Fundamental Frequency Modelling: An Articulatory Perspective with Target Approximation and Deep Learning

Hao Liu

A thesis submitted in partial fulfilment

of the requirements for the degree of

Doctor of Philosophy

Department of Speech, Hearing and Phonetic Sciences

Division of Psychology and Language Sciences

Faculty of Brain Sciences

University College London

January 2017

Declaration

I, Hao Liu , confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Hao Liu

January 2017

Abstract

Current statistical parametric speech synthesis (SPSS) approaches typically aim at state/frame-level acoustic modelling, which leads to a problem of frame-by-frame independence. Besides that, whichever learning technique is used, hidden Markov model (HMM), deep neural network (DNN) or recurrent neural network (RNN), the fundamental idea is to set up a direct mapping from linguistic to acoustic features. Although progress is frequently reported, this idea is questionable in terms of biological plausibility.

This thesis aims at addressing the above issues by integrating dynamic mechanisms of human speech production as a core component of F0 generation and thus developing a more human-like F0 modelling paradigm. By introducing an articulatory F0 generation model – target approximation (TA) – between text and speech that controls syllable-synchronised F0 generation, contextual F0 variations are processed in two separate yet integrated stages: linguistic to motor, and motor to acoustic.

With the goal of demonstrating that human speech movement can be considered as a dynamic process of target approximation and that the TA model is a valid F0 generation model to be used at the motor-to-acoustic stage, a TA-based pitch control experiment is conducted first to simulate the subtle human behaviour of online compensation for pitch-shifted auditory feedback.

Then, the TA parameters are collectively controlled by linguistic features via a deep or recurrent neural network (DNN/RNN) at the linguistic-to-motor stage. We

trained the systems on a Mandarin Chinese dataset consisting of both statements and questions. The TA-based systems generally outperformed the baseline systems in both objective and subjective evaluations. Furthermore, the amount of required linguistic features were reduced first to syllable level only (with DNN) and then with all positional information removed (with RNN). Fewer linguistic features as input with limited number of TA parameters as output led to less training data and lower model complexity, which in turn led to more efficient training and faster synthesis.

Acknowledgements

First of all, I would like to thank my principal supervisor, Yi Xu, for his enduring support and patience in the past years. I am very grateful for his insightful advice and encouragement during the tough time of my study. Thanks to my subsidiary supervisor, Mark Huckvale, for helping me solve various technical problems. Thanks to Bob Ladd (Edinburgh) for encouraging me to pursue a PhD in prosody modelling. I will never forget my first visit to his office. Especially, I would also like to thank Chunyu Kit (CityU HK). Without him, I probably wouldn't get the chance to study in the UK.

Many thanks to my friends and colleagues at UCL and Edinburgh. Special thanks to Heng Lu and Xu Shao for accepting my visit to Nuance in California last summer and offering my current job at Nuance in Shanghai. Thanks to my parents and my wife Xiaoxuan Liu for everything.

Lastly, I would like to express my deepest gratitude to my motherland for financially supporting my study at UCL.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Speech is the most common means of human-human communication, and speech science research aims at achieving a clear understanding of human speech production and perception. In recent decades, along with the advances of computer science and machine learning, research and development in speech science are not only paramount for understanding human speech production and perception, but also able to facilitate human-machine interaction. This thesis follows the philosophy that '*better speech technology will come from better speech science*' (Huckvale, 2002, p. 1261).

The goal of the work presented in this thesis is to achieve a more human-like fundamental frequency (F0) modelling paradigm for speech synthesis and address the frame-by-frame independence issue in its typical statistical approaches. Focusing on the application of a recently developed articulatory F0 production model — target approximation (TA), the contribution of the thesis will be twofold. First, the online compensation response to the pitch-shifted auditory feedback will be simulated with underlying articulatory pitch control based on the TA model. Second and more importantly, by linking the TA model to state-of-the-art deep

learning techniques, a two-stage articulatory-based F0 modelling system will be implemented and its performance will be compared with the baseline systems.

In this chapter, we will start with a brief introduction to current speech synthesis approaches and then go deep into the issues in F0 modelling.

1.1 Speech Synthesis

Speech synthesis, also known as **text-to-speech** or **TTS**, is a process of automatically converting written text to synthetic speech voice by computer. It has a broad range of applications in human-machine interaction. Some important applications are clinical, which may include reading systems for the blind, where a system would read some text from a book and convert it to speech; or speaking assistive systems for people who cannot speak, where a system would speak out what the disabled person wants to say. In recent years, with remarkable advances in speech synthesis and other complementary technologies such as speech recognition, natural language understanding and generation, TTS systems are more commonly used in intelligent personal assistants (e.g. Siri on iPhone), call-centre automation, reading of news, weather reports, travel directions and a wide variety of other human-machine interactive applications (Taylor, 2009).

A TTS system normally consists of two parts — 'frontend' and 'backend'. The frontend converts the written text of an utterance to contextual linguistic specifications, which may include words, phone sequence, part-of-speech (POS) tags, phrase boundaries, and so on. The frontend is a natural language processing (NLP) component, which is usually a combination of linguistic rules and statistical models, and is usually language dependent. The backend generates speech waveform based on the linguistic specifications given by the frontend and usually can be language independent. When comparing speech synthesis approaches, we normally refer to

methodological differences between backends since different backends often share the same frontend with a few modifications. In general, besides the old-fashioned waveform synthesis (e.g. MITalk (Allen, Hunnicutt, Klatt, Armstrong & Pisoni, 1987)) which employed a rule-based source-filter model to produce speech waveform (Holmes, Mattingly & Shearme, 1964), there are currently three major speech synthesis approaches: concatenative synthesis, statistical parametric synthesis and articulatory synthesis.

1.1.1 Concatenative synthesis

The concatenative approach emerged earlier than the statistical parametric approach and a number of commercial applications are still based on this approach. Today when we mention concatenative synthesis we normally refer to the **unit-selection synthesis** (Hunt & Black, 1996). The most renowned unit-selection speech synthesis system *Festival* was created in 1997 for research purposes at the Centre for Speech Technology Research, University of Edinburgh (Taylor, Black & Caley, 1998).

In this approach, segmented speech sound *units* (e.g. diphones) are selected from a recorded speech database and concatenated to synthesise novel utterances. Since the units used for synthesis are real examples of human speech, good naturalness can be expected. Usually, there are multiple examples of a unit in the database subject to contextual differences so that the process of synthesis can be seen as a process of finding the most suitable unit sequence. The selection of a particular unit is jointly determined by two functions, one is *target cost*, which measures the contextual feature mismatch between a unit that we need and the units that we have in the database; the other is *join cost*, which measures the acoustic mismatch between potentially concatenated units. The challenge of unit-selection synthesis is to balance between the two cost functions with search algorithms (e.g. Viterbi

search (Forney, 1973)) and find a unit sequence which requires minimal signal processing to eliminate the glitches (join effects) between adjacent units. In the perfect situation, for example, to resynthesise a recorded utterance in the database, the original units of it will be selected with no signal processing needed at all and the synthetic utterance sounds quite natural.

This kind of unit-selection TTS system can bring us the highest naturalness of synthetic speech on the one hand, but on the other hand it requires us to build and maintain a large database to achieve the goal. Specifically, practical issues include how to optimise a recording script in order to obtain a cost-effective database covering a reasonable number of multi-phone or even multi-syllable tokens, how to segment and label units, how to efficiently search the database, how to deal with the missing units, and so on (Clark, Richmond & King, 2007). Furthermore, since the units are recorded beforehand and so cannot be manipulated to any great extent, we do not have any real control over the synthetic speech. Namely, we can only control the unit selection process and how selected units are glued together (e.g. alleviating glitches with signal processing techniques), but cannot control the actual content and effect of a unit. In the case of multi-speaker synthesis, the situation would be exacerbated since most of the recorded units are highly speaker-specific and can hardly be used interchangeably.

1.1.2 Statistical parametric synthesis

Statistical parametric speech synthesis (SPSS) (Yoshimura, Tokuda, Masuko, Kobayashi & Kitamura, 1999) has dominated the field of TTS research in the last decade. The success of SPSS mainly relies on the use of hidden Markov models (HMMs) and Gaussian mixture models (GMMs) (Tokuda, Nankaku, Toda, Zen, Yamagishi & Oura, 2013) to learn acoustic features from a large speech dataset and then generate and output parameter sequence for unseen text during synthesis.

In its conventional approach, spectral and F0 features are first extracted frame-wise from training data. Then linguistic context-dependent phone HMMs, which represent nonstationary acoustic feature distributions with a sequence of hidden states (usually five states per phone model), are trained on the extracted acoustic features via the maximum likelihood (ML) estimation by using the Expectation-Maximisation (EM) algorithm (Dempster, Laird & Rubin, 1977). State-level single Gaussian or GMM conditional probability density functions (PDFs) are computed accordingly. Binary decision trees are then constructed to cluster and tie contextually similar states together and set up a mapping from contextual linguistic features (obtained from text analysis via a frontend) to HMM states. At the synthesis stage, acoustic parameters are generated from decision-tree-selected HMM sequences based on the maximum likelihood parameter generation (MLPG) algorithm with both static and dynamic acoustic features (Tokuda, Yoshimura, Masuko, Kobayashi & Kitamura, 2000). Finally, generated acoustic parameters are sent to a vocoder (e.g. STRAIGHT by Kawahara, Masuda-Katsuse & de Cheveigne (1999)) for synthesising waveforms. Although this HMM-based approach can generate highly intelligible speech with flexible controllability (Zen, Tokuda & Black, 2009), the synthesis tends to be over-smooth and is still not as good as the best-quality concatenative systems (Hunt & Black, 1996). This critical issue has been alleviated but not fully resolved by a number of techniques, such as considering global variance (GV) during synthesis (Toda & Tokuda, 2007), minimum generation error (MGE) training (Wu & Wang, 2006), trajectory HMM modelling (Zen, Tokuda & Kitamura, 2007) and modulation spectrum post-filtering (Takamichi, Toda, Neubig, Sakti & Nakamura, 2014).

Recently, along with its successful application in automatic speech recognition (ASR), deep neural network (DNN) has shown its power to improve the accuracy of statistical acoustic modelling in speech synthesis (Ling, Deng & Yu, 2013a; Lu,

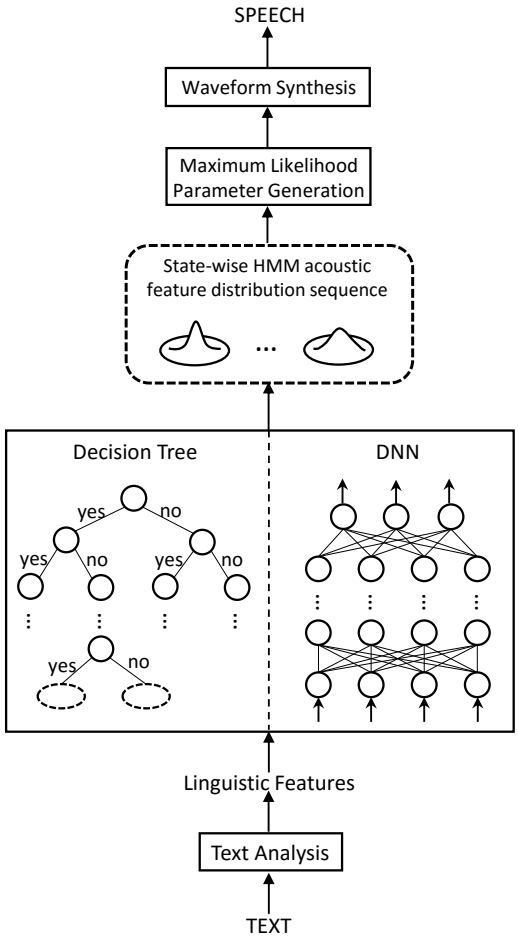


Figure 1.1 Overview of statistical parametric speech synthesis (SPSS) based on decision tree or DNN.

King & Watts, 2013; Zen, Senior & Schuster, 2013; Zen & Senior, 2014; Qian, Fan, Hu & Soong, 2014). In general, it overcomes some problems (e.g. complexity limit, training data fragmentation) faced by decision-tree-based approaches by offering a highly complex and nonlinear yet efficient mapping between linguistic features and state-level acoustic features via a compact hierarchical structure. Figure 1.1 shows a unified overview of such decision tree or DNN-based SPSS. Some more recent studies demonstrate even better results by embedding the parameter generation process inside long short-term memory[1] (LSTM) based recurrent neural networks

[1]In contrast to 'short-term memory' (i.e. typical RNN), which can theoretically model sequential events but is practically not very successful due to the gradient-vanashing problem when minimal time lags between inputs and outputs are long, the 'long' version here overcomes such problem. More details please refer to Hochreiter & Schmidhuber (1997).

(RNNs) and directly predicting vocoder-ready acoustic parameter sequence (Fan, Qian, Xie & Soong, 2014; Fernandez, Rendel, Ramabhadran & Hoory, 2014; Zen & Sak, 2015). Some more innovative research investigated the possibility of directly modelling speech at the waveform level with recurrent or convolutional neural networks. Namely, contextual linguistic features are mapped directly to time domain waveform signals with RNN or CNN modelling. The pilot experiments (Tokuda & Zen, 2015; Tokuda & Zen, 2016; van den Oord, Dieleman, Zen, Simonyan, Vinyals, Graves, Kalchbrenner, Senior & Kavukcuoglu, 2016) show that these methods are able to generate raw speech waveforms which mimic any human voice. However, directly predicting waveform sample-by-sample is an extremely computationally expensive and time-consuming task, which makes their drawback very clear in real applications. More details concerning the SPSS approaches will be reviewed in Chapter 5 when we need to build SPSS baseline systems for F0 modelling.

The DNN/RNN-based SPSS approaches have become state-of-the-art in speech synthesis nowadays, and some researchers believe that these approaches simulate the process of human speech production (Fan et al., 2014; Zen, 2015). However, comparing them to the articulatory synthesis introduced below, we can find that the success of SPSS approaches should be mainly owed to the advance of machine learning techniques. From a methodological perspective, what SPSS approaches have achieved so far is not even close to the reality of speech production.

1.1.3 Articulatory synthesis

The basic methodology of articulatory synthesis is to produce speech by simulating principles of speech production. Articulatory synthesis is a direct reflection of our current knowledge in speech science with aims of building bionic vocal tract models and replicating the articulation process. Theoretically, this approach has the potential to simulate every aspect of human speech production. Important issues involved

in articulatory synthesis include the difficulties in accurately measuring vocal tract and real articulatory process (e.g. through X-ray, magnetic resonance imaging (MRI) or electromagnetic articulography (EMA)) as well as the computational complexity needed in controlling the artificial articulators to produce speech.

A typical articulatory synthesiser is composed of three main parts:

- a vocal tract model constructed based on the measurements of real vocal tract
- a mechanism to control the artificial articulators
- an acoustic model of glottal excitation (sound source) and vocal tract resonance (filter)

Different from the waveform synthesis, which directly specifies sound source parameters, formant frequencies and bandwidths, articulatory synthesis allows us to actually adjust the shape of vocal tract and control the movements of different artificial articulators by means of a small set of articulatory parameters. In a 3D vocal tract model, an *area function* is used to describe how the cross sectional area of the vocal tract tube varies between the glottis and lips. During production, temporal and spatial variations of the articulatory parameters change the area function so that speech waveform can be calculated by the acoustic model based on the sequence of area functions.

VocalTractLab developed by Birkholz, Jackèl & Kröger (2006) is a representative of such articulatory synthesisers. It consists of a detailed 3D model of the vocal tract (Figure 1.2) that can be configured to fit the anatomy of any specific speaker, an advanced self-oscillating model of the vocal folds and an accurate method for the aeroacoustic simulation of the speech signal. VocalTractLab is capable of generating a full range of speech sounds (e.g. voiced vowels; voiceless consonants including fricatives and plosives; consonant-vowel coarticulation) by controlling vocal tract shapes, aerodynamics and voice quality (Birkholz, 2013).

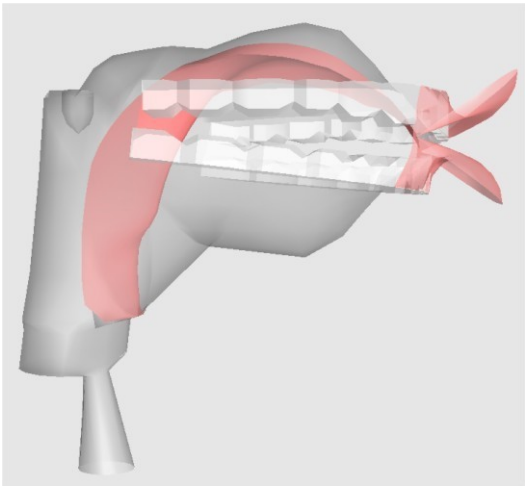


Figure 1.2 A 3D vocal tract model in VocalTractLab (Birkholz, 2013).

What is directly relevant to this thesis is that, inspired by the target approximation (TA) model (Xu & Wang, 2001; Prom-on, Xu & Thipakorn, 2009) in F0 production, VocalTractLab recently adopted the concept of *sequential target approximation* as the mechanism to control the dynamics of the articulators (Birkholz, Kröger & Neuschaefer-Rub, 2011). In their experiment of reproducing EMA-measured movement trajectories of the constrictors (lower lip, tongue tip and tongue dorsum) and the jaw, the model-generated trajectories closely matched the observed ones, which supported the hypothesis that articulatory movement can be considered as a dynamic process of target approximation.

Based on this finding, Prom-on, Birkholz & Xu (2013) successfully trained a TA-based VocalTractLab to produce several simple continuous utterances for the first time. In that experiment, surface acoustics of natural speech and manually annotated segmental boundaries were used and underlying targets were optimised with bounded 'analysis-by-synthesis'. Namely, candidate targets were used to actually synthesise speech sounds round by round and were continuously optimised by correcting errors found between synthetic samples and their natural counterparts. Figure 1.3 displays the workflow of this experiment.

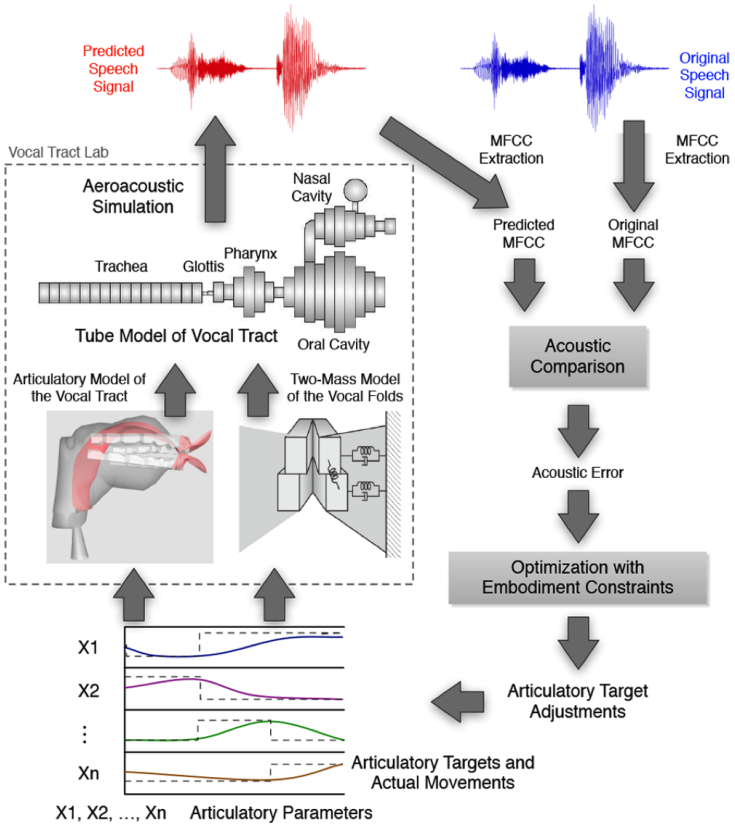


Figure 1.3 The workflow of training a TA-based VocalTractLab to produce continuous speech (Prom-on, Birkholz & Xu, 2013).

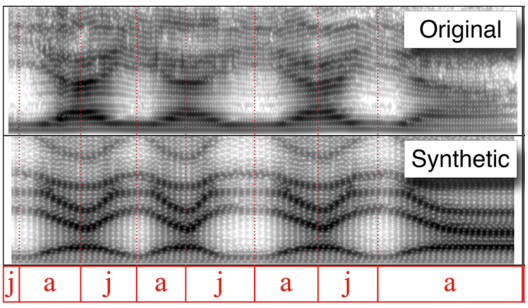


Figure 1.4 The original and synthetic spectrograms of the utterance /ˌjaja'jaja/ (Prom-on, Birkholz & Xu, 2013).

The learned articulatory targets were able to generate utterances that approximate to the original both acoustically (Figure 1.4) and articulatorily (Figure 1.5). From the latter we can see that the trajectories generated by the TA model with the learned articulatory targets are very close to the EMA trajectories.

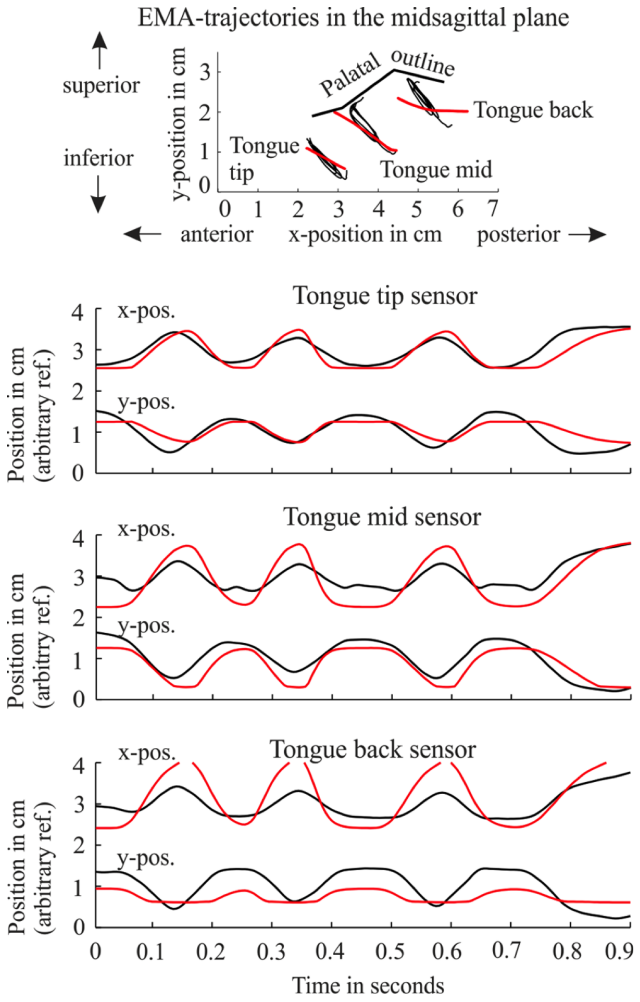


Figure 1.5 EMA trajectories of the three tongue sensors for the utterance /ˌjaja'jaja/. Original EMA trajectories are shown as black curves and TA-generated trajectories based on the learned articulatory targets are shown as red curves (Prom-on, Birkholz & Xu, 2013).

Speech production is a rather complex motor control process and our knowledge about it is still limited. Therefore, articulatory synthesis remains as a very difficult task at present. While progress is being made slowly and so far articulatory systems are still impractical for general use, some advances of this approach may already be useful to other speech synthesis approaches. For example, studies by Toda, Black & Tokuda (2004), Ling, Richmond, Yamagishi & Wang (2008), Ling, Richmond, Yamagishi & Wang (2009), Ling, Richmond & Yamagishi (2012) and Ling, Richmond & Yamagishi (2013b) tried to integrate EMA features into HMM-based

speech synthesis to achieve articulatory control. The work presented in this thesis, which intends to make direct use of the sequential target approximation model for F0 modelling in TTS, is another attempt.

1.2 F0 Modelling

1.2.1 Intonation

The frequency of vocal fold vibration is called *fundamental frequency* (F0), which is a physical property of speech. *Pitch*, as another commonly used term, is the psychophysical (perceptual) correlate of F0. These two terms are often used interchangeably in many contexts. Fundamental frequencies are normally described as F0 contours, and patterns of such contours are referred to as *intonation*. Intonation plays an important role in human-human communication. The functions of intonation include but are not limited to (van Santen, Mishra & Klabbers, 2008):

- Structuring utterance and resolving syntactic ambiguities
- Conveying pragmatic information, e.g. emphasis, contrast, focus, etc.
- Providing cues of the emotional state of the speaker
- Serving as a continuity guide in noisy environments

A phonological defination of intonation is given in Ladd (2008, p. 4): 'Intonation, as I will use the term, refers to the use of *suprasegmental* phonetic features to convey "post-lexical" or *sentence-level* pragmatic meanings in a *linguistically structured* way.' There are at least two points that we should note from this definition. One is *suprasegmental*, which indicates that intonation cannot be accounted for strictly based on the segmental structure of an utterance. Instead, it should be seen as a property of larger units of speech (e.g. syllable, word and phrase). The other is

'post-lexical', by which it means that intonation as a phonological concept excludes lexical features (e.g. tones as used in tonal languages, in which pitch contours are used to distinguish one word from another). However, since this thesis is intended to achieve an articulatory modelling of pitch contour that can be used universally for speech synthesis, we adopt the expression 'F0 modelling' more often instead of the restricted 'intonation modelling' to indicate that the modelling is for both lexical and post-lexical pitch patterns.

With the progress of speech synthesis, the intelligibility of synthetic speech nowadays is satisfactory in most cases. The naturalness of synthetic intonation, which directly affects the overall quality of synthetic speech, has become the biggest issue in speech synthesis and has attracted extensive research.

