Controlling the voice quality dimension of prosody in synthetic speech using an acoustic glottal model

Andrew Murphy

Thesis for the Degree of Doctor of Philosophy

Phonetics and Speech Laboratory

School of Linguistic, Speech and Communication Sciences

Trinity College Dublin

2021

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish.

Signature of author: 

Andrew Murphy

November 2020

“We were away a year ago.”

– Anon.

Summary

Statistical parametric speech synthesis (SPSS) offers a means of generating synthetic speech without the need for complex and extensive rules. One way in which this approach is sometimes lacking is through the use of simple excitation models that may result in unnatural, robotic sounding synthetic speech. A further limitation is that these systems tend to lack prosodic variation. For example, they do not capture the expressive nature of human spoken interaction. This is something that would be highly desirable for applications utilising speech synthesis, such as educational games or synthetic voices for people with disordered speech. The use of a more complex excitation model could offer the flexibility in the voice source that could provide a basis for more adequate modelling of prosody. However, acoustic models of the voice source entail many potentially important parameters and controlling these could be a challenge.

The main aims of this work were to: investigate how an acoustic glottal model could be used to manipulate aspects of linguistic and paralinguistic prosody of synthetic speech using a minimal set of control parameters; implement the knowledge gained from this investigation into an analysis-and-synthesis system; use this system in SPSS; and conduct pilot tests to demonstrate how the system can be used to explore the voice source correlates of prosody, through user-driven manipulation tasks.

To achieve the first goal, experiments were carried out to explore how the global waveshape parameter, R_d (Fant, 1995), could be used to control aspects of linguistic and paralinguistic prosody. This parameter can be used to generate glottal pulse shapes that result in voice qualities ranging from breathy to tense. As the tense-lax dimension of voice quality is important in prosodic modulation, R_d appears to be ideal for minimising the number of control parameters needed to transform voice quality.

To allow control of this parameter, the glottal source, and the vocal tract filter that shapes it, must be modelled using the principles of the source-filter model of speech production. These speech components must be separated effectively to do this. Inverse filtering was used to obtain estimates of the source signal by removing the effects of the vocal tract transfer function from the speech signal. An acoustic glottal model, the Liljencrants-Fant (LF) model (Fant *et al.*, 1985), was then used to parameterise the glottal source.

Three experiments were carried out, using manually inverse filtered data, to investigate how R_d could be used as a control parameter for linguistic and paralinguistic prosody, even in the absence of f_0 modulation. Experiment 1 examined how manipulating R_d could be used to control where focal prominence (an aspect of linguistic prosody) occurs in an utterance. Experiment 2 explored how R_d could be used as a control parameter for perceived affect. Experiment 3 built upon the results of Experiment 1 to optimise the implementation of the R_d parameter contour.

The results of these experiments confirmed that R_d can serve as a control parameter to generate linguistic prominence as well as paralinguistic modification of affective colouring. The results

confirmed, and elaborated on, the findings of earlier research, suggesting that tense-lax modulation of voice quality is important in prosodic expression. They indicated that a more tense phonation on the focally accented item can be used to signal prominence, while laxer phonation of post-focal material provides source deaccentuation that further enhances the perceived prominence.

These experiments provided information concerning R_d ranges and settings that fed into the development of the second goal of this work, i.e. an analysis-and-synthesis system, called GlórCáil, for the control of parameters for prosodic variation in synthesis. The system also allows for some speaker characteristic transformation, letting the user manipulate both voice source and vocal tract parameters before resynthesis. This provides the means to alter the prosodic pattern and speaker characteristics of an utterance. The interface allows the user to listen to any changes they make after resynthesis, to see if they have the desired effect or if further manipulations are required.

The third goal was achieved by integrating the finished system into a DNN-based speech synthesis framework so that it could be used to generate unseen synthetic speech.

The final goal of this work was achieved by demonstrating the system's ability to transform linguistic and paralinguistic prosody, as well as speaker characteristics, in copy synthesis. Two manipulation tasks were carried out using purpose-built interfaces developed for these experiments. Experiment 4 involved participants modifying an utterance so that it sounded like an appropriate response to a given question by moving sliders that controlled the R_d parameter. Experiment 5 involved participants manipulating parameters to make an utterance sound like it was being spoken by a particular speaker in a particular affective state. The responses were then used to modify the default parameters generated by the DNN-based speech synthesis system, by multiplying them by a scaling factor, to create a set of stimuli. These stimuli were used in the listening test of Experiment 6, where participants were asked to identify the speaker, the emotion of the speaker, and rate the magnitude of the emotion and naturalness of the utterance on five-point scales. Although participants identified sad stimuli successfully, this was not the case for happy stimuli. It is likely that additional modifications of the vocal tract and f_0 contour are needed to improve identification rates.

The last two experiments were intended as pilot demonstrations: given that the focus of the thesis is on the voice quality dimension, and the inclusion of vocal tract and f_0 modulation is beyond the intended scope of the work. Using the GlórCáil system, future work is planned which will explore how the voice quality dimension combines with f_0 in both linguistic and paralinguistic prosodic expression. How these combine in speaker voice transformation is another area which will be of interest.

The GlórCáil system experiments reported here are seen as a contribution not only towards better control of the voice quality dimension of prosody in speech synthesis, but also towards research methodologies that will enhance our understanding of this vital dimension of human communication.

Acknowledgements

I would like to start by deeply thanking my supervisor, Dr. Christer Gobl. He provided an incredible amount of insight and knowledge into the world of speech. I would also like to thank Prof. Ailbhe Ní Chasaide for acting like a second supervisor throughout the period of time over which this work was carried out. Go raibh míle míle maith agat! I would also like to offer special thanks to Dr. Irena Yanushevskaya for offering a huge amount of support and advice regarding not only the PhD, but also cats.

I'm very grateful to An Roinn Cultúir, Oidhreachta agus Gaeltachta (the Department of Culture, Heritage and the Gaeltacht) for funding me through the ABAIR project, without which this work would never have been carried out.

I would like to thank everyone in the Phonetics and Speech Lab (PhLab). You're the guys who make it such an enjoyable place to be, and without you who'd be there to do crosswords or laugh at grumpy cat pictures?

To Mum, Dad, Kieran and Sonia. Thank you so much for the love, moral support, kindness and understanding. It means so much to me.

And finally, to my grandparents. Even though none of you are around to see me finish this, you were always so supportive and happy to see me doing well. This is for you!

Table of Contents

Abbreviations

SPSS – Statistical parametric speech synthesis

DNN – Deep neural network

LPC – Linear predictive coding

DAP – Discrete all-pole

IAIF – Iterative adaptive inverse filtering

FIR – Finite impulse response

IOP – Iterative optimal pre-emphasis

GFM – Glottal flow model

QCP – Quasi-closed phase

WLP – Weighted linear prediction

AME – Attenuated main excitation

GCI – Glottal closure instants

LF – Liljencrants-Fant

TTS – Text-to-speech

FL – Fujisaki-Ljungqvist

OQ – Open quotient

SQ – Speed quotient

ClQ – Closing quotient

NAQ – Normalised amplitude quotient

QOQ – Quasi-open quotient

HRF – Harmonic richness factor

MSPD – Mean squared phase difference

EKFM – Extended Kalman filtering method

HMM – Hidden Markov model

MRI – Magnetic resonance imaging

PSOLA – Pitch-synchronous overlap and add

HTS – HMM-based speech synthesis system

CNN – Convolutional neural network

RNN – Recurrent neural network

GPU – Graphical processing unit

ANN – Artificial neural network

LSTM – Long short-term memory

BLSTM – Bidirectional Long short-term memory

BRNN – Bidirectional recurrent neural network

MLSA – Mel-log spectrum approximation

MBE – Multi-band excitation

STRAIGHT - Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum

FFT – Fast Fourier transform

HNM – Harmonic-plus-noise model

MVF – Maximum voiced frequency

DSM – Deterministic plus stochastic model

MGC – Mel-generalised cepstral

LSF – Line spectral frequencies

GMM – Gaussian mixture models

GSS – Glottal spectral separation

DC – Direct current

SVLN – Separation of the Vocal-tract with the Liljencrants-Fant model plus Noise

AM – Autosegmental-metrical

ToBI – Tones and break indices

GUI – Graphical user interface

ANOVA – Analysis of variance

VUV – Voiced/unvoiced

DGF – Differentiated glottal flow

NCCF – Normalised cross-correlation function

LSP – Line spectral pair

MLPG – Maximum likelihood parameter generation

VQ – Voice quality

ML – Maximum likelihood

ICC - Indicative of the correlation of the items within a cluster

CI – Confidence interval

VT – Vocal tract

List of Tables

List of Figures

Chapter 1. Introduction

1.1 Rationale for this work

Statistical parametric speech synthesis (SPSS) currently offers the highest degree of parametric control without the need for expertly derived rules. If an overly simple excitation model is used the resulting synthetic speech can sound rather robotic and unnatural. This is also an issue when it comes to transforming aspects of speech such as, the type of phonation, speaker characteristics, or affective state. These are important elements to control when using synthetic voices in educational software, video games, or other applications where expressive speech is required. While parametric flexibility is important for the above-mentioned applications, the naturalness of the synthesised speech is an important factor to consider due to how sensitive listeners are to the slightest distortions in speech. In order to further improve the expressiveness of these synthesis systems, an excitation model of higher complexity must be used. A downside to using a more complicated excitation model is that controlling and transforming it also becomes more complicated. This becomes an issue when these transformations are being carried out by a nonexpert user, or in time sensitive cases where synthetic speech material needs to be generated quickly.

One major goal of this work is to develop a system that will allow for the transformation of the voice in synthetic speech, by utilising our understanding of speech production. One way of integrating knowledge of speech production into a synthesis system would involve the use of an acoustic glottal model-based excitation that more closely models the natural glottal source. Many state-of-the-art speech modelling methods that implement improved excitation models focus on either improving the naturalness of synthetic speech with little control of the source parameters or implementing full parametric control at a cost to the naturalness of the synthesised speech.

The ability to transform the quality of the voice in synthetic speech in a way that is guided by knowledge of speech production would offer many advantages. Firstly, it would allow us to take a given synthetic voice and from it generate multiple new voices (sounding like different speakers) by altering the production characteristics. Currently, the most widely

used method for transforming voices in SPSS is through speaker adaptation, covered in detail in Yamagishi *et al.* (2009). This method involves the transformation of acoustic models between speakers to impart characteristics of one speaker to the other. To reduce the specificity of a particular voice an average voice is created that is trained using data from many speakers. This method is not however based on our knowledge of speech production, and for many languages, especially minority and low resource languages, it requires specific databases of spoken language that are typically not available.

As mentioned, even if the baseline voice in a speech synthesis system is quite natural, the output tends to be monotonous as it lacks the kinds of prosodic modulations that are a defining feature of natural speech in human interactions. These prosodic modulations are often seen as being of two kinds, linguistic and paralinguistic.

- **Linguistic prosody:** this is the aspect of prosody that has been studied within linguistics for many decades, described mostly in terms of the intonational, melodic modulation of running speech. Linguistic prosody is seen to serve a number of different functions: it helps cue grammatical structure; it helps cue the listener on the information structure, i.e. which parts of an utterance they need to particularly attend to; it helps regulate discourse so that listeners know whether the speaker's turn is coming to an end and whether the floor is being yielded or not.

  Although the vast majority of work in this field has been focussed on f_0 modulation, recent research has indicated that other parameters of the voice, pertaining to the quality of the voice, are also modulated as part of linguistic prosody. In the present work, it is this voice quality dimension of the voice source that is focussed on.

- **Paralinguistic/affective prosody:** this is the aspect of prosody that carries the crucial affective and attitudinal information of the message. When we decode speech, we decode not just the string of words, but also the emotion and mood of the speaker (angry, sad, bored) and their attitudes and relationship to us (friendly, condescending, sarcastic, polite). This aspect of prosody has frequently tended to be regarded as outside the scope of linguistic investigation and has been a topic pursued more in the field of psychology – with little cross-reference to the linguistic aspects of prosody. Although the voice quality dimension of the voice source is appreciated to be fundamentally important to the affective, paralinguistic prosody, it is nonetheless true

that most of the empirical research carried out to date has been focussed on f_0, intensity and timing of speech.

Another goal of this work is to develop a system that will allow control of the voice source in speech synthesis. It is intended as a contribution towards more flexible control of the voice source in synthesis. Based on our current understanding of the voice source and speech production theory, some recent research on aspects of voice modulation in prosody will be taken as a starting point. The system is also intended as a tool for future exploration that will extend our understanding of the voice source, on how voice source modulations contribute to the prosodic expression of utterances and how such modulations might be incorporated into speech synthesis systems.

The system should allow voice source modifications that can serve both to transform the speaker-specific baseline voice quality (an extralinguistic aspect of voice) and the prosodic characteristics of utterances (including linguistic and/or paralinguistic features of the voice). In this work, the main focus is on the on the latter, prosodic dimension of voice source modulation. Based on studies on specific aspects of prosodic variation, a number of experiments are undertaken to see how such modulations may be controlled in speech synthesis. In later pilot experiments, modifications of the speaker-baseline characteristics are also included, but are not treated in depth.

The need for a flexible system to control voice quality in speech synthesis can be illustrated in terms of parallel research that is ongoing in the ABAIR Irish synthesis project in Trinity College Dublin (Ní Chasaide *et al.*, 2011a, 2017) to which this work is linked. This project is concerned with the development of speech technology for Irish, an endangered minority language. Synthetic voices have been developed for the three main dialects of Irish, and these are currently being used in applications for users with disabilities (such as visual impairment) as well as in Irish language learning applications, such as interactive multimodal educational games. A means to flexibly and easily control the voice source and filter characteristics of the synthetic speech output would be of major benefit in this context. The ability to build many voices for a particular dialect by transforming an available voice is important. It would enable for the population of educational games with a cast of different characters who have sufficiently differentiated voices. To date, the project has generated such differentiated characters using relatively simple transforms. Results are often rather poor, and while such transformations have

served the purpose up to a point, they are never going to come close to modelling the diversity and natural quality of different human voices. The ideal would be a system that allows for control of both the source and filter, thus allowing for the generation of, from first principles of speech production, any number of voices. These voices could differ in subtle or major ways from each other. It would also provide a means to personalise synthetic voices for use in communication systems for those with disordered speech.

The other aspects where flexibility in the generation of synthetic speech is important concerns the linguistic and paralinguistic/affective prosody. Currently, they are both difficult to approximate using mainstream synthesis techniques. Gaining full control of the voice source is an important step towards this goal. The characters in interactive games need expression in their voices that matches the contexts of the scenarios within the games. For example, at certain times the characters in a game need to sound angry, friendly, or secretive, to suit the storyline.

Likewise, in many contexts where people with disabilities use synthetic speech, the need to control these same features will have a major impact. Being able to personalise and fine-tune a voice for a user is an important goal, which it is hoped that this work will contribute to. Furthermore, the ability to control linguistic prosody and the ability to modulate the voice to carry the affective dimension of a message would greatly enhance the use of speech synthesis as a communication device.

All these factors show that there are still many improvements that can be made to create a fully flexible system, with parametric control over the voice source signal and vocal tract filter and narrow the gap between synthesised and natural speech.

Another issue is the vast number of voice source parameters to choose from. A problem with rule-based formant synthesisers is the need for extensive expert knowledge in order to produce speech with natural sounding prosody. It is very desirable to simplify the control parameter scheme of expressive speech synthesis so that the manipulation of a minimal set of values equates to useful changes in prosody. For this reason, this work will investigate the use of the global waveshape parameter, R_d (Fant, 1995, 1997), as a means to control the voice source, and therefore, voice quality. This parameter captures the covariance of other model parameters, while reducing the overall number of parameters needed to generate a synthetic source signal.

1.2 Aims and objectives

The aims of this work were to:

- Investigate how a small set of parameters could be used to control aspects of linguistic and paralinguistic prosody. Specifically, the use of the R_d parameter is explored as the parameter that might effectively capture the tense-lax dimension of voice quality, found to vary in previous studies with specific aspects of prosodic expression.
- Implement a speech analysis-and-synthesis system that allows a user to easily control R_d and other voice source parameters, as well as vocal tract characteristics.
- Integrate the system into an SPSS system.
- Demonstrate how the system can be effectively used through a set of user-driven manipulation tasks.

The research carried out aimed to answer the following question:

> ***To what extent can control of both linguistic and paralinguistic prosody be achieved by manipulating the voice source and vocal tract in statistical parametric speech synthesis utilising an acoustic glottal source model with a minimal set of control parameters?***

Answering this question is intended to contribute to the area of linguistic and paralinguistic prosody, to a voice control system for voice modulation/transformation in synthesis with its basis in speech production theory and towards a more prosodically adequate, expressive speech synthesis that is needed for many speech technology applications.

To address the research question proposed above, a set of experiments was devised that would demonstrate different dimensions of control that might be offered by the system.

1.3 Outline

This thesis is organised in the following way: Chapter 2 outlines the principles of speech production, including descriptions of the components involved. It also introduces the acoustic theory of speech production, which acts as the basis for many of the analysis and synthesis methods discussed in the work. Chapter 3 describes methods of analysing and

parameterising the voice source, with a particular focus on acoustic glottal models. Chapter 4 describes several different approaches to speech synthesis and introduces the concept of neural networks and how they are utilised in speech synthesis systems. Chapter 5 describes and discusses a number of different speech modelling techniques that are used in speech synthesis. Chapter 6 is concerned with the role of voice quality in signalling prosody and considers ways in which prosodic variation might be incorporated in speech synthesis. Chapter 7 details three experiments that were carried out to explore how voice source manipulations linked with voice quality shifts along the tense-lax continuum could signal prominence, as well as perceived affect, in synthetic speech. Chapter 8 details the development of an analysis-and-synthesis system, GlórCáil, including control interfaces for analysis, synthesis and carrying out perception experiments. This chapter also describes how the system was integrated into a DNN-based speech synthesis framework. Chapter 9 describes three experiments that were carried out to demonstrate the functions of the GlórCáil system in copy-synthesis applications and when it is used in an SPSS framework. Finally, Chapter 10 presents and discusses the conclusions made about this research and lists possible avenues of future work that could be carried out.

Portions of the work carried out as part this thesis were reported in the following list of publications[1]:

1) Yanushevskaya, I., Murphy, A., Gobl, C. and Ní Chasaide, A. (2016) 'Perceptual Salience of Voice Source Parameters in Signalling Focal Prominence', in *Proceedings of INTERSPEECH*. San Francisco, USA, pp. 3161–3165. doi: 10.21437/Interspeech.2016-1160.
2) Murphy, A., Yanushevskaya, I., Ní Chasaide, A. and Gobl, C. (2017) 'R_d as a control parameter to explore affective correlates of the tense-lax continuum', in *Proceedings of INTERSPEECH*. Stockholm, Sweden, pp. 3916–3920. doi: 10.21437/Interspeech.2017-1448.
3) Murphy, A., Yanushevskaya, I., Ní Chasaide, A. and Gobl, C. (2018) 'Voice Source Contribution to Prominence Perception: R_d Implementation', *in Proceedings of INTERSPEECH*. Hyderabad, India, pp. 217–221. doi: 10.21437/Interspeech.2018-2352.
4) Murphy, A., Yanushevskaya, I., Ní Chasaide, A. and Gobl, C. (2019) 'The Role of Voice Quality in the Perception of Prominence in Synthetic Speech', *in*

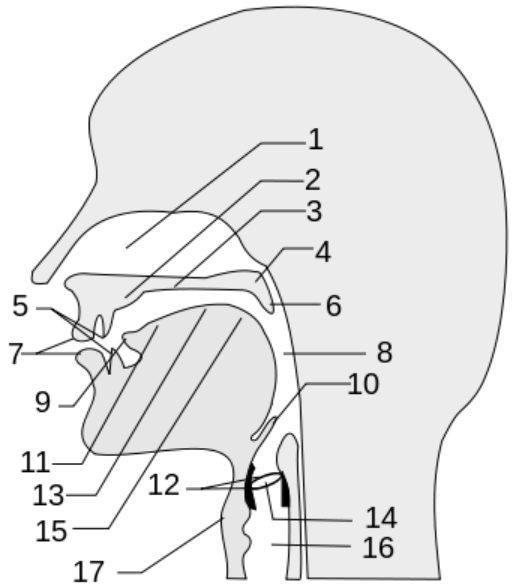
[1] The design and execution of these studies was carried out principally by the author of this thesis. This work built on earlier research by the co-authors, and they were also involved in the planning and interpretation of the data.

Proceedings of INTERSPEECH. Gratz, Austria, pp. 2543–2547. doi: 10.21437/Interspeech.2019-2761.

Chapter 2. Speech production, acoustics and voice source modelling

2.1 Speech production

From a physiological point of view it is useful to describe the speech production system as comprising of three parts; the subglottal respiratory system, the larynx, and the supra-glottal vocal tract (Lieberman and Blumstein, 1988, pp. 3–13). The subglottal respiratory system contains the lungs and respiratory control muscles. The tensing and relaxing of these muscles cause the lungs to expand and contract, producing a flow of air through the trachea, which not only sustains life but also can also act as the power source for speech. The next part of the system is the larynx, positioned in the neck, between the respiratory system and the vocal tract. Muscles within the larynx control the tension and position of two folds of tissue, the vocal folds. These folds can be employed to modulate the flow of air coming from the lungs into what becomes a source excitation for speech. The next element in the speech production system is the vocal tract. The vocal tract (see Figure 2.1) is comprised of the pharynx and oronasal cavities. Here, the source excitation, generated by the vocal folds, is modified by the resonances and antiresonances created by the varying geometries of the cavities. Finally, speech is emitted through the lips and nostrils as an acoustic signal.



1. Nasal cavity
2. Oral cavity
3. Hard palate
4. Soft palate
5. Teeth
6. Uvula
7. Lips
8. Pharynx
9. Tip of tongue
10. Epiglottis
11. Blade of tongue
12. Vocal cords
13. Front of tongue
14. Glottis
15. Back of tongue
16. Trachea
17. Larynx

Figure 2.1: Midsagittal cross-section of the head, including the larynx and vocal tract. By Tavin, licensed under CC-BY-3.0

The two main categories of speech sounds are voiced and unvoiced. Voiced speech is produced with the modulated airflow excitation, generated through the vibratory action of the vocal folds. Unvoiced speech is produced in the absence of quasi-periodic vibration of the vocal folds. The airflow is constricted within the vocal tract, generating a noise excitation with no periodicity.

2.1.1 The larynx, vocal folds and glottal flow

The larynx is an adjustable cartilaginous tubular organ at the top of the trachea (Reetz and Jongman, 2008). Figure 2.2 illustrates the muscles and cartilages that form the larynx. The main structure of the larynx is made up of the thyroid and cricoid cartilages, which are attached together through the cricothyroid muscle. Within the larynx lie the vocal folds.

The vocal folds are two protruding folds of tissue that stretch between the thyroid and arytenoid cartilages (see Figure 2.2). The open area between the vocal folds, allowing air to flow through from the lungs, is called the glottis. Interaction between the glottis and the pulmonic airstream can cause the vocal folds to vibrate (Reetz and Jongman, 2008). Intrinsic laryngeal muscles can control the tension of the vocal folds laterally and longitudinally, changing their vibratory behaviour and the dimensions of the glottis.

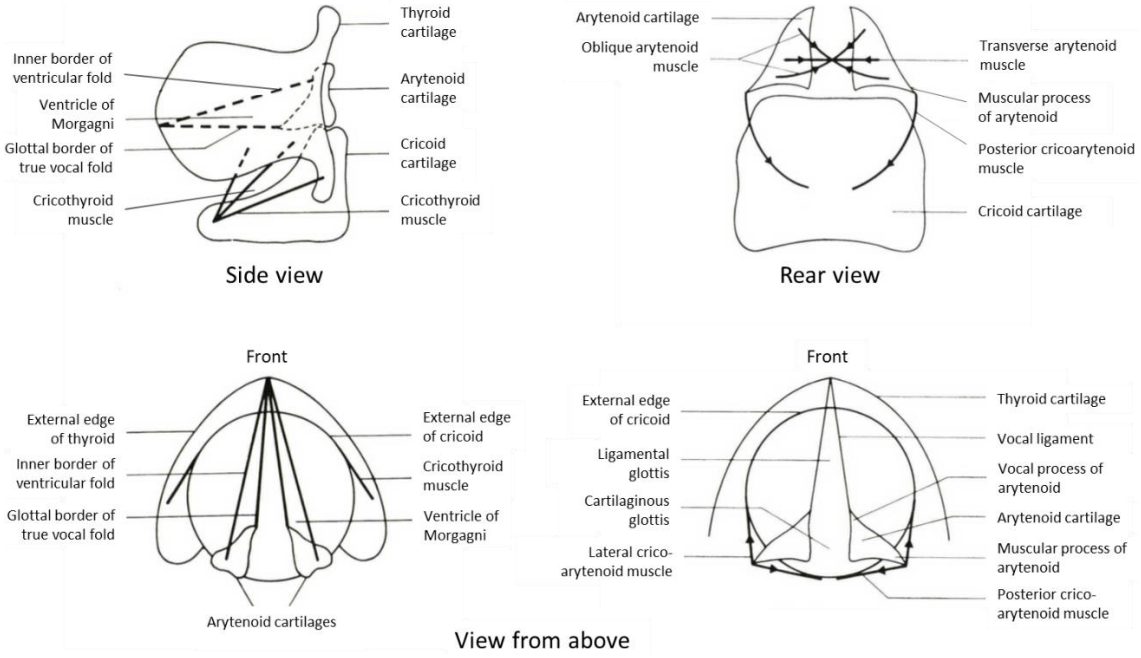


Figure 2.2: Diagram of the larynx and vocal folds as viewed from the side, rear and above. Adapted from Laver (1980)

The oscillatory action of the vocal folds when they are employed for voiced speech, referred to as phonation (Ladefoged and Johnson, 2010), modulates the size of the glottis and converts the steady airflow from the lungs into a quasi-periodic series of pulses. These glottal pulses are known as the glottal source or the voice source.

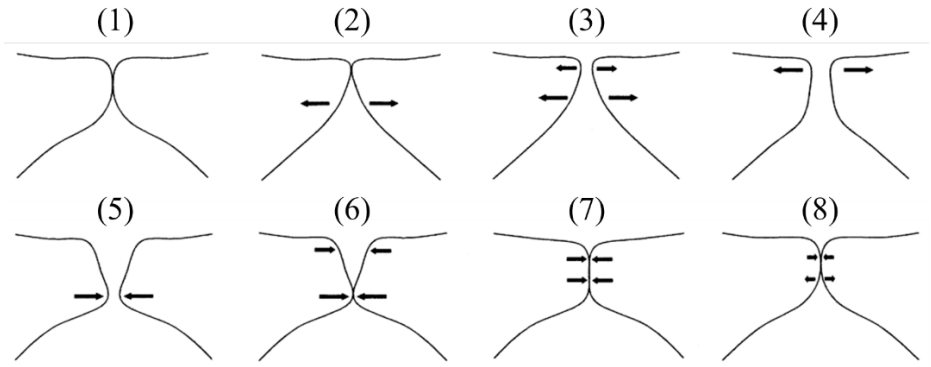


Figure 2.3: Diagram of an idealised single cycle of vocal fold vibration. Adapted from Story (2002)

The myoelastic-aerodynamic theory (van den Berg, 1958) is regularly used to explain the mechanism of vibration of the vocal folds. This theory combines knowledge of the elastic properties of the vocal folds with the aerodynamic characteristics of the glottis to account for some of the behaviour of the vocal folds. In order to explain how the folds are able to completely close, the two-mass model (Ishizaka and Flanagan, 1972) must also be

considered. Additionally, the muco-viscose (Broad, 1979) and flow-separation (Ishizaka and Matsudaira, 1968) theories account for other details in the oscillatory behaviour of the vocal folds. Figure 2.3 illustrates the vocal folds in the coronal plane during an idealised single cycle of their vibration (Story, 2002):

1. At the beginning of the glottal cycle the vocal folds are adducted (i.e. brought together) by muscular tension acting upon them (Laver, 1980). This closes off the airway resulting in an increase in subglottal pressure below the vocal folds.
2. The lower portions of the folds begin to move away from each other laterally due to the increased subglottal pressure (Hardcastle, 1976).
3. The lateral movement continues until the folds separate, opening the airway.
4. They continue to move until maximum lateral travel has been reached.
5. The folds begin to move closer together, due to elastic forces and the Bernoulli effect, caused by the increase in airflow velocity (Lieberman and Blumstein, 1988).
6. The lower portions of the folds come together, closing off the airway.
7. The upper portions of the folds also come together so they are fully adducted.
8. The subglottal pressure increases, and the cycle begins again.

The rate of vocal fold vibration determines the fundamental frequency (f_0), which correlates closely with the perceived pitch. This rate is governed by the mass, length and tension of the vocal folds. These dimensions differ for males and females, with men tending to have longer vocal folds of a higher mass and thus a lower f_0. The average f_0 values for males and females are approximately 120 Hz and 220 Hz respectively (Fant, 1956). Children have even shorter vocal folds and therefore a higher average f_0 (approximately 260 Hz (Peterson and Barney, 1952)).

Laryngeal adjustments (e.g., increased tension of the vocal folds) not only affect the pitch, but also the mode of vibration of the vocal folds. These adjustments can be described, as in Laver (1980) in terms of three main parameters illustrated in Figure 2.4:

1. Adductive tension – this tension brings the arytenoid cartilages together.
2. Medial compression – this tension closes the ligamental glottis.
3. Longitudinal tension – this controls the tension along the length of the vocal folds.

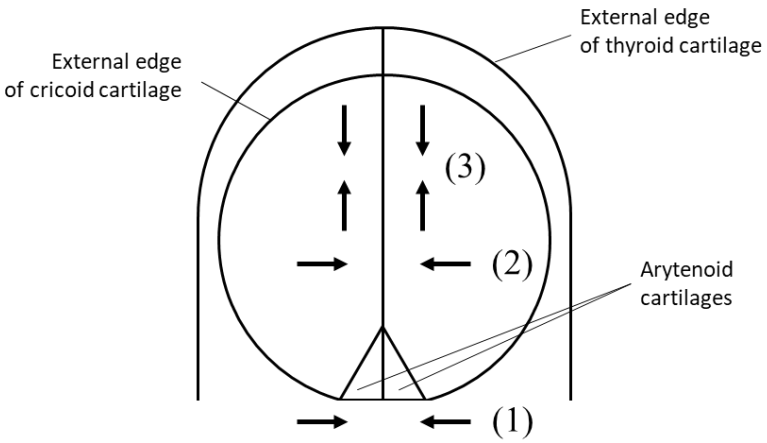


Figure 2.4: Schematic of the main dimensions of vocal fold tension: (1) Adductive tension. (2) Medial compression. (3) Longitudinal tension. Based on Laver (1980)

2.1.2 Phonatory settings and voice quality

To avoid confusion, this thesis uses the descriptive framework proposed by Laver (1980), where voice quality is described as auditory colouring of an individual speaker's voice. Short and long term changes in voice quality can serve a linguistic purpose such as differentiating between speech units (Laver and Trudgill, 1979), or making a particular syllable or word stand out in a sentence. Voice quality can also communicate information about a speaker, such as their age, sex, size and affective state. These features are often referred to as being extralinguistic, or paralinguistic in the case of affect, as they are beyond the normal scope of linguistics (Gobl, 2003). This work will be mainly focus on the laryngeal contribution to voice quality, brought on by changes in muscular adjustments within the larynx.

Modal voice can be defined as the neutral phonation setting. The vocal folds function in an efficient manner, using moderate adductive tension, medial compression and longitudinal tension. The complete closure of the glottis produces a quasi-periodic glottal flow with minimal aspiration noise (Gobl, 1989). This voice quality can be used as a baseline from which to describe the range of voice qualities relevant to this work along a lax to tense continuum with regards to laryngeal muscle tension.

Whisper is generated with a triangular opening in the glottis at the arytenoid cartilages. This opening is formed with low adductive tension, moderate medial compression of the vocal folds. The opening produces a turbulent flow of air with no periodic component. This quality can be combined with modal voice to produce whispery voice, which includes a periodic component introduced by vibration along parts of the vocal folds (Laver, 1980).

Breathy voice involves less efficient vibration of the vocal folds with a higher rate of airflow when compared to modal voice. It includes strong aspiration noise with incomplete closure of the glottis due to low muscular effort in the larynx. This incomplete, and more gradual compared to modal voice, closure leads to more rounded glottal pulses. These more rounded pulses exhibit a steeper spectral slope resulting from lower amplitudes of higher frequency harmonics.

Tense voice results from higher adduction of the vocal folds when compared to modal voice. The higher tension results in sharper pulses passing through the glottis, in which the vocal folds close more abruptly than in modal voice. This sharpness leads to increased harmonic amplitudes at higher frequencies and a shallower spectral tilt.

Harsh voice is produced by applying extreme adductive tension and medial compression to the vocal folds. This tension causes the vocal folds to vibrate in an irregular manner, introducing aperiodicities into the glottal flow. This produces an unpleasant, raspy sound.

Creaky voice is another voice quality that will be mentioned in the later experiments. Catford (1964) describes creaky voice as sounding like a rapid series of taps. It is thought to originate from a small section of the ligamental glottis, near the thyroid, when high adductive tension, medial compression, and low longitudinal tension are applied to the vocal folds (Fónagy, 1962) in combination with low subglottal pressure (Monsen and Engebretson, 1977).

Modulations in voice quality can be used by speakers for several communicative purposes. Some languages use voice quality to signal phonemic differences (Laver and Trudgill, 1979; Gordon and Ladefoged, 2001). In several tonal languages, different voice qualities can be used to signify certain tones (Huffman, 1987; Jianfen and Maddieson, 1992). Voice quality can serve a paralinguistic function by communicating the affective state of a speaker (Ladd *et al.*, 1985; Gobl and Ní Chasaide, 2003).

Chapter 4 contains a more details description of how voice quality contributes to linguistic and paralinguistic prosody.

2.1.3 The vocal tract

The glottal flow is an excitation signal that acts as the basis for speech. If we were to hear the excitation signal in its raw form, we would perceive it as a harmonically rich, but

unintelligible buzz. In order for this signal to be shaped into the sounds that humans use in spoken communication it must be filtered by the supra-glottal vocal tract.

The supraglottal vocal tract – comprised of the laryngeal cavity, the pharynx, the oral cavity and the nasal cavity – acts as an acoustic filter. The glottal source is filtered by the vocal tract filter, and the resonances of this filter produce peaks in the speech output spectrum known as formants.

The geometry of the vocal tract changes when a speaker changes the position of their articulatory organs, namely, the jaw, tongue and lips. Furthermore, the larynx may be raised or lowered, effectively changing the length of the vocal tract. The velum can also be controlled, opening and closing airflow into the nasal cavity. These changes in the geometry result in changes in the acoustic properties of the vocal tract, which modify the source signal into the different sounds that make up speech. For voiced speech, the glottal source signal is modified by the resonances of the vocal tract. Fricatives are produced by a noise source of turbulent airflow caused by constrictions in the vocal tract. If vocal fold vibrations occur alongside the noise source, voiced fricatives are produced. Voiceless fricative are produced when only the noise source is used (Flanagan, 1972). Stop consonants are produced when the vocal tract is fully closed at a point, allowing for the build-up and sudden release of air pressure. The location of the closure determines the sound produced. Nasal consonants are produced when the velum, also known as the soft palate, is opened, while the vocal tract is closed off by the tongue.

Just as with the vocal folds, age and gender affects the properties of the vocal tract. Fitch and Giedd (1999) found the average vocal tract length to be 16.1 cm for adult males, 14.6 cm for adult females, and the average values for children between ages 2 and 12 ranged between 10.4 cm and 13.4 cm. The variation in vocal tract dimensions between people is what, along with features of the voice source, gives each person their own characteristic timbre or recognisable quality to their speech.

2.2 The acoustic theory of speech production

Although, in reality, speech production is carried out by a series of continuous physiological processes, from an engineering point of view it can be very useful to describe these processes individually. Doing so allows for the examination of each component independently of others. It is for this reason that it is favourable to view speech

production in terms of a source and filter. The source component, produced by the vocal folds, is filtered, by the vocal tract in this case, and the resulting output is speech (see Figure 2.5).

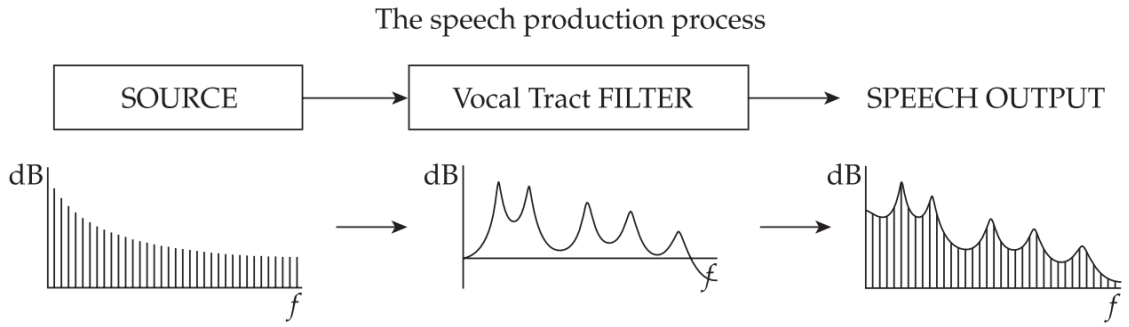


Figure 2.5:A schematic illustration of the source-filter model of speech production. From (Gobl and Ní Chasaide, 2010)

This theory forms the basis of Fant's (1960) "Acoustic Theory of Speech Production". This theory describes speech production in a way that is closer to the concepts usually encountered in digital signal processing. It is for this reason that many speech synthesis systems employ this theory in their speech modelling methods. This theory can be represented schematically (see Figure 2.6) as the voice source signal being filtered by the transfer function of the vocal tract filter, resulting in the speech signal.

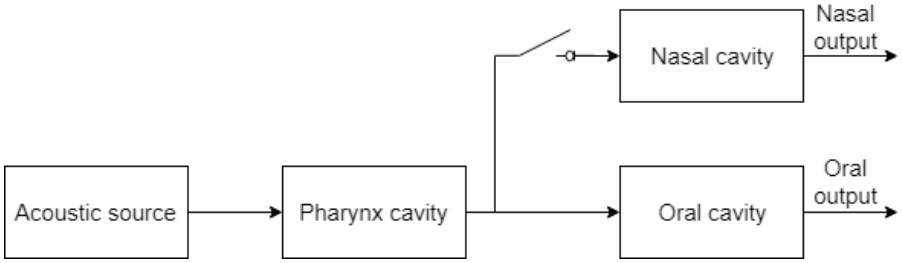


Figure 2.6. Schematic of the production of speech as a filtering process. Based on Fant (1960)

The system can be described in the z-domain as seen in Equation (2.1), where $G(z)$ is the spectrum of the glottal source signal, $g(t)$, $V(z)$ is the transfer function of the vocal tract filter, $R(z)$ is the radiation characteristics of the lips, and $S(z)$ is the spectrum of the speech signal output, $s(t)$.

$$S(z) = G(z)V(z)R(z) \qquad (2.1)$$

Separating speech production into these separate parts allows for much easier analysis of each individual component. Application of this theory allows us to separate the source component from the filter component (discussed in Section 2.3).

A limitation of this model is that it does not take any subglottal interactions into account. It assumes that the vocal tract is a tube, closed at the glottis and open at the mouth, and the source signal originates at the closed end of this tube. Also, the source and filter are assumed to be completely independent of each other with no interactions between them. More detail on these interactions can be found in Gobl (2003). The following two sections will discuss the individual components of the source-filter theory.

2.2.1 The source

The source, written in the time-domain as $g(t)$, acts as the excitation signal to the vocal tract filter in the source-filter model of speech production. It is the approximation of the glottal airflow waveform that is produced by air travelling from the lungs and shaped by the opening and closing of the glottis. The source signal can be treated as an individual component of the source-filter system, independent of changes made to the vocal tract filter. The source can be divided into two main categories, voiced and unvoiced. When the source is voiced it consists of quasi-periodic pulses with a harmonic structure that usually sounds like a coarse buzz. This is the glottal source used to produce sounds like vowels and voiced consonants (e.g., [m] and [l]). The unvoiced source consists of noise produced by turbulent airflow at the glottal constriction (Lieberman and Blumstein, 1988) or another location within the vocal tract, e.g. the labiodental and dental constrictions for [f] and [θ] respectively.

2.2.2 The filter

The vocal tract filter, $V(z)$, is what shapes the glottal source into the sounds we hear as speech. It can be modelled using a number of complex-conjugate poles and zeros. An all-pole model is usually used to model vowel sounds, which consists of a set of resonators, called formants. In theory, the number of formants is infinite (Kent and Read, 1992, chap. 2; Stevens, 1998), but only the first few formants are perceptually relevant. Each formant has two characteristics, the centre frequency (often referred to as the formant frequency) and bandwidth. The centre frequency of a formant defines the point of strongest resonance in the frequency domain, while the formant bandwidth is a measure of the damping rate in the time domain. In combination, these formants make up the transfer function of the vocal tract.

2.2.3 Radiation

The final component of the source-filter model to mention is the radiation characteristic of the lips and nostrils, $R(z)$. This has the approximate effect of raising the spectral amplitude of the output speech signal by 6 dB per octave, which can be represented as a first-order differentiator as shown in Equation (2.2) (Rabiner and Schafer, 1978), where α is usually set to a value just below 1 (Deller *et al.*, 1999, chap. 3), corresponding to a real zero just inside the unit circle.

$$R(z) = 1 - \alpha z^{-1} \qquad (2.2)$$

This differentiator effect can be incorporated into the glottal source so that the source-filter model can be expressed in the form shown in Equation (2.3), where $G'(z)$ is the spectrum of the glottal source derivative.

$$S(z) = G'(z)V(z) \qquad (2.3)$$

This can be convenient as several glottal source models are, in fact, modelling the glottal flow derivative (Fant, Liljencrants and Lin, 1985).

2.3 Voice source analysis

The previous sections in this chapter detailed how the speech production system can be described using the source-filter theory. The following section describes several commonly used techniques for voice source analysis that exploit the principles of this theory to separate speech into its source and filter components. This section also includes descriptions of the processes and models used to parameterise the voice source and vocal tract once they have been estimated.

2.3.1 Glottal inverse filtering

The process of glottal inverse filtering involves estimating the vocal tract filter transfer function and removing its effects from a speech signal to obtain an estimate of the glottal source. The coefficients of the vocal tract transfer function can be estimated both automatically and manually. Automatic approaches use techniques such as linear predictive coding (LPC) to approximate the shape of the vocal tract transfer function. These approaches can be very inaccurate when carried out on speech recorded using inadequate equipment (which may introduce phase distortion) or in an unideal environment (too much noise, reverberation, etc.)(Alku, 2011), but can be used to analyse large amounts of data in a relatively short amount of time. Manual approaches usually involve an

automatic or semi-automatic analysis stage to obtain initial estimates of the vocal tract transfer function. These values are then manually corrected by an expert. Although systems that use these methods (Gobl and Ní Chasaide, 1999a; Kane and Gobl, 2013) can potentially yield very accurate estimates of the vocal tract transfer function and the glottal source signal, they are very slow and laborious, and so cannot be used on large data sets. Even using state-of-the-art approaches, obtaining an accurate estimation can be difficult. The following sections briefly describe several commonly used methods of glottal inverse filtering including some of their variants.

2.3.1.1 Inverse filtering with pre-emphasis

When standard LPC analysis is performed on speech it is likely to estimate the spectral envelope of the whole speech signal rather than just the vocal tract transfer function. This envelope includes the effects of the vocal tract, radiation at the lips and nostrils, plus the spectral characteristics of the glottal source. The signal obtained when the inverse of this spectral envelope is applied to a speech waveform is not an accurate representation of the voice source, but rather just an error-based residual with an approximately flat spectrum, as long as the LPC order is sufficiently high.

One technique that can be used to help obtain an estimation of the voice source signal is to pass the speech signal through a pre-emphasis filter before it is inverse filtered. This process acts to boost the higher frequency components of the speech spectrum – in effect compensating for the drop in the higher frequency components due to the combined spectral contributions of the voice source spectrum and the lip radiation. A discrete-time filter of the following format is usually used for pre-emphasis,

$$H(z) = 1 - \alpha z^{-1} \qquad (2.4)$$

where α is typically in the range of 0.95 to 0.99. A problem with using a pre-emphasis filter in this way is that although it improves the accuracy of glottal inverse filtering, due to its stationary nature it will not properly compensate for the varying glottal source spectrum of continuous speech. Methods have been described to vary this filter across time in order to improve upon this e.g., (Schnell and Lacroix, 2007).

A discrete all-pole (DAP) filter (El-Jaroudi and Makhoul, 1991) can be used in place of the usual all-pole filter for the inverse filtering. This type of modelling has been shown to provide more accurate estimates of the formants of the vocal tract than an all-pole model,

reducing the amount of harmonic interference, especially at higher fundamental frequencies (Alku *et al.*, 1998).

2.3.1.2 Closed-phase inverse filtering

Another way of estimating the voice source is through the use of closed-phase inverse filtering introduced by Strube (1974) and Wong *et al.* (1979). In this case, rather than estimating the vocal tract filter of an analysis window of several glottal pulses, only regions where the glottis is closed are used. This is advantageous because when the glottis is closed there is minimal interaction between the subglottal areas and the vocal tract transfer function. If these closed regions can be accurately detected, a more accurate estimation of the vocal tract transfer function can be obtained using the covariance method of LPC analysis (Rabiner and Schafer, 1978).

This method of analysis can still be problematic. One problem is when this method is used for speech with a high f_0. The closed-phase duration in such speech may be too short to accurately estimate the transfer function of the vocal tract. Methods have been proposed to overcome this problem by carrying out analysis over several glottal cycles (Brookes and Chan, 1994; Yegnanarayana and Veldhuis, 1998; Plumpe *et al.*, 1999). Another problem with this technique is the difficulty in detecting the closed-phase to begin with. Alku *et al.* (2009) proposed a method that constrained inverse filtering parameters prior to optimization to account for errors introduced by analysis frame position and to provide a model of the vocal tract that more accurately follows the source-filter theory. Improvements in estimating the location of the closed-phase have been proposed in Thomas *et al.* (2012) and Drugman *et al.* (2012), but this still remains a difficult task. Laxer voice qualities (e.g., breathy voice) also pose a problem for this form of inverse filtering as the vocal folds may not completely close, making detection of a fully closed-phase next to impossible.

2.3.1.3 Iterative and adaptive inverse filtering

Iterative adaptive inverse filtering (IAIF) is an automatic method of estimating the glottal flow waveform by cancelling the effects of the vocal tract and lip radiation.

Figure 2.7 shows a flowchart of all the stages of the IAIF method. The process starts by high-pass filtering the speech signal using a linear phase finite impulse response (FIR)

filter with a cut-off frequency of 30 Hz. This removes any low-frequency distortions from the signal. This is followed by several iterations of LPC analysis, inverse filtering, and integration. The first LPC analysis step (block 2) estimates the combined effect of the vocal tract and lip radiation (H_{g1}) of the speech signal using 1st order LPC analysis. This estimation is then used to inverse filter the signal in the next step (block 3). This signal is then analysed using pth order LPC (block 4). The LPC order, p, can be calculated using the equation,

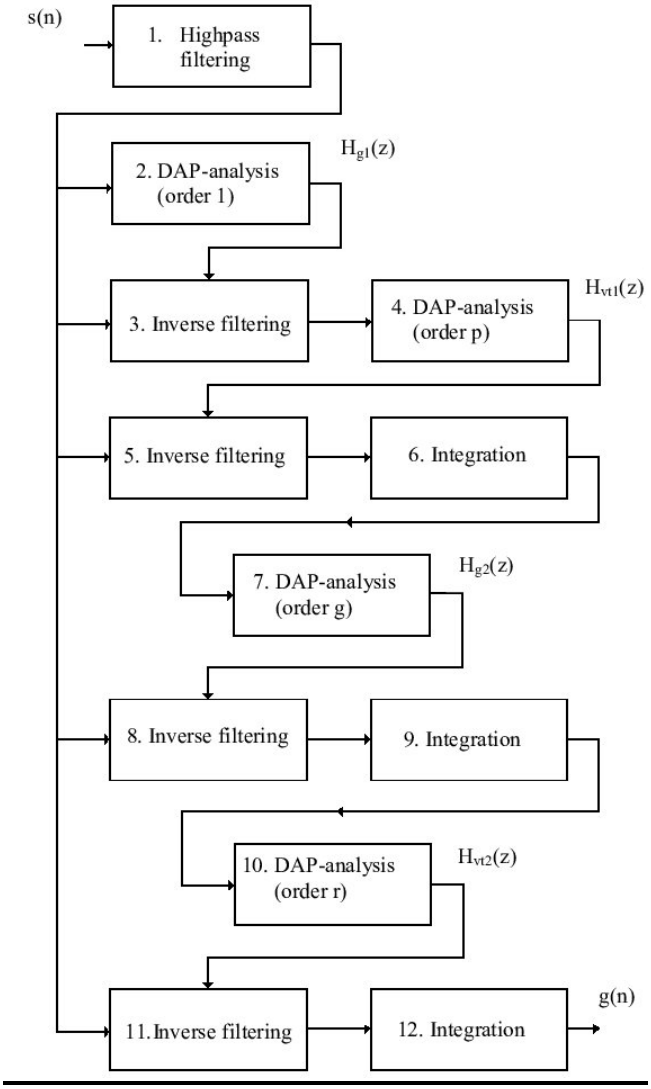


Figure 2.7: Flowchart of IAIF method for estimating the voice source excitation g(n) from an input speech signal s(n). p = (fs/1000) + 4 and g = (fs/2000), where fs is the sampling frequency. From Alku *et al.*(1999)

$$p = \left(\frac{fs}{1000}\right) + 4 \qquad (2.5)$$

where *fs* is the sampling frequency of the signal. This obtains an estimate of the effects of the vocal tract (H_{vt1}), which is then removed from the signal in the next inverse filtering

step (block 5). A glottal flow estimate is then acquired from the resulting signal by integrating (block 6), thus removing the lip radiation effect. The updated glottal flow estimate is then analysed using a g^{th} order LPC (block 7), where g is calculated as,

$$g = \left(\frac{fs}{2000}\right). \qquad (2.6)$$

An updated estimate of the vocal tract transfer function (H_{vt2}) is then created by removing the glottal flow effect (block 8), lip radiation (block 9), and performing a p^{th} order LPC analysis (block 10). A final inverse filtering (block 11) and integration (block 12) gives the final estimate of the glottal waveform.

