For speakers wanting to bank their voice to personalise a speech synthesiser, this technique seems to be appropriate.

The technique can be altered to account for dysarthric speech, identifying the well articulated parts of speech and using those for adaptation data. To reconstruct the personalised voices where dysarthria has affected various features of the speech, HTS's use of separate streams of spectral information, log F0 and aperiodicity plus a separate duration model and access to the voicing decisions, energy component and global variance measures, there is a possibility of selectively substituting models taken from another voice which may be able to compensate for impairment captured by adapting the models using disordered data.

Chapter 5

Voice banking using HMM-based synthesis for speech data pre-deterioration

5.1 Introduction

Previous chapters have motivated the use of HMM-based synthesis as an appropriate technique for voice banking for speakers with non-deteriorated speech and speakers whose speech has started to deteriorate. This chapter describes the use of HTS for voice banking and establishes the upper limit of performance by using a large amount of professional speaker data to build a personalised voice. It reports an evaluation of the voices built using human listener responses and measures the amount of data needed to provide a voice resembling a target speaker.

To investigate which acoustic features are used by human listeners to make judgements of similarity between the synthesised speech and the target speaker, this chapter also describes an experiment to train a neural network to replicate listener responses. The results for this experiment are analysed to determine which acoustic features sets contributed most to making these judgements. The experiments in this chapter have been partially reported in [47].

5.2 Background

As discussed in chapter 2, the task of building personalised voices for individuals for use with VOCAs should be applicable to both those individuals who are able to bank voice recordings pre-deterioration and for those individuals whose speech has already started to deteriorate. The task for those whose speech is still intelligible can be simulated by building a voice for an individual who has no speech impairment. This provides insight into how much data would be required for voice banking and whether this approach is fit for purpose. This information can then be used to inform subsequent experiments as to the suitability of this approach for dysarthric speech data and how much of this type of data is required to build an acceptable personalised voice.

Previous experiments have shown that to build a speaker-dependent voice using HTS, approximately 80 minutes or 1200 sentences of speech are required [11, 236]. To build a voice using the adaptation procedure with HTS requires a significantly smaller amount of data from the target speaker, approximately 100 sentences or 6-7 minutes of speech data [223, 228].

In building personalised voices for individuals to use with VOCAs, it has already been established that the voices should not only be intelligible and natural-sounding but that they are also accurate representations of the individual using the voice. This evaluation concentrates on whether it is possible to fully capture the characteristics of an individual's speech with this amount of data and this experimental set-up using listening tests with human judgement responses.

5.3 Evaluation

The following describes experiments to determine how accurately this method is able to build personalised synthetic voices with non-deteriorated data. The hypothesis tests whether using 100 sentences of speaker data is an appropriate amount of data to capture the speaker characteristics of an individual as shown in [223, 228] and whether this result is applicable to other speakers.

5.3.1 Stimuli

5.3.1.2 Participant speakers

The participant speakers for this evaluation were speaker 1 and speaker 2. Speaker 1 was a professional broadcaster from Barnsley, South Yorkshire, UK, male, aged 51 at the time of recording. His voice is distinctive with a definite accent indicative of the Barnsley area and well-known through his broadcasting work. Speaker 2 was a university lecturer from north-west England, aged 32 at the time of recording. He has a Northern British English accent and was well-known to some of the participant listeners in the experiment. Examples of the speech of the participant speakers are available on the CD as examples 5.1 and 5.2.

5.3.1.3 Data collection

Recordings were made in a single-walled Industrial Acoustics Company (IAC) acoustically isolated chamber, using a Bruel and Kjaer (B&K) type 4190 0.5 inch microphone located approximately 30 cm in front of the speaker. The signal was pre-amplified using a B&K Nexus model 2690 conditioning amplifier prior to digitisation at 16 kHz using a Tucker Davis Technologies System 3 RP2.1 processor.

The recorded data were sentences taken from the Arctic dataset A [117]. This set of utterances is used in the Blizzard challenge voice building evaluations and consists of 593 sentences taken from a set of out-of-copyright books in English. The sentences are between 5 and 15 words in length to ensure ease of readability. The set covers all diphones of US English as it was originally designed to use for building voices using concatenative synthesis. The participants completed the recordings in one sitting, although they took frequent breaks. The prompts of the Arctic sentences were read from a computer screen, presented one at a time to maintain consistent recording conditions. The speakers were asked to produce the sentences as naturally as possible. The process involved the participants self-monitoring their reading and they were asked to re-record a sentence if they noticed an error in their production.

5.3.1.4 Test sentences

A test set of sentences comprising 25 sentences that had 8 or fewer words, with no embedded clauses, no complicated words or words with ambiguous pronunciation were selected from the Arctic dataset A. This ensures that the sentences are easy to listen to, reducing the complexity of the experimental procedure for the listeners. No sentences were included where the quality of output could be affected by errors in the text analysis stage of the synthesis. For example:

"I only read the quotations." (arctic a0243)

was excluded because the pronunciation of "read" is ambiguous.

"But Johannes could, and did." (arctic a0534) was also excluded because of its complicated phrasing and the potential for mismatch between the pronunciation of the proper noun by the speaker and the pronunciation dictionary. Selecting the shortest sentences was an attempt to ensure that the time required to complete an evaluation was minimised. Sentences in the test set were not used as adaptation data in any of the voices built. This allows the voice output to be evaluated for quality in output production rather than focussing on how well the adaptation procedure works with seen data.

Once selected, the test set was checked to ensure that it had a broad spread of phoneme coverage. The test set sentences are available in appendix C.

5.3.1.5 Voices

The evaluated voices were built using HTS version 2.1 (internal) using a 138-dimensional feature vector containing: 40 STRAIGHT mel cepstral coefficients (including the zeroth coefficient representing energy), deltas and delta-deltas; log F0, its delta and delta-delta; 5 band aperiodicity values, deltas and delta-deltas (see section 4.3).

Voices were built using the adaptation procedure as detailed in section 4.2.2. The average voice consisted of full Arctic data sets (1132 sentences) as spoken by 6 male speakers: 4 US English speakers, 1 Canadian English speaker and 1 Scottish English speaker. Examples of the average voice are available on the CD as examples 5.3 and 5.4.

Mel cepstral coefficients, log F0 and aperiodicity features were extracted using the following parameters: 25 ms window, 5 ms frame shift, F0 minimum 70 Hz and F0 maximum 280 Hz for participant speaker 1 and F0 minimum 70 Hz and F0 maximum 200 Hz for participant speaker 2. These values were derived by visually inspecting the F0 trace with the F0 minimum and maximum parameters set to a wider range and altering these values to more accurately represent the speaker's F0 range. This procedure minimises any pitch extraction errors (see section B.2.2).

The stimuli were presented using one speaker per experiment. The participants were asked to rate each stimuli for similarity to the target speech (see section 5.3.3 for more detail). Each listener heard a total of 375 stimuli: 3 presentations of 25 sentences in 5 conditions (average voice, voices adapted with 10, 100 and 500 sentences and the resynthesised original). The participant speaker voices were built with the first 10, 100 and 500 sentences from the data set (all distinct from the test set). 500 sentences is the majority of the recorded data set with a removed test set. 100 sentences is referred to in the literature as an appropriate amount of data with which to produce an HTS adapted voice. 10 sentences produces a voice that retains a certain amount of average voice qualities but provides an idea of the success of this technique using a very small amount of adaptation data. Having examples of the speaker's original speech and the average voice in the test set ensured that the participant was continually reminded of the extremes of the rating scale and allowed a continual calibration of their rating throughout the experiment. The target speakers' original speech stimuli were resynthesised, where the features were extracted and then directly resynthesised. This was to ensure that the listeners did not use the different sound quality of the stimuli as a cue for identifying the original recordings. Examples of the stimuli for both speakers are available as examples 5.3-5.7 for speaker 1 and 5.8-5.12 for speaker 2.

The stimuli were all of equal loudness with no unnecessary silence at the beginning and end of the stimuli sentences, to keep the length of the experiment to a minimum and therefore make it easier for the listener to maintain a reasonable level of concentration and interest to provide accurate ratings.

5.3.2 Participants

The participant listeners for the evaluation were native British English speakers between the ages of 23-48. The participants reported no hearing, language or speech difficulties. For the first experiment (using participant speaker 1 as the target speaker) there were 7 participants (6 male, 1 female) and for the second (with participant speaker 2 as the target speaker) there were 10 participants (8 male, 2 female). There was some overlap between the participants for both experiments, although the two experiments took place with an interval of two months. All subjects were students or employees of the University of Sheffield and were not paid for their participation.

5.3.3 Procedure

The experiments took place in a single-walled IAC acoustically isolated chamber and the stimuli were presented over Sennheiser HD 515 headphones. The participants were initially presented with a set of four examples of the original speech from the target speaker, randomly chosen from the recorded set but different from the test set. Having distinct sentences from the test set to use as reference sentences ensured that the participants were not just making a direct comparison between segments of speech in each example and instead focussing on the speaker characteristics and overall impression of the voice. These sentences were not resynthesised. The participants were asked to listen to all four examples initially and were told that this was the target speaker to which they would be comparing the experimental stimuli. They could listen to each reference example an unlimited number of times and at any point throughout the experiment if they so wished. The participant was asked to make sure that they had listened to the whole of the sentence before rating it and was encouraged to move through the stimuli quickly and not worry about the correctness of their rating.

Once the experiment had begun, the participant was presented with a stimulus and asked to rate it on a Likert-type scale of 1-7 for similarity where 1 meant 'it sounds like a totally different speaker' to 7 'it sounds like the same speaker'. This phrasing was chosen over 'how similar is this voice to the reference voice?' as this allows the listener to have a clearer idea of what 'similarity' means. This type of scale was used in the Blizzard Challenge 2007 [71] and 2008 [106] where approximately 73% (2007) and 71% (2008) of the participants said that they found the task easy to do. A scale of 1-7 was chosen so that the listeners would be more discriminating than a 1-5 scale in their evaluation. The rating scale was visible at all times to the participant with the 1 button labelled 'different speaker' and the 7 button labelled 'same speaker' to ensure that the individual was always aware of the correct order of the scale. Participants during the speaker 1 experiment could use either the mouse or the number pad on the keyboard as this was felt to be quicker by the participants. Due to some technical changes for the speaker 2 experiment, only the mouse was to be used to click on the appropriate rating button.

The stimuli were presented randomly for each participant listener and they could only listen once to the stimulus they were rating. The next presentation began immediately once a rating had been given. The stimuli were split into three blocks with opportunity to pause and rest after each section if the participants so wished. The overall length of the experiment was approximately 30 minutes for each participant.

5.3.4 Results

Results for each condition for both speakers are shown in the boxplots in figures 5.1 and 5.2. These boxplots show the median value and interquartile ranges comprising the middle 50% of responses to the stimuli, as shown by the central box. The full set of ratings for both experiments are shown in figures 5.3 and 5.4.

Increasing the amount of adaptation data used to build the voices showed an increase in the similarity rating to the target speaker for both sets of speaker data. The highest median rating for the conditions using adaptation data was 5 for the 500 sentence voice for speaker 1 and 4 for speaker 2. Speaker 1's increase is greater than that of speaker 2 using the same amount of data. The voices built using 100 sentences of data showed a wider range of responses than the other voices built across both speakers.

Both experiments show that the average voice is consistently rated as sounding like a different speaker (rating 1). The target speech was rated as sounding like the same speaker (rating 7) for both speaker experiments, although it was more consistently rated as that for speaker 1 (~89%) than for speaker 2 (~57%).

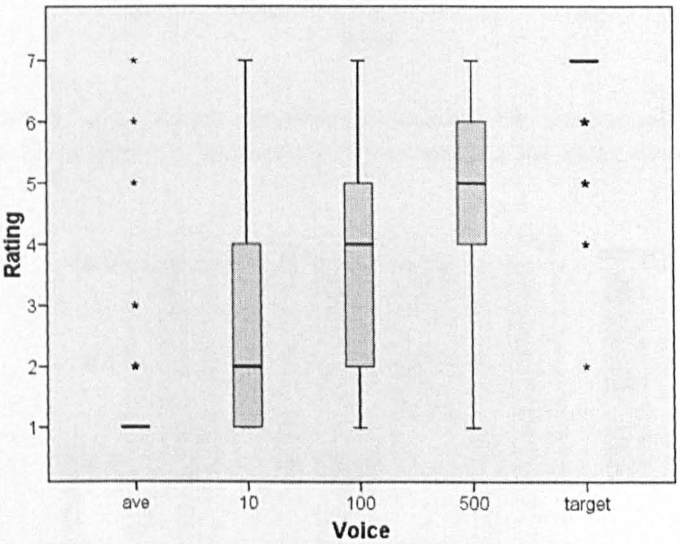


Figure 5.1: *Results of the listener experiments showing the median rating for speaker 1 voices. Rating 1 (sounds like a different speaker) to 7 (sounds like the same speaker) is used.*

The probability of listeners rating a synthetic voice as more similar to the target increases with the amount of data used to adapt the voice. The responses of listeners were judged to be significantly different for the voices under test for both speakers using Friedman's ANOVA (Speaker 1: $\chi^2(4) = 27.47, p < 0.001$; speaker 2: $\chi^2(4) = 38.73, p < 0.001$). Wilcoxon signed rank test results for a comparison of the different conditions are shown in

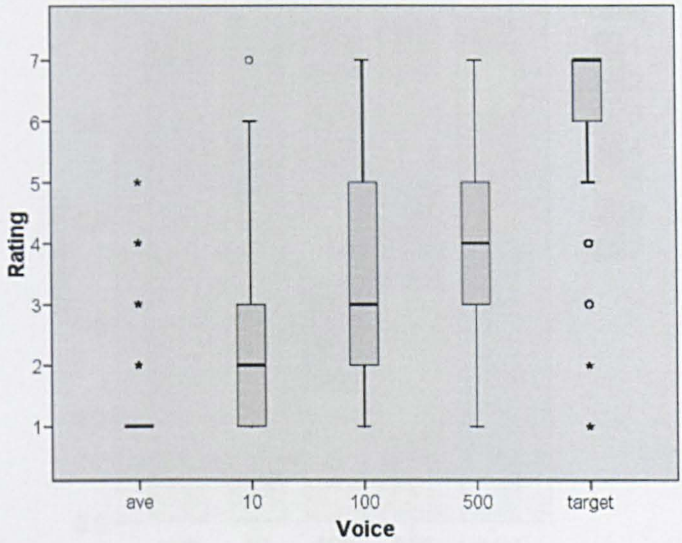


Figure 5.2: *Results of the listener experiments showing the median rating for speaker 2. Rating 1 (sounds like a different speaker) to 7 (sounds like the same speaker) is used.*

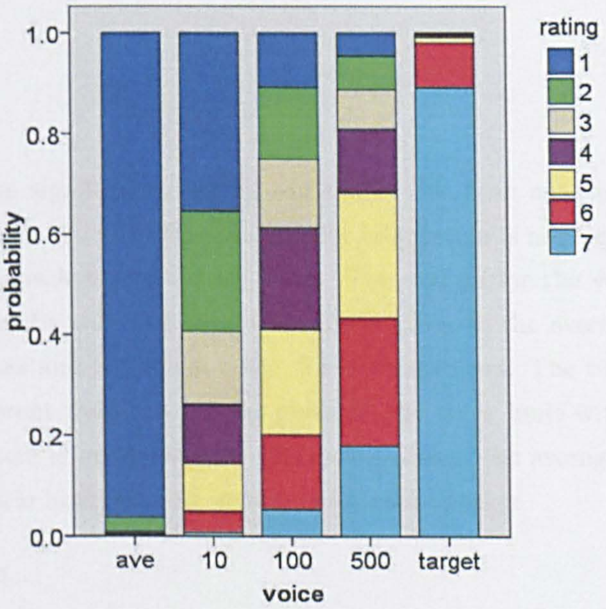


Figure 5.3: *Results of the listener experiments for speaker 1. Rating 1 (sounds like a different speaker) to 7 (sounds like the same speaker) is used.*

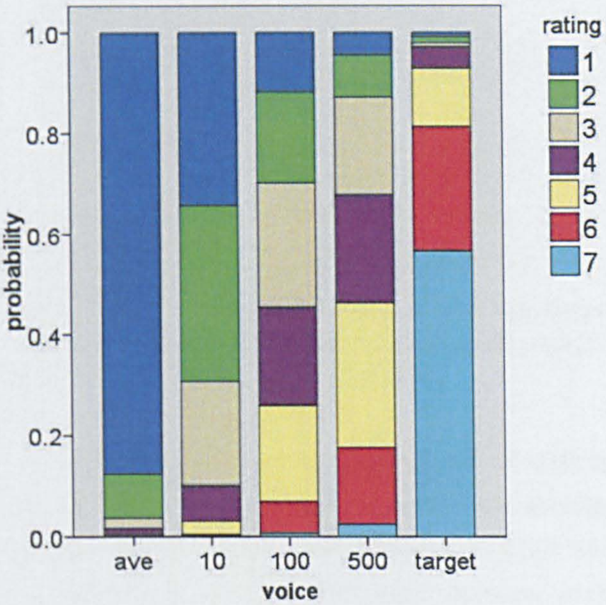


Figure 5.4: *Results of the listener experiments for speaker 2. Rating 1 (sounds like a different speaker) to 7 (sounds like the same speaker) is used.*

table 5.1.

The results of the significance testing are similar for both speakers except that the difference between using 100 and 500 sentences for adaptation is not significant for speaker 1 but significant for speaker 2 at the 5% level. The ratings for the voice built with 100 sentences are significantly different from the ratings given to the average voice, the voice built with 10 sentences and the target voice, for both speakers. The target speech ratings are significantly different from the ratings given to the voice built with 500 sentences of adaptation data. There is no significant difference between the average voice ratings and those given to the voice built with 10 sentences for each speaker.

5.3.5 Discussion

The results of these experiments support previous work that, with this technique, 100 sentences of adaptation data is the minimum required to produce a synthesised voice that

	Significance	
Comparison	Speaker 1	Speaker 2
ave-10	0.063	0.016
ave-100	**0.008**	**0.001**
10-100	**0.008**	**0.002**
100-500	0.016	**0.008**
500-target	**0.008**	**0.001**
100-target	**0.008**	**0.001**

Table 5.1: *Results for comparisons of the different voices for significance. Bonferroni corrections applied for 6 comparisons means that the difference is significant at the 5% level where* $p < 0.00833$. *Significant conditions are in bold face.*

is distinct from the average voice with some similarity to the target speaker. The results also suggest that this is dependent on the speaker data and the average voice used.

The results suggest that using 10 sentences for adaptation is not sufficient to produce a voice distinct from the average voice, but the information contained in the additional 90 sentences of data in the voice built with 100 sentences captures enough speaker characteristics to produce a significantly different voice.

For speaker 1, the difference between the voice built with 100 sentences and that built with 500 sentences is not significant. This result suggests that the additional 400 sentences of adaptation data used has no further significant similarity information to the target speech as perceived by the listeners. This result is different from that of speaker 2 that shows a significant increase with using the increased amount of data. This result reveals that this increase in the amount of data used for adaptation does not guarantee a significant increase in the similarity to the target speaker. It also shows that the amount of data required for adaptation to achieve a voice high in similarity to a target is speaker-dependent.

For both speakers, the significant differences between the target voice and the conditions using 500 sentences for adaptation data suggest that even with this large amount of adaptation data, the voices do not capture sufficient similarity to the target speech to be recognisable as the speaker. This result can be attributed to the average voice used in this experiment. The majority of the speakers that contributed to the average voice database were North American. Any influence of that accent on the voices built introduces more distance from the target speaker. The North American average voice used for these experiments was pre-built for the software demonstration scripts and used accurate hand-corrected labels. At the time of starting the experiments, there was no pre-built British

English equivalent and to build a British English average voice of the same high quality would have been too time-consuming to build for the time-scale of the thesis.

Using a more appropriate average voice is particularly relevant when the amount of data used to build the voices is small. The adaptation procedure adapts globally using CSMAPLR and adapts those observed models with sufficient data using MAP (see appendix B.5.1). This means that where there is enough information to make a robust estimate of certain models, these sections of speech match closely to that of the target speaker. Where the other sections are transformed globally or not at all, they retain more characteristics of the average voice model. Where the average voice model has different characteristics from the target speaker, the synthesised voice is not perceived as being as similar to the target. Using a more appropriate average voice for the target speakers' voices is likely to improve the similarity judgement to the target speaker and reduce the amount of data needed to produce a voice with high similarity. It is also likely to make the judgements show less clear distinctions between the voices. Having an average voice that is not so distinct from the target speakers may not introduce its own characteristics which has to compete with those of the target speakers.

Speaker 1's voice is more distinctive and therefore more recognisable than that of speaker 2, which could account for the difference in results between the two speakers. Speaker 2 was also known to some of the participants and some commented on the difference between the read reference sentences and their recollection of the individual's voice when speaking spontaneously. This lack of consistency could account for the difference between the ratings in similarity between the target voice ratings and the reference speech for each speaker. Speaker 1's professional speaking experience means that he sounds more natural when reading prompts and more consistently like the voice that participants may have been exposed to on television and radio.

5.4 Acoustic features affecting listener judgements

5.4.1 Introduction

The results of the above experiment have shown that as more adaptation data is used, the models built show a closer representation of the speech of the target speaker. The following section details an experiment to investigate what features are perceptually relevant to identify a speaker's voice and increase the assigned similarity rating. Understanding more about which acoustic features listeners are using to make their similarity judgements better informs any attempt to build an objective measure of similarity, which is also discussed below.

If human listener judgements can be replicated using an objective measure it could provide empirical evidence on which to base minor decisions or parameter changes without having to perform expensive and time-consuming human listener experiments at every stage of the voice building process. As discussed in section 3.2, objective measures are currently used for each feature separately (F0, durations and spectral features) to evaluate the performance of a voice conversion technique. Human listeners evaluate similarity perceptually using these features in combination but there is currently no individual objective measure for speaker similarity that can replicate this behaviour.

The following sections detail an experiment which attempts to investigate whether it is possible to build an objective measure which assesses the similarity of a synthesised sentence to a target sentence incorporating the acoustic features combined together. The experiment uses a neural network, more specifically a *multi-layer perceptron (MLP)*, trained with a set of acoustic features to predict human similarity judgements using the same set of test sentences as used in the above experiment. MLPs are used as they can take multiple features as input and combine them together to be modelled non-linearly. The non-linear combined modelling reflects more closely the perception of a human listener of the similarity of speech samples. Acoustic features were extracted from the test set, comparing feature values for each target speaker utterance T and corresponding synthesised approximation S for each condition: average voice and 10, 100 and 500 sentences of adaptation data.