1.2.2 Issues

Regardless of F0 modelling approaches, a shared technical issue in this field is the difficulty in F0 measurement. The accuracy of F0 can be affected by the phonatory state of a speaker (e.g. creaking and breathiness) and the environmental noise during recording. In some TTS systems, multiple pitch trackers are often used to offer more than one pitch estimate to the same utterance so that a relatively reliable pitch contour can be ultimately obtained by averaging across multiple estimations. Besides that, there are many other voicing and segmental perturbation problems that are not easy to tackle. However, even if a pitch contour is accurately extracted, it is still difficult to tell which feature of it should be seen as a segmental effect and which should be taken as an intonational variation intended by the speaker to convey information.

Regarding the issues in terms of F0 modelling methodology, the approaches can be divided into two types: one is model-based and the other is statistical data-driven.

Model-based

As the name suggests, this type of F0 modelling tries to apply existing intonation models to TTS systems. While intonation has been studied intensively in the past decades, different schools of thought have emerged and consensus has never been reached. The controversies among the schools actually show that our theoretical understanding of intonation is still incomplete. And this has actually limited the success of speech synthesis. Over the time, the influential theories and models include, but are not limited to, the Autosegmental-Metrical (AM) theory (Pierrehumbert, 1980) with ToBI annotation scheme (Silverman, Beckman, Pitrelli, Ostendorf, Wightman, Price, Pierrehumbert & Hirschberg, 1992), the Fujisaki model (Fujisaki, 1983), the IPO approach ('t Hart, Collier & Cohen, 1990), the Tilt model (Taylor, 1992; Taylor, 2000), the STEM-ML model (Kochanski & Shih, 2003), the SFC model (Bailly & Holm, 2005) and the TA model (Xu, 2005; Prom-on et al., 2009). Before we bring up more details about these theories and models in Chapter 2 and Chapter 3, some issues can be discussed here.

One of the major issues for the *phonological* theories (e.g. AM and IPO) is that they are qualitative and symbolic so their usage in speech synthesis is heavily dependent on accurate annotations. For example, when the ToBI labels are adopted to mark intonational events defined in the AM theory (*pitch accents*, *phrase accents* and *boundary tones*), the consistency among annotators is critical for training TTS models. Practically, although the place of accents can usually be agreed by different annotators, making distinctions between different accent types is not always as easy as the originally illustrated examples (Syrdal & McGory, 2000). Although some tools have been developed to automate the annotation process (e.g. AuToBI by Rosenberg (2010)), manual correction is still needed (the reported accuracy for pitch accent type classification was around 70%).

Compared to the phonological theories, the *phonetic* models (e.g. Tilt and SFC) may look less problematic. An advantage of them over the phonological ones is that they are quantitative and parametric (using continuous parameters rather than imposing categorical classification on the intonational events). So, they are capable of directly modelling most of the surface intonational events. However, a main drawback of these models is also clear: no matter linear or superpositional, they are only concerned with the intonational events, with the underlying mechanism of F0 contour formation largely ignored. As a consequence, in the Tilt model, for example, the pitch accents and boundary tones are modelled piecewise and then inter-connected by linear interpolation. Moreover, as the Tilt model was developed for English, it may not be directly applied to other languages with much more complex pitch patterns (Taylor, 1992).

In terms of the *articulatory* models (e.g. Fujisaki and STEM-ML), while the physiological process of F0 production has been taken into consideration, they are mathematically and physically too complex with potentially redundant degrees of freedom (DOF), which directly leads to training difficulty. Although they are able to produce some example utterances with careful tuning, large-scale application in TTS is still not feasible. As pointed out by Hirose, Sato, Asano & Minematsu (2005), the Fujisaki model is actually a variation of dynamic-system model such as the one proposed by Ross & Ostendorf (1999). The difference is that the Fujisaki model made several physiological assumptions to constrain the DOF. However, the constraints are not enough and the Fujisaki model still suffers from the DOF problem with many interactive variables involved and which makes it difficult to train. In Hirose et al. (2005), for example, further constraints were applied on the location of commands in the Fujisaki model in order to enable an application. A valuable comment can be found in Taylor (2009, p. 253) when introducing the dynamic-system model by Ross & Ostendorf (1999): 'The dynamic-system model

is a natural choice for statistical generation of F0 contours since it is well suited to the job of generating continuous trajectories. If it has any weaknesses, we can point to the facts that the state trajectories are limited to being those of a first-order filter, ..., and the training process can be quite intricate'. The TA model, which will be introduced and applied later in this thesis, has addressed these issues by providing a high-order dynamical system with very limited DOF based on our latest understanding of speech production.

Statistical data-driven

Satisfactory results in F0 modelling have not been achieved in statistical parametric speech synthesis (Ling, Kang, Zen, Senior, Schuster, Qian, Meng & Deng, 2015). The major issue in this approach is that the F0 contours are modelled frame-by-frame independently (i.e. observation probabilities suddenly change when moving from one state to another), which is mainly due to the discrete nature of the hidden state space in the HMM model. As a consequence, the generated F0 contours would be piecewise and unnatural if no further processing is provided. The aforementioned MLPG algorithm (Tokuda et al., 2000) was targeted exactly at this problem and managed to output smooth contours by taking into account the learned dynamics of F0. However, this solution is not perfect since it requires going though the whole utterance again after the acoustic features are predicted, which causes much delay, let alone the smoothing process is implicit and uncontrollable.

Besides that, all previous studies tried to consider numerous contextual prosodic factors (e.g. phone position in phrase/sentence) in an attempt to better represent longer-term F0 patterns. Hierarchical and additive constraint methods have also been developed either by modelling prosodic components at different phonetic levels (Qian, Liang & Soong, 2008; Wang, Ling, Zhang & Dai, 2008; Zen & Braunschweiler, 2009; Lei, Wu, Soong, Ling & Dai, 2010; Qian, Wu, Gao &

Soong, 2011; Wu & Soong, 2012) or by relying on discrete cosine transform (DCT) to capture suprasegmental F0 patterns (Teutenberg, Watson & Riddle, 2008; Wu, Qian, Soong & Zhang, 2008; Yin, Lei, Qian, Soong, He, Ling & Dai, 2014; Yin, Lei, Qian, Soong, He, Ling & Dai, 2016).

Furthermore, sequential training criteria minimising errors between utterance-level F0 trajectories rather than independent F0 frames were developed for both the conventional HMM-based (Wu & Wang, 2006) and the DNN-based approaches (Fan, Qian, Soong & He, 2015; Wu & King, 2015; Wu & King, 2016a) to ensure that the inter-frame relationship is not lost at the training stage. Recent RNN-based approaches, on the other hand, offer a direct solution of sequence-to-sequence mapping so that the correlations between contiguous frames are not ignored and the dynamic process of speech production is always implicitly embedded.

What is common in these approaches, however, is that they treat articulatory mechanisms of F0 production only implicitly. Even in studies that try to integrate articulatory features into speech synthesis (Toda et al., 2004; Ling et al., 2008, 2009, 2012, 2013b), articulatory mechanisms are treated as unknown. A fundamental idea of these approaches is to set up a direct mapping from linguistic features to acoustic features, and such acoustic features can be multi-level with considerable hierarchical constraints. This idea, however, is questionable in terms of its 'biological plausibility'. Namely, do we human really produce speech with so much detail considered at the same time? Or, do we actually only consider intended linguistic meanings and the final speech outputs are natural realisations that are physically and physiologically constrained? The answer should be obvious.

Furthermore, from our perspective, overusing machine learning techniques is becoming a trend in the field. This should by no means be encouraged, although sometimes it may bring us the best results. Recent CNN-based waveform modelling approach (e.g. WaveNet by van den Oord et al. (2016)) is such an example. While

it is indeed able to produce very natural speech that captivates our ears, the computation cost of it can be unreasonably high.[2] Despite that acceleration methods are being explored (Paine, Khorrami, Chang, Zhang, Ramachandran, Hasegawa-Johnson & Huang, 2016), it is difficult to call it the final solution of TTS. After all, the human speech production process is by no means as straightforward as 'text to speech'.

1.3 Proposed Articulatory F0 Modelling

In this thesis we explore an articulatory approach of F0 modelling that aims at achieving a more human-like complete speech production paradigm and addressing the frame-by-frame independence issue in SPSS approaches. By introducing an articulatory F0 generation model — target approximation (TA) — between text and speech that controls dynamical F0 generation, contextual F0 variations are processed in two separate yet integrated stages:

- I. *Linguistic-to-motor*: contextual linguistic features are associated to motor parameters of the TA model
- II. *Motor-to-acoustic*: F0 contours are dynamically generated by TA-based articulatory simulation

With the TA model, the syllable is considered as the basic F0 modelling unit instead of frame or phone as commonly used in other approaches. As mentioned in Section 1.2.2, the TA model is superior to other similar articulatory models in that it is implemented as a high-order dynamical system but with only a limited number of DOF. This design is based on the understanding that basic human speech movement can be considered as a dynamic process of target approximation

[2]The authors of WaveNet reported that it took about 90 minutes to synthesise 1 second of speech on a laptop.

with syllable-synchronised control (Xu & Wang, 2001; Prom-on et al., 2009). More importantly, unlike common dynamical system models, those F0 movements not occurring in natural utterances are largely avoided in the TA model. While some other articulatory synthesis studies have found TA to be a valid mechanism to drive the actual movements of articulators (Section 1.1.3), in this thesis we further supply a TA-based pitch control experiment based on the online auditory feedback compensation behaviour found in empirical research. The purpose of this experiment is not only to demonstrate the capability and flexibility of TA, but also to show its validity to be used at the motor-to-acoustic stage of our proposed system.

We simulate the human-like two-stage F0 production process by linking the TA model to a DNN, which learns the 'linguistic-to-motor' mapping given the 'motor-to-acoustic' mapping provided by TA (Figure 1.6). By predicting syllable-synchronised TA motor parameters instead of frame-by-frame acoustic features, the unnatural sudden fluctuations in F0 trajectories are avoided. The possibility of adopting a gated recurrent unit (GRU) based RNN to help TA capture higher than the syllable level articulatory dynamics of F0 production is also explored.

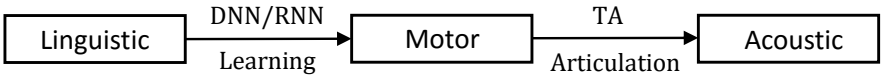


Figure 1.6 A two-stage speech production process.

Some previous studies have also experimented with the TA model before (Zhang, Wang, Yu & Wu, 2010; Pang, Wu & Cai, 2012; Na & Garner, 2013; Gao, Ling, Chen & Dai, 2014). However, the model was either incorporated into a hierarchical structure or used in a post-filtering way but seldom used on its own. While its efficacy has been demonstrated, the potential of the TA model was not fully utilised. This thesis aims to offer a canonical implementation of the TA model and

apply it directly to TTS to test its potential to further enhance synthesis quality in combination with the latest DNN and RNN learning methods.

The rest of the thesis is organised as follows. A selection of influential intonation theories and models are reviewed in Chapter 2, followed by Chapter 3 dedicated to introducing the TA model. Chapter 4 describes the TA-based simulation for online auditory feedback compensation in pitch, which contributes to the motor-to-acoustic stage. Chapter 5 reviews existing F0 modelling approaches in SPSS first and then completes the two-stage F0 modelling process by linking the TA model to DNN/RNN and building systems to evaluate the performance of the proposed articulatory F0 modelling approach. Chapter 6 is a general conclusion of the thesis.

Chapter 2

Intonation Theories and Models – The Literature

In this chapter, we will review some of the most influential theories and models of intonation proposed over the decades. Based on methodological differences, they are divided into *phonological*, *phonetic* and *articulatory*.

2.1 Phonological Models

2.1.1 The AM theory and ToBI labels

Originating from Liberman (1975) and Liberman & Prince (1977), the Autosegmental-Metrical (AM) theory (Goldsmith, 1990; Ladd, 2008) is a phonological theory of intonation. It first formally appeared in Pierrehumbert's PhD thesis (Pierrehumbert, 1980), so that is also known as the Pierrehumbert model. Together with its later extensions (Beckman & Pierrehumbert, 1986; Pierrehumbert & Hirschberg, 1990), the AM theory is treated as the first established intonation research framework.

The AM theory describes intonation as a linear sequence of high and low tones. The fundamental intonation units in the theory, as argued by Pierrehumbert, are simply **H** (high) and **L** (low) tones defined with respect to the current F0 range. These two types of tones are building blocks of *pitch accents*, *phrase accents* and *boundary tones*. Pitch accents can be formed by single tone (**H***, **L***), or by double tones (**H*+L**, **H+L***, **L+H***, **L*+H**). The star sign '*' in each pitch accent marks the tone that actually align with the stressed syllable. An *intermediate phrase* is composed of one or more pitch accents. An *intonational phrase*, which is the largest prosodic unit, is composed of one or more intermediate phrases. Boundary tones are additional tones defined at intonational phrase boundaries, which are single tones marked with '**%**' (**%H**, **%L**, **H%**, **L%**) to indicate the alignment to the onset or offset of the intonational phrase, respectively. The path of pitch movement between the last pitch accent to the boundary tone in an intonational phrase is represented by phrase accent, which is a single tone followed by a '**-**' sign (**H-**, **L-**). A collection of finite state grammar rules were further defined by Pierrehumbert to specify the possible combinations of pitch accents, phrase accents and boundaries tones. Figure 2.1 shows a labelled example F0 contour.

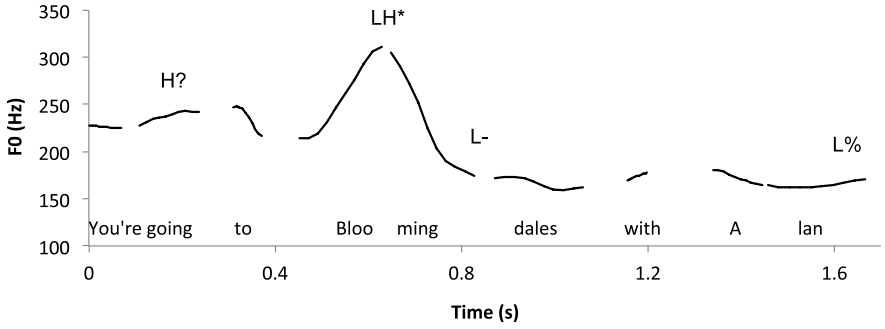


Figure 2.1 A labelled F0 contour based on the AM theory (Xu, 2015).

Early applications of this phonological model were rule-based, which require some heuristics definined by experts. In Pierrehumbert (1981), some *phonetic realization rules* (interpolation rules) were developed to generate F0 contours based

on the model. Given the tonal representation of an utterance determined by the finite state grammar, the target F0 values of tones can be specified first based on the metrical prominence of the associated syllables and the F0 values of the preceding tones. Then, the interpolation rules can be applied to connect these target F0 values to generate an F0 contour (van Santen et al., 2008). Later in the unit-selection system developed by Black & Hunt (1996), statistical data-driven approach was used to generate F0 contours based on the ToBI labels (see below).

It is worth mentioning that there is a *downdrift* phenomenon commonly observed in F0 contours across languages ('t Hart & Cohen, 1973; 't Hart & Collier, 1975; Ladd, 1984; Liberman & Pierrehumbert, 1984; Fujisaki & Kawai, 1988; Cooper & Sorensen, 2012). Namely, there is sometimes an overall left-to-right downward trend of the F0 contour in an utterance. While this phenomenon is regarded by most theories as an automatic physiological effect arising from the reduction of sub-glottal pressure during production, Pierrehumbert attributed it to a phonological effect *downstep*, which can be controlled by the speaker. The effect was first proposed to be triggered by the use of **H L H** sequence in Pierrehumbert (1980), and then rectified as by **H+L** pitch accent alone in Pierrehumbert & Hirschberg (1990).

ToBI (Tones and Break Indices) is a standard annotation scheme based on the Pierrehumbert model for transcribing intonation of English (Silverman et al., 1992). There are three parallel labelling tiers in ToBI including a *tone tier*, a *break index tier* and a *miscellaneous tier*. The tone tier allows users to specify the intonation events defined in the Pierrehumbert model. The break index tier is used to mark breaks ranging from 0 to 4, which indicates the strength of association between adjacent words at phrase boundaries (e.g. 0 for no boundary, 3 for intermediate phrase boundary and 4 for intonational phrase boundary) (van Santen et al., 2008). The

miscellaneous tier is an assistive tier for marking unintended pauses, disfluencies and so on.

ToBI is very influential since it provides a complete framework for intonation labelling for the first time, which allows different researchers to follow the same standard to conduct experiments based on the Pierrehumbert model. Moreover, ToBI also enables statistical modelling of intonation in speech synthesis. For example, in unit-selection synthesis, Black & Hunt (1996) used linear regression to predict three target F0 values (start, mid-vowel and end) for every syllable based on the intonation features (e.g. stress and syllable position) represented by ToBI labels. To the present, some DNN-based TTS systems are still using ToBI annotations as input features to assist intonation modelling.

2.1.2 The IPO approach

The IPO approach ('t Hart & Cohen, 1973; 't Hart & Collier, 1975; 't Hart et al., 1990), developed at the Institute of Perception Research (IPO), is a perception-oriented model of intonation. An influential notion made by this approach is that not all aspects of intonation are perceptually important to the human ear, and only those important are worth modelling. The important aspects of intonation in the IPO approach are *perceptually relevant pitch movements* rather than pitch levels, so that the intonation unit here is different from the level tones considered in the AM theory. There are two important procedures in the IPO approach, *stylisation* and *standardisation*.

In the stylisation procedure, the original F0 contours are fitted linearly by a series of piecewise straight lines known as *close-copy contours*, which represent the perceptually relevant pitch movements. Note that this procedure is interactive. A human listener is required to replace the F0 contours with a minimum number of straight lines and compare the intonation between the resynthesised utterance

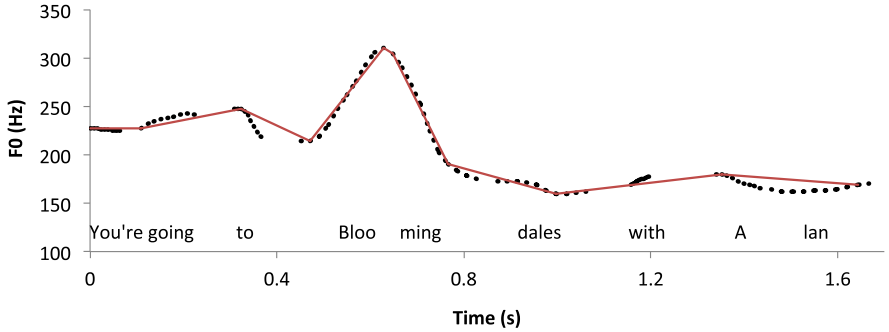


Figure 2.2 An example of a stylised F0 contour in the IPO approach (Xu, 2015).

and the original. The close-copy contours can be accepted once the resynthesised utterance is perceptually equal to the original. Figure 2.2 shows an example result of this procedure, in which the red straight lines are close-copy contours.

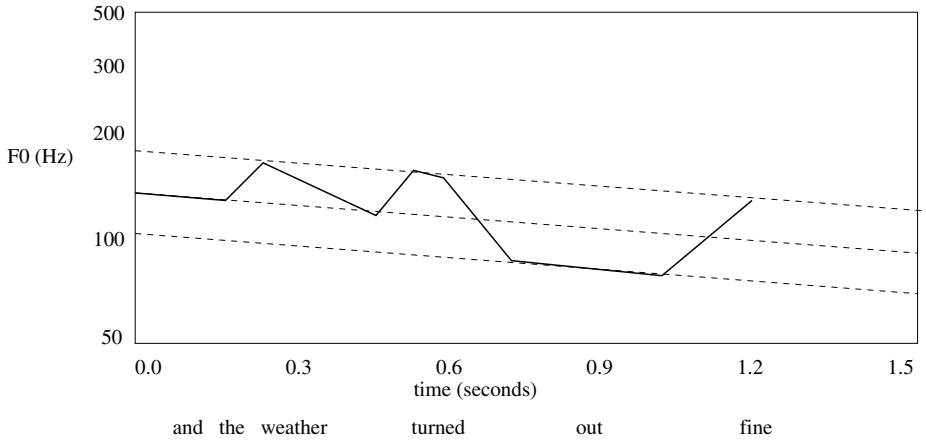


Figure 2.3 An example of a standardised F0 contour in the IPO approach (Willems, 1983).

In the standardisation procedure, the close-copy contours obtained in the stylisation procedure are collectively analysed and classified into an inventory of basic patterns, which are discrete, phonetically defined types of F0 rises and falls. Then an *intonation grammar* can be defined to specify possible and permissible combinations of the F0 rises and falls and used for speech synthesis. Besides that, downdrift is also considered in the IPO approach by specifying three declination lines (high, middle and low). The rising and falling F0 contours are realised between these lines.

Figure 2.3 shows an example of a standardised contour, in which the dashed lines are the three declination lines and the solid lines between them are the standardised F0 contours.

The assumption of perceptual importance of the IPO approach is a valuable contribution to our knowledge. However, the modelling process defined by the IPO approach is not as straightforward as the AM theory, which makes it rarely be used in real TTS applications.

2.1.3 Discussion

There are two shared properties between the two phonological models. The first is that they are both qualitative, symbolic and defined as descriptive schemes of intonation. This leads to an issue that both of them have to rely on accurate annotations. However, annotation is always an error-prone task, and it is also very difficult to achieve high consistency among human annotators. For example, in the case displayed in Figure 2.1 for the AM theory, the question mark after the first '**H**' indicates that the annotator is not sure about this particular decision so that further negotiations and checks are required. As for the IPO approach, the situation is even worse since the stylisation procedure which involves resynthesis and comparison so that it is more time-consuming, labour-intensive and error-prone. Although there are automatic tools available (e.g. AuToBI by Rosenberg (2010)), the tools need to be trained first on 'correctly' annotated training data. The second is that they both only consider the downdrift of F0 contour over a sentence. However, F0 contours can be much varied with various pragmatic meanings conveyed (e.g. the downdrift can be reversed at the end of yes/no questions). And the inclusion of complex pragmatics is exactly what current TTS systems lack.

A major difference between the two phonological models is that, the level tones are used as targets by the AM theory to mark intonation events, whereas more

interpretable pitch movements are favoured by the IPO approach. As a consequence, the AM theory has to rely on phonetic realization rules and linear interpolation to generate continuous F0 contours, whereas rules used by the IPO approach are more focused on controlling overall shape of F0 contours.

Rule-based modelling approach is easy to implement, efficient and able to produce consistent F0 contours. However, its disadvantages are also clear. It is highly dependent on expert rules, the naturalness is not sufficient and the variability of generated contours is not rich. On the one hand, there is always a lack of rules in rule-based systems. But on the other hand, designing and maintaining sophisticated rules is a rather challenging task. Nevertheless, rule-based systems can be used to directly test our theories so that are helpful for a better understanding of intonation.

2.2 Phonetic Models

2.2.1 The Tilt model

Developed by (Taylor, 1992; Taylor, 2000), Tilt is a phonetic model of intonation. The aim of the model is to provide a parameterised representation of intonation events for practical engineering use in speech synthesis. It is purely descriptive, with linguistic concerns (e.g. the AM theory) and biological plausibility (e.g. the Fujisaki model) ignored (Taylor, 2009). Instead of developing a fixed set of categories for intonation events as in the AM theory, the Tilt model represents the intonation events with a set of continuous parameters. Taylor (2009, p. 242) argues that the evidence for the particular categories defined in the AM theory is weak: 'With verbal language, phonetically we have a continuous space, either articulatory or acoustic, but cognitively this is divided up into a discrete set of phonological units, i.e. phonemes. The AM model follows the same policy with intonation, but ... there was no equivalent to the minimal-pair test to decide how the phonetic space

should be divided up' (in terms of intonation). This argument is also held by many other intonation models with a direct parametric representation of F0 contours, including the target approximation model (Chapter 3) used in this thesis.

Similar to the AM theory, the types of intonation event considered in the Tilt model are pitch *accent* and *boundary tone*, and they also occur in a linear fashion. Between events, *connections* are made by linear interpolations. For each event, the amplitude (Eq. (2.1)) and duration (Eq. (2.2)) of its rise and fall are parameterised (Figure 2.4). This leads to a sub-model of Tilt called the rise/fall/connection model (Taylor & Black, 1994).

$$\text{Amplitude} = A = |A_{\text{rise}}| + |A_{\text{fall}}| \tag{2.1}$$

$$\text{Duration} = D = |D_{\text{rise}}| + |D_{\text{fall}}| \tag{2.2}$$

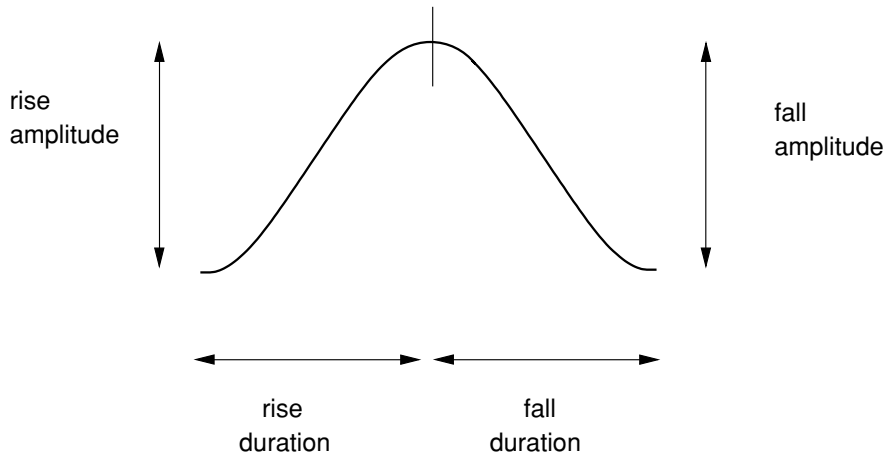


Figure 2.4 A pitch accent split into two parts, rise and fall. And the amplitude and duration of them are parameterised (Taylor, 2009).