As with the other inverse filtering methods described in this section, DAP modelling can be used in place of the LPC analysis to improve the modelling of the vocal tract filter. The study by Alku and Vilkman (1994) demonstrated that using DAP modelling in place of LPC improved the accuracy of modelling the vocal tract transfer function of voices with higher f_0 values. This resulted in estimated glottal pulses containing less formant ripple in their closed phases when compared to conventional LPC analysis.

The iterative optimal pre-emphasis (IOP) (Mokhtari and Ando, 2017) algorithm is an adaptation to the IAIF method which, when combined, forms the IOP-IAIF method of inverse filtering.

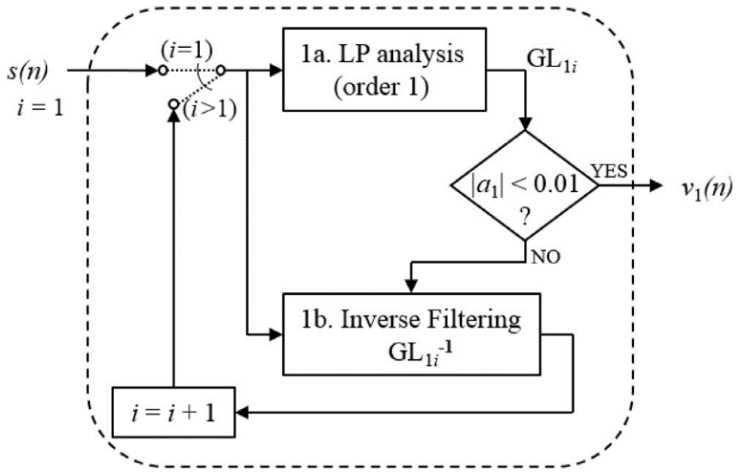


Figure 2.8 Block diagram of the IOP-IAIF method. From Mokhtari et al. (2018)

It replaces the initial LPC estimation and inverse filtering stage (blocks 2 and 3) of the standard IAIF process with an iterative algorithm that aims to provide a spectrally flat input to the first vocal tract modelling stage. It achieves this by performing a 1^{st} order linear prediction on the signal and checking the absolute value of the prediction coefficient, α_1, against a threshold. If that threshold is not met, then the signal is inverse filtered using

the value of α_1. This process is repeated until the threshold value is reached, which should mean that the output signal passing to the first vocal tract modelling stage is spectrally flat. The threshold value is set to 0.01, which is small enough to ensure that any further values would have an insignificant effect on the estimated spectrum, while keeping the number of iterations within a reasonable limit. Due to its design, this process attributes most of the spectral tilt of a speech signal to the glottal spectrum, while treating the vocal tract as a resonator with little to no spectral tilt. A block diagram of this process is shown in Figure 2.8.

Another proposed improvement to IAIF is the Iterative Adaptive Inverse Filtering with Glottal Flow Model (GFM-IAIF) approach (Perrotin and McLoughlin, 2017, 2019). Similar to IAIF-IOP, this method replaces the initial source modelling step of IAIF.

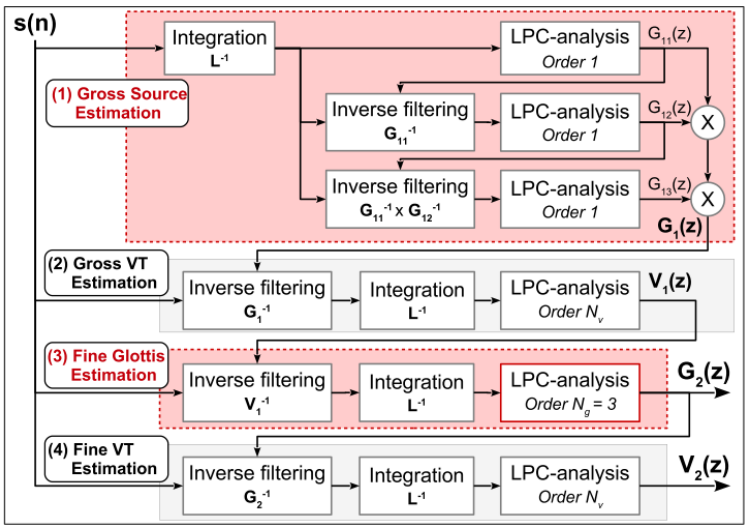


Figure 2.9: Flowchart of the GFM-IAIF method. The stages that differ from standard IAIF are highlighted in red. From Perrotin and McLoughlin (2019)

Figure 2.9 shows the architecture of this method. The speech signal is first integrated to remove the effects of lip radiation. The integrated signal is then passed through 3 successive 1st order LPC analysis and inverse filtering iterations, rather than just the one used in standard IAIF or the threshold dependent number in IOP-IAIF (Although the first integration step would be effectively undone by the first 1st order LPC analysis and inverse filtering stage, it was included so that the overall process matched the description of speech production presented in Section 2.2.). The signal is then passed to the initial vocal tract modelling step. During the second source estimation ($G_2(z)$) step, a 3rd order LPC analysis is carried out to account for the characteristics of the glottal flow spectrum. The results presented by Perrotin and McLoughlin (2019) show this method to perform equally well

as IAIF and IOP-IAIF at estimating the low frequency regions of the glottal source spectrum, but to outperform the other methods at higher frequencies and relative to a range of voice quality variations.

2.3.1.4 Quasi-closed phase inverse filtering

The Quasi-closed phase (QCP) method of inverse filtering (Airaksinen *et al.*, 2014) involves performing closed phase analysis over several glottal pulse frames using weighted linear prediction (WLP) (Ma *et al.*, 1993). The effects of the open portion of the glottal pulse are lessened by applying an attenuated main excitation (AME) weight function (Alku *et al.*, 2013), producing an estimate of the vocal tract transfer function with fewer effects of the source excitation. This approach works on the assumption that, during the closed phase, there is minimal coupling between the source and the speech signal (Plumpe *et al.*, 1999), thus, the speech signal is mainly defined by the resonances of the vocal tract.

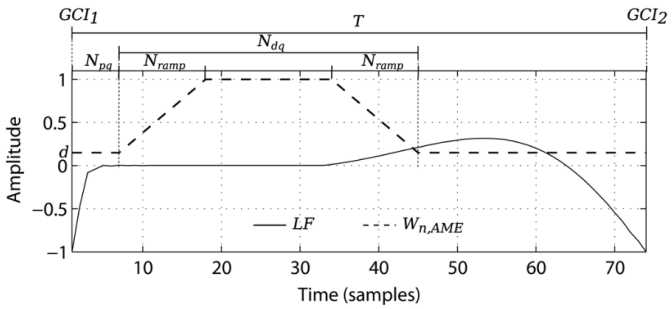


Figure 2.10: LF-model waveform and AME weight function between two GCIs. From Airaksinen et al.(2014)

The AME function (see Figure 2.10) is constructed for each glottal cycle and consists of a region of N_{dq} samples with values of 1 and a region of N_{pq} samples close to the GCI where it reaches a small positive value, *d*, close to zero. To avoid discontinuities, a gradual linear ramp, of N_{ramp} samples, is used to transition between these regions. When this function is applied to the speech signal, it attenuates the effects of the open phase, while leaving the closed phase mostly unaffected. This allows for the autocorrelation method of LPC analysis to be used, which requires a longer analysis window, but yields more consistent results when compared to the covariance method (Rabiner and Schafer, 1978). Figure 2.11 shows a block diagram of the QCP analysis process.

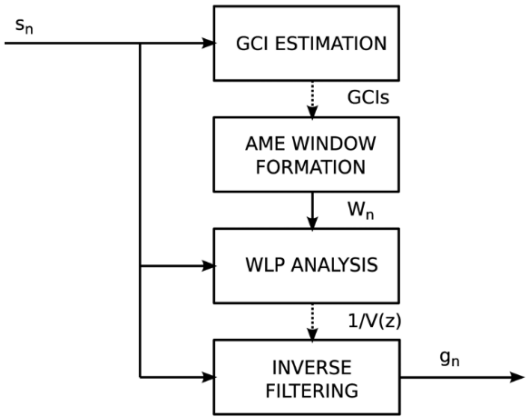


Figure 2.11: Block diagram of the QCP method. From Airaksinen et al.(2014)

The QCP process consists of four steps. The first step is GCI detection. The estimated GCIs are then used to construct the AME functions. WLP is then carried out on the pre-emphasised speech signal to estimate the vocal tract model. The vocal tract model is then used to obtain an estimate of the differentiated glottal flow by inverse filtering the speech signal. The results presented in Airaksinen *et al.* (2014) show that QCP outperforms four commonly used methods of inverse filtering, including IAIF and closed-phase inverse filtering, in several test cases, although its dependence on GCI positioning make it less reliable in real-world scenarios, where GCI estimation might be less accurate.

Although the methods described in this section may be sensitive to external noise and distortions, they provide a reasonably robust method of estimating the voice source without being invasive. These attributes are vital when processing large amounts of data for applications in speech technology, such as the work carried out in this thesis.

2.3.2 Direct measures of the glottal source

Some methods for analysing the voice source do not estimate the signal, but directly measure it. Although these methods can give very accurate measures of the voice source, they are impractical for collecting large amounts of data. Even still, applying these approaches can improve our understanding of the voice source in ways that inverse filtering techniques simply cannot, due to the fact that they provide estimates rather than direct measures of the voice source.

Sondhi Tube

The technique described by Sondhi (1975) involves a speaker producing a vowel into a 1.8 metre long brass tube with a fibreglass wedge at the other end to suppress reflections. This

long tube reduces the resonant effects of the vocal tract on the source signal without the need to perform inverse filtering. This method can also be carried out in a room with no acoustic treatment because the recording microphone is enclosed within the tube, making it robust to outside noise. Downsides of this technique include the facts that it is only suitable for static vowel sounds, and obviously cannot be applied to pre-recorded audio.

Pressure transducers

Another way with which the voice source signal can be measured directly is by using pressure transducers introduced into the larynx in several locations using a catheter or by some other means. This method can be considered quite invasive as it is an uncomfortable process for the participant, involving medical supervision and local anaesthetic. One such study that employed this technique is Cranen and Boves (1985). Another method involves measuring oesophageal pressure using a small balloon placed in the oesophagus directly below the vocal folds (Ladefoged, 1972), although this method yields an approximation of the voice source rather than a direct measure.

2.3.3 Glottal source models

Once the glottal source is estimated using inverse filtering, or another approach, it is useful to fit a parametric model to this waveform so that it can be quantified by easily representable numerical values. It also allows for the production of an excitation signal that closely matches the natural voice source, for resynthesis purposes, while also giving the option to transform this signal. This allows for modifications of voice quality and speaker characteristics. The following section briefly describes several commonly used glottal source models.

2.3.3.1 Rosenberg model

The Rosenberg B glottal pulse (Rosenberg, 1971)(see Figure 2.12) is comprised of the polynomials,

$$g_R(t) = \begin{cases} \alpha\left[3\left(\frac{t}{T_p}\right)^2 - 2\left(\frac{t}{T_p}\right)^3\right] & for\ 0 \le t \le T_p \quad opening\ \ phase \\ \alpha\left[1 - \left(\frac{t - T_p}{T_N}\right)^2\right] & for\ \ T_p < t \le\ T_p + T_N \quad closing\ \ phase \end{cases} \qquad (2.7)$$

where T_{p} is the percentage of the pulse with a positive slope that corresponds to the opening phase, T_{N} is the percentage of the pulse with a negative slope that corresponds to the closing phase, and α is the peak amplitude of the pulse.

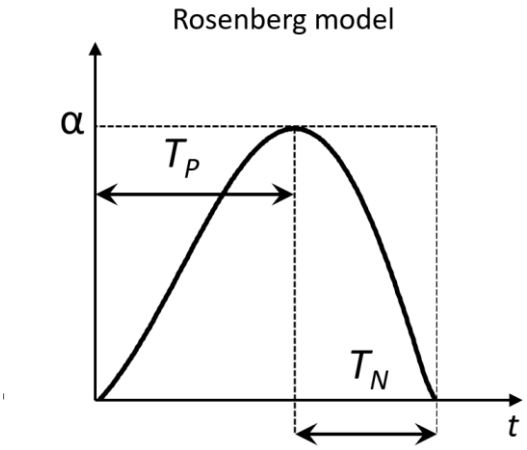


Figure 2.12: Illustration of the Rosenberg model

This model was ranked the highest, in terms of preference scores, among a set of difference pulse shapes, including triangular, trigonometric and trapezoidal shapes (Rosenberg, 1971). This model has no return phase and a set relationship for deriving T_{p}, which heavily limits its flexibility (Veldhuis, 1998).

2.3.3.2 LF-model

The Liljencrants-Fant (LF) model (Fant *et al.*, 1985) is a four-parameter model of the differentiated glottal flow. Figure 2.13 shows an example of two LF-model pulses (bottom) and their corresponding integrated forms (top).

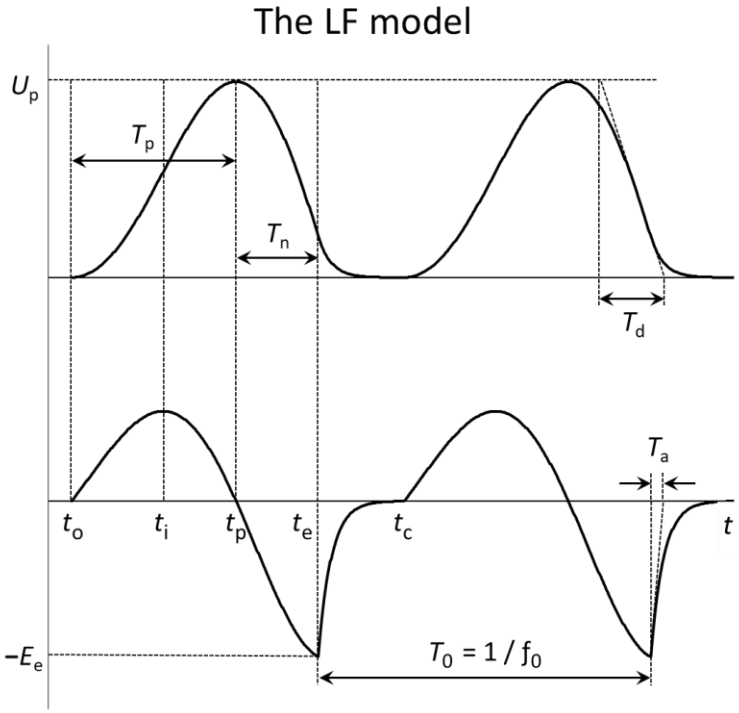


Figure 2.13: Example of two LF-model pulses (bottom) and their corresponding integrated forms. Adapted from Gobl (2003)

The model is calculated in two phases, the open phase (an exponentially growing sinusoidal function) and the return phase (an exponential function) using the following equations,

$$g'_{LF}(t) = \begin{cases} E_0 e^{\alpha t} sin\omega_g t & for\ t_0 \le t \le t_e \quad open\ phase \\ \frac{-E_e}{\varepsilon T_a}\left(e^{-\varepsilon(t-t_e)} - e^{-\varepsilon T_b}\right) & for\ \ t_e < t < t_c\ \ return\ phase \end{cases} \qquad (2.8)$$

where ω_g is $\frac{\pi}{T_\mathrm{p}}$ and $T_\mathrm{b} = t_c - t_e$. E_0 is a scaling factor that is needed to achieve area balance (the absolute area of the two phases are equal) in the model, and α is a coefficient that determines the rate of amplitude increase in the open-phase sinusoid. The constant ε in the return-phase can be derived iteratively using Equation (2.9).

$$\varepsilon = \frac{1}{T_a}(1 - e^{-\varepsilon T_b}). \qquad (2.9)$$

A detailed description of calculating α and ε can be found in Gobl (2003, 2017), see also Fant *et al.* (1985a).

Another way in which the LF-model pulse can be characterised is by using the three glottal R-parameters: R_a, R_k, and R_g. R_a describes the return-phase and the degree of high-frequency attenuation in the model spectrum. R_a is linked to the amount of *dynamic*

leakage, or the airflow present during the return-phase (Gobl, 1988) and is calculated using,

$$R_a = \frac{T_a}{T_0} \tag{2.10}$$

where T_a is the duration of the return phase. R_g is the f_0 normalised glottal frequency and is defined as,

$$R_g = \frac{T_0}{2T_p} \tag{2.11}$$

where T_p is the duration from the glottal opening, t_o, to the time point of the peak amplitude of the LF glottal flow, t_p. R_k is a measure of the glottal skew and is defined as,

$$R_k = \frac{t_e - t_p}{t_p} \tag{2.12}$$

where t_e is the time point of the main excitation.

Through analytical studies, it was found that many of these voice source parameters tend to covary (Gobl, 1988; Pierrehumbert, 1989; Fant, 1993) which led to the development of an alternate method of expressing and controlling the LF-model using a global waveshape parameter, R_d (Fant, 1995). This parameter describes the overall shape of an LF-model pulse, and is related to the effective pulse declination time (see Figure 2.13),

$$T_d = \frac{U_p}{E_e} \tag{2.13}$$

where U_p is the peak glottal flow and E_e is the maximum negative value of the LF-model pulse. R_d can be expressed as,

$$R_d = 1000 \cdot \frac{U_p}{E_e} \cdot \frac{f_0}{110} = \frac{1}{0.11} \cdot \frac{T_d}{T_0} \tag{2.14}$$

where T_0 is the pulse period and the scaling factor of 1000/110 means that the value of R_d is equal to T_d in milliseconds when f_0 is 110 Hz. The R_d value can then be used to predict values of R_a and R_k (i.e. R_{ap} and R_{kp}) with the following equations derived using statistical analysis performed on data extracted from a set of vowels and voiced consonants (Fant and Kruckenberg, 1994),

$$R_{ap} = (-1 + 4.8R_d)/100 \tag{2.15}$$

$$R_{kp} = (22.4 + 11.8R_d)/100. \tag{2.16}$$

Based on a geometrical simplification of the LF waveform, the relationship between R_d, R_a, R_k and R_g can be expressed approximately as (Fant, 1997):

$$R_d = \frac{1}{0.11} \cdot (0.5 + 1.2R_k)\left(\frac{R_k}{4R_g} + R_a\right) \tag{2.17}$$

This equation can then be rewritten to predict R_g,

$$R_{gp} = \frac{R_{kp}}{4\left(\frac{0.11R_d}{0.5 + 1.2R_{kp}} - R_{ap}\right)}. \qquad (2.18)$$

The normal range for R_d is between 0.3 and 2.7, while values above 2.7 produce a very abducted pulse shape suitable for modelling pre-pause voice terminations (Fant, 1995). R_d has been shown to be a very effective parameter for describing voice qualities and simplifying control of the voice source in text-to-speech (TTS) synthesis (Fant, 1995). A more detailed description of how this parameter can be implemented in speech synthesis can be found in Section 2.3.5.

2.3.3.3 FL model

The Fujisaki-Ljungqvist (FL) model (Fujisaki and Ljungqvist, 1986)(see Figure 2.14), $g_{FL}(t)$, consists of 4 polynomial segments (see Equation (2.19)).

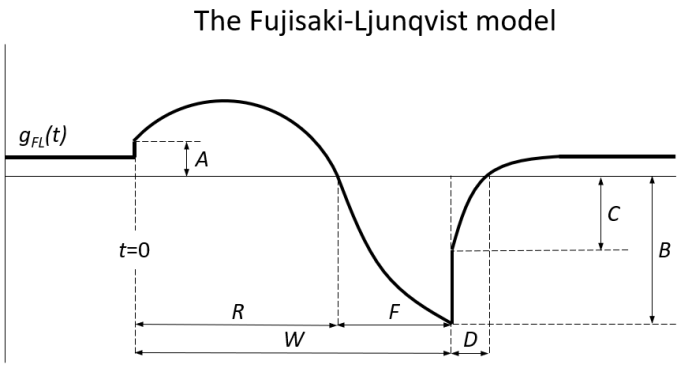


Figure 2.14: Illustration of the Fujisaki-Ljunqvist model. Adapted from Fujisaki and Ljungqvist (1986).

It is controlled using 3 timing parameters: open phase duration, *W*, pulse skew, *S*, and the time from glottal closure to the maximum negative flow, *D*, and 3 amplitude parameters controlling slope at glottal opening, *A*, slope prior to closure, *B*, and slope following closure, *C*.

$$g_{FL}(t) = \begin{cases} A - \frac{2A + R\alpha}{R} + \frac{A + R\alpha}{R}t^2, & 0 < t \le R \\ \alpha(t-R) + \frac{3B - 2F\alpha}{F^2}(t-R)^2 - \frac{2B - F\alpha}{F^3}(t-R)^3, & R < t \le W \\ C - \frac{2(C-\beta)}{D}(t-W) + \frac{C-\beta}{D^2}(t-W)^2 & W < t \le W + D \\ \beta & W + D < t \le T \end{cases} \qquad (2.19)$$

$$where\ \alpha = \frac{4AR - 6FB}{F^2 - 2R^2}\ and\ \beta = \frac{CD}{D - 3(T - W)}, T = fundamental\ period.$$

What makes this model different from the others described in this section is that one of the amplitude parameters controls the slope of the waveform at glottal opening. This allows for the inclusion of discontinuities at the glottal opening and closing.

2.3.3.4 Other glottal source models

The Rosenberg++ (R++) model (Veldhuis, 1998) is a computationally more efficient alternative to the LF-model, based on an extension of the Rosenberg model. It uses the same time and R-parameters as the LF-model and can be used to produce perceptually equivalent synthetic speech. Although it is more computationally efficient the LF-model is used by researchers more frequently.

Shue and Alwan (2010) proposed a glottal flow model (see Figure 2.15), based on observations made from imaging of the glottis.

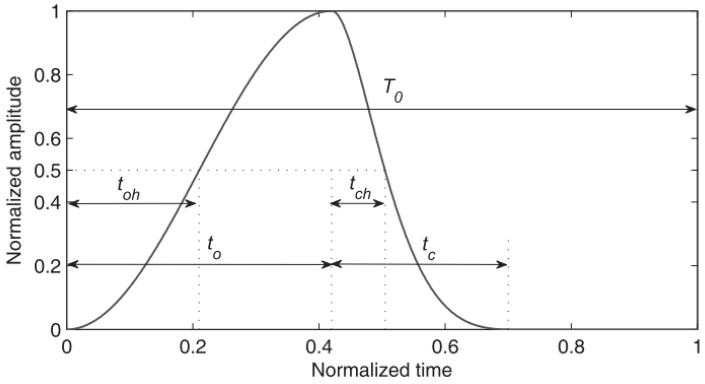


Figure 2.15: Glottal model proposed by Shue and Alwan. From Shue and Alwan (2010)

The model is derived from the integrated LF-model and includes a higher level of flexibility in defining the opening and closing phase durations. It has five control parameters: the fundamental period, T_0, open quotient, OQ, asymmetry coefficient (α), speed of opening phase, S_{op}, and speed of closing phase, S_{cp}. The values are derived from the times shown in Figure 2.15. $OQ = \frac{t_o t_c}{T_0}$, $\alpha = \frac{t_o}{t_o + t_c}$, $S_{op} = \frac{t_{oh}}{t_o}$ and $S_{cp} = 1 - \frac{t_{ch}}{t_c}$, where t_{ch} and t_{oh} are at 50% of the amplitude of the model pulse. The model is defined as:

$$g_{SA}(t) = \begin{cases} f\left(\beta_o t, \lambda_{S_{op}}\right), & 0 \le t \le \alpha OQ \cdot T_0 \\ f\left(\beta_c (OQ \cdot T_0 - t), \lambda_{S_{cp}}\right), & \alpha OQ \cdot T_0 < t \le OQ \cdot T_0 \\ 0, & OQ \cdot T_0 < t \le T_0 \end{cases} \qquad (2.20)$$

where

$$f(x, \lambda^*) = A(\lambda^*)\left[e^{\lambda^* x}(\lambda^* \sin(\pi x) - \pi \cos(\pi x)) + \pi\right] \qquad (2.21)$$

and

$$\lambda^* = arg_\lambda min A(\lambda^*) \left| \frac{e^{\lambda s}(\lambda\, sin(\pi s) - \pi\, cos(\pi s))}{\pi(e^\lambda + 1)} + \frac{1}{e^\lambda + 1} - \frac{1}{2} \right| \quad (2.22)$$

with $A(\lambda^*) = \left(\pi\left(e^{\lambda *} + 1\right)\right)^{-1}$, $s = S_{op}$ or S_{cp}, $\beta_o = (\alpha OQ \cdot T_0)^{-1}$, and $\beta_c = \left((1-\alpha)OQ \cdot T_0\right)^{-1}$. Results from a model fitting test showed that this model outperformed the LF-model, in terms of a mean squared error criterion, on the test material used (Shue and Alwan, 2010).

A perceptual evaluation of voice source models carried out by Kreiman *et al.*(2015) found that synthetic stimuli generated using the LF-model, and Shue and Alwan models were perceptually closer to target stimuli than the FL and Rosenberg models. This, in addition to its extensive use in the literature and lower number of control parameters are reasons that the LF-model was selected for use as the basis for synthetic voice source signals in this work. The latter reason, of having few control parameters, aligns with one of the main goals of this work, to investigate further reduction of model parameter dimensionality for voice quality control leading to possible control of prosodic features.

2.3.4 Parameterising the voice source

There are several approaches that have been explored when it comes to parameterising the glottal source signal. These approaches usually involve either directly measuring the glottal source signal or fitting a parametric model to the glottal source signal and taking measurements from that. This section will describe several parameterisation methods proposed in the literature.

2.3.4.1 Direct measures

Direct measurements of the time domain glottal source can offer important temporal information about different events within the glottal cycle. The measurements are usually expressed in terms of ratios or quotients relative to the glottal cycle period. The three main time-domain parameters that are used are the open quotient (OQ) and speed quotient (SQ), both proposed by Timcke *et al.* (1958), and the closing quotient (ClQ) (Monsen *et al.*, 1978). OQ is the ratio between the open phase and the total period of the glottal pulse. ClQ is the ratio between the closing phase and the glottal period. SQ is the ratio between the opening and closing phases of the glottal pulse. These measurements are sensitive to formant effects still present in the signal after inverse filtering, and possible noise

introduced by recording equipment or environment, making it difficult to ascertain exact timing locations. A more robust approach has been found to use amplitude-based measurements that relate to the time domain parameters. The normalised amplitude quotient (NAQ) (Alku *et al.*, 2002) is defined as,

$$NAQ = \frac{f_{ac}}{d_{peak} \cdot T_0} \qquad (2.23)$$

where f_{ac} is the peak amplitude of the glottal pulse and d_{peak} is the maximum negative amplitude of the differentiated glottal pulse. This measure correlates closely with the ClQ and OQ. It can also be seen that this parameter is very similar to R_d: note that U_p is equivalent to f_{ac} and the value of E_e is in most cases the same as d_{peak} (see Equation (2.14).

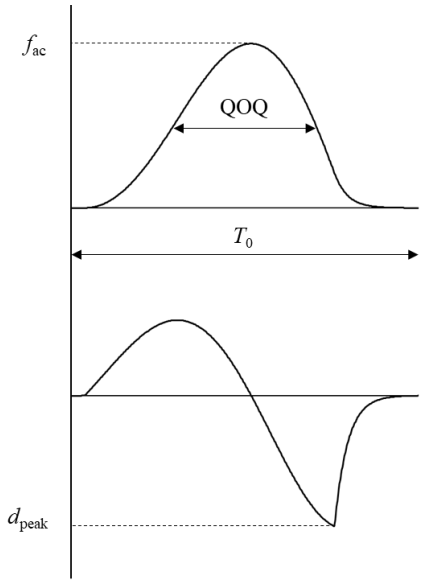


Figure 2.16: Glottal pulse and differentiated glottal pulse illustrating NAQ and QOQ measurements.

Another measure used is the quasi-open quotient (QOQ) (Hacki, 1989). It is measured by finding the peak amplitude of a glottal pulse and locating points to either side of it with 50% of the peak amplitude.

A frequency-domain measure often used involves calculating the difference in amplitude level between the first and second harmonic (H1-H2) in the spectrum of the differentiated glottal source.

Another frequency domain measurement is the so-called, harmonic richness factor (HRF) (Childers and Lee, 1991). This parameter is the ratio of the summed amplitudes of harmonics above the fundamental (usually 10) and the amplitude of the fundamental.

The main problem with taking direct measurements from the voice source is that they rely on accurate estimation of the voice source. Recording conditions and the method of inverse filtering used can degrade the accuracy of these measures.

2.3.4.2 Phase minimisation

Degottex *et al.* (2011) proposed a method of estimating R_d directly from the speech signal without the need for inverse filtering. This method is based on minimising a phase-based criterion called the mean squared phase difference (MSPD), independent of glottal pulse position. R_d can be derived by minimising the equation,

$$MSPD^2(\theta, N) = \frac{1}{N}\sum_{k=1}^{N}\left(\Delta^{-1}\Delta^2 \angle R_k^{\theta}\right)^2 \qquad (2.24)$$

where $N = |f_{lim}/f_0|$, f_{lim} is the Nyquist frequency, k is the harmonic number and θ is R_d. Computing this objective function is carried out using the second-order phase difference, Δ^2, and the anti-difference operator, Δ^{-1}, with the convolutive residual, R_k^{θ}. R_k^{θ} is the ratio of the spectrum of a speech segment, S_k, with that of a model spectrum. The model spectrum consists of an LF pulse spectrum, G_k^{θ}, and a minimum phase vocal tract model $\varepsilon_{\left(S_k/G_k^{\theta}\cdot jk\right)}$, where ε is a minimum phase version of the model spectrum and *jk* is the filter corresponding to lip radiation.

This method was further developed in Huber *et al.* (2012) and Huber and Roebel (2015), with the latter describing how the method was incorporated into a speech analysis-and-synthesis system.

A significant advantage to this approach is that it avoids the need to inverse filter the speech signal before carrying it out, avoiding possible errors in glottal flow estimation. A disadvantage is its reliance on the phase characteristics of the speech signal. Phase errors or distortions introduced by the recording environment or the equipment could lead to inaccuracies in the parameter estimation.

2.3.4.3 Model matching

Another method of parameterising the voice source is to match and fit models to each glottal pulse. The method described in Strik *et al.* (1993) and Strik (1998) fits LF-model to differentiated glottal flow pulses. This method involves first low-pass filtering the signal and then estimating GCI values. For each GCI a search is made for a nearby maximum

negative peak. The amplitude of this peak and the timepoint it occurs at are taken as E_e and t_e respectively (see Figure 2.13). Then, the closest zero-crossing point to the left of t_e is directly measured and taken as being t_p. The next point that is found is t_o, which is taken as the point to the left of t_p that falls below a certain threshold. T_a is determined by fitting an LF-model derived from the other already estimated parameters. An optimisation procedure is then carried out to refine the parameter estimates. Similar model-fitting approaches are used in the systems described in Kreiman *et al.* (2006) and Airas (2008).

The extended Kalman filtering method (EKFM), described in Li *et al.* (2011), estimates LF-model parameters (T_p, T_e, T_a, α and ε) by dividing a differentiated glottal pulse into an open and closed phase and applying a Kalman filter (Kalman, 1960) iteratively to each phase. The parameters are then optimised by minimising the mean square error between the open and closed phases of an LF pulse generated using the parameters and the original differentiated glottal pulse. The results of the evaluations carried out in Li *et al.* (2011) show that EKFM outperforms a standard time-domain approach (Airas, 2008) in terms of estimation accuracy and computational efficiency.

Another method of model matching, called DyProg-LF, is described in Kane and Gobl (2013). This approach consists of an exhaustive search of R_d parameter values, followed by a dynamic programming step, and finally, an optimisation step is carried out. The exhaustive search involves taking a three-pulse long GCI centred frame of the differentiated glottal source and calculating its amplitude spectrum. GCI centred synthetic source segments of three concatenated LF pulses for a range of R_d values (0.3-5 in steps of 0.1) are generated using f_0 and E_e values measured from the differentiated glottal source. The amplitude spectrum of each synthetic segment (corresponding to each R_d step) is calculated and an error measure is derived using the following:

$$spec_err = \{1 - |cor\{h_{g'}(l), h_{LF}(l)\}|\} \cdot w_s, 1 \leq l \leq L \in [0,1] \qquad (2.25)$$

Where $h_{g'}(l)$ and $h_{LF}(l)$ are the harmonic amplitudes of the differentiated glottal source and LF pulse segment respectively, L is the number of harmonics up to a maximum frequency of 3 kHz, $cor\{\cdot\}$ is the Pearson correlation between the harmonic amplitudes, and w_s is a weight constant. A time-domain error is also calculated using:

$$time_err = \{1 - |cor\{g', g_{LF}\}|\} \cdot w_t, \in [0,1] \qquad (2.26)$$

Where g' and g_{LF} are the current frame of the differentiated glottal flow signal and the synthetic LF segment respectively, and w_t is a weight constant. The two weight constants,

w_s and w_t, determine the relative importance each error value has in the dynamic programming step. The five candidates saved for the dynamic programming step are those that minimise the total error function:

$$total_err = \frac{spec_{err} + time_{err}}{2}, \in [0,1] \qquad (2.27)$$

The dynamic programming stage (Ney, 1983) is carried out to find the optimum path of R_d values across the speech signal. The target cost, *d(i,j)*, is the *total_err* calculated in the exhaustive search using Equation (2.27) for each of its LF candidates, where $1 \leq i \leq M$, $1 \leq j \leq N$, and M and N are the number of analysis frames and candidates respectively. The transition cost is expressed as:

$$\delta_{i,j,k} = \left\{1 - cor\left\{seg_{i.j}, seg_{i-i,k}\right\}\right\} \cdot w_{tr} \cdot ss, \in [0,1] \qquad (2.28)$$

where $seg_{i,j}$ is a single LF pulse generated using the R-parameters from the j^{th} R_d candidate of analysis frame *i*, and $seg_{i-1,k}$ is an LF pulse generated using the k^{th} R_d candidate of the previous analysis frame. w_{tr} and *ss* are the transition weight constant and the inverse of the Itakura distortion (Itakura, 1975b), referred to as the spectral stationarity. This *ss* measure tends towards 1 when adjacent frames are spectrally similar and goes towards 0 when they are different. This lessens the effect of the transition costs, $\delta_{i,j,k}$, in regions of rapid change, such as consonant to vowel transitions, or voice onsets and offsets.

Incorporating the target and transition costs, a function can be defined as,

$$D_{i,j} = d_{i,j} + \min_{k \in N}\left\{D_{i-1,k} + \delta_{i,j,k}\right\}, \quad 1 \leq j \leq N \qquad (2.29)$$

initialised with,

$$D_{o,j} = 0, \quad 1 \leq j \leq N \qquad (2.30)$$

The R_d candidates that minimise $D_{i,j}$ are saved.

Once the closest matching R_d values have been selected they are used to derive values for the R-parameters: R_a, R_k and R_g, which are optimised using a simplex-based method (Nelder and Mead, 1965). This method allows for multi-variable optimisation. The R-parameters are allowed to vary so that they minimise the error functions shown in Equations (2.25) and (2.26). The DyProg-LF method was found to outperform the traditional time-domain based methods (e.g. (Strik, 1998)) and a phase minimisation method of estimating R_d (Degottex *et al.*, 2011a) when estimating OQ of reference data.

2.3.4.4 Interactive glottal source analysis

A different approach to improving the accuracy of inverse filtering and voice source parameterisation involves manually adjusting values obtained from automatic analysis through an interactive interface. Several systems have been developed that employ this method (Ní Chasaide *et al.*, 1992; Gobl and Ní Chasaide, 1999b; Kreiman *et al.*, 2010; Kane and Gobl, 2013; Dalton *et al.*, 2014). These systems employ some kind of automatic inverse filtering e.g. closed-phase inverse filtering (Ní Chasaide *et al.*, 1992; Gobl and Ní Chasaide, 1999b) or IAIF (Dalton *et al.*, 2014) to provide an initial estimate of the vocal tract transfer function. The formant frequencies and bandwidths can then be manually tuned to cancel their effects using visual feedback from frequency and time-domain plots.

Once the inverse filtering process has been completed an automatic method initially fits the LF-model to the pulses of the estimated differentiated glottal waveform. The user can then manually optimise the model fit by adjusting its amplitude and time parameters.

Although this method of glottal source analysis can provide very accurate results, it is incredibly time consuming. Each glottal pulse must be analysed individually where even a short one second utterance will usually contain at least one hundred pulses. Another problem with this approach is that the user needs to have experience of this type of analysis, and even with this, there is a certain level of subjectivity involved when a user is presented with an ambiguous signal to filter or model to fit.

2.3.5 R_d as a control parameter for speech synthesis

As described in Section2.3.3.2, the shape of the LF-model may be described by three *R*-parameters R_a, R_k, and R_g. The global waveshape parameter, R_d, decreases the number of parameters required to define an LF pulse's shape to one (Fant, 1995). This reduction in dimensionality is of benefit when trying to parametrise a glottal source pulse as it is easier to determine a single best matching parameter than several (Gobl, 2003). Another advantage in reducing the parameter set that describes the source is in the development of simple, robust systems for allowing naive user control of synthetic glottal source signals, which is an aim of this work.

The equations for mapping R_d to the other *R*-parameters (Fant *et al.*, 1994; Fant, 1995), were determined using linear regression analysis on *R*-parameters values extracted from real speech in Gobl (1988) and Karlsson (1990). These analyses captured the natural

covariation of the glottal shape parameters allowing them to be controlled with the R_d parameter.

The voice qualities resulting from the pulse shapes determined by R_d tend along a continuum of tense to lax voice. Lower values of R_d correspond to tenser voice, high values correspond to laxer voice, while values in the middle of the range correspond to modal voice. This makes it a very useful parameter for controlling prosodic features such as focus, where, studies have suggested, speakers can use shifts to tenser phonation to signal focally accented syllables and laxer phonation for post-focal material (Yanushevskaya *et al.*, 2010, 2016b). Focus, in this case, refers to the highlighting of a particular unit in a sentence, through some level of prominence, an accent for example.

The study in Degottex *et al.* (2012) describes how the R_d parameter can be successfully used in the modification of breathiness in a HMM-based synthetic voice, while another study used variations in R_d to produce voice qualities from very tense to very lax (Huber and Roebel, 2015). The R_d parameter has also been used to modify voice qualities in concatenative synthesis systems. The work described in Sorin *et al.* (2017) used R_d manipulations as a means to transform the voice quality of synthetic speech by adjusting the source signal of voiced speech in a concatenative speech synthesis system, so that it became, what they called, semi parametric. Yanushevskaya *et al.* (2017) carried out a principle component analysis on various voice source parameters and found that R_d was important in describing cross-speaker differences in voice quality.

The results of the studies outlined in the above paragraph reinforce the idea of using R_d as the main control parameter in a minimal set for controlling voice quality in an interactive system. Minimising the set of control parameters is helpful in systems that will be used by non-experts, disabled users or in instances where the application is being used through a simple or constrained interface.

The concept of using R_d as a control parameter for linguistic and paralinguistic prosody will first be tested through a set of perception experiments based on manually inverse filtered data where R_d is the only parameter that is being manipulated. This will provide knowledge of the way R_d must be manipulated in order to elicit perceived prosodic changes, or whether or not this can even be achieved in the absence of other parameter effects, especially f_0.

The knowledge gained from the results of the perception experiments will then be used to develop a speech analysis-and-synthesis system that allows a user to easily control the prosody of synthetic speech manually by manipulating R_d and other voice source parameters, as well as vocal tract characteristics.

In order to generate unseen synthetic speech outputs for applications such as educational games or expressive TTS programmes/screen readers, the analysis-and-synthesis system needs to be integrated into an SPSS system. This is another advantage to using a small set of parameters because it is favourable to keep the input parameter sets of such system as small as possible to reduce analysis, training and synthesis times.

The effectiveness of using the R_d parameter in the analysis and synthesis, and SPSS systems will also be demonstrated through a set of perception tests, including user driven tasks, with control interfaces developed as part of this work.

2.3.6 Selection of model and methods

Based on the review of the literature the following approaches were used for the work carried out in this thesis.

The GFM-IAIF approach of inverse filtering with DAP was selected due to its robustness and greater accuracy at estimating the frequency response of the vocal tract. GFM-IAIF has been shown to outperform other methods of IAIF (Perrotin and McLoughlin, 2019).

The LF-model was chosen to parameterise the differentiated glottal flow because of its wide use in literature and the fact that it can be expressed using the global waveshape parameter, R_d. This is beneficial when it comes to implementing control interfaces for voice quality transformation, where a more complex multidimensional description would be confusing to a user and might require expert knowledge to operate.

The transformed LF-model was used in this work due to the fact that using a single waveshape parameter minimises the complexity of parameter selection and user control of voice quality in speech synthesis applications.

The model matching approach described in Kane and Gobl (2013) was used to estimate this parameter from the differentiated glottal flow signals obtained from inverse filtering because it has been proven to perform better than traditional time-domain based and phase

minimisation methods and is better suited to parameterise large speech corpora than interactive methods.

2.4 Chapter conclusions

This chapter described aspects of the speech production system that are important in the context of understanding topics discussed in further chapters. The chapter began by outlining the anatomy related to the speech production system, then going into greater detail for the main structures involved. The process of vocal fold vibration, the resulting voice qualities, their function in linguistic and paralinguistic contexts and the vocal tract were discussed. The acoustic theory of speech production was also introduced, describing how speech can be represented as a source signal being passed through a vocal tract filter with the additional effects of lip radiation. Applying this model of speech production means that the individual components may be separated and analysed using digital signal processing techniques.

This chapter also reviewed some of the main approaches for analysing and parameterising the voice source. The chapter also included a description of several glottal source models and how the waveshape parameter of the LF-model, R_d, can and has been used as a control parameter for voice quality in speech synthesis. The chapter finishes by listing the selected methods and model to be used in this work, based on their benefits and how well they fit this work's aims.

Chapter 3. Speech synthesis

With the advancements in technology over the past several decades, it has been possible to use analogue and digital electronics to synthesise speech. While vast improvements have been made to the naturalness, intelligibility and quality of synthesised speech, it has still yet to reach a level of expressivity comparable to the human voice. It is an important factor when considering the main applications of speech synthesis, such as assistive screen-readers and text-to-speech programs for individuals with sight, speech or reading impairments, educational programs and games, and language processing programs on mobile devices, when being used in combination with speech recognition software. In all these applications the ability to synthesise speech with an expressivity and naturalness comparable to human speech would be a huge advantage. It is for this reason that the ultimate goal of speech synthesis research is to develop a method of synthesis that can closely replicate the qualities of natural speech, while still remaining practical in terms of computational and memory costs. While achieving this high level of natural expressiveness is of utmost importance, another crucial element to improve the functionality of speech synthesis is the implementation of a way with which to control this expression. This can only be achieved through a high degree of parametric flexibility.

3.1 Methods of speech synthesis

The following sections contain a brief description of the speech synthesis methods that are relevant to the research being carried out. The approaches discussed in this chapter can be separated into two main categories, rule-based and corpus-based. Rule-based approaches rely on understanding the principles of how the speech production system operates and modelling these principles in order to generate speech. Formant and articulatory synthesis are examples of rule-based approaches. On the other hand, corpus-based approaches utilise recordings of natural speech, directly using the audio waveforms, or parameterising and modelling them to train a statistical framework for example. Concatenative, statistical parametric and end-to-end speech synthesis are examples of corpus-based approaches. Each method tackles the problem of accurately synthesising natural speech in quite a different way and some of the advantages and disadvantages will be discussed.

3.1.1 Formant synthesis

Formant synthesisers rely on a set of rules that describe the acoustic parameters of speech and the way that they vary over time. These rules must first be derived by experts through analysis of speech data and knowledge of the natural limits of speech production. These rules describe the way in which the acoustic parameters, such as f_0 and formant/anti-formant values, change over time. For voiced speech, a periodic signal is used to represent that glottal source. This can take the form of a simple pulse train, but more sophisticated glottal source models have been utilised to obtain much more natural sounding synthetic speech (Klatt, 1980). Noise is used as the source signal for unvoiced speech.

Formant synthesis can produce perfectly intelligible speech, but the form of parameterisation used leads to somewhat 'processed' sounding speech that is easily identified as being synthetic. The rules that control the parameters also require expert linguistic knowledge when being defined. This is not always feasible for applications where many voices are required, due to time and cost constraints.

3.1.2 Articulatory synthesis

Articulatory speech synthesis is the method of speech synthesis that most closely models the human speech production process. The very first speech synthesiser, created by von Kemplen in the 18th century (von Kempelen, 1791), was an example of an articulatory synthesiser. It used a bellows in place of the lungs, and tubing, boxes and pipes in place of the rest of the speech production system. It could not produce actual speech, but speech-like sounds. More recent synthesisers using this method model the physical processes involved in the articulation of speech and how they change with time digitally (see examples in (Rubin *et al.*, 1981; Birkholz, 2005; Birkholz *et al.*, 2015). In theory, this method should produce the most natural sounding speech, and it does so very well for static vowel sounds, but difficulties arise when it comes to obtaining the data to describe the dynamic movements and processes of the speech production system. This task is made easier with modern imaging techniques, such as magnetic resonance imaging (MRI) and ultrasound. There are still difficulties when trying to synthesise continuous speech due to problems when attempting to model intricate coarticulation effects. These lead to an overall lowering of naturalness in generated speech and make it less viable for implementation in commercial TTS systems. There is, however, a very vital need for this

method in the study of speech production, articulator physiology and audio-visual synthesis (Bailly *et al.*, 2003; Taylor, 2009).

3.1.3 Concatenative synthesis

Concatenative speech synthesis builds speech by combining short segments of recorded speech selected from a corpus. One form of concatenative speech synthesis is diphone synthesis. This method involves constructing a collection of all the diphones (pairs of phones) that exist in a language, and concatenating them at synthesis time using some form of digital signal processing (e.g., linear prediction, pitch-synchronous overlap and add (PSOLA) (Hunt *et al.*, 1989; Moulines and Charpentier, 1990; Moulines and Verhelst, 1995; Compernolle and Wambacq, 2000; Hamon *et al.*, 2003)).

Another method that is commonly used today is unit selection synthesis (Black and Campbell, 1995; Hunt and Black, 1996; Black and Taylor, 1997). Rather than using diphones, this method utilises a corpus containing a whole range of speech units, from whole sentences and phrases, down to sub-phone segments. These units are then clustered according to similar acoustic characteristics and linguistic contexts. At synthesis time, the best units are selected from the corpus by minimising a target cost associated with each unit and a concatenation cost between the selected units. The target cost is related to how well a unit fits in terms of acoustic and linguistic context. The concatenation cost is related to how well a particular unit matches the other units in the utterance to be synthesised.