The first experiment investigated the abilities of an MLP to predict the listener judgements as reported for the above experiments using a speaker-dependent approach where the network was trained with and tested on data taken from the same speaker. The MLP was then trained using different sets of input acoustic features to determine which set of features best replicate the listener judgements and therefore which features contributed most to the perception of similarity.

The speaker-specific MLP was then used to predict the responses for the other speaker's data. This aim of this experiment was to determine whether an MLP had the potential to be used as an objective measure of similarity for new speakers. Further to this, a speaker-independent MLP was then trained using data taken from both speakers and tested on unseen data from both speakers. This more closely replicates the process that would take place to build an objective measure, where multiple speakers' data would contribute to the speaker-independent MLP and then used on new speakers. This is not possible to do with the amount of data available at this time but is the next step for further work on this topic.

5.4.2 Multi-layer perceptrons

Multi-layer perceptrons are types of artificial neural networks, which are models consisting of a group of interconnected artificial neuron units. The neural network system is a framework for representing non-linear mappings from a set of input variables to particular output variables [16]. This mapping is modelled using a set of mathematical functions containing adjustable parameters or weights. The values for these weights are derived and optimised using a training procedure for the provided data set. In *supervised training* the target output values for each set of input values are provided and the network learns an input-output mapping which minimises the error between the target outputs and the actual outputs.

For this experiment, the input variables were perceptually relevant acoustic features extracted using the PRAAT toolkit [24]. The output variables were the similarity scale ratings (1-7) as described in the listener judgement experiment above. The MLP configuration is illustrated in figure 5.5.

5.4.3 Feature extraction

The first stage of the experiment was to extract the acoustic features from the data. The features considered were the 18 spectral and prosodic features in table 5.2. Kominek, Schultz and Black [118] have used mel cepstral distortion to calibrate synthesised voice quality and Yamagishi and Kobayashi [223] have advocated this feature for evaluation of voice conversion together with F0 error and vowel duration error.

The mel cepstral distortion (feature 1) was obtained from the global cost of an asymmetric Dynamic Time Warping (DTW) alignment between 12 cepstral coefficients for T and S, excluding overall energy, scaled and normalised across the coefficients. The cepstral coefficients were extracted from endpointed files using the Rastamat toolkit [64] using a 10 ms frame shift and a 25 ms size window.

The remaining features for each T and S were extracted and compared by averaging point-by-point absolute differences between the target and synthesised frames along the DTW alignment path. Features 5 to 18 were measured only in sections where both T and S were voiced. Feature 4 is the fraction of frames in T and S which are both voiced or both

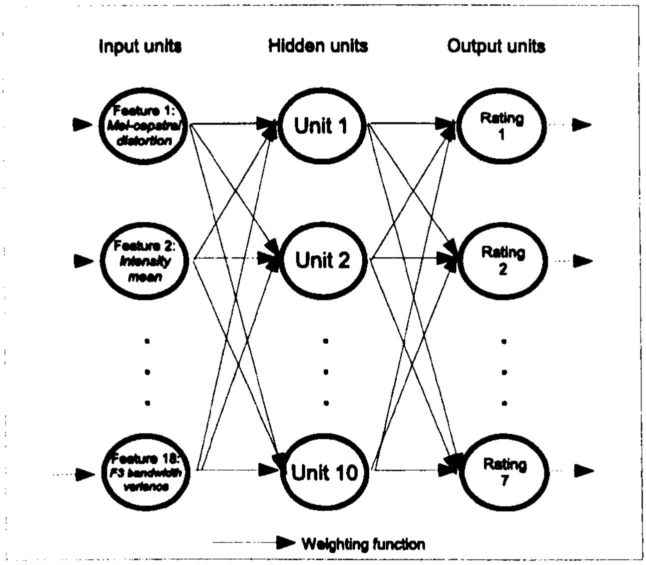


Figure 5.5: *Multi-layer perceptron showing layers of input units, hidden units and output units. Each connection has an associated weight, optimised during training, illustrated by a solid line arrow. The input in this experiment is the difference between a target and synthesised sentence for 18 acoustic features and the output is a probability estimate of a rating 1-7*

unvoiced according to the PRAAT pitch extraction algorithm. This measure is similar to the vowel duration error as used for evaluation of voice conversion in [223].

Feature	Description
1	Mel cepstral distortion
2	Intensity mean
3	Intensity variance
4	Fraction of voicing agreement
5	F0 mean
6	F0 variance
7	F1 mean
8	F1 variance
9	F1 bandwidth mean
10	F1 bandwidth variance
11	F2 mean
12	F2 variance
13	F2 bandwidth mean
14	F2 bandwidth variance
15	F3 mean
16	F3 variance
17	F3 bandwidth mean
18	F3 bandwidth variance

Table 5.2: *Description of the acoustic features used to train the MLP*

5.4.4 Training

The training for the multi-layer perceptrons used 10 hidden units to model the listeners' judgements given the extracted features. The number of hidden units was arbitrarily selected to maintain a balance between the complexity it can capture and time taken for training. 7 output units corresponded to the listener's judgement on the similarity of the stimuli to the target speaker j, in the range 1-7. Each pair of sentences T-S, each listener and each stimulus presentation generated an MLP dataset item in which the input layer received the acoustic feature values. The output unit target value was 1 for the jth output unit and 0 for the remainder. The trained MLP therefore estimated the probability that a listener's rating of the similarity between the synthetic utterance S and the target utterance T would take each of the 7 allowed scores.

5.4.4.1 Speaker-dependent MLPs

For the speaker-dependent MLPs, 50% of the dataset was randomly chosen to form the training set. The remainder was used for testing.

Different combinations of the extracted features were used to train the MLPs to identify those features which contributed the most to accurately predicting the listener responses. These combinations were derived empirically excluding each feature one at a time from the training to determine which contributed to the overall judgement replication accuracy.

5.4.4.2 Speaker-independent MLPs

The speaker-independent MLPs were trained on 50% of the dataset from each speaker's data. To provide data that was balanced between the two experiments for speaker 1 and speaker 2, where there were 7 participants for the speaker 1 experiment and 10 participants for the speaker 2 experiments, 3 of the 10 listeners' responses for the speaker 2 experiment were randomly removed from the training dataset. This ensured that the data associated with each speaker experiment contributed equally to the MLP training.

The remainder of the data (still excluding the extra 3 speakers for the speaker 2 experiments) was used for testing.

5.4.5 Results: speaker-dependent MLPs

Figures 5.6 and 5.7 show that the MLP can reproduce listeners' judgements (shown in figures 5.3 and 5.4) accurately for an unseen test set, given all 18 acoustic features. Note that MLPs produce probability estimates not probabilities, which accounts for why the probability estimate does not exactly total 1.

Listener and MLP results are compared quantitatively using the mean squared point-by-point error in the 5x7 condition/judgement matrices (termed the Frobenius norm). This is illustrated in table 5.3, which gives the MLP prediction error for various feature combinations. The acoustic features to which these numbers correspond are listed in table 5.2. These results show the mean prediction error of the results over 10 trials. Statistical significance is determined by whether the mean error over the 10 trials of one set of features is 2 standard deviations away from the mean of a comparison feature set. A pairwise comparison of the acoustic feature set errors in table 5.2 determined that all features sets displayed are significantly different from each other.

The most important features for the MLP as determined by dropping each individual feature out of the MLP training procedure, are mel cepstral distortion and fraction of voicing agreement. The formant bandwidth features have least influence for the MLP modelling. However, the results show that the modelling of many features in combination is required for accurate prediction.

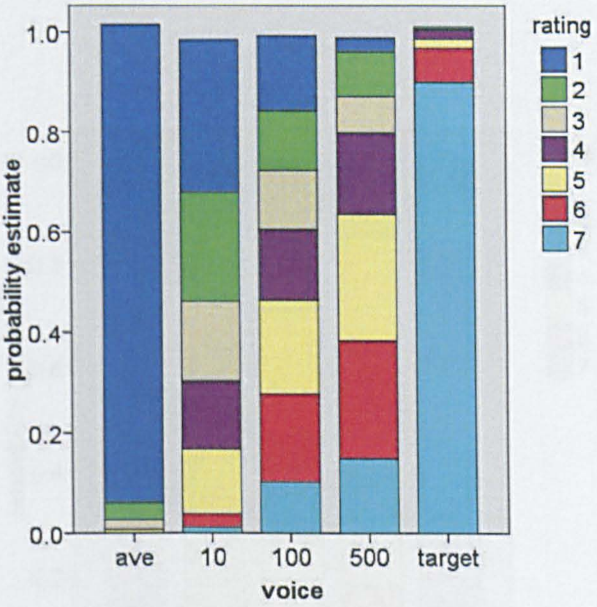


Figure 5.6: *Results of the speaker-dependent MLP experiment for speaker 1. Rating 1 (sounds like a different speaker) to 7 (sounds like the same speaker) is used.*

5.4.6 Results: speaker-independent MLPs

To use this technique as an objective measure for evaluating similarity of synthesised speech without training a speaker-dependent MLP for that data, listener responses were generated using the MLP trained on data from one speaker on the synthesised speech of the other speaker. The MLP technique showed to be a promising method of objectively measuring similarity for an unseen speaker's data, with a prediction error of ~0.01. This result indicates that it may be possible to use a pre-trained MLP to replace listening tests when assessing voice similarity for a new speaker.

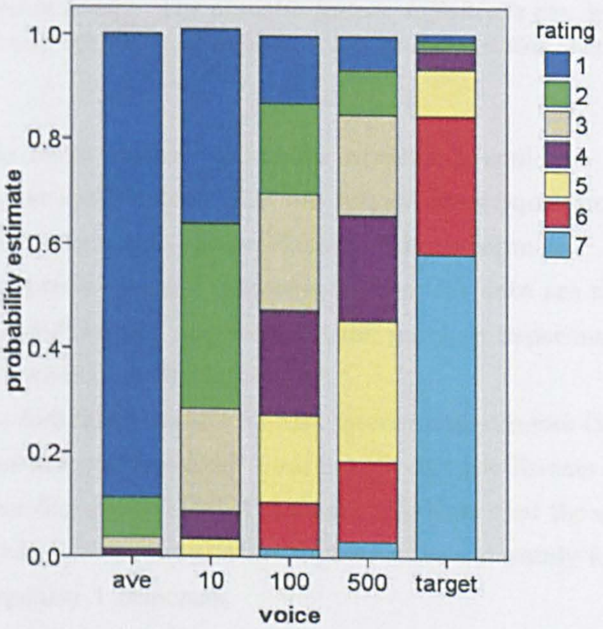


Figure 5.7: *Results of the speaker-dependent MLP experiment for speaker 2. Rating 1 (sounds like a different speaker) to 7 (sounds like the same speaker) is used.*

Input features	speaker 1	speaker 2
1	0.0031	0.0046
4	0.0041	0.0038
1,4	0.0016	0.0016
1,2,4	0.0013	0.0012
1,2,4,5	0.0011	0.0011
1,2,4,5,7,11,15	0.0008	0.0009
1,2,3,4,5,6,7,8,11,12,15,16	0.0007	0.0005
all	0.0006	0.0002

Table 5.3: *MLP prediction error for different input feature combinations (see section 5.4.5 for definition of prediction error). The acoustic feature definitions can be found in table 5.2. Mean values over 10 trials are given. The feature sets shown are statistically significant from each other.*

To investigate this result further and assess whether it could be more generalisable across speakers, a speaker-independent MLP was trained using equal amounts of data from speaker 1 and speaker 2 taken from the previous speaker experiments. The results of how well this approach can predict human responses for the test data are shown in figure 5.8, which shows the combined listener responses for the previous experiments and figure 5.9, which shows the MLP predictions on the test set.

Comparison of the individual results for the listener experiments (figures 5.3 and 5.4) shows that the speaker-independent MLP does not predict the listener responses as accurately as using a speaker-dependent MLP. The results also show that the speaker-independent MLP predicts the human listener judgements slightly more accurately for the speaker 2 responses than for the speaker 1 responses.

5.4.7 Discussion

The results show that listeners' judgements can be modelled closely using the extracted acoustic features of the synthesised and target sentences. Mel cepstral distortion has been used to calibrate synthesised voice quality in voice conversion together with F0 error and vowel duration error [118, 223]. The results shown in table 5.3 show that similarity measures produced using just these features in isolation do not provide an accurate perceptual response to how similar the synthesised and target sentences are to each other. The comparison shows a much more complex picture, specifically looking at the effects of the features in combination and taking advantage of the non-linear classification capability of the MLP.

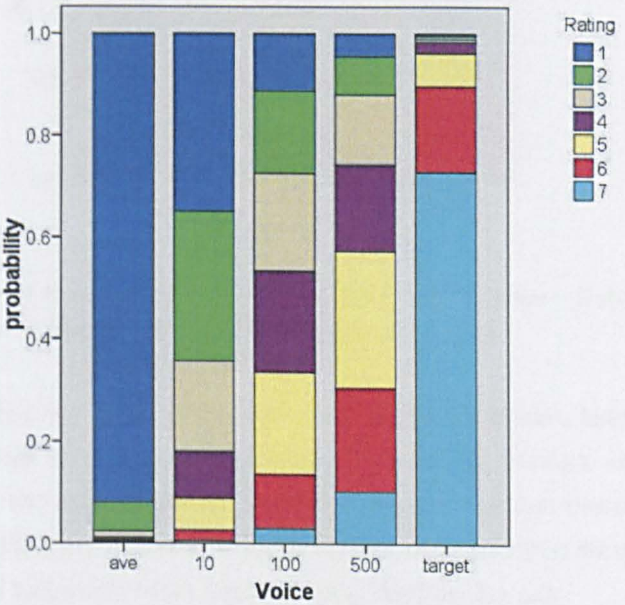


Figure 5.8: *Results of the listener experiments with both speakers combined (taking each speaker as an equal contribution). Rating 1 (sounds like a different speaker) to 7 (sounds like the same speaker) is used.*

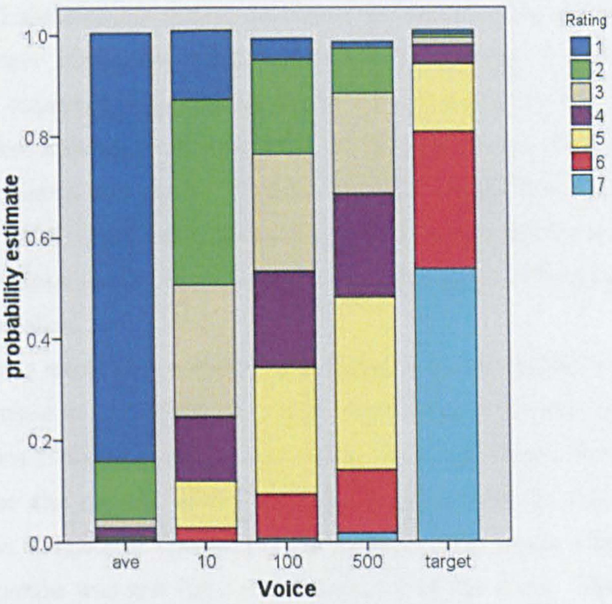


Figure 5.9: *Results of the speaker-independent MLP experiments. Rating 1 (sounds like a different speaker) to 7 (sounds like the same speaker) is used.*

The results show that the MLP more accurately replicates human listener responses if all the extracted features are used in combination to train the network and make the judgement. To fully test the importance of the contribution of the non-linear capabilities of the MLP on these results, there would have to be further investigations done. This may include comparisons with a technique combining acoustic features linearly.

The results of the experiment with MLPs trained on speaker-independent data and applied to the test set show that this procedure shows potential to be used to build a listener response predictor for assessing similarity throughout a voice building process.

The predictions displayed by the speaker-independent MLP seem to match more closely to speaker 2 responses than speaker 1 responses. This suggests that the data for speaker 2 is more predictable and therefore more consistent with the training set than the results for speaker 1. These results, however, have not been tested for significance and without significance testing the results are difficult to interpret. These experiments are designed purely as a pilot study to see if this technique holds potential to be used as an objective measure for similarity between synthesised and target sentences. There are many limitations in this procedure which would need to be taken into account for future work, including this lack of statistical significance testing for this experiment. Comparison of the predictor results with the listener responses indicate the direction of the best method of response prediction where, as expected, the speaker-dependent systems are the most accurate predictors and the speaker-independent predictors are more accurate than using one speaker's data to predict the other's responses. To fully test the capabilities of this technique, the speaker-independent MLP should also be used to predict responses for a new speaker's data. Further to this, more data should be collected to make a more robust speaker-independent listener response predictor.

A limitation in the experiment which determined the extracted acoustic feature sets was that each trial used to calculate the mean prediction error was conducted using the same randomly chosen training and test sentences. This calculated the mean and variance over the 10 trials for the output error, which took into account the randomly assigned starting point for the weights in the MLPs. However it also meant that the variance used in the statistical measure was not fully representative of the data. The assumption taken was that the training set was a good representation of the distribution of the entire data set. For a more in depth investigation into whether this technique could be used for the purpose of providing an objective measure, a cross-validation technique should be used in the experiment where both the training and test sets are varied over the multiple trials.

A further limitation is that the error measurement only takes into account the high level distribution of results, rather than testing the data on an individual utterance level. The assumption made here is that if the test set distribution matches the distribution of the overall set of results then the MLP is a good representation of the responses. However, to show that the MLP is always able to accurately predict the results of individual utterance responses then further work may involve a more detailed utterance level error metric.

A more general limitation of using this approach for the objective evaluation of similarity to a target speaker is that it requires the availability of synthetic and target versions of the same sentences. This is not always possible in practice, particularly with respect to speakers with dysarthria who frequently do not have 'target speech' recordings. This is discussed in more detail in chapter 6.

5.5 Conclusions

The listening experiment confirmed that at least 100 sentences are needed to build a voice that is distinct from the average voice and resembles a target speaker, although using more data increases the similarity. The similarity also depends on the speaker and the average voice used. Using an average voice which is closer to the target speaker is likely to produce a better result. Differences were found between the two target speakers which may be related to the proficiency of the speaker to read naturally as one speaker was a professional broadcaster.

The experiments of section 5.4 show that listeners' judgements can be modelled closely by comparing acoustic features. The performance of speaker-independent MLPs shows potential to be a listener response predictor for assessing similarity throughout a voice building process. These experiments, however, have many limitations and provide only the suggestion that this technique has the potential to be used as an objective measure of similarity where there is an availability of synthetic and target versions of the same sentences. Further work in this area is required to make more robust conclusions about its capabilities.

It has been shown that mel cepstral distortion and voicing agreement have the largest influence on similarity judgement (see table 5.3). This experiment suggests that mel cepstral distortion alone is not sufficient for measuring voice similarity and a measure which combines other acoustic features is needed to provide an objective measure of similarity between synthesised and target speech. This experiment shows that the non-linear combination of factors could prove to be beneficial in providing such an objective measure, but this requires further investigation.

These experiments can also inform work on building voices for speakers with speech impairments. The 'voice banking' approach as detailed above has to be altered if the adaptation data is affected by the individual's condition. As will be reported in the following chapter, the approach with dysarthric speakers involves capturing the individual's speaker characteristics without replicating the impairment in the synthesis. Results from the above experiments suggest that spectral information, F0 and voiced passage duration contribute significantly to perceptual similarity judgements. It follows therefore that these features need to be retained as far as possible from the speaker data to be able to produce a synthesised voice that captures the identity of the individual.

Chapter 6

Building voices using dysarthric data

6.1 Introduction

This chapter reports experiments done using HMM-based synthesis to build personalised synthetic voices using dysarthric speech data. The results of chapter 5 showed that it was possible to build reasonable voices with non-disordered data using the HTS adaptation system. This chapter uses the same technique with dysarthric speech adaptation data and describes the alterations made in the process to deal with the differences in the data. This chapter evaluates the synthetic voices reconstructed for three individuals with different types of dysarthria. The results of evaluations of these voices by the speakers themselves and listeners who know them are presented and discussed. The results in this chapter have been partially reported in [48].

6.2 Evaluation

To evaluate whether these substitutions make using HMM-based synthesis a possibility for building voices for individuals with dysarthric speech, the next section describes an experiment to evaluate voices built using these modifications. The aim of this evaluation is to investigate whether appropriate synthesised voices can be built for individuals whose speech has already begun to display dysarthric characteristics using the procedures stated above. This is implemented for three individuals with different pathologies.

This evaluation poses three questions:

1. *Can the individual recognise themselves in the voices built and which features contribute to this recognition?*

2. *Which features affect the quality of the voice output for the different participant speakers?*

3. *Can features be altered to make the voices more appropriate for that speaker?*

Question 1 aimed to provide information about which features should be used in the output model to capture the individual's speaker characteristics, using the results from the experiment in chapter 5. It also aimed to provide a measure of how well or if the output model captures the speaker's identity in the synthesis, following the requirement of similarity of speaker as set out in chapter 2. Question 2 aimed to see which of the possible substitution of features would affect or improve the output voice synthesis, where quality can be associated with its potential use in a communication aid, and whether these feature combinations provide a synthetic voice that would be usable in such a situation. This follows the requirements of intelligibility and naturalness of the voice quality, required for acceptability of a VOCA as set out in chapter 2. Question 3 aimed to look at whether there are features that could be altered to suit the requirements of the user, providing more information on the flexibility of the system. This addresses the requirement of manipulability of prosodic output as defined in the list of requirements for the output voice in chapter 2.

6.2.1 Method

The method for the evaluation of the voices built with data in which the speech has begun to deteriorate does not follow that of chapter 5 because there is no target speech available with which to compare the synthesised voices. The target voice in this experiment is one that is recognisable as the individual who provided the original speech but that has been reconstructed to provide an intelligible synthesised voice without the dysarthric features. To know whether the output speech is appropriate for that speaker for potential use in a communication aid, the evaluation participants should be people who are able to make that type of judgement without hearing target speech for comparison. The target speech for those participants exists in some form in the memory of the participants. For this reason, the evaluation takes a qualitative approach and uses the speakers themselves (defined as *participant speakers*) and people who know the speakers (defined as *participant listeners*) as participants in the evaluation.

6.2.1.1 Participants

The participant speakers in this experiment are identified as speakers 3, 4 and 5. Speaker 3 was male and 80 years old at the time of recording, two years post cerebrovascular accident (CVA), with moderate flaccid dysarthria. In his speech, overall energy varied, with imprecise and slow movement of the articulators resulting in a slow rate of production. An example of his original speech is available as example 6.1 on the attached CD. Speaker 4 was male, 69 years old at the time of recording and had been diagnosed with Parkinson's disease six years previously. He showed symptoms of mild hypokinetic dysarthria. His speech was quiet, with variable energy. There was little variation in pitch and a high perceived rate of articulation. An example of his original speech is available as example 6.2 on the attached CD. Speaker 5 was male, 80 years old at the time of recording with severe primary progressive apraxia of speech and dysphasia, which had onset six years previously. His dysarthria was classed as moderate at the time of recording. His speech was telegraphic, it contained many insertions, with imprecise and slow movement of the articulators resulting in a slow rate of production. An example of his original speech is available as example 6.3 on the attached CD.