A *tilt* parameter can be calculated to represent the shape of an intonation event. Two interim tilt parameters, tilt_{amp} (Eq. (2.3)) and tilt_{dur} (Eq. (2.4)), are calculated first and then combined into the final tilt parameter (Eq. (2.5)).

$$\text{tilt}_{\text{amp}} = \frac{|A_{\text{rise}}| - |A_{\text{fall}}|}{|A_{\text{rise}}| + |A_{\text{fall}}|} \tag{2.3}$$

$$\text{tilt}_{\text{dur}} = \frac{|D_{\text{rise}}| - |D_{\text{fall}}|}{|D_{\text{rise}}| + |D_{\text{fall}}|} \tag{2.4}$$

$$\text{tilt} = \frac{\text{tilt}_{\text{amp}} + \text{tilt}_{\text{dur}}}{2} \tag{2.5}$$

The range of the tilt parameter is $[-1, 1]$. -1 indicates a pure fall, 1 indicates a pure rise and 0 indicates a symmetrical bell shape. Figure 2.5 shows an array of pitch accents with different tilt values.

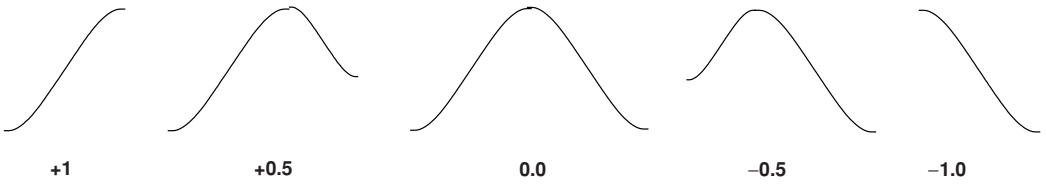


Figure 2.5 Five pitch accents with different tilt values (Taylor, 2009).

Besides the parameters describing the shape of an intonation event, there are two extra parameters describing the event position in the time-F0 plane when applying the Tilt model (van Santen et al., 2008). Moreover, in order to practically apply the Tilt model for TTS, the tilt events need to be labelled in the dataset so that automatic tilt analysis can apply to extract tilt parameters. Dusterhoff & Black (1997) used the Tilt model to predict F0 contours based on the tilt events derived from ToBI labels. Reddy & Rao (2013) and Reddy & Rao (2016), in their use of the Tilt model for neural network based concatenative synthesis, treated each syllable as an intonation event and used the extracted tilt parameters as an input feature in addition to other linguistic and production constraints. Their studies show that, with the inclusion of tilt parameters, the system performance significantly improved in both objective and subjective evaluations.

2.2.2 The SFC model

The Superposition of Functional Contours (SFC) model (Bailly & Holm, 2005) is a functional model of intonation. The functional approach of intonation modelling assumes that the defining properties of intonational patterns are *communicative meanings* rather than surface forms (phonetic properties). In other words, unlike the phonological models that categorise intonation events based on their acoustic similarity (e.g. high and low in the AM theory), the SFC model classifies F0 contours into *functional contours* (FC) based on different communicative meanings they convey. The communicative meanings are regarded as *metalinguistic functions* including segmentation, hierarchisation, emphasis and attitude. Each functional contour directly encodes a particular metalinguistic function. Especially, the functions are independent and may have various *scope*, so that it is possible that different functions affect the same part of an utterance. In this situation, the functional contours need to be combined together to form a final surface F0 contour.

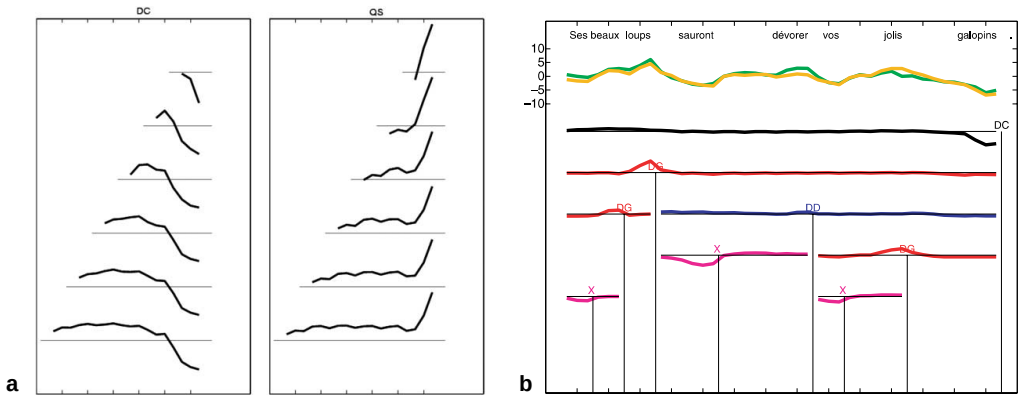


Figure 2.6 Examples of the SFC model. **a**: Synthesised functional contours with different durations by trained functional generators. **b**: A final F0 contour (yellow) is predicted by a superposiiton of multiple functional contours. The original F0 contour is in green. (Figures are digested and reproduced from Bailly & Holm (2005).)

The realisation of surface F0 contours in the SFC model is achieved by *superposition* of learned functional contours. Superposition means that a final F0

contour is obtained by overlaying one functional contour onto another, which is different from previous linear models. Syllable is the basic modelling unit in the SFC model, and each syllable is described by three F0 values (at 10%, 50%, and 90%) and a duration value. In Bailly & Holm (2005), each metalinguistic function is assigned a neural network as generator. The generators are trained on a database labelled with functions and able to produce the globally learned functional contours during synthesis. Figure 2.6a shows two sets of contours synthesised by generators of 'assertion' (DC) and 'questions' (QS), respectively. To produce an utterance, the synthesised functional contours are superposed one by one subject to different scopes. Figure 2.6b shows such an example. The SFC model has been applied for F0 modelling in TTS systems for several languages including French (Bailly & Holm, 2005), German (Bailly & Gorisch, 2006) and Mandarin Chinese (Chen, Bailly, Liu & Wang, 2004).

2.2.3 Discussion

In contrast to phonological models, phonetic models are quantitative and parametric, and they make F0 modelling a direct data-driven task for the first time so that statistical learning algorithms can apply. The Tilt model, for example, actually stands between categorical representations of phonology (e.g. tones in the AM theory) and surface F0 contours. Therefore, the F0 modelling process for TTS also becomes two-staged. At the first stage, phonological and other linguistic representations can be mapped to tilt parameters. At the second stage, tilt parameters can be mapped to F0 contours through F0 generation formulation defined by the Tilt model. However, it might be risky to transform F0 contours to sets of model parameters. On the one hand, parameters of a phonetic model may become meaningless if they can not be clearly clustered and linked to phonological representations. And on the other hand,

if the mapping between model parameters and linguistic features is too complex, it might be difficult for a statistical algorithm to learn successfully (Sun, 2002).

The phonetic models start to benefit from the data-driven machine learning approach and are able to generate more natural intonation with more variations. However, this approach critically relies on labelled datasets. The labelling process of intonation events in the Tilt model is still a tedious and error-prone task, despite that an automated process is available (Taylor, 1998). Relying on statistical learning algorithms is also risky, very strange F0 contours may also be generated if the dataset is not balanced.

Beside that, F0 modelling based on phonetic models may not be accurate enough. For example, the Tilt model only cares about the recognised intonation events in an F0 contour with transitions between them either ignored or simply filled by linear interpolations, which makes it lose a great amount of details in the original contour. Moreover, it is reported that we are perceptually sensitive to the alignment between pitch movements and syllables or segmental boundaries (Kohler, 1990; d'Imperio & House, 1997; van Santen et al., 2008), but the Tilt model is not strictly aligned to these boundaries. Although the SFC model is syllable-based, its superpositional modelling manner makes it difficult to manipulate surface F0 contours effectively or achieve very high modelling accuracy.

2.3 Articulatory Models

2.3.1 The Fujisaki model

The Fujisaki model (Fujisaki & Hirose, 1982; Fujisaki, 1983; Fujisaki & Hirose, 1984) is also a superpositional model of intonation. However, the aim of the model is not only to describe F0 contours accurately, but also to simulate some aspects of the physiological process of speech production, i.e. the control mechanism of vocal

fold vibration. The Fujisaki model is composed of a phrase component controlled by *phrase commands* and an accent component controlled by *accent commands*. And the final F0 contour of an utterance is considered to be the result of accent components overlaying onto phrase components on a logarithmic scale.

Figure 2.7 shows an overview of the Fujisaki model. The model works on the logarithmic domain of F0.[1] The overall F0 contour shape of an utterance is characterised by the phrase commands, which are modelled as pulses leading to local F0 maxima followed by slow decays. The contours resulted from phrase commands are connected sequentially. Detailed pitch accents (rises and falls) are realised by accents commands, which are modelled as step functions. The resulting accent contours are then added on to the phrase contours to form the final F0 contour.

The Fujisaki model was originally developed for Japanese, but soon became popular and used for many other languages with improved controllability (Mixdorff, 2004). The determination of command positions is a persistent issue in the Fujisaki model, which prevents the Fujisaki model from being effectively trainable or used for large scale TTS systems. Several attempts (Mixdorff, 2000; Hirose et al., 2005; Kameoka, Yoshizato, Ishihara, Kadowaki, Ohishi & Kashino, 2015; Torres & Gurlekian, 2016) were made to learn the parameters in the Fujisaki model from large corpus automatically and reliably.

2.3.2 The STEM-ML model

The STEM-ML (Soft TEMplate Markup Language) model (Kochanski & Shih, 2000; Kochanski & Shih, 2003) is also an intonation model built from the perspective of physiology. However, its primary goal is not to simulate the physical mechanism of vocal fold vibration. Instead, the model strives to find the relationship

[1] This is because human hearing perception (and all other senses) works on a logarithmic scale (Fechner, 1966). As a consequence, most quantitative F0 models work on the logarithmic domain.

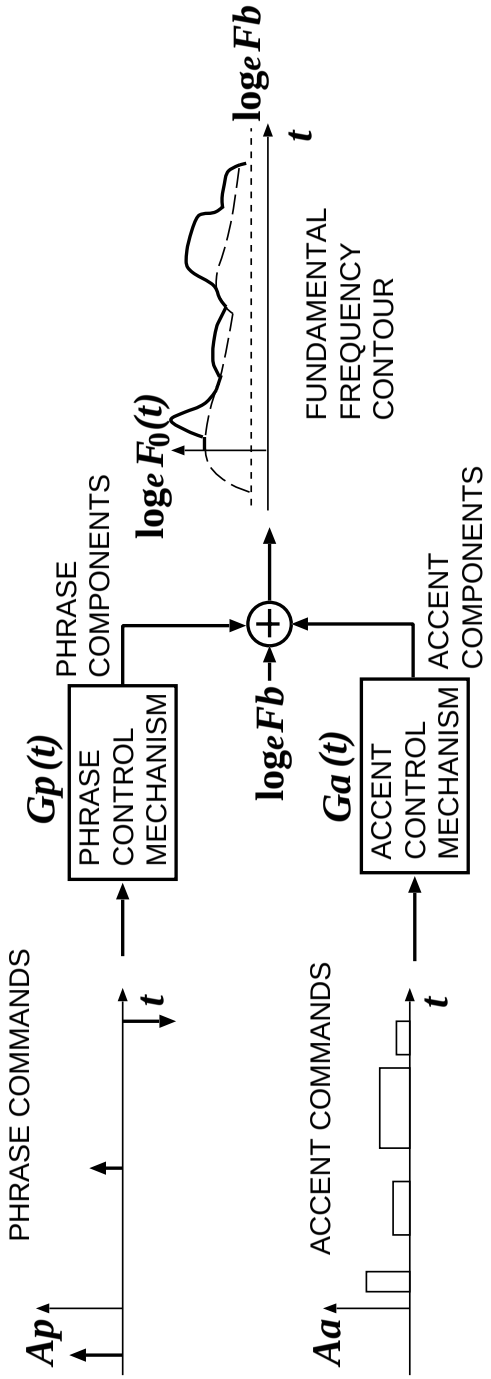


Figure 2.7 Overview of the Fujisaki model (Fujisaki, 2004).

between articulatory effort and prosodic ambiguity based on the assumption that there is always a balance between the two. The key assumptions made by the model can be summarised as (van Santen et al., 2008):

- Syllables are preplanned during speech production and there are *soft templates* underlying them.
- Speakers are able to balance between the articulatory effort they are making and the possible prosodic ambiguity caused. Namely, at critical intonation events a speaker may maximise his articulatory effort because these events convey important information that should be delivered clearly (high cost of ambiguity), whereas at less important intonation events a speaker may lower his articulatory effort to a certain degree because that is harmless to communication (low cost of ambiguity).

The overall F0 contour of an utterance is considered as a sequence of connected local pitch accents, represented as the abovementioned soft templates that can be learned from a dataset. 'Soft' implies that certain distortions are allowed in the case that some 'interactions' occur at the joins of connected templates. Note that a template can be affected by both the preceding and the following templates so that both carry-over and anticipatory coarticulations are considered. A tagging system is proposed in Kochanski & Shih (2003) to label pitch accents for the model. A key parameter in the model is *strength*, which effectively controls pitch accent shape. Strength is a correlate of articulatory effort. If its value is high, which suggests high cost of ambiguity so that high articulatory effort is made, the accent template remains unchanged and is fully realised as the surface F0 contour; if its value is low, which suggests the cost of ambiguity is low so that a compromise is made and a surface F0 contour that deviates from the template is realised. In other words, the STEM-ML model simulates F0 contours as deviations from underlying accent

templates under the influence of surrounding accents (Xu, 2015). The STEM-ML has been experimented in TTS systems for Mandarin Chinese (Kochanski & Shih, 2001), Cantonese (Lee, Kochanski, Shih & Li, 2002) and English (Shih & Kochanski, 2003).

2.3.3 Discussion

Our knowledge about the physiological process of speech production is rather limited, which makes current articulatory models immature. On the one hand, observations on glottal excitation and vocal tract resonance are already well studied. But on the other hand, the motor control and coordination mechanisms of laryngeal movements still remain unclear. Under this background, the Fujisaki model is actually an incomplete articulatory model with vocal fold vibration well thought and simulated but with speech timing less considered and open to be explored. As summarised by Kameoka et al. (2015), automatic estimating parameters of the Fujisaki model from raw F0 contour has been a long-term difficulty since both levels and timings of its phrase and accent commands need to be explored at the same time. To address this issue, they approximated the deterministic formulation of the Fujisaki model with an HMM-based discrete-time stochastic translation so that automatic parameter estimation becomes possible through the iterative expectation-maximisation (EM) algorithm (Dempster et al., 1977; Feder & Weinstein, 1988). Nevertheless, this valuable solution can only be seen as a bypass with the core timing issue faced by the Fujisaki model unchanged.

In terms of the template based STEM-ML model, its main advantage is the ability to model the anticipatory coarticulation while most other models can only handle the carry-over coarticulation. Namely, when a template is produced, it is not only affected by the past template but also by the future as long as there is a following template available. However, it is usually not easy to define a reasonable

number of templates and successfully relate them to phonological representations at the same time. Moreover, from our perspective, the STEM-ML model cannot really be seen as an articulatory model since it only focuses on the effort and resulting contour variations of F0 production with the whole physiological process largely ignored.

2.4 Summary

A number of F0 models were introduced in this chapter. The phonological models are theoretically successful in the sense that they inspired extensive research in the field and helped us gain a better understanding of intonation. The phonetic models are practically influential since they are relatively easy to implement and have been shown useful in real TTS applications. The articulatory models are not only useful but also can be seen as milestones along our way towards complete articulatory speech synthesis systems.

From a modelling point of view, whether these models are necessary in real TTS applications is currently debatable. The main reason is that all these models are either problematic to train or not accurate enough to be reliably used to handle F0 synthesis alone. Given that state-of-the-art machine learning techniques can already synthesise speech at the frame level without relying on any speech production model, nowadays F0 contours generated by these models can only be considered as constraints to aid F0 modelling in TTS.

However, from our perspective, human speech production involves systematic motor movements and surface speech signals are simply acoustic realisations of underlying motor movements. Instead of being used to constrain and aid modelling, a mature model of speech production should be articulatory-oriented, faithful to the reality and accurate enough. It should be capable of dominating the modelling

process and generating reliable speech signals on its own. This thesis on articulatory F0 modelling is an initial attempt in this direction.

Chapter 3

The Target Approximation Model

3.1 The PENTA Framework

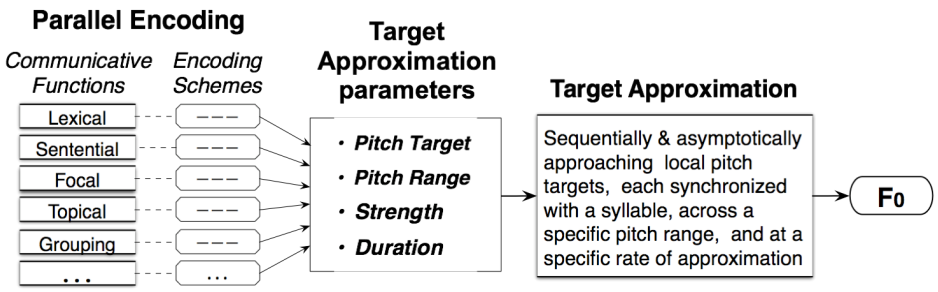


Figure 3.1 The PENTA framework (Xu, 2005).

The target approximation (TA) model is the core component of the parallel encoding and target approximation (PENTA) framework (Xu, 2005), serving as its dynamic F0 generator. As illustrated in Figure 3.1, the PENTA framework treats F0 modelling from an articulatory-functional view of speech, in which multiple communicative functions can be encoded in parallel as joint contributors to a single invariant underlying pitch target for each syllable. Note that this functional approach is different from the one in the SFC model (Section 2.2.2), where each function has an independent contour.

As shown in Figure 3.1, the PENTA framework assumes that the input into the F0 production process should be well-defined communicative functions with specific meanings but no phonetic specifications. Figure 3.2 displays an example of functional annotations in the PENTA framework, in which *essential function combinations* are obtained and their parameters can then be learned through stochastic learning (Xu & Prom-on, 2014).

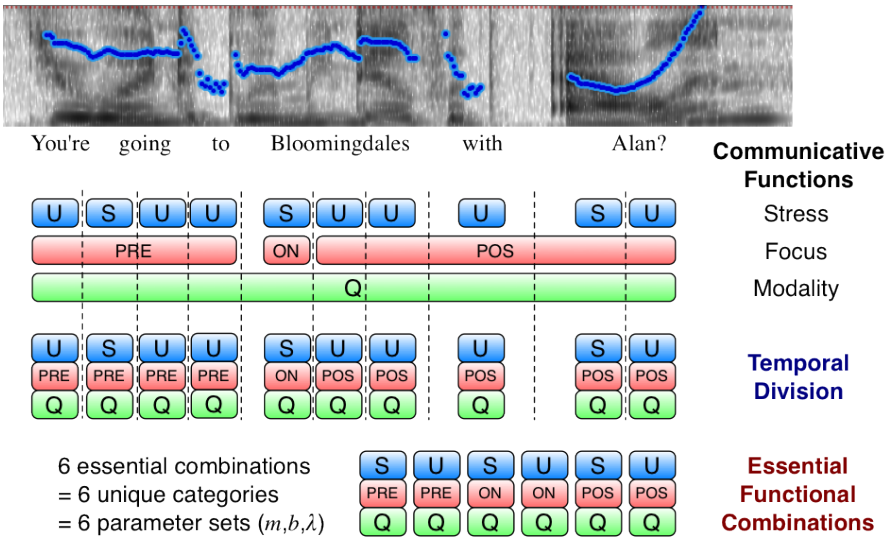


Figure 3.2 Functional annotations in the PENTA framework (Xu & Prom-on, 2014). In the 'Stress' layer, S and U denote stressed and unstressed syllables, respectively. In the 'Focus' layer, PRE, ON and POS denote pre-focus, on-focus and post-focus, respectively. Q denotes question in the 'Modality' layer.

In current TTS systems, however, only some of the common contextual linguistic features are clear-cut communicative functions, while many others are purely contextual. In order to set up a universal platform to compare the proposed TA-based approach with other typical SPSS approaches, the application of the TA model in this thesis does not exactly follow PENTA's strict functional assumptions, but rather makes use of all the common contextual linguistic features in current TTS systems, at least as the starting point. Later on in this thesis, the reduction of the number of contextual linguistic features needed turned out to a move in the direction of getting closer to the functional assumption of PENTA.

3.2 Syllable-based Modelling

In conventional SPSS approaches, all the acoustic features (e.g. MGC, F0 and duration) are jointly modelled at the HMM state level. For example, an F0 contour is formed by a sequence of signal frames generated separately subject to learned state-level Gaussian distributions. In recent deep learning approaches, frame-level acoustic features are treated as the targets to be learned and predicted by DNN or RNN. In a recent study, although hierarchical F0 modelling was implemented with DNN, frame-level F0 residual modelling was still used (Yin et al., 2016).

However, human speech production is unlikely to be coordinated at such a small time scale. Rather, as a complex motor control behaviour, speech production is organised on a serial gesture basis with much longer host units (MacNeilage, 1970; Saltzman & Munhall, 1989; Levelt, Roelofs & Meyer, 1999; Xu & Liu, 2006). It is further explained in the frame/content theory (MacNeilage & Davis, 1993; Davis & MacNeilage, 1995; MacNeilage, 1998) that the constitution of syllable is associated with the cyclicity of continual rhythmic mouth open-close alternation, which actually evolves from ingestive cyclicities (e.g. chewing). And 'much of the patterning of infant babbling is a direct result of production of syllabic "frames" by means of rhythmic mandibular oscillation', 'intra-syllabic and inter-syllabic "content" of the syllable-like cycles' are gained through social interaction at a later stage. Furthermore, prosody studies suggest that characteristics of F0 are suprasegmental (e.g. lexical tones hosted by syllables in tonal languages, stress, focus and other intonation patterns), which are encoded either superpositionally (Fujisaki, 1983) or in a parallel manner (Xu, 2005; Xu, 2007).

Among the existing F0 models, a number of them are syllable-based. For example, in concatenative synthesis Black & Hunt (1996) used linear regression to predict three target F0 values (start, mid-vowel and end) for every syllable with ToBI annotations (Pierrehumbert & Hirschberg, 1990; Silverman et al., 1992)

representing stress and syllable position as features. The SFC model (Section 2.2.2) assumes that each syllable-level F0 contour can be obtained from superposition of multiple simple functional contours and its practical application in Mandarin Chinese showed positive results (Chen et al., 2004). Reddy & Rao (2013, 2016), in their use of the Tilt model (Taylor, 1992; Taylor, 2000) in concatenative synthesis treated each syllable as an intonation event with the tilt parameters extracted from each syllable. Empirically, models using syllable as the basic modelling unit often produce more natural sounding F0 contours (Sun, 2002; Raidt, Bailly, Holm & Mixdorff, 2004).

As discussed by Xu & Prom-on (2015), from the perspectives of both motor control of articulatory movements and the acquisition of speech production skills, syllable plays an important role of reducing the degrees of freedom (DOF) by synchronising multiple articulatory movements. Xu & Liu (2006) proposed that, because of its much slower speed than segmental movement, pitch articulation has to use the entire syllable as its temporal domain of execution. This also means that within each syllable, F0 in adjacent time frames are highly correlated with each other, with a very strong right-to-left dependency that could be captured by a mechanical model such as target approximation, as will be explained next.

3.3 Formulation

The development of the TA model was inspired by empirical findings about tonal dynamics. The basic concept of the TA model (Xu & Wang, 2001), shown in Figure 3.3, is that continuous surface F0 contours are the results of successive, non-overlapping articulatory (laryngeal) movements, each approaching an underlying target associated with a host syllable.

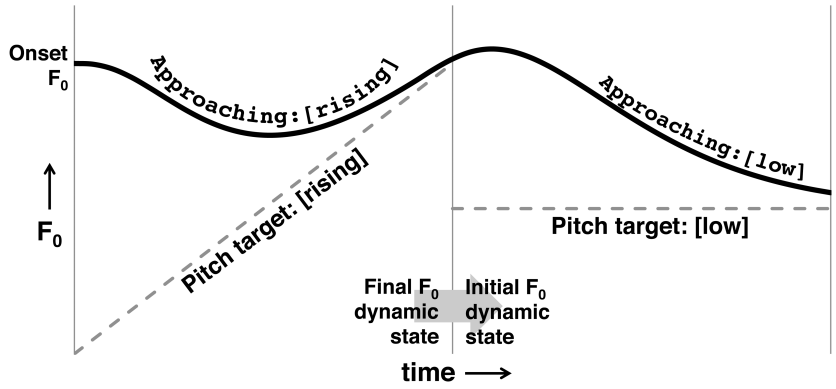


Figure 3.3 The target approximation model.

The TA model was qualitative at the time of its proposal (Xu & Wang, 2001). The concept of the TA model is then algorithmically implemented as the quantitative target approximation (qTA) model by Prom-on et al. (2009). In this model, a target can be either static or dynamic, which can be represented by a simple linear equation

$$x(t) = mt + b, \tag{3.1}$$

where m and b represent the spatial properties of the target in terms of target height and slope, respectively, and t is time relative to the onset of the host syllable.