The relatively small amount of processing of the speech samples used in this method results in the most natural-sounding synthesised speech, but this naturalness comes at a cost. The use of recorded audio samples as the base unit for this synthesis means that it requires a large, memory-heavy database from which to work from. This can be somewhat of a restriction when it comes to applications based on embedded or mobile devices that only possess a limited amount of internal memory. Another issue with this synthesis method is that there is very limited control of the voice characteristics. The characteristics of the speech being synthesised will generally be similar to that of the database being used. A further issue is that although the speech may sound very natural, distortions may occur at the concatenation points between the speech units. This is caused by the differences between the phonetic units taken from different words, which can lead to prosodic and acoustic mismatches when the units are placed in the wrong context or come from different

utterances. These distortions can be quite jarring to a user and are undesirable in applications where smooth speech at a high playback speed might be required (e.g. screen-readers). On the other hand, applications that require more natural-sounding speech (e.g. video games) can take full advantage of the high-quality of concatenative speech synthesis when implementing character voices with little concern over playback speed.

3.1.4 Statistical parametric speech synthesis

Statistical parametric speech synthesis (SPSS) uses statistical models to produce a best-fit speech sound by averaging out a set of similar sounding sections of speech (Zen *et al.*, 2009). This approach is a relatively recent development in speech synthesis and can produce high quality speech that is both intelligible and fully parametric. Up until recently the statistical models used in this form of synthesis were based on hidden Markov models (HMM), but now more commonly deep neural networks (DNN) are used in their place.

Most SPSS systems consist of three stages. The first stage involves analysing and extracting acoustic parameters from a speech corpus using a vocoder. These parameters are then used in the second stage, where statistical models are trained (e.g. HMMs, DNNs, etc.). These models can then generate parameters corresponding to the desired sequence of phonemes specified by the synthesiser front-end (text to phoneme interpreter). Some systems then use a smoothing method to produce more natural parameter trajectories (Tokuda *et al.*, 2000). These parameters are then used by the vocoder to generate the synthetic speech waveform.

SPSS systems require less speech material than unit-selection systems and can produce speech not present in the corpora they are built with, due to the modelling methods used. They also have a much smaller footprint, meaning they can be used on embedded devices with low resources. The approach taken with SPSS also means that there is the possibility of transforming the synthetic speech. Yamagishi *et al.* (2009) describes a method that allows for the transformation of a synthetic voice so that it sounds like a different speaker. Speaker styles and emotions can also be transformed as described in Yamagishi *et al.* (2004) and using DNNs as described in Takaki *et al.* (2016).

SPSS can be used to produce more natural-sounding speech than some rule-based synthesisers and is now coming close to comparing with state-of-the-art unit-selection concatenative speech synthesis. The level of naturalness is mainly dictated by the way the

speech is modelled. Earlier speech modelling techniques, such as the Pulse-train or noise model (see Section 3.3.1), produced very *buzzy* sounding speech due to the simplistic way in which the source is represented. More sophisticated modelling methods have been proposed that improve the naturalness of speech synthesised using SPSS (see Sections 3.3.2-3.3.5). The HMM-based Speech Synthesis System (HTS) (Tokuda *et al.*, 2002) and Merlin (Wu *et al.*, 2016) are widely used for SPSS.

3.1.5 End-to-end speech synthesis

End-to-end speech synthesis involves directly modelling a speech waveform from text. Systems based on this method seek to reduce or remove the need for expert intervention in defining linguistic rules and annotation and allow deep learning approaches to take care of these matters. WaveNet (Oord *et al.*, 2016) is a deep neural network architecture for generating raw audio and speech signals. It consists of convolutional neural networks (CNN) that can be conditioned (i.e. trained) from acoustic or linguistic features of speech and can be integrated into TTS systems. Other end-to-end systems use some kind of intermediate representation of acoustic features. Char2Wav (Sotelo *et al.*, 2017) uses traditional vocoder parameters as the output of a recurrent neural network (RNN), which are then used by a *neural vocoder* (Mehri *et al.*, 2016) to produce the speech waveform. Tacotron (Wang *et al.*, 2017) and Tacotron 2 (Shen *et al.*, 2018) both use spectrograms as their intermediate features, with Tacotron 2 adding a WaveNet based vocoder to convert the spectrograms into audio signals. These systems can produce very natural sounding speech but require a large amount of speech to train them on. The direct mapping from text to speech also makes it difficult to allow for parametric control of voice quality.

3.1.6 Conclusions

Each of these methods of speech synthesis have their own advantages and disadvantages. While both formant and articulatory synthesis are incredibly flexible in terms of parametric control, defining these parameters, and the rules that govern how they change over time is very difficult. These synthesisers tend to have a ‘processed’ characteristic to their voices that can easily distinguish them from natural speech.

Concatenative-based unit selection synthesis systems can produce very natural sounding speech but are very limited when it comes to controlling speaker characteristics and voice quality. These systems can also suffer from concatenation errors between speech units that

can be detrimental to their use in applications where a high speech rate is required. The speech database required for unit-selection voices is also quite large, which can limit their application.

The end-to-end approach to speech synthesis can produce extremely natural speech, but the implementations are less interested in implementing control of voice quality characteristics, and more in producing a very natural and intelligible speech.

SPSS systems offer a high level of parametric control, but also the ability to produce natural sounding synthetic speech without the need for complex previously designed rules. They also require a very small amount of memory in comparison to unit-selection systems. That being said, there are still many improvements that could be made to narrow the gap between the somewhat buzzy voices that they produce, and natural speech. The main benefit of this method in the context of the current work is that the parametric control allows for the manipulation of voice source and vocal tract filter parameter with relative ease.

3.2 DNN-based statistical parametric speech synthesis

Up until quite recently, most SPSS systems used HMMs as their main statistical modelling component. With the increases in computational power available to researchers and the use of graphical processing units (GPUs) to speed up calculations, DNNs are now a viable alternative that can yield better results.

3.2.1 Artificial neural networks

This section will introduce the concept of artificial neural networks (ANNs). ANNs are made up of units called neurons. Figure 3.1 shows the structure of one of these neurons. The process within a neuron involves multiplying each input, x_n, by a weight, w_n, and summing them with a bias term, b. Next, an activation function, $f(\cdot)$, is applied, such as a hyperbolic tangent or sigmoid function. This whole process can be expressed as,

$$y = f\left(\sum_n w_n x_n + b\right) \qquad (3.1)$$

where y is the output of the neuron. The activation function acts as a kind of classifier, distributing the output values according to its shape.

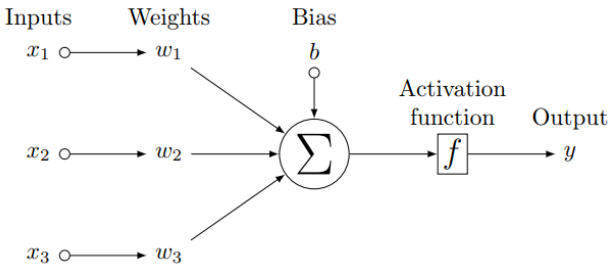


Figure 3.1: Artificial neuron.

Combinations of neurons can be arranged in an interlinked network to be used to model complex patterns and problems. The first layer of neurons is referred to as the input layer and the final layer is the output layer. In the example shown in Figure 3.2, the input layer is connected directly to the output layer.

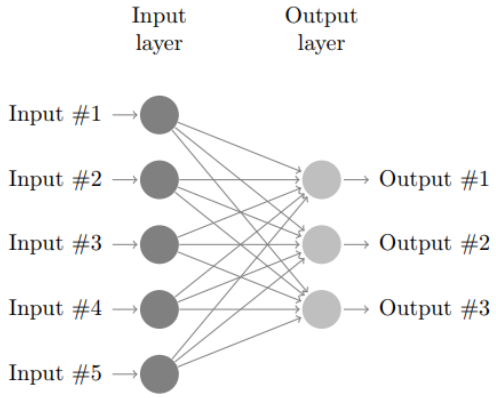


Figure 3.2: Neural network with two layers of neurons.

For more complex tasks additional *hidden* layers of neurons can be added to the network. Figure 3.3 shows an example of an ANN with two hidden layers. The outputs of each layer feed into the next layer in the network until the output layer is reached. ANNs with several hidden layers are referred to as deep neural networks (DNNs).

The usual way DNNs are trained is using an iterative process called back-propagation. This involves fine-tuning the weights of the neural network by considering the error in the output of previous training steps. The first step in the process is to feed inputs forward through the network, obtaining the weighted sum with the applied bias, and then passing these values through the activation function. The error calculated from this output is then fed back through the network to adjust the weights in the network so that it will output a more accurate value with the next training step (epoch). This cycle repeats until either a

pre-set maximum number of epochs is reached, or the error of the output falls below a set threshold (see Goodfellow *et al.* (2016, chap. 6) for a detailed account).

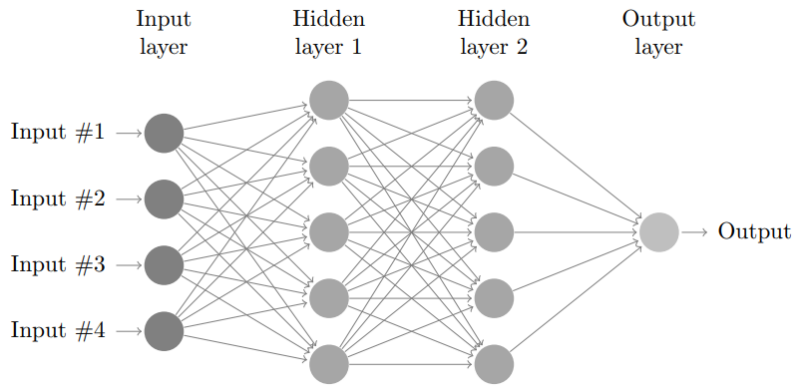


Figure 3.3: Neural network with two hidden layers.

When dealing with temporal sequences, such as speech, it can be useful to include information from previous epochs when calculating the weights of the current state. Recurrent neural networks do just that by taking into account inputs from the past few epochs (Goodfellow *et al.*, 2016, chap. 10). Although this additional temporal feature improves modelling of short time sequences, when it is applied to longer sequences problems can occur. A way to circumvent these problems is to use an improved neural network unit like long short-term memory cells.

3.2.2 Long short-term memory

Long short-term memory (LSTM) cells can model long term dependencies very well which makes them perfect for modelling speech. A full description of LSTM cell structure and their underlying mathematics can be found in Hochreiter and Schmidhuber (1997). In addition to the input and output of a usual neuron, LSTM cells also have three gates: an in-gate, a forget-gate, and an out-gate. The in-gate determines when an input is relevant enough to the problem to be remembered. The forget gate controls how long an input should be remembered or when it should be forgotten. The out gate determines when the cell should output a value.

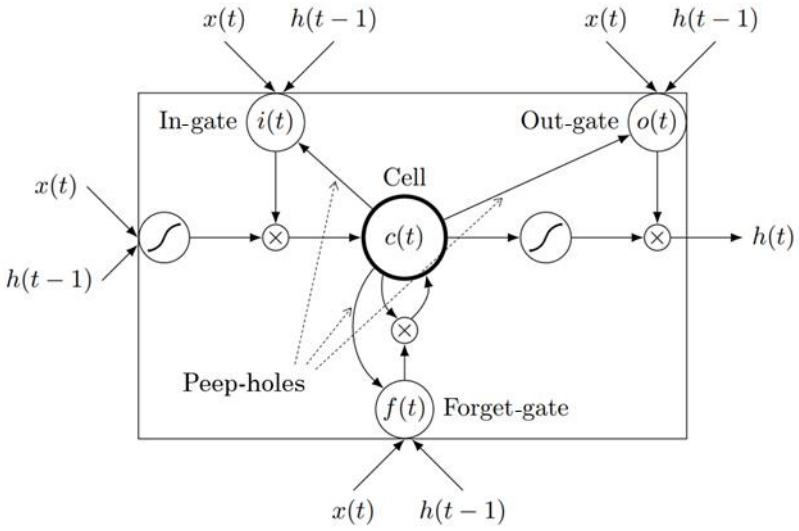


Figure 3.4: Schematic of an LSTM cell

These operations can be formulated as:

$$i(t) = sig\left(W_i x(t) + R_i h(t-1) + p_i \odot c(t-1) + b_i\right) \quad (3.2)$$

$$f(t) = sig\left(W_f x(t) + R_f h(t-1) + p_f \odot c(t-1) + b_f\right) \quad (3.3)$$

$$c(t) = f(t) \odot c(t-1) + i(t) \odot tanh(W_c x(t) + R_c h(t-1) + b_c) \quad (3.4)$$

$$o(t) = sig\left(W_o x(t) + R_o h(t-1) + p_o \odot c(t) + b_o\right) \quad (3.5)$$

$$h(t) = o(t) \odot tanh(c(t)) \quad (3.6)$$

Where $i(\mathrm{t})$,$f(\mathrm{t})$, and $o(\mathrm{t})$ are the in, forget and out gates respectively; $c(\mathrm{t})$ is the memory cell; $h(\mathrm{t})$ is the hidden activation at time t; $x(\mathrm{t})$ is the input; W_* and R_* are the weights applied to inputs and recurrent hidden units respectively; p_* and b_* are the peep-hole connections and biases respectively; $\mathrm{sig}(\cdot)$ and $\tanh(\cdot)$ are the sigmoid and hyperbolic tangent activation functions respectively. The $\odot$ symbol denotes an element-wise product. The DNN architecture of particular interest in this work is that of the bidirectional LSTM RNN (BLSTM-RNN)(Graves and Schmidhuber, 2005). This neural network structure is a combination of a bidirectional recurrent neural network (BRNN)(Schuster and Paliwal, 1997) with LSTM cells.

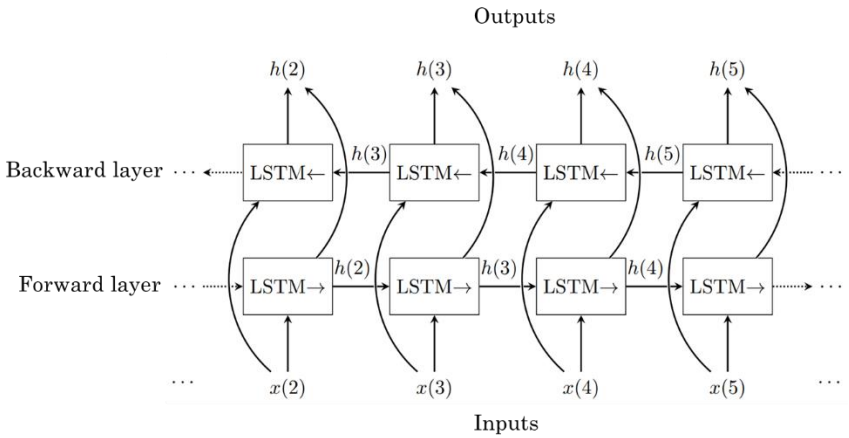


Figure 3.5: Architecture of a bidirectional-LSTM RNN

This type of network can model features over a long time span, is capable of modelling signals with combinations of low and high frequency components, and outperforms both HMM and regular DNN in a TTS system (Fan *et al.*, 2014).

3.2.3 DNNs in SPSS

Zen *et al.* (2013) proposed an SPSS system that used DNNs to model acoustic features. As shown in Figure 3.6, this approach replaces the HMMs by DNNs in the training stage of the system. Although in this case phoneme durations from natural speech are used, the authors state that a separate DNN could be used to model these.

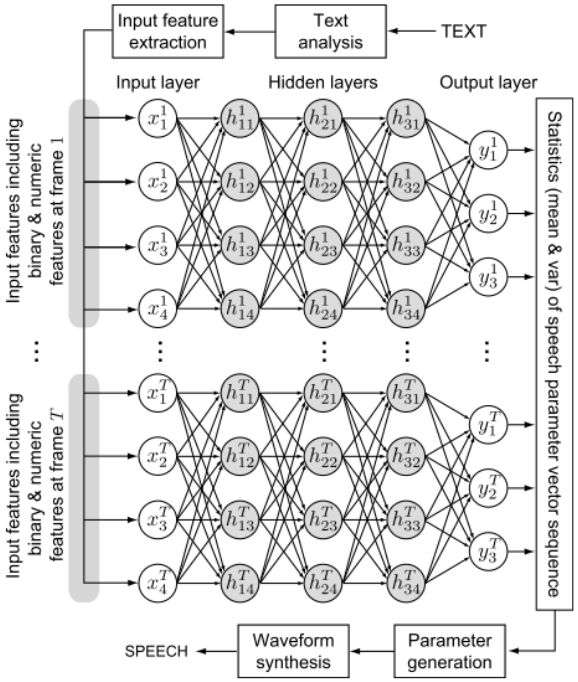


Figure 3.6: Speech synthesis system using a DNN. From Zen et al. (2013)

DNNs are able to model much more complex relationships when compared to HMMs. Other DNN-based speech synthesis systems include those described in Qian *et al.* (2014), Hu *et al.*(2015), Hashimoto *et al.* (2015), and Wu *et al.* (2016). As the benefits of using DNNs have been demonstrated by the systems listed above, they were chosen as the method of statistical modelling for the SPSS framework that the analysis-and-synthesis system developed in this work would be integrated into.

3.3 Speech modelling in speech synthesis

The following section describes several state-of-the-art methods for modelling speech in statistical speech synthesis that have been developed in recent years. The following sections will highlight the advantages and differences between each, as well as some of their limitations.

3.3.1 Pulse-train or noise

The most basic method used for modelling speech in SPSS involves filtering a simple excitation. This excitation consists of an impulse train for voiced segments of speech and white noise for unvoiced segments. A Mel-log spectrum approximation (MLSA) filter is most commonly used to filter this excitation and produce synthetic speech.

The analysis stage of this method involves extracting f_0 and spectral parameters, as well as carrying out voiced/unvoiced detection.

The synthesis stage of the pulse-train or noise vocoder is shown in Figure 3.7. To create the excitation, the f_0 parameter is used to determine the period between generated impulses for voiced speech. The amplitudes of the impulses are scaled so that their overall energy is equal to that of the noise being generated. This excitation is then filtered using the extracted spectral parameters.

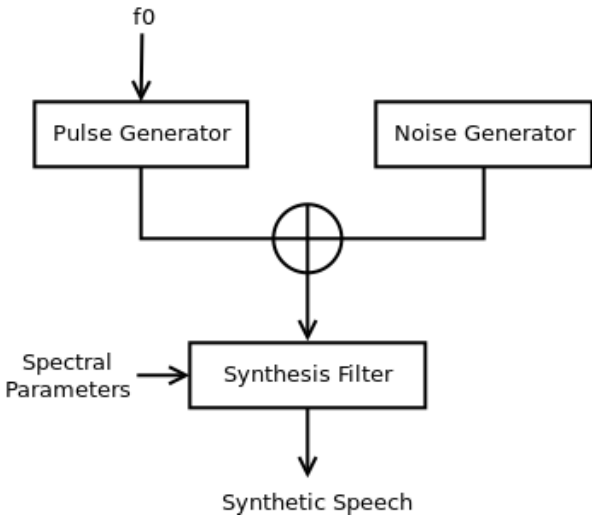


Figure 3.7: Flowchart of the pulse-train or noise vocoder

Although this method is simple, robust and produces intelligible speech, it has a certain 'buzzy' characteristic. This is due to the flatness of the spectrum of the impulse train excitation. Some elements of the higher harmonics of the source signal are still present in the speech output. This may be due to differences in the spectral shape of the original voice source and the pulse train excitation. Another shortcoming of this model is its inability to create an accurate excitation for mixed sounds that contain a periodic element with added noise (i.e. voiced fricatives). It also does not allow for any voice quality transformation.

3.3.2 Multi-band excitation

Multi-band excitation (MBE) models are used in SPSS as a method of reducing the inherent 'buzziness' that comes with using a simple impulse/noise model vocoder. This is achieved by combining spectral regions of a periodic signal with noise. This section will review several relevant MBE methods that are used for SPSS.

3.3.2.1 STRAIGHT

Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum (STRAIGHT), is a set of procedures developed by Kawahara *et al.* (1999) for use in speech modification applications. It is also well suited as an improvement on the basic impulse/noise vocoder for SPSS systems such as HTS. A modification of the original STRAIGHT programme, which we will call STRAIGHT-HTS, uses a mixed excitation model that includes aperiodicity parameters that act to scale the periodic and noise component (Zen *et al.*, 2007).

STRAIGHT-HTS has been used for parameter extraction and resynthesis in several successful HMM-based synthesis systems, for example, NITECH-HTS-2005 (Zen *et al.*, 2007) Nitech-Naist-HTS-2006 (Zen *et al.*, 2008) and HTS-2007 (Yamagishi *et al.*, 2009). The original implementation of STRAIGHT uses fast Fourier transforms (FFT) coefficients to represent spectral and aperiodicity parameters. The high-dimensionality of these coefficients means that they are not suited to the modelling methods used in HMM-based synthesis. The STRAIGHT-HTS vocoder, used in the NITECH-HTS-2005 system, solved this problem by first extracting the FFT coefficients using the standard STRAIGHT method and then converting the spectral values to Mel-cepstral coefficients. The aperiodicity values are grouped and averaged across a set of five frequency sub-bands (0-1, 1-2, 2-4, 4-6, and 6-8 kHz). This band aperiodicity is then calibrated against a table-look-up based on results from known aperiodic signals.

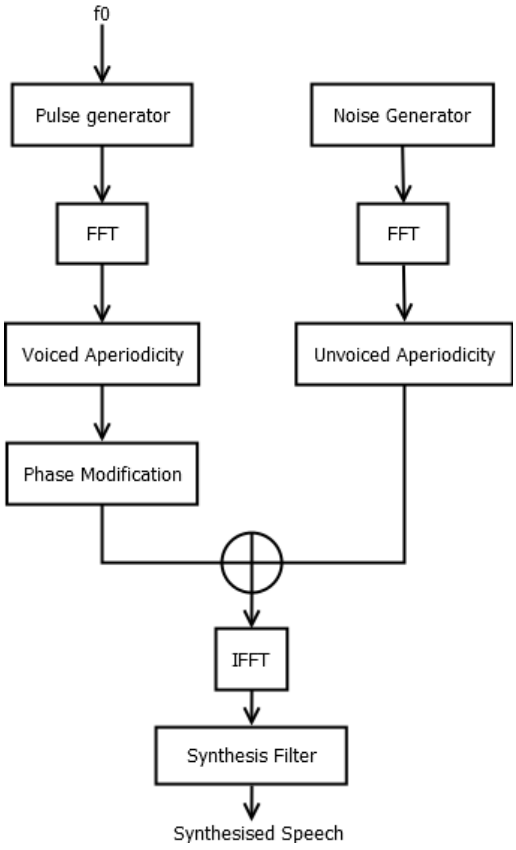


Figure 3.8: Flowchart of the synthesis stage of the NITECH-HTS-2005 system

Figure 3.8 shows a flowchart of the synthesis stage of the NITECH-HTS-2005 that uses the STRAIGHT-HTS vocoding method. For voiced segments, a mixed excitation is generated by combining an impulse with a noise component. Frames are generated pitch-synchronously with a length equal to twice that of the pitch period for both the impulse and

noise components. The amplitude spectra of the separate frames are then weighted by multiplying them by two different frequency dependent stepwise functions. These functions are determined by the corresponding aperiodicity parameters for each of the five sub-bands. The impulse frame is then phase manipulated to compensate for the delay introduced by the system as described in Kawahara *et al.* (2001). The impulse and noise component are then combined and filtered using a Mel-log spectrum approximation (MLSA) synthesis filter, controlled by generated Mel-cepstral coefficients. This use of an MLSA filter and Mel-cepstral coefficients significantly reduces the computational cost compared to the FFT-based processing used in standard STRAIGHT. Finally, the filter frames are concatenated together using pitch synchronous overlap and add (PSOLA) to produce the synthetic speech waveform.

3.3.2.2 WORLD vocoder

WORLD (Morise *et al.*, 2016) is vocoder-based speech synthesis system similar to STRAIGHT (see Section 3.3.2.1). It consists of algorithms for extracting f_0, spectral envelope and aperiodicity values from speech, as well as a resynthesis algorithm. The f_0 extractor, called Harvest (Morise, 2017), works by first band-pass filtering speech with a number of filters, each with a different centre frequency. A fundamental component is obtained from these filters and f_0 candidates are selected based on the distances between zero-crossing points, adjacent waveform maxima and adjacent waveform minima in a three pitch-period frame. Selected candidates are refined by removing rapid changes above a set threshold, removal of very short voiced regions, with interpolation and smoothing as a final step. The spectral envelope estimation, CheapTrick (Morise, 2015), is made up of three steps: f_0-adaptive windowing using a three pitch-period Hanning window, frequency domain smoothing of the power spectrum, and finally, liftering in the quefrency domain to remove frequency fluctuations introduced by quantisation. The band aperiodicity is calculated using the Definitive Decomposition Derived Dirt-Cheap (D4C) estimator (Morise, 2016). D4C estimates the power ratio between the periodic and aperiodic components of speech for different frequency bands using a temporally static parameter calculated on the basis of group delay. The extracted parameters related to the source signal are then used to generate an excitation which is convolved with a minimum phase response derived from the extracted spectral envelope. WORLD was found to perform

better than a STRAIGHT based vocoder in terms of computation time and speech quality (Morise *et al.*, 2016).

3.3.2.3 Harmonic-plus-noise model

The speech synthesiser detailed in Erro *et al.* (2011, 2014) uses a harmonics plus noise model (HNM)-based vocoder to extract three features from analysed speech signals: f_0, maximum voiced frequency (MVF) (Stylianou, 2001; Kim and Hahn, 2007), and spectrum. MVF is the frequency that defines the split between the harmonic and upper noise components of voiced speech segments. The spectrum is represented by cepstral coefficients. Analysis is performed on voiced frames that determine the amplitudes of the harmonics at integer multiples of the detected f_0 value. Unvoiced frames are analysed using fast Fourier transform (FFT). Cepstral coefficients are then extracted and warped to the Mel scale.

For resynthesis, each frame is constructed individually and then concatenated by overlap-and-add. The noise component for both voiced and unvoiced speech frames is generated by sampling the cepstral envelope. For voiced frames, the noise is then high-pass filtered using the MVF as the cut-off frequency. The harmonic components of voiced frames are then obtained using harmonic amplitudes determined by cepstral envelope values at multiples of f_0. A minimum phase approach is used to determine the phase, and phase coherence is maintained across frames by the addition of a linear-in-frequency phase term.

3.3.3 Residual modelling

An improvement to the simple pulse-train excitation is to use the residual signal obtained from an inverse filtered speech signal. This removes the effects of the vocal tract filter from the speech signal. The remaining signal is the residual, and it may contain more source information when compared to the impulse train, depending on the method of inverse filtering carried out. This includes variations in energy, that represent features of the natural source to a better extent, as well as including phase information.

3.3.3.1 Mixed excitation by residual modelling

Maia *et al.*(2007a) describes a speech synthesis system that utilises an excitation model based upon the input of pulse train and white noise into two state dependent filters. This

model follows a similar approach to the Code Excited Linear Prediction (CELP) speech coding algorithms (Guerchi and Mermelstein, 2000), but focuses on the linear prediction residual rather than speech, and takes the error of the system to be the unvoiced component of waveform generation. A flowchart of the excitation model for this system is shown in Figure 3.9. The voiced part of the excitation, *v(n)*, is constructed by filtering a pulse train of varying positions and amplitudes, *t(n)*, by the voiced filter, $H_v(z)$. The output of the voiced filter, *v(n)*, is meant to resemble the residual as closely as possible. The unvoiced part of the excitation, *u(n)*, is constructed by filtering white noise, *w(n)*, by the unvoiced filter, $H_u(z)$.

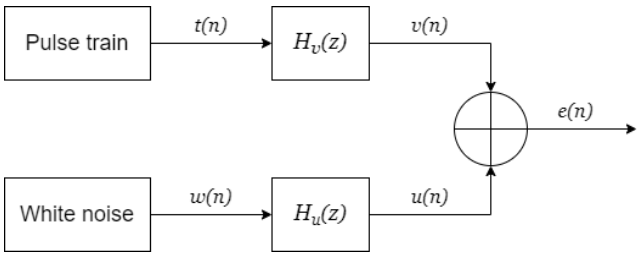


Figure 3.9: Flowchart of the Mixed excitation generation model

Figure 3.10 shows the system used to determine the filter coefficients and optimise the pulse train, *t(n)*. In this case, the inputs of the system are the voiced excitation, *v(n)*, and the residual, *e(n)*, and the error of the system is *w(n)*. Therefore, the error can be minimized through the systematic modification of the filter coefficients and the positions, $\{p_1,\ldots,p_z\}$, and amplitudes, $\{a_1,\ldots,a_z\}$, of pulses within the pulse train, *t(n)*.

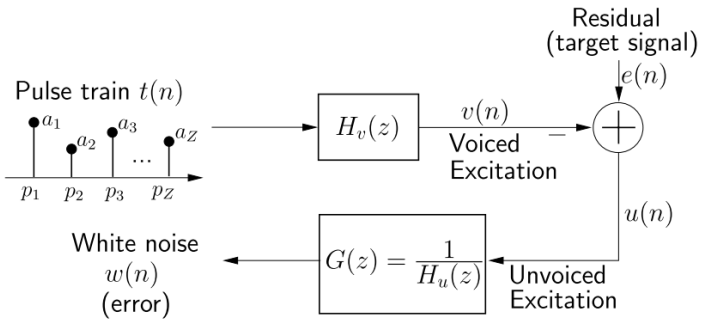


Figure 3.10: System for determining the most likely residual using the given excitation model. From Maia et al.(2007b)

As seen in Figure 3.9, at synthesis, pulses and white noise are filtered by the voiced and unvoiced filters respectively. The white noise component is fist high-pass filtered with a cutoff of 2 kHz to account for limitations in the excitation model and improve the naturalness of the resultant synthetic speech. The voiced and unvoiced excitations are then

added together and filtered by an MLSA synthesis filter. This system was later augmented by including a multipulse model at synthesis time (Maia *et al.*, 2011).

3.3.3.2 Deterministic-plus-stochastic model

The deterministic plus stochastic modelling (DSM) of the residual signal was developed by Drugman *et al.* (2009; 2012). This method was shown to significantly improve the naturalness of statistical parametric speech synthesis, particularly in terms of reducing the 'buzziness'. Since then it has been demonstrated that this can be adapted to provide a natural rendering of creaky voice (Raitio *et al.*, 2013). Although this method does not allow for full parametric control over the glottal source it does offer a successful example of voice characteristic transformation.

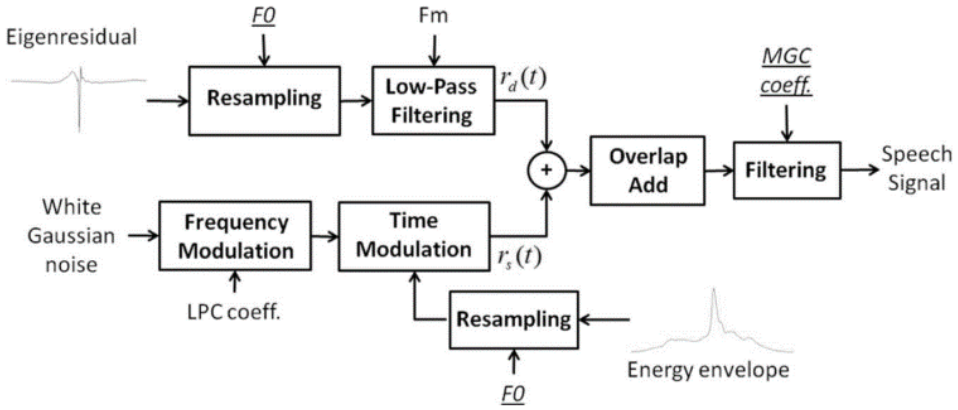


Figure 3.11: Flowchart of the synthesis stage of the DSM vocoder. From Drugman and Dutoit (2012)

A flowchart of the DSM vocoder can be seen in Figure 3.11. This vocoder creates residual pulses to act as the voiced excitation. These pulses contain two elements, a deterministic component made by performing principal component analysis of a set of residual frames, and a stochastic component made by applying an energy envelope and filtering noise. The deterministic component has been coined as the *eigenresidual*. Several of the parameters used in the vocoder are precomputed from a training data set of speech and then used at synthesis time. These are the MVF (represented as F_m), the *eigenresidual*, the estimated pitch of the residual, and the energy envelope and LPC coefficients of the noise component. This means that only two parameter streams are used as external inputs at synthesis time: f_0 values of the source and Mel-generalised cepstral (MGC) coefficients for the synthesis filter. The pre-computed features are used to compose residual frames which

are then joined, using pitch-synchronous overlap-and-add (PSOLA), to produce the voiced excitation. White Gaussian noise is used as the excitation for unvoiced segments.

3.3.4 Formant modelling

The formant modelling approach to statistical speech synthesis integrates elements of the old formant synthesisers with the newer statistical parametric method. Formant values are extracted and directly modelled using HMMs, along with information about the source and several other speech descriptors that were used in the original formant synthesiser implementations.

3.3.4.1 KlaTTStat

The so-called KlaTTStat method (Anumanchipalli *et al.*, 2010) is also close to the research interests presented in this thesis. This method involves parametrising speech (i.e. both source and filter components) using a similar method to that in the KLSYN88 speech synthesiser (Klatt and Klatt, 1990). This parametrisation then replaces the standard MGC coefficients or line spectral frequencies (LSF) modelling in the building of the statistical speech synthesis system. Although this method shows considerable promise, it still produces a distinct *"processed"* quality to the synthesised speech. Furthermore, the glottal source model used does not offer the same level of flexibility as, for instance, the LF-model.

This system first extracts the formant values from windowed speech samples. These values are then used to obtain formant amplitudes from an FFT magnitude spectrum of the same speech sample. The original KLYSYN88 proposed the use of coefficients to describe nasality, aspiration and frication. The KlaTTStat system uses a discriminative approach to classify these features in the data. Gaussian mixture models (GMMs) are used to find the probability of these articulatory phenomena occurring. This classification is performed on the appropriate phonemes for each feature (i.e. nasality, [n, m, ŋ]; frication, [f, h, s, ʃ, θ, v, z, ʒ]; and aspiration, [h]). The output of this detector can then be used to determine whether, and the degree to which one of these features occur. The original synthesiser also included parameters for *gain*, *skew* and *aturb* (amplitude of turbulence). These were set in such a way that synthesised speech sounded as similar as possible to the original speaker.

The system trained voices using the 40 parameters suggested for the KLYSYN88, with statistical modelling being carried out by the CLUSTERGEN synthesiser (Black, 2006).

Once the voice has been trained, a C implementation of the original KLSYN88 synthesiser is used to resynthesise from generated parameters. Although this method produced intelligible speech, the authors state that it retains the robotic sound of the original Klatt-based synthesisers.

3.3.5 Glottal source modelling

Another method to improve the naturalness of synthetic speech is to try and replicate the exact details of the glottal source. This can be achieved by either using estimated samples of the real glottal flow or by closely modelling it. By using estimates of the real glottal flow, a high level of naturalness can be achieved, but this approach lacks the level of parametric flexibility offered by a model. This section describes three methods that utilise both of these approaches.

3.3.5.1 HTS-LF

The HTS-LF synthesis system involves parametric modelling of the glottal source signal, using the LF-model (Cabral, 2010; Cabral *et al.*, 2011, 2014). The effects of the estimated glottal source signal are then removed from the speech signal and the remainder is modelled using the STRAIGHT spectral envelope.

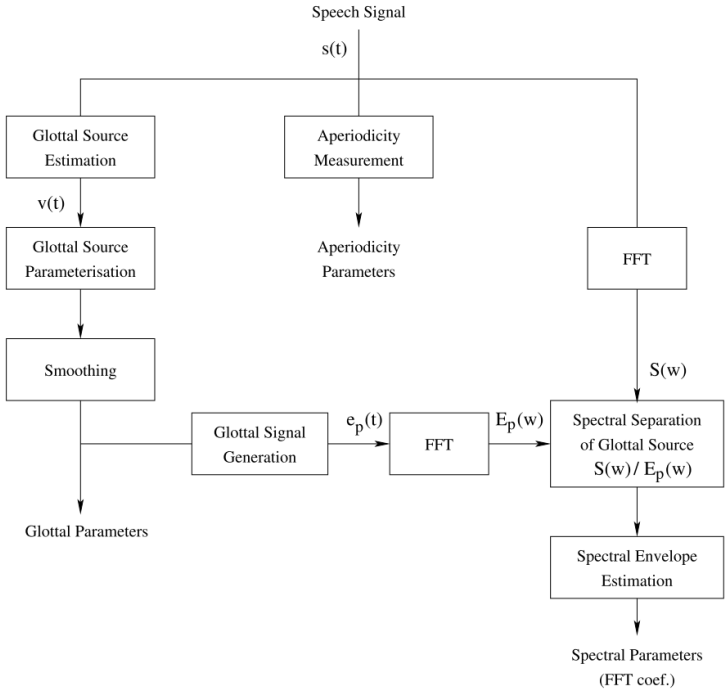


Figure 3.12: Flowchart of the analysis stage of the HTS-LF vocoder. From Cabral (2010)

The basis of the vocoder used in this synthesiser is a process coined glottal spectral separation (GSS). The first step of GSS is to estimate the glottal flow derivative by performing inverse filtering, using the IAIF method, on the speech signal. LF-model pulses are then estimated, using an optimization method, to the glottal cycles of the derivative signal. Next, the spectral parameters are estimated. A cycle of the estimated LF-model is then calculated, and its amplitude spectrum is used to remove the source model effects from the speech spectrum. STRAIGHT is then used to extract the spectral envelope of the signal resulting from the separation of the estimated LF-model spectrum from the speech spectrum.

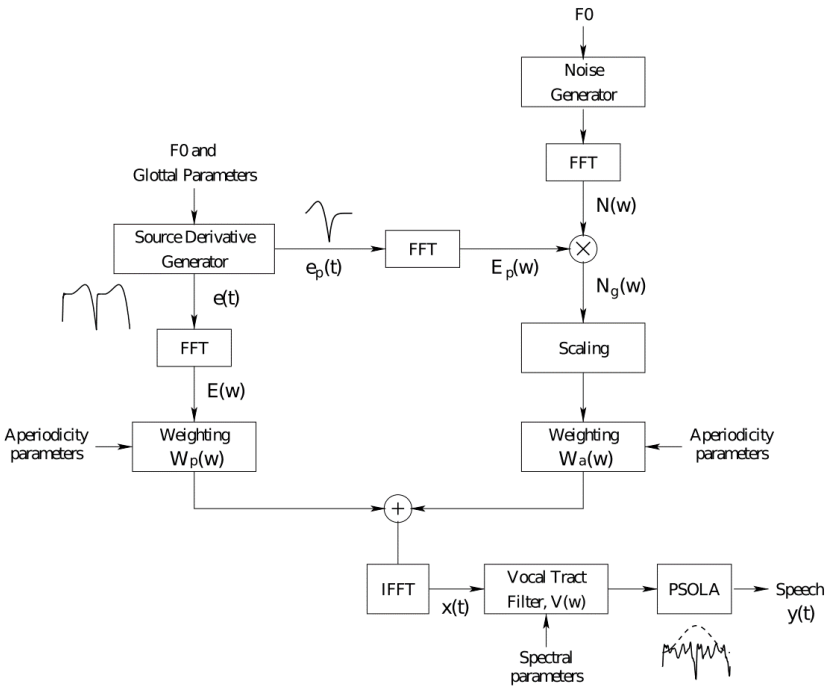


Figure 3.13: Flowchart of synthesis stage of the HTS-LF vocoder. From Cabral (2010)

For the synthesis stage, each frame of voiced speech uses two pitch cycles, starting from the time of the main excitation of the first cycle. This signal is then multiplied by a Hamming window and its spectrum calculated using fast Fourier transform (FFT). This spectrum is then multiplied by the amplitude spectrum obtained from the spectral parameters extracted during the analysis stage. The waveform is then generated by calculating the inverse FFT of the resulting spectrum and phase of the excitation signal, along with removing the DC offset and effects of the Hamming window. The final frames are multiplied by asymmetric windows and concatenated using PSOLA.

3.3.5.2 GlottHMM

The GlottHMM synthesis system (Raitio, 2008; Raitio *et al.*, 2011) uses inverse filtering to obtain an estimate of the glottal source, similar to the HTS-LF method. However, unlike the HTS-LF method, this method involves storing glottal source frames. At synthesis time these frames are resampled to meet the f_0 specifications and combined with a vocal tract filter. This approach has been shown to produce very high-quality synthesised speech, with subjective evaluation showing better levels of naturalness than both standard vocoders and the state-of-the-art STRAIGHT-HTS vocoder. Although glottal pulse frames can be modified at synthesis time, as they are explicitly modelled, comprehensive parametric flexibility is not available.

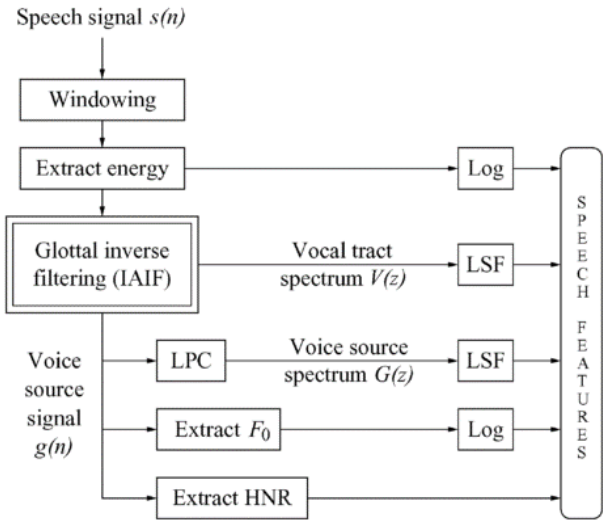


Figure 3.14: Flowchart of the analysis stage of the GlottHMM vocoder. From Raitio et al. (2011)

The analysis stage involves parametrisation of the training data. The glottal source and the vocal tract models are separated using IAIF. The signal is first high-pass filtered with a cut-off frequency of 60Hz to remove any low frequency components that might interfere with the inverse filtering of the glottal source. IAIF is used to remove the spectral effects of the vocal tract and lips from the speech waveform. The output of this filter is the estimated glottal flow signal and a linear predictive coding (LPC) representation of the vocal tract filter.

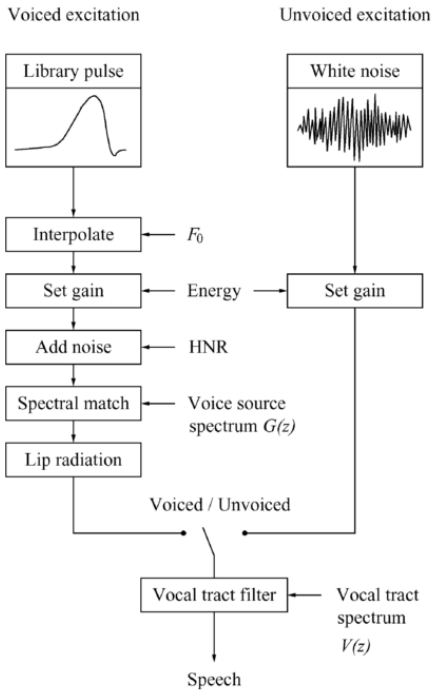


Figure 3.15: Flowchart of the synthesis stage of the GlottHMM vocoder. From Raitio et al. (2011)

During analysis, LPC representations of the glottal flow and unvoiced parts of the training data are also extracted. In order for these LPC models to be more robust to the statistical modelling that takes place within HTS, they are converted to Line Spectral Frequency (LSF) representations (Soong and Juang, 1984), as well as being adapted to the Mel-scale, a frequency scale which more closely represents the perceived pitch scale. Autocorrelation is used to calculate the f_0, and FFT is used to calculate the spectral energy of speech frames to generate the correct unvoiced excitation at synthesis time. This set of extracted parameters is then used in the training stage as described below.

An improved version of this vocoder was introduced in Airaksinen *et al.* (2016) called GlottDNN. As the name suggests, this vocoding method generates a glottal excitation signal using a DNN-based approach, and also uses the QCP method of inverse filtering (see Section 2.3.1.4) in place of IAIF.

3.3.5.3 SVLN

Glottal source modelling is also utilised in the procedure called Separation of the Vocal-tract with the Liljencrants-Fant model plus Noise (SVLN) (Degottex, 2010; Degottex *et al.*, 2012). This procedure involves analysis and parameterisation of the glottal source and estimation of the vocal tract. Resynthesis involves concatenating LF pulses, along with an amplitude modulated noise component, for voiced speech, and windowed frames of noise for unvoiced.

During analysis, this method assumes that the glottal source spectrum can be divided into a deterministic part and a random part. The split of voiced/unvoiced takes place at the MVF, which can be estimated using established methods (Stylianou, 2001; Kim and Hahn, 2007). R_d values for each glottal source pulse are estimated using phase minimisation (Degottex *et al.*, 2011a)(see also Section 2.3.4.2). The noise components of each glottal source pulse are also estimated along. The vocal tract filter is represented in two parts split at the MVF, a True-Envelope (Villavicencio *et al.*, 2006) cepstral envelope for the harmonic component, and a cepstral envelope for the noise component. For unvoiced speech the MVF is reduced to zero, meaning that the vocal tract filter is represented purely by a cepstral envelope.

At synthesis the signal is reconstructed in the spectral domain by mixing filtered and modulated noise components with generated LF-model pulses. Each pulse is then convolved with the estimated vocal tract filter and radiation characteristics before being

transformed to the time domain using envelope inverse fast Fourier transform. The filtered pulses are then concatenated using PSOLA.

This method has been found to successfully transform the pitch and breathiness of synthetic speech (Degottex *et al.*, 2011b). Although it was originally found to be lower quality than STRAIGHT when used for HMM-based speech synthesis (Degottex, 2010), a later adaptation of this method did provide better results (Degottex *et al.*, 2018).

3.3.6 Glottal source modelling in concatenative synthesis

All the speech modelling methods discussed in the previous sections are designed to be used in SPSS systems that generate synthesis parameters from statistical models trained on a speech corpus. The following methods are designed to be used with concatenative systems to allow for the manipulation of the voice source, and in the case of the method described in Section 3.3.6.1, manipulation of filter parameters also. These methods have similarities to the method proposed in this thesis but differ in the fact that their aim is to allow flexible control in concatenative synthesis rather than in SPSS.

3.3.6.1 IBM TTS - Semi Parametric Concatenative TTS

Researchers at IBM have integrated a glottal vocoder into their Concatenative synthesis system that allows for the transformation of speech at synthesis time (Sorin *et al.*, 2017). After a pitch contour has been extracted, analysis is performed pitch synchronously on each voiced frame, while unvoiced frames are ignored. For each voiced frame, the differentiated glottal pulse is estimated using IAIF. This pulse is then parameterized by fitting to it an LF-model pulse that most closely correlates to it. Aspiration noise is calculated by subtracting the LF-model pulse from the differentiated glottal pulse. The vocal tract transfer function is obtained by removing the combined effects of the LF-model pulse and aspiration noise from the original speech and using LPC analysis to model the resulting vocal tract estimation.

3.3.6.2 CerePulse

CerePulse (Buchanan *et al.*, 2018) is another example of a glottal model being used to modify speech for unit selection synthesis. In this case, voice quality modifications are

carried out on suitable regions of voiced speech to increase the expressive range of a speech corpus.

The process uses pitch-synchronous IAIF to obtain the raw differentiated glottal flow estimate. Areas deemed suitable for modification are found by finding the likelihood of them matching LF-model parameter contours.

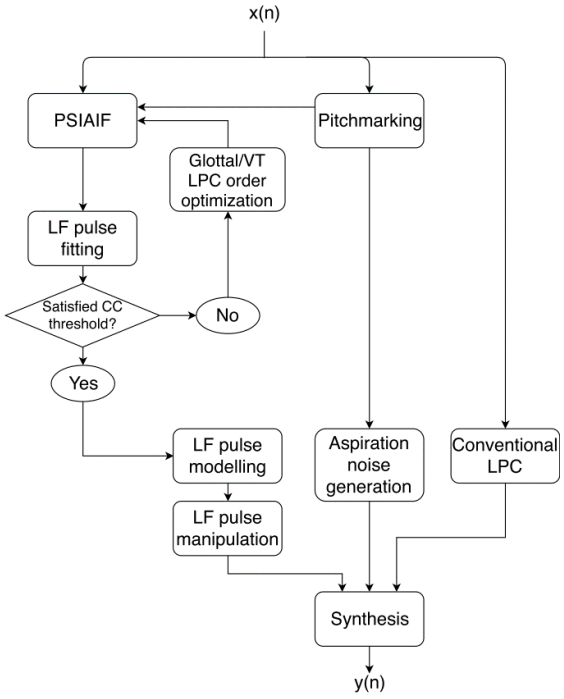


Figure 3.16 Flowchart of the CerePulse speech modification system. From Buchanan et al. (2018)

Once the suitable regions have been found, optimal LF pulses are found for each pitch frame by using a set of LF parameters that maximises the correlation between them. These LF-model parameters can then be modified and used to generate pulses that would result in different voice qualities. This allows for the generation of several different voice quality sub-corpora, in this case for tense and lax voice. Aspiration noise is also added to the pulse frames by high-pass filtering white noise and weighting it according to the specific voice quality.