These speakers show a range of different pathologies and severity of dysarthria. To fully investigate the possibilities of using HTS for voice building for speakers with dysarthria a much wider population and many more speakers would be required to make quantitative and stastically significant claims. For a proof of concept study such as this, three speakers were deemed to be sufficient to see if this is a technique which can be explored more fully.

The evaluation is conducted using speaker participants 3 and 4 themselves, evaluating their own voice. Speaker 5 did not participate in the evaluation as it was felt that it would be too distressing for him in the current stage of his condition. For speakers 3 and 4, two listeners who knew the speaker also participated in the evaluation. One participant listener was able to participate in the evaluation of speaker 5's synthetic voices. The participant listeners were student or staff members of the University of Sheffield. They were not only participating as listeners familiar with the speakers but also as expert listeners due to their training or work as speech and language therapists, familiar with speakers with these conditions and situations in which communication aids are used. The listeners reported no hearing or speech impairment and were not paid for their participation.

6.2.1.2 Data collection

Data was collected for speakers 3 and 4 in a quiet clinic room in the Department of Human Communication Sciences, University of Sheffield using a Morantz PMD670 audio recorder with a Shure SM80 microphone. Speaker 5's data was recorded onto a laptop computer using the internal microphone in his own home.

The recorded material was taken from the Arctic dataset A, the same set which was recorded for the experiments in chapter 5. The sentences were presented to speakers 3 and 4 on separate sheets of paper in a folder to avoid any listing intonation effects in the reading and to maintain consistent recording conditions. The participants were asked to read the sentences as naturally as possible. The participants completed the recordings in one sitting but were encouraged to take breaks with a drink of water at least every 50 sentences or as often as they felt necessary. In these conditions, speaker 3 recorded the first 200 sentences of the Arctic set A and speaker 4 recorded the first 150 of the same set.

For speaker 5, the sentence prompts were displayed one at a time on the computer screen using Prorec 1.01 Speech Prompt and Record system [94]. This displays the prompts in sequence and records the speech data as it is produced. The sentences were recorded in sections of 20 per session and he had aimed to do two recording sessions a day. Although it is generally considered best to record at the same time of day for voice building recordings [21], it was left to the speaker to decide when during the day he felt able and motivated to do the recordings. Speaker 5 completed the recording of the first 379 of the set A sentences before it was felt that the recordings should not continue as his voice was deteriorating faster than expected.

6.2.1.3 Building voices

The voices were built using HTS version 2.1 (internal), 138-dimensional feature vectors and the same average voice set up as in the voice banking experiments (see section 5.3.1.5).

For each speaker, two voices were built: one with all the unedited data the participants had recorded and the other with data manually selected for intelligibility. The editing was done by listening to the utterances and selecting those sections which were intelligible, matching the labels extracted from the expected orthographic transcription of the original prompts. Any sections with noise, unlabelled pauses or perceived incorrect articulations were removed from the adaptation data. A protocol was designed to ensure as consistent a process as possible. The protocol used a minimum selection size as one syllable and required the surrounding context of the articulations at each edge of the selection to coincide with the labels. Further details of the data selection protocol are detailed in appendix D.

Figure 6.1 shows the amount of data used for the adaptation process for each speaker and particularly shows the importance of this data selection technique for speaker 5. Speaker 5 is the speaker with the most severe dysarthria. His unedited speech data was rejected at a high rate, leaving only 21% of the data remaining for use as adaptation data. After data selection, 55% of the total data was selected as intelligible and 88% of that was used during the adaptation procedure. In total, 48% of the data could be used for adaptation when the data had been edited for intelligibility, an increase of 27% from the unedited data. As an example, 6.4 on the CD is an example of speaker 5's voice (see section 6.2.1.1 for more information) built with unedited data and using all his own voice features. Example 6.5 is an example of the same speaker's voice built with data edited for intelligibility and all his own features. Example 6.3 on the CD is as example of his originally recorded data for comparison.

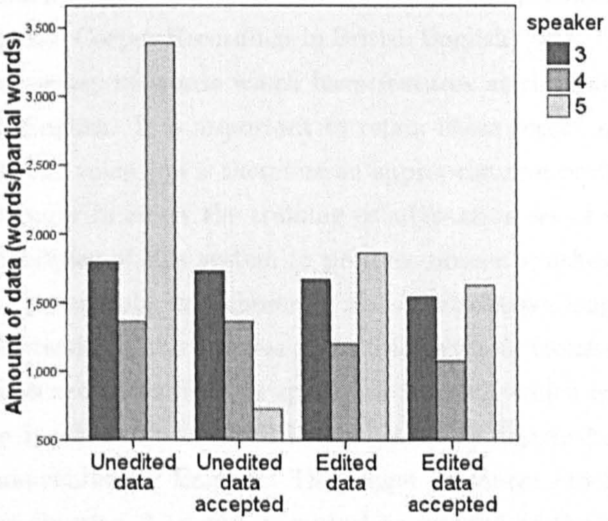


Figure 6.1: *Amount of data accepted by the system in the case of unedited original recorded data and data edited for intelligibility. The difference visible for Speaker 5's data shows the importance of this procedure for severely dysarthric speech.*

For speakers 3 and 4, the data selection technique reduces the amount used by the adaptation process but improves the quality of that data in terms of removal of segments with unwanted noise, inappropriate insertions or unintelligible sections. In total, 86% of the total speech was used for adaptation for speaker 3 and 79% of the total speech data was used for adaptation for speaker 4. The evaluation compares the voices built with each set of data and assesses whether the trade off between amount and quality is worth the data selection procedure.

Speakers with dysarthria may find that they are unable to produce certain articulations consistently, which would result in an unbalanced data set across phonemes. This was not entirely the case, however, for these speakers who although were more consistent at producing certain articulations, all had full coverage of phonemes in their edited data set. Further work into using unbalanced phoneme coverage for building voices could be pursued particularly for speakers with more severe dysarthria who have a very limited repertoire of articulations.

6.2.1.4 Stimuli

The stimuli presented to the participants were synthesised sentences and paragraphs taken from SCRIBE (Spoken Corpus Recordings in British English) [95]. The SCRIBE paragraphs contain a high frequency of words which have features attributable to different regional accents of British English. It is important to retain these accent-specific features to fully personalise a synthetic voice and is therefore an appropriate set of data for this task. These sentences did not occur in either the training or adaptation set of sentences and therefore fully test the capabilities of this system to produce unseen synthesised output. Sentences extracted from the paragraphs were chosen to make the passages long enough for the listener to get a general impression of the features of the voice without focussing on individual errors. The paragraph sentences were used for questions 1 and 2, which consisted of 37-54 words depending on the speaker. The SCRIBE sentences were designed as a whole set to cover the range of demi-syllables of English. The longer sentences (15-18 words) were used in the evaluation for question 3 to give a general impression of the voice while being short enough to make the length of the whole evaluation appropriate. The test set is available as appendix E.

6.2.1.5 Procedure: speakers

The evaluations took place in a quiet room at the University of Sheffield. The stimuli were presented to the participants individually using a laptop computer with external speakers.

The research was introduced as building voices for a computer to use to speak for that individual on days where their own voice was not clear. An example of the average voice was played and introduced as a starting point from which the voice was changed to an approximation of the participant's voice, based on the data that they recorded previously. An original recording of two non-disordered voices built with 500 sentences was played, followed by the synthesised version of the same sentence and a sentence for which they had not heard an original recording. This was to make the participants aware of the capacity of the performance of this system. For each voice the participants were asked to rate the similarity of the synthesised output to the original recordings on a 1 (sounded like a different person) - 5 (sounded like the same person) scale. This attempted to gauge their reaction to the synthesised voices whilst getting them used to the task ahead. It also provided an opportunity to attune their hearing to synthesised speech. Where they responded with two separate ratings for each sentence, the rating given for the example without an original recording is displayed. This better represents whether the speaker characteristics have been captured rather than just performing a direct comparison of the synthesis and original version.

An example of their own speech from the original recording was also played to the individual to make them aware of the sound of their own voice as played through the computer. The voice as heard by the speaker during production sounds different to the voice when heard from a different direction or played back to the individual. This phenomenon is explained by the sound transmission medium. When recorded, the voice is captured by the microphone, which is usually placed in front of the speaker and when played back, the sound is transmitted to the auditory system via air conduction. When hearing one's own voice in production, the ears, as receivers of the sound through air conduction are in a different position and the sound is not only transmitted through the air but also through bone conduction [217]. This difference makes it more difficult for the speaker participants in the experiment to recognise themselves in the output speech. It does, however, allow them to hear the output speech as listeners would hear it, including themselves, which is relevant to the evaluation of the voices for the potential application of using the voices with communication aids.

1. Is the participant speaker recognisable in the voices built and which features contribute to this recognition?

To answer the first evaluation question, comparisons were made between the average voice and voices synthesised with average voice components introducing features that display speaker characteristics [211] taken from the participant speaker model. The results of chapter 5 showed that the largest contributions to the perceptual judgements of similarity were mel cepstral distortion and fraction of voicing agreement between frames for the synthesised and target speech. This indicates that spectral information is important for speaker similarity, with formant frequency information also contributing to a significant difference in reducing prediction error when introduced into the MLP training. Duration of segments, as approximated by the fraction of voicing agreement, and intensity were not introduced into this part of the evaluation despite the contribution to predicting listener similarity responses. These are features commonly affected by dysarthria and are likely to introduce distortion into the output synthesis, which may sound more like the speaker but only in terms of reproducing their impairment, which is not the aim of the evaluation. Fundamental frequency contributed to reducing the prediction error in the similarity experiments and is regarded as a contributing feature to speaker identity.

The conditions in this stage of the evaluation were: average voice, average voice with participant log F0 features, average voice with participant spectral information and average voice with participant log F0 and spectral information. Only edited data was used to build these voices. The participants were asked to rate the difference between the original recording and the synthesis on a 1 (does not sound like me) - 5 (sounds like me) scale. One stimulus per condition was presented, where the length of the stimuli aimed to provided enough information to give a generalised idea of the condition synthesised, rather than provide a number of different examples producing a number of different ratings. There are too few participants in this evaluation to analyse the rating results quantitatively to show anything other than trends. It was therefore judged that one long paragraph would provide a more general overall reaction to the condition than multiple shorter sentence ratings.

Using the average voice condition created a 'speaker line-up' situation in which the participant became aware that not all the voices would sound like them and they would have to identify those that did sound like them or that may contain some of their speaker characteristics. The same paragraph was played for each condition, which was different to the original recording.

2. Which features affect the quality of the voice output for the different participants?

To answer the second evaluation question, a choice was presented between the average voice with participant spectral and log F0 features and the same voice with one additional feature of the participant speaker's model substituted. The question asked was "For each pair, which voice do you think sounds best?". Conditions evaluated were: use of the participant's durations, use of the participant's global variance for spectral features, use of the participant's energy and using the full set of unedited data to build the voice. These conditions were chosen for evaluation as they had a perceived effect on the output for at least one of the participants. The participant could indicate that they perceived no difference between the two samples. The pairs were randomly ordered and could be listened to as many times as was required by the speaker. The use of the word 'best' allowed the individual an interpretation using whatever criteria they deemed appropriate to answer the question. As for the previous section, one paragraph was used for each condition, again allowing the choice to be made based on the overall reaction to more information.

3. Can features be altered to make the voices more appropriate for that participant?

The third evaluation question dealt with appropriateness of synthetic speech output for that participant and their preferences for the customisable features: speech rate of utterance and global variance for log F0. A pairwise comparison was made for three different sentences. For rate, the comparison was between the average voice durations and a slowed down version of the average voice durations. For global variance for log F0, the two options were that of the average voice or that calculated from the participant's adaptation data. For each pair the question was asked "Can you tell a difference and if so, which one do you prefer?". Only edited adaptation was used to build these voices.

Follow up questions to access the overall acceptability of the voice were then posed as follows:

- Do you like the voice? For the one you liked the best, can you give a rating of 1 (do not like the voice) - 5 (like the voice)?
- On days when you felt your voice was not clear, would you be happy to use that synthesised voice instead?
- Is there anything you would change about the voice?
- If you could choose between using this voice or an alternative voice (an example of a commercially available voice from Acapela (Peter), example 6.6 on the CD), which would you choose?

6.2.1.6 Procedure: listeners

The procedure for the participant listeners experiment was very similar to that designed for the participant speakers. For the listeners who knew speakers 3 and 4, the stimuli were presented to both participants at the same time in each evaluation to allow for discussion although their responses were recorded separately.

The listeners were only presented with one of the non-disordered speech voices. One example was deemed to be sufficient for the listeners to get used to the task and they had all had experience listening to synthesised speech before. The listeners were not presented with original recordings from the speaker's data set allowing responses to the stimuli based only on their perception of whether the output could be associated with the speaker themselves rather than a direct measure of similarity to the initial recordings. This was particularly important for the 'speaker line-up' situation so they could indicate when they recognised any of the participant speaker's characteristics in the presented conditions based on their memory of the speaker's voice.

The questions asked during the presentation were for section 1 "Does this voice sound like the speaker?", for section 2 "Which of these voices sounds best for the speaker?" and for section 3 "Can you tell a difference and if so, which one is most appropriate for that speaker?". Follow up questions were not asked to the participant listeners.

6.3 Results

6.3.1 Speaker 3

The results of the introductory part of the experiment for speaker 3 and listeners who knew speaker 3 (listeners 3A and 3B) are displayed in table 6.1. Non-disordered voice 1 was the same voice used as speaker 1 (voice built with 500 sentences) in the experiments in chapter 5 and only the speaker participants heard non-disordered voice 2. The ratings were very similar for each participant, all agreeing that the synthesised versions sounded very like the original speakers. This correlates with the results shown in chapter 5.

1. Is the participant speaker recognisable in the voices built and which features contribute to this recognition?

Voice	Speaker 3	Listener 3A	Listener 3B
Non-disordered voice 1	5	4.5	4.5
Non-disordered voice 2	4	-	-

Table 6.1: *Ratings from speaker 3 and listeners 3A and 3B evaluating voices built for voice banking on a 1(does not sound like that speaker) - 5(sounds like that speaker) similarity scale. Note that only the speaker participant heard non-disordered voice 2.*

The results for the first question in the evaluation are shown in table 6.2. After exposure to the stimuli, speaker 3's rating of the average voice was high, showing that he perceived the average voice as sounding similar to his own. The rating increased to 5 for all other conditions containing components of his model substituted into the output voice. Listeners 3A and 3B note more discriminating differences between the voices, agreeing that introducing speaker log F0 alone is insufficient to recognise speaker characteristics in the output synthesis. The similarity increases as the speaker's own spectral features are used and further increases when using both speaker spectral features and log F0. In discussion the listeners stated that using just the speaker's own log F0 with all other features being those of the average voice was rated by the listeners as having the lowest similarity to the speaker as they stated that the average voice was smoother in quality.

Voice	Speaker 3	Listener 3A	Listener 3B
Ave	4	1	1
Ave + speaker logF0	5	1	1
Ave + speaker mel cep	5	2	1.5
Ave + speaker logF0 + mel cep	5	3	3

Table 6.2: *Ratings from speakers 3 and listeners 3A and 3B evaluating voices built from average voice models with different participant speaker model components introduced. Ratings are on a 1(does not sound like me/him) - 5(sounds like me/him) similarity scale.*

2. Which features affect the quality of the voice output for the different participants?

The results for this section are summarised in table 6.3. The unedited data version was preferred by listener 3B only, with listener 3A agreeing with speaker 3 in that there was no difference. Listener 3A agreed with speaker 3, preferring the average voice energy over the speaker's own although 3B preferred the speaker's own energy, even though it was noted that it was difficult to listen to because of the fluctuations in the output. Speaker 3 preferred the voice using his own durations but both listeners preferred the average voice durations. There was agreement between the listeners and speaker 3 that there was no difference between the voices when using the different global variance for mel cepstra.

Feature (conditions)	Speaker 3	Listener 3A	Listener 3B
Data (Unedited/Edited)	No diff.	No diff.	Unedited
Energy (Speaker/Ave)	Ave	Ave	Speaker
Durations (Speaker/Ave)	Speaker	Ave	Ave
GV for mel cep (Speaker/Ave)	No diff.	No diff.	No diff.

Table 6.3: *Preferences for quality shown by speaker 3 and listeners 3A and 3B for output synthesised in two difference conditions, shown in brackets. All other features remained constant.*

3. Can features be altered to make the voices more appropriate for that participant?

The results for this question are shown in tables 6.4 for output rate and 6.5 for global variance for log F0. The results show that differences between the conditions are discernible and preferences can be made for both rate of utterance and global variance for log F0. Speaker 3 noted no difference between the different durations for all three comparisons. Listener 3A preferred the average voice durations and Listener 3B preferred the slowed down versions.

The listeners preferred the average voice global variance for log F0 whereas speaker 3 preferred his own global variance for log F0.

	Speaker 3			Listener 3A			Listener 3B		
Conditions	ave	slow	none	ave	slow	none	ave	slow	none
Preferred	0	0	3	2	1	0	0	2	1

Table 6.4: *Number of utterances out of 3 preferred for different rates of speech for each participant. The conditions compared average voice durations (ave) and average voice durations slowed down (slow). None indicates the participants had no preference.*

	Speaker 3			Listener 3A			Listener 3B		
Conditions	ave	sp.	none	ave	sp.	none	ave	sp.	none
Preferred	0	2	1	3	0	0	2	1	0

Table 6.5: *Number of utterances out of 3 preferred for different global variances for log F0 for each participant. The conditions compared average voice gv-lF0 (ave) and the speaker's own gv-lF0 (sp.). None indicates the participants had no preference.*

For the rating of likeability of the voice, from 1 (do not like the voice) - 5 (like the voice), speaker 3 rated his output voice as 5. He stated that he would be happy to use that voice on days when his own was not clear and showed no preference between the choice of using his own reconstructed voice or the Acapela voice, Peter.

6.3.2 Speaker 4

The results of the introductory part of the experiment for speaker 4 and listeners who know speaker 4 (listeners 4A and 4B) are displayed in table 6.6.

Voice	Speaker 4	Listener 4A	Listener 4B
Non-disordered voice 1	1	4	4
Non-disordered voice 2	1	-	-

Table 6.6: *Ratings from speaker 4 and listeners 4A and 4B evaluating voices built for voice banking on a 1(does not sound like that speaker) - 5(sounds like that speaker) similarity scale. Note that only the speaker participant heard non-disordered voice 2.*

Again, non-disordered voice 1 was the same voice used as speaker 1 (voice built with 500 sentences) in the experiments in chapter 5 and only the speaker participant heard non-disordered voice 2. The ratings from the listeners were the same and the high rating correlates with the results shown in chapter 5. Their rating does not agree with that of speaker 4 who rated the voice with no similarity to the original speech, which does not agree with the results from chapter 5.

1. Is the participant speaker recognisable in the voices built and which features contribute to this recognition?

The results for the first question in the evaluation are shown in table 6.7. Speaker 4's ratings remained at 1 for each condition, stating that when speaker information is substituted into the models, he did not recognise himself in the voice. The similarity rating was not high from the listeners but increased when information taken from the speaker's models was introduced. The average voice with the speaker's spectral information obtained the highest rating when the listeners discussed ordering the voices in terms of similarity to the speaker. Introducing the speaker's spectral features was vital to capture some of the speaker's characteristics in the voice as indicated by the rating increase. The speaker's log F0 information was less important.

Voice	Speaker 4	Listener 4A	Listener 4B
Ave	1	1	1
Ave + speaker logF0	1	1	1
Ave + speaker mel cep	1	2	2
Ave + speaker logF0 + mel cep	1	2	2

Table 6.7: *Ratings from speakers 4 and listeners 4A and 4B evaluating voices built from average voice models with different participant speaker model components introduced. Ratings are on a 1(does not sound like me/him) - 5(sounds like me/him) similarity scale.*

2. Which features affect the quality of the voice output for the different participants?

The results for the second evaluation question are summarised in table 6.8. The results show that for the data type, speaker 4 noted a difference, preferring the unedited data. Listener 4A noted no difference and listener 4B preferred the edited data version. Speaker 4 and listeners 4A and 4B all agree that the use of the average voice energy produces a better output than using the speaker's own energy information. Using the speaker's own durations made no difference for speaker 4 but the listeners preferred the average voice durations. Using the speaker's global variance for spectral features made a difference distinguishable by speaker 4 preferring his own global variance for mel cepstra whereas the listeners preferred using the average voice global variance for mel cepstra output.

3. Can features be altered to make the voices more appropriate for that participant?

The results for this question are shown in tables 6.9 for rate of output and 6.10 for global variance for log F0. The results show that differences between the conditions are

Feature (conditions)	Speaker 4	Listener 4A	Listener 4B
Data (Unedited/Edited)	Unedited	No diff.	Edited
Energy (Speaker/Ave)	Ave	Ave	Ave
Durations (Speaker/Ave)	No diff.	Ave	Ave
GV for mel cep (Speaker/Ave)	Speaker	Ave	Ave

Table 6.8: *Preferences for quality shown by speaker 4 and listeners 4A and 4B for output synthesised in two difference conditions, shown in brackets. All other features remained constant.*

discernible and preferences can be made for both rate of utterance and global variance for log F0.

Speaker 4 and listener 4B noted differences between the two rates of production and showed a preference for the average voice durations. Listener 4B showed no real preference across the three comparisons. The results for global variance for log F0 showed that a difference was detectable between the two conditions, with all participants preferring that of the average voice.

	Speaker 4			Listener 4A			Listener 4B		
Conditions	ave	slow	none	ave	slow	none	ave	slow	none
Preferred	2	0	1	1	1	1	2	0	1

Table 6.9: *Number of utterances out of 3 preferred for different rates of speech for each participant. The conditions compared average voice durations (ave) and average voice durations slowed down (slow). None indicates the participants had no preference.*

	Speaker 4			Listener 4A			Listener 4B		
Conditions	ave	sp.	none	ave	sp.	none	ave	sp.	none
Preferred	2	0	1	2	1	0	3	0	0

Table 6.10: *Number of utterances out of 3 preferred for different global variances for log F0 for each participant. The conditions compared average voice gv-lF0 (ave) and the speaker's own gv-lF0 (sp.). None indicates the participants had no preference.*

For the rating of likeability of the voice, from 1 (do not like the voice) - 5 (like the voice), speaker 4 rated his output voice as 1. He stated that he would not want to use that voice on days when his own was not clear and showed a preference for the Acapela voice, Peter over the presented reconstructed versions of his own voice.