The realisation of the target is through a third-order critically damped linear system defined by the following equation

$$f_0(t) = x(t) + (c_1 + c_2 t + c_3 t^2)e^{-\lambda t}, \tag{3.2}$$

where $f_0(t)$ is the complete form of the fundamental frequency in semitones, $x(t)$ is the forced response and the polynomial and the exponential are the natural response. λ is the rate of target approximation, i.e., how rapidly the target is approached, which controls the strength of target approximation movement. The transient coefficients c_1, c_2 and c_3 are jointly determined by the initial F0 dynamic state of the syllable, consisting of F0 level, velocity as well as acceleration transferred from

the offset of the preceding syllable (as such they are not free parameters):

$$c_1 = f_0(0) - b, \tag{3.3}$$

$$c_2 = f_0'(0) + c_1\lambda - m, \tag{3.4}$$

$$c_3 = (f_0''(0) + 2c_2\lambda - c_1\lambda^2)/2. \tag{3.5}$$

At the end of the syllable, the final F0 dynamic state is transferred to the next syllable to become its initial state, which results in a smooth and continuous F0 trajectory across the syllable boundary (Figure 3.3).

In short, the process of F0 production is simulated by the TA model at the syllable level by controlling just three motor parameters (m, b and λ), and this process forms a deterministic 'motor-to-acoustic' mapping.[1]

Previously, TA as an articulatory model has been shown to be highly effective in a variety of F0 modelling tasks, including synthesising utterance F0 contours with stochastic learning (Prom-on et al., 2009; Xu & Prom-on, 2014) and simulating speakers' online compensation in response to the pitch-shifted auditory feedback (Liu & Xu, 2015). Beyond F0 modelling, the core idea of the TA model has been successfully used in simulating consonant-vowel articulatory trajectories recorded by electromagnetic articulography (EMA) (Birkholz et al., 2011) and training an articulatory synthesiser to learn invariant underlying targets of vowels and glides, without speaker normalization, and use them to generate highly natural synthetic speech (Birkholz, 2013).

[1]Note that TA is only an F0 model with inertia, carry-over coarticulation and some other aspects of F0 production effectively simulated. This does not imply a fully active 'motor-to-acoustic' process.

3.4 A Dynamical System

The key feature of the TA model is that it implements an inertia-based deterministic dynamical system with finite dimensions so that it is capable of neatly simulating the dynamical process of in-syllable F0 production as well as cross-syllable transient effects. This means that TA is fundamentally different from mathematical transformation functions, e.g. discrete cosine transform (DCT) as used in other studies (Teutenberg et al., 2008; Wu et al., 2008; Yin et al., 2014, 2016).

As a dynamical system, on the one hand TA exhibits a strong stateful generation nature (i.e. with built-in time dependency), and on the other hand it can be easily controlled by syllable-specific parameters (i.e. underlying targets). The stateful nature of the TA model differs from the independent HMM state based modelling in that it implements a physically plausible, systematic and syllabically uniform modelling. The easy controllability makes it possible to directly manipulate the dynamic process of F0 production on-the-fly (Liu & Xu, 2015). Also, it implies a crucial possibility that TTS systems using the TA model as F0 generator can be effectively and systematically controlled in real time during synthesis. Although this feature will not contribute much to this thesis, it will probably be useful in various adaptation tasks for speech synthesis in the future.

3.5 Case Study: Target Distributions of Mandarin Tones

From a modelling point of view, a model can be considered valuable only if its parameters are learnable. Namely, in our case, the prerequisite that different tones and intonations can be accurately learned and predicted by machine learning algorithms is that the underlying targets of them are able to exhibit clear patterns. To

this end, TA target parameters of a Mandarin Chinese spontaneous speech dataset are extracted and examined in this section.

The dataset consisted of 641 statements and 387 questions recorded from a female native Mandarin speaker, and which was also used to train and test different F0 modelling systems in Chapter 5. Syllable segmentations were derived from phone segmentations obtained through force alignment. Syllables between 100 ms and 350 ms in length were used in this study since they were seen to be more stably produced. Shorter or longer syllables may contain more creaky or breathy effects so that interpolation and smoothing are necessary, but such interventions should be avoided in this study. Note that what we want to see here are the targets extracted directly from raw F0 contours. Optimal target parameters were extracted through the Levenberg-Marquardt nonlinear least-squares method (Moré, 1978) provided by the LIMFIT Python package (Newville, Stensitzki, Allen & Ingargiola, 2014). More details of the extraction process are provided in Section 5.5.2.

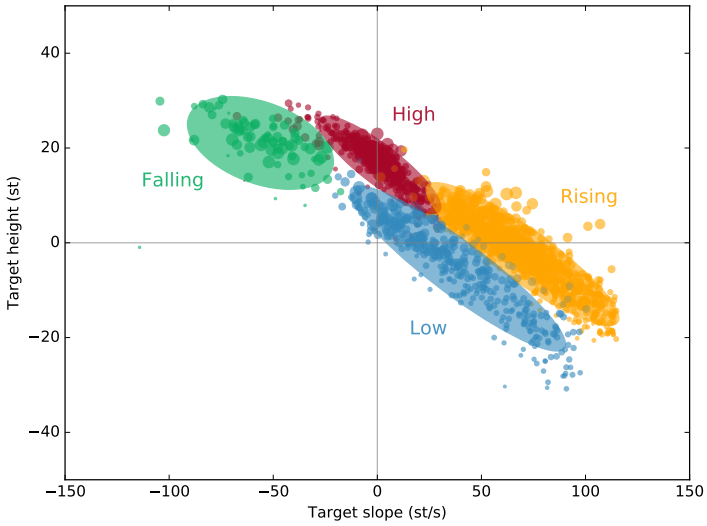


Figure 3.4 Two-dimensional display of the extracted target parameters of the four tones in the **statement** set. The X and Y axes represent target slope and target height. Circle size represents target strength. Tone clusters are represented by covariance error ellipses with 95% confidence interval.

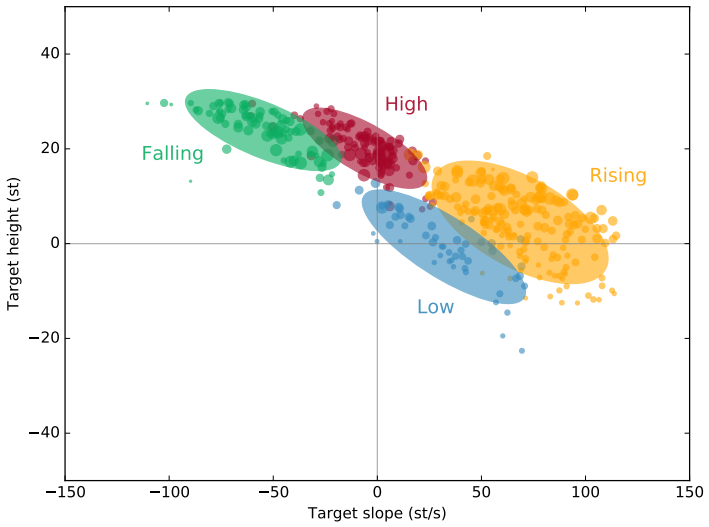


Figure 3.5 Two-dimensional display of the extracted target parameters of the four tones in the **question** set. The X and Y axes represent target slope and target height. Circle size represents target strength. Tone clusters are represented by covariance error ellipses with 95% confidence interval.

Four basic Mandarin tone types were considered in this study, which are High, Rising, Low and Falling. The number of the four tones labelled as statement are 600, 1094, 390 and 77, respectively. And the number of the four tones labelled as question are 106, 176, 51 and 80, respectively. Note that only part of the syllables in the question set were considered as question-related and labelled as question. As shown in both Figure 3.4 and Figure 3.5, target parameters of each tone are clearly clustered and the tone clusters are generally separated from each other. This demonstrates that there are clear TA target patterns with respect to the four tones in Mandarin Chinese.

In terms of intonation difference, tone targets extracted from the statement set and those from the question set were compared. All the three target parameters (slope, height and strength) were compared in pairs between statement and question. Table 3.1 shows statistical results of the comparison, in which p values were obtained through the Wilcoxon rank-sum test (Mann & Whitney, 1947). We found that target heights of all the four tones in the question set are statistically higher than

Table 3.1 Statistical comparison of tone targets between the statement (S) set and the question (Q) set. Parameters in the question set that are significantly different from those in the statement set are in bold.

Tone	Target Slope			Target Height			Target Strength		
	S-Mean	Q-Mean	p	S-Mean	Q-Mean	p	S-Mean	Q-Mean	p
High	1.2	$\mathbf{-6.1}$	2.0×10^{-5}	16.3	**20.1**	2.4×10^{-14}	28.0	**32.6**	1.2×10^{-5}
Rising	69.7	68.6	0.69	-2.4	**4.5**	1.9×10^{-29}	25.6	**32.9**	1.5×10^{-7}
Low	38.6	32.0	0.18	-5.5	$\mathbf{-0.7}$	7.3×10^{-4}	23.3	24.4	0.18
Falling	-56.1	-54.9	0.70	21.1	**24.0**	3.1×10^{-4}	40.3	37.7	0.27

those in the statement set. This suggests that TA targets of all the four tones need to be raised to a higher level in order to realise higher intonation contours to convey interrogative meanings, which stays in line with previous studies on tone-intonation relations in Mandarin Chinese (Shen, 1990; Yuan, 2006; Liu & Xu, 2006). Besides that, target approximation strength values of High and Rising tones in the statement set are also significantly higher than those in the statement set, which suggests more strength is needed for these two tones when producing questions. The significant decrease of High tone slope in the question set is unexpected. However, according to Liu & Xu (2014), target slope is a less sensitive (or stable) parameter than the other two and such change in it may not lead to much surface pitch contour deviation.

3.6 Summary

Details of the quantitative implementation of the TA model were introduced in this chapter. Compared to other existing models, the TA model has at least three unique features (Xu, Lee, Prom-on & Liu, 2015): a) unitary dynamic targets, which are different from contour targets as in the SFC or STEM-ML models; b) unidirectional sequential target approximation (no overlap of movements as in the task dynamic model (Saltzman & Munhall, 1989) or return phase in a movement as in the Fujisaki model); c) high-order state transfer across target approximation movements, which overcomes one of the weaknesses of the dynamic-system model proposed by Ross & Ostendorf (1999) in producing complex contours. Besides that, full synchrony of pitch targets with syllables serves as the core assumption about speech articulation of the entire PENTA framework, which should always be highlighted.

The TA model is based on our latest understanding of the physical process of pitch control and has made a salient contribution to the field. However, it might

still be ambitious to say that the model is correct on every aspect or can be directly applicable to other languages. There were a serial debates between the TA model and the phonological school of intonation modelling (Arvaniti & Ladd, 2009; Xu et al., 2015; Arvaniti & Ladd, 2015), in which several arguments were raised and a number of caveats of the TA model were discussed.

First of all, Arvaniti & Ladd (2009) held the assumption that 'not every syllable has to have specification for pitch', which was their major disgreement with the TA model. And similar assumptions can also be found in other models like Fujisaki (1983) and Hirst (2005). This assumption is based on the fact that in non-tonal languages (e.g. English and Greek), many syllables appear unspecified for pitch. Namely, there are syllables in these languages that are neither stressed nor prominent, but exhibit high F0 variability. The Fujisaki model addresses this issue by assigning a single command to the time interval of a string of unstressed syllables. Whereas the TA model treats every syllable fully specified with its own target. Although no consensus has been made, the TA model actually benefits from this syllable-synchrony assumption and effectively escapes from the unfixed timing issue as practically faced by other models.

The second challenging argument is that for utterances with very different lengths their linguistic functions can be very similar, which might be problematic for the TA model. At this point, there are two possible situations. One is that different utterance lengths are caused by different number of syllables, and the other is that the long utterances contain prolonged syllables. For the former situation, contextual information such as syllable position in phrase/sentence turned out to be very helpful in our previous implementations (e.g. Xu & Prom-on (2014)), although such information seems not 'functional'. For the latter situation, while we have to admit this is a drawback and the TA model cannot handle it well with its original

design, a temporary solution is to treat the prolonged syllables as special cases and assign them two consecutive targets.

The third is the lack of additional articulatory mechanisms in the current TA model. As also mentioned by Xu & Prom-on (2014), coarticulation mechanisms like anticipatory raising (Gandour, Potisuk & Dechongkit, 1994; Potisuk, Gandour & Harper, 1997; Xu, 1999), post-low bouncing (Chen & Xu, 2006) and consonantal perturbation (Silverman, 1986) require further research and extra components may need to be introduced.

In addition to the above mentioned, we also found some other practical issues when applying the TA model to real TTS systems. Discussions of them are provided in Chapter 6.

Chapter 4

Motor-to-Acoustic: Simulating Online Auditory Feedback Compensation with TA

Speech production is one of the most complex human motor behaviours involving highly precise coordination of various articulatory movements. From a great number of empirical studies, many interesting phenomena have been found and some of which exhibit a very high complexity in terms of spatiotemporal variations. As introduced in the background section below, online feedback compensation is one of such cases. The aim of this chapter is to investigate how articulation is controlled online in response to the shifted auditory feedback during pitch production. Different from other studies targeting at complete sensorimotor control mechanisms (e.g. the state feedback control framework by Houde & Nagarajan (2011)), this study focuses on the details of resulting motor changes when control commands are given. Namely, our focus is to study how reactive surface pitch perturbations are formed based on the assumption that pitch movements are realisations of underlying target approximation processes.

4.1 Background on Auditory Feedback

4.1.1 Offline learning and the DIVA model

The study of auditory feedback in speech production can be traced back to early studies of Lombard in 1911. Lombard (1911) and later research by Lane & Tranel (1971) showed that when speakers' hearing ability was blocked by noise, they were still able to produce normal speech. If the volume of noise was increased, speakers would increase their volume accordingly. Later, Cowie, Douglas-Cowie & Kerr (1982) and Lane & Webster (1991) reported their findings that even after years of deafness, adult speakers were still able to produce intelligent speech.

As a contrast, children who were pre-lingually deaf could not manage to learn how to speak (Ross & Giolas, 1978; Oller & Eilers, 1988; Raphael, Borden & Harris, 2011). Through extensive comparison of vocal development in deaf and hearing infants, Oller & Eilers (1988) showed that the deaf infants' ways of babbling are different from those of the hearing infants. This study showed that the traditional belief that auditory feedback plays only a minor role in the babbling stage of speech acquisition is erroneous. Instead, auditory feedback is critical for children's development of speech acquisition through its interaction with babbling. This notion is also supported by other early studies (Smith, 1975; Borden, 1979; Osberger & McGarr, 1982). More recently, Koopmans-van Beinum, Clement, Den Dikkenberg-Pot et al. (2001) classified early infant vocalizations by using canonical babbling as the cue and traced the lack of auditory perception in deaf infants. Their analysis revealed that auditory feedback is a prerequisite for the coordination of articulatory system and the subsequent development of speech acquisition. Nevertheless, this finding does not support that auditory feedback plays an equally important role in adulthood. However, as we will see in the next section, auditory feedback does affect adult speech production in a subtle way.

Later, based on these findings some articulatory synthesis models were proposed to simulate the early stage of human speech acquisition. For example, Kröger, Kannampuzha & Neuschaefer-Rube (2009) trained a neurocomputational model of speech production and perception with a computer-implemented vocal tract, capable of producing and perceiving isolated vowels, vowel-consonant (VC) and consonant-vowel (CV) syllables. The virtual infant KLAIR (Huckvale, Howard & Fagel, 2009; Huckvale, 2011) that simulates spoken language acquisition through imitative interactions between a virtual infant and its caregiver. The Elija model (Howard & Messum, 2011) that avoids the imitative mechanism by simply associating Elija's discrete motor actions to caregiver's natural responses with a designed rewarded exploration. What is common in these models is that the produced speech signals are perceived via auditory feedback by the infant so that they can be paired with corresponding speech actions to train a feedforward speech production model.

A more complete and influential model developed based on the babbling-feedback interaction is the DIVA[1] model (Guenther, 1995; Guenther & Perkell, 2004; Guenther, Ghosh & Tourville, 2006; Tourville & Guenther, 2011). Taking the speech-related brain activation patterns observed in the functional magnetic resonance imaging (fMRI) experiments into consideration, the DIVA model tries to simulate the process of learning speech sound production through babbling and imitation. It is designed to reflect the integration of auditory, somatosensory and motor information represented in the cerebral cortex during speech production.[2]

The DIVA model originates from a concept proposed in Guenther (1995) called *speech sound map*, which is the key component in this neural model. Phoneme, syllable, or even syllable sequence are all accepted as 'speech sound', among which syllable is treated as the most typical unit represented by a single 'model neuron' in

[1] DIVA stands for 'Directions Into the Velocities of the Articulators' according to Cai (2012).

[2] Strictly speaking, bone-conducted feedback is missing in all the above-mentioned models. Please refer to Section 4.1.3 for more detailed discussion.

the speech sound map. Note that in DIVA experiments, text scripts are carefully designed and produced speech sounds are normally controlled to a fixed length, so that syllable boundaries remain relatively stable and the segmentation issue is avoided. Therefore, the model neurons can always be mapped to its corresponding speech signal correctly. The model neurons are abstract, but they are assumed to correspond to a small population of real neurons in the cortex. The activation of such model neurons leads to motor commands driving two control subsystems.

- *Feedforward control system*: learning to project directly from the articulatory control units to the speech sounds
- *Feedback control system*: mapping detected sensory errors into corrective motor commands during learning

At the babbling stage, the feedforward system simply generates random and reduplicated motor, auditory and somatosensory information, by which the feedback control system is activated to tune the projections in the speech sound map according to the detected sensory errors. The two control systems are weighted and trained interactively at the same time during the babbling stage. Namely, when one system is more weighted than the other, the babbling trial will be dominated by it and work more like a training process for production or for motor commands correction. The purpose of this stage is to simulate the infant babbling stage, during which the motor-sound pairs obtained through random self-production can be collected as training data to train both the feedforward and feedback control systems.

At the imitation stage, syllable-specific learning occurs when an infant is presented with a new speech sound to learn. The model first learns an *auditory target* for that new sound, represented as a time-varying acoustic signal. This leads to an activation of previously unused speech sound map neurons for that sound. Then, the projections between the model neurons and corresponding motors are tuned via the

feedback control system. As described by Tourville & Guenther (2011), the readout of the feedforward system will result in auditory errors, and the system must employ the auditory feedback control system to transform auditory errors into corrective motor commands via the feedback control map. This map (the transformation from auditory errors to corrective motor commands) will be less functional when the DIVA model reach a 'mature' state. As highlighted in Guenther et al. (2006), after the babbling stage the DIVA model is capable of producing a typical syllable with perceptually negligible amount of acoustic errors within as few as 6 imitation trials.

An important note is that the DIVA model is based on the hypothesis that speech movements are planned. As a consequence, the learning process of the DIVA model is totally *offline*, i.e. it adopts the feedback control map to tune the feedforward map based on the errors in the auditory feedback collected after each production.

The learning paradigm established by the DIVA model is believed to be close to the reality of infant speech learning (Guenther & Vladusich, 2012). However, the actual performance of the DIVA model in terms of perceptual quality is not very impressive. Currently, it can only produce steady-state vowels with acceptable intelligibility but low naturalness. The reason is that the DIVA model lacks a dynamic speech production module to make multisyllabic articulation. More discussions on the DIVA model are provided in the following sections.

4.1.2 Online compensation

After speaking skill is gained, people can speak normally without relying on immediate auditory feedback. However, there is evidence showing that auditory feedback still affects adult speech production. In the past decades, a number of studies have shown that auditory feedback plays an important role in the online control of speech F0. These studies investigated feedback response in various tasks, including singing (Natke, Donath & Kalveram, 2003), glissando (Burnett & Larson, 2002), sustained

vowels (Hain, Burnett, Kiran, Larson, Singh & Kenney, 2000; Larson, Burnett, Bauer, Kiran & Hain, 2001; Bauer & Larson, 2003; Sivasankar, Bauer, Babu & Larson, 2005), prolonged vowels (Jones & Munhall, 2000; Jones & Munhall, 2002), nonsense syllables (Natke & Kalveram, 2001; Donath, Natke & Kalveram, 2002) and normal speech (Elman, 1981; Kawahara, 1993; Burnett, Freedland, Larson & Hain, 1998; Larson, 1998; Xu, Larson, Bauer & Hain, 2004; Chen, Liu, Xu & Larson, 2007; Chang, Niziolek, Knight, Nagarajan & Houde, 2013). In the experiments of these studies, unexpected pitch shifts were applied *online* to the voice of the human subjects before being fed back to their ears. In response to the pitch shifts, subjects involuntarily made compensatory F0 adjustments (in the opposite direction of pitch shift) with short latencies (100-150 ms on average according to Xu et al. (2004) and Liu & Larson (2007)). Meanwhile, other studies found similar feedback compensation in formants (Gracco, Ross, Kalinowski & Stuart, 1994; Houde & Jordan, 1998; Houde & Jordan, 2002; Purcell & Munhall, 2006a; Purcell & Munhall, 2006b; Villacorta, Perkell & Guenther, 2007; Tourville, Reilly & Guenther, 2008; Munhall, MacDonald, Byrne & Johnsrude, 2009; MacDonald, Goldberg & Munhall, 2010; Katseff, Houde & Johnson, 2012; Cai, 2012; Niziolek, Nagarajan & Houde, 2013).

It is worth noting that, based on experimental findings, Houde & Jordan (1998), Houde & Jordan (2002), Jones & Munhall (2000) and Jones & Munhall (2002) further argued that a long-term effect is retained beyond the course of such online compensation, and speakers' sensorimotor mapping between articulatory motor space and acoustic space is adapted upon it. Their research findings suggest that there is an online learning mechanism through auditory feedback compensation. And such online learning not only applies to adults, but also should occur when infants learn to speak. That is, online learning occurs when they are imitating their parents. However, recent studies on auditory feedback compensation in children

by van Brenk, Terband & Cai (2014), Terband, van Brenk & van Doornik-van der Zee (2014) and Terband & van Brenk (2015) suggest that the proportion of subjects showing compensatory response is smaller in the child group (aged 4-9 years) than in the young adult group (aged 19-29 years). Moreover, the experiments on 2-year-olds (toddlers) and 4-years-olds by MacDonald, Johnson, Forsythe, Plante & Munhall (2012) showed that the 2-year-olds did not response to the auditory perturbations at all. The 4-years-olds, on the other hand, responded to the perturbations but with a larger token-to-token variability than adults. Although limitations remain (e.g. experiment results on very young children may not be always reliable), these findings suggest that the online compensation behaviour may be developed just around the age of 4 years. More importantly, online compensation is possibly not involved in the early learning stage of speech acquisition (babbling) and instead it is only used by more mature speakers as a mechanism of vocal behaviour maintenance (Brainard & Doupe, 2000).

Characteristics of auditory feedback compensation can be summarised as follows:

- **General existence**: Auditory feedback compensation not only occurs for fundamental frequency (F0) but also for formants. In terms of F0 studies: Cowie et al. (1982) first provided clinical evidence showing that without the presence of auditory feedback, the control of pitch deteriorates soon after patients' deafness; from the perspective of neuroscience, Guenther et al. (2006) and Chang et al. (2013) showed that there is a sensorimotor cortical network underlying auditory feedback-based control of vocal pitch; Liu, Russo & Larson (2010) and Liu, Chen, Jones, Huang & Liu (2011) reported more findings about age-related differences in compensation of F0 feedback perturbation and Chen, Liu, Jones, Huang & Liu (2010) reported sex-related differences. In terms of formant studies: Gracco et al. (1994) demonstrated that the speak-

ers will change their articulation to shift the spectrum in speech according to the spectral shifts introduced in their auditory feedback; furthermore; Houde & Jordan (1998) showed that control of the production of vowels adapts to perturbations of auditory feedback; Cai (2012) reported differences between stuttering and normal speakers in formant compensation.

More recently, Niziolek et al. (2013) provided evidence showing that auditory feedback compensation not only exists in altered speech, but also during normal speech production. Through magnetoencephalographic imaging (MEG-I) experiments, they reported findings suggesting that compensatory mechanism is also employed in natural, unaltered speech, and less-prototypical utterances. This kind of utterances actually make up a large proportion of natural speech, which are processed as containing potential errors. And feedback-driven speech error correction is occurring to correct these potential errors constantly on a small scale.

- **Online vs. delay**: There is a debate about whether online compensation can actually happen since the delay of auditory feedback may exceed the temporal domain of the target linguistic units (e.g. syllables). In Mandarin speech, Xu (1997) and Xu (1999) reported that the mean duration of a simple consonant-vowel (CV) syllable is about 180 ms. In some studies, the observed compensation response latencies were about 150 ms in experiments when German speakers compensate for the mismatch between intended and feedback pitch during production of nonsense syllables (Natke & Kalveram, 2001; Donath et al., 2002; Natke et al., 2003), which is shorter than the reported mean syllable duration. However, Jones & Munhall (2002) found that Mandarin speakers' compensatory changes in voice F0 are around 200 ms, which is longer than the mean syllable duration. Therefore, some researchers were convinced that these latencies were too slow to allow speakers to control

F0 effectively within single syllables. Instead, the compensations should occur on a suprasegmental level in the context of prosody.

On the other hand, in studies for nonspeech tasks, reported response latencies were as short as 76 ms for pitch compensation (Burnett & Larson, 2002), 114 ms (Hain et al., 2000) and 130 ms (Larson et al., 2001) for sustained vowel compensation. Xu et al. (2004) analysed the reasons that had caused the discrepancies in previous experiments and designed improved F0 compensation experiment in Mandarin demonstrating that the majority of the compensatory changes occurred significantly sooner (143 ms) than the mean syllable duration (180 ms). In some conditions, latencies were short enough (130 ms) for the response to correct for perturbations within single syllables. In addition, in the above-mentioned MEG-I experiment, Niziolek et al. (2013) provided evidence showing that during normal speech production the auditory feedback is used strictly *online* (e.g. compensation occurs within single vowels) by speakers to correct errors, and they found an online 'vowel centring' phenomenon that spontaneous compensation constantly occurs during vowel production. Note that this vowel centring behaviour was found to be fairly irregular. While its underlying mechanism is still unknown, a reasonable account is that there is a rectification behaviour against the tiny perturbations caused by various factors during speech production.