It was found that although this approach is useful for generating sub-corpora of differing voice qualities where otherwise they would not be available, the lower level of naturalness means that it cannot fully replace the approach of recording sub-corpora for unit-selection voices.

3.4 Chapter conclusions

This chapter briefly described the methods of speech synthesis relevant to the research carried out in this work. Formant and articulatory synthesis are examples of rule-based approaches, while concatenative, SPSS and end-to-end speech synthesis are examples of corpus-based approaches. Each of these approaches attempts to solve the problem of synthesising natural speech in quite a different way. The approach adopted in this work was based on SPSS due to the following points:

- It offers a high level of parametric control.
- Its ability to produce natural sounding synthetic speech without the need for complex previously designed rules.
- It requires a very small amount of memory in comparison to unit-selection systems.
- The ability to incorporate a more sophisticated source model to allow for the transformation of voice quality.

While up until recently HMMs were used as the basis for statistical modelling within SPSS systems, they have now been mostly replaced by DNN-based approaches. This is mainly due to their ability to model more complex relationships than HMMs. This chapter included an introduction into the concepts of artificial neural networks, long short-term memory and systems that utilise DNNs within SPSS.

This chapter also included a review of several of the main methods of speech modelling in speech synthesis. Each of the methods presented approaches the problem of synthesising speech in a different way. Some can create very natural speech, while others offer a high level of parametric flexibility that can be used to transform aspects of the voice. The methods of speech modelling that generate synthetic glottal source signals are the most promising methods when it comes to controlling voice quality and speaker characteristics. GlottHMM/DNN produce voices with a very high level of naturalness but require a library of glottal pulses rather than allowing for full parametric control of the source signal. The HTS-LF and SVLN approaches offer both high levels of naturalness and a degree of parametric flexibility, but they lack a direct interface for straightforward user control of their underlying source parameters. The concatenative speech synthesis methods require large speech corpora and have a much larger memory footprint. The proposed system will include a user interface, allowing for intuitive control of parametric features related to the

voice source and voice quality. It is hoped that by controlling these dimensions the user will also be able to control the prosody of synthetic utterances. The next chapter will go on to discuss the links between voice quality and prosody.

Chapter 4. Voice quality and prosody

One of the main goals of this work is to investigate how a small set of parameters can be used to control linguistic and paralinguistic prosody in speech synthesis. This chapter will introduce the aspects of prosody that are focused on in this thesis, discussing how the voice source contributes to it, and how it is implemented in speech synthesis.

Prosody can be thought of as the music-like modulations of speech. Just as music has dynamic modulations of pitch and timbre, along with temporal/rhythmic properties, spoken language also entails modulations of pitch, voice quality (of which intensity is one component) and timing/rhythmic properties. In speech research, there has been an overwhelming emphasis on the melodic dimension, i.e., the modulation of f_0. Prosody operates at the suprasegmental level and contributes to the information content of an utterance in various ways, some of which are dealt with here.

Broadly speaking, prosody can be seen as having two types of communicative function, linguistic and paralinguistic. The term 'linguistic prosody' includes the melodic variations that are associated with different sentence modes (statements, questions, etc.); the location of stress (which in English differentiates word classes, e.g., REBel *vs.* reBEL, (Gay, 1978)), and in languages with fixed stress, helps the listener identify word boundaries. Focal accentuation on a given syllable in a phrase (WE were away a year ago *vs.* we were AWAY a year ago) (Ladd and Morton, 1997), enables the listener to identify the most important item in the utterance. Prosody also enables the listener to chunk and segment the phrase (Jusczyk *et al.*, 1992).

The term 'paralinguistic prosody' is often used to refer to quite a different signalling function of prosody, which includes information of the speakers' emotional state, his/her attitude and relationship to the interlocutor. Prosody contains vital parallel strands of information about the way in which a speaker wishes their speech to be interpreted, whether it be providing specific information to a listener, and/or carrying an emotional message.

In human interaction, prosody is very important to the meaning the listener takes from the message. In synthetic speech, inadequate prosody can not only have an impact on the

meaning of the utterance, but it can also reduce its intelligibility and naturalness, making it harder to listen to for extended periods of time. For these reasons it important to generate convincing prosodic patterns when synthesising speech.

Given that the vast majority of linguistic research has been directed at understanding f_0 and timing aspects of prosody, research on speech synthesis technology has likewise focused on analysing and modelling f_0 modulations as a means of introducing prosody into synthetic speech. However, research such as that reviewed above, highlights the fundamental importance of all dimensions of the voice source and not only f_0. Relative to the other parameters involved in prosody, voice quality has the least amount of associated research. This is mainly due to the difficulty in analysing the voice source when compared to analysing the parameters associated with the other dimensions of prosody. For the purposes of this work, the term voice prosody will be used to describe aspects of prosody related to the voice source.

One of the goals of this work is to investigate the role of voice quality modulation in prosody and to explore how voice source modulation might be controlled, in order to incorporate it into speech synthesis. This work focuses on the voice quality dimension and sets out to develop an easy to use analysis-and-synthesis system that allows a user to manually control this feature. Most speech synthesis methods focus on extracting information about prosody from text, but it is the goal of this work to build a system that allows for the production of prosodic patterns that are defined by the user through a graphical user interface.

One point to note is that the present work deals with voice quality mostly in the absence of f_0 variation. Clearly both melodic and voice quality dimensions work together, but the focus here is on expanding our understanding of the 'missing dimension' of voice quality. Detailed investigation of how these two dimensions work synergistically can then follow in future work.

4.1 Voice quality in linguistic prosody

The term linguistic prosody is used here to refer to the prosodic features that are described in mainstream linguistic research. These prosodic modulations serve a number of different functions: they help cue grammatical structure; they aid the listener in decoding the stream

of speech into words, phrases, etc., they help cue the listener on the information structure, i.e. which parts of an utterance they need particularly to attend to; they help regulate discourse so that listeners know whether the speaker's turn is coming to an end and whether the floor is being yielded or not (Cutler *et al.*, 1997).

There are many aspects of this vast area that could be examined, but in this work the focus is on one particularly important feature, prominence. Prominence can be defined as a prosodic property where a linguistic element is perceived as standing out from its environment (Terken and Hermes, 2000). Obtaining control of this prosodic dimension is important in the implementation of speech synthesis systems, as it allows for particular, distinct linguistic meanings to be derived from an originally neutral sentence, by the addition of emphasis to a syllable, word or phrase. As with other aspects of prosody, the majority of research into prominence has been focused on the contribution of f_0 (Terken, 1991, 1994; Gussenhoven *et al.*, 1997; Hermes, 2006; Vainio and Järvikivi, 2006; Knight, 2008), and until relatively recently the contribution of voice quality modulation was largely overlooked. The present work looks mainly at the influence of voice quality on prominence, and thus, the following section reviews studies concerned with voice quality and how it is connected with prosodic prominence.

A study by Epstein (2003) found there to be differences in modal voice quality used to distinguish between prominent and non-prominent words. This study also found that speakers used tense voice to signal prominence, and that changes in voice quality tend to occur at phrase boundaries and on accented words. The statistical analysis of voice source parameters, carried out by Iseli *et al.* (2006), found that stressed syllables have less steep spectral slope, indicating relatively greater amplitude of higher harmonics (and suggesting tenser phonation). Pierrehumbert (1989) found that high pitch accents were distinguished by differences in voice source parameters that were not usually associated with high pitch.

A study carried out by Yanushevskaya *et al.* (2010) describes an investigation into the role of source parameters in signalling focal prominence in English. The results indicate that both prominence-lending voice source adjustments (tenser phonation) of the focally accented item and deaccentuation in post-focal material, indicated by changes in several source parameters linked to a laxer voice quality, are used in signalling focal prominence in English. A further analysis of source parameters in utterances with varying focal placement and with falling and rising pitch was carried out in Ní Chasaide *et al.* (2011b).

This study again describes shifts to tenser phonation on the focally accented syllable, while post-focal material underwent a complementary deaccentuation shift towards laxer phonation.

Results from production-based analyses of voice source correlates of accentuation led to the *Voice Prominence Hypothesis* (Ní Chasaide *et al.*, 2013). This hypothesis proposes that different dimensions of the source – f_0 and voice source settings governing phonatory quality – together contribute to our perception of prominence, whether on accented or on focally accented syllables. It further suggests that these dimensions allow a certain degree of freedom in realisations: a greater use of, say, tense voice to signal prominence 'allows' for less use of f_0 cueing: it is proposed that the extent to which melodic or voice quality cues are used may depend on the speaker, or on the location of focal prominence in the utterance, and most likely other factors.

A further production-based study (Ní Chasaide *et al.*, 2015) demonstrated how declination, typically treated as a phenomenon involving f_0, is similarly a matter of complex source changes, involving shifts in several source parameters, indicative of laxer phonation, along with the lowering of f_0. This study also highlights the interaction of prominence and declination. Many of these source parameter adjustments take place along the tense-lax dimension of voice quality.

The study of accentuation in Irish (Ní Chasaide *et al.*, 2013) showed that, while there were consistent voice quality correlates of accentuation, f_0 involvement was found to be a frequent, but not a necessary, associated feature, at least in the case of pre-nuclear position. As in focal accentuation, the prominence level of the accented syllable involves not only the source 'boosting' of the accented syllable but also the relative attenuation of adjacent syllables.

Analytical studies exploring cross-speaker variation in the signalling of focus (Yanushevskaya *et al.*, 2017) and source correlates of focal prominence across different voice qualities (Vainio *et al.*, 2010; Yanushevskaya *et al.*, 2016b) showed that baseline (speaker-specific, habitual) voice quality is likely to have an impact on speaker strategies in signalling focus. Although most recent similar studies appear to point to tense phonation as being associated with the focally accented syllable, a study carried out on Finnish found that prominence was associated rather with a *laxer* voice quality, as indicated by a higher

NAQ value (Vainio *et al.*, 2010). This finding runs contrary to the findings for English and illustrates the fact that the use of a more tense voice quality in signalling prominence cannot be assumed to be a universal tendency. This points to a need for language specific 'rules' when it comes to implementing prosodic variation in speech synthesis.

Given the findings of the studies listed above, it can be seen that voice quality has a very important role to play in the production and perception of prominence. It is important to exploit our knowledge of the voice source to establish the effect of differences in voice quality on signalling prominence in synthetic speech so that it may be utilised in the system developed in this thesis.

As part of this work, experiments were conducted to explore perceptual salience of voice quality modulations. In particular, it was of interest to establish if manipulations of the R_d parameter, without f_0 salience, can be used to generate focal prominence in synthetic speech. Two experiments were conducted involving listening tests with synthetic stimuli. The first experiment used Irish English speech data and the second used Irish speech data. An additional experiment was also carried out to investigate if voice quality manipulations could be used to signal changes in paralinguistic prosody, and the background to this experiment is discussed in Section 4.2.

4.2 Voice quality in paralinguistic prosody

Paralinguistic, or affective, prosody can be thought of as the characteristic of the voice that carries the crucial affective and attitudinal information of a message. When we decode speech, we not only decode the string of words, but also other information about the speaker's emotional state and mood (angry, sad, bored), along with their attitudes and relationship to us (friendly, condescending, sarcastic, polite) (Scherer, 2003). Until recently, this aspect was not included in most linguistic accounts of prosody and was more the domain of psychologists such as Scherer. The research has tended to be largely directed at f_0, intensity and timing, although the contribution of voice quality is widely acknowledged as crucial, little information on this aspect has been available.

Incorporating expressive, or affective prosody into speech synthesis is one of the most difficult problems that researchers encounter when trying to generate natural sounding synthetic speech. There are many reasons for this. One such reason is that speech synthesis

is normally based on a text input (text-to-speech), and while some (but by no means all) of linguistic prosody can be derived on the basis of text, paralinguistic prosody is a feature of spoken language which is not notated in written forms. A second reason is as mentioned, most research on this aspect of prosody was directed at global modifications of f_0, intensity and speech rate – but perception studies by Scherer and colleagues report limited success in the modelling of paralinguistic prosody by incorporating just these modulations.

Other approaches to implementing expressive prosody in speech synthesis involve complex speech parameter manipulations that require expert knowledge (Cahn, 1990; Murray and Arnott, 1993), the use of large emotional speech corpora (Douglas-Cowie *et al.*, 2003; Iida *et al.*, 2003), or utilizing model adaptation techniques and transformation methods (Tsuzuki *et al.*, 2004; Yamagishi *et al.*, 2005). These methods may require extensive resources, such as expressive speech corpora, that may not be available for many languages. This means that an alternative method is needed to achieve expressive speech synthesis.

The need to explicitly incorporate voice quality modulation has been emphasised for many years, e.g., by Banse and Scherer (1996). This point is highlighted by Scherer (1996, p. 1811):

> *…much of speech synthesis is flawed by the lack of appropriate affective variation in prosody and voice quality which seems to be required for both intelligibility and acceptability.*

Studies carried out in the Phonetics and Speech Laboratory in Trinity College Dublin (Gobl *et al.*, 2002; Gobl and Ní Chasaide, 2003; Ryan *et al.*, 2003; Yanushevskaya *et al.*, 2011, 2018) showed that synthetic stimuli of basic voice quality types generated using the KLSYNN88 synthesizer (Klatt and Klatt, 1990) were consistently associated by listeners with particular affective states. Yanushevskaya *et al.* (2018) further points to a language dependency in the mapping of voice quality to affect: results of perception tests on speakers with different language backgrounds showed that languages could differ considerably in how a specific voice quality might be associated with affective states (although there were also many cases of cross-language similarity). Despite a number of studies pointing out the crucial role of voice quality in signalling speaker affect (Gobl and Ní Chasaide, 2003; Scherer, 2003; Patel *et al.*, 2011; Sundberg *et al.*, 2011), the data (from production and perception studies) are still very limited.

Control of the voice source is, therefore, an important aspect of building expressive speech synthesis systems. Being able to readily manipulate voice source parameters in synthetic speech is also important, offering a means to explore how this aspect of prosody works using perception tests. It should be noted that although several of the above studies have found that even though f_0 and voice source parameters tend to covary to an extent, this is not always the case. This suggests that having the ability to control these parameters independently is necessary.

For this reason, the goal of this work was to develop an interface that operates independently of specific affective labelling, that would give the user the freedom to manipulate and control voice source parameters in synthetic speech. Several of the studies mentioned above were based on synthetic stimuli generated using complex parameter manipulations e.g. (Gobl *et al.*, 2002; Gobl and Ní Chasaide, 2003; Ryan *et al.*, 2003; Yanushevskaya *et al.*, 2018). This kind of approach may not be suited to real-time speech technology applications. An optimal solution could be to implement a control system that requires changing a minimal set of voice source parameters that translates to a change in the perceived affective colouring of the synthesized speech, which is one of the main goals of this work. This would have applications in developing realistic personalized and expressive voices, generating characters for educational games, and other software such as screen readers and spoken dialogue systems for low resource languages (Ní Chiaráin and Ní Chasaide, 2016a, 2016b).

4.3 Voice quality in speech synthesis

Although most models of prosody do not account for any voice source parameters other than f_0, it is important to review some of the main methods used and how they have been implemented in TTS in order to understand how voice quality may be included in future models.

The following studies are some examples of applications in which voice quality manipulations have been implemented into speech synthesis with the goal of producing more expressive synthetic voices.

D'Alessandro and Doval (2003) discuss the importance of including voice quality modifications in emotional speech synthesis. They also suggest a way in which speech

units within a concatenative speech synthesiser could be modified by manipulating the magnitude spectrum of the periodic source components, therefore changing the spectral tilt and glottal formant of the source. The aperiodic component may be modified using the method described in Richard and D'Alessandro (1996).

Cabral and Oliveira (2006) describes a method of expressive speech synthesis that uses pitch-synchronous time-scaling to modify the LPC residual of speech in order to transform f_0 and other source parameters related to voice quality. Neutral speech samples were then resynthesised using voice source parameters derived from emotional speech data of several different affective states. This system lacked the flexibility of one that is fully parametric and performs global transformations of voice source parameters rather than allowing for fine-tuning of parameter contours.

Another method, developed by Cabral *et al.* (2011) (see also 3.3.5.1), allows for the control and transformation of the voice source in SPSS by removing the effects of the source from speech, and then replacing it with a synthetic LF-model based source signal. This parameterised signal can be transformed to change the voice quality. This method still requires a level of expert knowledge to carry out any transformations as it does not include a control interface. In addition, the synthetic speech produced by this system contained distortions and was not rated as highly as the authors expected. It was their opinion that these discrepancies were due to errors in the analysis of the source.

Buchanan *et al.* (2018) describes a system that allows for the control of voice quality in concatenative speech synthesis. This system substituted LF-model pulses in place of the voice source and it was found that it could transform the voice quality of synthetic speech along the tense-lax dimension, but the transformations lead to a substantial drop in perceived naturalness.

This thesis sets out to develop a system that allows the user to manually manipulate the voice quality dimension of prosody of a synthetic utterance using a graphical interface. There are two main reasons for doing this. Firstly, by including a control interface, users will have the ability to easily alter synthetic utterances for their desired application. Secondly, the interface is intended as a research tool that allows perceptual testing of the role of the voice source in prosody, and the optimal settings that might be required to generate specific prosodic targets.

To be useful to a wide range of users, it was felt that the interface should be as simple as possible. The production and perception studies that led to this work (Gobl *et al.*, 2002; Gobl and Ní Chasaide, 2003; Ryan *et al.*, 2003; Yanushevskaya *et al.*, 2011, 2018) involved the analysis and synthesis of many complex voice source parameters, many of which covary in natural speech production. Controlling such an array of complex parameters would make control in synthesis very difficult, and would render the system unusable by all but experts in the field. The approach taken in this work is to explore the use of the R_d parameter to control voice quality modulations in synthetic speech.

The following chapter (Chapter 5) explores the use of the R_d parameter to control for the tense-lax dimension of voice quality in prosodic variation through perception experiments. These experiments also serve to perceptually test the production findings, reviewed in the above sections, concerning the tense-lax source modulations associated with accentuation/focalisation, and with certain affective states. It should be noted that voice quality alone is manipulated in these experiments. This is in order to demonstrate the importance of this dimension of prosody – even in the absence of the f_0 cues with which voice quality is modulated in real speech. Clearly in real speech production, both f_0 and voice quality dimensions work together (Ní Chasaide *et al.*, 2013), but a deeper investigation of the complementary use of both dimensions will need to form part of future explorations.

The findings of the experiments in Chapter 5 guide the building of the analysis-synthesis GUI, which is then outlined in the following Chapter 6. The effectiveness of this system is then illustrated through further experiments in which it is deployed, and these are described in Chapter 7.

4.4 Chapter conclusions

This chapter discussed the challenges of modelling linguistic and paralinguistic prosody for incorporation into text-to-speech synthesis. The vast majority of research on prosody has been limited to linguistic prosody and to the modulation of the f_0 contour. It is therefore not surprising that most models of prosody for synthesis have been directed at generating the f_0 contour. Nonetheless there is growing recognition of the need to incorporate the voice quality dimension for the generation of more adequate linguistic and paralinguistic prosody.

Having the ability to control the voice source will be a key factor in providing speech synthesis that meets the needs of real-world applications, such as, language learning games or communication systems for people with disordered speech. By including voice quality in the description and implementation of prosodic models the goal is that another degree of detail and control is made available.

The topics covered within this chapter act as a background for the experiments described in Chapter 5, where perception experiments explore how effectively complex source parameters can be simply controlled by using the R_d parameter, while demonstrating the perceptual relevance of the tense-lax dimension of the voice in signalling aspects of linguistic and paralinguistic prosody in synthetic speech.

Chapter 5. Exploring R_d as a control parameter for voice prosody

This chapter describes three experiments that were carried out to examine how the R_d parameter may be used for controlling aspects of linguistic and paralinguistic voice prosody in speech synthesis. The aim was to establish whether the tense-lax modulation described in earlier production studies could be approximated by using the R_d parameter, and to explore its effectiveness in cueing the prosodic effects described.

The first experiment explores the perceptual salience of R_d modulation to cue focal prominence in English. Experiment 2 examines the correlation between the tense-lax continuum (controlled by R_d) and affect. Experiment 3 builds upon the findings of Experiment 1, to further explore how and where focal prominence may be achieved through R_d manipulation, and optimising the way in which R_d is recalculated after the manipulations have been carried out. Experiment 3 used Irish speech data.

5.1 Experiment 1: R_d in the signalling of focal prominence

One of the goals of this work was to investigate how an acoustic glottal model could be used to manipulate linguistic prosody of synthetic speech using a minimal set of control parameters. This experiment focused on the feature of focal prominence, where earlier production studies highlighted important voice source modulations related to voice quality in the tense-lax domain.

This experiment examined whether the manipulation of R_d can achieve the source variation capable of signalling focal prominence[2]. Earlier analytical studies of production data (see Section 4.1 for a description of relevant studies) found that voice quality plays an important role in the perceived prominence of focally accented syllables in combination with f_0 salience. In this experiment the perceptual importance of voice source adjustments

[2] Presented in Yanushevskaya *et al.* (2016)

made using a minimal set of control parameters is explored. These adjustments were made based on observations from previous studies (Yanushevskaya *et al.*, 2010, 2016b), in sentences with variable locations of focal accent. Further exploration was carried out as to whether such voice source adjustments on their own might be capable of shifting the perception of the location of focal accent within the sentence without f_0 salience.

In this experiment, a recording of the sentence 'We were away a year ago', produced with broad focus, was analysed and subsequently manipulated so that the two accentable syllables WAY and YEAR were (subjectively) deemed to have the same degree of prominence. Broad focus can be seen as when an utterance is produced where no particular part of the utterance stands out when compared to the other parts (Ladd, 1980). This served as the baseline stimulus. Voice source characteristics were then further manipulated in ways that should enhance the prominence of one or other of these syllables. Stimuli were constructed in which the voice source was manipulated (i) in the potentially accentable syllables WAY and YEAR as well as (ii) in the portion of the utterance following the manipulated syllable. The manipulations of (i) and (ii) were carried out individually and in combination. The questions that were set out to be answered were:

- Can such source manipulations induce the perception of focal accent on one or other syllable? (This essentially covers the question as to whether voice-quality changes alone might suffice to cue focal prominence, and the question of whether R_d is effective in controlling the tense-lax dimension of voice quality.)
- Which of the source manipulations (or which combinations of source manipulations) were most effective in cueing focal accentuation?

In these tests, f_0 did not vary across the stimulus set. This is not to suggest that f_0 does not play a major role in cueing focus, but rather represents an attempt to explore how voice source features other than f_0 might be contributing towards perceived prominence, and to see whether source variations alone (without f_0 variation) can alter the perception of where the focal accent lies in a phrase.

5.1.1 Materials

The stimuli for this experiment were based on an all-voiced utterance 'We were away a year ago' produced by a male speaker of Irish English. The utterance was elicited with broad focus, and was recorded as part of an earlier study (Gobl *et al.*, 2015). The utterance

was manually inverse filtered using an interactive inverse filtering system (see Section 2.3.4 and Ní Chasaide *et al.* (1992); Gobl and Ní Chasaide (1999a)). The system first carried out automatic closed-phase inverse filtering to obtain estimates of formant frequencies and bandwidths. These estimates were then manually optimised in order to fully cancel the effects of the formants from the speech spectrum, resulting in an estimate of the voice source signal. The system was then used to carry out voice source parameterisation by first automatically approximating and then manually fitting the LF-model (Fant *et al.*, 1985) to each differentiated glottal pulse. In the construction of the synthesised stimuli, only the R_d parameter was varied.

As explained in Section 2.3.3.2, variation in R_d tends to reflect voice source variation along the tense-lax continuum; the values typically range between 0.5 (tense voice) to 2.5 (breathy voice). As mentioned, earlier analyses of speech data of the speaker used[3] suggested that the speaker shifted towards tenser phonation in focally accented syllables (Yanushevskaya *et al.*, 2010, 2016b) and towards laxer phonation in the post-focal material. These effects were mimicked in the stimuli by lowering the values of R_d in the potentially accented syllables and raising them in the post-focal part of the utterance respectively. Increased phonatory tension tends to correspond to a drop in R_d, but that for the purpose of the illustration in Figure 5.1, R_d drops are referred to as peaks, as it seems intuitively easier for a reader to associate increased tenseness of the voice with positive values.

In the synthesis manipulations, f_0 and U_p were kept constant, which means that changes in R_d were reflected by changes in E_e. By changing R_d, other parameters of the glottal source such as R_a and R_k also vary, and these changes can be predicted from R_d. To synthesize the LF-model waveform, data for the full set of parameters are required. In this case, the transformed LF-model parameters were obtained from R_d using the parameter correlations described in Section 2.3.3.2. Once the source signal had been generated it was filtered by a set of formant filters, arranged in cascade, with formant and bandwidth values obtained in the inverse filtering step.

For the 'baseline' stimulus, the values of f_0, R_d and E_e were set to their average values across the utterance (f_0 = 120 Hz, R_d = 0.86, E_e = 69.8 dB). As this produced a stimulus

[3] Renditions of the utterance 'We were away a year ago' in which the focus placement was varied.

that sounded rather tense, f_0 and R_d were adjusted to make it sound laxer and improve the naturalness: f_0 was increased by 5 percent to 127 Hz and R_d was increased by 50 percent to 1.3. These changes resulted in an E_e value of 67.2 dB. f_0 was flattened in order to reduce, or remove, its contributions to the perceived prominence of syllables so that the effects of R_d manipulations could be examined in isolation. This stimulus served as the baseline for further manipulations of the magnitude and timing of peaks (located relative to the midpoint of the vowels in the syllables WAY and YEAR) as well as deaccentuation in the post-focal material. These manipulations are described in the following paragraphs (see also Figure 5.1 and Table 5.1). The ranges of values used in the manipulations were based on analysis of the speaker in earlier studies (Yanushevskaya *et al.*, 2010, 2016b; Ní Chasaide *et al.*, 2011b). Note that f_0 values were not manipulated and were kept constant in all the syllables of the stimuli.

Table 5.1: Manipulation combinations used in the construction of stimuli.

	N	Peak magn.	Peak timing	Deaccent.
Baseline	0	0	0	0
Peak	1	Low (LP)	0	0
	2	High (HP)	0	0
Peak + timing	3	Low	Early	0
	4	Low	Late	0
	5	High	Early	0
	6	High	Late	
Deaccent.	7	0	0	Shallow
	8	0	0	Steep
Peak + deaccent.	9	Low	0	Shallow
	10	Low	0	Steep
	11	High	0	Shallow
	12	High	0	Steep
Peak + timing + deaccent.	13	Low	Early	Shallow
	14	Low	Late	Shallow
	15	High	Early	Shallow
	16	High	Late	Shallow
	17	Low	Early	Steep
	18	Low	Late	Steep
	19	High	Early	Steep
	20	High	Late	Steep

Peak height (magnitude) in focal syllables, Figure 5.1 (a)

Three levels of peak magnitude were used: no peak (indicated as 0 in Table 5.1), low peak and high peak. The R_d values were set as follows: no peak, R_d = 1.3; low peak, R_d = 1.1; high peak, R_d = 0.9. These changes in R_d resulted in the following E_e values: no peak, E_e = 67.2 dB; low peak, E_e = 68.6 dB; high peak, E_e = 70.3 dB.

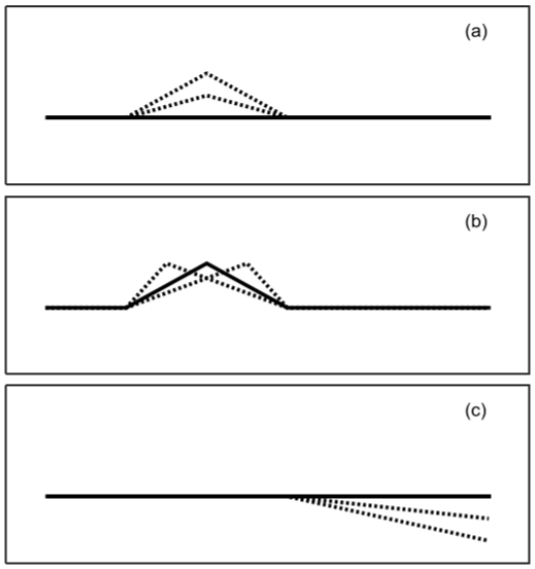


Figure 5.1: Schematic of parameter manipulation in the synthesized stimuli: (a) peak height; (b) peak timing (rate of change); (c) post-focal deaccentuation.

Peak timing, Figure 5.1 (b)

Stimuli were also generated where peak timing was changed relative to the vowel midpoints in the syllables WAY and YEAR. Two peak timing settings were used, early peak and late peak; the values of the peak time instants were negatively and positively shifted by 20% of the duration of the vowel for early and late peak respectively. Early peak corresponds to faster increase to the peak value and slower decrease of parameter values within the syllable; late peak corresponds to a slower rate of change of parameter values to the peak and a faster decrease of the values after the peak. These manipulations were added, as earlier studies of focal accentuation (Gobl, 1988; Yanushevskaya *et al.*, 2010; Ní Chasaide *et al.*, 2011b) suggested that source dynamics are heightened at the edge of the focally accented syllable.

Source deaccentuation in post-focal material, Figure 5.1 (c)

Three levels of deaccentuation in the post-focal material were used: no deaccentuation, shallow deaccentuation and steep deaccentuation. Note that for the WAY-manipulated

sentences, deaccentuation pertains to the entire sequence ‘a year ago’, whereas, for the sentence where YEAR is manipulated, deaccentuation is necessarily limited to the syllables of ‘ago’.

The R_d values were as follows: no deaccentuation, final R_d value = 1.3; shallow deaccentuation, final R_d value = 1.1 (deaccentuation rate 0.6 units/s); steep deaccentuation, final R_d value = 0.9 (rate 1.3 units/s). These changes in R_d resulted in following changes is E_e: no deaccentuation, final E_e value = 67.2 dB; shallow deaccentuation, final E_e value = 65.6 dB (5.1 dB/s); steep, final value = 64.3 dB (9.4 dB/s).

Peak magnitude, peak timing and deaccentuation were manipulated individually and in combinations. The combinations of parameters are shown in Table 5.1. Overall, 20 combinations were synthesized for each of the two syllables WAY and YEAR. The total number of stimuli used in the listening test was 41 (2 syllables x 20 combinations + 1 baseline stimulus). It should be noted that in perceptual terms, differences in peak magnitude and peak timing are registered in terms of the rate of signal change.

5.1.2 Listening tests

Two online listening tests were carried out where the 41 synthesized stimuli were presented to the participants in random order. Participants were instructed to use high quality headphones and to complete the test in a quiet environment. The participants were told that they would hear several utterances in which the syllables WAY or YEAR may or may not be prominent. The participants were asked to listen to each stimulus as many times as they wish and to complete a number of tasks. Two listening tests were carried out to examine how participants responded to two different tasks. Listening test 1 involved simply identifying and rating the level of prominence of a syllable in an utterance. Listening test 2 was a more complex task that required participants to rate the prominence of a syllable relative to adjacent syllables.

In listening test 1 (see Figure 5.2), the participants’ tasks were as follows:

1. Select the prominent syllable: “**Which word is the most important**” (WAY, YEAR, None).
2. For the prominent syllable, indicate the magnitude of prominence: “**Please mark the prominence of this word**” (using a slider on a continuous analogue visual scale).

3. Indicate how confident they were: "**How confident are you in your decision?**" (on a continuous visual analogue scale, 'not at all confident – very confident').
4. Indicate how natural the utterance sounded: "**How natural did the audio sound?**" (on a continuous visual analogue scale, 'not at all natural – very natural').

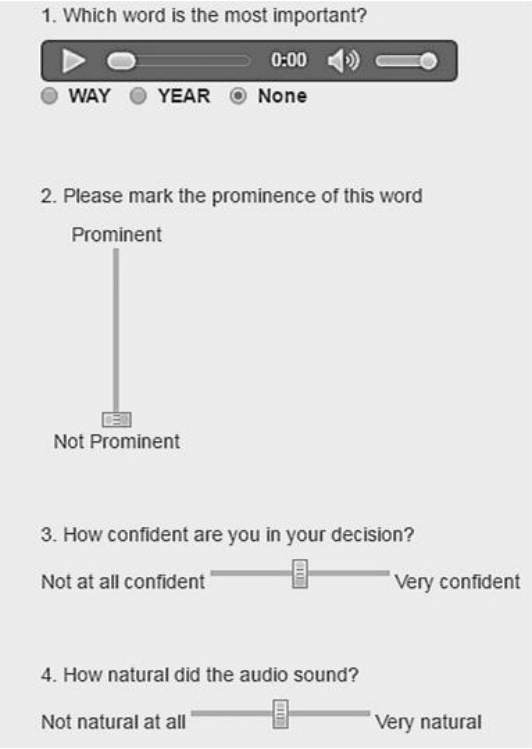


Figure 5.2: Interface of listening test 1.

In listening test 2 (see Figure 5.3), the participants were asked to mark the relative prominence of each syllable in the utterance by adjusting sliders on a continuous analogue visual scale (any value between 0 and 100 could be selected, although participants were not shown numbers). They were also asked to rate the naturalness of the stimuli and to indicate how confident they were in their judgment on a continuous analogue visual scale.

The first experiment was completed by 29 participants; the second experiment was completed by 18 participants.

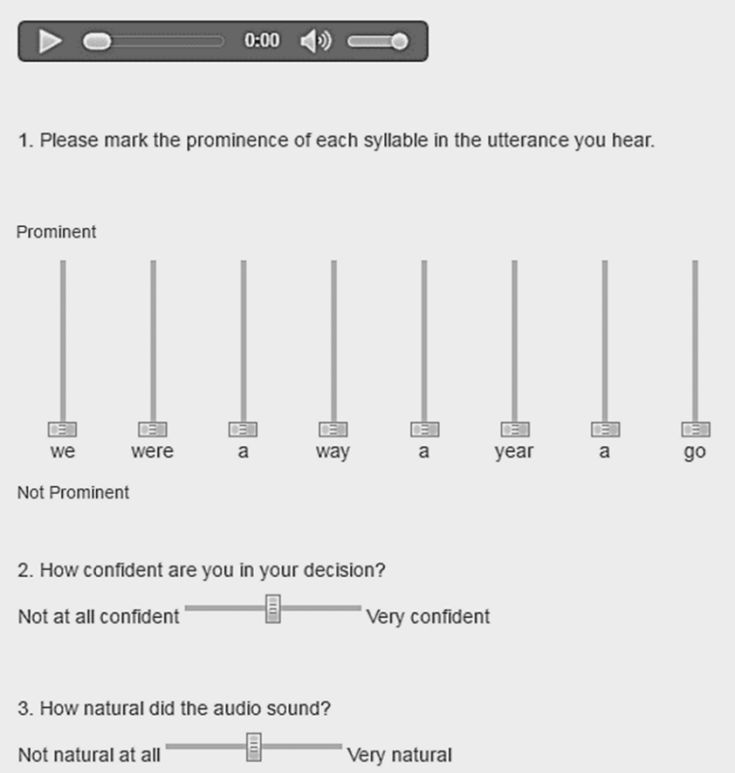


Figure 5.3: Interface of listening test 2.

5.1.3 Results

Listening test 1

It was expected that the syllables WAY and YEAR in the sentences where the voice source for those syllables and for the following material was manipulated would tend to be identified as more prominent. It was also expected that the degree of prominence perceived on the targeted syllable would correlate with the magnitude of the source manipulation carried out. The results show a clear difference in how the two syllables were rated. The overall confusion matrix is given in Table 5.2. In most cases, the WAY-manipulated sentences (those in which the WAY syllable and following material were manipulated) were identified as having prominence on WAY (64%). For the YEAR sentences (those where the YEAR syllable and subsequent material were similarly manipulated), listeners were as likely to hear prominence on WAY as on YEAR – in other words, these sentences were heard to be much the same as the baseline stimulus.

Table 5.2: Overall confusion matrix of perception of the stimuli in the listening test.

		Modified sentences and baseline		
		WAY	**YEAR**	**BASELINE**
Perceived Prominent	**WAY**	64%	37%	38%
	YEAR	19%	39%	38%
	Neither	17%	23%	24%

The full set of results for Listening Test 1 are shown in Figure 5.4. This heatmap illustrates the frequency of which each syllable was marked as prominent for each stimulus (WAY stimuli on the left and YEAR on the right, with the baseline frequencies at the bottom of both). The results are grouped according to the manipulations carried out on the stimuli. There is a clear trend that more stimuli were selected as having a prominent WAY syllable in the WAY sentences than YEAR in the YEAR sentences. For the WAY sentences, the stimuli in which WAY was selected most frequently correspond mainly to stimuli with high peak magnitude and steep post-focal deaccentuation. On the other hand, the stimuli with shallow deaccentuation alone, or manipulations involving a low peak, were identified as prominent by fewer participants. For the stimuli containing manipulations to the YEAR syllable and following material, results were very different. There were only relatively minor shifts from the baseline stimulus results.

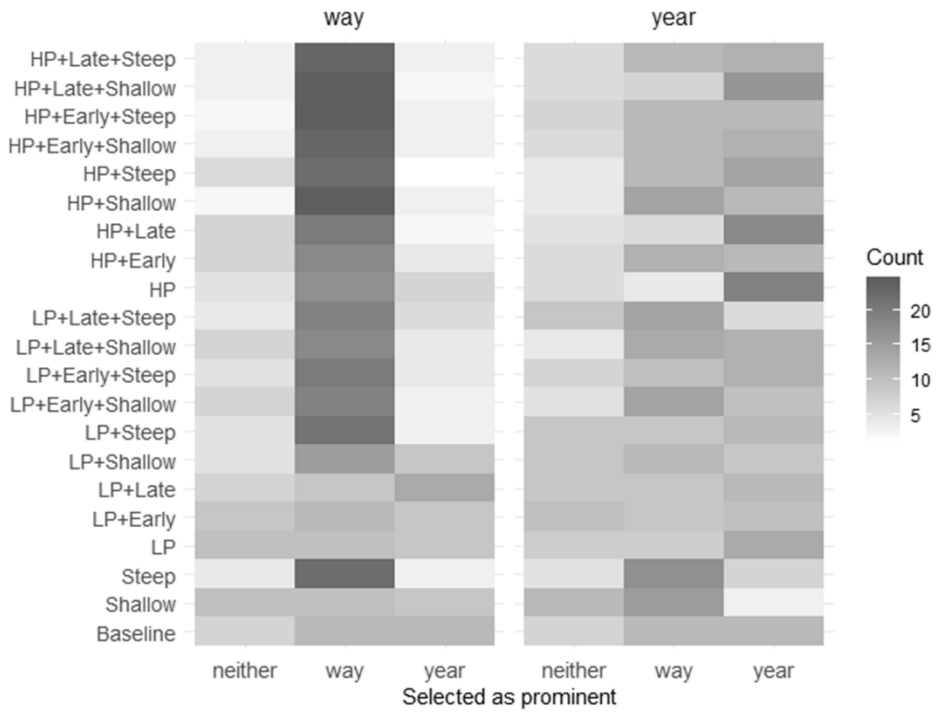


Figure 5.4: Frequencies with which WAY, YEAR and Neither selected as prominent for WAY (left) and YEAR (right) stimuli. HP = high peak and LP = low peak.

Figure 5.4 shows the frequencies with which WAY, YEAR or Neither were selected as being prominent for the WAY and YEAR stimuli. It can clearly be seen that WAY is mostly selected as being prominent for the WAY stimuli, with the only exception being for the stimulus with a low peak and late peak timing. It can also be seen that the high peak stimuli were the most effective at signalling prominence in the WAY syllable, followed by the low peak stimuli with post-focal deaccentuation. YEAR was much less frequently selected as being prominent for the YEAR stimuli, with the exception being three of the high peak stimuli that had slightly higher counts.

Figure 5.5 shows the mean ratings and 95% confidence intervals of prominence magnitude as well as confidence and naturalness ratings for those cases where the syllable WAY was deemed prominent by 70% or more of listeners (grey bars).

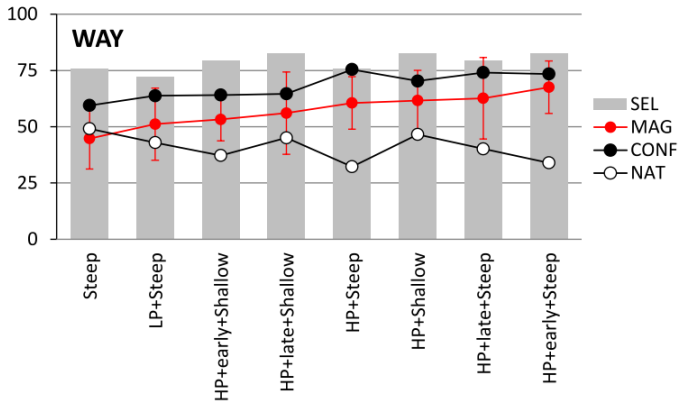


Figure 5.5: Prominence magnitude (red), confidence (black) and naturalness (white) for cases where WAY deemed prominent by 70% or more (grey bars).

The peak height appears to be the main determinant of perceived prominence. There is a significant positive correlation between listeners' judgments on the degree of prominence and their confidence in such judgments (r = 0.91, n = 8, p < 0.05) and a less consistent negative correlation with naturalness (r = -0.59, n = 8, p = 0.06). The naturalness ratings were rather low. This is hardly surprising, considering that the f_0 was constant throughout the phrase, something that does not occur in natural productions. Another factor that could have lead to low naturalness ratings was the loss in signal detail caused by the analysis and synthesis method. Auditory analysis carried out by the author and other expert listeners did not find any other factors that may have affected the naturalness of the stimuli.

A 3 x 3 (peak height, peak timing, deaccentuation slope) factorial analysis was conducted to establish the contribution of the type of parameter manipulation to the prominence magnitude rating. Results indicated a significant main effect of peak $F(2,583) = 15.31$, $p < 0.01$ and deaccentuation $F(2,583) = 10.35$, $p < 0.01$. There was also a weak but significant interaction effect of peak and deaccentuation $F(4,583) = 2.45$, $p = 0.046$. The effect of peak timing was not significant $F(2,583) = 1.46$, $p = 0.23$.

Listening test 2

For this test, participants marked the relative prominence of all the syllables in the utterance. Figure 5.6 illustrates the results, in terms of the difference in the perceived magnitude of the WAY and YEAR syllables within each of the stimulus sentences. (Positive values = WAY perceived as more prominent; negative values = YEAR perceived as more prominent. Blue and red bars indicate sentences with manipulations that should in principle enhance prominence of WAY and YEAR respectively. The cases where the

difference in the magnitude of perceived prominence of WAY and YEAR is significant are shown by asterisks.)

The results, once again, show a clear difference in how prominence is rated in the two cases. In the case of the WAY sentences, most manipulations did enhance the relative prominence of the WAY syllable. The most striking (and statistically significant) effects are found when there is both a high peak on the syllable WAY, alongside deaccentuation of the post-focal material. The steepness of the deaccentuation (shallow or steep) in these cases does not appear to matter. Where the peak on WAY is lower, the effects on magnitude difference are less, and only achieve significance when the low peak combines with deaccentuation, steep or shallow. Manipulation to the height of the WAY peak, on its own, does increase that syllable's relative prominence. This increase only makes it significantly different in prominence from YEAR when the peak magnitude is high, and the timing is early or late. Manipulating the deaccentuation on its own (without adjusting the WAY peak height) is effective only when a steep deaccentuation slope is used.

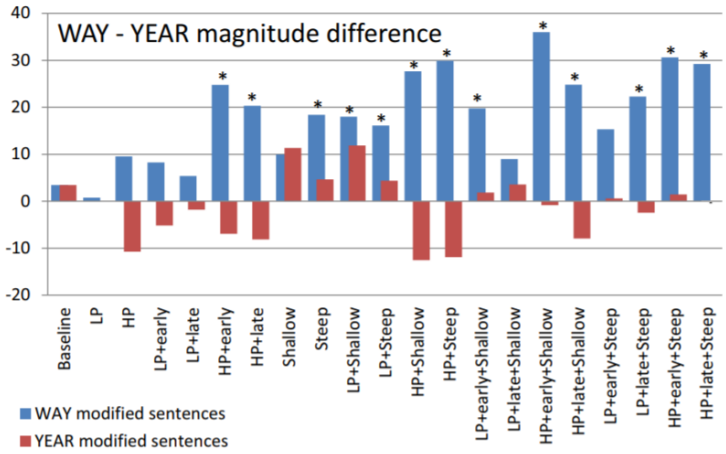


*Figure 5.6: The WAY-YEAR difference in magnitude of perceived prominence. Stimuli where there is a significant difference between WAY and YEAR in the same utterance are marked with *.*

Ratings for the sentences where manipulations should in principle lead to enhancing the prominence of YEAR are very different (red bars). Although a few stimuli shifted the balance somewhat (e.g., some cases where YEAR had a high peak) there was not a single case where such a shift was significant. In most cases, the relative prominence of the two syllables was rather similar to the baseline stimulus.

5.1.4 Discussion

It is clear in these data that the cueing of focal accentuation can vary depending on its location in the utterance. For the non-final syllable, WAY, even relatively small changes in the source parameter values appear to make a difference and can tip the balance in terms of where focal accent is likely to be perceived. It is also clear that there is a synergy between the local prominence on the syllable and deaccentuation in the post-focal material.

In the final accentable syllable (YEAR) the findings were not symmetrical. A low peak has a negligible effect: a high peak can raise the perceived prominence, but the effects are not significant. Furthermore, post-focal source deaccentuation does not appear to play a role. The lack of a deaccentuation effect here may simply reflect the fact that there are only two unstressed syllables for deaccentuation to play out, and that this is insufficient, and not comparable to the case of WAY.

It is likely that the differences observed here between the final (YEAR) and non-final (WAY) syllables have to do with what was not included in these tests, i.e. manipulations to f_0. The f_0 was kept constant in these stimuli as the objective was to ascertain the role of other voice source effects. However, in normal speech production f_0 movement occurs alongside the kinds of source effects implemented here and it is very likely that f_0 movement is far more crucial in final than in non-final syllables. In a production study of focus (Ní Chasaide *et al.*, 2011b) an f_0 fall was found in both WAY and YEAR syllables when focally accented, but the fall was greater and more rapid in YEAR. A further study of source correlates of accentuation (Ní Chasaide *et al.*, 2013) indicated that while accented syllables in non-final position may, but need not, exhibit f_0 movement, a sharp f_0 fall always characterized the final accented syllable. To the extent that this fall is missing in the present stimuli, it is likely to militate strongly against the perception of greater prominence on YEAR, regardless of the other voice source (voice quality) changes that occur.

5.1.5 Conclusions

These tests indicated that voice source modulations of the type observed in earlier production data can cue focal prominence. They also suggest that having a source prominence peak on the focally accented syllable may work together with a degree of source deaccentuation in the post-focal material.

The manipulations that signalled the perception of focal accentuation in the non-final syllable had much less effect on the final syllable, where focal accentuation was not as well cued. The cueing of focal prominence may depend on its location in the utterance, and that in the case of the final accented syllable f_0 movement (which was not included) is a necessary component. Further research is necessary to examine the interplay of source parameters with f_0 in final and non-final syllables, and also the effects of deaccentuation when the post-focal portion of the utterance is longer. Control for vowel quality in focally accented syllables is also required (comparing like vowel with like by using the same vowel in all accentable syllables).

5.2 Experiment 2: R_d as a control parameter for affective prosody

In addition to being able to signal prominence, tenser phonation is also associated with certain affective states, both of positive valence (e.g., elation, joy) and negative valence (e.g., fear, anger)(Scherer, 1986). Similar results are reported in Gobl and Ní Chasaide (2003) and Yanushevskaya *et al.* (2018). The same is true in the opposite direction, towards laxer phonation. These two opposite directions form a tense-lax continuum with a multitude of possible voice quality settings along its length, each with one, or possibly many, associated affective states. These associated affective states are by no means universal across languages. Yanushevskaya *et al.* (2018) demonstrated cases where subjects with different language backgrounds associated the same voice quality with different affects. As one of the goals of this work was to investigate a means of controlling paralinguistic prosody using a minimal set of voice source parameters, control along this continuum is an important element to investigate in terms of controlling the affective colouring of synthetic speech.