6.3.3 Speaker 5

The results of the introductory part of the experiment for the participant listener who knew speaker 5 (listener 5A) are displayed in table 6.11. The rating showed that the synthesised example of speaker 1 from 5 sounded very like the original speaker. This correlates with the results shown in chapter 5.

Voice	Listener 5A
Non-disordered voice 1	4

Table 6.11: *Ratings from listener 5A evaluating a voice built for voice banking on a 1(does not sound like that speaker) - 5(sounds like that speaker) similarity scale.*

1. Is the participant speaker recognisable in the voices built and which features contribute to this recognition?

The results for the first question in the evaluation are shown in table 6.12. The listener judged that there was no similarity of the voices to the speaker until both the speaker spectral features and log F0 were introduced.

Voice	Listener 5A
Ave	1
Ave + speaker logF0	1
Ave + speaker mel cep	1
Ave + speaker logF0 + mel cep	2

Table 6.12: *Ratings from listener 5A evaluating voices built from average voice models with different participant speaker model components introduced. Ratings are on a 1(does not sound like him) - 5(sounds like him) similarity scale*

2. Which features affect the quality of the voice output for the different participants?

The results for the second evaluation question for speaker 5 are summarised in table 6.13. The results show that the listener preferred the edited data version and the durations and global variance for mel cepstra that were taken from the average voice model. No difference was noted between the speech with the speaker's own energy and that of the average voice.

Feature (conditions)	Listener 5A
Data (Unedited/Edited)	Edited
Energy (Speaker/Ave)	No diff.
Durations (Speaker/Ave)	Ave
GV for mel cep (Speaker/Ave)	Ave

Table 6.13: *Preferences for quality shown by listener 5A for output synthesised in two difference conditions, shown in brackets. All other features remained constant.*

3. Can features be altered to make the voices more appropriate for that participant?

The results for this question are shown in tables 6.14 for rate of output and 6.15 for global variance for log F0. The results show that differences between the conditions were not discernible for this speaker and preferences were not made for both rate of utterance and global variance for log F0.

	Listener 5A		
Conditions	ave	slow	none
Preferred	0	0	3

Table 6.14: *Number of utterances out of 3 preferred for different rates of speech for the participant. The conditions compared average voice durations (ave) and average voice durations slowed down (slow). None indicates the participant had no preference.*

	Listener 5A		
Conditions	ave	sp.	none
Preferred	1	0	2

Table 6.15: *Number of utterances out of 3 preferred for different global variances for log F0 for the participant. The conditions compared average voice gv-lF0 (ave) and the speaker's own gv-lF0 (sp.). None indicates the participant had no preference.*

6.4 Discussion

6.4.1 Limitations of the evaluation

One limitation of this evaluation is the number of participants involved. Using participant listeners who know the individual provides a better evaluation of whether the reconstructed voices with a reduced impact of the effects of dysarthria are still recognisable as that individual. The target speech in this case does not have a tangible reference as was present in the procedure in chapter 5. This therefore limits the number of available evaluation participants to those who know the speaker well enough to be able to recognise aspects of their voice within the output synthesis. Ideally evaluation responses would be collected from the individuals' family members or friends who would be able to provide this type of judgement and for whom the capture of the vocal identity may also be important. This was not available in this study due to various reasons for each speaker, including distress this may cause to family members and difficulty finding appropriate participants. The participant speakers used in this study had the advantage of being familiar with communication aids and synthesised speech and could provide a more emotionally detached assessment of the practicality of the voices built with respect to using them in a communication aid. Using a qualitative approach attaches more value to context and information provided with the responses over the quantities of judgements. The structured interview format provided reasons for responses and allowed reactions to be recorded and discussed. Having the judgement from the speakers themselves is the most important information in this evaluation. It is these individuals who would use these voices in communication aids and who can best determine whether they are suitable to represent them.

A limitation of the voice output quality evaluation was that due to the parameter generation algorithm, this section of experiments was not completely scientifically controlled. It is not just one factor that is being changed in the analysis, with all other remaining the same, the values input to the algorithm affect all the parameters being generated. However, this represents how the substitution of features method would work as part of the HTS system rather than providing a test of how these features individually contribute to the output speech.

These results should be viewed in the context that neither speaker 3 or 4 are at the point in their conditions where a communication aid is necessary. Their speech is generally intelligible although it is affected by their conditions. The idea of being in the position of having to use a communication aid to be able to communicate is a potentially emotional and worrying prospect. This was addressed in the evaluation by introducing the idea of using a communication aid as a temporary solution when trying to get a message across on a day when the individual felt their voice was not clear. However, these results should be viewed mindful of the individual's particular experience of their conditions and the implications that are presented during this evaluation.

Setting the results within the context of the reactions to the introductory examples played to the participants, the results from chapter 5 and the rating from all other participants were contradictory to that rating given by speaker 4. This suggests that speaker 4's ratings were affected by other factors than purely judging similarity between the stimuli.

6.4.2 Capturing speaker characteristics

With sufficient non-impaired data, it has been shown to be possible to produce voices with high similarity to an original recording (see chapter 5). The ratings in the evaluation of the factors involved in capturing the speaker characteristics in the output synthesis suggests, unsurprisingly, that speaker spectral information is vital to produce a likeness to the original speaker. Using the speaker's own log F0 contributes to capturing speaker characteristics only in combination with the spectral information. For both speakers 3 and 4, the participant listeners noted that using the speaker's own log F0 reduced the quality of output slightly introducing a less smooth output in comparison to using the average voice log F0. This effect contributed to their ratings. Examples of the voices built with average voice features and the speakers' own spectral features and log F0 are available as examples 6.7, 6.8 and 6.9 on the CD.

The results suggest that 150 or 200 sentences is not enough data to fully capture the likeness of the speakers' voices using that particular average voice model. Speaker 3's output voice was received more positively in terms of similarity by the listeners than speaker 4 and 5's output. A hypothesised explanation based on the results of chapter 5 is that this is related to the extra sentences of data used for adaptation for speaker 3 where the more adaptation data used, the more the process captures the speaker characteristics. Although speaker 5 used slightly more edited adaptation data than speaker 3, the data used to build speaker 5's voice was of significantly lower quality, as it was recorded using a laptop computer internal speakers. His speech was also more severely impaired than that of the other two speakers and although the data was edited for intelligibility, it is likely that the inconsistency of manual data selection introduced more variable quality sections into the adaptation data.

The influence of the average voice becomes more apparent when using less adaptation data and the US English dominated average voice prevented speaker 4 from recognising himself in the output voices. Listeners 4A and 4B also made this observation and emphasised that the English quality conveyed in speaker 4's voice was important to display his character, not only his voice output. Participant listener 5A stated that the voice sounded like an American version of speaker 5 and then emphasised that this was not his identity. With a less intrusive average voice, that is closer in similarity to the voices being modelled, it is hypothesised that using the same amount of data would produce a voice with a better likeness to the speaker which therefore may be more acceptable to them.

The participant listeners for speakers 3 and 4 noted that there were sections of the output with a strong likeness to the voice of the speaker participants, but this was usually at the syllable level and the US English influence on the voice made it sound disjointed and less like the speakers. For speaker 3, the listeners agreed that those sections that did sound like speaker 3 had captured his voice well. Participant listener 5A noted that the voice with speaker log F0 and mel cepstra sounded somewhere between a generic speech synthesiser voice and the speaker's voice.

6.4.3 Voice output quality

For the evaluation of features contributing to the quality of the output, it was expected that for all speakers the preference would be for the edited data versions. Speaker 4's preference for the unedited data version is difficult to interpret, although the perceptual difference between the two stimuli was small, also accounting for the mixed response by the listeners. For both speakers 3 and 4, a large amount of original data was retained in the editing process, as demonstrated in figure 6.1, where the amount of data selected from the original recordings for use as adaptation data was approximately 93% for speaker 3 and 88% for speaker 4. For speakers with this level of severity, it may therefore not be as essential to carefully select the data to remove extraneous noise or unintelligible sections, than for speakers with more severe dysarthria. Listener 5A's preference for the edited version suggests that this selection of data could contribute to increasing the quality of the synthesised output for a speaker with more severe impairment. Example 6.10 on the CD is an example of speaker 5's voice using unedited data and speaker spectral features and log

F0 (for comparison with example 6.9). All other features were used from the average voice.

The different factors influencing the quality of the voice output were dependent on the individual and the effects of dysarthria on their speech. Where there were large perceptual differences, the voices containing factors that improved the output quality and intelligibility were perceived as best, except when it was perceived as more accurately representing the speech of that individual, seemingly confusing quality with similarity. This occurred for both speaker and listener judgements. Most of the evaluators for speakers 3 and 4 preferred the smoother output provided by the average voice energy information, although listener 4B noted that they could hear more of speaker 4's accent in the version using his own energy, which influenced the judgement. Example 6.11 shows the substitution of the speaker's own energy for speaker 3 (for comparison with example 6.7) and example 6.12 shows the substitution of the speaker's own energy for speaker 4 (for comparison with example 6.8). Listener 5A heard no difference between the speech with the energy taken from the average voice and the speech using the speaker's own energy, this difference could have been masked by the overall lower quality of the output for speaker 5. Speaker 3 preferred the voice where his own durations were used as he identified his own voice clearly in that example. Listeners 3A and 3B noted that that example sounded more like speaker 3 but in their judgements noted that they preferred the example with the average voice durations because it was more intelligible and therefore more suitable for use with a communication aid. This was also true for listener 5A who noted that the speaker's own durations contributed to the identity of the speaker but who preferred the average voice durations because of the perceptual reduction of the impairment in the voice if it was to be used in a communication aid. For speaker 4, the duration information was not very different to his original speech and although both listeners preferred the average voice durations, they did note that the two outputs were very similar. Speaker 4 noted that although for one particular voice, the global variance for spectral information from the average voice made the output clearer, he preferred the voice with his own global variance for spectral information. This output produced a slightly muffled percept but this preference could be related to the perceived softness in the voice quality that it introduced, which speaker 4 noted was missing in other examples. This was also noted by the listeners but they chose the average voice example as they recognised the need for the output to be clear and intelligible. The preferences for the global variance for spectral features across all participants suggests that it positively contributes to the output synthesis quality for these speakers.

As reported in the results of chapter 5, the durational aspect of an individual's speech contributes significantly to the similarity of the synthesised speech to the original speaker. These results mean that to retain speaker characteristics in the output synthesis, the duration distributions of the target speaker should be retained. However, this is a feature which is likely to be affected by the individual's condition and therefore would replicate dysarthric sounding synthesis. The results of the evaluation showed that the durations did contribute to the identification of the speaker although the decision by the listeners in preferring the durations from the average voice, suggest that where the durations severely affect the clarity of the output, the substitution should be made. To compensate for this, an average voice with the same regional accent would ideally be used to impose the durations for the dysarthric speaker's models to capture the duration aspects of the accent of the individual [219]. An individual local donor would not offer that same level of robustness that can be found in the models taken from an average voice. If a choice of regionally-appropriate average voices were available, it would offer a more appropriate set of duration characteristics to more closely replicate the accent of the speaker.

6.4.4 Manipulation of prosodic features

The differences in output rate could be perceived by some of the participants, although there is a limited extent to which the rate can be slowed until it starts to reduce intelligibility, as observed during the production of the stimuli by the author. The rate of output was therefore only slowed slightly, which may not have been sufficient for all participants to observe. Where the difference was perceived, this contributed to the individual preferences along with observing where certain speaker characteristics were more strongly perceived in certain stimuli.

The change of global variance for log F0 could also be perceived by some of the participants. Speaker 4, who had a relatively narrow range of log F0 preferred to have a wider range than his own in the output. Speaker 3's range was closer to the average voice and the preference showed it was more appropriate for him. Where speaker 3 did notice a difference in what he heard for the global variance for log F0 stimuli, he said that the difference was that one was easier to understand than the other. This was supported by the evaluators who also used intelligibility to make their judgements but were also listening more closely to identify bits of speaker 3's accent and the Americanised output. Listener 5A identified a difference in the output of one stimulus, preferring the average voice global variance for log F0. The difficulty in recognising a difference between the stimuli for these parameter changes could be related to the overall quality of the output synthesis for this speaker, although it is difficult to draw conclusions based on the limited amount of results for this speaker. Examples 6.13, 6.14 and 6.15 are examples of the voices built with the speakers' own spectral features, log F0 and global variance for log F0, for comparison with examples 6.7, 6.8 and 6.9 respectively.

6.4.5 Speech reconstruction

In relation to the pathologies of the speakers, both 3 and 4 had variable energy in their speech and both preferred voices with normalised energy output. Speaker 4's monopitch output was reconstructed to have a preferred wider variability in pitch. This factor could also be altered for speaker 5 to widen the log F0 variability found in his speech data. Imprecise articulations present in all speakers' data were handled by using the average voice model durations and global variance for spectral features. Selecting data for adaptation also contributed to the reconstruction quality, particularly for the more severely impaired speech of speaker 5.

6.4.6 Speaker acceptability

Speaker 3's priority seemed to be clarity of output whereas speaker 4 did not want to be represented by a voice which he regarded as sounding nothing like his own and with which he had non-neutral associations. The alternative voice played to the speakers was an example of Acapela's British English male, Peter, which has a standard Southern British English accent. Speaker 4 stated that he would prefer to use that voice rather than any of the voices he had previously heard. Speaker 3 also stated that he would not mind being represented by this voice as long as it was intelligible and clear. Speaker 3 did not appear to associate the voices built using his own data with anything other than himself and therefore was happy to be represented by any clear and intelligible output. All listeners noted the intrusion of the American-sounding average voice as being a negative contribution as it altered the identity of the speaker. The listeners for speakers 3 and 4 particularly regarded accent as being very important to the speakers and rated higher the examples which showed more features closer to their accent and had more evidence of their speaker characteristics. This was sometimes at the expense of the intelligibility of the output.

The reactions to the voices built can show some insight into what is required in voice reconstruction for individuals with speech impairment. Speaker 4's reactions to the voices suggested that for some individuals, if a voice is to be personalised to match that of an individual, then the reconstruction must match that voice very closely to be acceptable to that person. This point of acceptability seems to be different between individuals from this evaluation, although this hypothesis should be more robustly tested with more participants.

6.5 Conclusion

These results point to more success being achieved and better similarity judged if the American influence on the voices was removed. Using a British English average voice would reduce the difference between the speaker characteristics of the average voice and the adaptation data, reducing the amount of discontinuity that was apparent in the synthesised output in the evaluation. This led to the percept of hearing more than one speaker in the voice as noted by all the participant listeners. Speakers with dysarthria find it more difficult to produce the amount of data needed to fully adapt all the characteristics contained in the average voice to their own. The average voice model should therefore contain only neutral associations which will not dominate or intrude on the participant speaker's voice characteristics if there is insufficient data to fully adapt all the models towards the target.

This chapter has shown that using HMM-based synthesis with data selection and imposition of information from the average voice model is a promising technique to reconstruct voices of these individuals with mild to moderate dysarthria.

Chapter 5 showed that spectral features were the most important feature for the accurate prediction of human responses to speaker similarity. The target speech in this chapter was not defined with examples but the results confirmed those of the previous chapter that without the spectral features, there was little or no recognition of the individual's identity in the output speech.

The limited number of participants means that it is difficult to draw conclusions based on quantitative analysis, but generalisations can be made from this data. Having good quality recordings is likely to improve the output synthesis and the more impaired the speech is, the more difficult the process to produce a good quality synthesised voice. Further work should be done to expand on this issue and more accurately define a target population for this process.

It is hypothesised that the use of a more regionally-appropriate British English average voice model would improve this process for this amount of data. Ongoing work with building HTS voices with British English data means that UK average voice models are now available along with multi-accented English speaking average voices [229]. This provides a more appropriate starting point for further work in this area.

These results hold for these speakers only and further work in this area is required to fully test the reconstructive abilities of this technique for people with different pathologies and severity of dysarthria.

In terms of acceptability, this evaluation shows that different people have different priorities for their VOCA use and this highlights the need to provide more choice and more customisation for voices that are provided with communication aids to fit the wants and needs of individuals. The evaluation also provided insight into the importance of some individuals' voices to them as their marker of identity. The reactions within the evaluation suggested that if a voice is said to be personalised to match that of an individual, then the point of acceptability of that voice reconstruction is dependent on the individual. What is also clear is that if the user is not accepting of the voice then they do not want it to represent them, again supporting the case that customisation, choice and adaptation to the individual is important for the acceptability of such devices.

These evaluations do not test whether a personalised voice makes a VOCA more acceptable, it aimed to evaluate the technique itself as a method of personalisation of VOCAs. Once more accurate voices can be built using this technique, then it is left to further work to test the hypothesis that this personalisation increases acceptability and encourages social interaction.

Chapter 7

Conclusions and further work

7.1 Introduction

Initial proof of concept experiments using HMM-based synthesis have shown that this method of synthesis shows promise as a technique for providing personalised synthetic voices for people with speech impairment. The technique can be used with banked speech data pre-deterioration and can be used with speech data of an individual that has begun to show the effects of dysarthria. The work set out in this thesis needs to be placed in the context of what further work needs to be done to assess this technique in wider terms, specifying the population for which this technique is suitable and its limitations for practical use.

This chapter summarises the results of the experiments conducted in the thesis and sets the results in the context of how this work should progress. Further work described aims towards providing a practical tool for clinicians to firstly assess the suitability of building a personalised voice for an individual, depending on their condition and stage of deterioration and then to provide a synthesised voice to suit that user's requirements.

7.2 Conclusions

Chapter 2 discussed the problems associated with using voice output communication aids and the implications of these problems for social interaction. It identified personalising the voice to sound like the VOCA user as a particular issue which needs to be addressed. It is hypothesised that this could be a contributing factor to encourage the user's social interaction. There is also currently no provision for building synthetic voices using dysarthric speech. This chapter also detailed the acoustics of dysarthric speech. A review of the acceptability issues defined the requirements for the output voice, the technique and the person. These were: the output has to be intelligible, natural-sounding, sound like the user before their speech deteriorated and allow access to manipulate the prosodic features; the technique has to be able to build voices with minimal input data and also must be able to deal with the specific acoustic problems associated with dysarthria as set out previously in the chapter; the person must have a state of emotional readiness to be able to accept the technology.

Following the requirements stated above, it was concluded in chapter 3 that HMM-based synthesis was the most suitable technique for personalising synthetic voices. Chapter 4 underlined the potential for using the HTS toolkit to produce high quality synthetic voices for speech data banked before deterioration due to a motor speech disorder. It also proposed a potential approach for dealing with speech data once deterioration had started.

Using speech data pre-deterioration, chapter 5 showed that voices distinct from the average voice could be built with around 100 sentences, although the quality of the output improves when using more data. This has also shown to be dependent on the speaker and the average voice used. The similarity to the target speaker is hypothesised to improve when using a more appropriate average voice model.

Chapter 5 also proposed a method of objectively measuring the similarity between a synthesised and target sentence using a multi-layer perceptron. The technique showed promising results for initial experiments, however, it depends on having pairs of the same sentences and a set of listener responses on which to train the MLP. The experimental results found that combining multiple acoustic features together non-linearly in a similar manner to the MLP, was required for accurate prediction of human responses, rather than using individual features to evaluate similarity of speakers. The experiments found that mel cepstral distortion and fraction of voicing agreement, which is related to the relative durations of segments, were important features contributing to listeners' ratings of speaker similarity.

Chapter 6 showed that HMM-based synthesis could be use for building personalised voices for speech data that had begun to deteriorate. This technique addressed the output acoustics rather than addressing the underlying production problems associated with dysarthria. Two processes were used to reconstruct voices that showed the effects of dysarthria. The first dealt with the data temporally, selecting data to be used for adaptation based on a human listener judgement of intelligibility. The second process approximated adapting only those non-disordered features of the dysarthric speech data from the average voice model to produce a reconstructed personalised voice. The results of chapter 5 inform the work done here by making sure that the vocal identity is preserved by always adapting the spectral information belonging to the participant speaker.

7.3 Contributions

The main contributions of this thesis are as follows:

1. Identifying a technique which is able to provide personalised voices for individuals with dysarthria or who are wishing to bank their voices pre-speech deterioration due to a motor speech disorder.

2. Providing evidence for the applicability of HMM-based synthesis for circumstances which require minimal data input.

3. Providing evidence for the ability of HMM-based synthesis to reconstruct voices and increase intelligibility of those speakers with dysarthria.

4. Providing supporting evidence for the ability of MLPs to be used as an objective listener response predictor for assessing similarity between synthesised and target speech.

5. Providing evidence that an accurate objective assessment of the similarity between synthesised and target speech requires more than just a spectral acoustic measure but could benefit from the non-linear combination of other acoustic features.

7.4 Further work

The thesis provides with these conclusions an awareness of the limitations of the work done. These limitations are discussed below in terms of continuing research in this area.

The further work section describes the research and knowledge required to aim towards providing a practical tool for clinicians to build personalised voices for individuals who want them. This involves assessing the suitability of the individual's voice for such a procedure, taking into account their condition and stage of deterioration and ultimately providing a synthesised voice to suit that user's requirements. This section suggests a preliminary user study to assess the value of using personalised voices for communication aids. Automation of the processes involved in building personalised voices is then discussed, followed by the need to define the target population for this application.

7.4.1 Value assessment of the application

The motivation for providing personalised synthetic voices set out in chapter 2 is based on evidence gathered from literature across disciplines combining together a theoretical view of reconstructing the functions of a voice rather than results derived from empirical research. One of the participants in the experiments himself initiated this line of research, wanting to be able to use a communication aid that would output his own synthesised voice. This shows that there is some demand for this provision. However, there has been no empirical study to assess the value or suitability of personalised voice output communication aids in actual usage. There is no evidence, therefore, to suggest that once this personalised voice is built, the individual will prefer to use it over any other generic voice that they might choose on a communication aid or that this use of a new synthetic voice with their own vocal identity is a positive experience for that individual. The lack of empirical evidence to date can be attributed to the unavailability of a voice building procedure able to do this type of experiment. This situation may change with the development of ModelTalker [28] and contributions made by this study.

This thesis does not attempt to investigate the possible emotional or psychological effects of losing a voice due to a progressive disorder and the effects of using a replacement voice. It highlights the need for an individual to have a voice that they are comfortable using and shows an awareness of potential problems that may arise when using what is deemed an inappropriate voice. It motivates the idea that an individual should have a choice in what voice represents them and that one of these choices should be something that attempts to match their identity. These psychological factors should be taken into consideration for further work. For example, there should be more investigation into the idea of what identity is to be represented by the voice. As discussed in chapter 2 the premise of this thesis is that a voice displays the identity of the individual using it. The personalised voice used in this case is that which represented the individual before their speech deteriorated. However, the dysarthria and condition of the individual now contributes to their identity and to what extent the individual wants or needs to display that part of their identity should also be investigated.