Despite these debates, it is still worthwhile to investigate how suprasegmental compensation (across syllable boundaries) works. In general, in previous studies as long as a response to perturbation occurs within the course of connected running speech, it is treated as an online compensation. Especially, from the perspective of motor control, the cross-syllable compensation behaviour which requires more complicated underlying sensorimotor adjustments involving longer term effects is more interesting than the in-syllable

compensation. Similarly, Cai, Ghosh, Guenther & Perkell (2011), in their investigation on spatiotemporal complexity of online formant control, studied both in-syllable and cross-syllable phenomena. More details on this research are introduced in the next section.

- **Partial compensation for altered speech**: Partial compensation phenomena can be widely found in Jones & Munhall (2000), Xu et al. (2004), Jones & Munhall (2005) and Liu et al. (2010) for F0, and Houde & Jordan (2002), Purcell & Munhall (2006a) and Pile, Dajani, Purcell & Munhall (2007) for formants. By using a stepwise feedback alteration design, Katseff et al. (2012) reported comprehensive results demonstrating that subjects only partially compensated for experimentally induced changes to their auditory feedback (they never compensate for formant shifts, on average, more than 40%, e.g. a subject whose F1 feedback is raised by 200 Hz will produce vowels with an F1 no more than 80 Hz lower than usual) and they compensated more for small feedback shifts than for larger shifts (compensation was approximately complete for small shifts (50 Hz) in auditory feedback and partial for all shifts greater than 50 Hz).

4.1.3 Stammering research and objections to 'feedback' control

In the field of stammering research, studies on the effects of altered auditory feedback (AAF) have grown in popularity in recent decades. The earliest research can be traced back to Lee (1951) on delayed auditory feedback (DAF) and Howell, El-Yaniv & Powell (1987) on frequency-shifted feedback (FSF). Researchers first found speech effects that are similar to stammering could be observed in fluent speakers when DAF was presented. Lee (1951) refers to this phenomenon as

'simulated' stammer. Later, a number of studies showed influential findings that stammering people were able to improve their speech control when DAF was presented (e.g. Soderberg (1960), Chase, Sutton & Rapin (1961), Neelley (1961), Ham & Steer (1967) and Curlee & Perkins (1969)). DAF-based stammering treatment was initialised by Lee (1951), Cherry & Sayers (1956) and Goldiamond (1965), and then evolved to a therapy programme by Ryan (1974). Howell et al. (1987) managed to present FSF to stammering speakers for the first time and found that it could lead to more fluent speech than DAF, which inspired many later studies on this topic (e.g. Kalinowski, Armson, Stuart & Gracco (1993), Armson, Foote, Witt, Kalinowski & Stuart (1997), Burnett, Senner & Larson (1997), Zimmerman, Kalinowski, Stuart & Rastatter (1997) and Natke, Grosser & Kalveram (2001)). From both scientific and clinical perspectives, tremendous contributions were made by all these empirical studies and which promoted the emergence of stammering treatment devices.

From a theoretical perspective, significant contributions were also made by studies in stammering research. One worth detailed discussion here is the strong objections to the view held by some production models that phonetic content of speech segments is perceived and monitored as 'feedback' by the brain to determine whether there are any errors between the actual speech and the intended so that corrective actions can then be taken. This view is exactly the one strictly followed by the DIVA model as introduced earlier in this section. At least three arguments against this view were summarised by Howell & Sackin (2002) and Howell (2004).

First, Borden (1979) argued against the 'feedback' point of view on two aspects. One is that it takes a great amount of time for the brain to recover and process the information delivered by the feedback. The processing time was estimated to be around 200 ms (more or less the length of a syllable). One question arises here is that, when a segment is being produced, whether this time lag is too long

for its feedback to be actually used to control its own production (as discussed earlier). The other is that adult speakers are still able to speak normally for a period of time after hearing loss, which suggests the absence of online feedback control mechanism.

Second, based on the study by von Békésy (1960) showing that bone-conducted sound is at approximately the same level as airborne sound, Howell & Powell (1984) argued that airborne sound would be masked by bone-conducted sound. And the latter is dominated by the fundamental frequency of speech output since formant details can be easily degraded by resonances of our body structures. As a consequence, the feedback signal may have lost plenty of phonetic details so that cannot provide sufficient information for feedback control.

Third, in the DAF experiment conducted by Howell & Archer (1984), they replaced phonetic content in the feedback with noise and showed that such non-speech noise could achieve equivalent performance as standard DAF. This finding strongly suggests that the feedback speech does not actually go through the speech comprehension system as a high level mechanism. The speech control process may be a low level mechanism instead.

In response to these empirical findings against the previous view of 'feedback', disruptive rhythm hypothesis (DRH) as a non-feedback account was then developed by Howell, Powell & Khan (1983). DRH interprets DAF effect from a rhythmic perspective by suggesting that synchronous activities are much easier to produce than asynchronous ones, no matter whether such activities contain useful information about speech. The disruptions observed in DAF can then be interpreted as consequences of asynchrony caused by the delay. Moreover, it was found that roughly a syllable-sized delay (200 ms) causes maximum disruption on speech control (Black, 1951), which suggests that syllable could be the basic synchronous unit used by speakers.

Subsequently, a complete stammering theory EXPLAN was developed by Howell (2002) and Howell & Au-Yeung (2002). Its basic idea is that cognitive-linguistic planning (PLAN) precesses are independent of motor execution (EX). It assumes that linguistic planning is processed utterance by utterance. It proposes that the cause of stammering is speech rate and disfluencies arise when an utterance needs to be produced but its plan is not ready. And this can be avoided by simply changing motor execution rate with a timekeeper, repeating a plan, interrupting speech to gain more time or direct advancing. AAF actually affects a timekeeping process that controls motor execution rate so that fluent speech can be achieved. Specially, a new term 'alterations to recurrent auditory information' (ARAI) started to be used (instead of AAF) to cover both feedback and non-feedback interpretations of the effects that occur when auditory environment is altered (Howell, 2004).

We need to admit that the underlying mechanism of speech control has not yet been investigated sufficiently. And all these controversial voices are helpful for us to gain a better understanding of speech production. While we continue to follow the view of feedback that the DIVA model relies on, we have to admit that this feedback concept is rather limited and possibly wrong. However, what interested us the most and motivated our study here is that there is just one study so far that computationally simulated the auditory feedback compensation behaviour. In Cai et al. (2011) and Cai (2012), they first simulated spatiotemporal control of articulation in response to formant-shifted auditory feedback of normal speakers with a modified DIVA model and received comparable results to the behaviour data. They then compared the modelling results between people who stutter (PWS) and normal speakers and reported that PWS speakers generally responded smaller in magnitude and slower in time to both spatial and temporal formant perturbations than normal speakers. To our knowledge, however, there is a lack of similar research on pitch control.

4.2 Introduction

To simulate the online control of pitch production, behavioural data need to be obtained first. Therefore, we developed a high performance speech signal processing programme to provide precise pitch manipulation with very low latency in the auditory feedback. With this programme, we conducted an empirical experiment to apply real-time pitch shift to auditory feedback to human subjects while they are producing meaningful bi-tonal (disyllable) Mandarin Chinese phrases. We analysed the collected behavioural data and verified previously reported online feedback compensation phenomenon. That is, short-delayed compensatory responses opposing the directions of pitch shifts were found.

Unlike the work by Cai (2012) which involved both spatial and temporal manipulations of F2, we only considered spatial changes of pitch in this study. Then, we simulated the behavioural data with an articulatory-based pitch controller. Our basic assumption for the simulation is that online auditory feedback compensation has to happen as part of the basic F0 production mechanism. For this mechanism, we adopted the target approximation (TA) model as the pitch controller, which has already been introduced in Chapter 3.

Through the simulation, we intend to achieve two goals. The first is to demonstrate that human speech movement (at least pitch movement) can be considered as a dynamic process of target approximation and the TA model is a valid F0 generation model at the motor-to-acoustic stage that can closely replicate human productions. Along this goal, the capability and flexibility of the TA model is shown. While we always introduce the model with the syllable as the basic modelling unit in the first place and have used it frequently in a syllable-synchronised way in other studies (and also in the next chapter), it can also be used with arbitrarily smaller units as long as they follow the asymptotic target approximation nature in speech production. Namely, a complete target approximation process can be seen as an

integration of multiple subprocesses of incomplete target approximation. And the boundary states of these subprocesses can be dynamically transferred from one to another resulting in a smooth final F0 trajectory. Therefore, online pitch control is naturally feasible with the TA model by simply manipulating the underlying pitch targets of such subprocesses. This is different from what Cai (2012) had to tackle with the DIVA model in their simulation experiments. As introduced in the background, the original DIVA model can only produce steady-state vowels, because it was designed primarily to produce single speech unit (e.g. syllable /ba/). In addition, some frequently used short words and utterances (e.g. 'I owe you a yo-yo' and 'good doggie' as commonly used by DIVA) can also be produced under the condition that they are treated as complex single units. As pointed out by Cai (2012, p. 96) 'it is highly unlikely that the speech motor system stores pre-learned motor trajectories for all possible utterances'. Indeed, in our daily life, many of the utterances that we produce are ones that we have never produced before. Our ability to produce complex utterances that are new to us implies that there must be a process that only learns the basic production units but is able to apply them dynamically in production.

As a consequence, in order to make more realistic multisyllabic articulation and enable a valid online formant control, a sequencing mechanism was developed by Cai (2012) to string the stored pre-learned syllables together to approximate the dynamic production of multisyllabic utterances. In order to simulate online formant compensation, those syllables were further divided into shorter independent 'epochs' (each consisting of a monotonic formant transition), which are then controlled separately. Without a dynamic model for trajectory generation, that was probably the best way to make the simulation possible. However, this epoch-by-epoch independence nature may lead to severe artefacts at the joints of produced signals,

which actually constrains the DIVA model to such motor control studies and does not allow it to be used in practical synthesis systems.

The second goal of this study is similar to the one in Cai (2012), that is to account for the spatiotemporal compensation patterns observed in the empirical experiment of online auditory feedback compensation. By comparing the simulated compensatory trajectories with the subject-produced ones, the sensorimotor mechanisms underlying the online control of TA-based F0 production in response to pitch-shifted auditory feedback can be revealed. That is, how do the TA targets react to the pitch shifts?

It is worth mentioning that there was a previously published mathematical model of pitch stabilisation using negative feedback and delays for sustained vowels (Hain et al., 2000), which was also later used for normal speech (Xu et al., 2004). However, that model lacked a critical F0 production model. It simply used control F0 signal as input and filtered it afterwards in response to perturbations found in the feedback. Hence, the underlying mechanism of F0 production and how it reacts to pitch-shifted feedback remains unclear. We aim to simulate feedback compensation in a way that is biomechanically plausible, so that it is also general enough to be extendable to other areas of motor control.

Our hypothesis for the simulation is: the surface compensatory responses observed in behavioural data is a result of temporary *adjustments* of underlying articulatory pitch targets of the TA model during speech production in reaction to the pitch-shifted auditory feedback. So there should be a momentary alternation of the originally planned target during the ongoing target approximation process. Following the behavioural data, the adjustment of pitch target should have a short latency and a small amplitude. Based on previous studies of the TA model (Liu & Xu, 2014; Prom-on et al., 2009), among the three TA target parameters, variation of target height directly affects the spatial displacement of surface F0 contour. So

in the current simulation, target height was chosen as the model parameter to be adjusted.

Generating pitch contours for speech synthesis is a demanding task. Given its previous influential achievements in synthesising designed lab speech of tonal (e.g. Chinese and Thai) and non-tonal (e.g English and German) languages (Prom-on et al., 2009; Xu & Prom-on, 2014), if the TA model can clearly reproduce the subtle pitch changes in this study with systematic feedback control, we have the reason to believe that the model has fulfilled its goal of replicating the physical process of F0 generation to a great extent. Therefore, it can be seen as qualified to be used in the motor-to-acoustic stage of our proposed speech synthesis approach for large scale spontaneous speech.

4.3 Behavioural Data

4.3.1 Subjects and stimuli

Eight paid subjects (four males and four females; age 22-27) speaking Beijing dialect of Mandarin Chinese participated in the experiment. Three of them were recruited at UCL and the rest five were recruited at the TechTemple startup camp[3] in Beijing. All subjects passed a hearing test and none of them reported history of neurological or speech disorders. In particular, none of them was aware of the purpose and methodology of the study before the experiment. In fact, a lot more subjects were recruited for the experiment, but due to technical reasons (e.g. recording quality, pitch tracking errors, etc.) only eight data sets were selected as qualified for further analysis and simulation.

[3] http://techtemple.cn/

Table 4.1 The four bi-tonal Chinese phrases used as stimuli in the experiment.

Phrase	Pronunciation	Pattern
妈妈	/mā mā/	High-High
妈麻	/mā má/	High-Rising
妈马	/mā mǎ/	High-Low
妈骂	/mā mà/	High-Falling

The stimuli used in the experiment are listed in Table 4.1. They consist of four meaningful disyllabic Chinese phrases. The first syllables are all in the High tone, whereas the second syllables are in the High, Rising, Low and Falling tones, respectively. The choice of these phrases was based on two reasons. First, a consistent first syllable provides a stable speech production condition that allows us to always apply pitch shifts to the same place of production and obtain clear compensatory patterns. Second, a varying second syllable can help us capture the most complex cross-syllable compensation effects. We limited the phrase length to only two characters to make sure they are simple, yet long enough to expose both effects and exceptional aftereffects of compensation. Short stimuli also meant that they would not exhaust the subjects and degrade their performance.

Before each trial of the experiment, a long random script was generated based on the phrases. The four phrases were repeated 75 times in the script and the order of the four phrases was shuffled in each repetition. Therefore, there were 300 phrases in total in each script. Also, the boundary phrases of each repetition were deliberately set as different to those of its adjacent repetitions so that throughout a script each phrase was guaranteed to be different from its neighbours. The purpose of doing this was to ensure that the speaker did not have to produce the same phrase twice successively, which may lead to lowered sensitivity in the second production. Furthermore, the magnitudes of applied pitch shifts were set to $-2.0, 0.0, 2.0$

semitones, which were also assigned randomly and evenly to the phrases in each script.

4.3.2 Experimental procedure

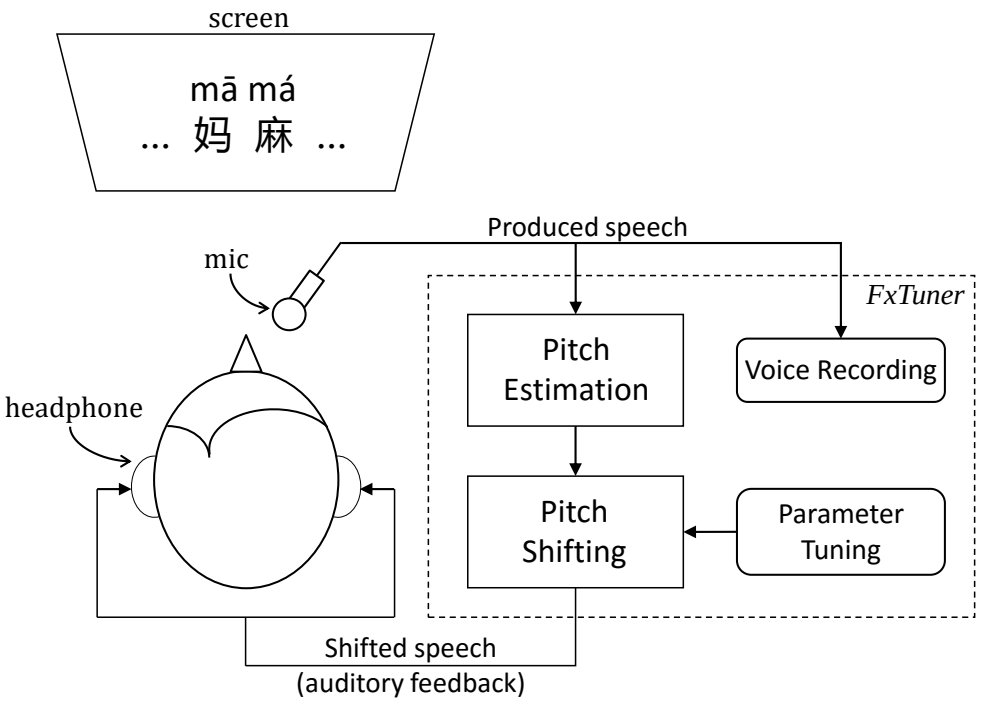


Figure 4.1 A schematic diagram showing the experimental settings and the workflow of behavioural data collection. The diagram is adapted and optimised from the ones displayed in Cai (2012) and van Brenk, Terband & Cai (2014).

Before the experiment, the subjects were trained to maintain their speech rate at 250 ms per character, i.e. 500 ms per phrase, as steadily as possible. This was to make sure that, on the one hand, the applied auditory pitch shifts occur roughly in the same time interval of the utterances, and on the other hand, the utterances are long enough to collect complete compensatory effects. In addition, the subjects were asked to practice reading their stimuli script aloud at about 70 dB SPL (sound pressure level) in order to sense the speaking effort that they need to maintain in the

experiment. Loudness was measured by a Brüel & Kjær 2203 sound level meter and monitored by the subjects themselves.

Figure 4.1 displays a schematic diagram showing the experimental settings and the workflow of this experiment. During the experiment, subjects were seated comfortably in a recording booth (at UCL or TechTemple) and asked to read aloud the phrase displayed on a screen in front of them. The phrases were displayed one by one and the progress was totally controlled by the subjects themselves by clicking the space key on a keyboard. The 300-phrase long script was loaded by our programme and then divided into 5 short scripts assigned to 5 experiment sessions. This was to make sure the speaker can have a rest after finishing each session and guarantee a stable performance.

4.3.3 Pitch shifting method and apparatus

The speech signals were first captured by a Countryman ISOMAX headset microphone and transmitted to the real-time pitch shifting programme, *FxTuner*[4], which loads the stimuli script, applies pitch shifts and records subjects' utterances to sound files. The programme relies on the PortAudio C/C++ library[5] and the CoreAudio driver on Mac OS X for low-latency playback and applies modified Short-time Fourier Transformation (STFT) for fast and accurate pitch shifting. Processed speech signals were instantly delivered back to both ears of the subject via a Beyerdynamic DT231 PRO headphone, with an added masking pink noise at 40 dB (Figure 4.1). The same masking noise was also used in many previous studies and has been shown to be effective in masking the bone-conducted feedback of self-production.

[4] Available at http://www.homepages.ucl.ac.uk/~uclyyix/tools.html

[5] A powerful cross-platform, open-source audio I/O library. Available at http://www.portaudio.com.

Generally speaking, playback latency exists in every electronic audio device nowadays. We can confidently say that no one can achieve zero latency as long as there is an electronic audio device used in their experiments. Nevertheless, for an experiment focusing on pitch-shift of auditory feedback, minimising such unwanted latency is the top priority. *FxTuner* achieved an overall latency of around 12 ms for the whole process of voice 'capture-manipulation-playback'[6] with original voice quality maintained to the maximum. A ring buffer was adopted to facilitate this process (Figure 4.2).

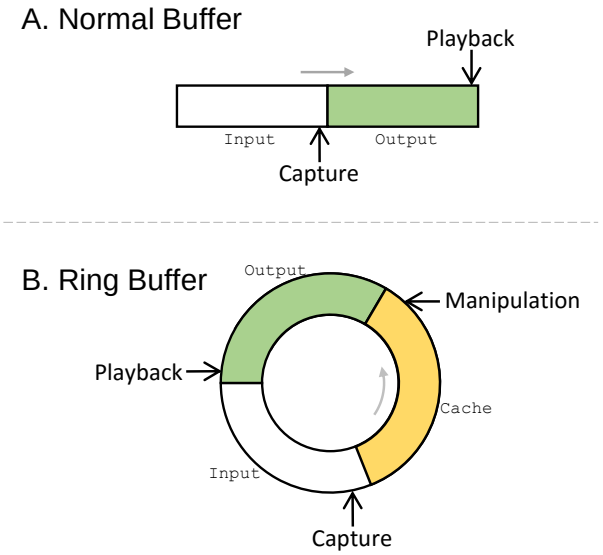


Figure 4.2 The type of ring buffer that we used in *FxTuner*. **A**: Normal buffer for comparison. **B**: Ring buffer with an extra 'manipulation' cache.

As comparison, Figure 4.2A shows a normal buffering structure that every electronic device has to use in order to accumulate certain amount of data for fluent playback. In this structure, an input buffer has to be filled first (Analog-to-Digital Conversion (ADC)) and then instantly copied to an output buffer for playback (Digital-to-Analog Conversion(DAC)). As long as an input buffer can be fully

[6]Cai (2012) reported 11 ms latency with a similar algorithm in their experiment. Howell et al. (1987) reported 5 ms latency with a speed-changing method. Note that those experiments were for formant or overall frequency change, whereas this experiment is for pitch change only and requires original formants largely retained.

filled before the last output buffer is used up, no glitch would be perceived by user. However, latency is unavoidable. In our case, with 44.1 kHz sampling rate and 256-sample buffer size, the latency would be $256/44100 = 0.005805$s (5.81 ms)[7]. When pitch shift is required, an intermediate buffer has to be added as a cache to accommodate this process. This leads to a ring buffer structure (Figure 4.2B). In this situation, when an output buffer starts to play, two buffers of data have been accumulated so that the latency is doubled to 11.62 ms. Note that *FxTuner* is optimised to be able to finish processing a cache buffer before it needs to be played, otherwise severe glitches would appear. In practice, extra delays caused by the ADC and DAC procedures are unavoidable. For our system, the sum of such delays was reported as 1.27 (±0.14) ms by the internal timer of PortAudio.

We also considered time-domain pitch manipulation techniques such as TD-PSOLA (Moulines & Charpentier, 1990) and WSOLA (Verhelst & Roelands, 1993), which normally introduce minimal formant change as long as the pitch change is within a small scale. However, these time-domain methods generally require at least two pitch periods (e.g. 20 ms for male voice) for signal processing, which is too slow for our low-latency purpose.

After the experiment, all subjects reported no awareness of distortion in the feedback other than the slightly higher loudness and the special masking noise in the auditory feedback. While we humbly request readers to consider that the latency achieved by *FxTuner* generally satisfied the requirement of 'imperceptible' temporal distortion and caused little distraction to speakers, we admit that this latency might impose extra effect on the feedback in addition to the designed pitch shift. However, since it is currently not possible to find a study that achieved genuine zero latency and applied similar pitch shift to the auditory feedback, we cannot make a comparison and assume how this extra latency may affect the

[7]The standard buffer size is normally 1024-sample or even larger, which would result in much higher latency.

results. Nevertheless, as we strictly followed the experiment protocol established by previous studies (Howell et al., 1987; Larson, 1998; Xu et al., 2004; Chen et al., 2007; Cai, 2012) and achieved one of the lowest latencies, we can say that this experiment is legitimate to the purpose of exposing real human behaviour in response to the pitch-shifted auditory feedback.

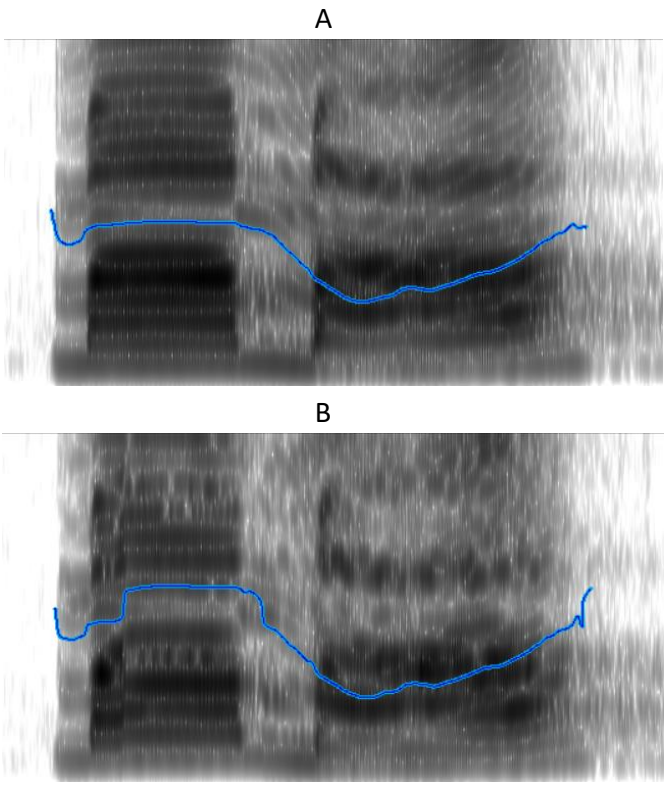


Figure 4.3 Example spectrograms and pitch tracking results (blue contour) of a production-feedback pair. **A**: the recorded /mā má/ production of a speaker. **B**: the recorded auditory feedback based on the production in A, in which an upward pitch shift was applied for 200 ms.

While we also admit that the use of typical STFT algorithm for pitch shifting also affects formants to some extent, with the help of engineers from Dolby Laboratories[8] we optimised the algorithm by trying to maintain the original envelope of power spectrum as much as possible in order to minimise formant deterioration (an example pitch shift is displayed in Figure 4.3).

[8]http://www.dolby.com

During the experiment, feedback fundamental frequencies were shifted upward or downward by 2 semitones (200 cents) for 200 ms, or left unchanged. The start of the pitch shift was 100 ms after the detection of vocalization onset. For this detection, we developed another algorithm basically based on a silence score accumulated within a window of captured sound samples. If the silence score exceeded a threshold for a certain number of times, the first time that it reached the threshold was treated as the vocalisation onset. Then, the timing (onset/offset) of applying pitch shift was calculated accordingly.