This experiment used the R_d parameter to simulate the phonatory tense-lax continuum and to explore its affective correlates in terms of activation and valence[4]. A range of synthetic stimuli were generated varying along the tense-lax continuum using R_d as a control parameter. These stimuli were based on a natural utterance which was inverse filtered and source-parameterised. Two additional stimuli were included, which were versions of the

[4] Presented in Murphy *et al.* (2017)

laxest stimuli with additional creak (lax-creaky voice). These last two stimuli were motivated by the findings of Gobl and Ní Chasaide (2003) and Yanushevskaya *et al.* (2018), where the creaky feature combined with lax voice yielded the most extreme low activation responses. The generated stimuli were presented in a listening test where participants chose an emotion from a set of affective labels and indicated its perceived strength for each stimulus. Participants also indicated the naturalness of each stimulus and their confidence in their judgment.

5.2.1 Materials

Synthetic stimuli

The stimuli were constructed on the basis of an all-voiced utterance 'We were away a year ago' spoken by a male speaker of Irish English in a modal voice. The utterance was produced with narrow focus on the WAY syllable. According to Ladd (1980), narrow focus can be defined as when one section of an utterance is more prominent than the rest of the utterance. The utterance was originally recorded for another study, where other versions of the sentence, with differing focal placement, were also obtained and their source parameters analysed (Gobl *et al.*, 2015). The utterance was inverse filtered using interactive manual inverse filtering software (Gobl and Ní Chasaide, 1999a). Parameterization of the voice source was then carried out using the LF-model (Fant *et al.*, 1985; Fant, 1995), on a pulse-by-pulse basis. Based on the analysed utterance, a modal voice stimulus was first generated. The synthetic stimuli, varying along the tense-lax continuum, as well as lax-creaky stimuli, were then produced using a range of values for the R_d parameter. R_d was manipulated within the range observed for the speaker analysed (0.49 – 2.84). This range was determined by taking the minimum and maximum values from the data presented in the original study (Gobl *et al.*, 2015).

Modal stimulus

A synthetic stimulus for modal voice was generated using parameter values obtained from the prior inverse filtering and source parameterization. The resulting stimulus sounded rather tense; the average parameter values for this stimulus (f_0 = 120 Hz, R_d = 0.8, E_e = 71 dB) also corresponded to a tense phonation type. To make the voice sound less tense and more like modal voice the R_d parameter values were increased by 25% to a mean value of 1.0. These parameters were then used to recalculate E_e, resulting in an average value of 63

dB. U_p was set to its average value and kept constant across the utterance. The contour of the original utterance was stylized by taking values of f_0, E_e and R_d at vowel midpoints in each syllable and interpolating between them. The resulting contour for the R_d parameter is shown in Figure 5.7 (black line). The stimulus was synthesized using cascade formant synthesis.

Lax and tense stimuli

Once the modal stimulus had been created, its parameter values were used as the basis for synthesizing the remaining stimuli. Due to the greater range between the modal and upper R_d limit (1 to 2.84) compared to the range between the modal and lower R_d limit (1 to 0.49), a logarithmic scale was used to establish four steps that would be approximately equidistant in the direction of both tense and lax voice. The steps are shown in Figure 5.7.

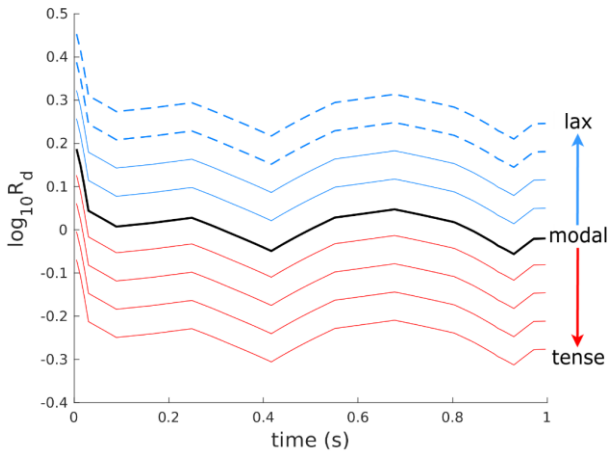


Figure 5.7: $log_{10}R_d$ values of the synthetic stimuli. The lax and tense stimuli are shown in blue and red respectively. The modal stimulus is shown in black. The values used for the two lax-creaky stimuli are indicated by the blue dashed lines.

For all the stimuli, f_0 values were kept the same as those of the modal stimulus. E_e was recalculated according to the new R_d values for each stimulus. The stimuli were labelled as tense1, tense2, tense3, tense4, lax1, lax2, lax3, and lax4, where tense4 and lax4 had the lowest and highest values of R_d respectively. To improve the naturalness of the stimuli aspiration noise was added to the source waveform using the method described in Gobl (2006). This method automatically determines the overall amplitude and modulation of the added aspiration noise based on the shape of LF-model waveform.

Lax-creaky stimuli

Two further stimuli were generated by adding aperiodicity to the source signal of lax3 and lax4. This effectively mimicked a lax-creaky voice quality. This was achieved by using a method similar to the one with which diplophonic double pulsing is produced in the KLSYN88 synthesizer. These lax-creaky stimuli were included here based on the findings of a previous study (Gobl and Ní Chasaide, 2003) where this voice quality was found to yield high ratings for low activation affective states such as bored, intimate, relaxed, and content. The addition of creak to an already lax synthetic stimulus would not require much more in terms of parameter input but could add an extra dimension of affective control.

To generate creak, the source signal was divided into frames so that each frame contained two glottal pulses. The first pulse of each frame was shifted towards the second pulse and its amplitude was attenuated by an equal percentage. This percentage was determined by normalizing and scaling the f_0 contour of the utterance to the range of 0 to 10%. These values were then inverted. This corresponded to a 10% shift and attenuation at the points in the synthetic stimuli with the lowest f_0 values, and no shift or attenuation at the points with the highest f_0. This method, along with a reduction in aspiration noise, produced satisfactory lax-creaky stimuli, labelled as lax3+c and lax4+c.

The total number of stimuli generated was 11 (modal, 4 x tense, 4 x lax, 2 x lax-creaky).

5.2.2 Listening test

The 11 synthetic stimuli, as well as the original natural utterance, were presented to 30 participants (all native speakers of English) in a listening test. The listening test was carried out in a quiet environment using high quality headphones, via an interactive GUI. The participants were presented with 60 synthesized stimuli (5 repetitions of the 12 stimuli) in random order. An additional 5 random stimuli were added at the beginning of the test so that the participants could become accustomed to the process involved. The results of these first 5 stimuli were discarded. The participants were informed that they were going to hear a number of different sound files and that they could listen to each file as many times as they wished in order to answer the following questions for each stimulus:

1. How does the speaker sound? [the subject chose from a selection of radio buttons with affective labels];

2. To what extent? [a continuous analogue visual scale ranging from 'Not at all' to 'A lot'];
3. How confident are you in your judgment? [a continuous visual analogue scale ranging from 'Not at all confident' to 'Very confident'];
4. How natural does the audio sound? [a continuous visual analogue scale ranging from 'Not at all natural' to 'Very natural'].

Participants were given eight affective labels to choose from for question 1. These were: *relaxed, angry, content, upset, happy, sad, excited* and *bored*. These emotional states were chosen to give a balanced set in terms of high and low activation, and positive and negative valence emotional dimensions (see Figure 5.8). Affect labels were chosen, rather than using an activation level scale, in order to simplify the task by using emotional categories that the listeners may be more familiar with. Participants were also given the options of *other* and *no emotion*. The continuous visual analogue scale (Streiner and Norman, 2008) used to measure the magnitude of the affective colouring present as well as the listener's confidence and the naturalness of the stimuli (questions 2-4) was interpreted as ranging from 1 to 100. The confidence score covered questions 1 and 2.

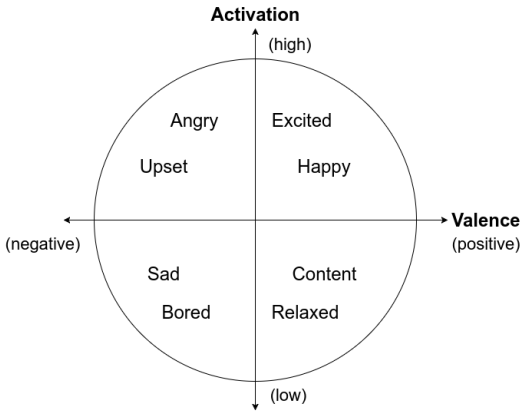


Figure 5.8: Affective labels used in the listening test in the two-dimensional model of affect.

There were no major expectations that a particular stimulus would be associated with a particular affect (prior studies showed that the same voice quality can cue a number of different affects, e.g., (Gobl and Ní Chasaide, 2003; Yanushevskaya *et al.*, 2011, 2018). Clear cut categorical differentiation of the selected affects was not expected either. Support was expected for earlier findings (Ryan *et al.*, 2003) that tenser voice quality would tend to be associated with high activation states and laxer and creakier voice quality with low activation states. Although positive high activation states (happy and excited) were

included to ensure a balanced representation, it was not expected that high percentages of stimulus to affect association would be achieved in these cases (happiness is an affect that has consistently been difficult to model in many, especially in the absence of f_0 modulation (Gobl and Ní Chasaide, 2003; Grichkovtsova *et al.*, 2012)). The type of manipulations used in the construction of the stimuli did not include large dynamic variation of f_0 and did not include any adjustments to the filter settings. These are known to be necessary in generating the impression of excitement and happiness (shorter vocal tract, higher frequency of the second formant for smiling speech)(Tartter, 1980).

5.2.3 Results and discussion

Table 5.3 shows the percentage of cases in which a particular stimulus was associated with each of the options in the listening test.

Table 5.3: Cases (%) in which the stimuli were associated with the affects in the listening test. The most frequent case for each stimulus is underlined and emboldened.

	Angry	**Upset**	**Excited**	**Happy**	**Content**	**Relaxed**	**Bored**	**Sad**	**Other**	**No emotion**
natural	**31**	7	8	12	17	7	1	5	3	9
tense4	**27**	6	21	11	13	5	1	3	3	12
tense3	**26**	8	12	9	16	6	3	1	2	17
tense2	**21**	10	10	10	17	4	5	5	2	16
tense1	15	12	4	6	**21**	9	5	4	2	21
modal	14	10	8	9	**23**	6	5	1	4	19
lax1	2	16	3	3	**21**	7	16	13	1	17
lax2	3	16	1	1	7	15	19	**22**	2	13
lax3	5	21	1	0	5	12	18	**23**	2	13
lax4	3	20	1	1	5	7	19	**34**	1	8
lax3+c	5	15	0	1	3	4	17	**30**	2	22
lax4+c	7	14	1	3	1	8	13	**29**	3	22

Overall, tense voices (except tense1) as well as the natural production were mainly associated with the *angry* affect, lax-creaky and lax voices (except lax1) were associated with the *sad* affect. Modal voice as well as the least extreme cases of lax and tense voices (lax1 and tense1) were associated with *content*. It is interesting to note that the natural recording was selected as *angry* in more cases than any of the tense stimuli. This could be due to some factor present in the original recording that was lost through the analysis and synthesis process.

Looking at the two most frequently selected affects (Figure 5.9), a clear trend can be observed. The degree of tenseness/laxness in the synthetic stimuli is correlated with the frequency with which a particular stimulus is associated with anger/sadness. No one-to-one mapping was found between voice quality and affect. For example, the most frequent response for the lax stimuli was *sad*, but *bored* and *upset* were also often chosen.

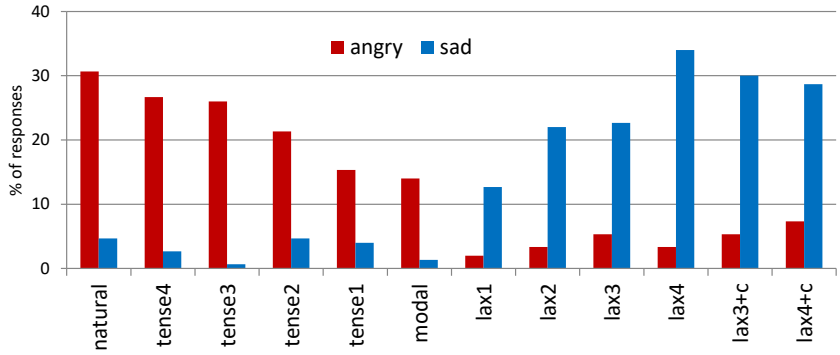


Figure 5.9: Frequency of responses associating the stimuli with the angry and sad affects.

Due to the wide distribution of affect selections per synthetic stimuli, affects with similar valence and activation were grouped (see Figure 5.8). The results are plotted in Figure 5.10.

Tense voice was most commonly associated with high activation affects and lax-creaky and lax voices with low activation affects. Modal voice was almost equally associated with high and low activation affects.

The natural utterance was more commonly associated with high activation states. This is most likely caused by the fact that it has a lower R_d value (between tense1 and tense2) than the modal synthetic stimulus.

In terms of valence, the lax voice stimuli were mainly associated with negative affects, whereas tense, natural and modal stimuli were equally associated with both negative and positive affects.

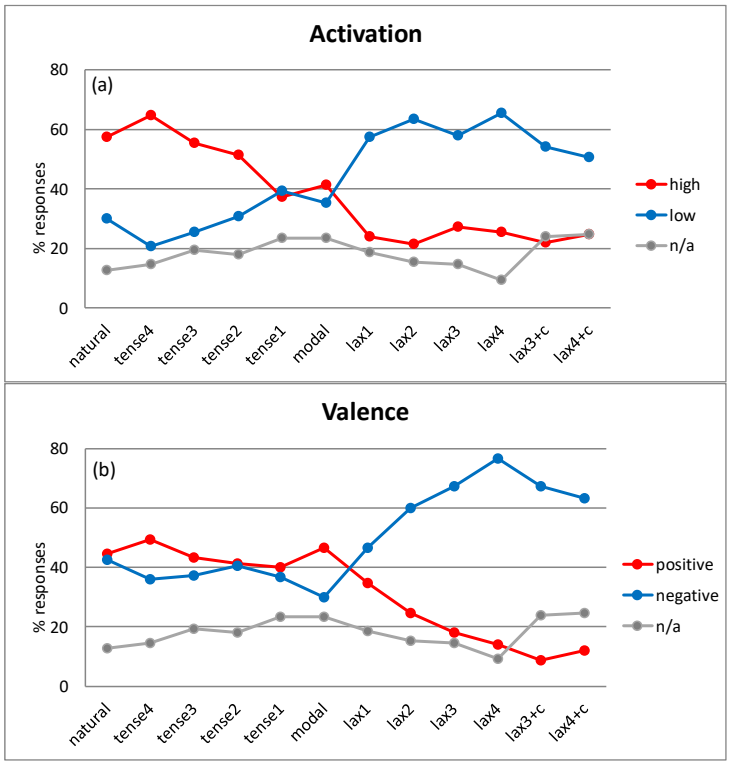


Figure 5.10: Percentage of instances where the synthetic stimuli were associated with affects of particular activation (panel a) or valence (panel b) group.

This even distribution agrees with the initial predictions that additional manipulations of the vocal tract are necessary to differentiate between positive and negative valence states for stimuli representing tenser phonation types.

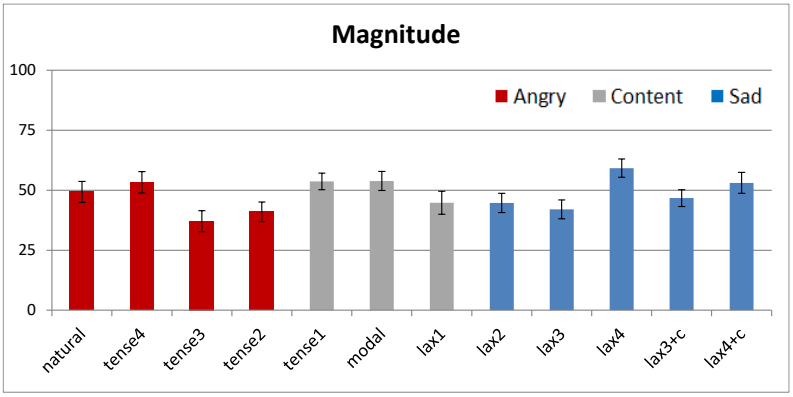


Figure 5.11: Magnitude of most frequently selected affects for each stimulus (mean and standard error).

The magnitude of the manipulation correlated with the perceived strength of affect, particularly with the stimuli at the limits of the tense-lax range. The addition of creak did not yield consistent results: it seemed to increase the magnitude slightly in the case of lax3 but reduced the magnitude in lax4.

The modal stimulus was associated with a different affect than the natural utterance. This is most likely due to the fact that the source parameter values used to generate the modal stimulus were altered to produce a less tense voice quality. The natural stimulus had parameter values between the tense1 and tense2 stimuli. It is not surprising that these stimuli were associated with affects of the same valence/activation, and that the magnitude of the perceived affects was also similar for these stimuli. The small discrepancy between them is perhaps due to artefacts introduced by the analysis and synthesis process (see Section 5.1.1 for a description of this system).

Table 5.4: Mean naturalness and confidence score for each stimulus (scale range 0-100).

	Naturalness	**Confidence**
natural	76	65
tense4	57	56
tense3	57	57
tense2	59	56
tense1	63	57
modal	62	54
lax1	62	55
lax2	61	55
lax3	58	56
lax4	58	61
lax3+c	42	58
lax4+c	43	58

The mean naturalness score for each stimulus can be seen in Table 5.4. The natural stimulus was found to have the highest naturalness at 76. This result would be expected to be higher, but may be due to participants not using the full range of the continuous analogue visual scale. Modal, tense1, lax1, and lax2 follow this with values between 61 and 63. As the stimuli become more lax or tense, there is a slight drop in naturalness, but the values are still well above 50. The two lax-creaky stimuli show the lowest naturalness score. This is likely due to the values used in the generation of these stimuli which may not have been optimal despite the fact that creak was varied dynamically. More research is needed to improve this feature.

All the mean confidence score values were above 54, with the natural stimulus having the highest confidence score of 65.

5.2.4 Conclusions

The study showed that stimuli ranging in the tense-lax continuum generated by manipulating the R_d parameter can evoke an affective response and that the response correlates with the magnitude of the manipulation. Thus, the tenser the voice, the higher the perceived magnitude of, say, anger.

Although manipulations of R_d were effective in evoking an affective response in terms of activation, they were not so successful in distinguishing between valence states. This is most likely due to the lack of vocal tract parameter manipulations. Future work will explore control of these parameters. It will also include improving the quality of the synthetic creak, as the synthetic creaky stimuli were perhaps too regular to be perceived as being natural.

Overall these results show that R_d can be used as a single control parameter to generate variation along the tense-lax continuum of phonation. This has potential uses in the development of expressive speech technology applications.

5.3 Experiment 3: R_d and focal prominence: further exploration

This experiment also examined the perceptual salience of voice source parameters in signalling focal prominence, acting as a follow up to Experiment 1 using Irish speech data[5]. The findings of Experiment 1 (see Section 5.1) led to considerations as to whether phrasal position is important to the realization of prominence. A phrase-final (default nuclear) accent may simply require f_0 salience (e.g., a falling tone) to be prominent, where other source boosting changes alone may suffice in other positions – something suggested by the production data in Ní Chasaide *et al.* (2013). Furthermore, in the above experiment, the short tail following 'year' might have reduced the potential for post-accent attenuation

[5] Presented in Murphy *et al.* (2018)

to have an effect. Other factors in Experiment 1, not related to phrasal position, may also have contributed, such as the difference in the vowel qualities in 'way' and 'year'.

To mitigate some of the above factors, this experiment uses a baseline sentence of Irish with three accented syllables (referred to as P1, P2 and P3), all with the same vowel. The source parameters of the baseline were initially 'flattened', and a slight declination imposed to improve overall naturalness (see details in Section 5.3.1). The manipulations (source boosting of the accented syllable and attenuation in preceding or following material) targeted only P1 and P2, avoiding the final accented syllable. The prediction was that the source manipulations would have roughly similar effects on their perceived prominence relative to each other.

The voice source manipulations for the individual stimuli were implemented using R_d to control the tense-lax dimension of voice variation. As the objective of this experiment was to examine the role of voice source parameters (other than f_0) in prominence perception, the manipulations for the individual stimuli did not include f_0 manipulation.

5.3.1 Materials

Baseline stimulus

The baseline stimulus was generated based on the following declarative sentence of Irish (the accented syllables are shown in bold caps):

*"Bhí **CÁIT** cúpla **LÁ** ar an **TRÁ**laer"*

[vʲi kɑtʲ kʊplˠə lˠɑ ərˠ ənˠ trˠɑlˠərˠ]

"was Kate couple days on the trawler" (word gloss)

"Kate was a couple of days on the trawler" (translation)

The sentence was produced by a male speaker of Irish (Kerry dialect) and was elicited with broad focus. The utterance was analysed, parameterised and synthesised using the same methods as described in Experiment 1 (see Section 5.1.1). Initially, in the baseline stimulus, all source parameters were flattened, i.e. the f_0 and R_d values were set to the averaged values across the original utterance. Following this, to improve naturalness, sentence declination effects were included, entailing a drop of 5% (3.24 Hz/s) in f_0 over the utterance, along with an increasingly laxer phonation (which corresponds here to a rise in

R_d of 20% or 0.114 units/s). An analysis of declination (Ní Chasaide *et al.*, 2015) shows that phonatory quality also alters with declination, and not simply f_0.

R_d is determined by E_e, U_p and f_0 and to effect variation in R_d changes to these parameters are required. Given that the intention was not to vary f_0 (other than for the sentence declination baseline as explained above), to implement the R_d variation in these stimuli, one can vary either E_e or U_p, or a combination of the two. In Experiment 1, only E_e was allowed to vary. For this experiment, two series were prepared. For the E_e stimuli, changes to R_d were implemented by varying E_e, while not modifying U_p. For the U_p stimuli, changes to R_d were implemented by varying U_p without modifying E_e.

Prominence-varying stimuli

Source manipulations for each series of stimuli entailed changes to the source that should boost the salience (through tenser phonation) of the accented syllable (i.e. targeting P1-Cáit or P2-Lá) and/or attenuate, reduce the salience (through laxer phonation) of those portions of the utterance before (Pre) or after (Post) the syllable in question (P1-Cáit or P2-Lá).

The labels used for the individual stimuli and the specific R_d adjustments involved are outlined in Table 5.5 and illustrated schematically in Figure 5.12.

Table 5.5: R_d stimulus manipulations and labels used.

Stimulus label	**R_d adjustments**
Baseline	No salience-lending adjustments
Peak	Tenser phonation (lower R_d) in P1 or P2
Post	Laxer phonation (raised R_d) after the targeted syllable (P1 or P2)
Pre	Laxer phonation (raised R_d) before the targeted syllable (P1 or P2)
Peak+Post	Combinations of above adjustments
Pre+Peak	
Pre+Post	
Pre+Peak+Post	

The R_d parameter was manipulated in three ways, using values prompted by earlier analytic studies (Yanushevskaya *et al.*, 2010, 2017; Ní Chasaide *et al.*, 2011b). For the Peak stimuli, R_d was lowered by 33% in the targeted syllable (whether P1 or P2). In the Peak stimuli targeting P1-Cáit, the R_d value for P1 was 0.8 (compared to the P1 baseline value of 1.14). For the Peak stimuli targeting P2-Lá, the R_d value for P2 was 0.85 (compared to its

baseline value of 1.27). Recall that lower R_d values correspond to tenser phonation. These values reflect the influence of the R_d declination already applied to the baseline stimulus, as explained above.

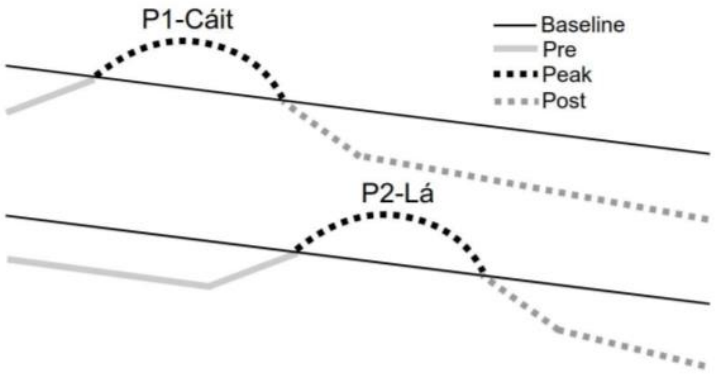


Figure 5.12: Phonatory shifts from the baseline (solid black): Peak (P1 or P2) = tenser phonation (dotted black); Pre (solid grey) or Post (dotted grey) = laxer phonation.

In the Post stimuli, R_d was raised by 30% relative to the baseline value in the syllable immediately following the peak target. This was followed by a further 20% rise over the remainder of the utterance. Pre-target attenuation involved raising R_d in pre-target material by 20% relative to the baseline, with a lowering of R_d across the final pre-target syllable. The combined stimuli simply involved the combination of these adjustments. Discontinuities were avoided in the R_d parameter contours by applying spline interpolation between the parameter transition points.

Two sets were produced, the E_e stimuli and the U_p stimuli, with 13 stimuli in each set. Each set included, in addition to the Baseline stimulus, a range of stimuli that would potentially entail enhancement to prominence on either the P1-Cáit target or the P2-Lá target: Peak, Post, Peak+Post, Pre+Peak, Pre+Post, and Pre+Peak+Post. The Pre stimulus on its own was not included.

5.3.2 Listening test

The listening test was carried out with 29 participants, speakers of Irish, over headphones. In the test, the U_p stimuli were first presented in random order, followed by the E_e stimuli, also in random order. Listeners heard each stimulus four times with a two second gap between each repetition, and there was a 15 second gap before the next stimulus was presented. The participants were asked to:

1. Indicate the perceived prominence of each syllable in the utterance on a continuous visual analogue scale (any value between 0 and 100 could be selected, although participants were not shown numbers).
2. Indicate how natural the audio sounded on a scale from 1 (not natural) to 5 (natural).

Each new stimulus was introduced with a beep followed by the number of the stimulus, as well as a beep 10 seconds before a new stimulus was introduced. Five random stimuli were included at the beginning of the test to allow the participants to get used to the procedure; these responses were not included in the analysis.

5.3.3 Results

The syllable prominence magnitude values obtained for each stimulus were min-max normalized per participant to account for the variation in the use of the scale range. Figure 5.13 shows the difference in perceived prominence of the two accents P1 and P2 (within the same stimulus) for the E_e stimuli (upper panel) and U_p stimuli (lower panel). Blue bars show the P1 minus P2 difference for those manipulations targeting P1-Cáit. Green bars show P2 minus P1 values for manipulations targeting P2-Lá. Thus, blue bars above zero indicate that P1 in the P1-Cáit stimuli is perceived as more prominent than P2; green bars above zero show where P2 (in the stimuli targeting P2-Lá) is deemed the more prominent. Black asterisks indicate significant differences in the magnitude of P1 and P2 within the same stimulus established by a one-way ANOVA.

Figure 5.13 shows the E_e stimuli (upper panel) to be much more effective in achieving prominence than the U_p stimuli (lower panel), and discussion will therefore focus on the former.

P1 is deemed more prominent than P2 in most cases: P2 is judged more prominent than P1 in only two P2-Lá targeted stimuli (E_e: Pre+Peak and Pre+Peak+Post).

It is also clear from responses that in the baseline stimulus, P1 was perceived to be more prominent than P2. Red asterisks show where the magnitude of the difference between P1 and P2 is significantly different from the magnitude of the difference between them in the baseline stimulus. If the perceived differences in prominence are examined relative to the baseline results it can be seen that the R_d manipulations have a greater and more significant effect on P2 perception – even if the baseline bias towards P1 prominence means that P2

still rarely emerges as more prominent than P1. When the baseline difference is considered, the 'enhancement' of P1 brought about by the P1-Cáit manipulations appear less dramatic. Despite this factor, P1 emerges as both intrinsically more prominent and more readily enhanced perceptually.

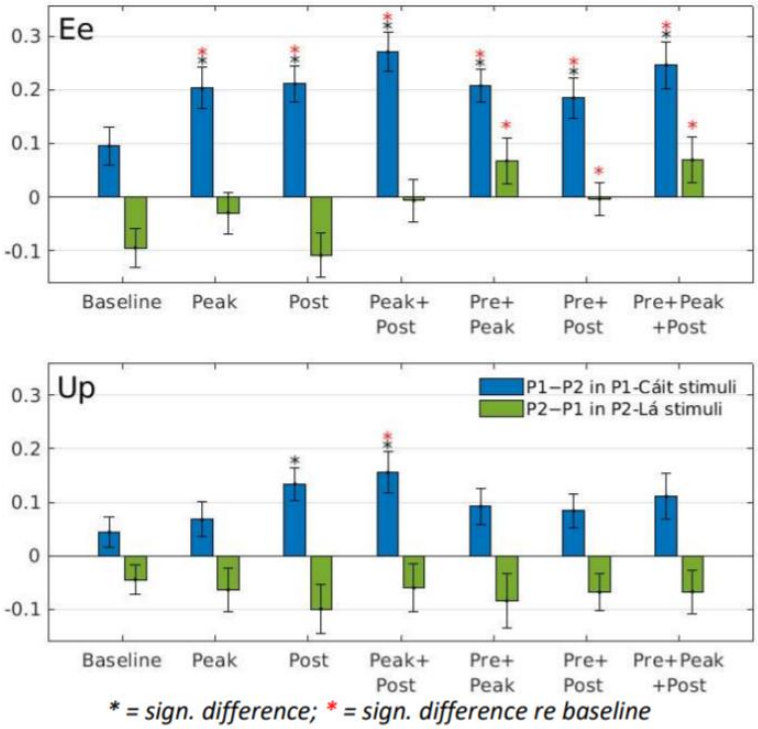


Figure 5.13: P1-P2 differences for P1-Cáit stimuli (blue), and P2-P1 differences for P2-Lá stimuli (green).

Figure 5.14 provides a more detailed view of the effects on the overall P1 P2 P3 contour of the P1-Cáit targeted manipulations (left panels) and the P2-Lá targeted manipulations (right panels).

Panel (a) of Figure 5.14 shows, relative to the baseline (black dashed line), values for those stimuli with peak-boosting, and those with post-attenuation. For P1, peak-boosting has the effect, not only of raising prominence on P1, but also of lowering it on P2. The Peak stimulus yields essentially the same contour as the Post stimulus. This suggests the functional equivalence of these two source adjustments in the utterance, the local and the global. In the case of the P2-Lá targeted stimuli, these two effects do not appear to be equivalent. Peak-boosting does raise the prominence of P2 relative to P1 and P3, but post-attenuation has little effect on the contour (though there is a drop in overall level).

Panel (b) of Figure 5.14 illustrates the additive effect of peak-boosting and post-attenuation. When targeting P1-Cáit, the effects are additive, increasing the perceived prominence of P1 beyond that achieved by either of the individual source adjustments. This is not so for the P2-Lá targeted stimulus: the contour is much like that of the peak-boosted P2, indicating again that post-attenuation is not contributing much.

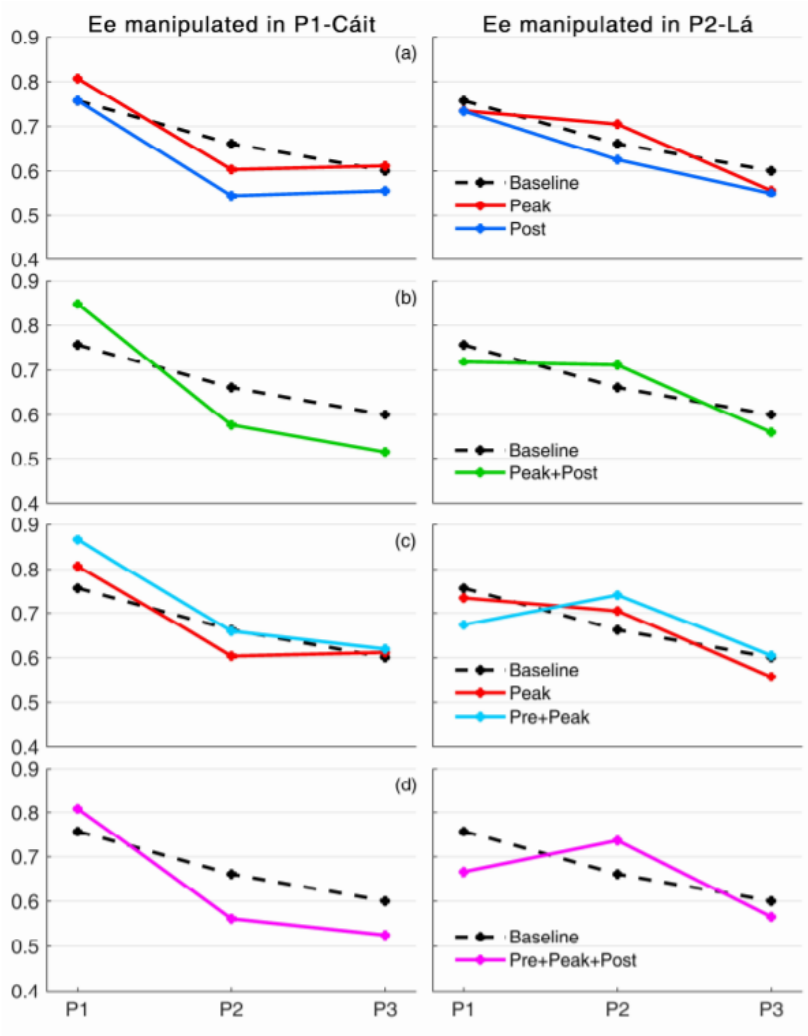


Figure 5.14: P1 P2 P3 prominence for stimuli targeting P1-Cáit (left) and P2-Lá (right).

Panel (c) compares Pre+Peak (pre-attenuation in combination with peak-boosting) and peak-boosting on its own. The effects on perceived prominence magnitude are apparent for both P1 and P2 targets. This was unexpected in the case of P1, given that the pre-attenuation for P1 pertained only to a single short unaccented syllable.

Panel (d) illustrates the combined effect of all three manipulations: Pre+Peak+Post (pre-attenuation, peak-boosting and post-attenuation). While this complex manipulation is effective for both P1 and P2, it looks like the combination of pre- and post-attenuation with

peak-boosting is not necessarily more effective than the simpler combination of Peak+Post (peak-boosting and post-attenuation) for P1, and only marginally more effective than the combination of pre-attenuation and peak-boosting for P2.

The naturalness of the E_e and U_p varying stimuli was not significantly different. The average naturalness of the E_e stimuli was rated 3.2 (range 2.89-3.54), the average naturalness of the U_p stimuli was 3.1 (range 2.78-3.26).

5.3.4 Discussion

Although there is extensive research into the contribution of f_0 to prominence, far less is known about the contribution of other source parameters. Production studies of accentuation and focalization (see Section 4.1), which demonstrated the involvement of the entire source signal in prominence, indicated that f_0 and other source parameters work together. The combined source effects involve local salience boosting to enhance prominence on the accented syllable, as well as more global, phrase-level source attenuations in other areas of the utterance. The results of this experiment illustrate how the phrase level pre- and post-attenuation may be perceptually equivalent to the peak-boosting source adjustments.

Initial expectations that the effects of source prominence manipulation on P1 and P2 would be the same were not borne out, as P1 was deemed to be the more prominent of the two. In all the P1-Cáit targeted stimuli and in all but two of the P2-Lá targeted stimuli, P1 was judged the more prominent. Furthermore, these two possible targets differed in the extent to which they were influenced by phrase-level attenuation. For P2, pre-attenuation emerged as being very effective: P2 was perceived as more prominent than P1 only in the two stimuli where pre-attenuation was present. Surprisingly, post-attenuation appeared to contribute relatively little. In the case of P1 it was very different in that both pre- and post-attenuation were effective in conferring prominence.

P1 is very clear more dominant in the baseline values. Two possible explanations exist. It may simply be that a phrase-initial accent is inherently more prominent, with some sort of perceptual or contextual bias favouring the phrase-initial location. On the other hand, it may be that the declination included in these stimuli introduced a P1 bias, conferring greater prominence in the order of P1>P2>P3. These two explanations could be complementary: an inherent prominence of the phrase-initial accent may be linked to the

declination across the phrase. To date, it was thought that declination in itself may not confer prominence, and that accentuation is judged relative to the declination line, but this assumption may need to be explored further. It may in fact be the case that the declination-prominence 'adds' a natural prominence bias to the initial prenuclear accent of a phrase. It is worth noting, however, that in Experiment 1 (see Section 5.1), there was no declination included in the stimuli, and judgements also indicated a similar disparity in the prominence ratings of consecutive accents.

Of the two R_d implementations used in this experiment, the one involving variation of E_e was considerably more effective in achieving prominence than the implementation involving variation of U_p. When considering the spectral correlates of U_p and E_e variation, this finding was expected. U_p is associated primarily with the level of the first harmonic, and the lower frequency components of the source spectrum, while E_e has a more pervasive influence on overall spectral levels above H1.

5.3.5 Conclusions

This experiment demonstrates how voice source modulations affecting phonatory quality are perceptually important in signalling prominence, operating both at a local level in boosting the prominence of the accented syllable and at a global level in attenuating the prominence of other portions of the utterance.

The results also suggest that the location of a syllable within the phrase may be important, revealing a bias towards greater prominence on the phrase-initial prenuclear accent, when compared to the following one. It is unclear from these data whether this bias results from the effects of voice source declination (described in Ní Chasaide *et al.* (2015)).

The results of this experiment support previous findings that R_d has potential as a useful parameter for controlling linguistic focus in speech synthesis, even in the absence of f_0 manipulation. Two implementations of the R_d variation were used in this study and it was found that the method where E_e was allowed to vary was much more effective at signalling prominence. A fuller elaboration of how to optimally control this parameter may also be the focus of future work beyond this thesis.

5.4 Chapter conclusions

The findings from the experiments discussed in this chapter suggest that R_d is a useful parameter for controlling both linguistic and paralinguistic prosody in synthetic speech. Due to the extensive amount of work required in performing manual inverse filtering and glottal source parameterisation the experiments focused on single speakers and utterances. Although this means that the results cannot be viewed as a complete generalisation, it is thought that they offer a deep insight into voice prosody control. Experiments 1 and 3 found that it was much easier to signal prominence in a syllable towards the beginning of an utterance by manipulating R_d alone, but it becomes more difficult with syllables towards the end of an utterance. This could be due to several factors, including: the lack of f_0 manipulation in the stimuli; the lack of, or lesser amount of, post-focal material in the stimuli; or the perceived prominence in the baseline stimuli overriding any effects of parameter manipulations. Experiment 2 showed that stimuli ranging in the tense-lax continuum generated by manipulating the R_d parameter can evoke an affective response, but that distinguishing between valence states may require additional parameter manipulation.

The findings from these studies acted as the basis for the design and development of an analysis-and-synthesis system described in the next chapter (see Chapter 6). This system has the ability to be integrated into a speech synthesis system, which allows users to manipulate voice source parameters in order to change perceived prosody in synthetic speech. These initial studies used manual analysis techniques and formant synthesis. The system outlined in Chapter 6 performs both analysis and synthesis automatically but incorporates a means to manually adjust voice source parameter contours to alter synthetic speech in a way desired by the user.

The main points that will be taken into account in the design of the system are that:

- Control of the voice source parameters – E_e and R_d in this case – is required across whole utterances to successfully elicit prominence in different locations.
- Control of E_e and R_d is sometimes sufficient to signal prominence, even without f_0 salience in some case, and to alter the perceived affect of an utterance.

- An additional important point that was taken into account in the design of the analysis-and-synthesis system is to allow for immediate feedback to the user, in experiments (see Sections 7.1 and 7.2) where they are asked to modify a speech signal. There is no use developing a system that allows for the modification of speech if the user is required to wait in order to hear the result.

Chapter 6. The GlórCáil analysis-and-synthesis system

One of the main goals of this work was to implement a speech analysis-and-synthesis system that allowed a user to easily control a minimal set of voice source parameters, as well as vocal tract characteristics, in order to transform linguistic and paralinguistic prosody, as well as speaker characteristics. These kinds of transformations would prove very useful in the generation of characters in educational games for low resource languages where large amounts of speech data are unavailable.

This chapter describes the setup and methods involved in the GlórCáil[6] analysis-and-synthesis system. The graphical user interfaces for the analysis and synthesis stages and a specific implementation of the synthesis interface for use in a manipulation task perception test outlined in Chapter 7 are also discussed.

The GlórCáil system is an application developed for the analysis and resynthesis of speech with a particular focus on control of the voice source parameters. It is implemented in the MATLAB environment (The Mathworks Inc., 2018). Vocal tract and voice source parameters are first estimated during the analysis stage. Speech is then resynthesised using an acoustic glottal source model in place of the original glottal source. This allows users to change the perceived voice quality of an utterance by manipulating voice source parameter contours in an interactive GUI. Changes in voice quality can lead to changes in the perceived prominence of words/syllables (see Experiment 1 and 3, Sections 5.1 and 5.3)(Yanushevskaya *et al.*, 2016a; Murphy *et al.*, 2018), and even changes in the perceived affect of an utterance (see Experiment 2, Section 5.2)(Murphy *et al.*, 2017). As mentioned previously, an earlier study carried out a principle component analysis on various voice source parameter and found that R_d was important in describing cross-speaker differences in voice quality (Yanushevskaya *et al.*, 2017). These studies suggest that R_d is an effective voice quality control parameter, and by reducing the dimension of source parameters to a

[6] The name comes from the Irish words for voice (glór [ˈɡ̪ʲ oː rʲ]) and quality (cáilíocht [ˈk ɑː lʲ iː x t̪ʲ]).

minimal set, the process of parameter manipulation is simplified. This is an important factor when considering usability and the integration of the system into an SPSS system.

Another main aim of this work was to integrate the GlórCáil system into a DNN-based SPSS system. This chapter also includes a description of this process, outlining the SPSS system that was used and how it was adapted to accommodate the GlórCáil system.

The findings from the studies detailed in Chapter 5 indicate that components of this system would be useful tools when exploring the correlates of linguistic and paralinguistic prosody. There are also possible applications in the generation of varied prosody including expressive speech for languages or voices where there are limited, or no, expressive speech corpora available. When integrated into an SPSS system, the tool could be used to create utterances with particular prosodic patterns that are missing from the original speech corpus.

6.1 Analysis Stage

A flowchart of the analysis stage is shown in Figure 6.1. The first step of the analysis stage is the polarity check. This is followed by the estimation of f_0, GCI locations and voiced/unvoiced (VUV) regions of a recorded speech signal. This is carried out using the REAPER programme created by Talkin (2015). The next step is to perform inverse filtering on voiced regions to obtain an estimate of the glottal source. This signal is then parameterised by fitting the LF-model to each differentiated glottal flow (DGF) pulse. LPC analysis is performed on unvoiced regions of speech to obtain an estimate of the vocal tract transfer function.

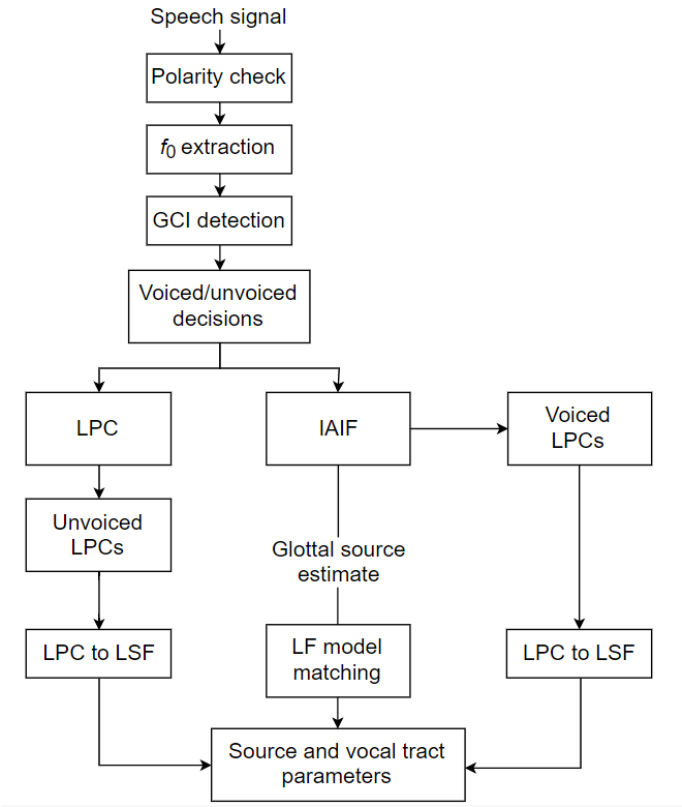


Figure 6.1: Flowchart of the GlórCáil analysis stage

6.1.1 Polarity check

The first step of the GlórCáil analysis stage is to check the polarity of the signal being analysed. The use of varying microphones and recording equipment results in speech data with different polarities. Correcting the polarity of a signal is important due to the fact that the accuracy and performance of many analysis techniques may be negatively affected if the polarity of the signal is inverted. In the case of this work, the LF-model matching process, described in Section 6.1.4, could provide less accurate results due to its use of polarity sensitive time domain measures. The polarity check is carried out using the method described in Drugman (2013). This method determines the polarity of a signal by first calculating the residual of the signal and a rough estimation of the glottal flow derivative. The skewness of the sample distributions of each signal are then calculate and subtracted from each other. The sign of the result is the polarity of the signal, with a negative value signifying an inverted or negative polarity meaning the signal must be multiplied by -1 to correct it.

6.1.2 f_0, GCI and voiced/unvoiced estimation

The next step is to estimate f_0, GCI locations, and VUV regions using the REAPER programme (Talkin, 2015). This programme is called from the MATLAB environment using a system call and it extracts the desired parameters to binary files which are then loaded into MATLAB. The process REAPER uses involves first removing low frequency noise and any DC component from the speech signal using high-pass filtering. A Hilbert transform can be applied that may reduce phase distortion in the signal by calculating its analytic representation. This analytic signal is a complex-valued function that has no negative frequency components (Smith, 2007). Next, the linear prediction residual is extracted and a set of normalised cross-correlation functions (NCCFs) are calculated from weighted combinations of the speech signal and the LP residual. Correlations are calculated for each possible fundamental period within the whole pitch range, defined by specified minimum and maximum f_0 bounds. A lattice of GCI candidates is constructed with associated costs derived from the NCCFs, and additional values representing the likelihood of a frame being voiced or unvoiced. These costs are then used in the dynamic programming step, where the candidates with the minimum cost are selected, considering the costs associated with the previous and following pulse candidates. The f_0 values are calculated by taking the maxima points of NCCF frames closest to the corresponding GCIs and computing the periods between the adjacent points.

6.1.3 Inverse filtering of voiced speech

Inverse filtering is then carried out on a frame-by-frame basis using the modified version of IAIF, GFM-IAIF (described in Section 2.3.1.3), with a frame length of 25 ms, a frameshift of 5 ms, and an order defined as $\lfloor fs/1000 \rfloor + 4$. DAP is used in place of LPC modelling in the GFM-IAIF process as it provides better estimations of the vocal tract transfer function (Alku and Vilkman, 1994). This analysis provides estimations of the DGF and the filter coefficients describing the vocal tract transfer function. The filter coefficients are converted to line spectral frequencies (LSFs) to make them less susceptible to distortions introduced by later processing steps and to make them more robust to statistical modelling in the cases where the vocoder is used in conjunction with an SPSS system.

LSFs are the roots of the line spectral pair (LSP) polynomial described in Itakura (1975a) and Soong and Juang (1984). The inverse filter resulting from an LPC analysis can be expressed in the *Z* domain by the m^{th} order polynomial:

$$A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + a_3 z^{-3} + \cdots + a_m z^{-m} \quad (6.1)$$

The LPC coefficients $(a_1, a_2, a_3 \ldots a_m)$ are very sensitive to quantisation errors, which can lead to filter instabilities. The *A(z)* polynomial can instead be expressed as:

$$A(z) = 1/2[P(z) + Q(z)] \quad (6.2)$$

Where *P(z)* and *Q(z)* are symmetric and anti-symmetric LSP polynomials respectively. These polynomials are defined as:

$$P(z) = A(z)\left(1 + z^{-(m+1)}\frac{A(z^{-1})}{A(z)}\right) \quad (6.3)$$

$$Q(z) = A(z)\left(1 - z^{-(m+1)}\frac{A(z^{-1})}{A(z)}\right) \quad (6.4)$$

The polynomials have the following properties when *A(z)* has all roots within the unit circle:

1. Their zeros all fall on the unit circle;
2. Their zeros are interlaced with each other;
3. The minimum phase property of *A(z)* is maintained after the zeros of *P(z)* and *Q(z)* are quantised.

The first two properties are valuable when calculating the zeros of *P(z)* and *Q(z)*, and the third property guarantees the stability of the filter derived from the polynomials (Soong and Juang, 1984).