For example, user feedback provides evidence that this personal decision of banking a voice is inextricably linked to complex emotional factors involved in the change of life initiated by a condition leading to speech loss. One individual who was about to undergo a laryngectomy procedure had downloaded the ModelTalker software and was preparing to begin the voice banking recordings when she changed her mind about following the procedure. She felt that her voice was representative of the person she was pre-laryngectomy and that she would prefer to separate that vocal identity from the person she was to become post-laryngectomy and instead use a generic speech synthesiser. The individual reported that thinking about the process involved in preparing for voice banking initiated the realisation and acceptance of the imminent change in her life.

The motivation presented in chapter 2 requires a grounding in empirical evidence in addition to an investigation of the emotional and psychological factors involved in this process of personalisation.

7.4.2 Automation of the procedure

One of the advantages of a data-driven approach such as HMM-based synthesis is the limited human input required. Human input is expensive, time-consuming and can be inconsistent. In the context of providing a toolkit to be used by clinicians or assistive technology specialists, having a more automated voice building procedure would make this technology much more acceptable and time- and cost-effective.

The most time-consuming part of this process is the data selection process. This required up to 10 minutes of editing time per utterance for the most severely dysarthric of the speakers' data. Providing an objective measure of similarity would also reduce the need for human involvement in the evaluation procedure and allow minor procedural changes to have some empirical basis rather than relying on the potentially biased assessment of those involved in the voice building process themselves. These processes are discussed below.

7.4.2.1 Data selection

As discussed in chapter 6, the data selection procedure selects an audio segment for use as adaptation data if that section is firstly a speech sound that can be associated with an appropriate sequence of labels at the syllable or word level and secondly, that it is intelligible to a human listener. The first criterion allows the data to be used by the algorithm and the second ensures that the data used contributes to building a model that synthesises intelligible speech.

One approach to this problem is to provide an interactive labelling tool which assists a human listener to isolate the usable parts of the speech data for adaptation. A line of investigation being followed currently as an MSc thesis project at the University of Sheffield Computer Science Department is to use Gaussian Mixture Models (GMMs) to model and automatically identify usable and non-usable sections of data.

The set of data with identified usable (speech) and non-usable (garbage) sections labelled is used to build a GMM for both categories. This data has been taken from the speakers with dysarthria who participated in this thesis. The aim is to use these models to assign estimated speech or garbage categorisations to new utterances not in the training set. Within an interactive graphical user interface, the user can then alter these labels to more accurately represent the categorisations of the new data. This accurately labelled data can then be used to re-estimate the GMMs and iteratively produce more accurate classification of usable adaptation data, eventually requiring less human intervention.

It is expected that this technique may require a larger number of categorisations, such as introducing a silence model and discriminating between speech and non-speech garbage sounds. A further limitation is that this technique does not take into account speech sounds that are well-articulated but do not match the labelling sequence as defined by the orthographic transcription. This technique may also have to take into account the label sequence as expected by the orthography in some way to make this process more accurate.

Carmichael and Green [33] attempted to address the problem of a human listener's inconsistency in their judgements of intelligibility as they adapt to the speech of an individual over a short period of time. To replicate the behaviour of a listener who has never been exposed to the speech of a particular individual before and can therefore provide a consistent judgement on the intelligibility of that individual's speech, they used a non-adaptive speech recogniser. They assessed the ability of a speaker-independent HMM-based word recogniser to predict the intelligibility of dysarthric speech using forced-alignment likelihood scores. As they note, the success of this technique is reliant on the reference model being an accurate match to the style and accent of the speaker being tested. The reference is therefore problematic in that it requires a regional and socio-economic specific speaker-independent model to provide a more accurate estimation of the likelihood measure. Without such a model, this technique does not distinguish between impairment and accent in its metric. Potential extensions of this technique using syllable or phone-level models may provide a possible way of identifying the more intelligible sections of speech by having an intelligibility measure based on this technique.

Relating dysarthria to the impairment of control and function of the articulatory system, Middag, Martens, van Nuffelen and De Bodt [147] attempted to produce automated analysis of the intelligibility of pathological speech using phonological features. Phonological features are defined as articulatory features derived directly from the waveform. They used pre-defined canonical phonological representations of a known transcription which is force-aligned with the same input taken from a pathological speaker producing an intelligibility measure at the phoneme level. Using an intelligibility measure such as this could contribute to the automation of the data selection process.

Both of the above processes are only applicable for use where the transcription of an utterance is known and the data can be force-aligned accurately. This type of measure would therefore only be able to be applied for the automation of the data selection once the speech-like sections of the dysarthric speech had been identified. This work would also have to incorporate identification of the intelligibility score which provides a reasonable threshold at which to use or reject the data for adaptation.

It remains as further work to see how far using either a phonological representation of speech or a more specific reference model for comparison would be able to distinguish between dysarthric impairment or accent.

In this work, one of the issues has been to try to maximise the use of the data and it has been shown that the more data used, the better the quality of the output voice. Further maximisation of data is possible if the data was selected without requiring the correct articulations from all the component features of a voice. For example, where elements such as energy or correct voicing decisions can be substituted from the average voice model, they are not required to be correctly produced in the data to ultimately produce a good quality synthesised voice. More data can therefore be selected which may contain incorrect values for such features. Using a feature or articulation based technique may allow certain features to not contribute to the intelligibility measure, making a decision solely based on what speaker features can be retained in the output model.

Further work in this area could pursue unsupervised adaptation techniques, where the correct transcription is unknown but provided by a speech recogniser for the adaptation process. Data selection could then be based on a confidence score, a measure of the correctness of the transcribed label. Experiments conducted using unsupervised adaptation techniques for HMM-based synthesis have proven to slightly reduce the quality of the output synthesis for non-disordered data in terms of intelligibility but the levels of naturalness and similarity to target speaker were not as severely affected [113]. This could be attributed to the initial transcription of the labels being incorrect but also being acoustically similar to those of the correct transcription. The parameter sharing that occurs in this technique means that the acoustically similar segments share adaptation transformations and therefore will be adapted in the same way still producing what is a reasonable output.

With this technique, the same issues would arise of having to match the input speech against a speaker-independent model but further investigations with accent or more appropriate region-specific average voice models may make this a viable option, particularly for more severely dysarthric speakers.

7.4.2.2 Objective measure of similarity

Use of an MLP to predict listener responses is a promising objective measure of similarity. As discussed in chapter 5, these experiments require much further work to fully test whether this technique can be robustly used for the assessment of similarity of synthesised and target speech. For example, more data should contribute to the speaker-independent MLP and this should be tested using multiple test speakers to see how well this procedure generalises across speaker data.

7.4.3 Specification of target population

The work reported here deals specifically with three individuals with particular manifestations of their conditions. Dysarthric speech is highly variable across conditions and across speakers and for this reason, further work should provide an opportunity to test the claims made in this thesis for a wider range of individuals with different conditions exhibiting different symptoms of dysarthria. Table 4.1 listed a possible set of symptoms of various types of dysarthria that have proposed solutions using both the data and feature selection technique. These have not been fully tested in the above study and further work should aim to cover the full area of dysarthric symptoms to help identify the target population for which this procedure is appropriate.

7.4.3.1 Severity of dysarthria and stage of deterioration

In addition to the types and symptoms of dysarthria, the severity of dysarthria is perhaps a more important issue to deal with. Further work should aim to determine the relationship between the effectiveness of proposed techniques and the severity of individuals' condition. The data selection technique relies on a certain level of intelligibility to provide adaptation data as used in the proposed way. For more severe speakers, potential unsupervised techniques could prove to be useful to provide more data for adaptation where the severity limits the amount of intelligible data. Work could also be done to design more appropriate data for collection, which may be easier for the individuals to produce. Depending on the condition, this could be words or phrases that contain less complex articulations, potentially tailoring the data set to firstly assess the availability of the individual's articulatory repertoire and then asking them to produce a certain set of data dependent on that. This could be achieved in combination with the intelligibility tests that are already part of speech and language therapy procedures, such as the Frenchay Dysarthria Assessment (FDA) [66].

One of the requirements set out in chapter 2 was that the individual had to have a level of emotional readiness to participate in a voice building process. If a stage in the deterioration can be identified from which it is more difficult to build a successful voice then a better awareness can be reached for what options are available to the speaker at the time they become ready to deal with this issue.

7.4.3.2 Average voice

As was shown in chapters 5 and 6, the average voice makes a contribution to the similarity of the output voice synthesis to the target speaker. The experiments in this thesis used an average voice previously built using a majority North American database of speakers. The results of the experiments showed that this characteristic of the average voice intruded into the personalised voices when adapting with the smaller amounts of data. A more regionally-appropriate average voice is hypothesised to produce better results for this type of procedure. Further work could follow this line of research, testing this hypothesis and determining how specific the average voice should be to the participant in question for an optimal personalisation, balancing quality of voice with the practical requirements of building average voice models. This process should then attempt to provide a way to automatically determine which of the average voice models is the most appropriate to use for a particular speaker.

7.4.3.3 Feature selection

As previously discussed, the feature selection technique can be investigated further using more participants with varying degrees of severity and combinations of dysarthric symptoms. Having a mapping between identifiable symptoms and which features to select for adaptation would provide a procedure for clinicians or assistive technologists to apply to build an optimal personalised voice for a client dependent on the diagnosis of their condition.

In addition to this, the customisable features of global variance for F0 and altering the rate of utterance output should be investigated in terms of appropriateness for the speaker and also in terms of the influence on intelligibility, naturalness and similarity measures of the voice. This could provide more guidelines for what values of these features are appropriate for that speaker dependent on various factors, including taking input from the speakers themselves to make these decisions.

7.5 Summary

This chapter has reiterated the conclusions reached and contributions made to the field during this thesis. It has provided areas of further work that look towards developing this technique as a toolkit for clinicians or assistive technologists to use to provide this service for those individuals who are in the position to bank their voice for future use as a voice prosthesis.

Appendix A

Speech production theory

The aim of this appendix is to introduce and compare the existing theories of speech production which are relevant to the thesis.

A.1 Classical phonetics and phonology

Classical phonetic theory makes the assumption that speech is composed of a sequence of discrete sounds or *phonetic segments.* Given this basic assumption, the theory then attempts to describe and classify phonetic segments in terms of their physical properties, both acoustic and articulatory. Further, the theory attempts to explain acoustic phenomena relating to groups of segments, the *prosody* of the language.

Classical phonetics concerns itself with the organisation and functionality of the underlying sound system in a language, termed the *phonology.* One aspect of phonology is the definition of what sounds exist as the finite set of underlying segments in the language termed *phonemes,* of which there are approximately 46 in British English, for example, depending on accent [103, 220]. The phoneme is defined as the smallest unit which distinguishes meaning in the language and from this basic unit, acoustic realisations or *phones* are derived. Phonology attempts to explain observations related to the grouping and sequencing of units at the sub-phonemic or above level and the representation of these units for conveying linguistic meaning.

A.1.1 International phonetic alphabet

One of the main contributions of classical phonetics was the provision of a formal classification of the phonetic segments of any language based upon the properties of the sound.

This classification system is known as the international phonetic alphabet (IPA). This system identifies common articulatory properties of sounds, for example, place and manner of articulation in the case of consonants, and categorises each individual sound as a unique combination of such properties.

A.1.2 Limitations

A major limitation of classical phonetics is that it is based upon the false assumption that speech is a concatenation of discrete segments. It has been shown that speech is a continuous signal, a consequence of the continuousness of the articulators which create the signal.

Further limitations include a lack of explanation of several phenomena, including why certain sounds are observed in natural speech and others are not. Classical phonetic theory also fails to explain constraints upon the acceptable sequences of sounds and provides little insight into observed relationships between phonetic segments, in particular coarticulation effects (see section A.2).

Despite some attention being paid to phonology, the scope of classical phonetic theories is mainly limited to an analysis of the surface realisation of sounds, either at the articulatory level (articulatory phonetics) or the acoustic level (acoustic phonetics). While this analysis is useful, classical phonetics makes no attempt to explain how speech is produced, unlike the speech production models and theories described below.

A.2 Coarticulation theory

Coarticulation is the influence of a phonetic segment upon its neighbouring segments. An example is the observed nasalisation of the start of the vowel in the word 'mad' due to the preceding nasal consonant.

The idea of coarticulation presupposes that each phonetic segment has a corresponding *target* specification (a set of acoustic and/or articulatory features). The occurrence of coarticulation in a segment is indicated by a difference between the target features and those observed. Like classical phonetic theory, the definition of coarticulation assumes that speech is composed of a sequence of discrete segments.

Models of coarticulation attempt to describe a mapping from a sequence of discrete symbols (representing the hypothesised discrete segments of speech, sometimes referred to as the utterance plan or cognitive specification of the planned sequence) to a continuous acoustic waveform or articulatory contour. Whereas classical phonetics focussed on the perception of speech, coarticulation theory takes a speech production approach, particularly looking at the more abstract planning stages of production. Coarticulation models focus on different aspects of the mapping between the cognitive plan for the utterance as it exists in the brain and the final acoustic realisation, for example:

- the units of the utterance plan.
- the relationship between the utterance plan and a physical representation of the utterance used as input to the physical speech processes (motor control system and articulators).
- the realisation of the acoustic signal via the motor control system and articulators, particularly the physical movements from one target specification to the next in the linear sequence.

An example of a model of coarticulation is target theory, described below.

A.2.1 Target theory

Target theory [139] assumes that the utterance plan is, at some point in the speech production process, translated to a sequence of physically-specified targets. This is labelled as the *spatial representation* in the graphical depiction of target theory in figure A.1.

The target theory model uses the phoneme as the unit of speech production, therefore, as with classical phonetics and coarticulation theory, taking a segmental approach to speech production. Each phoneme is mapped to a set of 3-dimensional articulator positions within the vocal tract, called the ***mental map*** of the articulatory space. This mental map is then translated into neuromotor system commands which in turn effect the movement of articulators between targets. Coarticulation is incorporated into the neuromotor system commands via the ***gamma motor system*** [143]. Note that, since coarticulation is specified prior to the execution of neuromotor system commands, the target theory model claims that the cognitive part of speech production is responsible for coarticulation effects. This contrasts clearly with the action theory model (see section A.3).

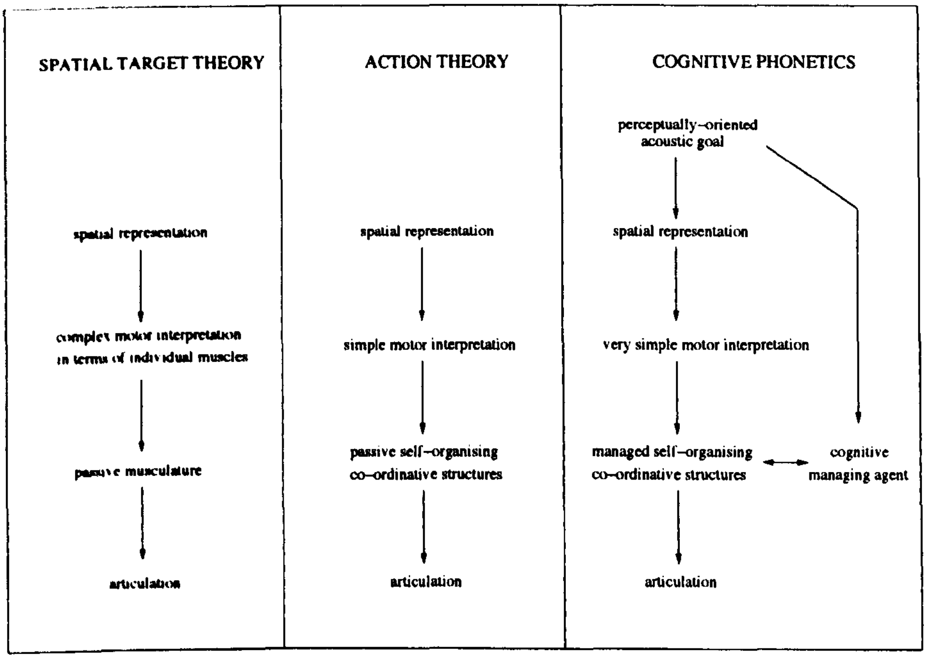

SPATIAL TARGET THEORY
spatial representation
complex motor interpretation in terms of individual muscles
passive musculature
articulation
ACTION THEORY
spatial representation
simple motor interpretation
passive self-organising co-ordinative structures
articulation
COGNITIVE PHONETICS
perceptually-oriented acoustic goal
spatial representation
very simple motor interpretation
managed self-organising co-ordinative structures
cognitive managing agent
articulation

Figure A.1: *Alternative models of speech production (from [195]).*

A.3 Action theory

Action theory [70] proposes that the cognitive level of speech production has no knowledge of highly detailed speech representations, for example, large sets of articulatory features. The cognitive processes, instead, consist of simpler instructions (i.e. spatial targets). These broad instructions are then executed via motor commands by a relatively knowledgeable articulatory system.

It is claimed that the target theory model places too much complexity upon the cognitive component of speech production to be plausible. Instead, it is said that the muscles involved in articulation are arranged in a *co-ordinative structure* which encodes working relationships between the muscles (see figure A.1). This co-ordinative structure is pre-programmed such that it is capable of interpreting and executing the details of the relatively detail-free instructions issued by the cognitive system.

The co-ordinative structures in action theory suggest that vowels and consonants have different modes of articulation and it is the vowels which represent an underlying continuous structure upon which consonantal gestures are imposed. For example, the production of a consonant will be affected by the demands placed on the articulators to form the following vowel. This view suggests that during speech production, it is not the properties of the gestures that change with context but that it is the temporal overlap with other gestures that causes the variability in the output. Coarticulation, in this theory of speech production, is therefore the consequence of the dynamic properties of the articulators moving in co-ordination.

A.4 Articulatory phonology

An *articulatory gesture* is a specification of an articulatory event, specified as a set of abstract articulatory parameters. A *gestural score* details the variation of these abstract articulatory parameters in time, in much the same way as musical scores specify how musical instruments (played in parallel) vary in time.

Articulatory phonology [26] attempted to bring together phonetics and phonology by the idea that the constraints of the physical articulatory system act as a basis of the phonological system. The unit of control at the planning level is the same as that used at the level of production, the articulatory gesture. The gesture is a discrete unit of the phonology but the gestural score relates the organisation of the articulations together in an overall plan.

One of the main contribution of articulatory phonology is the introduction of a parametric model for phonetic realisation, including temporal specifications and sequencing (i.e. start and end times of events). This model, which outputs a gestural score, a set of tiers representing the end articulator that executes the gesture, predicts temporally overlapping articulatory features, explaining different types of phonological variation including allophonic variation and coarticulation, and agrees with the action theory interpretation of coarticulation [70].

A.5 Task dynamics

Like action theory, the theory of task dynamics [178] is a physical model which focusses on the output of articulatory trajectories. The gestural score used in articulatory phonology acts as input to the task dynamic model. It has two functionally distinct levels: the interarticulator level, which defines which articulator is moving and on which dimension it is moving, and the intergestural level, which is defined by a geometric co-ordinate system. The intergestural level deals with the strength of the articulation and the co-ordination between articulators at a point in time.

Gestures have a functional task, as an underlying gestural plan. The task is achieved using co-ordinative structures which are groups of articulators or the musculature involved in the physical control of the articulators. It is the task itself which is focussed on in this theory of production rather than the individual articulators used to fulfil the tasks.

Task dynamics has been coupled with articulatory phonology (see section A.4) to create a speech synthesis engine (the CASY configurable articulatory synthesiser, [175]).

A.6 Cognitive phonetics

In articulatory phonology, constraints on physical processes are known and used before the utterance plan, or gestural score is set out. The knowledge of these constraints is therefore unchangeable information. The *cognitive phonetic* model [193, 195] suggests that there are some constraints that are static in nature and unchangeable, but that there are also optionally controllable constraints which are changeable throughout the speech production process.

Action theory (see section A.3), depends on the idea that phonetic objects have some internal inherent physical properties that dictate its phonetic realisation. The cognitive phonetic model extends this idea that the phonetic object also has some inherent cognitive properties that contribute to the realisation. For example, using feedback to control the optional variability in the utterance plan.

Several types of feedback exist during speech production, which are listed below.

- Auditory feedback. The auditory and cognitive systems process the speech signal produced during speech production.
- Tactile feedback. Pressure sensors within the vocal tract deliver information back to the brain during speech production. For example, the sensation of the tongue touching the teeth.
- Intramuscular feedback. The muscles involved in speech production inter-communicate using the relevant neural pathways and the spinal cord. The muscles also feedback to the cognitive system via the spinal cord.

While some of the above theories of speech production have incorporated information feedback within their component processes, for example the physical co-ordinative structure used in action theory, none have proposed feedback to the cognitive component of speech production. One of the main features which differentiates the cognitive phonetic speech production model from those mentioned above is the inclusion of a cognitive phonetic agent (CPA) which supervises the speech production process. The CPA is the *cognitive managing agent* in figure A.1, which contrasts the cognitive phonetic model with the action theory and target theory models of speech production.

The CPA predicts articulator (or abstract spatial) positions, acoustics and perceptual outcomes (for example, a listener response) during the speech production process. It then uses these predictions to modify the speech production process. The CPA may then be viewed as a model of auditory and muscular feedback, as discussed above. By integrating listener behaviour information into the process, the cognitive phonetic model goes further than the previously mentioned models of speech production.

Appendix B

HMM-based speech synthesis: HTS further details

B.1 Hidden Markov Models

Hidden Markov Models can be used to probabilistically model sequences of *feature vectors*: a compact representation of speech characterising the acoustics of the signal. HMMs are not only able to successfully characterise sequences of feature vectors, as exploited in the field of speech recognition, but they are generative models and can therefore generate feature vectors dependent on the probabilistic modelling, from which speech waveforms can be synthesised.

An HMM models a stochastic process, for example, speech. The temporal variation of speech is modelled with a Markov chain of states with associated transition probabilities between these states. Associated with each state is a statistical model of the acoustics of a particular segment of speech. This model is usually a continuous probability distribution. Figure B.1 shows a diagrammatic representation of an HMM, where the circles represent states and arrows represent transitions with associated transition probabilities. To estimate this statistical representation, a training process is performed. The model is exposed to multiple examples of the unit being modelled and its parameters are re-estimated such that the likelihood of the model, given the examples, is maximised.