4.3.4 Behavioural results

The produced pitch contours were estimated by the autocorrelation algorithm (Rabiner, 1977) provided by Praat (Boersma, 2002) and sampled at 100 Hz. The contours were then transformed from Hertz to cent ($cent = 100 \times (39.86 \times \lg(f_0/195.997))$ where f_0 equals F0 in Hertz (Xu et al., 2004; Liu & Larson, 2007). Similar to the findings in previous studies, under the condition of pitch shift, not all productions were compensatory, as there were a small number of non-responses and following responses (i.e., the reactive pitch change followed the direction of the feedback pitch shift). The distinction between compensation, following and non-response was determined statistically by point-by-point serial two-tailed *t*-tests. Only if *p*-values within the pitch-shift window of a trial were continuously lower than the significance level of 0.05, the trial could be considered as a genuine compensatory or following response. In this study, only compensatory trials were selected for modelling and the following trials were considered as merely symmetric situations of the compensatory. In terms of non-response trials, we believe they imply that there was no underlying motor change triggered in reaction to pitch shift (i.e. normal production as there was no pitch shift at all). The selection was done by a preprocessing procedure manually by the author with an interactive programme.

During the data preprocessing, mispronounced and disfluent production trials were discarded first. Then, the remaining trials were screened and compared one by one to remove the following and non-responsive ones. Approximately, 57 % of the response trials[9] (with pitch-shifted feedback) and 92 % control trials (with normal feedback) survived this preprocessing, which were then used for analysis and simulation. Also, we would like to make it clear that all the three response types (compensatory, following and non-response) were found in every subject, which means that no subject stuck to one or two particular responses.

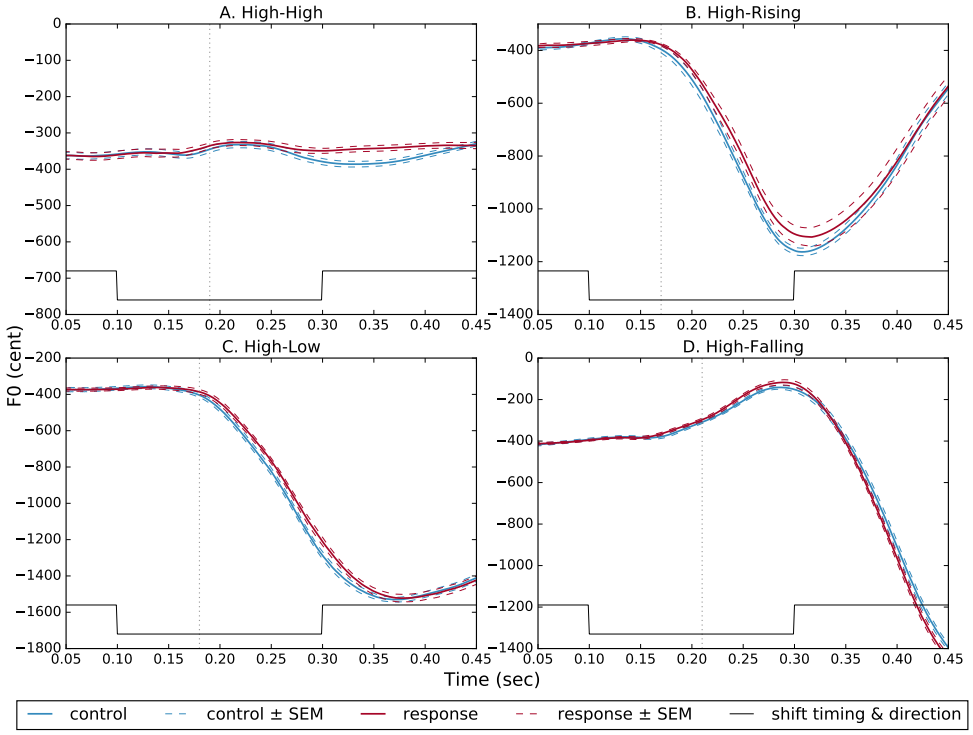
After data preprocessing as the first-pass, the pitch contours were then time-aligned, filtered to remove significant sample outliers, and smoothed with a five-point Hamming window. The purpose of this process was mainly to reduce the dispersion of temporal variations in the pitch contours so as to avoid confounding the pitch changes due to temporal misalignment and spatial compensatory adjustments. Similar process can also be found in previous studies (Xu et al., 2004; Liu & Larson, 2007; Patel, Niziolek, Reilly & Guenther, 2011; Cai, 2012).

Generally, the collected behavioural results showed similar patterns as those reported in previous studies. We will start with a representative case by an individual subject as illustration and then report grouped results of the experiment.

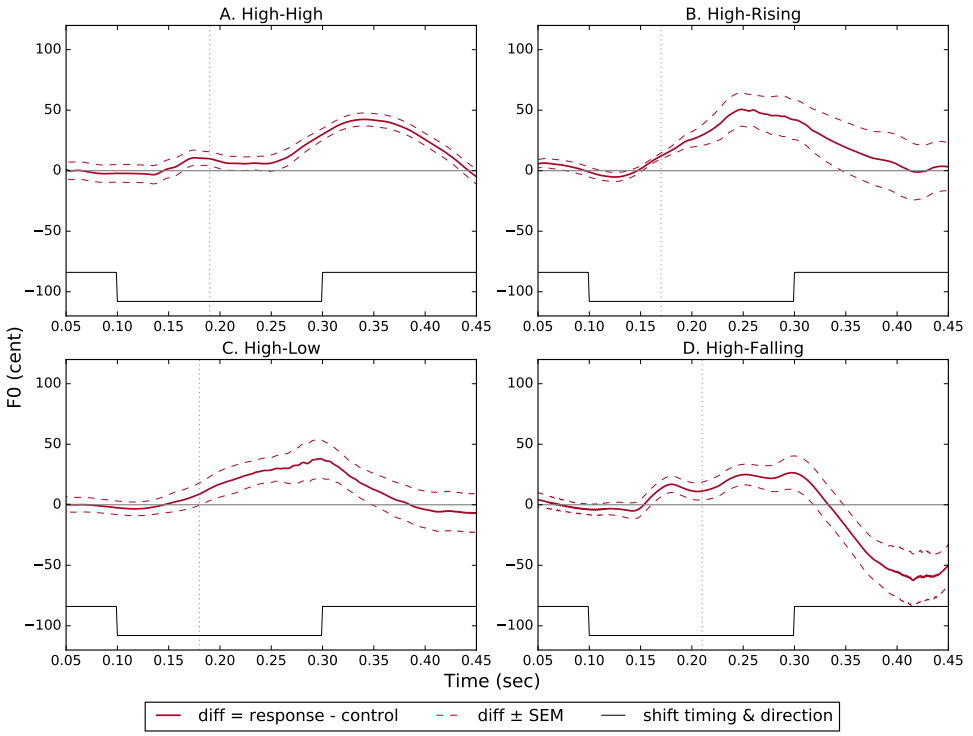
Individual results

A representative case by a male speaker (MS01, Figure 4.4 and Figure 4.5) is shown in this section. Figure 4.4a displays the production data with pitch shifted downward in the feedback, and which includes all the four bi-tonal phrases (H-H, H-R, H-L and H-F). The stability of compensation is demonstrated by the means (solid curves) and standard error of the means (SEM, dashed curves) plotted in the graph. These compensatory patterns are what to be modelled in the simulation

[9]We can see that compensatory trials are the majority (non-response trials were 17 % and following trials were 26 %).

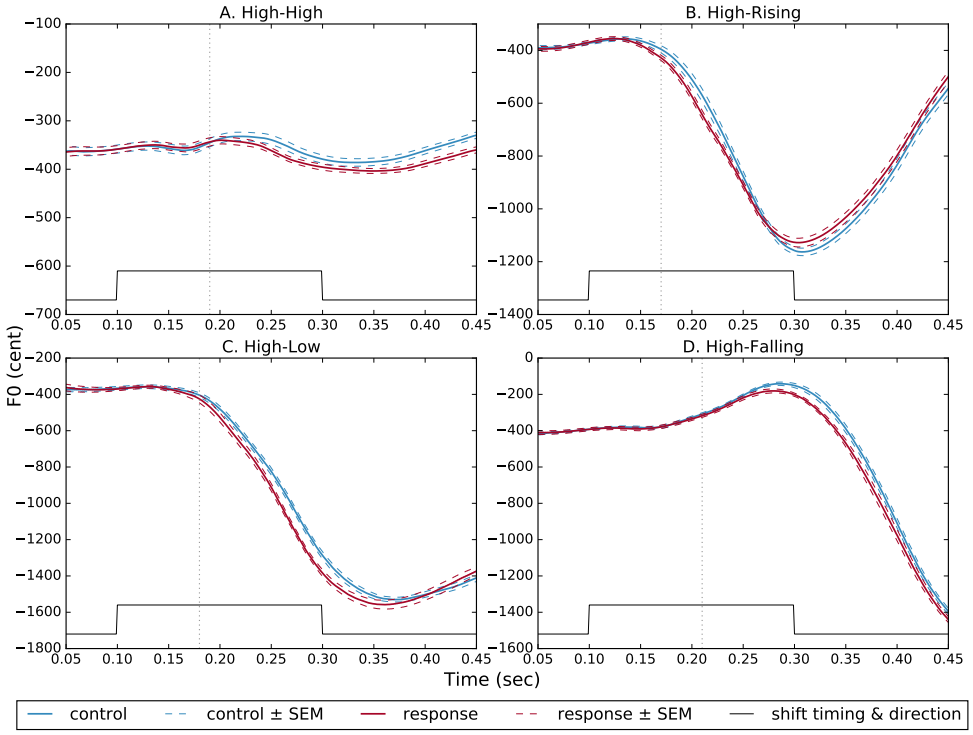


(a) The averaged F0 trajectories for the four bi-tonal phrases. The compensatory trajectories are in red, and the control ones are in blue.

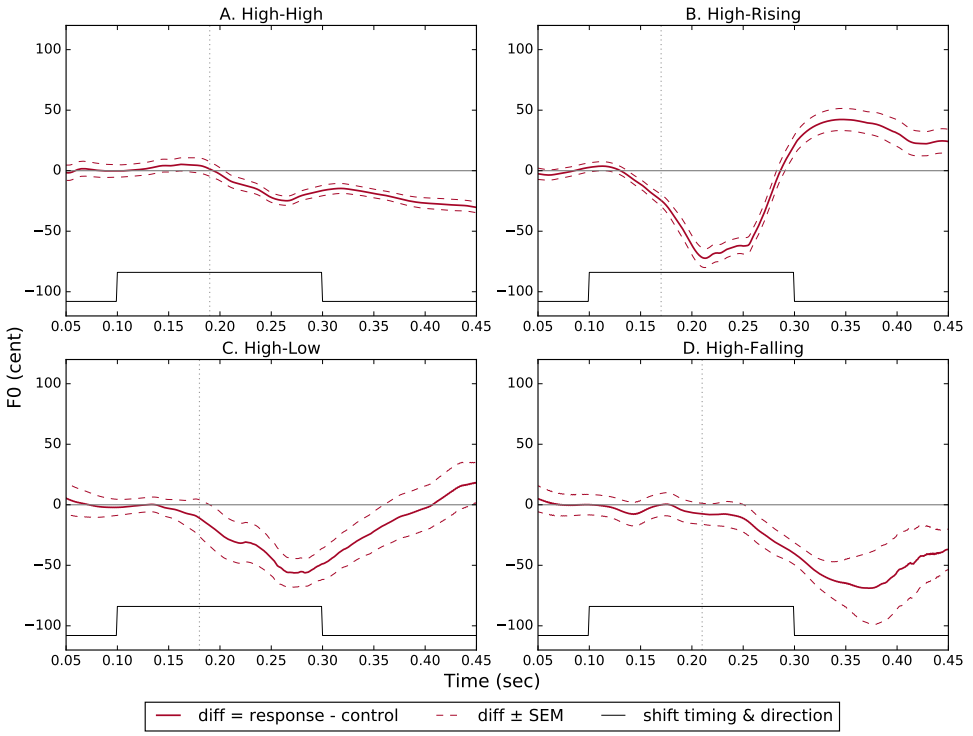


(b) Error view between the compensatory and the control trajectories in (a). The errors are plotted as red solid curves accompanied by red dashed curves indicating ± 1 SEM.

Figure 4.4 The productions by subject MS01 under the **downward** pitch-shifted feedback. Grey dotted lines are syllable boundaries.



(a) The averaged F0 trajectories for the four bi-tonal phrases. The compensatory trajectories are in red, and the control ones are in blue.



(b) Error view between the compensatory and the control trajectories in (a). The errors are plotted as red solid curves accompanied by red dashed curves indicating ± 1 SEM.

Figure 4.5 The productions by subject MS01 under the **upward** pitch-shifted feedback. Grey dotted lines are syllable boundaries.

to be discussed in the next section. Figure 4.4b provides an error view of such productions showing the average discrepancies between the compensatory and the control productions (solid red curves, accompanied by SEM as dashed red curves).

It is worth noting that, as shown in panel D of Figure 4.4b, negative aftereffects were found in the productions. Namely, when pitch shifts were removed from feedback, the subjects not only just returned towards the control level, but also overshot that level and produced lower F0s. Similar cases can also be seen in other productions (e.g. panel B in Figure 4.5b). This finding is consistent with the summary made by Purcell & Munhall (2006a) that negative aftereffects exist in F0 but are hardly observed in formants adaptations (Houde & Jordan, 2002). Additionally, as they also mentioned, this kind of negative aftereffects can also be found in other areas of motor control research such as arm control (Lackner & DiZio, 2005) and somatosensory speech control (Tremblay, Shiller & Ostry, 2003).

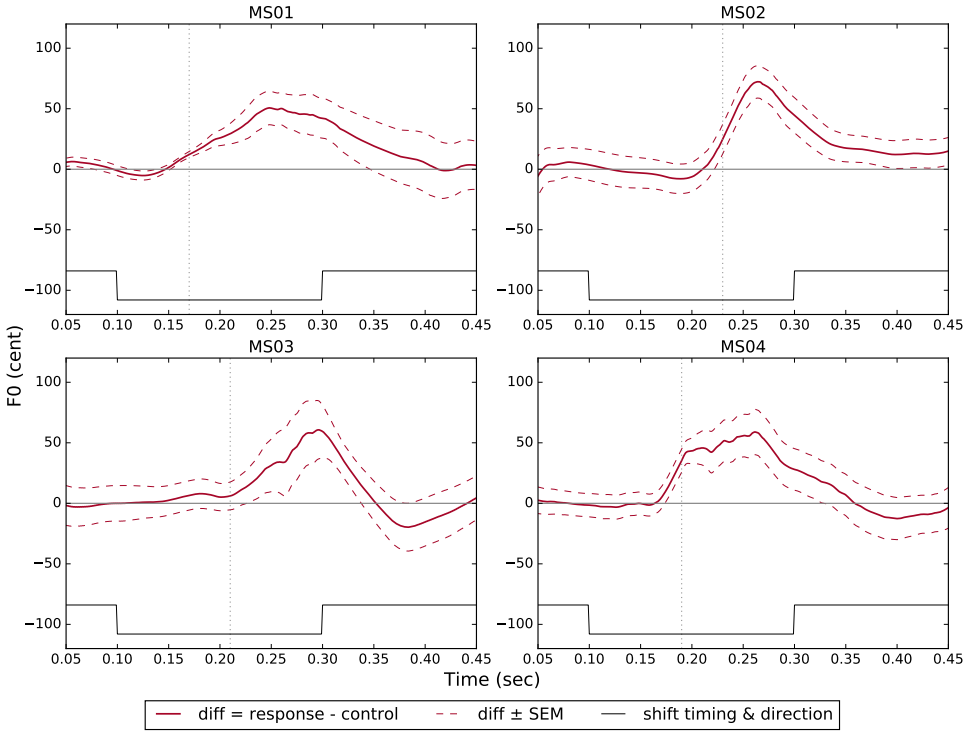
As we only introduced spatial pitch change in the experiment, responsive adjustments by subjects in the time domain will not be taken into consideration here. We have to admit it is possible that syllable duration may also be affected by pitch-shifted feedback, which might be a minor factor causing such overshooting phenomenon (Patel et al., 2011). Nevertheless, due to the fact that the number of productions in the experiment showing this overshooting pattern are limited (<25%), we chose not to discuss much on the possibility of temporal adjustment. More importantly, in terms of simulation, although temporal adjustment is not difficult to simulate, it would introduce time as an extra degree of freedom and cause interference with the spatial adjustment. Therefore, in this study we treat syllable durations as preplanned and unchanged. A more comprehensive study can be found in Cai (2012), as they also conducted an extra dedicated empirical experiment for temporal formant perturbation only and found significant difference between

PWS and normal speakers in response to such perturbations through DIVA-based simulation.

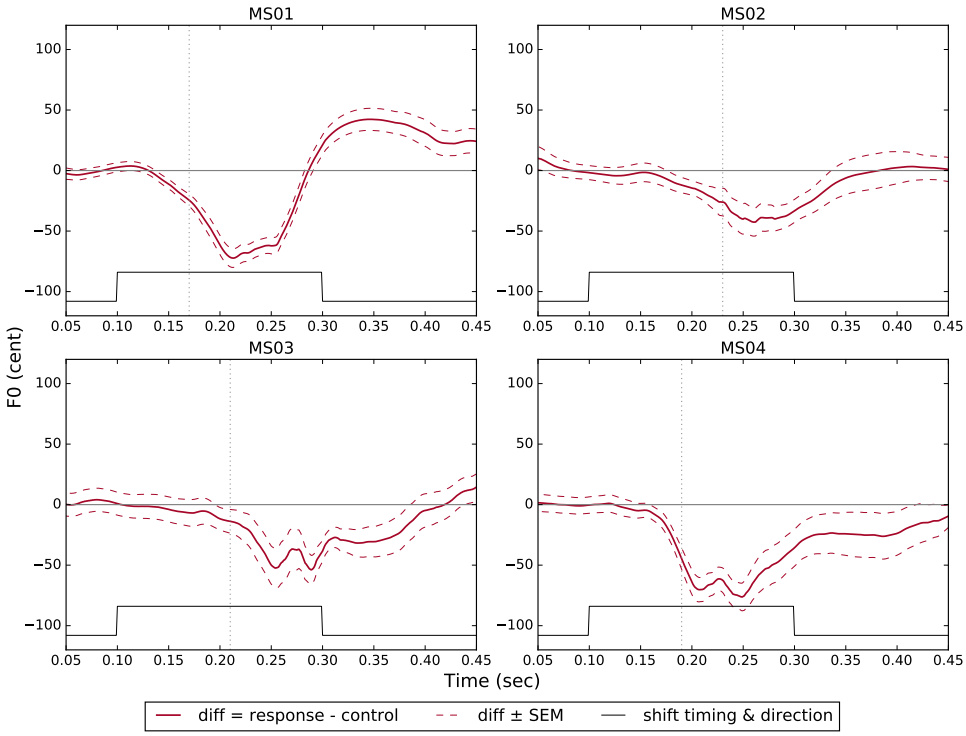
Group results

The inter-subject variability of compensatory patterns in response to both downward and upward pitch-shifted feedbacks can be seen in Figure 4.6 (male subjects) and Figure 4.7 (female subjects). The figures are based on subjects' productions of the High-Rising bi-tonal phrase. It can be easily seen that, on the one hand, the compensatory patterns generally existed in all subjects' productions; on the other hand, such patterns varied considerably across subjects. Namely, subjects compensated for the same pitch shift pattern in substantially individualised ways. For example, subject MS01 is the only one in Figure 4.6b showing the above-mentioned overshooting behaviour in response to upward pitch-shifted feedback, while no subject in Figure 4.6a shows such behaviour in response to downward pitch-shifted feedback. In contrast, most female subjects in Figure 4.7 overshoot the control contour after normal compensation except FS02 and FS04, in Figure 4.7a when compensating for downward pitch-shifted feedback. This kind of variability among subjects is quite normal in this research area and can be found in most previous studies.

For quantitative analysis, point-by-point serial two-tailed *t*-tests were run between the averaged compensatory and control F0 trajectories. Statistically, the observed mean onset of compensation is 101 ms after the onset of pitch shift across subjects, and the observed mean offset of compensation is 231 ms. The compensation onsets and offsets were determined by the *p*-values obtained from *t*-tests. Specifically, a window consisting of a consecutive series of *p*-values that are lower than the significance level of 0.05 was considered as genuine compensatory response. According to Xu et al. (2004), the compensation window has to be longer

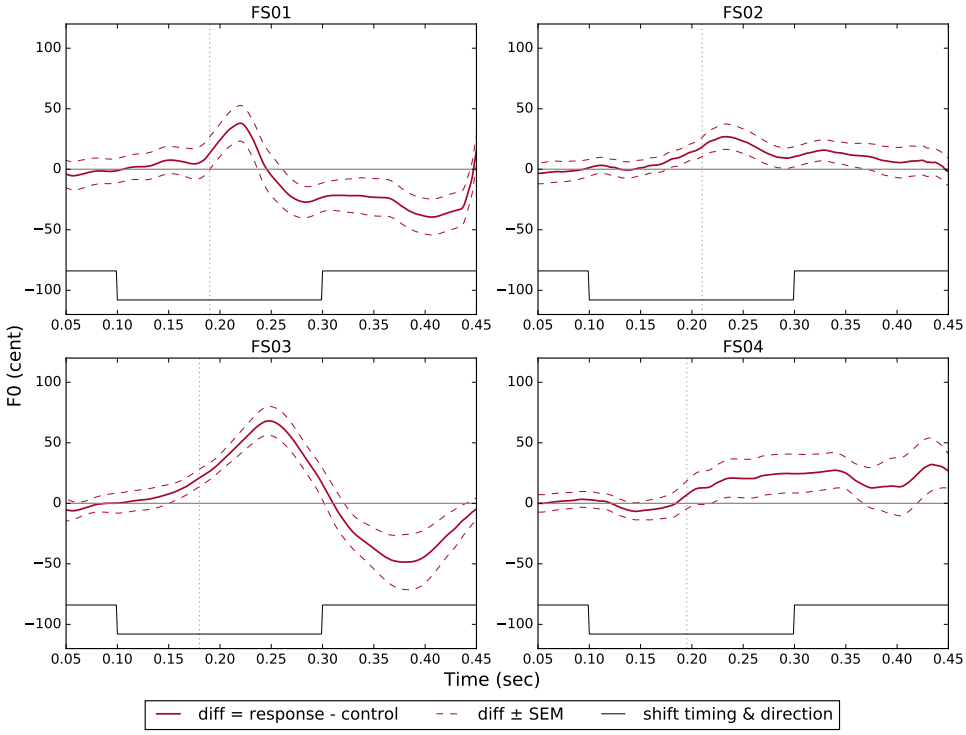


(a) Error view of the productions under the **downward** pitch-shifted feedback.

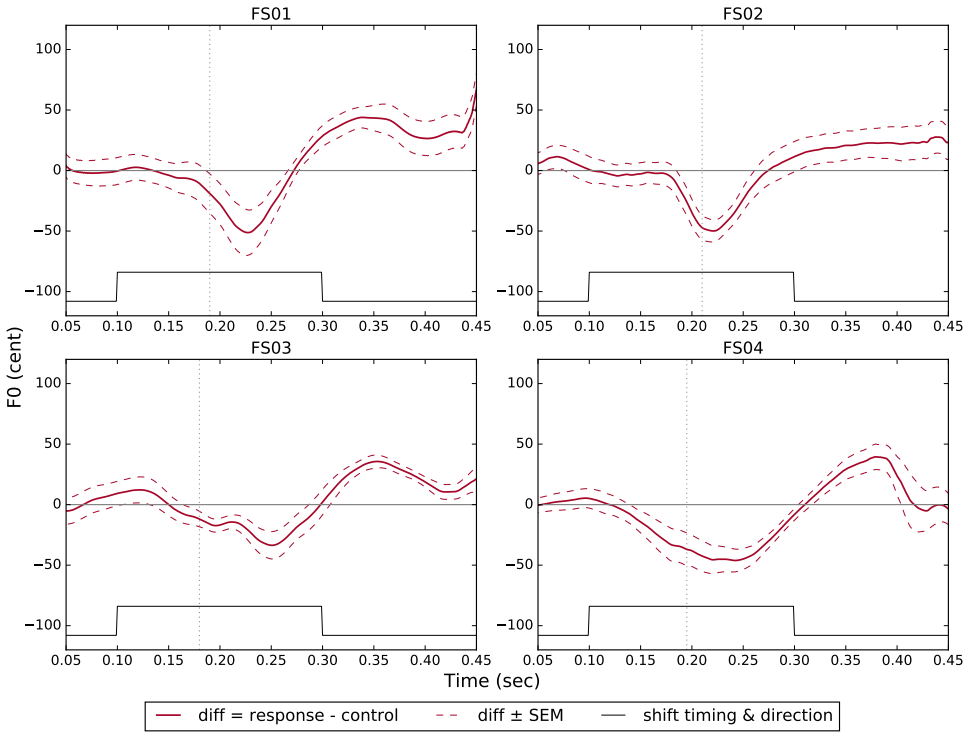


(b) Error view of the productions under the **upward** pitch-shifted feedback.

Figure 4.6 The averaged errors between the compensatory and the control trajectories of the High-Rising phrase produced by *male* subjects. The errors are plotted as red solid curves accompanied by red dashed curves indicating ± 1 SEM. Grey dotted lines are syllable boundaries.



(a) Error view of the productions under the **downward** pitch-shifted feedback.



(b) Error view of the productions under the **upward** pitch-shifted feedback.

Figure 4.7 The averaged errors between the compensatory and the control trajectories of the High-Rising phrase produced by *female* subjects. The errors are plotted as red solid curves accompanied by red dashed curves indicating ± 1 SEM. Grey dotted lines are syllable boundaries.