6.1.4 Glottal source parameterisation

Once the differentiated glottal waveform has been obtained, model fitting is performed to parameterise each pulse using a method similar to that described in Kane and Gobl (2013) (see Section 2.3.4.3). Each analysis frame consists of a GCI(n) centred glottal pulse that extends from GCI(n-1) to GCI(n+1) and is windowed using a Hann window.

A range of pulses are generated using R_d values between 0.3 and 2.7 in 0.05 increments and an E_e value measured as the maximum negative value of the DGF pulse. E_e, as well as f_0 are used to calculate a set of default R-parameter and generate the corresponding LF pulse for each incremental value of R_d within the range. The time domain and spectral (amplitude and phase) correlation coefficients of the DGF pulses, when compared with the

generated LF pulses, are calculated and subtracted from 1 to give an error value for each of these measures. These error values are then sorted according to increasing error. The top ten R_d candidates that minimise the error values are selected for the next stage of the process. The next stage involves using a dynamic programming method (Ney, 1983) to calculate the optimum path of R_d values through the utterance. This is achieved by selecting the values that minimise target, transition and spectral stationarity costs (see Section 2.3.4.3). This method considers the context of a particular R_d value in relation to previous values as well as the effect of transitional regions of speech like vowel onsets and offsets. The optimal R_d values are then combined with their corresponding E_e and f_0 values as the source parameters. The extracted voice source parameters are then smoothed using a 5^{th} order median filter followed by a 5^{th} order moving average filter before being saved.

6.1.5 Analysis of unvoiced speech and parameter output

LPC analysis is carried out for unvoiced frames to approximate the unvoiced spectrum. The gain of the unvoiced frames is measured by taking the root-mean-square of the prediction error. The unvoiced LPCs are also converted to LSFs.

6.2 Synthesis Stage

Once parameters have been extracted from an utterance in the analysis stage, the GlórCáil synthesis stage (see Figure 6.2) can be used to resynthesise it. This stage includes an option to manipulate the voice source parameter contours of an utterance so that it can be resynthesised, and the results of the manipulations can be listened to straight away. Users can also compare a manipulated utterance with the original unmanipulated version directly in the interface. The process of parameter control will be discussed in more detail in Section 6.3.2.

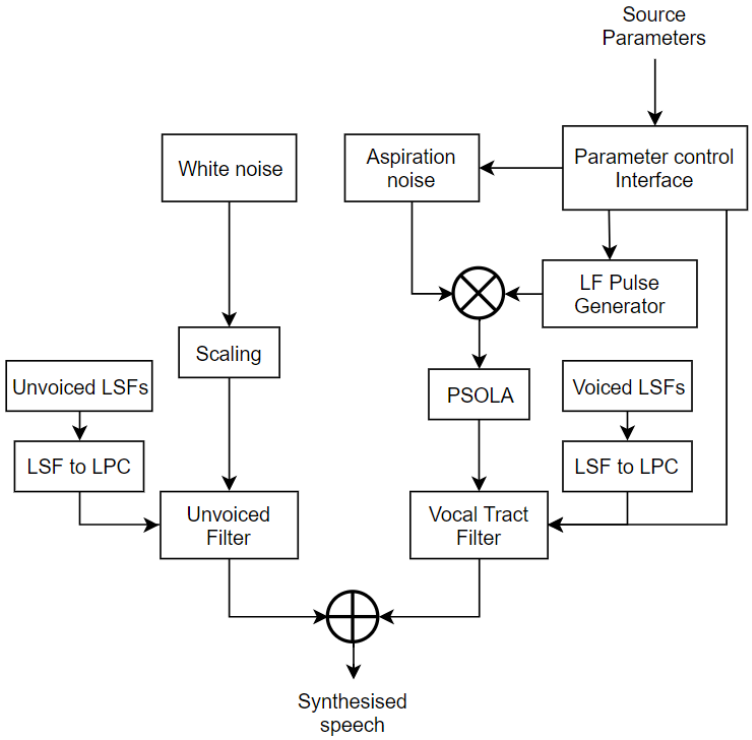


Figure 6.2: Flowchart of the GlórCáil synthesis stage

6.2.1 Excitation generation

The process of resynthesis begins by generating the voiced and unvoiced excitations. Voiced regions are defined by frames with f_0 values above and below a minimum and maximum frequency (default 50 and 500 Hz respectively). The voiced excitation consists of LF pulses with the addition of amplitude modulated Gaussian white noise.

The f_0 value from the first voiced frame is used to calculate the period of the first LF pulse using the relationship of $f_0 = 1/T_0$. The period is used to define the start and end points of the pulse frame within the voiced excitation. The parameters E_e, R_d and f_0 are used to generate the full set of LF-model parameters using the correlations outlined in Fant (1995), which are in turn used to generate the corresponding LF pulse. Amplitude modulated noise is added to the pulse according to the method described in Gobl (2006). The level of this noise is modulated according to the shape of the pulse. A pulse corresponding to a tense voice quality will have little to no added noise, while a pulse corresponding to a lax voice quality will have a considerable amount of aspiration noise added to it. There is an additional scaling factor that allows the noise level to be increased or decreased independently of pulse shape. Note, however, that this scaling factor is mainly used to calibrate the overall noise level. Once set, further adjustments should normally not be

necessary, as the noise level is automatically modulated depending on the shape and amplitude of the pulse.

This first pulse is placed within the start and end indexes that have already been defined, and the next pulse is calculated using parameter values from the closest frame to the end of the previous pulse. This process continues until the current voiced region has been filled with pulses. The pulses start and finish at values of zero, so no windowing is required to prevent discontinuities in the signal.

Frames with values of f_0 below the minimum f_0 threshold are treated as unvoiced regions. White Gaussian noise is used as the excitation signal in these areas.

6.2.2 Filtering of excitation

Once the excitation signals have been generated, they are filtered by a filter that represents the vocal tract transfer function. The voiced excitation is filtered using the LPC filter coefficients that are converted back from the LSFs that were output from the synthesis stage. A scaling factor may be applied to the filter coefficients to effectively shorten or lengthen the vocal tract. The frequencies of the poles are warped by performing a bilinear transformation in the z-domain. This approach involves substituting the original filter coefficients with first order all-pass elements using the mapping shown in Equation (6.5) (Härmä *et al.*, 2000),

$$z^{-1} \rightarrow \tilde{z}^{-1} = \frac{z^{-1} - \lambda}{1 - \lambda z^{-1}} \qquad (6.5)$$

where z^{-1} is the original filter coefficient, $\tilde{z}^{-1}$ is the transformed coefficient, and λ is the warping factor. The range of the warping factor is limited to between -0.1 and 0.1, where negative and positive values effectively shorten or lengthen the vocal tract respectively. This method is based on the implementation provided by Ellis (2004). This ability to scale the resonant peaks of the vocal tract adds an extra dimension of control and allows for further transformations to be made to the synthetic speech. The unvoiced excitation is filtered using filter coefficients and gain values obtained from LPC analysis. The final speech signal is then created by overlap-adding each of the filtered frames.

6.3 GlórCáil interfaces

An important element of this system is the user control interface. It allows a user to easily analyse speech and extract parameters which can then be used to resynthesise speech with or without voice quality manipulations. The GUI was created using the MATLAB GUI Design Editor (GUIDE) and is split into two stages, analysis and synthesis. Key points in the design process of these interfaces, particularly the synthesis interface, were gleaned from the results of the experiments described in Chapter 5. These points were that:

- Full and flexible control of the voice source parameters is required across a whole utterance to successfully elicit prominence in different locations.
- E_e and R_d were found to be effective in signalling prominence.
- Immediate feedback for the user is important when modifying a speech signal.

Each of these points were considered important features for inclusion in the synthesis interface that will be discussed in Section 6.3.2.

This section also details a modified version of the synthesis stage interface that was used in an experiment to test the effectiveness of the parameter control in signalling prominence in synthetic speech. The experiment is discussed in Section 7.1.

6.3.1 Analysis interface

The analysis interface (see Figure 6.3)(layout based on Dalton *et al.* (2014)) lets the user pick the directory containing the speech data that they wish to analyse. There is an option to check the polarity of either only the first file, or of each file in the directory individually. This option is useful if the user knows that all the speech data were recorded under the same conditions, as it saves time.

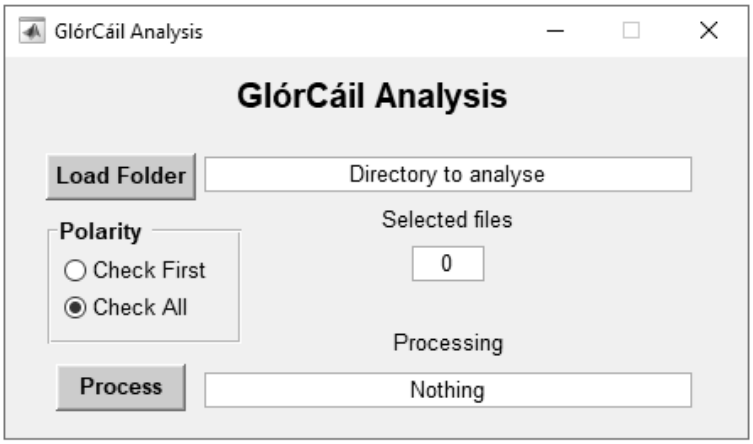


Figure 6.3: GlórCáil analysis GUI

Once the user has selected the correct directory and options, they can click the *Process* button to analyse the speech. A file containing the values of analysed parameters for each file is generated and saved in the same directory as the original speech data.

This interface is simple and easy to use, enabling users with any level of experience to analyse the speech data that they wish and obtain estimates of a small set of voice source parameters, R_d, E_e and f_0.

6.3.2 Synthesis interface

Once the analysis files have been generated the user can resynthesise speech using the synthesis interface (see Figure 6.4). After an analysis file is loaded into the interface, the user can manipulate R_d and E_e – two voice source parameters that were found to be effective at signalling prominence and changes in affective state in the experiments described in Chapter 5, as well as f_0. This level of control enables a user to manipulate the voice quality along the tense-lax continuum.

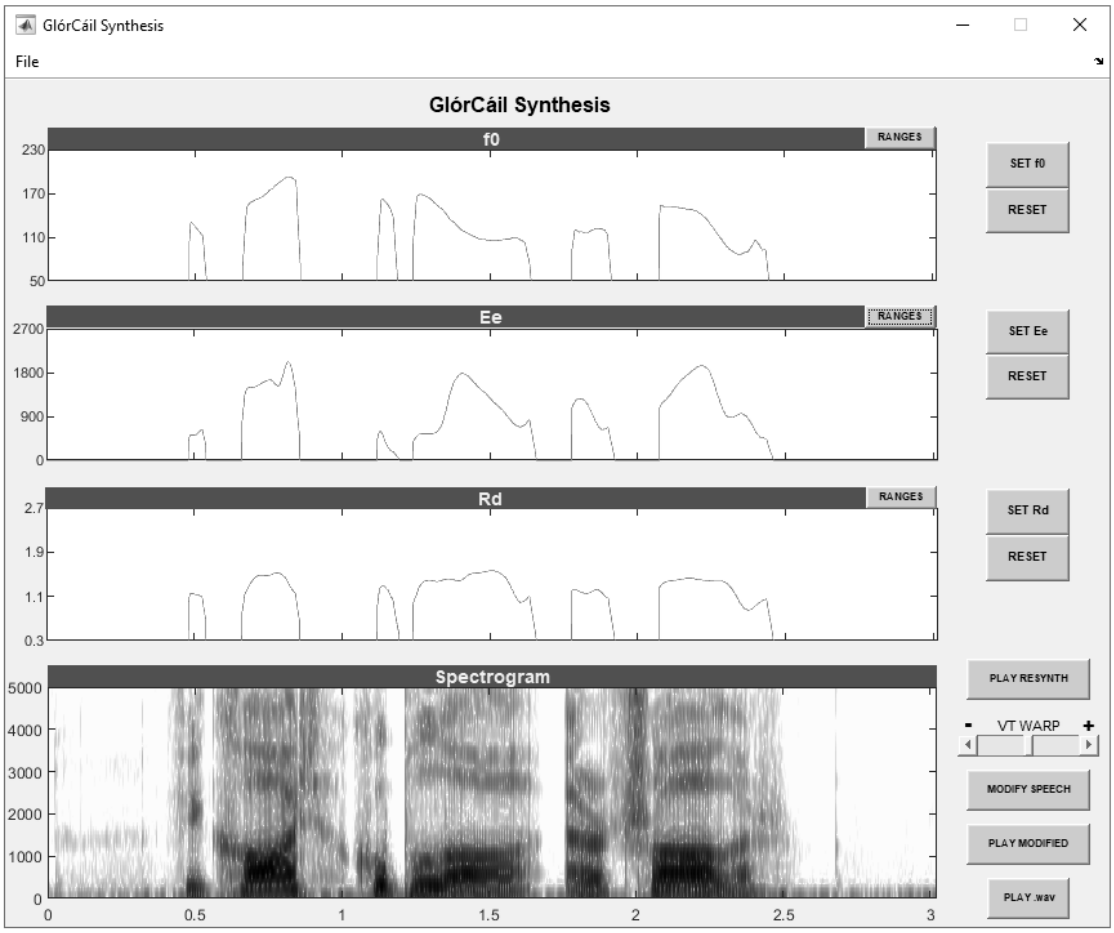


Figure 6.4: GlórCáil synthesis GUI

The user is presented with three frames, each containing a parameter contour that they can manipulate to create a desired output. The parameter contours can be manipulated by first clicking the *SET* button associated with each parameter, then left clicking within the parameter box along the length of the utterance, creating a stylised parameter contour. Once the chosen contour has been created, the user right clicks at the final point they wish to define. These points are used to generate a new parameter contour using linear interpolation. Areas of the original contour that have not been manipulated by the user remain unchanged.

An example of how a parameter contour, R_d in this case, may be set is shown in Figure 6.5.

(1) R_d is lowered within a region marked by the blue circles in an attempt to make this unit more prominent. The circles are defined by the user by clicking with the mouse. This has the result of making this region more tense.

(2) R_d is increased in the region after the section defined in step (1). This has the effect of making this region progressively more lax.

(3) A slope is defined preceding the section defined in step (1), lowering R_d and resulting in the voice quality transitioning from lax to tense.

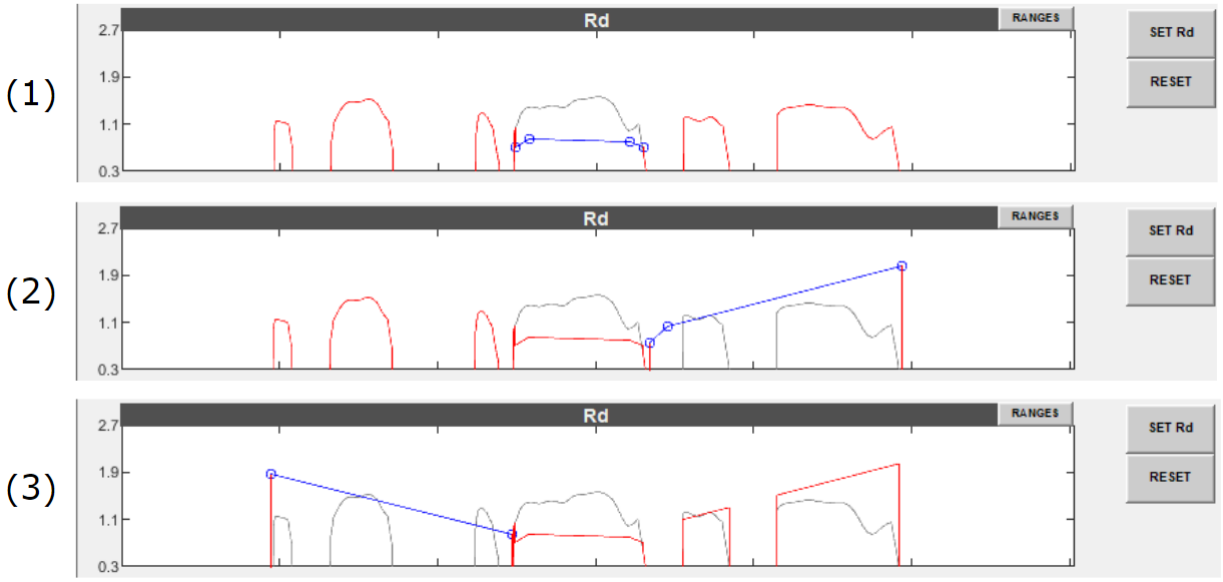


Figure 6.5: Illustration of contour setting in the GlórCáil synthesis GUI. The original parameter contour is shown in black, the set contour in red and the user defined points are circled in blue.

The ranges of each of the parameter frames can be changed by clicking the *RANGES* button at the top right of the respective frame. At the bottom of the window is a spectrogram so that the user can establish their position within the utterance. The user can also control the vocal tract warping factor using the slider marked *VT WARP*. By moving this slider from its default position, it is possible to simulate the acoustic effect of a change in vocal tract length.

The user can apply the changes they have made to synthesise a modified utterance. This is achieved by pressing the *MODIFY SPEECH* button. This utterance can then be listened to by pressing the *PLAY MODIFIED* button in order to compare it to the unmodified resynthesised utterance (*PLAY RESYNTH*), or the original speech file (*PLAY .wav*). This feature is useful when attempting to create a particular transform or shift in the modified utterance. The user can also choose to save the modified and unmodified resynthesised utterances to WAV files for later use through a dropdown menu at the top left of the window.

This interface allows the user to easily control R_d and the other voice source parameters, E_e and f_0, as well as vocal tract characteristics. It accomplishes this by enabling the user to define parameter contours using computer mouse controls and interactive frames.

6.3.3 Interface for perception experiments

One of the aims of this work was to demonstrate the effectiveness of the GlórCáil system using user-driven manipulation tasks. These are perception tests where the participants are given an objective and manipulate a parameter/parameters until they reach that objective. Another variant of the synthesis interface was developed for use in these manipulation-based perception tests. It is hoped that these types of test will prove useful, not only in the context of this work, but also in future work that is carried out in this area. Data obtained through user-driven tasks may give a more accurate representation of what the users are perceiving. It is also hoped that these types of interactive tests will be more enjoyable and less fatiguing for participants than some traditional perception tests, due to their level of involvement and engagement with the tasks.

Any desired utterance can be parameterised using the GlórCáil analysis stage and passed into this interface, along with temporal information, such as word or vowel boundaries. This temporal information is used to indicate the regions in which the researcher wishes to manipulate a voice source parameter/parameters. The temporal information is currently notated using Praat (Boersma and Weenink, 2019), but it is hoped that it will be integrated into the analysis stage of the system as future work.

The sliders used to control each region can be customised using any rectangular block images. The example in Figure 6.6 uses a set of sliding blocks corresponding to each word in an utterance "*We were away a day ago"*. The blocks in this example can be moved up and down in order to manipulate the R_d parameter contour.

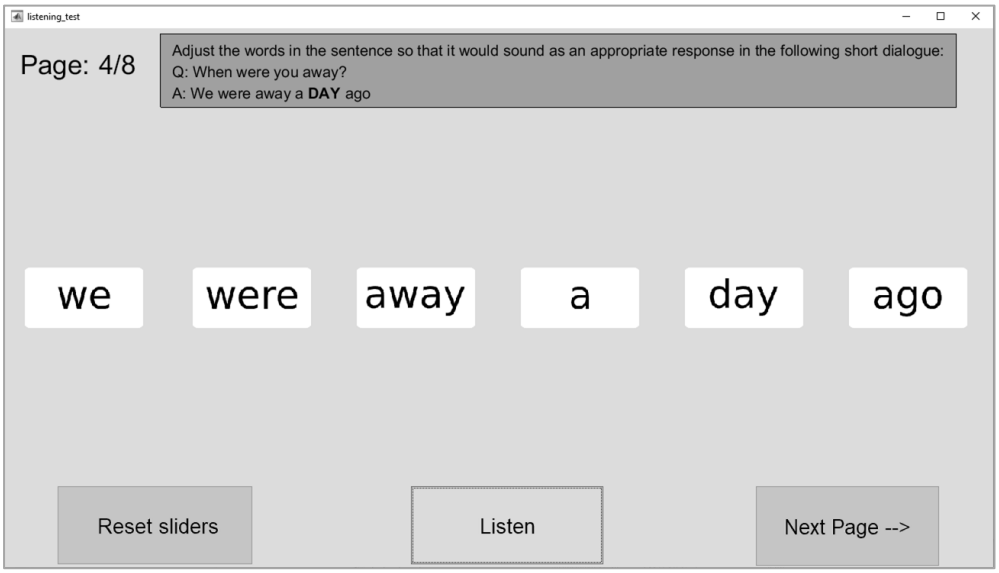


Figure 6.6: GUI for manipulation task perception task. Participants are presented with a synthetic sample to manipulate using the sliding blocks.

The responses of participants and the number of times they listen to each utterance are recorded each time the *Next page* button is pressed. This also moves onto the next sample and resets the slider blocks to their original positions.

6.4 GlórCáil-DNN

The previous sections in this chapter have discussed the GlórCáil system in the context of copy-synthesis or copy-synthesis with feature transformation. This section will describe how the system was integrated into a DNN-based SPSS framework.

6.4.1 Baseline system

The baseline SPSS system used in this work was built using the nnmnkwii python library (Yamamoto, 2017). This library makes prototyping of speech synthesis systems fast and easy due to its simplicity and transparent design. Some of the system's scripts are based on Merlin (Wu *et al.*, 2016) speech synthesis system demo scripts.

Speech corpora of Irish and Irish-English audio with a sampling rate of 16 kHz and a bit depth of 16 bits were used to train this system. Table 6.1 lists the corpora used, their biographical information and the number of utterances they contain.

Table 6.1: Speech corpora used to train DNN-based speech synthesis system

Language	Variety	Speaker	Gender	Number of utterances
Irish	Kerry	CMG	M	1452
Irish-English	Donegal	ANB	F	1122

The WORLD vocoder is used to extract log f_0, spectral and aperiodicity parameters. The spectral parameters are in the form of 60-dimensional MFCCs. An aperiodicity value is also extracted. These features, along with their first and second derivatives, are combined with a voiced/unvoiced measure to form the full acoustic feature set of 187 parameters. Full context linguistic labels in the HTS format are converted into vectors suitable as neural network inputs. The acoustic and linguistic features are then normalised to make them compatible with neural network training. The linguistic features are min-max normalised, while the acoustic features are normalised to zero mean and unit variance.

Two BLSTM networks are used by this system, one to model the duration and one to model acoustic features. Each BLSTM network consists of three hidden layers with 512 neurons in each layer implemented using PyTorch (Paszke *et al.*, 2017). The networks are trained by feeding utterance-wise sequences into the network and the Adam optimizer (Kingma and Ba, 2014) is used to update the network's weights after each training iteration.

Once the networks have been trained, maximum likelihood parameter generation (MLPG) (Tokuda *et al.*, 2000) is used to produce a smooth parameter contour from the values and their corresponding dynamic features of the neural network output. The generated parameters are then used by the WORLD vocoder to construct the synthetic speech waveform as described in Section 3.3.2.2.

6.4.2 Proposed system

The analysis in the proposed system is carried out in the MATLAB environment using the GlórCáil analysis interface (see Section 6.3.1). The extracted parameters are loaded and used to train the BLSTM network in place of the parameters extracted by the WORLD vocoder in the baseline system. The DAP and LPC coefficients are converted to LSFs, and f_0 values are converted to log f_0 and when combined with their first and second derivatives forms an acoustic feature set with a dimension of 249. The same network architecture is

used as in the baseline system with a network to model duration and one to model acoustic features, each containing three BLSTM layers with 512 cells in each layer.

Parameters are generated using MLPG in the same manner as the baseline system. The parameter generation stage remains the same as the baseline system except for slight modifications made to account for the different parameter dimensions used.

The parameters are then input into the GlórCáil synthesis stage and synthetic speech is generated using the method described in Section 6.2, where a user has the ability to manipulate voice source parameters prior to synthesis using the synthesis interface described in Section 6.3.2.

6.5 Chapter conclusions

This chapter described the GlórCáil system which allows for the manipulation of voice source and vocal tract parameters of synthetic speech. The analysis stage separates speech into its voice source and vocal tract filter components using established methods. The voice source signal is modelled using a glottal model controlled by a minimal set of parameters, which simplifies integration into SPSS systems. This reduced parameter set also means that direct control of voice quality manipulations is simplified for users of the GlórCáil system.

The synthesis stage reconstructs the source signal using a glottal source model and aspiration noise for voiced speech and scaled Gaussian noise for unvoiced speech. The source signal is then filtered by a filter that approximates the effects of vocal tract resonances.

The control interfaces created to control the two stages were also explained. The synthesis interface allows users to manipulate the voice source parameters and scale the vocal tract filter coefficients before resynthesis. The development of an interface for manipulation task perception experiments was also outlined and its use will be explained in more detail in Chapter 7.

This chapter also described how the analysis-and-synthesis system were integrated into a DNN-based speech synthesis framework.

Chapter 7. Transforming voice prosody and speaker characteristics of synthetic speech

This chapter outlines three experiments that were carried out to test the usability of the GlórCáil analysis-and-synthesis system as well as the user interfaces described in Chapter 6. These experiments are intended to illustrate how the system can be used to derive user-defined settings for prosodic signalling. They also demonstrate how the source might be controlled in speech synthesis and how the system can be integrated with a DNN-based SPSS system.

7.1 Experiment 4: Deriving user-defined settings for linguistic prosody

An experiment was carried out to investigate how voice source parameters would be manipulated to elicit focal prominence in a synthetic utterance[7]. An additional aim of this experiment was to demonstrate the effectiveness of the GlórCáil system at transforming linguistic prosody through the manipulation task interfaces described in Section 6.3.3.

The aim of this experiment was to obtain optimal user-defined settings for R_d that would signal focal prominence. This was carried out on an utterance synthesised with three voice qualities (breathy, modal and tense). It was hypothesised that the baseline voice quality of the utterance would influence the extent of R_d variation required to signal prominence.

The approach adopted in this experiment differs from previous studies: rather than rating premade stimuli, participants were asked to actively manipulate the R_d parameter, to generate the appropriate response to a question, i.e. with perceived prominence on a specific item in the utterance. The methodology is similar (though not identical) to an adjustment task used in Kreiman *et al.* (2007), where the listeners were asked to adjust

[7] Presented in Murphy *et al.* (2019)

jitter, shimmer and Harmonics-to-Noise ratio in synthetic stimuli to match natural voice samples or to the quasi-adjustment task used by Gussenhoven *et al.* (1997), involving judgment of the relative prominence of f_0 peaks. A manipulation task rather than an adjustment task was used here, that is, the participants were not given a naturally produced example sentence to match. Instead, they were presented with mini-dialogues constructed to elicit narrow focus and were asked to manipulate utterances synthesised with different voice qualities so that they sounded acceptable/natural as a response to a given question. The objective is to establish what degree of R_d salience is required to make a particular syllable prominent, and whether the R_d values required differ for the three voice qualities. By giving subjects direct control over acoustic parameters, it is hoped that one is obtaining indicators as to the actual R_d values that are optimal for signalling prominence.

The stimuli for the perception test were based on a recording of an all-voiced sentence 'We were away a day ago' spoken by a male Irish English speaker. The vowel quality in the potentially accentable syllables WAY and DAY was the same; in the original recording the duration of the vowels in these syllables was approximately the same (162 ms and 170 ms respectively). The reasoning for selecting just this single utterance was to allow for direct comparison between the results of this experiment and those described in Chapter 5 (see Sections 5.1 and 5.3).

The utterance was analysed and parameterised using the GlórCáil system. Three baseline sentences were created representing breathy, modal, and tense voice quality. The R_d values in each of these sentences were kept constant and were set to 1.6 for breathy, 1 for modal, and 0.7 for tense voice (based on the production data in Yanushevskaya *et al.*, (2016b) and auditory analysis). f_0 was set to its average value (104 Hz) with the addition of a degree of declination (8.5 Hz/s) and was kept the same in all three sentences. Amplitude modulated aspiration noise was added to each sentence according to the method described in Gobl (2006). Duration was not manipulated.

An informal auditory analysis, carried out by the author, confirmed that there was no prominence on the potentially accentable syllables in the resulting sentences; in other words, they were both equally non-prominent.

As the intention was not to vary f_0 and given the results of Experiment 3 (see Section 5.3), which suggested that a more perceptually effective R_d implementation involves fixed U_p and varying E_e, these were the settings used in the current experiment.

Word boundaries within the utterance were annotated in Praat (Boersma and Weenink, 2019) along with vowel midpoints and vowel boundaries. The R_d values were kept constant at word boundaries but were allowed to vary across the vowel segments during the manipulation task. Vowel midpoints were used to extract R_d values obtained as a result of the listening test described in Section 7.1.2. R_d was allowed to vary within ranges set for each phonation type and using scaling factors that were applied across vowel segments in manipulated words. Note that in *away* and *ago* only the initial vowels were manipulated.

7.1.1 Synthesis user interface

A user interface (introduced in Section 6.3.3) was designed for the manipulation task (see Figure 7.1). Each word was represented by a sliding block which, when dragged up and down, controlled a scaling factor. This factor was then multiplied by the corresponding region of the original (baseline) R_d contour. The scaling factors ranged between 0.5 and 2, so that the baseline R_d value could be halved or doubled at either end of the scale.

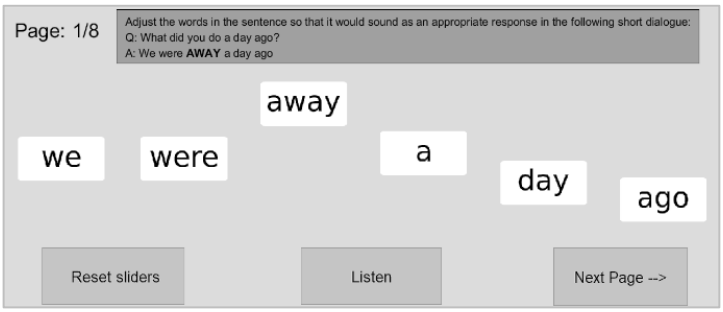


Figure 7.1: User interface for the prominence manipulation task.

Constraints were also applied so that the R_d values remained inside the ranges set for each voice quality: 1-2.3 for breathy voice, 0.6-1.4 for modal voice, and 0.35-1.4 for tense voice. These values were derived from data examined in Yanushevskaya *et al.* (2016b).

7.1.2 Listening test

42 participants (all native speakers of English) took part in the test. The participants were asked to manipulate the blocks/words in the utterances so that the resulting sentence could be an acceptable/natural sounding response in a mini-dialogue involving narrow focus, e.g.

Dialogue 1:	Q: What did you do a day ago? A: We were a**WAY** a day ago
Dialogue 2:	Q: When were you away? A: We were away a **DAY** ago.

There were eight utterances to manipulate. These included three instances of each of the dialogues, one for each of the three different voice qualities, in random order. Two more utterances were included at the beginning of the test to allow the participants to familiarise themselves with the procedure; the results from these were discarded. The participants were allowed to listen to the results of their manipulation and make changes as many times as they wished. The stimuli were presented through high quality closed back headphones in a quiet environment. The test took approximately 10 minutes to complete.

The specific hypotheses were that:

- Perceptually salient prominence can be generated by manipulating R_d (in the absence of f_0 variation).
- The magnitude of R_d excursions in the accented syllables would be different for different baseline qualities.
- Manipulations of R_d would differ depending on the location of the focally accented syllable in the utterance.

7.1.3 Results and discussion

The R_d values at vowel midpoints were extracted from each response (42 participants x 6 utterances = 252 responses). The average R_d contours and 95% confidence intervals of response sentences in which WAY and DAY were made prominent are shown in Figure 7.2. Note that the increased values of R_d correspond to laxer/breathier phonation and lower values of R_d to tenser phonation. Thus, the peaks in Figure 7.2 show an increase in phonatory tension.

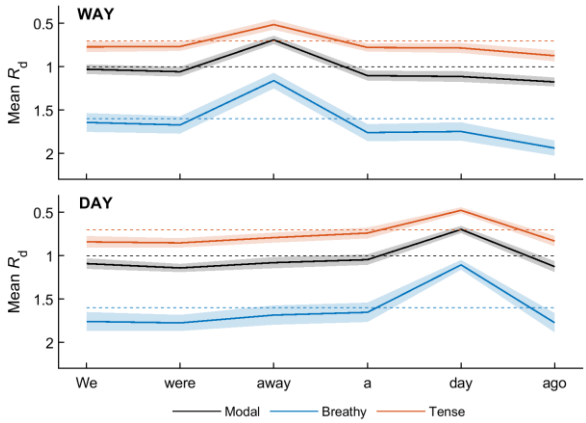


Figure 7.2: Mean R_d contours for WAY and DAY responses. Baseline values are shown by dotted lines. Shaded regions indicate 95% confidence intervals.

It is clear in Figure 7.2 that:

- The direction of R_d manipulation is the same across the three phonation types – towards lowering R_d /increasing tension relative to the baseline.
- The magnitude of R_d manipulation/excursion from the baseline varies with voice quality when plotted on a linear scale: the largest for breathy voice and the smallest for tense.
- R_d manipulation in pre- and post-focal material corresponds to changes towards breathier/laxer phonation relative to the original baseline.
- There is a considerable drop in phonatory tension/rise in R_d in the post-focal word *ago* when the focal syllable is WAY.

The magnitude of R_d manipulation/excursions in the focal syllables was measured as local 'protrusions' relative to the adjacent unaccented syllables (calculated as the difference between the focal R_d value and the average of the R_d values in the adjacent unaccented syllables and expressed as percentages relative to the ranges of R_d values).

Linear mixed-effect model analysis was used to test if the magnitude of R_d excursion is affected significantly by the baseline voice quality (VQ) and the location of the focal syllable (Focus) in the utterance. Analyses were conducted in the R environment (R Core Team, 2019) using the *lme4* [version 1.1-20] package (Bates *et al.*, 2015) for model fitting. The *lmerTest* package (Kuznetsova *et al.*, 2017) was used for step-down model simplification by eliminating non-significant effects and for calculating denominator degrees of freedom using Satterthwaite's approximations. The models were fitted using the

maximum likelihood (ML) method. The initial model included VQ and Focus as the main predictor variables (fixed effects) as well as their interaction; random effects included by-subject random intercepts and slopes: [Rd~VQ*Focus +(1+VQ|Participant)]. The final reduced model included VQ as the only fixed predictor and by-subject random intercept [Rd~VQ+(1|Participant)]. The location of the focal syllable (Focus) and the VQ*Focus interaction were not significant and were excluded from the model. ICC (indicative of the correlation of the items within a cluster) as well as marginal and conditional R-squared statistics (Nakagawa *et al.*, 2017) were obtained using the *sjPlot* package (Lüdecke, 2018). Marginal R-squared describes the proportion of the variance explained by the fixed effects; conditional R-squared indicates the variance explained by both fixed and random effects.

The summary of the estimated coefficients of the mixed effect model fitted to the R_d values (calculated as local 'protrusions' relative to the voice quality specific R_d ranges) obtained in the listening test is given in Table 7.1 (see also Figure 7.3). Based on marginal and conditional R^2 values, the amount of variance explained by the random effects amounted to about 40% of the variance. Fixed effect of baseline voice quality accounts for about 14% of the variance. Analysis of the fixed effects suggests a statistically significant association between perceptually salient R_d values in the focal syllables and baseline voice quality. While there is a drop in R_d across all voice qualities in the focal syllable, the magnitude of this drop is significantly less for tense voice compared to modal and breathy voice (β = 21.62, p<0.001 and β = 17.73, p<0.001 respectively, where β is the intercept and p is the p-value). The magnitude of perceptually salient R_d lowering is not significantly different in modal and breathy voice. As mentioned earlier, the location of the focal syllable in the utterance (earlier or later) has no significant effect on the magnitude of R_d excursions associated with that syllable within the same voice quality type.

Table 7.1: Estimated coefficients, confidence intervals and t-values (β, CI and t respectively) for the mixed effect model fitted to obtained R_d local protrusion values.

Mean (Intercept)	***β*₀**	***CI***	***t***	***p***
breathy	-44.71	-50.68 – -38.74	-14.67	<0.001
modal	-48.60	-54.57 – -42.63	-15.95	<0.001
tense	-26.98	-32.96 – -21.01	-8.85	<0.001
Contrasts	***β*₁**	***CI***	***t***	***p***
breathy vs. modal	-3.89	-9.07 – 1.29	-1.47	0.142
breathy vs. tense	17.73	12.55 – 22.90	6.71	**<0.001**
modal vs. tense	21.62	16.44 – 26.80	8.18	**<0.001**
Random effects				
ICC Participant	0.45			
Observations	252			
Marginal R^2 / Conditional R^2	0.142/ 0.531			

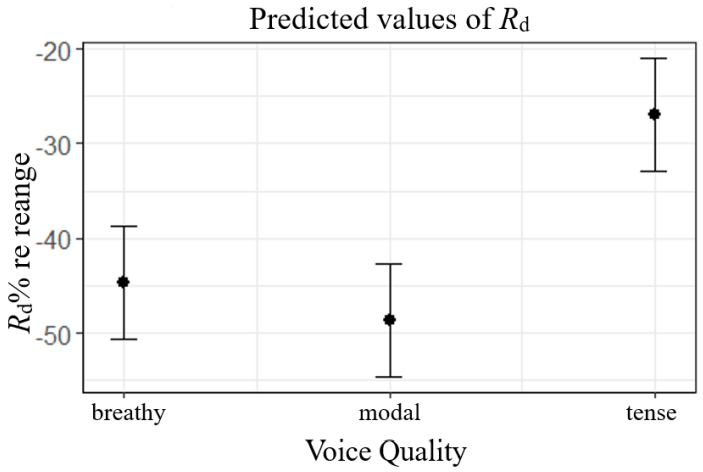


Figure 7.3: Predicted values of R_d and 95% CI (local protrusions relative to adjacent unaccented syllables expressed as % relative to voice quality R_d range).

Based on the analysis and given the initial difference in the R_d ranges across voice qualities, it appears that a higher degree of manipulation is required for modal and breathy voice than for tense voice. However, applying a logarithmic scale to the data would result in a closer representation of perceptually meaningful distances between voice qualities (see Figure 7.4). Mixed model analysis of the log-transformed data confirmed that the difference in the magnitude of manipulation across different voice qualities is not statistically significant. This suggests that, when using R_d as a control parameter in synthesis, logarithmic manipulations will be more appropriate.

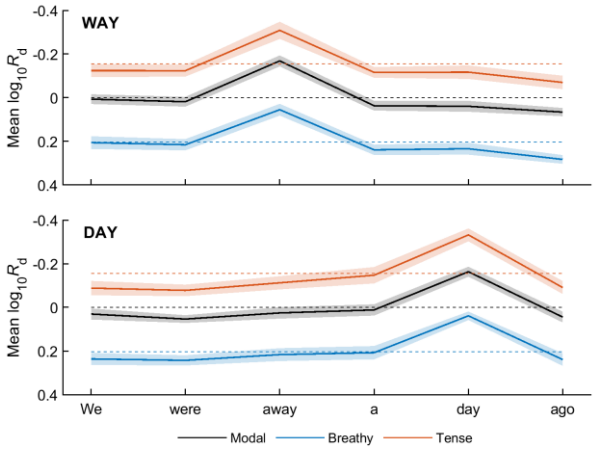


Figure 7.4: Mean $log_{10} R_d$ contours for WAY and DAY responses. Baseline values are shown by dotted lines. Shaded regions indicate 95% confidence intervals.

The results support the initial hypothesis that perceptually salient prominence can be generated by manipulating R_d (in the absence of f_0 variation). It appears that this manipulation is not confined to the focal syllable alone but affects the pre- and post-focal material and as a result, the overall contour of the utterance. The direction of the excursions was the same across synthetic qualities: the changes of the control parameter made by the participants in the focal syllable were all towards lowering R_d (perceptually tenser voice). Earlier studies with ready-made, predefined stimuli showed similar findings – that manipulating R_d towards tenser settings result in signalling prominence (at least for speakers of English and Irish who were participants in those earlier studies). Production data in Gobl (1988) also support these findings. However, this trend may not necessarily be independent of language, as is illustrated by the results in Vainio *et al.* (2010) for Finnish where focal syllables were characterized by higher NAQ values (laxer breathier phonation) (Alku and Vilkman, 1996).

The hypothesis that the magnitude of R_d excursions in the accented syllables would differ for different baseline qualities was supported when the values were normalized to the voice quality specific range and a linear scale was used. This is in keeping with the fact that the range of R_d values for laxer, breathier voice relative to modal (1-2.5) is larger than that for tense voice (0.3-1), meaning there is more room for adjustment when the R_d is in the lax/breathy range. This non-linear relationship was accounted for in Experiment 2 (see Section 5.2) by log-transforming the R_d values. In this study, the differences in the magnitude of R_d excursions were not statistically significant when the values were scaled logarithmically.

Manipulations of R_d did not differ depending on the location of the focally accented syllable in the utterance. The modelling of prominence using R_d in Experiments 1 and 3 (see Sections 5.1 and 5.3) was carried out by adding R_d excursions on top of an overall R_d declination, and the results of perception tests showed a significant difference in the perceptual prominence of a syllable depending on its phrase location (an earlier focal syllable was perceived as relatively more prominent than the one located later in the utterance). This trend did not emerge in the present results. The nature of task in this study was quite different from the previous studies insofar as the participants were tasked with generating an utterance with a prominent syllable rather than rating perceived prominence of predefined stimuli. It might be the case that additional R_d declination used in Experiments 1 and 3 (see Sections 5.1 and 5.3) plays a role in this difference, and it requires further exploration.

7.1.4 Conclusions

This experiment involved a user-driven manipulation task experiment in an effort to obtain and analyse perceptually salient R_d parameter contours that signal focal prominence. This experiment also aimed to demonstrate the capability of the GlórCáil system to transform the linguistic prosody of an utterance and demonstrate the manipulation task interface that was also developed in this work.

This type of experimental setup has not been widely used. Since the results represent the output of active manipulation of acoustic parameters, they can potentially give a more accurate picture of the extent of manipulation that would be required. Experiments 1 and 2 (see Section 5.1 and 5.3) found that pre and post-focal deaccentuation of R_d may increase the perceived prominence of a syllable. The findings of this experiment support these earlier results as the participants chose to manipulate not only the focal syllables but the pre- and post-focal material as well. The results clearly show that it is not enough to simply scale the overall contour linearly; as would be expected, the magnitude of the focal R_d peak needs to be scaled in proportion to the baseline R_d value. Although it has not been directly assessed in this experiment, it would be interesting to test and compare the relative importance of the extent of R_d decrease in the focal syllable (tenser phonation) and of the R_d increase (laxer phonation) in the pre-/post-focal material (deaccentuation) in signalling focal prominence.

The results of this experiment demonstrated that the GlórCáil system can be successfully used to transform linguistic prominence of a synthetic utterance. Additionally, the manipulation task interface developed in this work has proven to be a useful tool for carrying out user-driven listening tests. Participants informally reported to the author that they found this kind of test much more engaging and less fatiguing than the type of perception test carried out in Experiments 1, 2 and 3. It is hoped that by providing a more engaging task to participants a clearer representation of their perception will be obtained. This in itself is an interesting topic worth further future research.

7.2 Experiment 5: User control of paralinguistic prosody and speaker transformation

Two other aspects of the GlórCáil system are demonstrated in this experiment: the possibility of transforming the paralinguistic prosody (affective colouring) of speech as well as the extralinguistic voice properties of the individual speaker. The need to be able to control for the expressive, paralinguistic dimension of prosody has been discussed in detail in chapter 6, and its potential use in applications, such as interactive educational games, where the characters need to sound appropriate to the context of the game (angry, sad, etc.).

As mentioned in the introduction to this chapter, although not the main focus of the thesis, the current system allows for speaker transformation, i.e., the manipulation of voice quality in a way that would alter the perceived identity of the speaker. As discussed, the possibility of generating multiple characters/speakers from a single synthetic voice would be invaluable for interactive educational games/scenarios requiring diverse speakers in dialects, where currently a single voice is available, as is the case for Irish (Ní Chasaide *et al.*, 2017).

Both these aspects were tested in the following experiment. This experiment involved a manipulation task in which participants were asked to modulate a set of parameters in order to transform a baseline utterance to sound like (i) a target speaker (e.g., male, female, child), who is (ii) exhibiting a particular affect (e.g., sadness, anger).

In this experiment, voice quality was one of three parameters that were manipulated. The others involved f_0 and a parameter that controlled vocal tract size. These latter parameters

are particularly important in allowing for the transformation of targeted speaker-characteristics.

7.2.1 Materials

The baseline stimulus for the manipulation task was based on a recording of an all-voiced sentence 'We were away a day ago' spoken by a male Irish English speaker. The utterance was analysed and parameterised using the GlórCáil analysis stage (see Section 6.1). The extracted f_0 and R_d parameter contours were not modified. The average f_0 and R_d values were 104 Hz and 0.94 respectively. The stimulus was then resynthesised using the GlórCáil synthesis stage (see Section 6.2 for a description of this process).

7.2.2 Synthesis user interface

A user interface (introduced in Section 6.3.3) was designed for the manipulation task (see Figure 7.5). This interface allowed users to alter the parameter contours of the baseline stimulus. Three parameters were represented by the sliding blocks, which when dragged up and down controlled scaling factors. These blocks were labelled *Voice*, *Pitch* and *Size*. The *Voice* block controlled a scaling factor for the R_d parameter contour, *Pitch* controlled an f_0 scaling factor and *Size* manipulated a scaling factor for the length of the vocal tract. The first two factors were multiplied by the original (baseline) R_d and f_0 contours. The scaling factor for R_d ranged between 0.5 and 2, so that the baseline values could be halved or doubled at either end of the scale. The scaling factor for f_0 was between 0.667 and 3, as this approximated the range of f_0 values from adult male to child.The vocal tract scaling factor was used to modify the vocal tract transfer function in order to simulate changes in vocal tract length (see Section 6.2.2). The vocal tract scaling factor ranged between -0.1 and 0.1, with negative values effectively shortening the vocal tract and positive values lengthening it.

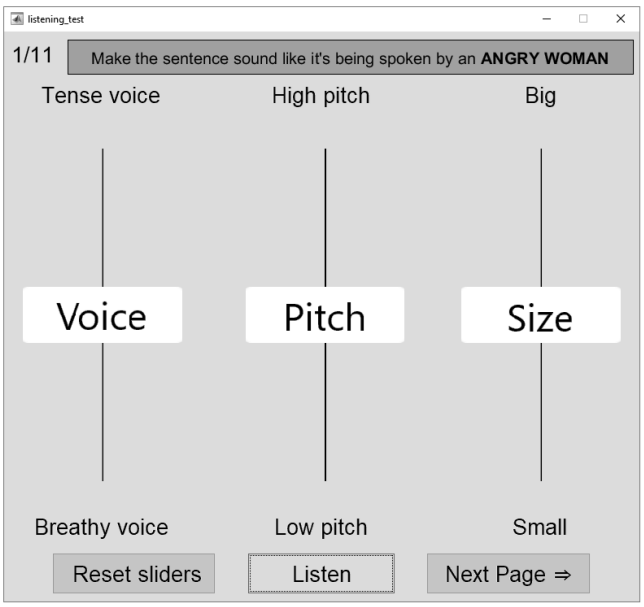


Figure 7.5: User interface for voice transformation task.

Constraints were also applied so that the R_d values remained inside its usual range of 0.3-2.7. Participants could listen to how their manipulations changed the utterance by pressing the *Listen* button, reset the sliders to their starting position by pressing the *Reset sliders* button, and move on to the next page by clicking the *Next Page* button.

7.2.3 Listening test

Eight participants, all native speakers of English, took part in the test. The participants were asked to manipulate the sliding blocks so that the resulting sentence sounded like it was being spoken by a particular speaker in a certain affective state, e.g., an angry woman, or a sad child. A full list of possible speaker and affect combinations is shown in Table 7.2.

Two more utterances were included at the beginning of the test to allow the participants to familiarise themselves with the procedure; the results from these were discarded. The participants were allowed to listen to the results of their manipulations and make changes as many times as they wished. The stimuli were presented through high quality closed-back headphones in a quiet environment. The test took approximately 10 minutes to complete.

Table 7.2: List of speaker and affect combinations for manipulation task

Speaker	**Affect**
man	*angry*
	sad
	no emotion
woman	*angry*
	sad
	no emotion
child	*angry*
	sad
	no emotion

A large body of research has established how male, female and child speech characteristics differ in terms of voice quality, f_0 and the size of the vocal tract. The goal of the test was to examine how the system parameter manipulations could be used to generate speech with gender/age specific characteristics and also with some degree of affect colouring.

Based on the descriptions of male, female and child speech characteristics in the literature, it was expected that:

- Participants would use higher f_0 scaling values when transforming to *child* than to *woman*.
- A negative vocal tract warping factor would be used for *woman* and *child*.
- A higher R_d scaling value would be used for *sad* voice.
- A lower R_d scaling value would be used for *angry* voice.