The different states capture subphonetic temporal variation. There should be enough states in an HMM to capture sufficient detail to model the sequence accurately while still accounting for natural variation in the acoustics. The unit modelled in HTS is the context-dependent phone, phone-sized units with contextual information and is modelled by five

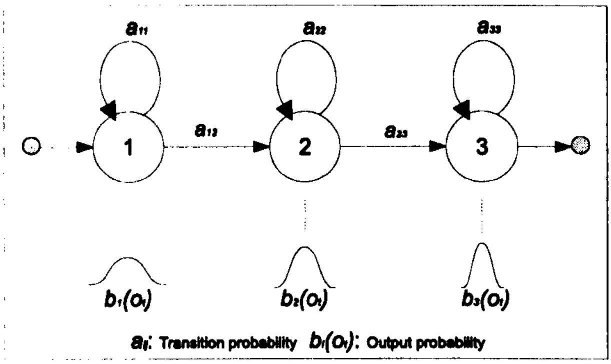


Figure B.1: *Hidden Markov Model. Emitting states are represented by circles and transitions are represented by arrows. There is a transition probability (a) associated with every transition and a Gaussian output probability (b) associated with every state.*

emitting states. This relatively high number of states allows acoustic information to be captured with high temporal resolution.

HMM transition probabilities do not provide an accurate model for duration. A geometric duration distribution is implied by standard HMMs, which is a poor model of actual phone durations. To combat this problem, HTS estimates a normally distributed state duration probability density for each state in each model during training, which is explicitly attached to the model for both training and synthesis. This alters some of the mathematical properties of the model and results in a *Hidden Semi-Markov Model* (HSMM) [237] as shown in figure B.2. The training corpus is used to estimate the parameters of the duration model.

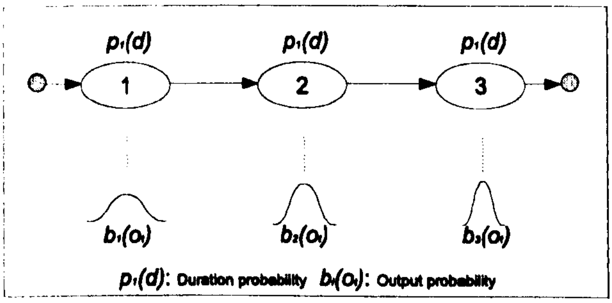


Figure B.2: *Hidden Semi-Markov Model (HSMM). Explicit duration probabilities (p) replace transition probabilities with a single component Gaussian output distribution (b) over the number of time frames spent in each state.*

In order to model speech with HMMs, assumptions have to be made to simplify the probability calculations. The conditional independence assumption states that, given a particular state, there is no dependency between previous and following feature vectors. This does not accurately represent the behaviour of the articulators whose configuration at one time-frame is highly dependent on their configuration at the previous and following time-frames. To compensate for this modelling deficiency, extra features are introduced into the feature vector which measure the rate of change of the static observations, called *deltas*, sometimes called *velocities*, and *delta-deltas*, sometimes called *accelerations*, which capture the rate of change in the deltas [72].

Unlike parametric synthesis, this data-driven technique does not demand human intervention for tuning any synthesis parameters; the variation is captured in the corpus of data on which the models are trained. Using HSMMs also creates the opportunity to use speaker adaptation techniques to personalise the voice of such a system, adapting from existing speaker-independent models towards the target speaker with a small amount of data.

B.2 HMM-based synthesis features

HTS uses STRAIGHT [110] vocoding to both extract features and resynthesise the waveform. Vocoding is the process of extracting features of speech that are perceptually relevant and then using those features to reconstruct the speech. STRAIGHT mel cepstra are used to capture the spectral information as they can also be used to reconstruct the waveform at the synthesis stage. The feature vectors are 138-dimensional and comprise separate streams: spectral features including energy, log F0 and band aperiodicity, which are described below.

B.2.1 Spectral features

The mel cepstral coefficients make up the spectral representation stream of the feature vector. As the number of coefficients increases, the detail captured in the representation becomes finer. Version 2.1 of HTS uses 40 STRAIGHT mel cepstra, including the zeroth coefficient, which is the overall energy captured in a frame.

These 40 coefficients only represent the static elements of the signal. It is necessary to capture the dynamics of speech to more accurately reconstruct the signal. The spectral stream of the feature vector is therefore 120-dimensional, consisting of 40 STRAIGHT mel cepstra (including energy), their deltas and delta-deltas.

B.2.2 Log F0

To produce an accurate output, the F0 must be accurately extracted. Pitch extraction algorithms are prone to producing errors where the F0 value is estimated as half or double the actual value. To minimise these errors, HTS extracts F0 from the database using three different pitch extraction algorithms: tempo [109], if_getf0 [7] and ESPS get_f0 [67]. The F0 is initially extracted within a defined wide fixed range. After visual inspection of the range of results, a more restrictive speaker-specific range is passed to the algorithms. The median of the three extracted values is selected per frame to ensure an accurate extraction with minimal halving or doubling error. This procedure relies on a maximum of one of the three algorithms producing an error in a frame to still get an accurate extraction.

The F0 is modelled by a *Multi-Space Probability Distribution* (MSD) [207]. Modelling F0 with HMMs is difficult due to the occurrence of both voiced and voiceless sounds in a speech sequence. Voiced sounds can be represented by continuous values but voicelessness cannot be modelled in the same way as there is no value for the fundamental frequency of voiceless sounds. The HTS system uses a multi-space probability distribution which allows a representation of the entire fundamental frequency sequence using two separate probability spaces, where voiced speech uses a continuous representation and voiceless sections use a zero-dimensional discrete symbol. The F0 stream of the feature vector consists of three dimensions: log F0, its delta and delta-delta.

B.2.3 Aperiodicity

A signal is rarely completely periodic and even in a voiced sound, aperiodicity is likely to occur at high frequencies due to breath moving through the glottis or other turbulence occurring in the vocal tract. The F0 stream defines whether a frame of speech is voiced or voiceless but the aperiodicity stream reflects the aperiodicity across different frequency bands in the frame. A measure of aperiodicity extracted by STRAIGHT contributes to an improved modelling of the excitation source and leads to a better quality output [167].

The aperiodicity value represents the relative energy of aperiodic components in the signal in different frequency bands: 0-1, 1-2, 2-4, 4-6 and 6-8 kiloHertz (kHz). The aperiodicity measure is the value of the upper envelope of the liftered power spectrum subtracted from the value of the lower envelope, which is then normalised. The upper envelope of the spectrum is defined by the overall resonant frequencies of the vocal tract at that point in time. The lower envelope of the spectrum at the troughs of the excitation source represents the aperiodic noise component which is also shaped by the same slowly changing vocal tract movements (see figure B.3).

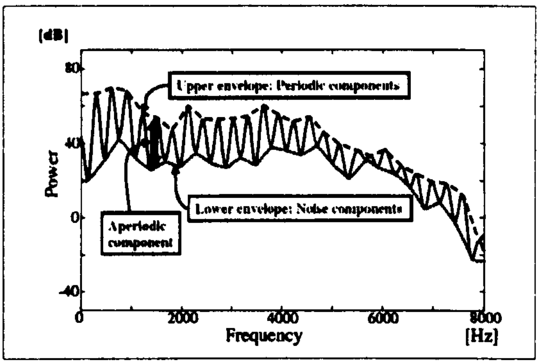


Figure B.3: *The aperiodicity component is extracted from the upper and lower envelope of the liftered power spectrum (figure taken from [167])*

The aperiodicity measure shows the relationship between these two components. As the distance between the upper and lower envelope increases the aperiodicity measure is minimised and as they get closer together, the aperiodicity increases. For each frequency band, the aperiodicity measure is extracted and calculated. The aperiodicity stream of the feature vector is 15-dimensional: 5 values for each frequency band, their deltas and delta-deltas.

B.3 Context-dependent modelling for HMM-based synthesis

The list of features included in the labelling of context-dependent phoneme models is as follows:

- Phoneme level:
  - current phoneme
  - preceding and following two phonemes
  - position of current phoneme in the syllable
- Syllable level:
  - number of phonemes in preceding, current and following syllable

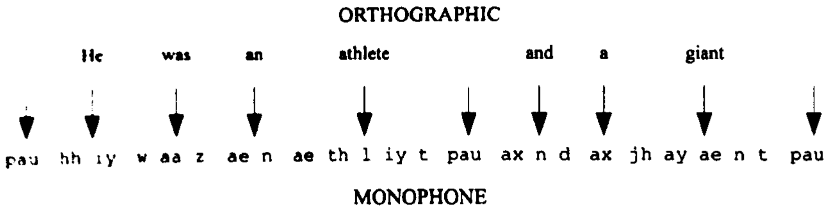
  - intonational accent of preceding, current and following syllable (as predicted by a CART model (see section 3.3.3.8)
  - lexical stress of preceding, current and following syllable (taken from the encompassing word entry in the lexicon)
  - position of current syllable in current word and phrase
  - number of preceding and following stressed syllables in current phrase
  - number of preceding and following accented syllables in current phrase
  - number of syllables from previous and to next stressed syllable
  - number of syllables from previous and to next accented syllable
  - vowel identity within current syllable

- Word level:
  - guessed part of speech of preceding, current and following word
  - number of syllables in preceding, current and following word
  - position of current word in current phrase
  - number of preceding and following content words in current phrase
  - number of words from previous and to next content word

- Phrase level:
  - number of syllables in preceding, current and following phrase
  - position in major phrase
  - ToBI (Tones and Break Indices) endtone of current phrase

- Utterance level:
  - number of syllables, words and phrases in current utterance

This information is extracted from the orthographic transcription of the input data using the text analysis component of Festival [23]. An example of the labelling from an utterance to an HTS label format is shown in figure B.4. The coding system is available below in section B.4 for reference. The labelling relies on the orthographic transcription of the data accurately matching the speech signal.



<table>
<tr><th>Orthographic Transcription</th><th>Monophone Transcription</th><th>Label</th></tr>
<tr><td></td><td>pau</td><td>x^x-pau+hh=iy@x_x/A:0_0_0/B:x-x-x@x-x&x-x#x-x$x-x!x-x;x-x|x/C:1+1+2/D:0_0/E:x+x@x+x&x+x#x+x/F:content_1/G:0_0/H:x=x@1=2|0/I:5=4/J:9+7-2</td></tr>
<tr><td>He</td><td>hh</td><td>x^pau-hh+iy=w@1_2/A:0_0_0/B:1-1-2@1-1&1-5#1-5$1-2!0-1;0-4|iy/C:1+0+3/D:0_0/E:content+1@1+4&1+1#0+3/F:aux_1/G:0_0/H:5=4@1=2|L-L%/I:4=3/J:9+7-2</td></tr>
<tr><td></td><td>iy</td><td>pau^hh-iy+w=aa@2_1/A:0_0_0/B:1-1-2@1-1&1-5#1-5$1-2!0-1;0-4|iy/C:1+0+3/D:0_0/E:content+1@1+4&1+1#0+3/F:aux_1/G:0_0/H:5=4@1=2|L-L%/I:4=3/J:9+7-2</td></tr>
<tr><td>was</td><td>w</td><td>hh^iy-w+aa=z@1_3/A:1_1_2/B:1-0-3@1-1&2-4#1-4$1-2!1-1;1-3|aa/C:1+0+2/D:content_1/E:aux+1@2+3&2+1#1+2/F:det_1/G:0_0/H:5=4@1=2|L-L%/I:4=3/J:9+7-2</td></tr>
<tr><td></td><td>aa</td><td>iy^w-aa+z=ae@2_2/A:1_1_2/B:1-0-3@1-1&2-4#1-4$1-2!1-1;1-3|aa/C:1+0+2/D:content_1/E:aux+1@2+3&2+1#1+2/F:det_1/G:0_0/H:5=4@1=2|L-L%/I:4=3/J:9+7-2</td></tr>
<tr><td></td><td>z</td><td>w^aa-z+ae=n@3_1/A:1_1_2/B:1-0-3@1-1&2-4#1-4$1-2!1-1;1-3|aa/C:1+0+2/D:content_1/E:aux+1@2+3&2+1#1+2/F:det_1/G:0_0/H:5=4@1=2|L-L%/I:4=3/J:9+7-2</td></tr>
</table>

...

Figure B.4: *The conversion of utterance level orthographic transcription to the phonetic and prosodic context-dependent labelling (see section B.4 for the label format).*

B.4 Label file format

Context-dependent label format for HMM-based synthesis for English, taken from lab_format.pdf (as part of HTS installation).

$$p_1\hat{}p_2 - p_3 + p_4 = p_5@p_6 - p_7/\mathrm{A}{:}a_1\_a_2\_a_3$$
$$/\mathrm{B}{:}b_1 - b_2 - b_3@b_4 - b_5\&b_6 - b_7\#b_8 - b_9\$b_{10} - b_{11}!b_{12} - b_{13};b_{14} - b_{15} \mid b_{16}$$
$$/\mathrm{C}{:}c_1 + c_2 + c_3/\mathrm{D}{:}d_1\_d_2/\mathrm{E}{:}e_1 + e_2@e_3 + e_4\&e_5 + e_6\#e_7 + e_8$$
$$/\mathrm{F}{:}f_1\_f_2/\mathrm{G}{:}g_1\_g_2/\mathrm{H}{:}h_1 = h_2@h_3 = h_4 \mid h_5/\mathrm{I}{:}i_1\_i_2/\mathrm{J}{:}j_1 + j_2 - j_3$$

p_1	identity of phoneme before the previous phoneme
p_2	previous phoneme identity
p_3	current phoneme identity
p_4	following phoneme identity
p_5	identity of phoneme after the following phoneme
p_6	position of current phoneme in the current syllable (forward)
p_7	position of current phoneme in the current syllable (backward)
a_1	whether the previous syllable is stressed or not (0: not stressed, 1: stressed)
a_2	whether the previous syllable is accented or not (0: not accented, 1: accented)
a_3	number of phonemes in the previous syllable
b_1	whether the current syllable is stressed or not (0: not stressed, 1: stressed)
b_2	whether the current syllable is accented or not (0: not accented, 1: accented)
b_3	number of phonemes in the current syllable
b_4	position of the current syllable in the current word (forward)
b_5	position of the current syllable in the current word (backward)
b_6	position of the current syllable in the current phrase (forward)
b_7	position of the current syllable in the current phrase (backward)
b_8	number of stressed syllables before the current syllable in the current phrase
b_9	number of stressed syllables after the current syllable in the current phrase
b_{10}	number of accented syllables before the current syllable in the current phrase
b_{11}	number of accented syllables after the current syllable in the current phrase
b_{12}	number of syllables from the previous stressed syllable to the current syllable
b_{13}	number of syllables from the current syllable to the next stressed syllable
b_{14}	number of syllables from the previous accented syllable to the current syllable
b_{15}	number of syllables from the current syllable to the next accented syllable
b_{16}	name of the vowel of the current syllable

c_1	whether the next syllable is stressed or not (0: not stressed, 1: stressed)
c_2	whether the next syllable is accented or not (0: not accented, 1: accented)
c_3	number of phonemes in the next syllable
d_1	gpos (guess part-of-speech) of the previous word
d_2	number of syllables in the previous word
e_1	gpos (guess part-of-speech) of the current word
e_2	number of syllables in the current word
e_3	position of the current word in the current phrase (forward)
e_4	position of the current word in the current phrase (backward)
e_5	number of content words before the current word in the current phrase
e_6	number of content words after the current word in the current phrase
e_7	number of words from the previous content word to the current word
e_8	number of words from the current word to the next content word
f_1	gpos (guess part-of-speech) of the next word
f_2	number of syllables in the next word
g_1	number of syllables in the previous phrase
g_2	number of words in the previous phrase
h_1	number of syllables in the current phrase
h_2	number of words in the current phrase
h_3	position of the current phrase in utterance (forward)
h_4	position of the current phrase in utterance (backward)
h_5	TOBI endtone of the current phrase
i_1	number of syllables in the next phrase
i_2	number of words in the next phrase
j_1	number of syllables in the utterance
j_2	number of words in the utterance
j_3	number of phrases in the utterance

B.5 Average voice building and speaker adaptation

For speech recognition tasks, the starting point for adaptation is a speaker-independent model. It is built from large amounts of data taken from multiple speakers and therefore provides a reliable, robustly-estimated model of the general characteristics of speech. This average voice model provides a well-informed prior probability distribution for the target speaker model. Use of this prior information enables robust estimation of the target speaker model when using a small amount of adaptation data.

In building an average voice model, speaker- and gender-dependent characteristics in the data are neutralised. The aim is to get a robust model of speech that captures the phonetic variation, not a model of inter-speaker variation. This is particularly important when using minimal training data where the average voice statistics could be skewed by the balance of speakers in the database. The inter-speaker variance produces a wider variance in the speaker-independent models and the acoustic characteristics of different phonemes becomes less well-defined.

This problem is overcome by using speaker adaptive training (SAT) for parameter re-estimation. The SAT framework [5] ensures that the acoustic variation due to the speaker population is reduced when estimating the variance of the acoustic model parameters. Speaker adaptive training separates the inter-speaker variation from the phonetic variation by estimating an affine transform of the speaker-independent model (estimated using all the data) using the data for each speaker. These transforms represent the characteristics of the individual speakers and therefore reduce inter-speaker variance when re-estimating speaker-independent acoustic model parameters.

B.5.1 Adaptation from the average voice

Adaptation from the average voice model towards the target speaker is done using a combination of constrained structural maximum a posteriori linear regression (CSMAPLR) and maximum a posteriori (MAP) techniques. This was empirically determined as the most successful technique using both objective and subjective measures to evaluate the distance between the original target speech and the synthesised output after adaptation [161].

CSMAPLR is derived from a combination of more standard adaptation techniques, the structure of which is illustrated in figure B.5 and described below.

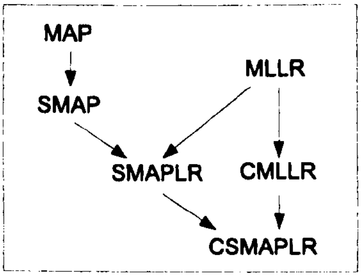


Figure B.5: *Diagram showing how CSMAPLR is built up from standard adaptation techniques*

Maximum likelihood linear regression (MLLR) [123] is where a linear transformation of the speaker-independent model is estimated to maximise the model likelihood, given the adaptation data. This method takes advantage of linear relationships between sounds, and uses this to adapt across related classes of sounds. For MLLR, the transformation of the means of the state output distributions is calculated. Constrained maximum likelihood linear regression (CMLLR) [73] simultaneously estimates the transformation of the variance of the state output distributions as well as the transformation of the means. This is important for all factors to more closely model the characteristics of the speech of individual speakers [161].

MAP is where prior information, the average voice model parameters in this case, is combined with the new observed data in a weighted sum to provide new parameter estimates of the model. Only those models that are observed in the adaptation data are adapted. Structured maximum a posteriori (SMAP) extends this technique and makes use of the context decision tree structure containing a hierarchical organisation of the distributions with a prior probability distribution determined at every node for its child cluster. The further extension to structured maximum a posteriori linear regression (SMAPLR) [182] is where the SMAP concept of structural organisation is used for adaptation but it is performed on the transformation matrices themselves that are used in MLLR, rather than the model parameters.

CSMAPLR combines CMLLR and SMAPLR techniques to improve robustness of parameter estimation and avoid overfitting to the adaptation data. Integrating CMLLR and SMAPLR therefore allows access to the advantages that they both provide: the use of the structured information made available via the context-dependent decisions trees and a way to adapt both the means and variances of the features.

An additional level of MAP adaptation is applied after the CSMAPLR stage to those clusters where there is enough speech data available to robustly re-estimate the model parameters.

Appendix C

Test set sentences

Test sentences extracted from Arctic set A, used in chapter 5 experiments.

arctic a0102 He will follow us soon.
arctic a0112 He was wounded in the arm.
arctic a0163 Philip made no effort to follow.
arctic a0183 And the air was growing chilly.
arctic a0192 He did not rush in.
arctic a0193 It was edged with ice.
arctic a0195 But a strange thing happened.
arctic a0205 From now on we're pals.
arctic a0287 Keep an eye on him.
arctic a0290 One by one the boys were captured.
arctic a0317 He was a wise hyena.
arctic a0333 This is no place for you.
arctic a0346 Get down and dig in.
arctic a0351 It was more like sugar.
arctic a0390 I'll go over tomorrow afternoon.
arctic a0399 And here's another idea.
arctic a0419 The Portuguese boy passed the Hawaiian.
arctic a0443 He was worth nothing to the world.
arctic a0473 The night was calm and snowy.
arctic a0489 They were artists, not biologists.
arctic a0498 The lines were now very taut.

arctic a0504 He was an athlete and a giant.

arctic a0578 But we'll just postpone this.

arctic a0580 This is my fifth voyage.

arctic a0586 We don't see ourselves as foolish.

Appendix D

Protocol for data selection process

Protocol:

1. An intelligible syllable is the minimum size of section to be extracted although a word is preferable. You can be more certain about how well something is articulated when there is more of it to listen to. This is also an easier size to select from the rest of the speech.
2. Only silences of length greater than 0.2 second are defined as pauses.
3. If there is a pau marker in the label file then that section can be taken for the speaker, but this is the only case where a pause should occur in a section.
4. Only segments that are surrounded by the expected segments (or a pause) should be selected. This includes coarticulation segments.
5. Cut off any audible breathiness at the end of a word or segment if it is easy and clear to do so.
6. Cut off any previous noise before the articulation/onset of the word.
7. Cut off any coarticulation fragments that occur at the end of words.

Appendix E

Test set for evaluation of voices built with dysarthric data

Test sentences extracted from SCRIBE [95] set, used in chapter 6 experiments.

Paragraphs used in questions 1 and 2:

- When a sailor in a small craft faces the might of the vast Atlantic Ocean today, he takes the same risks that generations took before him. But, in contrast to them, he can meet any emergency that comes his way with a confidence that stems from a profound trust in the advances of science.
- Boats are stronger and more stable, protecting against undue exposure; instruments are more accurate and more reliable, helping in all weather and conditions; food and drink are better researched and easier to cook than ever before.
- We have no means of measuring, of course, but the truth is, none of the commanders of the ships which accompanied Francis Drake are remembered today - no more than the type of sail, the make of radio or navigation instrument supplied to our modern adventurers will be remembered in four hundred years time.

Sentences used in question 3:

- 013. The government triumphed four years ago and we have every reason to believe that it will triumph again.
- 019. We have proof that the regime wields sufficient power in the North to exploit the entire population.

- 056. Doctor Philips raised a number of points about the professor's article in the recent journal.
- 063. Clara went through a phase when she always served Hungarian goulash followed by rhubarb crumble.
- 088. She had scarcely divulged the scandal before it was splattered over the front pages of the tabloids.