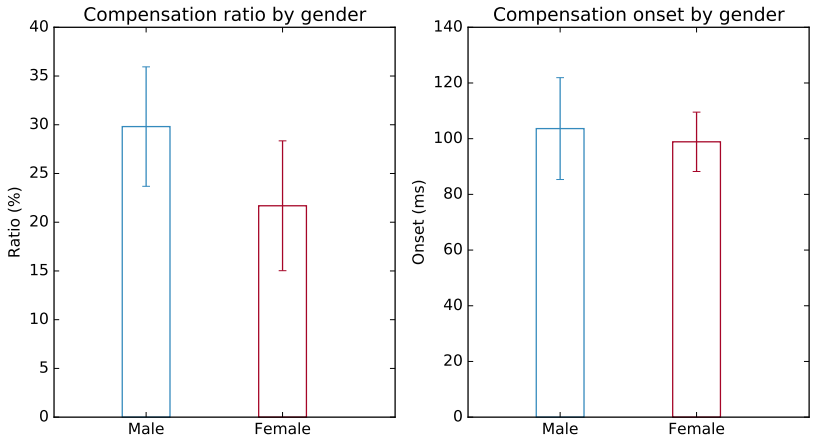


Figure 4.8 By gender comparisons of the observed compensation ratio (left panel) and onset (right panel).

than 50 ms in order to be qualified as a neuromuscular event. The first and last sample of such window were recognised as the onset and offset of compensation. The mean magnitude of peak compensation (defined as the largest displacement from control trajectory within the compensation window) collected in this experiment is 51.69 cents, i.e. 25.84 % if expressed as *compensation ratio* (divided by 200 cents, the magnitude of pitch shift). A comparison of gender difference in compensation ratio is shown in the left panel of Figure 4.8, which supports the finding reported in Chen et al. (2010) that male speakers produced larger compensatory responses than female speakers (ANOVA: $F_{1,8} = 5.64, p < 0.05$). However, the difference of compensation onset in gender (right panel, Figure 4.8) is not significant in our data (ANOVA: $F_{1,8} = 0.35, p = 0.56$). According to Chen et al. (2010), male speakers were significantly slower than female speakers when reacting to the pitch-shifted feedback. Xu & Sun (2002), in their study on maximum speed of pitch change, attributed this kind of gender differences to the physiological differences between men and women. A plausible explanation provided by them is that 'female speakers have less laryngeal mass and hence less laryngeal inertia, thus needing

less time than male speakers to initiate and end a pitch shift'[10] (Xu & Sun, 2002, p. 1405). Other physiological differences, such as thickness and length of vocal fold, may also responsible for such gender differences in reaction speed (Titze, 1989). The non-significance aspect of our data in terms of compensation onset difference is probably due to the physiological differences between the selected male and female subjects are not quite salient.

Based on the large number of empirical studies on this topic, we are confident that the data used here are consistent with the findings shown in other studies, and more importantly, faithful to the reality. Throughout the experiment, it can be seen that productions of the High-Rising bi-tonal phrase exhibit the most complex variations compared to others. Therefore, our simulation experiment will be based on productions of this phrase.

4.4 Simulation

To model the behavioural data, the TA model was first trained with the normal productions (the control trials), for which simple exhaustive search implemented by PENTAtrainer1 (Xu & Prom-on, 2010–2012) was used, to find the optimal target parameters of the averaged normal productions. These pre-learned TA targets were then used as the control targets for the simulation of compensation.

As discussed earlier in this chapter, the TA model is superior to the DIVA model in that it is a sequential model with an elaborate design to replicate the dynamic process of F0 production. Therefore, according to the established control protocol in Cai (2012), the TA model requires little modification for this simulation experiment. Specifically, when a syllable is being produced, its underlying pitch target can be freely changed at any time as needed, which merely shifts the current target

[10]Note that they only considered maximum speed of pitch change so that this cannot account for magnitude difference directly. To the present, no study has accounted for the reported gender difference in magnitude of pitch change.

so that a new target is temporally approached. For each compensatory response, the new target shifts against the direction of the pitch shift. For each following response, the new target shifts along the direction of the pitch shift. It is as if a multi-target syllable were produced. Importantly, the continuity of such multi-target approximation is guaranteed to generate F0 contours that are still smooth and natural, which may neatly fit the compensatory patterns produced by human subjects.

In Appendix A, we first provide a canonical implementation of the TA model in Python and then supply a 'multi-TA' function to demonstrate how to practically implement a multi-TA process to produce such multi-target syllables. Basically, a multi-TA process is composed of a series of subprocesses of target approximation. It is worth particular mentioning that, for each subprocess, the underlying pitch target is independent of those of other subprocesses. Especially, as we only consider spatial manipulation in this experiment, the height of each target needs to be locally calculated for its host unit. Moreover, once a new TA process is initiated, the time used for F0 generation during the process is always converted to relative time to its current onset.

In the following two subsections, simulation strategies regarding the cross-syllable compensation and the exceptional post-compensation overshooting behaviour are described in detail.

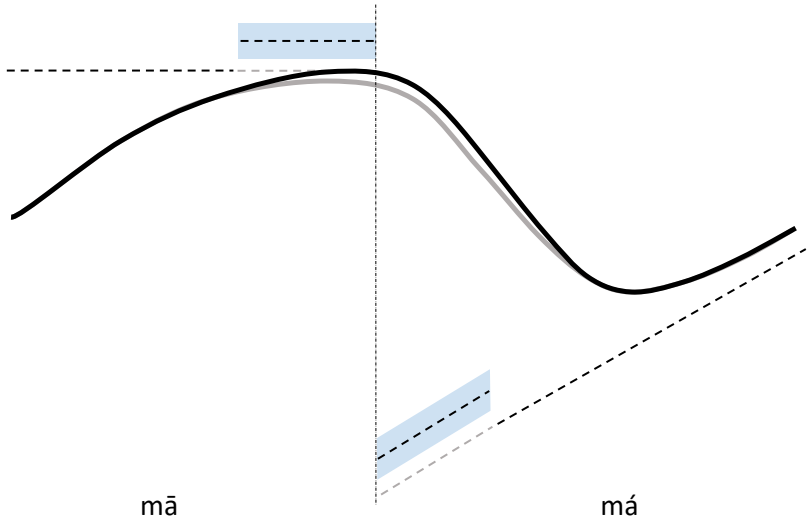
4.4.1 Cross-syllable compensation

As the pitch shifts in the empirical experiment were applied at 100 ms after vocalisation onset and lasted for 200 ms, given that speech rate was maintained at 250 ms per syllable, the compensations occurred in the experiment were mostly cross-syllable.

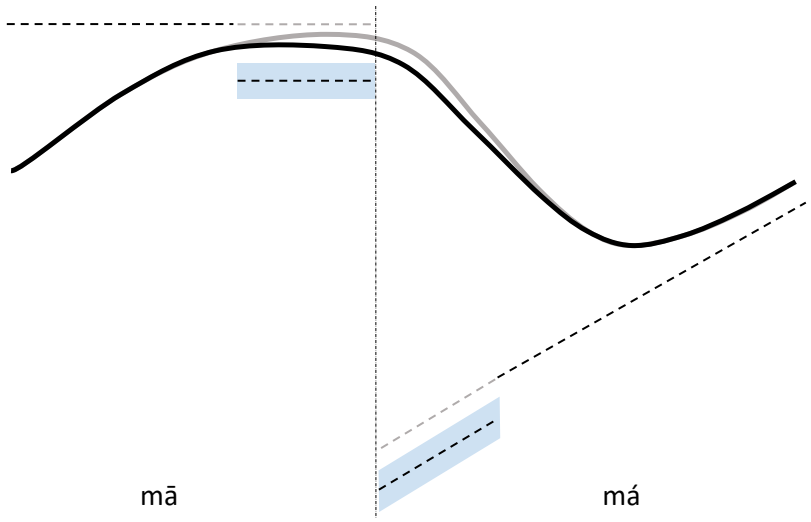
Three variables controlling the TA-based compensation were explored in this simulation: compensation onset, offset, and magnitude of target adjustment. These variables were generally set free within empirically reasonable ranges covering all the possible values of the variables when the best fits were searched for over the behavioural data.

For example, based on previous studies (Burnett & Larson, 2002; Donath et al., 2002; Natke et al., 2003; Tourville et al., 2008; Chen et al., 2010), the latency that surface F0 reacts to such pitch shifts is normally between 60 to 200 ms. That is, the earliest valid compensation onset can be at 60 ms after the first shifted F0 sample and the latest can be at 200 ms, i.e. the compensation onset should be within the 160-300 ms range after the vocalisation onset of an utterance. However, it is noteworthy that all those findings were based on surface F0 observations and it is actually not reasonable to consider that the change of articulatory movements can instantly give rise to significant surface acoustic change. Properties of physical movement like inertia should affect surface trajectory formation considerably and thus by no means can be ignored. As a consequence, the actural articulatory adjustments should occur earlier than corresponding acoustic changes in the observation. Cai (2012) was not able to take this into consideration due to the limit of the DIVA model, in which any changes on its motor commands lead to immediate changes on formants. With the TA model, the lower bound of this delay can be relaxed so as to leave some space for the target adjustments to take effect.

Figure 4.9 consists of two schematic diagrams showing how the pitch targets can be manipulated online in response to pitch-shifted feedback in TA-based pitch production. In Figure 4.9a, the F0 trajectory of syllable /mā/ first approaches the control target. Then, with the abrupt upward target shift before the syllable ends, the trajectory approaches the adjusted target and subsequently reaches a higher production than the control. More importantly, due to the persistence of pitch shift



(a) The case that pitch targets are temporarily adjusted upward.



(b) The case that pitch targets are temporarily adjusted downward.

Figure 4.9 Two schematic diagrams illustrating: 1. how underlying pitch target defined by the TA model can be temporarily adjusted downward or upward on-the-fly; 2. continuous target adjustment affecting two successive syllables results in cross-syllable compensation. The black and grey curves indicate the compensatory and the control F0 trajectories, respectively. And the black and grey dashed lines indicate the corresponding pitch targets of the trajectories. The intervals that pitch targets can be adjusted in are indicated as blue shadow.

in the feedback, this target change does not stop together with the first syllable, but also is extended to the beginning of the second syllable. Therefore, the second control target is also shifted upward for a while in the beginning of syllable /má/ before going back to normal. As a consequence, the trajectory continues to approach a higher pitch target in the second syllable at first and then goes down with the following target return. Figure 4.9b shows a similar case with the targets shifted to the opposite direction. Formally, the simulated cross-syllable compensation process can be expressed as below. In syllable 1,

$$F_0^1(t) = \begin{cases} A_1^1(t), & \text{if } 0 \leq t < T_{\text{on}} \\ \rho \cdot A_2^1(t - T_{\text{on}}), & \text{if } T_{\text{on}} \leq t < D^1 \end{cases}, \tag{4.1}$$

and in syllable 2,

$$F_0^2(t) = \begin{cases} \rho \cdot A_1^2(t - D^1), & \text{if } D^1 \leq t < T_{\text{off}} \\ A_2^2(t - T_{\text{off}}), & \text{if } T_{\text{off}} \leq t \leq D^2 \end{cases}, \tag{4.2}$$

in which all superscripts denote syllable index, $F_0(t)$ denotes the produced F0 at time t, $A(t)$ denotes TA-based production at t with pre-learned target parameters of its host syllable, subscripts of A denote the index of TA process, ρ applies height adjustment to the pitch target, whereas T_{on}, T_{off} and D denote compensation onset, offset and syllable duration, respectively. An optimisation process was then designed to find the corresponding underlying target movement when a pitch shift occurred in the auditory feedback. It is worth noting that, instead of being considered as parameterising the observed response contour, the optimisation process should be treated as an exploration process to look for the most appropriate target movement in reaction to the pitch-shifted auditory feedback and its best result closely simulates the motor behaviour of compensation.

The optimisation process was automated by a combination of both exhaustive exploration (for compensation onset and offset search) and nonlinear least squares optimisation (for target adjustment magnitude search). The latter was implemented by the LMFIT python package (Newville et al., 2014). Note that there are other nonlinear optimisers available, such as the Ceres Solver by Google (Agarwal, Mierle et al., 2010) and the *fmincon* function in the MATLAB Optimisation Toolbox (MathWorks, 2015), which can also be used here. The reason for including exhaustive exploration instead of entirely relying on nonlinear optimisation is that the time step for the experiment was set to 10 ms, which is challenging for most above-mentioned optimisers since they all work with floating numbers and assume that solutions can be found by initially making very small changes on them. Such discrete time series easily makes an optimiser to conclude that changing that time variable has no effect on the fit. Although we could make the time step as small as possible, it would cause some efficiency problem of the simulation process. The range of compensation onset was set to $[100, 300]$ ms, whereas the lower bound of compensation offset was dependent on the onset with an additional 50 ms (valid compensation should be longer than 50 ms) and the upper bound of which was set to 500 ms (the compensation can possibly reach the end of production). For each onset-offset timing pair, the best target adjustment magnitude can be found with the nonlinear optimiser, and a root mean square error (RMSE) score calculated between the simulated compensatory trajectory and its subject-produced counterpart can be obtained. By comparing all the collected RMSE scores, the best compensation onset T_{on}, offset T_{off} and target adjustment magnitude ρ can be found.

4.4.2 Post-compensation overshooting

As described in Section 4.3.4, consistent with other studies the overshooting phenomenon was observed in our behavioural data. Apparently, the above-mentioned

cross-syllable compensation strategy does not take this into consideration so that after compensation the simulated trajectories failed to replicate the quick return and overshooting observed in some subject-produced ones. Figure 4.10 is a real example of such situation. While compensatory adjustment of pitch target height alone does offer a better fit to the natural than the no-adjustment control, it fails to simulate the overshooting in the *post-compensation* interval. Therefore, an additional mechanism is needed to achieve fully satisfactory results.

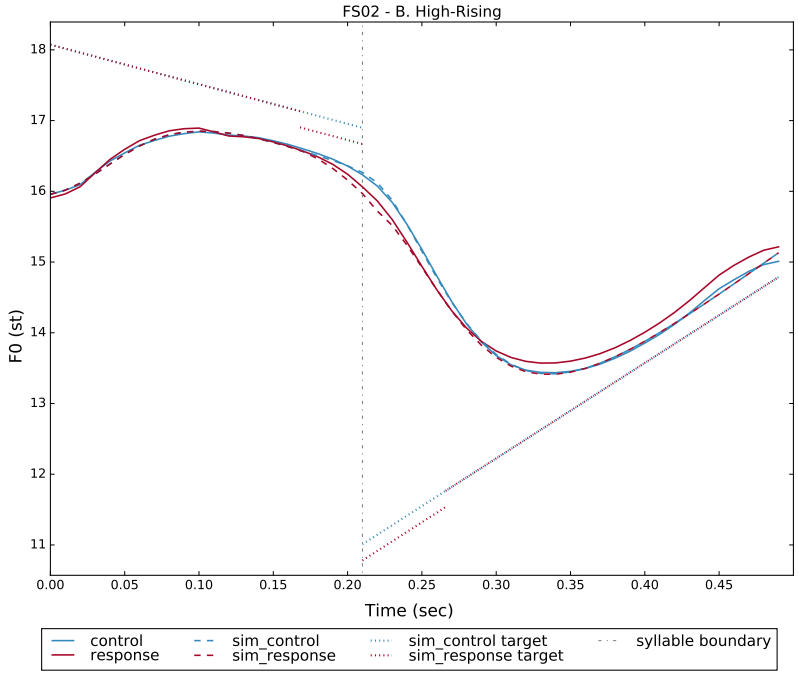
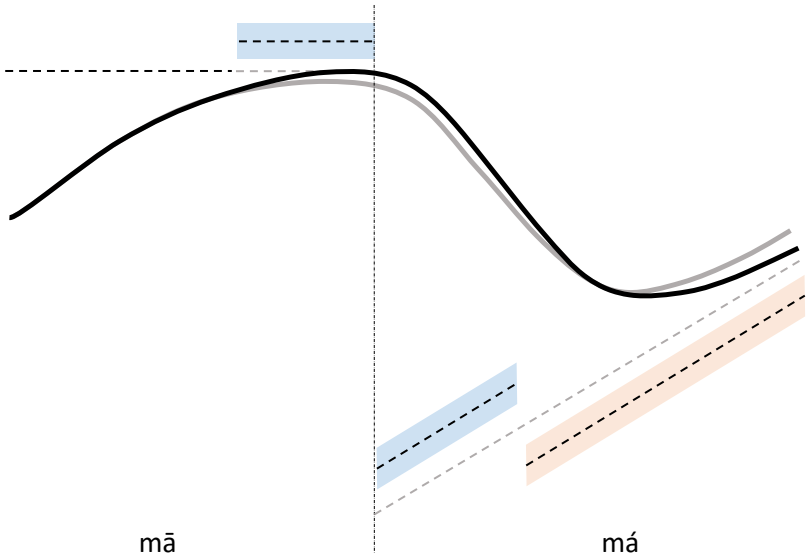


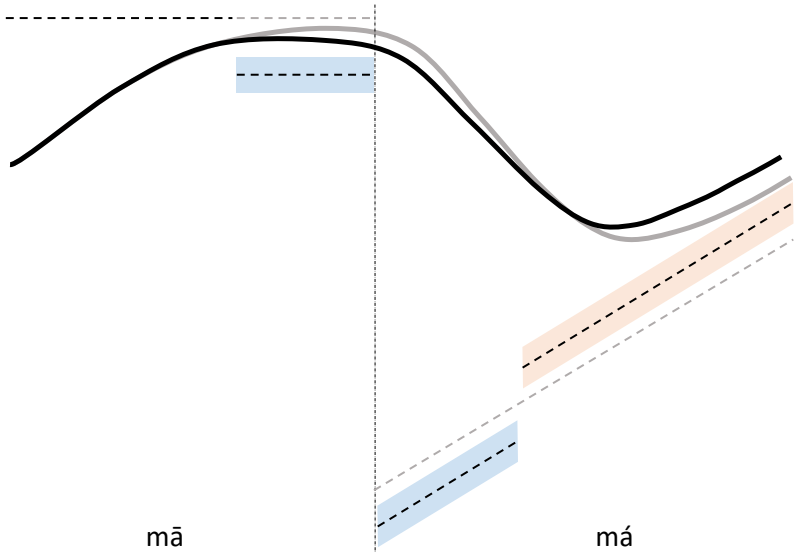
Figure 4.10 A real example showing that the compensation is effective within the normal compensation interval. But the simulated trajectory fails to replicate the subject-produced one in the post-compensation interval.

To address this issue, another target adjustment parameter ε is added in the post-compensation interval as an extra variable to be explored. We assume that this post-compensation target adjustment persists till the end of production.[11] Figure 4.11 shows two schematic diagrams illustrating this new strategy. Specifically, in Figure 4.11a, after the same compensation as in Figure 4.9a, the F0 trajectory

[11]This assumption relies on the fact that the post-compensation interval is short in this experiment. We admit that if the production is long enough (i.e. there are more syllables to be produced), such post-compensation target adjustment may stop at some point.



(a) The case that pitch targets are temporarily adjusted upward first and then downward in the post-compensation interval.



(b) The case that pitch targets are temporarily adjusted downward first and then upward in the post-compensation interval.

Figure 4.11 Additional post-compensation overshooting is added to Figure 4.9. Pitch target in the post-compensation interval becomes adjustable (pink shadow).

does not return to approach the control target of syllable /má/, instead it starts to be driven by a new target shifted downward to a lower level than the control. As a consequence, the simulated compensatory trajectory eventually overshoots the control. Similarly, Figure 4.11b shows the opposite case when the simulated production compensates downward during normal production and overshoots upward in the post-compensation interval. Formally, the production process in syllable 2 is modified as following,

$$F_0^2(t) = \begin{cases} \rho \cdot A_1^2(t - D^1), & \text{if } D^1 \leq t < T_{\text{off}} \\ \varepsilon \cdot A_2^2(t - T_{\text{off}}), & \text{if } T_{\text{off}} \leq t \leq D^2 \end{cases}, \tag{4.3}$$

in which the variable ε applies height adjustment to the control target in the post-compensation interval.

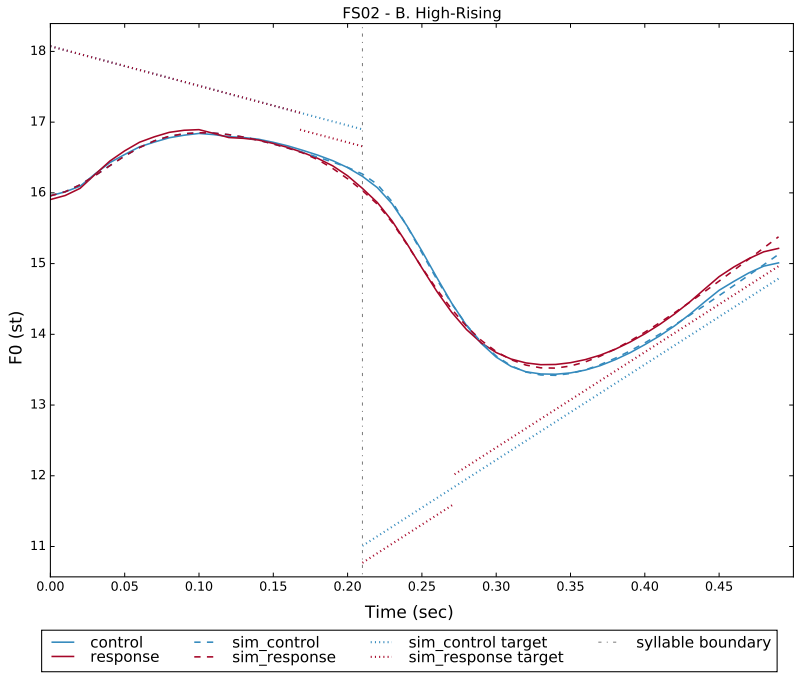
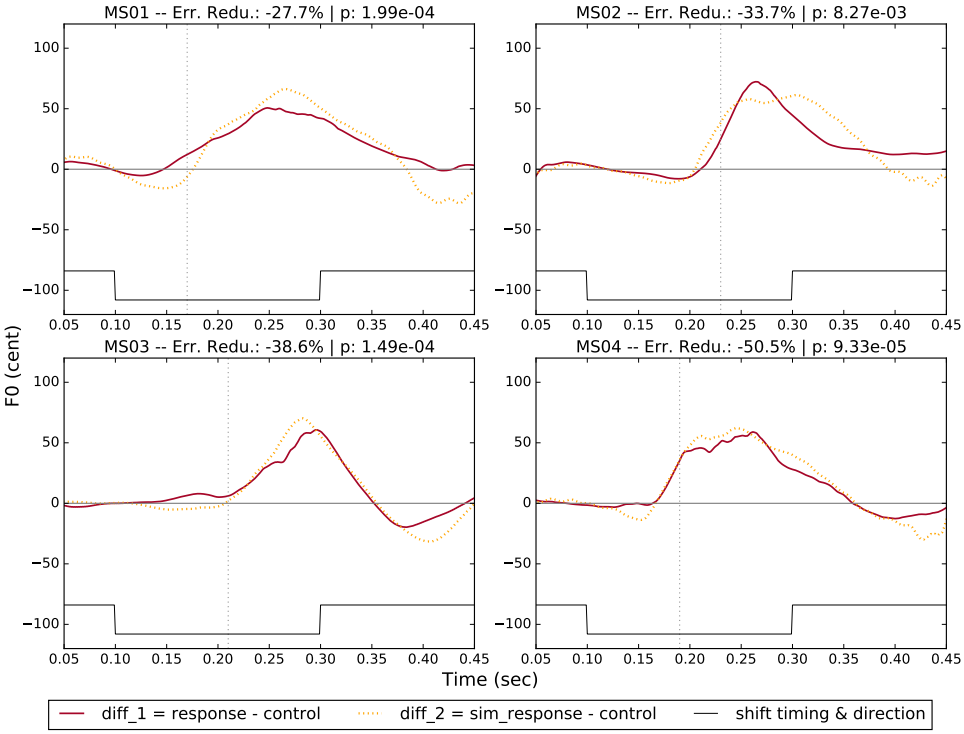


Figure 4.12 An improved example based on Figure 4.10 with extra overshooting simulation in the post-compensation interval.

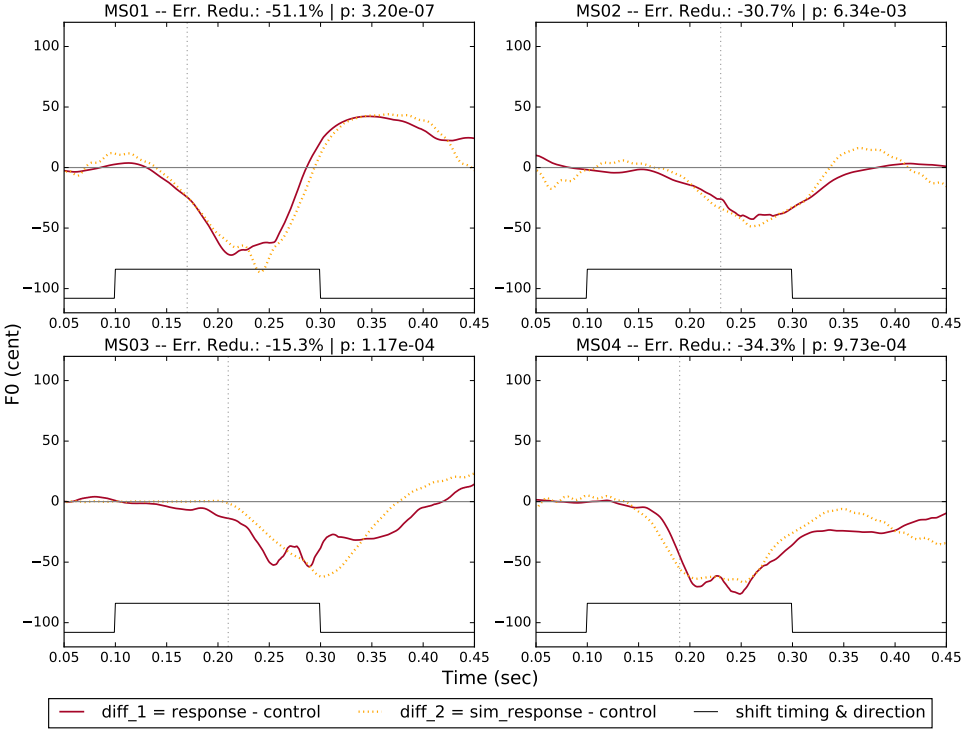
In contrast to Figure 4.10, a representative output of this improved simulation strategy is shown in Figure 4.12. While the compensation settings remain approximately the same as those in the original cross-syllable compensation (Figure 4.10), adding post-compensation target adjustment further improves the fitting of this interval. In general, the optimisation method remains unchanged in this simulation except that now two variables (ρ and ε) need to be considered simultaneously during the process, which is trivial for most nonlinear optimisers.