7.2.4 Results and discussion

The results of all adjustments are shown in Table 7.3 and Figure 7.6. It can be seen that, as expected, a larger increase in f_0 scaling is used to transform the speaker to *child* for each affect, but the values were not as high as expected. The case of *woman, no emotion* had the lowest scaling factor applied, whereas *angry* and *sad woman* had very similar f_0 scaling values. These values were also very close to *sad child* and *child* with *no emotion*. The highest f_0 scaling factor used was for *angry child*. The f_0 scaling values for *man* were very similar across affects.

Table 7.3: Average scaling factor adjustments.

Speaker	Affect	R_d scaling factor	f_0 scaling factor	VT warping factor
child	*angry*	0.66	2.28	-0.057
	sad	1.65	2.11	-0.062
	no emotion	1.09	1.97	-0.06
woman	*angry*	0.77	1.98	-0.05
	sad	1.63	2.01	-0.014
	no emotion	1.13	1.59	-0.047
man	*angry*	0.63	0.98	0.059
	sad	1.56	0.94	0.029
	no emotion	1.02	0.98	0.012

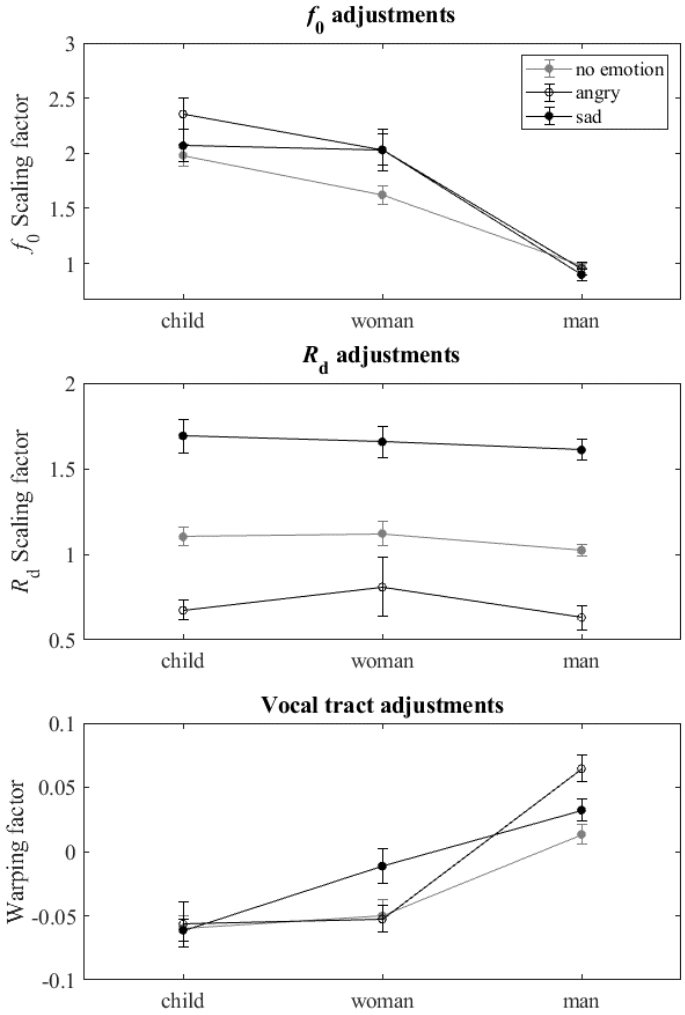


Figure 7.6: Average scaling factor adjustments of f_0 (top panel), R_d (middle panel), and vocal tract length warping factor (bottom panel) for each speaker type and emotion. Vertical bars show the standard error of the mean.

The vocal tract warping factor values exhibited a similar grouping. Similar values were used for both *woman* and *child*, excluding *sad woman*, where the warping factor was only

slightly lower than the baseline setting. One surprising result was that the highest vocal tract warping factor (longest vocal tract length) was seen for *angry man*. The author had predicted that the tenser voice normally associated with anger would lead to an effective shortening of the vocal tract due to an increase in muscular tension resulting in a slightly raised larynx, but the exact opposite was observed in this experiment. However, this could make sense if the proposed link between a larger size being perceived as more dominant and angrier (Ohala, 1984; Chuenwattanapranithi *et al.*, 2008) is considered.

The results for the R_d adjustments were just as expected, with higher scaling used to produce *sad*, and lower scaling used to produce *angry*. These values were largely independent of gender, with just *angry woman* showing a relatively high standard error of the mean, indicating a wider spread of values being used to produce this voice.

7.2.5 Conclusions

The goal of this experiment was to examine how the GlórCáil system parameter manipulations can be used generate speech with gender/age specific characteristics and affective colouring. The range of target speakers and affects that were specified for this voice transformation task was small, but was deemed sufficient as a proof of concept in the initial demonstration of the system.

The parameters appear to work in combination, and manipulations performed by the participants were mainly in the expected direction.

The results show that transformation of affective colouring was mainly dependent on the voice quality related parameter, R_d. Speaker gender/age transformation were mainly dependent on f_0 and vocal tract length.

As this experiment was an exploratory pilot demo of the system, conclusions based on the results could only be made tentatively. However, the results do indicate that participants usually made manipulations in the expected directions. It was therefore deemed acceptable to use the scaling factors gained from this experiment as the values for stimuli generation in Experiment 6, where the proposed system was tested when integrated into a DNN-based speech synthesis system.

7.3 Experiment 6: Voice quality control in a DNN-TTS framework

This experiment implemented the parameter manipulations found in Experiment 5 (see Section 7.2) into utterances synthesised using the proposed analysis and resynthesis system integrated into a DNN-based speech synthesis framework. This was in order to demonstrate how the system can be used to transform the paralinguistic prosody of synthetic speech, as well as speaker characteristics.

7.3.1 Materials

The stimuli for this experiment were generated by the baseline and proposed systems discussed in Sections 6.4.1 and 6.4.2 respectively. Both systems were trained using a corpus of 1452 utterances spoken by a male speaker of Irish (Kerry dialect) in his mid-20s. The audio was originally recorded in a semi-anechoic chamber using a B & K microphone at 44.1 kHz with a bit depth of 16, then resampled to 16 kHz before being used to train the speech synthesis systems.

Two sentences that were not used in the training of the systems were selected randomly and used as the text input to generate the stimuli for this experiment:

Sentence 1:

"Tá rang feadóg stáin ag tosnú ar an Máirt."

[t̪ˠaː rˠauŋg fʲad̪ˠoːg sˠt̪ˠaːnʲ əɟ t̪ˠosˠnˠuː ərˠ ənˠ mˠaːrˠtʲ]

"Is class whistle tin starting on the Tuesday." (word gloss)

"A tin whistle class begins on Tuesday." (translation)

Sentence 2:

"Tá cúrsa nua á sheoladh i gColáiste na Tríonóide, Baile Átha Cliath."

[t̪ˠaː kuːrˠsˠə nˠoː aː çoːlˠə I golˠaːʃtʲə nˠə tʲrʲiːnˠoːdʲə bˠalʲə aːhə clʲiə]

"Is course new being launched in College of Trinity, Dublin." (word gloss)

"A new course is being launched in Trinity College Dublin." (translation)

18 stimuli were generated using the proposed system (nine versions of each sentence). The stimuli were synthesised using the scaling factors obtained in Experiment 5 (see Section 7.2) to modify the voice source parameter contours of sentences 1 and 2 generated by GlórCáil-DNN. These corresponded to three different speakers (*child*, *woman* and *man*) in three different affective states (*angry*, *sad* and *no emotion*). The parameter contours that were scaled were R_d and f_0, as well as an overall vocal tract warping coefficient. Table 7.4 shows the full list of stimuli categories along with each of the scaling factors used. Two additional versions of sentences 1 and 2 were synthesised using the baseline system (see Section 6.4.1) making the total number of stimuli 20.

Table 7.4: Stimuli categories and their scaling factors

Speaker	**Affect**	***R*d scaling factor**	***f*0 scaling factor**	***VT warping factor***
baseline	-	-	-	-
child	angry	0.66	2.28	-0.057
	sad	1.65	2.11	-0.062
	no emotion	1.09	1.97	-0.06
woman	angry	0.77	1.98	-0.05
	sad	1.63	2.01	-0.014
	no emotion	1.13	1.59	-0.047
man	angry	0.63	0.98	0.059
	sad	1.56	0.94	0.029
	no emotion	1.02	0.98	0.012

7.3.2 Listening test

40 synthetic stimuli (two repetitions of the set of 20 stimuli) were presented to 20 Irish speaking participants in a listening test. The listening test was carried out using online survey software (Limesurvey GmbH., 2012) and participants were asked to take the test in a quiet environment using headphones. The stimuli were presented in a random order. The participants were informed that they would hear a number of different utterances, and that they could listen to each file as many times as they wished in order to answer the following questions for each stimulus:

1. Who is the speaker? [the participants chose from a selection of radio buttons with the options *child*, *woman* or *man*]
2. How does the speaker sound? [the participants chose from a selection of radio buttons with the affective labels *angry*, *sad* and *no emotion*];

3. To what extent emotion is present in the utterance? [a five-point scale ranging from 'Not at all' to 'A lot'];
4. How natural is the speech? [a five-point scale ranging from 'Not at all natural' to 'Very natural'].

Through auditory analysis carried out by the author, and another expert researcher, it was expected that:

- The stimuli generated with a particular target affect would be perceived as being coloured with that affect.
- The participants would be less likely to distinguish between the woman and child voices due to the similarity in the scaling factors.
- The participants would be less likely to label stimuli as angry than sad and no emotion.

7.3.3 Results and discussion

The responses from the listening test were divided into two main categories, related to speaker and affect identification. Speaker identification just involved the responses from question 1 of the listening test. Affect classification not only involved labelling of the perceived affect (question 2), but also the magnitude of the affect they perceived in the utterance (question 3). The naturalness rating (question 4) is an overall measure that incorporates both the speaker and affect transformation.

The results of the speaker identification are shown in the form of a confusion matrix in Figure 7.7. The results are presented as percentages of the total correct and incorrect identification of each group of stimuli (i.e. across each row). As was expected, the *man* stimuli were mostly correctly identified as being male. Both *woman* and *child* stimuli were often confused with each other. This was expected from the author's own auditory analysis of the stimuli. Even though the transformation scaling values were obtained from the manipulation task described in Section 7.2, they were often very similar. This resulted in producing stimuli that were difficult to categorise as being spoken by either a female or child speaker.

Based on the author's experience this appears to be the case when male voices are transformed using this method. This tendency is an interesting point in itself and could be

useful in the development of voices for applications where gender fluid voices are required, or a single voice could be used to represent characters with different genders and ages. Based on informal listening test carried out by the author transforming a female voice to male and child does seem to produce voices that are more readily identified 'correctly' in terms of target age and gender, but that was not the task of this experiment and will be explored in future work.

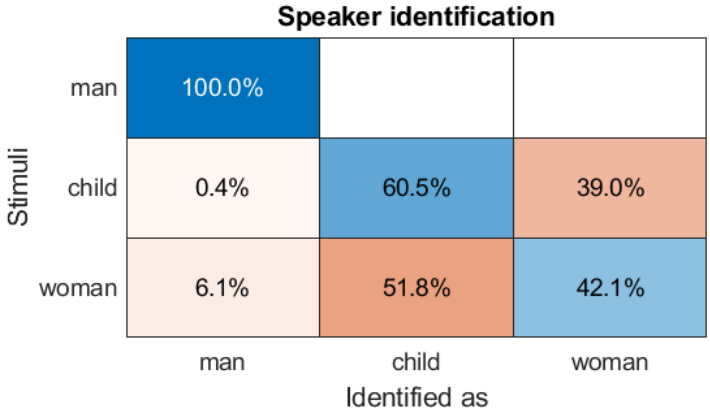


Figure 7.7: Confusion matrix of speaker identification

The results of the affect classification are shown in Figure 7.8 for the cases where the speaker was also correctly labelled. The *no emotion* stimuli were correctly identified 45% of the time, while *sad* stimuli were identified in 64.5% of the cases. The *angry* stimuli were only correctly classified 32.1% of the time.

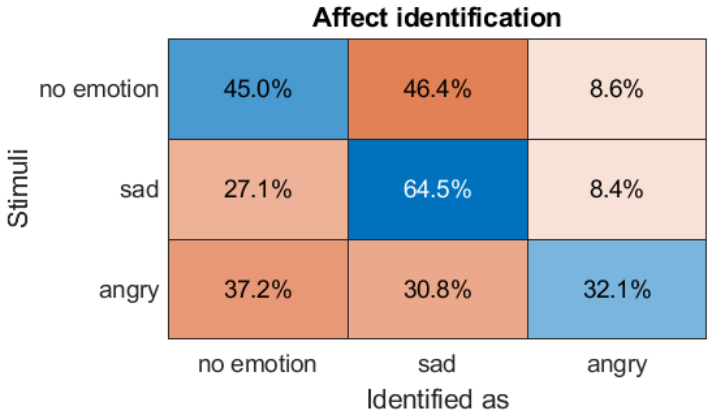


Figure 7.8: Confusion matrix of affect predictions for cases where the speaker was also identified correctly.

There are two factors that may account for the poor rate of identification of the affect *angry* in this experiment. First of all, the present implementation was of a simple, global kind of manipulation over the entire utterance. Past analytic work (Yanushevskaya *et al.*,

2009) indicates that within-utterance dynamic range of source modulations are also different for *angry* voice, and context-sensitive manipulations that capture these will need to be considered for future attempts. The very simple global application of a linear scaling across the utterance is clearly not enough.

A further factor that may have influenced the identification rates of *angry* may be the underlying meaning of the sentences. An *angry* rendition of these sentences requires considerable imagination.

Misidentification of one or more of the speaker (man, woman, child) stimuli could have contributed to lowering the rate of 'correct' affect identification. To see if this was the case the affect identification results were examined on a per-speaker basis. Figure 7.9 shows the results of the affect identification task for each speaker stimulus set.

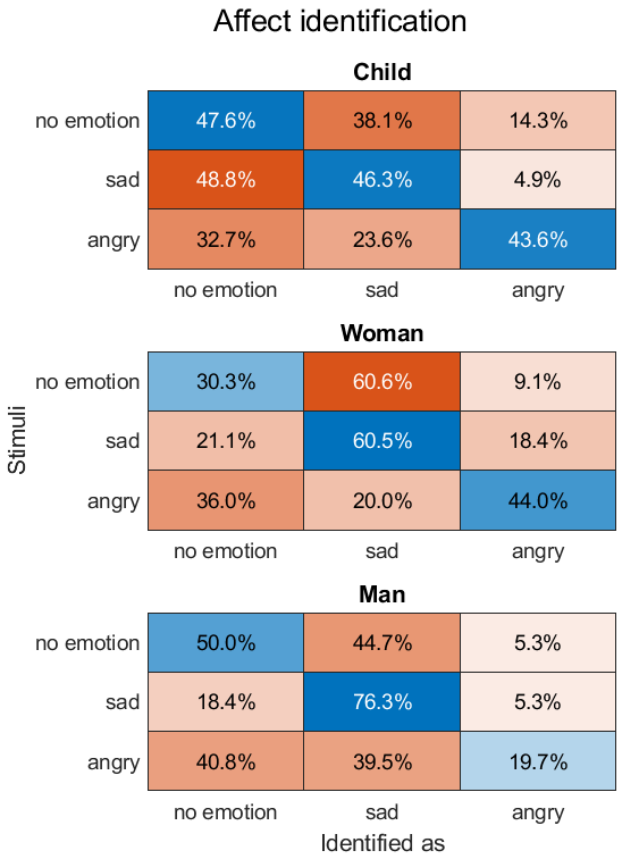


Figure 7.9: Confusion matrix of affect predictions for each target speaker set

***Child* stimuli**

The *sad child* targeted stimuli were identified an almost equal percentage as *no emotion* and *sad*, but almost never as angry. The *no emotion child* stimuli were often correctly

identified (47.6%), but also frequency identified as *sad* (38.2%). The *angry child* stimuli were most often identified 'correctly' (43.6%): these were often (32.7%) identified as having *no-emotion*, but infrequently (23.6%) as *sad*.

Woman stimuli

The *no emotion* and *sad woman* stimuli were both identified an equal percentage as *sad*, with a lower identification of *no emotion*. The *angry woman* stimuli were correctly identified 44% of the time.

Man stimuli

Angry man stimuli had a very poor rate of correct classification. This poor rating is what brought down the overall rate of classification of *angry* when the data is pooled across speakers, as in Figure 7.8. This suggests that one or more of the scaling parameters applied to these stimuli was far from optimal in signalling *angry*. *Sad man* stimuli had a very high rate of 'correct' identification (76.3%). The *no emotion man* stimuli were almost as often identified as *sad* as they were with their target affect.

Table 7.5: Mean magnitude for cases where speaker and emotion were correctly identified.

Speaker	Emotion	Magnitude	N	Std. Deviation
child	angry	3.13	24	.900
	sad	2.84	19	1.167
woman	angry	2.36	11	.924
	sad	3.00	23	.905
man	angry	2.47	15	.834
	sad	3.05	58	.963

The results of the affect magnitude rating for the responses where both speaker and affect were identified correctly are shown in Table 7.5 and Figure 7.10. The *angry child* stimuli were rated as having the highest affect magnitude, while *angry woman* and angry *man* had the lowest magnitude ratings. Conversely, *sad woman* and *sad man* were rated as having higher magnitudes than *sad child*.

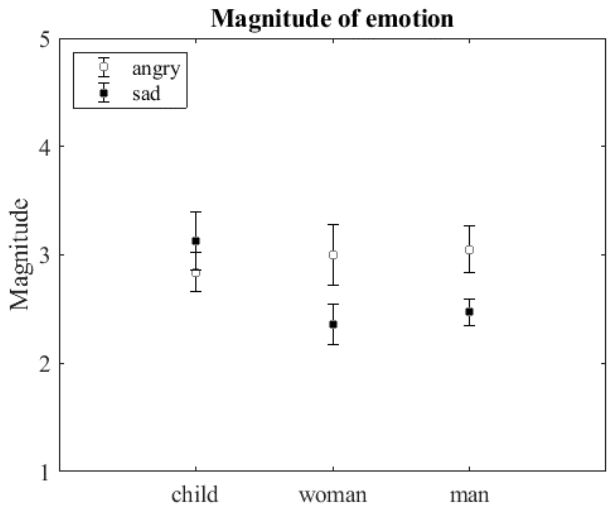


Figure 7.10: Mean magnitude of emotion ratings and their standard errors.

The results of the five-point naturalness rating for the responses where both speaker and affect were identified correctly are shown in Table 7.6 and Figure 7.11. The baseline stimuli had the highest naturalness rating at 3.96. This was closely followed by *angry man* at 3.93. The other *man* stimuli also had higher naturalness ratings than the *woman* and *child* stimuli; 3.81 and 3.74 for *sad* and *no emotion* respectively.

Table 7.6: Mean naturalness for cases where speaker and emotion were correctly identified.

Speaker	Emotion	Naturalness	N	Std. Deviation
baseline	no emotion	3.96	53	.999
child	angry	2.75	24	.608
	no emotion	3.05	20	.605
	sad	3.47	19	.841
woman	angry	3.00	11	.632
	no emotion	3.40	10	.516
	sad	3.13	23	.968
man	angry	3.93	15	.594
	no emotion	3.74	38	.921
	sad	3.81	58	.805

Both *child* and *woman* voices had lower naturalness ratings, with *angry* stimuli having the lowest ratings: 2.75 and 3 for *child* and *woman* respectively.

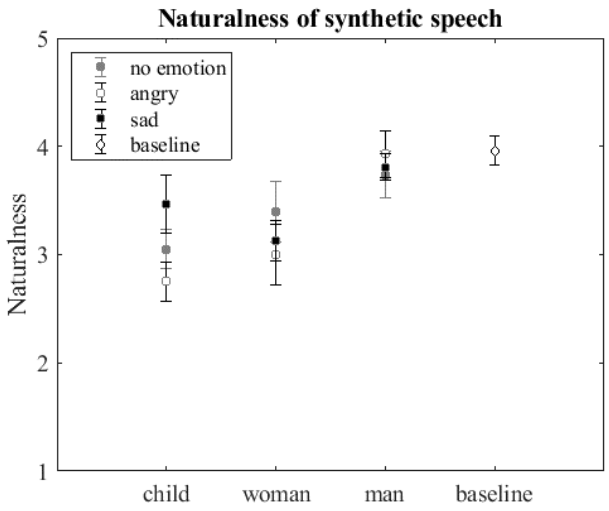


Figure 7.11: Mean naturalness rating of the stimuli

7.3.4 Conclusions

The aim of this experiment was to demonstrate the GlórCáil system in transforming a speaker into another target speaker with a different age/gender and affective state. Scaling factors were applied to voice source parameter contours as well as to the vocal tract filter, changing its perceived length. These scaling factors were obtained in the experiment described in Section 7.2. Some of the user-defined parameter scaling factors applied to the baseline utterances were not enough to differentiate between a woman and child speaker. This could be beneficial in applications where a voice is needed for a character with an undefined gender or age.

When all the responses where speaker identification was correct were examined it was found that there is some confusion between the *no emotion* stimuli being identified as both *no emotion* and *sad* an equal percentage of times.

The *sad* stimuli were identified correctly more often than not, and the *angry* stimuli were correctly identified at a rate just above chance.

When the responses are looked at for each individual speaker, a different pattern can be seen. Both *angry child* and *angry woman* were correctly identified a much higher percentage of the time when compared to the overall response. The *angry man* stimuli were only identified correctly 19.7% of the time.

The magnitudes of affect for *angry man* and *angry woman* were much lower than that of *sad man* and *sad woman*. The opposite trend is observed in the *child* stimuli.

The baseline system was found to produce the most natural utterances, closely followed by the male voice in all affective states produced by the proposed system. The woman and child voices were found to be less natural, but only by a small margin for certain affects. The *angry child* stimuli were rated as being the least natural. This could be due to high degree of vocal tract warping applied. Transforming the speaker gender, age and emotion tended to lower the perceived naturalness of the stimuli. Overall, these results are very promising and future research will be invested in increasing the level of naturalness so that it becomes more in line with the baseline example.

Although some of the samples produced by the proposed system were not correctly identified as being produced by the intended target speaker and/or with the intended target affect, they were identified as being different from the baseline. The present results should be viewed as an initial pilot test, that serves to illustrate how the system can be used to explore these aspects of speaker and affect transformation in synthetic speech. Future research will need to employ more context-sensitive transformations with the guidance of more detailed analysis of speech production data.

7.4 Chapter conclusions

This chapter described three experiments that were aimed at:

(i) Illustrating the use of the GlórCáil system as an experimental tool to explore the use of, and to obtain user-defined settings for, voice quality modulation in prosodic expression.

(ii) Exploring how the system might be used to transform the voice source in synthesis, both in the modulation of affect and in the transformation of speaker characteristics.

Experient 4 further investigated the findings of two of the experiments carried out in Chapter 5 concerning R_d settings for focal prominence. A manipulation task was created where participants controlled the R_d contour of an utterance and attempted to create focal prominence as appropriate in a previously given (written) scenario. This task was carried out using utterances with breathy, modal and tense voice qualities. One of the advantages of this type of user-driven experiment is that it can potentially provide a more accurate

estimation of the optimal manipulations required for signalling focal prominence in utterances with different voice qualities. Results indicate that it is not enough to simply scale the overall contour linearly, but to scale the magnitude of the focal R_d peak in proportion to a given baseline R_d value. Using this approach, it is hoped that it will be possible to devise initial recommendations regarding the types of scaling factors that will be required to generate the appropriate shifts in focal accentuation in speech synthesis.

Experiment 5 was a pilot experiment, aimed at illustrating how the users could readily manipulate parameters with which to transform speaker type and speaker affect using the GlórCáil system. This involved participants carrying out a task where parameter scaling factors were manipulated in order to transform the speaker and affective characteristics of a baseline utterance, produced by an adult male speaker of Kerry Irish. Instructions to participants were to vary acoustic parameters (R_d, f_0 and a vocal tract scaling factor). The results of this experiment were mostly as expected. A higher R_d scaling factor was used to transform the baseline to *sad*, a lower R_d was used for *angry*, and minimal scaling was used to produce *no emotion*. f_0 and vocal tract scaling were mostly used to transform the speaker: higher f_0 and shorter vocal tract scaling were used for producing *woman* and *child*. The exception to this was for *sad woman* where the vocal tract was not shortened as much as for *woman* in the other affective conditions. The *angry man* condition also produced unusual results, with the mean value of the vocal tract scaling factor acting to lengthen rather than shorten the vocal tract as had been expected

Experiment 6 was a further pilot exploration of how the speaker affect and speaker type might be altered in the context of a DNN-based speech synthesis system. It consisted of participants categorising stimuli constructed to target differences in terms of speaker and affect, while rating the magnitude of affect and naturalness of the stimuli. The stimuli were generated using the results of Experiment 5 to scale the R_d, f_0 and vocal tract parameters. The *child* and *woman* stimuli were most often confused with each other. The *sad* and *no emotion* stimuli were also frequently confused with each other for the stimuli targeted as *child* and *woman*. In the case of *man* targeted stimuli, *sad* was correctly identified a high percentage of the time, but *angry* was poorly identified. Overall, the naturalness ratings were quite high, with the baseline and *man* stimuli being rated the highest. The results indicate that the types of transforms that will be required to signal affect will probably

need to be more complex, entailing context-sensitive differences of voice source dynamics, as suggested in earlier research (Gobl and Ní Chasaide, 2002; Yanushevskaya *et al.*, 2009).

Nonetheless these experiments establish how the system can be deployed as a research tool to investigate the voice source correlates of prosodic features such as focal prominence, in a way that can be directly implemented in synthesis. They also illustrate how this system can be used to control affect and speaker characteristics in the generation of synthetic speech. Although the experiments focused on only male speakers and a small number of utterances, this allowed for an in-depth examination of voice prosody control in these cases.

It is hoped that this system will provide a basis for the research that will be required for the provision of the complex prosodic realisations needed in future speech technology applications. By being explicitly based on a voice source model and on our understanding of speech production, this approach also has the advantage of permitting the experimental basis for future exploration into speaker-specific characteristics.

Chapter 8. Discussion and conclusions

This thesis explores how linguistic and paralinguistic prosody might be controlled in speech synthesis using the global waveshape parameter, R_d, as the primary control parameter. This parameter may be used to modify the voice quality of synthetic speech along the tense-lax continuum.

The background chapters (Chapters 2-6) introduced and described the topics of speech production, voice source analysis, speech synthesis, speech modelling in speech synthesis and the role of voice source in prosody. Chapter 5 describes a set of analytical experiments involving synthetic stimuli, and perception tests, carried out to test various control schema for the R_d parameter. Chapter 6 outlines the analysis and resynthesis system that was developed. The knowledge gained from the results of the experiments described in Chapter 5 provided the basis for several design features. This system allows a user to control linguistic and paralinguistic prosody in synthetic speech. The system includes a visual interface where users can easily manipulate voice source parameters by manually "drawing" contours. Interfaces were also created for use in the experiments described in the subsequent chapter, which allow for intuitive user control in order to determine optimal R_d settings for voice source manipulation. Chapter 6 also describes how the GlórCáil analysis-and-synthesis system was integrated into a DNN-based speech synthesis framework. Chapter 7 described three experiments that were carried out, demonstrating how the GlórCáil system may be used to modify linguistic and paralinguistic prosody, as well as demonstrating the voice quality control capabilities of the system when integrated into an SPSS framework.

This work began by carrying out three experiments exploring how the global waveshape parameter, R_d, could be used to control linguistic and paralinguistic prosody. Experiment 1 explored the perceptual importance of voice source adjustments, namely in R_d, which have been observed in sentences with variable location of focal accent. It also examined whether such voice source adjustments on their own might be capable of shifting the perception of the location of focal accent within the sentence. The results of Experiment 1 (see Section 5.1.3) indicated that R_d was effective at signalling focal prominence but suggested that

phrasal position is important to the realisation, as the signalling of prominence was relatively ineffective for the phrase-final accent. It was considered that a phrase-final (default nuclear) accent may simply require f_0 salience (e.g., a falling tone) in order to be perceived as prominent, whereas source boosting changes alone (without f_0 salience) may suffice in other positions. It was also suggested that, the source deaccentuation (shift towards a lax voice quality) in the tail is an important aspect of signalling focus, and that the shorter tail available in the phrase-final condition may have been a factor in the reduced signalling effect of the source manipulations. The possibility of other factors (not related to phrasal position) having influenced results is also discussed – such as the difference in the vowel qualities (of 'way' and 'year') in the targets for focal accentuation.

Experiment 2 involved the modification of R_d parameter contours to simulate the phonatory tense-lax continuum and to explore its affective correlates in terms of activation and valence. A range of synthetic stimuli were generated varying along the tense-lax continuum using R_d as a control parameter. These stimuli were based on a natural utterance which was inverse filtered and source-parameterised. The generated stimuli were presented in a listening test where participants chose an emotion from a set of affective labels and indicated its perceived strength for each stimulus. Participants also indicated the naturalness of each stimulus and their confidence in their judgment. This experiment demonstrated that stimuli ranging in the tense-lax continuum generated by manipulating R_d can evoke an affective response and that the magnitude of the response corresponds to the magnitude of the manipulation. Even though manipulations of R_d were effective in evoking an affective response in terms of activation, they were not so successful in distinguishing between valence states. It is thought that this is most likely due to the lack of vocal tract parameter manipulations.

Experiment 3 further examined the potential of R_d to lend the perceptual salience that can signal focal prominence. It followed on from Experiment 1 and looked at the extent to which R_d might signal focal shifts within non-final elements of an utterance. It also explored the different ways of calculating R_d changes. This was achieved by generating the two series of stimuli incorporating the two possible alternatives to test the effectiveness of each implementation. A series of stimuli where U_p was kept constant and E_e could vary and another series where E_e was kept constant and U_p could vary were generated. The results of this experiment once again demonstrated that R_d can be used as an effective way of

controlling of focal prominence. However, it also showed that it again depends on the phrasal location of the focal accent. It was also found that, of the two possible implementations of R_d, the implementation where E_e was allowed to vary was far more effective at signalling prominence.

Overall, the results of these experiments suggest that R_d is a useful control parameter to generate linguistic prominence and affective colouring. The results also confirmed the findings of earlier studies, that suggested that tenser phonation is used to signal focal prominence, and that source deaccentuation of post-focal material towards laxer phonation also has strong prominence-lending effects. They also pointed to the likelihood that there is an interaction with the overall phrase prosody, in that the location of the focal accent in the intonational phrase appears to be important.

The knowledge gained from the results of the initial three experiments fed into the development and construction of an analysis-and-synthesis system, GlórCáil, that could be integrated into a text-to-speech synthesis framework, and that allows for the control of parameters for prosodic and speaker characteristic transformation. To gain control of these parameters, the glottal source, and the vocal tract filter that shapes it must first be modelled. These speech components must be separated effectively to do this. Inverse filtering was used to remove the effects of the vocal tract transfer function from the speech signal, thus giving the source signal, plus the filter coefficients of the vocal tract. An acoustic glottal model was then used to parameterise the periodic component of the source signal in voiced speech.

A set of GUIs were developed that allow the user to analyse and resynthesise speech. Before resynthesis is carried out, the user has the ability to manipulate the voice source and vocal tract parameters of an utterance to alter its prosodic pattern and perceived speaker characteristics. The interface also allows the user to instantly listen to any changes they make, to see if they had the desired effect or if further manipulations deemed necessary. This finished system was integrated into a DNN-based speech synthesis framework so that it could be used to generate unseen utterances. Unlike other approaches of speech synthesis incorporating acoustic glottal models, this system includes a user interface that allows for the transformation of voice source parameters that may be used to alter the linguistic and paralinguistic baseline of an utterance. It also focuses on simplifying the control scheme by

using a minimal set of control parameters to allow the system to be integrated into applications used by non-experts.

The system's ability to transform linguistic and paralinguistic prosody, as well as speaker characteristics, in copy-synthesis was demonstrated in Chapter 7, through two manipulation task listening tests. These manipulation tasks used interfaces developed as part of this work. In the first test, participants were asked to modify an utterance so that it sounded like an appropriate response (in terms of the location of focal accentuation) to a given question by moving sliders that controlled the R_d parameter contour of the utterance. For the second experiment, participants were asked to move three sliders to make an utterance sound like it was being spoken by a given speaker (child, woman or man) in a given emotional state (angry, sad or no emotion). The sliders controlled scaling factors that were applied to the R_d and f_0 contours, as well as the vocal tract filter coefficients that determined the perceived length of the vocal tract. The responses from this task were then used to modify the default parameters generated by the DNN-based speech synthesis system to create a set of stimuli. These stimuli were used in a listening test, where participants were asked to identify the speaker, the emotion of the speaker, and rate the magnitude of the emotion and naturalness of the utterance on five-point scales.

The results of the experiments discussed in Chapter 7 indicated that the GlórCáil system can be used to transform the linguistic and paralinguistic prosody of speech. The results of transforming the speaker characteristics of synthetic speech were less successful in terms of the rated naturalness of, and differentiation between female and child voice. The affect-transformations for a male speaker are promising but indicate that further exploration of parameter configurations is necessary. It should be noted that the last two experiments concerning affective and speaker transformation are intended as initial explorations, using a limited amount of test data, with the goal of illustrating the use of the system rather than as in-depth studies of these transformations. As such they provide a direction for how future studies may be possible using the system developed here.

The original aims of this work were to:

- Investigate how a global waveshape parameter, R_d, might allow for control of linguistic and paralinguistic prosody in speech synthesis.

- Implement a speech analysis-and-synthesis system that would allow a user to easily control R_d and other voice source parameters, as well as vocal tract characteristics.
- Integrate the system into an SPSS system.
- Demonstrate the effective use of this system through a set of user-driven manipulation tasks.

All of these tasks were accomplished within this work in the pursuit of answering the following question:

> ***To what extent can control of both linguistic and paralinguistic prosody be achieved by manipulating the voice source and vocal tract in statistical parametric speech synthesis utilising an acoustic glottal source model with a minimal set of control parameters?***

One of the main findings of this work is that the R_d parameter does affect the voice quality modulations in the tense-lax dimension that impact on the baseline prosodic pattern. It is a useful control mechanism in working towards a better control of linguistic and paralinguistic (expressive) prosody in speech synthesis. The fact that changes in perceived linguistic and paralinguistic prosody can be elicited through manipulation of this single voice parameter – even in the absence of f_0 modulation – is very promising in terms of being able to implement very minimalistic and simple control schema for linguistic and paralinguistic prosody in speech synthesis.

Future work will be needed to explore the voice quality dimension of other aspects of prosody not addressed in this thesis. It will also be important to examine how voice source variations combine with the (relatively well understood) melodic aspects of prosody (f_0 modulation). Ideally, we seek to define a model of prosody that will encompass the different dimensions of the voice source, as well as methods that will enable them to be effectively controlled within speech synthesis systems. Ultimately a better understanding of human prosody will pave the way towards speech technology applications that are more adapted to users' needs and the contexts of use.

Unlike some other examples of analysis-and-synthesis systems that employ acoustic glottal source models (Cabral *et al.*, 2011; Degottex *et al.*, 2011b; Huber and Roebel, 2015), GlórCáil offers an intuitive and easy to use interface where users may make changes to a

small set of voice source parameters that can result in prosodic changes and some level of affective colouring being applied to synthetic speech produced by an SPSS system. Although the system described in Sorin *et al.* (2017) does offer voice quality and speaker characteristic transformation through a web-based GUI, it uses proprietary software and synthetic voices, as well as being based upon unit selection speech synthesis rather than SPSS.

Although synthetic speech produced by the system was rated as being comparatively natural to a baseline DNN-based speech synthesis system, the naturalness dropped relative to the magnitude of vocal tract and f_0 transformations. This may be due to the derived transformation scaling factors being inadequate, the transformation process introducing distortions or artefacts that reduced the perceived naturalness – or to a number of other unknown factors. Another point to mention is that, while the synthesis interface of GlórCáil does offer a reasonably high level of control over a minimal set of voice source parameters, improvements could be made to the way in which the parameter contours are generated. One such improvement would be to form an integrated control scheme, controlling the parameters using a two-dimensional spidergram configuration. The user would be presented with a multi axis spidergram, defining the control parameters for a particular frame or region of an utterance. By moving a cursor around the spidergram the user could control multiple parameters at once. This method has been used to visualise voice source parameters (Yanushevskaya *et al.*, 2010), but it also offers an interesting approach to parameter control that warrants future investigation.

In its current iteration, the manipulation tasks using the present system and interface still rely on external timing annotations from Praat in order to be set up. Integrating this process into the system would be far more convenient. The system also lacks the means for applying global dynamic control of parameter contours, which may be necessary to quickly apply affective colouring to a synthetic utterance. A user may apply these transformations manually, but a semiautomated method would be optimal. The addition of these features and functions will be considered in future work.

8.1 Contributions of this work

8.1.1 Knowledge of voice quality and prosody

Experiments were carried out to establish how modifications of voice quality, using a small set of voice source parameters, can be used to control linguistic and paralinguistic prosody.

Experiments 1 (see Section 5.1) indicated that R_d was an effective control parameter in cueing focal prominence in the example utterance used, and that peaks in R_d (towards tenser phonation) on focally accented syllables may work with source deaccentuation in the post-focal material to enhance perceived prominence. Another point indicated by the experiments was that the magnitude of perceived prominence may depend on location within an utterance. When attempting to cue focal prominence towards the end of an utterance, an f_0 contribution, along with an R_d peak, may also be required.

Experiment 2 (see Section 5.2) showed that R_d could also be used as a control parameter for paralinguistic prosody. A set of R_d values were used to generate stimuli ranging along the tense-lax continuum. It was found that these stimuli could evoke an affective response and that the response correlated with the magnitude of the R_d modification relative to the baseline. Thus, the tenser the voice, the higher the perceived magnitude of, say, anger

Experiment 3 (see Section 5.3) demonstrated how voice source modulations affecting phonatory quality are perceptually important in signalling prominence, operating both at a local level in boosting the prominence of the accented syllable and at a global level in attenuating the prominence of other portions of the utterance. Two implementations of the R_d variation were used in this study and it was found that the method where E_e was allowed to vary was much more effective at signalling prominence. A fuller elaboration of how to optimally control this parameter may also be the focus of future work beyond this thesis.

The knowledge gained from these experiments is extremely useful in furthering our understanding of how the voice source contributes to prosody, and how important a factor it is when attempting to include expression in speech synthesis. Using only modifications of a small set off voice source parameters, and with minimal f_0 salience, prominence and affective colouring were successfully evoked in synthetic utterances.

8.1.2 Analysis-and-synthesis system and interfaces

The GlórCáil analysis-and-synthesis system was developed as part of this work. This system includes a graphical interface that allows the user to easily and directly control voice source parameter contours of synthetic speech. Manipulating these voice source parameter contours gives control of dimensions of voice quality (tense-lax voice) to the user through an intuitive mouse-controlled environment. The experiments in Chapter 5 have demonstrated that this sort of voice quality control may be used to produce prominence or impart affective colouring on a synthetic utterance. Users also have the ability to listen to the changes they have made to an utterance immediately. This allows the user to compare their synthetic utterance to the original unmodified utterance, or in the case of copy-synthesis, the original audio file. The system also has the capability of transforming the vocal tract transfer filter. This means that the system can be used to transform speaker characteristics to generate new voices and characters that can be used in applications such as computer assisted language learning games. This is especially useful for low resource languages that might not have access to a large amount of speech data. The system was also integrated into a DNN-based SPSS framework, allowing for the generation of almost any utterance.

8.1.3 Manipulation task interface

This work included the development and implementation of a manipulation task interface for carrying out user-driven perception experiments and listening test. Versions of this interface were applied successfully in Experiments 4 and 5 (see Sections 7.1 and 7.2). It is hoped that this style of perception experiment will prove useful in future work that is carried out in this area. Data obtained through user-driven tasks may give a more accurate representation of participants true perception, as giving them a manipulation task may prove more engaging and therefore more enjoyable. These factors would make the task less fatiguing than traditional perception tests that involve listening to countless stimuli.

8.1.4 Identification of future research in voice quality and prosody

The findings of this research have led to the identification of several avenues of future research into voice quality manipulation using a limited set of control parameters.

In terms of perceived focal prominence and position within an utterance, it was shown that voice quality without f_0 salience was capable of producing perceived prominence when the focal syllable was towards the front of an utterance. This was not the case when the focal prominence was towards the end of the utterance. Further study into the interplay of f_0 and other voice source parameters is necessary.

8.2 Future work

Despite the extensive research into expressive speech synthesis, there still remains a plethora of unanswered research questions and goals to achieve. Within the scope of this work there are several possible improvements and avenues of future research that will need to be investigated.

In terms of analysis of the glottal source, the main problem that still exists concerns the inaccuracy of performing automatic inverse filtering on continuous speech. It is hoped that future work will be carried out in order to reduce or negate this problem using a combination of modern methods of machine learning and knowledge-based approaches. It is also hoped to improve the naturalness of the synthetic speech produced by this system by experimenting with the use of different representations of the vocal tract filter. Although the modelling approaches used in this work simplified the process of obtaining an estimate of the voice source and vocal tract transfer function and suited the purposes of this research, using a different spectral model may improve the perceived naturalness of the synthetic speech. Another area of research that will need to be investigated is that of developing a comprehensive predictive prosodic model for voice quality similar to the Fujisaki (Fujisaki and Hirose, 1984) or Tilt (Taylor, 1992) models for f_0.

As mentioned in the previous section, the current versions of the system and manipulation task interfaces rely on timing annotations being carried out using the Praat programme. Ideally this step should be carried out within the system, and this is intended in the future further development of the system. When it has been used more, and user feedback obtained, other refinements to the controls of the interface are intended. A feature that will be added at a later date is that of batch processing – to allow the user to set certain global settings to either make synthetic speech more lax or more tenser, or transform the vocal tract characteristics of multiple utterances in a single step.

The results of Experiment 6 (see Section 7.3) indicated that the parameter manipulations made were not enough to effectively transform a baseline utterance into one that is perceived as angry. This is mostly likely due to factors such as a lack of dynamic voice source parameter manipulation, or indeed, inadequate vocal tract transformations. Future development of the GlórCáil system will attempt to rectify these problems by including dynamic global scaling of parameter contours and improving the manner in which vocal tract transformations are carried out.

These improvements should also address the problem of effectively transforming a male voice to a female or child's voice. This matter will lead to additional future work into investigating the most optimal voices to use as baselines for transformation.

Although the system functions well and the control interface is straightforward to use, being bound to a proprietary environment and programming language such as MATLAB is not ideal. Reprogramming the system in language such as Python or C would be beneficial, not only in terms of a possible increase in computational speed, but also in the open use of a compiled system written in these languages.

It is thought that elements of this work will be helpful in applications developed in the Phonetics and Speech Laboratory in Trinity College Dublin, where the ABAIR project is engaged in the development and delivery of speech technology for Irish. Currently, voices have been produced for the three main dialects of Irish. It is hoped that the GlórCáil analysis-and-synthesis system will allow for the generation of multiple voices/characters from a single synthetic voice, as well as the generation of some expressive speech material. Both of these features are important for applications such as educational interactive games in the Irish language.

References

Airaksinen, M., Bollepalli, B., Juvela, L., Wu, Z., King, S. and Alku, P. (2016) 'GlottDNN — A Full-Band Glottal Vocoder for Statistical Parametric Speech Synthesis', in *Proceedings of INTERSPEECH*, pp. 2473–2477. doi: 10.21437/Interspeech.2016-342.

Airaksinen, M., Raitio, T., Story, B. and Alku, P. (2014) 'Quasi closed phase glottal inverse filtering analysis with weighted linear prediction', *IEEE Transactions on Audio, Speech and Language Processing*, 22(3), pp. 596–607. doi: 10.1109/TASLP.2013.2294585.

Airas, M. (2008) 'TKK Aparat: An environment for voice inverse filtering and parameterization', *Logopedics Phoniatrics Vocology*, 33(1), pp. 49–64. doi: 10.1080/14015430701855333.

Alku, P. (2011) 'Glottal inverse filtering analysis of human voice production — A review of estimation and parameterization methods of the glottal excitation and their applications', *Sadhana*, 36(5), pp. 623–650. doi: 10.1007/s12046-011-0041-5.

Alku, P., Bäckström, T. and Vilkman, E. (2002) 'Normalized amplitude quotient for parametrization of the glottal flow', *The Journal of the Acoustical Society of America*, 112(2), pp. 701–710. doi: 10.1121/1.1490365.

Alku, P., Magi, C., Yrttiaho, S., Bäckström, T. and Story, B. (2009) 'Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering', *The Journal of the Acoustical Society of America*, 125(5), p. 3289. doi: 10.1121/1.3095801.

Alku, P., Pohjalainen, J., Vainio, M., Laukkanen, A.-M. and Story, B. H. (2013) 'Formant frequency estimation of high-pitched vowels using weighted linear prediction', *The Journal of the Acoustical Society of America*, 134(2), pp. 1295–1313. doi: 10.1121/1.4812756.

Alku, P., Tiitinen, H. and Näätänen, R. (1999) 'A method for generating natural sounding speech-stimuli for cognitive brain research', *Clinical Neurophysiology*, 110(8), pp. 1329–1333.

Alku, P. and Vilkman, E. (1994) 'Estimation of the glottal pulseform based on discrete all-pole modeling', *Proc. Int. Conf. on spoken language processing, Yokohama, September 1994*, (September), pp. 1619–1622.

Alku, P. and Vilkman, E. (1996) 'A Comparison of Glottal Voice Source Quantification Parameters in Breathy, Normal and Pressed Phonation of Female and Male Speakers', *Folia Phoniatrica et Logopaedica*, 48(5), pp. 240–254. doi: 10.1159/000266415.

Alku, P., Vilkman, E. and Laukkanen, A. M. (1998) 'Estimation of amplitude features of the glottal flow by inverse filtering speech pressure signals', *Speech Communication*, 24(2), pp. 123–132. doi: 10.1016/S0167-6393(98)00004-1.

Anumanchipalli, G. K., Cheng, Y.-C., Fernandez, J., Huang, X., Mao, Q. and Black, A. W. (2010) 'KlaTTStat: Knowledge-based Statistical Parametric Speech Synthesis', *7th*

ISCA Workshop on Speech Synthesis.

Bailly, G., Bérar, M., Elisei, F. and Odisio, M. (2003) 'Audiovisual Speech Synthesis', *International Journal of Speech Technology*, 6(4), pp. 331–346. doi: 10.1023/A:1025700715107.

Banse, R. and Scherer, K. R. (1996) 'Acoustic profiles in vocal emotion expression.', *Journal of Personality and Social Psychology*, 70(3), pp. 614–636. doi: 10.1037/0022-3514.70.3.614.

Bates, D., Mächler, M., Bolker, B. and Walker, S. (2015) 'Fitting Linear Mixed-Effects Models Using lme4', *Journal of Statistical Software*, 67(1), pp. 1–48. doi: 10.18637/jss.v067.i01.

van den Berg, J. (1958) 'Myoelastic-Aerodynamic Theory of Voice Production', *Journal of Speech and Hearing Research*, 1(3), pp. 227–244. doi: 10.1044/jshr.0103.227.

Birkholz, P. (2005) *3D-Artikulatorische Sprachsynthese*. University of Rostock.

Birkholz, P., Martin, L., Willmes, K., Kröger, B. J. and Neuschaefer-Rube, C. (2015) 'The contribution of phonation type to the perception of vocal emotions in German: An articulatory synthesis study', *The Journal of the Acoustical Society of America*, 137(3), pp. 1503–1512. doi: 10.1121/1.4906836.

Black, A. W. (2006) 'CLUSTERGEN: A Statistical Parametric Synthesizer Using Trajectory Modeling', in *Proceedings of INTERSPEECH*, pp. 1762–1765.

Black, A. W. and Campbell, N. (1995) 'Optimising selection of units from speech databases for concatenative synthesis', in *Proc. Eurospeech '95*. Madrid, Spain, pp. 581–584.

Black, A. W. and Taylor, P. (1997) 'Automatically Clustering Similar Units for Unit Selection in Speech Synthesis', in *Proc. Eurospeech '97*. Rhodes, Greece, pp. 601–604.

Boersma, P. and Weenink, D. (2019) 'Praat: doing phonetics by computer [Computer program]'. Available at: http://www.praat.org/ (Accessed: 18 October 2019).

Broad, D. J. (1979) 'The New Theories of Vocal Fold Vibration', in Lass, N. (ed.) *Speech and language: Advances in basic research and practice*. 2nd edn. New York, US: Academic Press, pp. 203–256. doi: 10.1016/B978-0-12-608602-7.50010-9.