Bibliography

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara. Voice conversion through vector quantisation. In *Proceedings of ICASSP*, pages 655–658, 1988. New York: NY, USA.

[2] A. J. Abrantes, J. S. Marques, and I. M. Trancoso. Hybrid sinusoidal modelling of speech without voicing decision. In *Proceedings of Eurospeech*, pages 231–234, 1991. Genova, Italy.

[3] J. Allen. Designing desirability in an augmentative and alternative communication device. *Universal Access in the Information Society*, 4:135–145, 2005.

[4] J. Allen, M. S. Hunnicutt, and D. Klatt. *From Text to Speech: the MITalk System.* Cambridge: Cambridge University Press, 1987. with Robert C. Armstrong and David Pisoni.

[5] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul. A compact model for speaker adaptive training. In *Proceedings of ICSLP*, pages 1137–1140, 1996. Philadelphia: PA, USA.

[6] D. H. Angelo, S. M. Kokosa, and S. D. Jones. Family perspective on augmentative and alternative communication: families of adolescents and young adults. *Augmentative and Alternative Communication*, 12(1):13–20, 1996.

[7] D. Arifianto, T. Tanaka, T. Masuko, and T. Kobayashi. Robust F0 estimation of speech signal using harmonicity measure based on instantaneous frequency. *IEICE Transactions on Information and Systems*, E87-D(12):2812–2820, 2004.

[8] L. M. Arslan and D. Talkin. Voice conversion by codebook mapping of line spectral frequencies and excitation spectrum. In *Proceedings of Eurospeech*, pages 1347–1350, 1997. Rhodes, Greece.

[9] J. L. Bedrosian, L. A. Hoag, and K. F. McCoy. Relevance and speech of message delivery trade-offs in augmentative and alternative communication. *Journal of Spech, Language and Hearing Research*, 46:800–817, 2004.

[10] L. Bell, J. Gustafson, and M. Heldner. Prosodic adaptation in human-computer interaction. In *Proceedings of ICPhS*, pages 2453–2456, 2003. Barcelona, Spain.

[11] C. L. Bennett. Large scale evaluation of corpus-based synthesisers: results and lessons from the Blizzard challenge 2005. In *Proceedings of Interspeech*, pages 105–108, 2005. Lisbon, Portugal.

[12] C. L. Bennett and A. W. Black. The Blizzard challenge 2006. In *Proceedings of the Blizzard Challenge Workshop*, 2006. Pittsburgh: PA, USA.

[13] C. Benoit, M. Grice, and V. Hazan. The SUS test: a method for the assessment of text-to-speech intelligibility using Semantically Unpredictable Sentences. *Speech Communication*, 18:381–392, 1996.

[14] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal. The AT&T next-gen TTS system. In *Joint meeting of ASA, EAA and DAGA*, pages 18–21, 1999. Berlin, Germany.

[15] P. Birkholz, D. Jackel, and B. J. Kröger. Construction and control of a three-dimensional vocal tract model. In *Proceedings of ICASSP*, pages 873–876, 2006. Toulouse, France.

[16] C. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford: Clarendon, 1995.

[17] A. W. Black. CLUSTERGEN: a statistical parametric synthesiser using trajectory modelling. In *Proceedings of Interspeech*, pages 1762–1765, 2006. Pittsburgh: PA, USA.

[18] A. W. Black, C. L. Bennett, J. Kominek, B. Langner, K. Prahallad, and A. Toth. CMU Blizzard 2008: Optimally using a large database for unit selection synthesis. In *Proceedings of the Blizzard Challenge Workshop*, 2008. Brisbane, Australia.

[19] A. W. Black and N. Campbell. Optimising selection of units from speech databases for concatenative synthesis. In *Proceedings of Eurospeech*, pages 581–584, 1995. Madrid, Spain.

[20] A. W. Black and K. A. Lenzo. Limited domain synthesis. In *Proceedings of ICSLP*, pages 411–414, 2000. Beijing, China.

[21] A. W. Black and K. A. Lenzo. Building synthetic voices. http://festvox.org/bsv/, January 2003. Last accessed 03 August 2009.

[22] A. W. Black and P. Taylor. Automatically clustering similar units for unit selection in speech synthesis. In *Proceedings of Eurospeech*, pages 601–604, 1997. Rhodes, Greece.

[23] A. W. Black, P. Taylor, and R. Caley. The Festival speech synthesis system. http://www.cstr.ed.ac.uk/projects/festival/. Last accessed 03 August 2009.

[24] P. Boersma and D. Weenink. Praat: doing phonetics by computer (version 5.1.04). Computer program. Last accessed 03 March 2009.

[25] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* London: Chapman and Hall, 1984.

[26] K. Browman and L. Goldstein. Towards an articulatory phonology. *Phonology Yearbook*, 3:219–253, 1986.

[27] J. S. Brumberg, P. R. Kennedy, and F. H. Guenther. Artificial speech synthesiser control by brain-computer interaction. In *Proceedings of Interspeech*, pages 636–639, 2009. Brighton, UK.

[28] H. T. Bunnell, C. Pennington, D. Yarrington, and J. Gray. Automatic personal synthetic voice construction. In *Proceedings of Interspeech*, pages 89–92, 2005. Lisbon, Portugal.

[29] J. E. Cahn. Generating expression in synthesised speech. Master's thesis, Massachusetts Institue of Technology, 1990.

[30] W. N. Campbell and A. W. Black. Prosody and the selection of source units for concatenative synthesis. In J. van Santen, R. Sproat, J. Olive, and J. Hirschberg, editors, *Progress in Speech Synthesis*, pages 279–292. New York: Springer-Verlag, 1997.

[31] R. Carlson, B. Granström, and I. Karlsson. Experiments with voice modelling in speech synthesis. *Speech Communication*, 10:481–489, 1991.

[32] J. Carmichael and P. Green. Devising a system of computerised metrics for the Frenchay Dysarthria Assessment intelligibility tests. In *Proceedings of the University of Cambridge First Postgraduate Conference in Language Research: CAMLING*, pages 473–479, Cambridge, 2003.

[33] J. Carmichael and P. Green. Revisiting dysarthria assessment intelligibility metrics. In *Proceedings of ICSLP*, pages 742–745, 2004. Jeju Island, South Korea.

[34] M. Carter. Communicative spontaneity of children with high support needs who ue augmentative and alternative communication systems ii: Antecedents and effectiveness of communication. *Augmentative and Alternative Communication*, 19(3):155–169, 2003.

[35] J. C. Catford. *A Practical Introduction to Phonetics.* Oxford: Clarendon, 1988.

[36] J. K. Chambers. *Sociolinguistic Theory.* Oxford: Blackwell, 1995.

[37] H. Clark. *Using Language.* Cambridge: Cambridge University Press, 1999.

[38] R. A. Clark, K. Richmond, and S. King. Festival 2 – build your own general purpose unit selection speech synthesiser. In *Proceedings of the 5th International Workshop on Speech Synthesis*, pages 173–178, 2004. Pittsburgh: PA, USA.

[39] C.Nass, Y. Moon, and N. Green. Are machines gender neutral? gender-stereotypic responses to computers with voices. *Journal of Applied Social Psychology*, 27:864–876, 1997.

[40] J. Coleman. Unification phonology: another look at "synthesis-by-rule". In *Proceedings of COLING*, volume 2, pages 79–84, 1990. Helsinki, Finland.

[41] A. Conkie and A. K. Syrdal. Expanding phonetic coverage in unit selection synthesis through unit substitution from a donor voice. In *Proceedings of Interspeech*, pages 1754–1757, 2006. Pittsburgh: PA, USA.

[42] A. D. Conkie and S. D. Isard. Optimal coupling of diphones. In J. van Santen, R. W. Sproat, J. P. Olive, and J. Hirschberg, editors, *Progress in Speech Synthesis*, pages 279–282. New York: Springer-Verlag, 1996.

[43] A. Copestake and D. Flickinger. Enriched language models for flexible generation in AAC systems. Technology and persons with disabilities conference CSUN-98, LA, CA, USA. www-csli.stanford.edu/~aac/csun.html Last accessed 15/12/09, 1998.

[44] R. Coulston, S. Oviatt, and C. Darves. Amplitude convergence in children's conversational speech with animated personas. In *Proceedings of ICSLP*, pages 2689–2692, 2002. Denver: CO, USA.

[45] M. Crabtree, P. Mirenda, and D. R. Beukelman. Age and gender preferences for synthetic and natural speech. *Augmentative and Alternative Communication*, 6(4):256–261, 1990.

[46] G. A. Creak. When HCI should be HHI. *Information Technology and Disabilities*, 6, 1999. Retrieved from: http://www.easi.cc/cd/itd/itdv06.htm, 22 October 2006.

[47] S. M. Creer, S. P. Cunningham, P. D. Green, and K. Fatema. Personalizing synthetic voices for people with progressive speech disorders: judging voice similarity. In *Proceedings of Interspeech*, 2009. in press.

[48] S. M. Creer, P. D. Green, S. P. Cunningham, and J. Yamagishi. Building personalised synthetic voices for individuals with dysarthria using the HTS toolkit. In J. W. Mullennix and S. E. Stern, editors, *Computer Synthesised Speech Technologies: Tools for Aiding Impairment*. Hershey, PA, USA: IGI Global, in press.

[49] F. L. Darley, A. E. Aronson, and J. R. Brown. Differential diagnostic patterns of dysarthria. *Journal of Speech and Hearing Research*, 12:246–269, 1969.

[50] C. Darves and S. Oviatt. Adaptation of users' spoken dialogue patterns in a conversational interface. In *Proceedings of ICSLP*, volume 1, pages 561–564, 2002. Denver: CO, USA.

[51] C. Delogu, S. Conte, and C. Sementina. Cognitive factors in the evaluation of synthetic speech. *Speech Communication*, 24:153–168, 1998.

[52] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, 39(1):1–38, 1977.

[53] S. Dickson, R. S. Barbour, M. Brady, A. M. Clark, and G. Paton. Patients' experiences of disruptions associated with post-stroke dysarthria. *International Journal of Language and Communication Disorders*, 43(2):135–153, 2008.

[54] R. E. Donovan and E. M. Eide. The IBM trainable speech synthesis system. In *Proceedings of ICSLP*, pages 1703–1706, 1998. Sydney, Australia.

[55] R. E. Donovan, A. Ittycheriah, M. Franz, B. Ramabhadran, E. Eide, M. Viswanathan, R. Bakis, M. Picheny, P. Gleason, T. Rutherfoord, P. Cox, D. Green, E. Janke, S. Revelin, C. Waast, B. Zeller, C. Guenther, and J. Kunzmann. Current status of the IBM trainable speech synthesis system. In *Proceedings of ESCA Tutorial and Research Workshop in Speech Synthesis*, volume 207, 2001. Perthshire, Scotland.

[56] R. E. Donovan and P. C. Woodland. Improvements in an HMM based speech synthesiser. In *Proceedings of Eurospeech*, pages 573–576, 1995. Madrid, Spain.

[57] R. E. Donovan and P. C. Woodland. A hidden Markov-model-based trainable speech synthesiser. *Computer Speech and Language*, 13(3):223–241, 1999.

[58] K. D. R. Drager, K. C. Justad, and K. L. Gable. Telephone communication: Synthetic and dysarthric speech intelligibility and listener preferences. *Augmentative and Alternative Communication*, 20(2):103–112, 2004.

[59] K. D. R. Drager and J. E. Reichle. Effects of discourse context on the intelligibility of synthesised speech for young adult and older adult listeners: applications for AAC. *Journal of Speech, Language and Hearing Research*, 44(5):1052–1057, 2001.

[60] H. Dudley. Remaking speech. *Journal of the Acoustical Society of America*, 11:169–177, 1939.

[61] J. Duffy. *Motor speech disorders: substrates, differential diagnosis and management.* St Louis, MO: Elsevier Mosby, 2nd edition, 2005.

[62] T. Dutoit and H. Leich. MBR-PSOLA: text-to-speech synthesis based on an MBE re-synthesis of the segments database. *Speech Communication*, 13(3-4):432–440, 1993.

[63] M. Eichner, M. Wolff, and R. Hoffmann. Voice characteristics conversaion for TTS using reverse VTLN. In *Proceedings of ICASSP*, pages 17–20, 2004. Montreal, Canada.

[64] D. P. W. Ellis. PLP and RASTA (and MFCC, and inversion) in Matlab. Online web resource, retrieved from http://www.ee.columbia.edu/ dpwe/resources/matlab/rastamat/, 2005. last accessed 03 February 2009.

[65] P. Enderby and L. Emerson. *Does speech and language therapy work?* London: Whurr, 1995.

[66] P. M. Enderby. *Frenchay Dysarthria Assessment.* Austin, TX: Pro-ed, 1983.

[67] Entropic Research Laboratory. ESPS programs version 5.0, 1993.

[68] G. Fant, editor. *Acoustic Theory of Speech Production.* The Hague, Netherlands: Mouton, 1960.

[69] L. J. Ferrier, H. C. Shane, H. F. Ballard, T. Carpenter, and A. Benoit. Dysarthric speakers' intelligibility and speech characteristics in relation to computer speech recognition. *Augmentative and Alternative Communication*, 11:165–174, 1995.

[70] C. Fowler. Coarticulation and theories of extrinsic timing. *Journal of Phonetics*, 8:113–133, 1980.

[71] M. Fraser and S. King. The Blizzard challenge 2007. In *Proceedings of the Blizzard Challenge Workshop*, 2007. paper 001, Bonn, Germany.

[72] S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(2):254–272, 1981.

[73] M. J. F. Gales. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, 16(3):5–24, 1998.

[74] C. W. Gorenflo, D. W. Gorenflo, and S. A. Santer. Effects of synthetic voice output on attitudes toward the augmented communicator. *Journal of Speech and Hearing Research*, 37:64–68, 1994.

[75] D. W. Gorenflo and C. W. Gorenflo. The effects of information and augmentative communication technique on attitudes toward non-speaking individuals. *Journal of Speech and Hearing Research*, 34:19–26, 1991.

[76] T. Hain. COM 4220/6460. Speech Technology course handout, Spring 2006.

[77] W. Hamza, R. Bakis, Z. W. Shuang, and H. Zen. On building a concatenative speech synthesis system from the Blizzard challenge speech databases. In *Proceedings of Interspeech*, pages 97–100, 2005. Lisbon, Portugal.

[78] H. M. Hanson and K. N. Stevens. A quasiarticulatory approach to controlling acoustic source parameters in a Klatt-type formant synthesiser using HLsyn. *Journal of the Acoustical Society of America*, 112:1158–1182, 2002.

[79] M. S. Hawley, S. Cunningham, F. Cardinaux, A. Coy, P. O'Neill, S. Seghal, and P. Enderby. Challenges in developing a voice input voice output communication aid for people with severe dysarthria. In G. Eizmendi, J. Azkoitia, and G. Craddock, editors, *Challenges for Assistive Technology*, pages 363–367. Amsterdam: IOS Press, 2007.

[80] C. Henton. Challenges and rewards in using parametric or concatenative speech synthesis. *International Journal of Speech Technology*, 5:117–131, 2002.

[81] O. E. Hetzroni and O. L. Harris. Cultural aspects in the development of AAC users. *Augmentative and Alternative Communication*, 12(1):52–58, 1996.

[82] D. J. Higginbotham, B. J. Moulton, G. W. Lesher, D. P. Wilkins, and J. Cornish. Frametalker: development of a frame-based communication system. In *Proceedings of Technology and Persons with Disabilities Conference*, 2000. California State University, Northridge.

[83] D. J. Higginbotham, H. Shane, S. Russell, and K. Caves. Access to AAC: Past, present and future. *Augmentative and Alternative Communication*, 23:243–257, 2007.

[84] D. Hill, L. Manzara, and C. Schock. Real-time articulatory speech synthesis by rule. In *Proceedings of AVIOS*, pages 27–44, 1995. San Jose: CA, USA.

[85] J. Högberg. Data driven formant synthesis. In *Proceedings of Eurospeech*, pages 565–568, 1997. Rhodes, Greece.

[86] E. Holmberg, K. Nordqvist, and G. Ahlström. Prevelance of dysarthria in adult myotonic dystrophy (m. steinert) patients; speech characteristics and intelligibility. *Logopedics Phoniatrics Vocology*, 21(1):21–27, 1996.

[87] J. Holmes. Formant synthesisers, cascade or parallel. *Speech Communication*, 2:251–273, 1983.

[88] J. Holmes and W. Holmes. *Speech Synthesis and Speech Recognition.* London: Taylor and Francis, 2nd edition, 2001.

[89] J. N. Holmes. Avoiding unwanted low-frequency level variations in the output of a parallel formant synthesiser. *Journal of the Acoustical Society of America*, 68:S18, 1980.

[90] J. N. Holmes, I. G. Mattingley, and J. N. Shearme. Speech synthesis by rule. *Language and Speech*, 7:127–143, 1964.

[91] J.-P. Hosom, A. B. Kain, T. Mishra, J. P. H. van Santen, M. Fried-Oken, and J. Staehely. Intelligibility of modifications to dysarthric speech. In *Proceedings of ICASSP*, 2003. Hong Kong, Hong Kong.

[92] A. S. House, C. Williams, M. H. L. Hecker, and K. D. Kryter. Psychoacoustic speech tests: a modified rhyme test. *Journal of the Acoustical Society of America*, 35(11):1899–1899, 1963.

[93] X. Huang, A. Acero, J. Adcock, H. Hon, J. Goldsmith, and J. Liu. Whistler: a trainable text-to-speech system. In *Proceedings of ICSLP*, pages 2387–2390, 1996. Philadelphia: PA, USA.

[94] M. Huckvale. Prorec 1.2. Computer program. http://www.phon.ucl.ac.uk/resource/prorec, Last accessed 24 March 2009.

[95] M. Huckvale. *SCRIBE manual version 1.0.* http://phon.ucl.ac.uk/resource/scribe/scribe-manual.htm, 2004. Last accessed 18 August 2009.

[96] A. J. Hunt and A. W. Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *Proceedings of ICASSP*, pages 373–376, 1996. Atlanta: GA, USA.

[97] M. Hunt-Berg. The bridge school: education inclusion outcomes over 15 years. *Augmentative and Alternative Communication*, 21:116–131, 2005.

[98] A. Iida and N. Campbell. Speech database design for a concatenative text-to-speech synthesis system for individuals with communication disorders. *International Journal of Speech Technology*, 6:379–392, 2003.

[99] A. Iida, J. Ito, S. Kajima, and T. Sugawara. Building an English speech synthesis system from a Japanese ALS patient's voice. In *Proceedings of Interspeech*, pages 1994-1997, 2006. Pittsburgh: PA, USA.

[100] K. Iskarous, L. M. Goldstin, D. H. Whalen, M. K. Tiede, and P. E. Rubin. CASY: The Haskins configurable articulatory synthesiser. In *Proceedings of ICPhS*, pages 185–188, 2003. Barcelona, Spain.

[101] F. Itakura. Line spectrum representation of linear predictive coefficients. *Journal of the Acoustical Society of America*, 57(Sup 1):S35, 1975.

[102] J. M. Johnson, E. Inglebret, C. Jones, and J. Ray. Perspectives of speech language pathologists regarding success versus abandonment of AAC. *Augmentative and Alternative Communication*, 22(2):85–99, 2006.

[103] D. Jones. *The Pronunciation of English.* Cambridge: Cambridge University Press, 1982.

[104] A. Kain and M. Macon. Personalizing a speech synthesiser by voice adaptation. In *Proceedings of the 3rd International Workshop on Speech Synthesis*, pages 225–230, 1998. Jenolan Caves, Australia.

[105] A. Kain, X. Niu, J.-P. Hosom, Q. Miao, and J. van Santen. Formant re-synthesis of dysarthric speech. In *Proceedings of the 5th International Workshop on Speech Synthesis*, pages 25–30, 2004. Pittsburgh: PA, USA.

[106] V. Karaiskos, S. King, R. A. J. Clark, and C. Mayo. The Blizzard challenge 2008. In *Proceedings of the Blizzard Challenge Workshop*, 2008. Brisbane, Australia.

[107] W. F. Katz. Anticipatory coarticulation and aphasia: implications for phonetic theories. *Journal of Phonetics*, 28(3):313–334, 2000.

[108] H. Kawahara. Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited. In *Proceedings of ICASSP*, pages 1303–1306, 1997. Munich, Germany.

[109] H. Kawahara, H. Katayose, A. Cheveigne, and R. Patterson. Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity. In *Proceedings of Eurospeech*, pages 2781–2784, 1999. Budapest, Hungary.

[110] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné. 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.

[111] B. Kemp. Quality of life while ageing with a disability. *Assistive Technology*, 11:158–163, 1999.

[112] R. Kent, G. Weismer, J. Kent, and J. Rosenbek. Toward phonetic intelligibility testing in dysarthria. *Journal of Speech and Hearing Disorders*, 54:482–499, 1989.

[113] S. King, K. Tokuda, H. Zen, and J. Yamagishi. Unsupervised adaptation for HMM-based speech synthesis. In *Proceedings of Interspeech*, pages 1869–1872, 2008. Brisbane, Australia.

[114] D. H. Klatt. The KlatTalk text-to-speech conversion system. In *Proceedings of ICASSP*, pages 1589–1592, 1982. Boston: MA, USA.

[115] D. H. Klatt. Text-to-speech conversion. *Journal of the Acoustical Society of America*, 82(3):737–793, 1987.

[116] D. H. Klatt and L. C. Klatt. Analysis, synthesis and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87:820–857, 1990.

[117] J. Kominek and A. W. Black. CMU Arctic databases for speech synthesis. http://festvox.org/cmu_arctic/cmu_arctic_report.pdf, 2003. Last accessed 20 April 2006.

[118] J. Kominek, T. Schultz, and A. W. Black. Synthesiser voice quality on new languages calibrated with mel-cepstral distortion. In *Online proceedings of SLTU*, 2008. http://www.mica.edu.vn/sltu/proceedings/papers/kominek_sltu_08 .pdf, Last accessed 02 April 2009.

[119] R. K. Koul. Synthetic speech perception in individuals with and without disabilities. *Augmentative and Alternative Communication*, 19(1):49–58, 2003.

[120] P. Ladefoged. *A Course in Phonetics.* Fort Worth: TX, Harcourt, Brace, and Jovanovich, 3rd edition, 1993.

[121] P. Ladefoged. *Elements of Acoustic Phonetics.* London: University of Chicago Press, 2nd edition, 1996.

[122] J. P. Lasker and J. L. Bedrosian. Promoting acceptance of augmentative and alternative communication by adults with acquired communication disorders. *Augmentative and Alternative Communication*, 17(3):141–153, 2001.

[123] C. Legetter and P. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech and Language*, 9(2):171–185, 1995.

[124] S. Lemmety. Review of speech synthesis technology. Master's thesis, Helsinki University of Technology, March 1999. www.acoustics.hut.fi/publications/files/theses/lemmetty_mst/contents.html, Last accessed 18 August 2009.

[125] S. C. Levinson. *Pragmatics.* Cambridge, Cambridge University Press, 1983.

[126] E. Lewis and M. A. A. Tatham. A new text-to-speech synthesis system. In *Proceedings of Eurospeech*, pages 1235–1238, 1991. Genova, Italy.

[127] A. M. Liberman, F. Ingemann, L. Lisker, P. Delattre, and F. Cooper. Minimal rules for synthesising speech. *Journal of the Acoustical Society of America*, 31:1490–1499, 1959.

[128] J. Light. Interaction involving individuals using augmentative and alternative communication systems: state of the art and future directions. *Augmentative and Alternative Communication*, 4(2):66–82, 1988.

[129] J. Light and K. Drager. AAC technologies for young children with complex communication needs. *Augmentative and Alternative Communication*, 23(3):204–216, 2007.

[130] J. Light, R. Page, J. Curran, and L. Pitkin. Children's ideas for the design of AAC assistive technologies for young children with complex communication needs. *Augmentative and Alternative Communication*, 23(4):274–287, 2007.

[131] M. Lilienfeld and E. Alant. Attitudes of children toward an unfamiliar peer using an AAC device with and without voice output. *Augmentative and Alternative Communication*, 18(2):91–101, 2002.

[132] B. Lindblom. Explaining phonetic variation: a sketch of the H&H theory. In W. Hardcastle and A. Marchal, editors, *Speech Production and Speech Modelling*, pages 403–439. Dordrecht: Kluwer, 1990.

[133] Z.-H. Ling, K. Richmond, J. Yamagishi, and R.-H. Wang. Articulatory control of HMM-based parametric speech synthesis driven by phonetic knowledge. In *Proceedings of Interspeech*, pages 573–576, 2008. Brisbane, Australia.

[134] Z.-H. Ling and R.-H. Wang. HMM-based unit selection using frame sized speech segments. In *Proceedings of Interspeech*, pages 2034–2037, 2006. Pittsburgh: PA, USA.

[135] J. L. Locke. Where did all the gossip go?: Casual conversation in the information age. *American Speech Language Hearing Association*, 40(3):26–31, 1998.

[136] D. Logan and D. B. Pisoni. Preference judgements comparing different synthetic voices. *Journal of the Acoustical Society of America*, 79(S1):S24–S25, 1986.

[137] J. S. Logan, B. G. Greene, and D. B. Pisoni. Segmental intelligibility of synthetic speech produced by rule. *Journal of the Acoustical Society of America*, 86:566–581, 1989.

[138] J. Lunn, J. Todman, P. File, and E. Coles. Making contact in the workplace. *Communication Matters*, 18(1):26–28, 2004.

[139] P. MacNeilage. Motor control of self-ordering of speech. *Psychological Review*, 77:182–196, 1970.

[140] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai. Speech synthesis using HMMs with dynamic features. In *Proceedings of ICASSP*, pages 389–392, 1996. Atlanta: GA, USA.

[141] P. Mathy, K. M. Yorkston, and M. L. Gutmann. AAC for individuals with amyotrophic lateral sclerosis. In D. R. Beukelman, K. Yorkston, and J. Reichle, editors, *Augmentative communication for adults with neurogenic and neuromuscular disabilities*, pages 183–229. Baltimore: MD, Paul H. Brookes, 2000.

[142] J. Matousek. Speech synthesis using HMM-based acoustic unit inventory. In *Proceedings of Eurospeech*, pages 2323–2326, 1999. Budapest, Hungary.

[143] P. Matthews. Muscle spindles and their motor control. *Physiological Review*, 44:219–288, 1964.

[144] C. Mayo, R. A. J. Clark, and S. King. Multidimensional scaling of listener responses to synthetic speech. In *Proceedings of Interspeech*, pages 1725–1728, 2005. Lisbon, Portugal.

[145] X. Menéndez-Pidal, J. B. Polikoff, S. M. Peters, J. E. Leonzio, and H. T. Bunnell. The Nemours database of dysarthric speech. In *Proceedings of ICSLP*, pages 1962–1965, 1996. Philadelphia: PA, USA.

[146] J. Metzner, M. Schmittfull, and K. Schnell. Substitute sounds for ventriloquism and speech disorders. In *Proceedings of Interspeech*, pages 1379–1382, 2006. Pittsburgh: PA, USA.

[147] C. Middag, J.-P. Martens, G. van Nuffelen, and M. D. Bodt. Automated intelligibility assessment of pathological speech using phonological features. *EURASIP Journal on Advances in Signal Processing*, 2009:1–9, 2009.

[148] N. Miller, E. Noble, D. Jones, and D. Burn. Life with communication changes in Parkinson's disease. *Age and Ageing*, 35:235–239, 2006.

[149] P. Mirenda and D. R. Beukelman. A comparison of speech synthesis intelligibility with listeners from three age groups. *Augmentative and Alternative Communication*, 3(3):120–128, 1987.

[150] P. Mirenda and D. R. Beukelman. A comparison of intelligibility among natural speech and seven speech synthesisers with listeners from three age groups. *Augmentative and Alternative Communication*, 6(2):61–68, 1990.

[151] R. Moore and A. Morris. Experiences collecting genuine spoken enquiries using WOZ techniques. In *Proceedings of the 5th DARPA Workshop on Speech and Natural Language*, pages 61–63, 1992. New York: NY, USA.

[152] A. T. Morgan and A. P. Vogel. Intervention for dysarthria associated with acquired brain injury in children and adolescents. *Cochrane Database Systematic Review*, 16(3):Art. No.: CD006279. DOI: 10.1002/14651858.CD006279.pub2, 2008.

[153] E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9:453–467, 1990.

[154] B. E. Murdoch and D. G. Theodoros. Ataxic dysarthria. In B. E. Murdoch, editor, *Dysarthria: a physiological approach to assessment and treatment*, chapter 8, pages 242–265. Cheltenham: Stanley Thornes, 1998.

[155] B. E. Murdoch and E. C. Thompson-Ward. Flaccid dysarthria. In B. E. Murdoch, editor, *Dysarthria: a physiological approach to assessment and treatment*, chapter 6, pages 176–204. Cheltenham: Stanley Thornes, 1998.

[156] J. Murphy. 'I prefer contact this close': perceptions of AAC by people with motor neurone disease and their communication partners. *Augmentative and Alternative Communication*, 20(4):259–271, 2004.

[157] I. R. Murray and J. L. Arnott. A tool for the rapid development of new synthetic voice personalities. In *Speech and Language Technology for Disabled Persons*, pages 111–114, 1993. Stockholm, Sweden.

[158] I. R. Murray and J. L. Arnott. Synthesising emotions in speech: is it time to get excited? In *Proceedings of ICSLP*, pages 1816–1819, 1996. Philadelphia: PA, USA.

[159] K. Nakamura, T. Toda, Y. Nakajima, H. Saruwatari, and K. Shikano. Evaluation of speaking-aid system with voice conversion for laryngectomees toward its use in practical environments. In *Proceedings of Interspeech*, pages 2209–2212, 2008. Brisbane, Australia.

[160] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano. Speaking aid system for total laryngectomees using voice conversion of body transmitted artificial speech. In *Proceedings of Interspeech*, pages 1395–1398, 2006. Pittsburgh: PA, USA.

[161] Y. Nakano, M. Tachibana, J. Yamagishi, and T. Kobayashi. Constrained structural maximum a posteriori linear regression for average voice based speech synthesis. In *Proceedings of Interspeech*, pages 2286–2289, 2006. Pittsburgh: PA, USA.

[162] S. Narayanan, A. Alwan, and Y. Song. New results in vowel production: MRI, EPG and acoustic data. In *Proceedings of Eurospeech*, pages 1007–1009, 1997. Patras, Greece.

[163] M. Narendranath, H. A. Murthy, S. Rajendran, and B. Yegnanarayana. Transformation of formants for voice conversion using artificial neural networks. *Speech Communication*, 16:207–216, 1995.

[164] J. Nurminen, V. P. J. Tian, Y. Tang, and I. Kiss. A parametric approach for voice conversion. In *TC-STAR Workshop on Speech-to-Speech Translation*, pages 225–229, 2006. Barcelona, Spain.

[165] R. Ogden. Parametric interpretation in Yorktalk. *York Papers in Linguistics*, 16:81–99, 1992.

[166] D. Öhlin and R. Carlson. Data-driven formant synthesis. In *Proceedings of FONETIK*, pages 160–163, 2004. Stockholm, Sweden.

[167] Y. Ohtani, T. Toda, H. Saruwaratari, and K. Shikano. Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation. In *Proceedings of Interspeech*, pages 2266–2269, 2006. Pittsburgh: PA, USA.

[168] B. M. O'Keefe, L. Brown, and R. Schuller. Identification and rankings of communication aid features by five groups. *Augmentative and Alternative Communication*, 14(1):37–50, 1998.

[169] Özgül Salor and M. Demirekler. Dynamic programming approach to voice transformation. *Speech Communication*, 48:1262–1272, 2006.

[170] P. Parette and M. B. Huer. Working with Asian American families whose children have augmentative and alternative communication needs. *Journal of Special Education Technology E-Journal*, 17(4), 2002. http://jset.unlv.edu/17.4T/parette/first.html, last accessed 25 October 2006.

[171] M. Parker, S. Cunningham, P. Enderby, M. Hawley, and P. Green. Automatic speech recognition and training for severely dysarthric users of assistive technology - the STARDUST project. *Clinical Linguistics and Phonetics*, 20(2–3):149–156, 2006.

[172] M. Plumpe, A. Acero, H. Hon, and X. Huang. HMM-based smoothing for concatenative speech synthesis. In *Proceedings of ICSLP*, pages 2751–2754, 1998. Sydney, Australia.

[173] L. R. Rabiner. A tutorial on HMM and selected applications in speech recognition. In *Proceedings of the IEEE*, volume 77, pages 257–286, 1989.

[174] A. Ratcliff, S. Coughlin, and M. Lehman. Factors influencing ratings of speech naturalness in augmentative and alternative communication. *Augmentative and alternative communication*, 18(1):11–19, 2002.

[175] P. Rubin, E. Saltzman, R. McGowan, L. Goldstein, M. Tiede, and K. Browman. CASY and extensions to the task-dynamics model. In *Proceedings of First ESCA Tutorial and Research Workshop on Speech Production Modelling*, pages 125–128, 1996. Autrans, France.

[176] A. I. Rudnicky, C. Bennett, A. W. Black, A. Chotomongcol, K. Lenzo, A. Oh, and R. Singh. Task and domain specific modelling in the Carnegie Mellon communicator system. In *Proceedings of ICSLP*, pages 130–134, 2000. Beijing, China.

[177] H. Sacks, E. A. Schegloff, and G. Jefferson. A simplest systematics for the organisation of turn-taking for conversation. *Language*, 50(4):696–735, 1974.

[178] E. Saltzman and K. Munhall. A dynamic approach to gestural patterning in speech recognition. *Ecological Psychology*, 1:333–382, 1989.

[179] M. M. Schepis and D. H. Reid. Effects of a voice output communication aid on interactions between support personnel and an individual with multiple disabilities. *Journal of Applied Behaviour Analysis*, 28:73–77, 1995.

[180] S. Schötz. F0 and segment duration in formant synthesis of speaker age. In *Proceedings of Speech Prosody*, 2006. Dresden, Germany.

[181] K. Shinoda and T. Watanabe. MDL-based context-dependent subword modelling for speech recognition. *Journal of the Acoustical Society of Japan (E)*, 21:79–86, 2000.

[182] O. Shiohan, T. A. Myrvoll, and C.-H. Lee. Structural maximum a posteriori linear regression for fast HMM adaptation. *Computer Speech and Language*, 16(3):5–24, 2002.

[183] M. M. Smith. The dual challenges of aided communication and adolescence. *Augmentative and Alternative Communication*, 21(1):76–79, 2005.

[184] S. E. Stern. Computer synthesised speech and perceptions of the social influence of disabled users. *Journal of Language and Social Psychology*, 27(3):254–265, 2008.

[185] S. E. Stern, J. W. Mullennix, C.-L. Dyson, and S. J. Wilson. The persuasiveness of synthetic speech versus human speech. *Human Factors*, 41:588–595, 1999.

[186] S. E. Stern, J. W. Mullennix, and S. J. Wilson. Effects of perceived disability on persuasiveness of computer synthesised speech. *Journal of Applied Psychology*, 87:411–417, 2002.

[187] S. E. Stern, J. W. Mullennix, and I. Yaroslavsky. Persuasion and social perception of human vs. synthetic voice across person as source and computer as source conditions. *International Journal of Human-Computer Studies*, 64:43–52, 2006.

[188] C. Stevens, N. Lee, J. Vonwiller, and D. Burnham. On-line experimental methods to evaluate text-to-speech (TTS) synthesis: effects of voice gender and signal quality on intelligibility, naturalness and preference. *Computer Speech and Language*, 19:129–146, 2005.

[189] K. N. Stevens and C. A. Bickley. Constraints among parameters simplify control of Klatt formant synthesiser. *Journal of Phonetics*, 19:161–174, 1991.

[190] R. L. Street and H. Giles. Speech accommodation theory: a social cognitive approach to language and speech. In M. Roloff and C. R. Berger, editors, *Social Cognition and Communication*, pages 193–226. Beverly Hills: Sage, 1982.

[191] T. Styger and E. Keller. Formant synthesis. In E. Keller, editor, *Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of the Art and Future Challenges*, pages 109–128. Chichester: John Wiley, 1994.

[192] Y. Stylianou, O. Cappé, and E. Moulines. Continuous probabilistic transform for voice conversion. In *IEEE Transactions on Speech and Audio Processing*, volume 6(2), pages 131–142, 1998.

[193] M. Tatham. Towards a cognitive phonetics. *Journal of Phonetics*, 12:37–47, 1986.

[194] M. Tatham and K. Morton. *Developments in speech synthesis.* Basingstoke, Palgrave Macmillan, 2005.

[195] M. Tatham and K. Morton. *Speech production and perception.* Basingstoke, Palgrave Macmillan, 2006.

[196] D. G. Theodoro. Mixed dysarthria. In B. E. Murdoch, editor, *Dysarthria: a physiological approach to assessment and treatment*, chapter 11, pages 337–372. Cheltenham: Stanley Thornes, 1998.

[197] D. G. Theodoros and B. E. Murdoch. Hyperkinetic dysarthria. In B. E. Murdoch, editor, *Dysarthria: a physiological approach to assessment and treatment*, chapter 10, pages 314–336. Cheltenham: Stanley Thornes, 1998.

[198] D. G. Theodoros and B. E. Murdoch. Hypokinetic dysarthria. In B. E. Murdoch, editor, *Dysarthria: a physiological approach to assessment and treatment*, chapter 9, pages 266–313. Cheltenham: Stanley Thornes, 1998.

[199] N. Thomas-Stonell, A.-L. Kotler, H. A. Leeper, and P. C. Doyle. Computerized speech recognition: influence of intelligibility and perceptual consistency on recognition accuracy. *Augmentative and Alternative Communication*, 14:51–56, 1998.

[200] E. C. Thompson-Ward. Spastic dysarthria. In B. E. Murdoch, editor, *Dysarthria: a physiological approach to assessment and treatment*, chapter 5, pages 205–241. Cheltenham: Stanley Thornes, 1998.

[201] T. Toda, A. W. Black, and K. Tokuda. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Audio, Speech and Language Processing*, 15(8):2222–2235, 2007.

[202] T. Toda and K. Tokuda. A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. *IEICE Transactions on Information and Systems*, E90-D(5):816–824, 2007.

[203] J. Todman. Rate and quality of conversations using a text-storage aac system: Single-case training study. *Augmentative and Alternative Communication*, 16(3):164–179, 2000.

[204] J. Todman, N. Alm, and L. Elder. Computer-aided conversation: a prototype system for non-speaking people with physical disabilities. *Applied Psycholinguistics*, 15:45–73, 1994.

[205] J. Todman, D. Rankin, and P. File. The use of stored text in computer-aided conversation: A single-case experiment. *Journal of Language and Social Psychology*, 18:287–309, 1999.

[206] K. Tokuda, T. Kobayashi, and S. Imai. Speech parameter generation from HMM using dynamic features. In *Proceedings of ICASSP*, pages 660–663, 1995. Detroit: MI, USA.

[207] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi. Multi-space probability distribution HMM. In *IEICE Transactions of Information and Systems*, volume E85-D(3), pages 455–464, 2002.

[208] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. Speech parameter generation algorithm for HMM-based speech synthesis. In *Proceedings of ICASSP*, pages 1315–1318, 2000. Beijing, China.

[209] K. Tokuda, H. Zen, and A. W. Black. An HMM-based speech synthesis system applied to English. IEEE Speech Synthesis Workshop, 2002. Santa Monica: CA, USA.

[210] K. van den Doel, F. Vogt, R. E. English, and S. S. Fels. Towards articulatory speech synthesis with a dynamic 3D finite element tongue model. In *Proceedings of ISSP*, pages 59–66, 2006.

[211] W. A. van Dommelen. Acoustic parameters in human speaker recognition. *Language and Speech*, 33(3):259–272, 1990.

[212] P. A. Vanderheyden and C. A. Pennington. An augmentative communication interface based on conversational schemata. In V. O. Mittal, H. A. Yanco, J. Aronis, and R. Simpson, editors, *Assistive Technology and Artificial Intelligence*, volume 1458 of *Lecture Notes in Computer Science*, pages 109–125. Berlin, Springer-Verlag, 1998.

[213] H. S. Venkatagiri. Effects of sentence length and exposure on the intelligibility of synthesised speech. *Augmentative and Alternative Communication*, 10(2):96–104, 1994.

[214] H. S. Venkatagiri. Segmental intelligibility of four currently used text-to-speech synthesis methods. *Journal of the Acoustical Society of America*, 113(4):2095–2104, 2003.

[215] J. Vepa, S. King, and P. Taylor. Objective distance measures for spectral discontinuities in concatenative speech synthesis. In *Proceedings of ICSLP*, pages 2605–2608, 2002. Denver: CO, USA.

[216] A. J. Viterbi. Error bounds for convolution codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):257–286, 1967.

[217] G. von Bekesy. The structure of the middle ear and the hearing of one's own voice by bone conduction. *Journal of the Acoustical Society of America*, 21:217–232, 1949.

[218] G. Weismer. *Motor Speech Disorders.* Oxford: Plural Publishing, 2007.

[219] J. C. Wells. *Accents of English: An Introduction.* Cambridge: Cambridge University Press, 1982.

[220] A. Wijk. *Rules for the Pronunciation of English.* Oxford: Oxford University Press, 1960.

[221] B. Wisenburn and D. J. Higginbotham. An AAC application using speaking partner speech recognition to automatically produce contextually relevant utterances: objective results. *Augmentative and Alternative Communication*, 24(2):100–109, 2008.

[222] J. Wouters and M. W. Macon. Perceptual evaluation of distance measures for concatenative speech synthesis. In *Proceedings of ICSLP*, pages 2747–2750, 1998. Sydney, Australia.

[223] J. Yamagishi and T. Kobayashi. Average voice based speech synthesis using HSMM-based speaker adaptation and adaptive training. *IEICE Transactions on Information and Systems*, E90-D(2):533–543, 2007.

[224] J. Yamagishi, T. Kobayashi, M. Tachibana, K. Ogata, and Y. Nakano. Model adaptation approach to speech synthesis with diverse voices and styles. In *Proceedings of ICASSP*, pages 1233–1237, 2007. Honolulu: HI, USA.

[225] J. Yamagishi, Z. Ling, and S. King. Robustness of HMM-based speech synthesis. In *Proceedings of Interspeech*, pages 581–584, 2008. Brisbane, Australia.

[226] J. Yamagishi, T. Nose, H. Zen, Z. Ling, T. Toda, K. Tokuda, S. King, and S. Renals. A robust speaker-adaptive HMM-based text-to-speech synthesis. *IEEE Transactions on Audio, Speech and Language Processing*, 17(6):66–83, 2009.

[227] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi. A context clustering technique for average voice models. *IEICE Transactions on Information and Systems*, E86-D(3):534–542, 2003.

[228] J. Yamagishi, H. Zen, T. Toda, and K. Tokuda. Speaker-independent HMM-based speech synthesis system – HTS-2007 for the Blizzard challenge 2007. In *Proceedings of the Blizzard Challenge Workshop*, 2007. Bonn, Germany.

[229] J. Yamagishi, H. Zen, Y.-J. Wu, T. Toda, and K. Tokuda. The HTS-2008 system: Yet another evaluation of the speaker-adaptive HMM-based speech synthesis system in the 2008 Blizzard challenge. In *Proceedings of the Blizzard Challenge Workshop*, 2008. Brisbane, Australia.

[230] H. Ye and S. Young. Perceptually weighted linear transformation for voice conversion. In *Proceedings of Eurospeech*, pages 2409–2412, 2003. Geneva, Switzerland.

[231] H. Ye and S. Young. Voice conversion for unknown speakers. In *Proceedings of ICSLP*, pages 1161–1164, 2004. Jeju Island, South Korea.

[232] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. Simultaneous modelling of spectrum, pitch and duration in HMM-based speech synthesis. In *Proceedings of Eurospeech*, pages 2347–2350, 1999. Budapest, Hungary.

[233] S. Young, G. Everman, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. *The HTK book version 3.2.1*, December 2002.

[234] 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 *Proceedings of the 6th International Workshop on Speech Synthesis*, pages 294–299, 2007. Bonn, Germany.

[235] H. Zen and T. Toda. An overview of Nitech HMM-based speech synthesis system for Blizzard challenge 2005. In *Proceedings of Interspeech*, pages 93–96, 2005. Lisbon, Portugal.

[236] H. Zen, T. Toda, M. Nakamura, and K. Tokuda. Details of the Nitech HMM-based speech synthesis system for the Blizzard challenge 2005. *IEICE Transactions on Information and Systems*, E90-D(1):325–333, 2007.

[237] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. Hidden semi-Markov model based speech synthesis. In *Proceedings of ICSLP*, pages 1397–1400, 2004. Jeju Island, South Korea.

[238] W. Ziegler. Apraxia of speech. In *Handbook of Clinical Neurology*, volume 88 of *Neuropsychology and behavioural neurology*, pages 269–286. Amsterdam: Elsevier Science, 2008.