Brookes, D. M. and Chan, D. S. (1994) 'Speaker characteristics from a glottal airflow model using glottal inverse filtering', *Proc. Institute of Acoustics*, 15, pp. 501–508.

Buchanan, C. G., Aylett, M. P. and Braude, D. A. (2018) 'Adding Personality to Neutral Speech Synthesis Voices', in *20th International Conference, SPECOM 2018, Proceedings*. Leipzig, Germany: Springer International Publishing, pp. 49–57. doi: 10.1007/978-3-319-99579-3_6.

Cabral, J. P. (2010) *HMM-based Speech Synthesis Using an Acoustic Glottal Source Model*. PhD Thesis. University of Edinburgh.

Cabral, J. P. and Oliveira, L. C. (2006) 'EmoVoice: a system to generate emotion in speech', in *Proceedings of INTERSPEECH*. Pittsburgh, Pennsylvania, USA, p. paper 1645-Wed2BuP.3.

Cabral, J. P., Renals, S., Yamagishi, J. and Richmond, K. (2011) 'HMM-based speech

synthesiser using the LF-model of the glottal source', *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 4704–4707. doi: 10.1109/ICASSP.2011.5947405.

Cabral, J. P., Richmond, K., Yamagishi, J. and Renals, S. (2014) 'Glottal spectral separation for speech synthesis', in *Proceedings of INTERSPEECH*, pp. 195–208. doi: 10.1109/JSTSP.2014.2307274.

Cahn, J. (1990) 'The generation of affect in synthesised speech', *Journal of American Voice I/O Society*, 8(July), pp. 1–19.

Catford, J. C. (1964) 'Phonation types: the classification of some laryngeal components of speech production', in Abercrombie, D., Fry, D. B., MacCarthy, P. A. D., Scott, M. C., and Trim, J. L. M. (eds) *In Honour of Daniel Jones*. London: Longman, pp. 26–37.

Childers, D. G. and Lee, C. K. (1991) 'Vocal quality factors: analysis, synthesis and perception', *Journal of the Acoustical Society of America*, 90(5), pp. 2394–2410. doi: 10.1121/1.402044.

Chuenwattanapranithi, S., Xu, Y., Thipakorn, B. and Maneewongvatana, S. (2008) 'Encoding Emotions in Speech with the Size Code', *Phonetica*, 65(4), pp. 210–230. doi: 10.1159/000192793.

Compernolle, D. Van and Wambacq, P. (2000) 'A unified view on synchronized overlap-add methods for prosodic modifications of speech .', in *Proceedings of INTERSPEECH*. Beijing, China, pp. 63–66.

Cranen, B. and Boves, L. (1985) 'Pressure measurements during speech production using semiconductor miniature pressure transducers: Impact on models for speech production', *The Journal of the Acoustical Society of America*, 77(4), pp. 1543–1551. doi: 10.1121/1.391997.

Cutler, A., Dahan, D. and van Donselaar, W. (1997) 'Prosody in the Comprehension of Spoken Language: A Literature Review', *Language and Speech*, 40(2), pp. 141–201. doi: 10.1177/002383099704000203.

D'Alessandro, C. and Doval, B. (2003) 'Voice quality modification for emotional speech synthesis', in *EUROSPEECH 2003 - 8th European Conference on Speech Communication and Technology*, pp. 1653–1656.

Dalton, J., Kane, J., Yanushevskaya, I., Ní Chasaide, A. and Gobl, C. (2014) 'GlóRí - the glottal research instrument', in *Speech Prosody 2014*. Dublin, Ireland, pp. 944–948.

Degottex, G. (2010) *Glottal source and vocal-tract separation*. PhD Thesis. IRCAM. Paris, France.

Degottex, G., Lanchantin, P. and Gales, M. (2018) 'A Log Domain Pulse Model for Parametric Speech Synthesis', *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(1), pp. 57–70. doi: 10.1109/TASLP.2017.2761546.

Degottex, G., Lanchantin, P., Roebel, A. and Rodet, X. (2012) 'Mixed source model and its adapted vocal tract filter estimate for voice transformation and synthesis', *Speech Communication*. Elsevier B.V., 55(2), pp. 278–294. doi: 10.1016/j.specom.2012.08.010.

Degottex, G., Roebel, A. and Rodet, X. (2011a) 'Phase Minimization for Glottal Model

Estimation', *IEEE Transactions on Audio, Speech and Language Processing*, 19(5), pp. 1080–1090. doi: 10.1109/TASL.2010.2076806.

Degottex, G., Roebel, A. and Rodet, X. (2011b) 'Pitch transposition and breathiness modification using a glottal source model and its adapted vocal-tract filter', *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. IEEE, 0(1), pp. 5128–5131. doi: 10.1109/ICASSP.2011.5947511.

Deller, J. R., Hansen, J. H. L. and Proakis, J. G. (1999) *Discrete-Time Processing of Speech Signals*. IEEE. doi: 10.1109/9780470544402.

Douglas-Cowie, E., Campbell, N., Cowie, R. and Roach, P. (2003) 'Emotional speech: towards a new generations of databases', *Speech Communication*, 40, pp. 33–60.

Drugman, T. (2013) 'Residual excitation skewness for automatic speech polarity detection', *Signal Processing Letters, IEEE*, 20(4), pp. 387–390. doi: 10.1109/LSP.2013.2249661.

Drugman, T. and Dutoit, T. (2012) 'The Deterministic plus Stochastic model of the residual signal and its applications', *IEEE Transactions on Audio, Speech and Language Processing*, 20(3), pp. 968–981. doi: 10.1109/TASL.2011.2169787.

Drugman, T., Thomas, M., Gudnason, J., Naylor, P. and Dutoit, T. (2012) 'Detection of Glottal Closure Instants From Speech Signals: A Quantitative Review', *IEEE Transactions on Audio, Speech, and Language Processing*. IEEE, 20(3), pp. 994–1006. doi: 10.1109/TASL.2011.2170835.

Drugman, T., Wilfart, G. and Dutoit, T. (2009) 'A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis', in *Proceedings of INTERSPEECH*, pp. 1779–1782.

El-Jaroudi, A. and Makhoul, J. (1991) 'Discrete all-pole modeling', *IEEE Transactions on Signal Processing*, 39(2), pp. 411–423. doi: 10.1109/78.80824.

Ellis, D. (2004) 'Spectral Warping of LPC resonance models'. Available at: https://www.ee.columbia.edu/~dpwe/resources/matlab/polewarp/ (Accessed: 12 June 2019).

Epstein, M. (2003) 'Voice quality and prosody in English', in *XVth International Congress of Phonetic Sciences*. Barcelona, Spain.

Erro, D., Sainz, I., Navas, E. and Hernáez, I. (2011) 'Improved HNM-based vocoder for statistical synthesizers', in *Proceedings of INTERSPEECH*. Florence, Italy, pp. 1809–1812.

Erro, D., Sainz, I., Navas, E. and Hernáez, I. (2014) 'Harmonics plus noise model based vocoder for statistical parametric speech synthesis', *IEEE Journal on Selected Topics in Signal Processing*, 8(2), pp. 184–194. doi: 10.1109/JSTSP.2013.2283471.

Fan, Y., Qian, Y., Xie, F. and Soong, F. K. (2014) 'TTS synthesis with bidirectional LSTM based Recurrent Neural Networks', in *Proceedings of INTERSPEECH*, pp. 1964–1968.

Fant, G. (1956) 'On the predictability of formant levels and spectrum envelopes from formant frequencies', in Halle, M., Lunt, H., and Maclean, H. (eds) *For Roman Jakobson*. The Hague: Mouton, pp. 109–120.

Fant, G. (1960) *Acoustic Theory of Speech Production*. 2nd edn, *The Hague*. 2nd edn. The

Hague: Mouton de Gruyter.

Fant, G. (1993) 'Some problems in voice source analysis', *Speech Communication*, 13(1–2), pp. 7–22. doi: 10.1016/0167-6393(93)90055-P.

Fant, G. (1995) *The LF-model revisited . Transformations and frequency domain analysis*, *STL-QPSR*. Stockholm, Sweden: Royal Institute of Technology.

Fant, G. (1997) 'The voice source in connected speech', *Speech Communication*, 22(2–3), pp. 125–139. doi: 10.1016/S0167-6393(97)00017-4.

Fant, G. and Kruckenberg, A. (1994) *Notes on stress and word accent in Swedish*, *STL-QPSR*. Stockholm, Sweden: Royal Institute of Technology.

Fant, G., Kruckenberg, A., Liljencrants, J. and Båvegård, M. (1994) 'Voice source parameters in continuous speech, transformation of LF-parameters', in *ICSLP'94*. Yokohama, Japan, pp. 1451–1454.

Fant, G., Liljencrants, J. and Lin, Q. (1985) *A four-parameter model of glottal flow*, *STL-QPSR*. Stockholm, Sweden: Royal Institute of Technology. doi: 10.1016/0167-6393(89)90001-0.

Fitch, W. T. and Giedd, J. (1999) 'Morphology and development of the human vocal tract: A study using magnetic resonance imaging', *The Journal of the Acoustical Society of America*, 106(3), pp. 1511–1522. doi: 10.1121/1.427148.

Flanagan, J. L. (1972) *Speech Analysis Synthesis and Perception*. Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/978-3-662-01562-9.

Fónagy, I. (1962) 'Mimik auf glottaler Ebene', *Phonetica*, 8(4), pp. 209–219. doi: 10.1159/000258130.

Fujisaki, H. and Hirose, K. (1984) 'Analysis of voice fundamental frequency contours for declarative sentences of Japanese.', *Journal of the Acoustical Society of Japan (E)*, 5(4), pp. 233–242. doi: 10.1250/ast.5.233.

Fujisaki, H. and Ljungqvist, M. (1986) 'Proposal and evaluation of models for the glottal source waveform', in *ICASSP '86. IEEE International Conference on Acoustics, Speech, and Signal Processing*. Institute of Electrical and Electronics Engineers, pp. 1605–1608. doi: 10.1109/ICASSP.1986.1169239.

Gay, T. (1978) 'Physiological and Acoustic Correlates of Perceived Stress', *Language and Speech*, 21(4), pp. 347–353. doi: 10.1177/002383097802100409.

Gobl, C. (1988) *Voice source dynamics in connected speech*, *STL-QPSR*. Stockholm, Sweden: Royal Institute of Technology.

Gobl, C. (1989) *A preliminary study of acoustic voice quality correlates*, *STL-QPSR*. Stockholm, Sweden: Royal Institute of Technology.

Gobl, C. (2003) *The Voice Source in Speech Communication.* PhD Thesis. KTH, Department of Speech, Music and Hearing. Stockholm, Sweden.

Gobl, C. (2006) 'Modelling aspiration noise during phonation using the LF voice source model', in *Proceedings of INTERSPEECH*. Pittsburg, PA, USA, pp. 965–968.

Gobl, C. (2017) 'Reshaping the Transformed LF Model: Generating the Glottal Source from the Waveshape Parameter Rd', in *Proceedings of INTERSPEECH*. ISCA: ISCA, pp. 3008–3012. doi: 10.21437/Interspeech.2017-1140.

Gobl, C., Bennett, E. and Ní Chasaide, A. (2002) 'Expressive synthesis: how crucial is voice quality?', in *IEEE Workshop on Speech Synthesis*. Santa Monica, California, USA, pp. 1–4.

Gobl, C. and Ní Chasaide, A. (1999a) 'Techniques for analysing the voice source', in Hardcastle, W. J. and Hewlett, N. (eds) *Coarticulation: Theory, Data and Techniques*. Cambridge: Cambridge University Press, pp. 300–321.

Gobl, C. and Ní Chasaide, A. (1999b) 'Voice source variation in the vowel as a function of consonantal context', in Hardcastle, W. J. and Hewlett, N. (eds) *Coarticulation: Theory, Data and Techniques*. Cambridge: Cambridge University Press, pp. 122–143.

Gobl, C. and Ní Chasaide, A. (2002) 'Dynamics of the glottal source signal', in Keller, E., Bailly, G., Monaghan, A., Terken, J., and Huckvale, M. (eds) *Improvements in Speech Synthesis*. John Wiley & Sons, Ltd., pp. 273–283.

Gobl, C. and Ní Chasaide, A. (2003) 'The role of voice quality in communicating emotion, mood and attitude', *Speech Communication*, 40(1–2), pp. 189–212. doi: 10.1016/S0167-6393(02)00082-1.

Gobl, C. and Ní Chasaide, A. (2010) 'Voice Source Variation and Its Communicative Functions', in Hardcastle, W. J., Laver, J., and Gibbon, F. E. (eds) *The Handbook of Phonetic Sciences*. 2nd edn. Oxford, UK: Willey-Blackwell Publishing Ltd., pp. 378–423. doi: 10.1002/9781444317251.ch11.

Gobl, C., Yanushevskaya, I. and Ní Chasaide, A. (2015) 'The relationship between voice source parameters and the maxima dispersion quotient (MDQ)', in *Proceedings of INTERSPEECH*. Dresden, Germany, pp. 2337–2341.

Goodfellow, I., Bengio, Y. and Courville, A. (2016) *Deep Learning*. MIT Press.

Gordon, M. and Ladefoged, P. (2001) 'Phonation types: a cross-linguistic overview', *Journal of Phonetics*, 29(4), pp. 383–406. doi: 10.1006/jpho.2001.0147.

Graves, A. and Schmidhuber, J. (2005) 'Framewise phoneme classification with bidirectional LSTM and other neural network architectures', *Neural Networks*, 18(5–6), pp. 602–610. doi: 10.1016/j.neunet.2005.06.042.

Grichkovtsova, I., Morel, M. and Lacheret, A. (2012) 'The role of voice quality and prosodic contour in affective speech perception', *Speech Communication*, 54(3), pp. 414–429. doi: 10.1016/j.specom.2011.10.005.

Guerchi, D. and Mermelstein, P. (2000) 'Low-rate quantization of spectral information in a 4 kb/s pitch-synchronous CELP coder', in *2000 IEEE Workshop on Speech Coding. Proceedings. Meeting the Challenges of the New Millennium (Cat. No.00EX421)*. IEEE, pp. 111–113. doi: 10.1109/SCFT.2000.878416.

Gussenhoven, C., Repp, B. H., Rietveld, A., Rump, H. H. and Terken, J. (1997) 'The perceptual prominence of fundamental frequency peaks', *Journal of the Acoustical Society of America*, 102(5), pp. 3009–3022.

Hacki, T. (1989) 'Klassifizierung von Glottisdysfunktionen mit Hilfe der Elektroglottographie', *Folia Phoniatrica et Logopaedica*, 41(1), pp. 43–48.

Hamon, C., Mouline, E. and Charpentier, F. (2003) 'A diphone synthesis system based on time-domain prosodic modifications of speech', in *International Conference on Acoustics, Speech, and Signal Processing*. IEEE, pp. 238–241. doi:

10.1109/ICASSP.1989.266409.

Hardcastle, W. J. (1976) *Physiology of Speech Production: An Introduction for Speech Scientists, Language*. London, UK: Academic Press.

Härmä, A., Karjalainen, M., Avioja, L., Välimäki, V., Laine, U. and Huopaniemi, J. (2000) 'Frequency-Warped Signal Processing for Audio Applications', *Journal of the Audio Engineering Society*, 48, pp. 1011–1031.

Hashimoto, K., Oura, K., Nankaku, Y. and Tokuda, K. (2015) 'The effect of neural networks in statistical parametric speech synthesis', in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 4455–4459. doi: 10.1109/ICASSP.2015.7178813.

Hermes, D. J. (2006) 'Stylization of pitch contours', in Sudhoff, S., Lenertová, D., Meyer, R., Pappert, S., Augurzky, P., Mleinek, I., Richter, N., and Schließer, J. (eds) *Methods in Empirical Prosody Research*. Berlin, Germany: Walter de Gruyter, pp. 29–61.

Hochreiter, S. and Schmidhuber, J. (1997) 'Long Short-Term Memory', *Neural Computation*, 9(8), pp. 1735–1780. doi: 10.1162/neco.1997.9.8.1735.

Hu, Q., Wu, Z., Richmond, K., Yamagishi, J., Stylianou, Y. and Maia, R. (2015) 'Fusion of multiple parameterisations for DNN-based sinusoidal speech synthesis with multi-task learning', in *Proceedings of INTERSPEECH*, pp. 854–858.

Huber, S. and Roebel, A. (2015) 'On glottal source shape parameter transformation using a novel deterministic and stochastic speech analysis and synthesis system', in *Proceedings of INTERSPEECH*, pp. 289–293.

Huber, S., Roebel, A. and Degottex, G. (2012) 'Glottal source shape parameter estimation using phase minimization variants', in *International Conference on Spoken Language Processing*. Portland, United States, pp. 1644–1647.

Huffman, M. K. (1987) 'Measures of phonation type in Hmong', *Journal of the Acoustical Society of America*, 81(2), pp. 495–504. doi: 10.1121/1.394915.

Hunt, A. J. and Black, A. W. (1996) 'Unit Selection in a Concatenative Speech Synthesis System using a Large Speech Database', in *Proc. ICASSP '96*. Atlanta, GA, pp. 373–376.

Hunt, M. J., Zwierzynski, D. A. and Carr, R. C. (1989) 'lssues in High Quality LPC Analysis and Synthesis', in *First European Conference on Speech Communication and Technology EUROSPEECH*. Paris, France, pp. 2348–2351.

Iida, A., Campbell, N., Higuchi, F. and Yasumura, M. (2003) 'A corpus-based speech synthesis system with emotion', *Speech Communication*, 40(1–2), pp. 161–187. doi: http://dx.doi.org/10.1016/S0167-6393(02)00081-X.

Iseli, M., Shue, Y.-L., Epstein, M. A., Keating, P., Kreiman, J. and Alwan, A. (2006) 'Voice source correlates of prosodic features in American English', in *Proceedings of INTERSPEECH*. Pittsburgh, Pennsylvania, USA, p. paper 1933-Thu1A3O.1.

Ishizaka, K. and Flanagan, J. L. (1972) 'Acoustic Properties of a Two-Mass Model of the Vocal Cords', *The Journal of the Acoustical Society of America*, 51(1A), pp. 91–91. doi: 10.1121/1.1981696.

Ishizaka, K. and Matsudaira, M. (1968) 'What makes the vocal cords vibrate?', in *the 6th International Congress on Acoustics*. Tokyo, Japan, pp. B9–B12.

Itakura, F. (1975a) ‘Line spectrum representation of linear predictor coefficients of speech signals’, *The Journal of the Acoustical Society of America*, 57(S1), pp. S35–S35. doi: 10.1121/1.1995189.

Itakura, F. (1975b) ‘Minimum prediction residual principle applied to speech recognition’, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23(1), pp. 67–72. doi: 10.1109/TASSP.1975.1162641.

Jianfen, C. and Maddieson, I. (1992) ‘An exploration of phonation types in Wu dialects of Chinese’, *Journal of Phonetics*, 20(September 2016), pp. 77–92.

Jusczyk, P. W., Hirsh-Pasek, K., Kemler Nelson, D. G., Kennedy, L. J., Woodward, A. and Piwoz, J. (1992) ‘Perception of acoustic correlates of major phrasal units by young infants’, *Cognitive Psychology*, 24(2), pp. 252–293. doi: 10.1016/0010-0285(92)90009-Q.

Kalman, R. E. (1960) ‘A New Approach to Linear Filtering and Prediction Problems’, *Journal of Basic Engineering*, 82(1), p. 35. doi: 10.1115/1.3662552.

Kane, J. and Gobl, C. (2013) ‘Automating manual user strategies for precise voice source analysis’, *Speech Communication*, 55(3), pp. 397–414. doi: 10.1016/j.specom.2012.12.004.

Karlsson, I. (1990) ‘Voice source dynamics for female speakers’, in *Proc. Internat. Conf. Spoken Language Proc.*

Kawahara, H., Estill, J. and Fujimura, O. (2001) ‘Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT’, *Proceedings of MAVEBA*, pp. 59–64.

Kawahara, H., Masuda-Katsuse, I. and De Cheveigné, A. (1999) ‘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(3), pp. 187–207. doi: 10.1016/S0167-6393(98)00085-5.

von Kempelen, W. (1791) ‘Mechanismus der menschlichen Sprache nebst Beschreibung einer sprechenden Maschine and Le Mchanisme de la parole, suivi de la description d’une machine parlante’. Vienna, Austria: J.V. Degen.

Kent, R. D. and Read, C. (1992) *The Acoustic Analysis of Speech*. Singular Publishing Group.

Kim, S. J. and Hahn, M. (2007) ‘Two-band excitation for HMM-based speech synthesis’, *IEICE Transactions on Information and Systems*, E90-D(1), pp. 378–381. doi: 10.1093/ietisy/e90-1.1.378.

Kingma, D. P. and Ba, J. (2014) ‘Adam: A Method for Stochastic Optimization’, pp. 1–15.

Klatt, D. H. (1980) ‘Software for a cascade/parallel formant synthesizer’, *Journal of the Acoustical Society of America*, 67, pp. 971–995.

Klatt, D. H. and Klatt, L. C. (1990) *Analysis, synthesis, and perception of voice quality variations among female and male talkers, The Journal of the Acoustical Society of America*. doi: 10.1121/1.398894.

Knight, R. A. (2008) ‘The shape of nuclear falls and their effect on the perception of pitch

and prominence: peaks vs. plateaux', *Language & Speech*, 51(3), pp. 223–244.

Kreiman, J., Antoñanzas-Barroso, N. and Gerratt, B. R. (2010) 'Integrated software for analysis and synthesis of voice quality', *Behavior Research Methods*, 42(4), pp. 1030–1041. doi: 10.3758/BRM.42.4.1030.

Kreiman, J., Garellek, M., Chen, G., Alwan, A. and Gerratt, B. R. (2015) 'Perceptual evaluation of voice source models', *The Journal of the Acoustical Society of America*, 138(1), pp. 1–10. doi: 10.1121/1.4922174.

Kreiman, J., Gerratt, B. R. and Antoñanzas-Barroso, N. (2006) *Analysis and synthesis of pathological voice quality*. The Regents of the University of California.

Kreiman, J., Gerratt, B. R. and Ito, M. (2007) 'When and why listeners disagree in voice quality assessment tasks', *Journal of the Acoustical Society of America*, 122(4), pp. 2354–2364. doi: http://dx.doi.org/10.1121/1.2770547.

Kuznetsova, A., Brockhoff, P. B. and Christensen, R. H. B. (2017) 'lmerTest Package: Tests in Linear Mixed Effects Models', *Journal of Statistical Software*, 82(13), pp. 1–26. doi: doi:10.18637/jss.v082.i13.

Ladd, D. R. (1980) *The Structure of Intonational Meaning: Evidence from English*. Indiana University Press (Books on demand).

Ladd, D. R. and Morton, R. (1997) 'The perception of intonational emphasis: continuous or categorical?', *Journal of Phonetics*, 25(3), pp. 313–342. doi: 10.1006/jpho.1997.0046.

Ladd, D. R., Silverman, K. E. A., Tolkmitt, F., Bergmann, G. and Scherer, K. R. (1985) 'Evidence for the independent function of intonation contour type, voice quality, and F 0 range in signaling speaker affect', *The Journal of the Acoustical Society of America*, 78(2), pp. 435–444. doi: 10.1121/1.392466.

Ladefoged, P. (1972) *Three areas of experimental phonetics: stress and respiratory activity, the nature of vowel quality, units in the perception and production of speech*. Oxford University Press (Language and language learning).

Ladefoged, P. and Johnson, K. (2010) *A Course in Phonetics*. 6th edn. Wadsworth, Cengage Learning.

Laver, J. (1980) *The Phonetic Description of Voice Quality, Cambridge studies in linguistics*. Cambridge: Cambridge University Press.

Laver, J. and Trudgill, P. (1979) 'Phonetic and linguistic markers in speech', *Social Markers in Speech*, pp. 1–26.

Li, H., Scaife, R. and Brien, D. O. (2011) 'LF model based glottal source parameter estimation by extended Kalman filtering', *Source*.

Lieberman, P. and Blumstein, S. E. (1988) *Speech Physiology, Speech Perception, and Acoustic Phonetics*. Cambridge University Press (Cambridge Studies in Speech Sc).

Limesurvey GmbH. (2012) 'LimeSurvey: An Open Source survey tool'. Hamburg, Germany. Available at: http://www.limesurvey.org.

Lüdecke, D. (2018) 'sjPlot: Data Visualization for Statistics in Social Science'. doi: doi: 10.5281/zenodo.1308157.

Ma, C., Kamp, Y. and Willems, L. F. (1993) 'Robust signal selection for linear prediction

analysis of voiced speech', *Speech Communication*, 12(1), pp. 69–81. doi: 10.1016/0167-6393(93)90019-H.

Maia, R., Toda, T., Zen, H., Nankaku, Y. and Tokuda, K. (2007a) 'A trainable excitation model for HMM-based speech synthesis', in *Proceedings of INTERSPEECH*, pp. 1125–1128.

Maia, R., Toda, T., Zen, H., Nankaku, Y. and Tokuda, K. (2007b) 'An Excitation Model for HMM-Based Speech Synthesis Based on Residual Modeling', *6th ISCA Workshop on Speech Synthesis*, pp. 131–136.

Maia, R., Zen, H., Knill, K., Gales, M. J. F. and Buchholz, S. (2011) 'Multipulse sequences for residual signal modeling', in *Proceedings of INTERSPEECH*, pp. 1833–1836.

Mehri, S., Kumar, K., Gulrajani, I., Kumar, R., Jain, S., Sotelo, J., Courville, A. and Bengio, Y. (2016) 'SampleRNN: An Unconditional End-to-End Neural Audio Generation Model', pp. 1–11.

Mokhtari, P. and Ando, H. (2017) 'Iterative optimal preemphasis for improved glottal-flow estimation by iterative adaptive inverse filtering', in *Proceedings of INTERSPEECH*, pp. 1044–1048. doi: 10.21437/Interspeech.2017-79.

Mokhtari, P., Story, B., Alku, P. and Ando, H. (2018) 'Estimation of the glottal flow from speech pressure signals : Evaluation of three variants of iterative adaptive inverse filtering using computational physical modelling of voice production', *Speech Communication*. Elsevier B.V., 104(August), pp. 24–38. doi: 10.1016/j.specom.2018.09.005.

Monsen, R. B. and Engebretson, A. M. (1977) 'Study of variations in the male and female glottal wave', *The Journal of the Acoustical Society of America*, 62(4), pp. 981–993. doi: 10.1121/1.381593.

Monsen, R. B., Engebretson, A. M. and Vemula, N. R. (1978) 'Indirect assessment of the contribution of subglottal air pressure and vocal-fold tension to changes of fundamental frequency in English', *Journal of the Acoustical Society of America*, 64(1), pp. 65–80.

Morise, M. (2015) 'CheapTrick, a spectral envelope estimator for high-quality speech synthesis', *Speech Communication*. Elsevier B.V., 67, pp. 1–7. doi: 10.1016/j.specom.2014.09.003.

Morise, M. (2016) 'D4C, a band-aperiodicity estimator for high-quality speech synthesis', *Speech Communication*. Elsevier B.V., 84, pp. 57–65. doi: 10.1016/j.specom.2016.09.001.

Morise, M. (2017) 'Harvest: A High-Performance Fundamental Frequency Estimator from Speech Signals', in *Proceedings of INTERSPEECH*. Stockholm, Sweden: ISCA, pp. 2321–2325. doi: 10.21437/Interspeech.2017-68.

Morise, M., Yokomori, F. and Ozawa, K. (2016) 'WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications', *IEICE Transactions on Information and Systems*, E99.D(7), pp. 1877–1884. doi: 10.1587/transinf.2015EDP7457.

Moulines, E. and Charpentier, F. (1990) 'Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones', *Speech Communication*, 9, pp. 453–468.

Moulines, E. and Verhelst, W. (1995) 'Time-Domain and Frequency-Domain Techniques for Prosodic Modification of Speech', in Kleijn, W & K. Paliwal, K. (ed.) *Speech Coding and Synthesis*. Netherlands: Elsevier Science B.V., pp. 519–555.

Murphy, A., Yanushevskaya, I., Ní Chasaide, A. and Gobl, C. (2017) 'Rd as a control parameter to explore affective correlates of the tense-lax continuum', in *Proceedings of INTERSPEECH*. Stockholm, Sweden, pp. 3916–3920. doi: 10.21437/Interspeech.2017-1448.

Murphy, A., Yanushevskaya, I., Ní Chasaide, A. and Gobl, C. (2018) 'Voice Source Contribution to Prominence Perception: Rd Implementation', in *Proceedings of INTERSPEECH*. Hyderabad, India, pp. 217–221. doi: 10.21437/Interspeech.2018-2352.

Murphy, A., Yanushevskaya, I., Ní Chasaide, A. and Gobl, C. (2019) 'The Role of Voice Quality in the Perception of Prominence in Synthetic Speech', in *Proceedings of INTERSPEECH*. Gratz, Austria, pp. 2543–2547. doi: 10.21437/Interspeech.2019-2761.

Murray, I. R. and Arnott, J. L. (1993) 'Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion', *Journal of the Acoustical Society of America*, 93(2), pp. 1907–1108. doi: 10.1121/1.405558.

Nakagawa, S., Johnson, P. C. D. and Schielzeth, H. (2017) 'The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded', *Journal of The Royal Society Interface*, 14(134), p. 20170213. doi: 10.1098/rsif.2017.0213.

Nelder, J. A. and Mead, R. (1965) 'A Simplex Method for Function Minimization', *The Computer Journal*, 7(4), pp. 308–313. doi: 10.1093/comjnl/7.4.308.

Ney, H. (1983) 'Dynamic programming algorithm for optimal estimation of speech parameter contours', *{IEEE} Trans Syst, Man and Cybernetics*, 13(3), pp. 208–214.

Ní Chasaide, A., Gobl, C. and Monahan, P. (1992) 'A technique for analysing voice quality in pathological and normal speech', *Journal of Clinical Speech and Language Studies*, 2, pp. 1–16.

Ní Chasaide, A., Ní Chiaráin, N., Wendler, C., Berthelsen, H., Kelly, A., Gilmartin, E., Ní Dhonnchadha, E. and Gobl, C. (2011a) 'Towards Personalised , Synthesis-based Content in Irish (Gaelic) Language Education', *SLaTE*, (January 2011), pp. 29–33.

Ní Chasaide, A., Ní Chiaráin, N., Wendler, C., Berthelsen, H., Murphy, A. and Gobl, C. (2017) 'The ABAIR Initiative: Bringing Spoken Irish into the Digital Space', in *Proceedings of INTERSPEECH*. Stockholm, Sweden: ISCA, pp. 2113–2117. doi: 10.21437/Interspeech.2017-1407.

Ní Chasaide, A., Yanushevskaya, I. and Gobl, C. (2011b) 'Voice source dynamics in intonation', in *XVIIth International Congress of Phonetic Sciences*. Hong Kong, China, pp. 1470–1473.

Ní Chasaide, A., Yanushevskaya, I. and Gobl, C. (2015) 'Prosody of voice: declination, sentence mode and interaction with prominence', in *XVIIIth International Congress of Phonetic Sciences*. Glasgow, UK, pp. 1–5.

Ní Chasaide, A., Yanushevskaya, I., Kane, J. and Gobl, C. (2013) 'The voice prominence hypothesis: The interplay of f0 and voice source features in accentuation', in

Proceedings of INTERSPEECH. Lyon, France, pp. 3527–3531.

Ní Chiaráin, N. and Ní Chasaide, A. (2016a) 'Chatbot technology with synthetic voices in the acquisition of an endangered language: Motivation, development and evaluation of a platform for Irish', in Calzolari, N. et al. (eds) *Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portoro, Slovenia, pp. 3429–3435.

Ní Chiaráin, N. and Ní Chasaide, A. (2016b) 'The Digichaint interactive game as a virtual learning environment for Irish', in Papadima-Sophocleous, S., Bradley, L., and Thouësny, S. (eds) *CALL communities and culture - short papers from EUROCALL 2016*. Limassol, Cyprus: Research-publishing.net, pp. 330–336.

Ohala, J. J. (1984) 'An ethological perspective on common cross-language utilization of f0 of voice', *Phonetica*, 41, pp. 1–16.

Oord, A. van den *et al.* (2016) 'WaveNet: A Generative Model for Raw Audio', pp. 1–15. doi: 10.1109/ICASSP.2009.4960364.

Paszke, A. *et al.* (2017) 'Automatic differentiation in PyTorch', in *Neural Information Processing Systems (NIPS) Autodiff Workshop*.

Patel, S., Scherer, K. R., Björkner, E. and Sundberg, J. (2011) 'Mapping emotions into acoustic space: The role of voice production', *Biological Psychology*, 87(1), pp. 93–98. doi: http://dx.doi.org/10.1016/j.biopsycho.2011.02.010.

Perrotin, O. and McLoughlin, I. (2019) 'A Spectral Glottal Flow Model for Source-filter Separation of Speech', in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 7160–7164. doi: 10.1109/ICASSP.2019.8682625.

Perrotin, O. and McLoughlin, I. V. (2017) 'On the Use of a Spectral Glottal Model for the Source-filter Separation of Speech', pp. 1–8.

Peterson, G. E. and Barney, H. L. (1952) 'Control Methods Used in a Study of the Vowels', *The Journal of the Acoustical Society of America*, 24(2), pp. 175–184. doi: 10.1121/1.1906875.

Pierrehumbert, J. B. (1989) *A preliminary study of the consequences of intonation for the voice source*, *STL-QPSR*. Stockholm, Sweden: Royal Institute of Technology.

Plumpe, M. D., Quatieri, T. F. and Reynolds, D. A. (1999) 'Modeling of the glottal flow derivative waveform with application to speaker identification', *IEEE Transactions on Speech and Audio Processing*, 7(5), pp. 569–585. doi: 10.1109/89.784109.

Qian, Y., Fan, Y., Hu, W. and Soong, F. K. (2014) 'On the training aspects of Deep Neural Network (DNN) for parametric TTS synthesis', in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 3829–3833. doi: 10.1109/ICASSP.2014.6854318.

Rabiner, L. R. and Schafer, R. W. (1978) *Digital processing of speech signals*, *London*. Prentice Hall: Prentice-Hall (Prentice-Hall signal processing series).

Raitio, T. (2008) *Hidden Markov Model Based Finnish Text-to-Speech System Utilizing Glottal Inverse Filtering*. MSc Thesis. Helsinki University of Technology. Helsinki, Finland.

Raitio, T., Kane, J., Drugman, T. and Gobl, C. (2013) 'HMM-based synthesis of creaky voice', in *Proceedings of INTERSPEECH*, pp. 2316–2320.

Raitio, T., Suni, A., Yamagishi, J., Pulakka, H., Nurminen, J., Vainio, M. and Alku, P. (2011) 'HMM-based speech synthesis utilizing glottal inverse filtering', *IEEE Transactions on Audio, Speech and Language Processing*, 19(1), pp. 153–165. doi: 10.1109/TASL.2010.2045239.

Reetz, H. and Jongman, A. (2008) *Phonetics: Transcription, Production, Acoustics and Perception*. Wiley Blackwell.

Richard, G. and D'Alessandro, C. (1996) 'Analysis/synthesis and modification of the speech aperiodic component', *Speech Communication*, 19(3), pp. 221–244. doi: 10.1016/0167-6393(96)00038-6.

Rosenberg, A. E. (1971) 'Effect of Glottal Pulse Shape on the Quality of Natural Vowels', *The Journal of the Acoustical Society of America*, 49(2B), pp. 583–590. doi: 10.1121/1.1912389.

Rubin, P., Baer, T. and Mermelstein, P. (1981) 'An articulatory synthesizer for perceptual research', *The Journal of the Acoustical Society of America*, 70(2), pp. 321–328. doi: 10.1121/1.386780.

Ryan, C., Ní Chasaide, A., Gobl, C., Chasaide, A. N. and Gobl, C. (2003) 'Voice Quality Variation and the Perception of Affect : Continuous or Categorical?', *XVth International Congress of Phonetic Sciences*. Barcelona, Spain, pp. 2409–2412.

Scherer, K. R. (1986) 'Vocal affect expression: a review and a model for future research', *Psychological Bulletin*, 99(2), pp. 143–165.

Scherer, K. R. (1996) 'Adding the affective dimension: a new look in speech analysis and synthesis', in *4th International Conference on Spoken Language Processing*. Philadelphia, USA, pp. 1811–1814.

Scherer, K. R. (2003) 'Vocal communication of emotion: a review of research paradigms', *Speech Communication*, 40(1–2), pp. 227–256. doi: 10.1016/S0167-6393(02)00084-5.

Schnell, K. and Lacroix, A. (2007) 'Time-varying pre-emphasis and inverse filtering of speech', in *Proceedings of INTERSPEECH*, pp. 761–764.

Schuster, M. and Paliwal, K. K. (1997) 'Bidirectional recurrent neural networks', *IEEE Transactions on Signal Processing*, 45(11), pp. 2673–2681. doi: 10.1109/78.650093.

Shen, J. *et al.* (2018) 'Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions', in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 4779–4783. doi: 10.1109/ICASSP.2018.8461368.

Shue, Y.-L. and Alwan, A. (2010) 'A new voice source model based on high-speed imaging and its application to voice source estimation', in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, pp. 5134–5137. doi: 10.1109/ICASSP.2010.5495030.

Smith, J. O. (2007) *Mathematics of the Discrete Fourier Transform*. BookSurge Publishing (W3K Publishing).

Sondhi, M. M. (1975) 'Measurement of the glottal waveform', *Journal of the Acoustical Society of America*, 57(1), pp. 228–232.

Soong, F. and Juang, B. (1984) 'Line spectrum pair (LSP) and speech data compression', in *ICASSP '84. IEEE International Conference on Acoustics, Speech, and Signal*

Processing. Institute of Electrical and Electronics Engineers, pp. 37–40. doi: 10.1109/ICASSP.1984.1172448.

Sorin, A., Shechtman, S. and Rendel, A. (2017) 'Semi Parametric Concatenative TTS with Instant Voice Modification Capabilities', in *Proceedings of INTERSPEECH*. ISCA: ISCA, pp. 1373–1377. doi: 10.21437/Interspeech.2017-1202.

Sotelo, J., Mehri, S., Kumar, K. and Kastner, K. (2017) 'Char2Wav: End - to - End Speech Synthesis', in *5th International Conference on Learning Representations (ICLR)*, pp. 1–6.

Stevens, K. N. (1998) *Acoustic Phonetics*. Cambridge, Massachusetts: The MIT Press.

Story, B. H. (2002) 'An overview of the physiology, physics and modeling of the sound source for vowels', *Acoustical Science and Technology*, 23(4), pp. 195–206. doi: 10.1250/ast.23.195.

Streiner, D. L. and Norman, G. R. (2008) *Health Measurement Scales*. 4th edn. Oxford, UK: Oxford University Press.

Strik, H. (1998) 'Automatic parametrization of differentiated glottal flow: Comparing methods by means of synthetic flow pulses', *The Journal of the Acoustical Society of America*, 103(5), pp. 2659–2669. doi: 10.1121/1.422786.

Strik, H., Cranen, B. and Boves, L. (1993) 'Fitting a LF-model to inverse filter signals.', *Eurospeech*, pp. 1–6.

Strube, H. W. (1974) 'Determination of the instant of glottal closure from the speech wave', *The Journal of the Acoustical Society of America*, 56(5), pp. 1625–1629. doi: 10.1121/1.1903487.

Stylianou, Y. (2001) 'Applying the harmonic plus noise model in concatenative speech synthesis', *IEEE Transactions on Speech and Audio Processing*, 9(1), pp. 21–29. doi: 10.1109/89.890068.

Sundberg, J., Patel, S., Bjorkner, E. and Scherer, K. R. (2011) 'Interdependencies among Voice Source Parameters in Emotional Speech', *IEEE Transactions on Affective Computing*, 2(3), pp. 162–174. doi: 10.1109/T-AFFC.2011.14.

Takaki, S., Kim, S. and Yamagishi, J. (2016) 'Speaker Adaptation of Various Components in Deep Neural Network based Speech Synthesis', in *9th ISCA Speech Synthesis Workshop*, pp. 153–159. doi: 10.21437/SSW.2016-25.

Talkin, D. (2015) 'REAPER: Robust Epoch And Pitch EstimatoR [Online]'. Available at: https://github.com/google/REAPER (Accessed: 17 October 2018).

Tartter, V. C. (1980) 'Happy talk: Perceptual and acoustic effects of smiling on speech', *Perception & Psychophysics*, 27(1), pp. 24–27. doi: 10.3758/BF03199901.

Taylor, P. (2009) *Text-to-Speech Synthesis*, *Text-to-Speech Synthesis*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511816338.

Taylor, P. A. (1992) *A Phonetic Model of English Intonation*. PhD Thesis. University of Edinburgh.

Team, R. C. (2019) 'R: A language and environment for statistical computing'. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.r-project.org/.

Terken, J. (1991) 'Fundamental frequency and perceived prominence of accented syllables', *Journal of the Acoustical Society of America*, 89(4), pp. 1768–1776.

Terken, J. (1994) 'Fundamental frequency and perceived prominence of accented syllables. II. Nonfinal accents', *Journal of the Acoustical Society of America*. Institute for Perception Research, Eindhoven, The Netherlands., 95(6), pp. 3662–3665.

Terken, J. and Hermes, D. (2000) 'The perception of prosodic prominence', in Horne, M. (ed.) *Prosody: Theory and Experiment. Studies presented to Gösta Bruce*. Dordrecht: Kluwer, pp. 89–127.

The Mathworks Inc. (2018) 'MATLAB'. Natick, Massachusetts: The MathWorks Inc.

Thomas, M. R. P., Gudnason, J. and Naylor, P. A. (2012) 'Estimation of Glottal Closing and Opening Instants in Voiced Speech Using the YAGA Algorithm', *IEEE Transactions on Audio, Speech, and Language Processing*. IEEE, 20(1), pp. 82–91. doi: 10.1109/TASL.2011.2157684.

Timcke, R., von Leden, H. and Moore, P. (1958) 'Laryngeal Vibrations: Measurements of the Glottic Wave: Part I. The Normal Vibratory Cycle', *Archives of Otolaryngology - Head and Neck Surgery*, 68(1), pp. 1–19. doi: 10.1001/archotol.1958.00730020005001.

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T. and Kitamura, T. (2000) 'Speech parameter generation algorithms for HMM-based speech synthesis', *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. Ieee, 3(3), pp. 1315–1318. doi: 10.1109/ICASSP.2000.861820.

Tokuda, K., Zen, H. and Black, A. W. (2002) 'An HMM-based speech synthesis system applied to English', *Proceedings of 2002 IEEE Workshop on Speech Synthesis*, pp. 227–230. doi: 10.1109/WSS.2002.1224415.

Tsuzuki, R., Zen, H., Tokuda, K., Kitamura, T., Bulut, M. and Narayanan, S. (2004) 'Constructing emotional speech synthesisers with limited speech database', in *Proceedings of INTERSPEECH*. Jeju Island, Korea, pp. 1185–1188.

Vainio, M., Airas, M., Järvikivi, J. and Alku, P. (2010) 'Laryngeal voice quality in the expression of focus', in *Proceedings of INTERSPEECH*. Chiba, Japan, pp. 921–924.

Vainio, M. and Järvikivi, J. (2006) 'Tonal features, intensity, and word order in the perception of prominence', *Journal of Phonetics*, 34(3), pp. 319–342. doi: http://dx.doi.org/10.1016/j.wocn.2005.06.004.

Veldhuis, R. (1998) 'A computationally efficient alternative for the Liljencrants–Fant model and its perceptual evaluation', *The Journal of the Acoustical Society of America*, 103(1), pp. 566–571. doi: 10.1121/1.421103.

Villavicencio, F., Robel, A. and Rodet, X. (2006) 'Improving Lpc Spectral Envelope Extraction Of Voiced Speech By True-Envelope Estimation', in *2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings*. IEEE, pp. I-869-I–872. doi: 10.1109/ICASSP.2006.1660159.

Wang, Y. *et al.* (2017) 'Tacotron: Towards End-to-End Speech Synthesis', in *Proceedings of INTERSPEECH*. Stockholm, Sweden, pp. 4006–4010. doi: 10.21437/Interspeech.2017-1452.

Wong, D., Markel, J. and Gray, A. (1979) 'Least squares glottal inverse filtering from the

acoustic speech waveform', *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(4), pp. 350–355.

Wu, Z., Watts, O. and King, S. (2016) 'Merlin: An Open Source Neural Network Speech Synthesis System', *9th ISCA Speech Synthesis Workshop*, 2016, pp. 202–207. doi: 10.21437/ssw.2016-33.

Yamagishi, J., King, S., Renals, S., Nose, T., Zen, H., Tokuda, K., Ling, Z. H. and Toda, T. (2009) 'Robust Speaker-Adaptive HMM-Based Text-to-Speech Synthesis', *IEEE Transactions on Audio, Speech and Language Processing*, 17(6), pp. 1208–1230. doi: 10.1109/TASL.2009.2016394.

Yamagishi, J., Masuko, T. and Kobayasih, T. (2004) 'HMM-Based Expressive Speech Synthesis - Towards TTS with Arbitrary Speaking Styles and Emotions', in *Proc. of Special Workshop*. Maui, Hawaii.

Yamagishi, J., Onishi, K., Masuko, T. and Kobayashi, T. (2005) 'Acoustic Modeling of Speaking Styles and Emotional Expressions in HMM-Based Speech Synthesis', *IEICE transactions on information and systems*, 88(3), pp. 502–509.

Yamamoto, R. (2017) 'nnmnkwii: Library to build speech synthesis systems designed for easy and fast prototyping'. Available at: https://github.com/r9y9/nnmnkwii (Accessed: 25 May 2019).

Yanushevskaya, I., Gobl, C., Kane, J. and Ní Chasaide, A. (2010) 'An exploration of voice source correlates of focus', in *Proceedings of INTERSPEECH*. Makuhari, Japan, pp. 462–465.

Yanushevskaya, I., Gobl, C. and Ní Chasaide, A. (2009) 'Voice parameter dynamics in portrayed emotions', in *6th International Workshop on Models and Analysis of Vocal Emissions for Biometrical Applications (MAVEBA 2009)*. Florence, Italy, pp. 21–24.

Yanushevskaya, I., Gobl, C. and Ní Chasaide, A. (2018) 'Cross-language differences in how voice quality and f0 contours map to affect', *Journal of the Acoustical Society of America*, 144(5), pp. 2730–2750. doi: https://doi.org/10.1121/1.5066448.

Yanushevskaya, I., Murphy, A., Gobl, C. and Ní Chasaide, A. (2016a) 'Perceptual Salience of Voice Source Parameters in Signaling Focal Prominence', in *Proceedings of INTERSPEECH*. San Francisco, USA, pp. 3161–3165. doi: 10.21437/Interspeech.2016-1160.

Yanushevskaya, I., Ní Chasaide, A. and Gobl, C. (2011) 'Universal and Language-Specific Perception of Affect From Voice', *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS 2011)*. Hong Kong, China, pp. 2208–2211.

Yanushevskaya, I., Ní Chasaide, A. and Gobl, C. (2016b) 'The interaction of long-term voice quality with the realization of focus', in *Speech Prosody 2016*, pp. 931–935. doi: 10.21437/SpeechProsody.2016-191.

Yanushevskaya, I., Ní Chasaide, A. and Gobl, C. (2017) 'Cross-speaker variation in voice source correlates of focus and deaccentuation', in *Proceedings of INTERSPEECH*. Stockholm, Sweden, pp. 1034–1038.

Yegnanarayana, B. and Veldhuis, R. N. J. (1998) 'Extraction of vocal-tract system characteristics from speech signals', *IEEE Transactions on Speech and Audio Processing*, 6(4), pp. 313–327. doi: 10.1109/89.701359.

Zen, H., Senior, A. and Schuster, M. (2013) 'Statistical parametric speech synthesis using deep neural networks', in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, pp. 7962–7966. doi: 10.1109/ICASSP.2013.6639215.

Zen, H., Toda, T., Nakamura, M. and Tokuda, K. (2007) 'Details of the nitech HMM-based speech synthesis system for the blizzard challenge 2005', *IEICE Transactions on Information and Systems*, E90-D(1), pp. 325–333. doi: 10.1093/ietisy/e90-1.1.325.

Zen, H., Toda, T. and Tokuda, K. (2008) 'The nitech-NAIST HMM-based speech synthesis system for the blizzard challenge 2006', *IEICE Transactions on Information and Systems*, E91-D(6), pp. 1764–1773. doi: 10.1093/ietisy/e91-d.6.1764.

Zen, H., Tokuda, K. and Black, A. W. (2009) 'Statistical parametric speech synthesis', *Speech Communication*, 51(11), pp. 1039–1064. doi: 10.1016/j.specom.2009.04.004.