4.5 Results

The simulation results for the High-Rising bi-tonal phrase are grouped and displayed as error views in Figure 4.13 (male) and Figure 4.14 (female). Note that the post-compensation overshooting strategy was applied to all the cases while producing these results. If an overshooting existed, the best ε controlling post-compensation pitch target adjustment could be found to guarantee a good fit, otherwise the variable would simply stay at a very small value. Compared to previously displayed group results (Figure 4.6 and Figure 4.7), the dotted red curves denoting SEMs are removed here for clarity, instead the yellow dotted curves indicating errors between the simulated compensatory trajectories and their non-compensatory counterparts are added in order to show their resemblance to those previously reported error patterns (red solid curves) observed in subject productions. From the displayed results, it can be clearly seen that the simulation was effective and it generated very similar compensatory trajectories to those produced by the individual subjects. The performance gains for each individual simulation are also indicated in the figures, which were obtained by comparing the RMSE scores between subject-produced compensatory trajectories and the TA-generated ones before and after the simulation. Statistically, the error reductions introduced by the simulation are

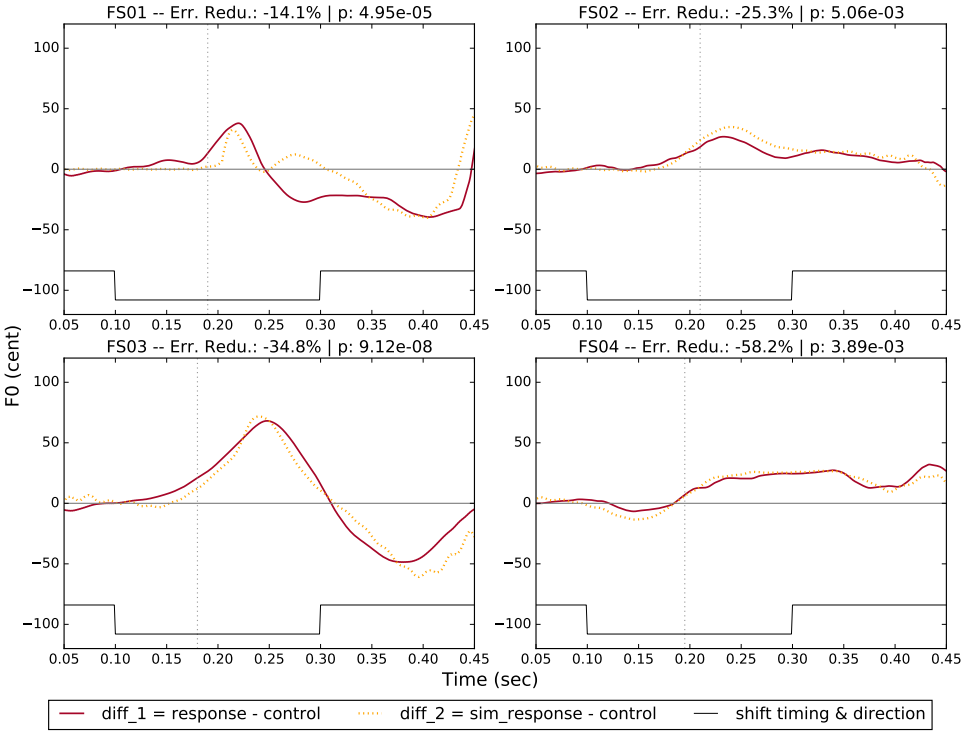


(a) Error view of the productions under the **downward** pitch-shifted feedback.

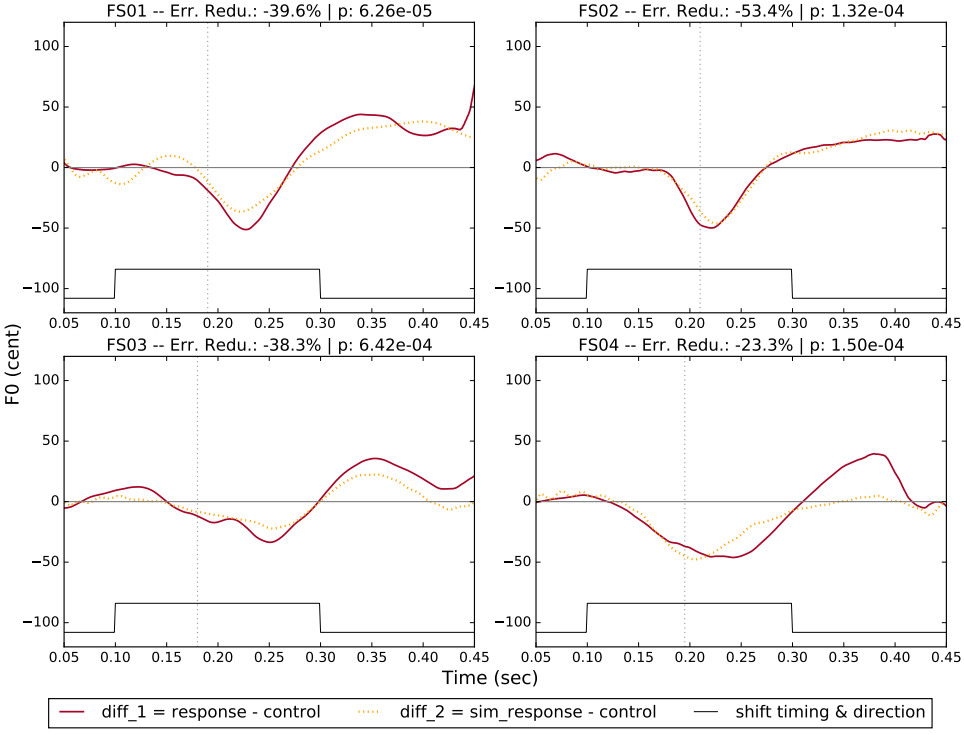


(b) Error view of the productions under the **upward** pitch-shifted feedback.

Figure 4.13 Error view comparisons showing the resemblance between the simulated and the observed compensatory patterns produced by *male* subjects. The observed errors are plotted as red solid curves, whereas the errors obtained by the simulation are plotted as yellow dotted curves. Grey dotted lines are syllable boundaries.



(a) Error view of the productions under the **downward** pitch-shifted feedback.



(b) Error view of the productions under the **upward** pitch-shifted feedback.

Figure 4.14 Error view comparisons showing the resemblance between the simulated and the observed compensatory patterns produced by *female* subjects. The observed errors are plotted as red solid curves, whereas the errors obtained by the simulation are plotted as yellow dotted curves. Grey dotted lines are syllable boundaries.

significant in all the displayed cases, as indicated by the p-values of paired t-tests shown in the figures.

Collectively, Figure 4.15 shows the fitting errors between the subject-produced trajectories and the TA-generated ones *before* and *after* the simulation. In this figure, the diagonal dashed line indicate equal fitting errors before and after the simulation, and the data points above the diagonal line indicates that the TA-based production achieved lower error in the simulation. On average, the performance gain of the TA-based simulation is 35.56%.

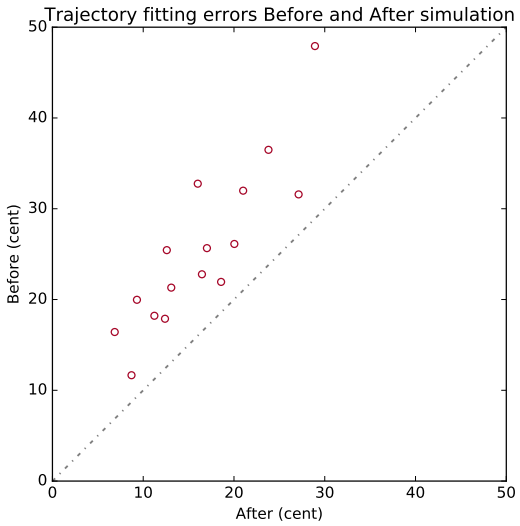


Figure 4.15 Performance of the TA-based simulation. The grey diagonal line indicates equal fitting errors *before* and *after* the simulation, and each circle corresponds to an individual subject.

In terms of target adjustment, Figure 4.16 shows that in all the simulated cases the target adjustments were made in the right direction (opposing the direction of pitch shift). However, the magnitude of target adjustment changes drastically from case to case, which can be as small as 0.37 semitones (37 cents) or as big as 4.1 semitones (410 cents), which is not actually proportional to the magnitude of pitch shift. Additionally, compared to the observed acoustic compensations, the gender difference of underlying motor control becomes statistically non-significant

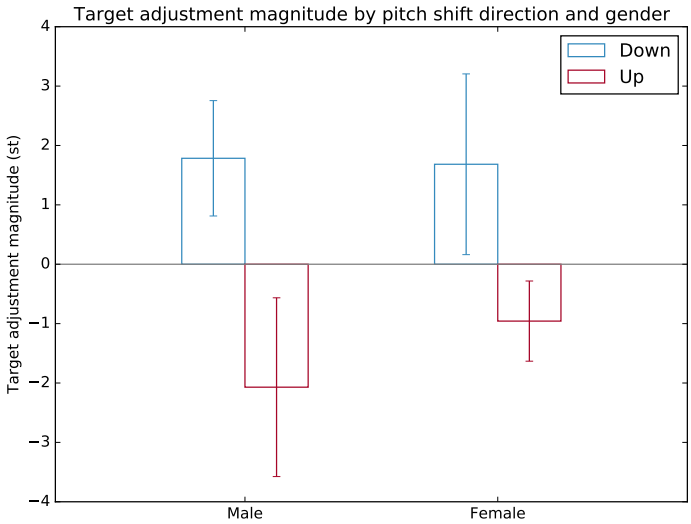


Figure 4.16 By gender comparison of target adjustment magnitude. Blue bars indicate adjustments in response to downward pitch shift, and the red bar indicate those in response to upward pitch shift.

(ANOVA: $F_{1,8} = 0.21, p = 0.65$). Multiple accounts are possible for this instability. For example, a local minimum reached in the nonlinear optimisation may lead to errors in target parameters; target parameters are subject-dependent and the rest two target parameters (target slope and rate of target approximation) may affect target height adjustment in different ways from case to case. Similarly, the target adjustment magnitudes in the post-compensation interval are also unstable.

Moreover, Figure 4.17 shows a box plot comparing the observed compensation onsets and offsets with the underlying target reaction onsets and offsets obtained through the simulation. Statistical analysis suggests that the underlying target compensation onset (mean: 53 ms) is significantly earlier than the observed compensation onset (mean: 101 ms) (paired *t*-test: $p = 6.19 \times 10^{-7}$). Meanwhile, the difference between underlying target compensation offset (mean: 169 ms) and the acoustic compensation onset (mean: 231 ms) is also statistically significant ($p = 0.023$), though the statistics is less confident about this. The intertwining between on-compensation and post-compensation target adjustments can be a pos-

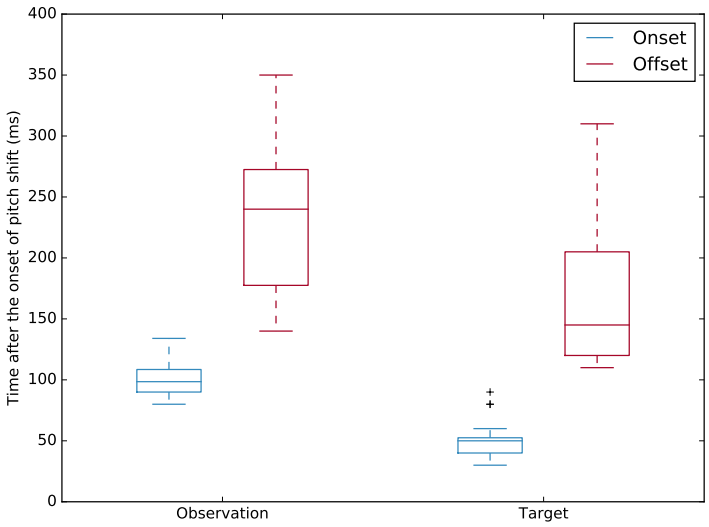


Figure 4.17 A box plot showing the difference of compensation timing between underlying articulatory pitch target and acoustic observation.

sible account for this, in which the target compensation offset is balanced due to the error-driven nature of the nonlinear optimiser. Nevertheless, the early target compensation onset of TA-based production found in the simulation suggests that in reality the underlying articulatory reaction to the pitch-shifted feedback may be earlier than the observed acoustic onset, and the mismatch between the two is possibly due to the time delay for articulatory movements to give rise to surface acoustic change.

The significant compensation results gained by the TA-based computational simulation support our hypothesis that human speech movement can be considered as a dynamic process of target approximation, and the online compensatory pitch production behaviour can be accounted for as a result of temporary adjustment of underlying articulatory pitch target.

4.6 Summary

Online compensation of pitch-shifted auditory feedback plays an important role in both normal speech production and childhood speech acquisition (Perkell, 2012). Computational simulation could help to achieve an understanding of how such online compensation works. In this chapter we adopted the TA model of dynamic pitch control in speech production and simulated feedback compensation by finding optimal adjustments to the original TA-based pitch targets learned from normal production data. We found that the best simulation results were obtained when the model applied both on-compensation target adjustment and post-compensation target overshooting. These findings have demonstrated the effectiveness of this compositional approach in simulating detailed dynamic pitch control, which could be linked to neural control mechanisms in the brain in future research (Perkell, 2012). This study provides further support for conceiving the basic human speech movement as a dynamic process of target approximation (Xu & Wang, 2001; Prom-on et al., 2009).

Chapter 5

Linguistic-to-Motor: Predicting TA Parameters with Deep and Recurrent Neural Networks

In this chapter, we complete the two-stage F0 modelling pipeline by providing the 'linguistic-to-motor' model, which sets up a mapping from linguistic features to TA motor parameters. While various machine learning techniques are available to build the model (e.g. stochastic learning, Bayesian methods, etc.), the state-of-the-art deep learning approach is used. We will start with a more detailed review of existing F0 modelling approaches and then dive into the details of our experimental neural network architectures. Besides that, in order to demonstrate the effectiveness of our TA-based 'linguistic-motor-acoustic' F0 modelling systems, we will compare them with the HMM and DNN baseline systems through both objective and subjective evaluations. The efficiency of the TA-based F0 modelling approach is also discussed.

5.1 Background on F0 Modelling Approaches in SPSS

5.1.1 HMM-based approach

HMM-based speech synthesis is one of the most renowned approaches in the field of TTS and it opened the research direction of SPSS (Figure 1.1). In its conventional formulation, spectral features, F0 and duration are modelled with a unified HMM framework. These features are first extracted frame-by-frame from each speech waveform and result in a corresponding feature vector sequence $\boldsymbol{o} = \left[\boldsymbol{o}_1^\top, \boldsymbol{o}_2^\top, ..., \boldsymbol{o}_T^\top\right]^\top$, in which an observed feature vector $\boldsymbol{o}_t$ consists of the static, delta and delta-delta components of spectral and F0 features of the tth frame. Contextual linguistic features are extracted through text analysis provided by a front end and aligned to acoustic features as sequence $\boldsymbol{w}$.

At the training stage, an HMM-based acoustical model is trained based on the maximum likelihood (ML) criterion with the expectation-maximisation (EM) algorithm (Dempster et al., 1977) to model the conditional distributions of an acoustic feature sequence given a linguistic feature sequence as

$$\hat{\Lambda} = \arg\max_{\Lambda} p(\boldsymbol{o} \mid \boldsymbol{w}, \Lambda), \tag{5.1}$$

where Λ denotes the HMM-based acoustic model. Especially, a multi-space probability distribution (MSD) modelling method was developed for F0 modelling to accommodate its voiced/unvoiced nature (Tokuda, Masuko, Miyazaki & Kobayashi, 1999).

A decision tree is then constructed based on a designed question set, which specifically depends on the extracted linguistic features. The decision tree clusters

contextually similar state models and enable them to share distribution parameters in order to avoid data-sparsity.

At the synthesis stage, given a contextual linguistic feature sequence $\boldsymbol{w}$ and trained HMM model $\hat{\Lambda}$, an acoustic feature sequence can be predicted as

$$\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{o}} p(\boldsymbol{o} \mid \boldsymbol{w}, \hat{\Lambda}). \tag{5.2}$$

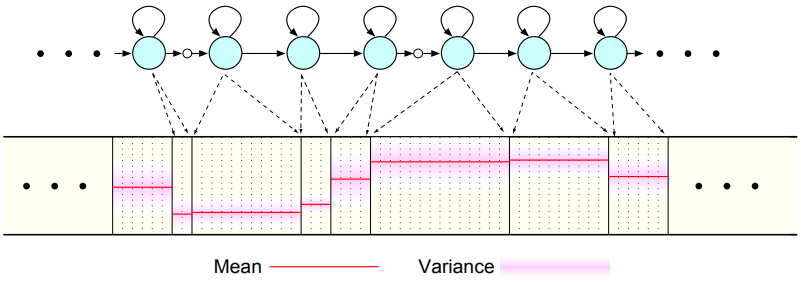


Figure 5.1 The piecewise static features output by a series of left-to-right discrete HMM states (Zen, 2015).

However, this will output piecewise constant parameter trajectories which change values abruptly at state transitions due to the conditionally independent state modelling nature of HMM (Figure 5.1). To address this issue, Tokuda et al. (2000) improved the parameter generation process with dynamic features considered so that Eq. (5.2) can be approximated as

$$\boldsymbol{c}^* = \arg\max_{\boldsymbol{c}} p(\boldsymbol{W}\boldsymbol{c} \mid \hat{\boldsymbol{q}}, \hat{\Lambda}), \tag{5.3}$$

in which

$$\hat{\boldsymbol{q}} = \arg\max_{\boldsymbol{q}} P(\boldsymbol{q} \mid \boldsymbol{w}, \hat{\Lambda}), \tag{5.4}$$

where $\boldsymbol{W}\boldsymbol{c}$ is equivalent to $\boldsymbol{o}$ with $\boldsymbol{c}$ denoting static acoustic features and $\boldsymbol{W}$ denoting the transformation matrix that applies dynamic features to $\boldsymbol{c}$; $\boldsymbol{q}$ denotes the state sequence and $\boldsymbol{c}^*$ denotes the output features. With this maximum likelihood

parameter generation (MLPG) algorithm, the original piecewise trajectories can be smoothed by considering the learned dynamics (Figure 5.2).

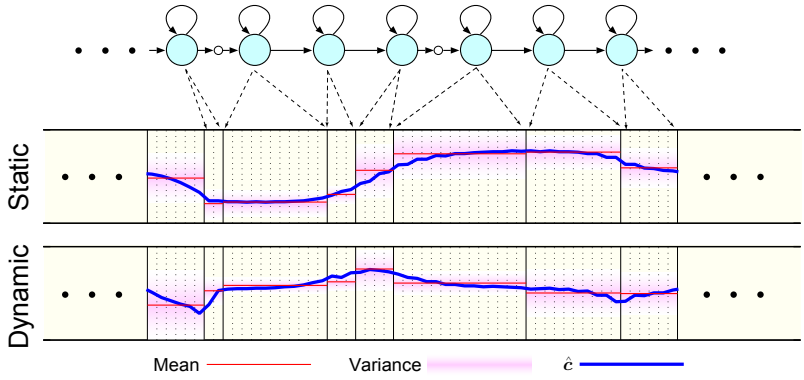


Figure 5.2 A smoothed trajectory by considering both static and dynamic features output by a series of left-to-right discrete HMM states (Zen, 2015).

The smoothed acoustic parameters are then sent to a vocoder (e.g. STRAIGHT by Kawahara et al. (1999), as widely used by researchers[1]) to reconstruct speech waveforms.

5.1.2 DNN-based approach

The DNN-based approach has attracted much attention in recent years with generally improved performance compared to the conventional HMM-based approach (Ling et al., 2013a; Zen et al., 2013; Zen & Senior, 2014; Qian et al., 2014). As investigated by Watts, Henter, Merritt, Wu & King (2016), replacing decision trees and moving from state-level prediction to frame-level prediction are the two major sources of improvements by DNN.

Generally, the DNN-based approach followed the same workflow as the HMM-based approach. For DNN-based F0 modelling, acoustic features are first extracted frame-by-frame from speech resulting in feature vectors, $\boldsymbol{y}_t = [y_t^1, y_t^2, y_t^3, y_t^4]^\top$, in which each component denotes F0 level, velocity, acceleration and binary

[1]However, STRAIGHT is slow so that it is actually rarely used as a vocoder in commercial applications.

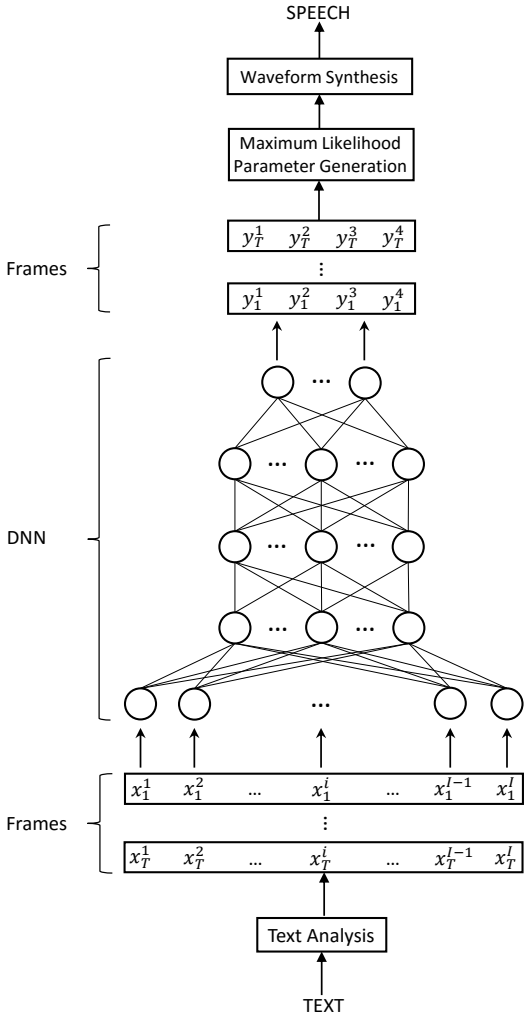


Figure 5.3 DNN-baseline system with frame-level linguistic to acoustic mapping.

voiced/unvoiced flag, respectively. As recently proposed by Yu & Young (2011) and widely used in other studies, the unvoiced parts of F0 contour are filled with interpolated values in order to preserve F0 continuity and assure ease of training. Figure 5.3 shows a diagram of this approach, in which a DNN replaces decision tree and sets up a frame-level mapping connecting acoustic feature vector $\boldsymbol{y}_t$ to its corresponding contextual linguistic feature vector $\boldsymbol{x}_t$ with dimension I. Note that there are extra linguistic features here compared to the HMM-based approach (e.g. frame position and amount in phone). Some of the linguistic features are categorical (e.g. phone identity, word POS, sentence type, etc.) which need to be

vectorised with one-hot encoding. Others are numeric which need to be normalised. The DNN is then trained with the backpropagation algorithm (Rumelhart, Hinton & Williams, 1986).

At the synthesis stage, the acoustic features are predicted by a trained DNN frame-by-frame according to given linguistic features which are then set as means of Gaussian distributions in the HMM scenario. Then the same parameter generation algorithm used in HMM-based approach is applied with the predicted dynamic features. The smoothed output acoustic features are then used for waveform synthesis.

While systems using DNN have reported significantly improved performance, there are two major drawbacks of the DNN-based approach as summarised by Zen (2015):

- Slow synthesis. Frame-by-frame matrix multiplication operation is way more computationally expensive than state-by-state traversing a decision tree to find acoustic feature statistics during synthesis.
- High parameter generation latency. The DNN-based approach still needs to use the same MLPG algorithm as in the HMM-based approach which needs to be performed over the whole utterance ($\mathcal{O}(T)$) in a post-filtering way after acoustic features are predicted by the DNN.

5.1.3 RNN-based approach

RNN-based approach is the frontier of SPSS at the time of writing this thesis, and it is still being actively explored. Basically, in this approach a recurrent neural network (RNN) is used to model the correlations between consecutive frames (which are ignored in a feedforward neural network). The use of RNN for speech synthesis is actually not a new invention, as the earliest application can be traced back to Tuerk

& Robinson (1993) and Karaali, Corrigan, Gerson & Massey (1997). However, due to the limits in computational power and algorithm at the time, the performance offered by the RNN was not very prominent. Owing to the development of deep leaning (LeCun, Bengio & Hinton, 2015) and especially to the emergence of long short-term memory (LSTM) (Hochreiter & Schmidhuber, 1997) based RNN, the latest applications of LSTM-RNN in speech synthesis are able to offer state-of-the-art performance.

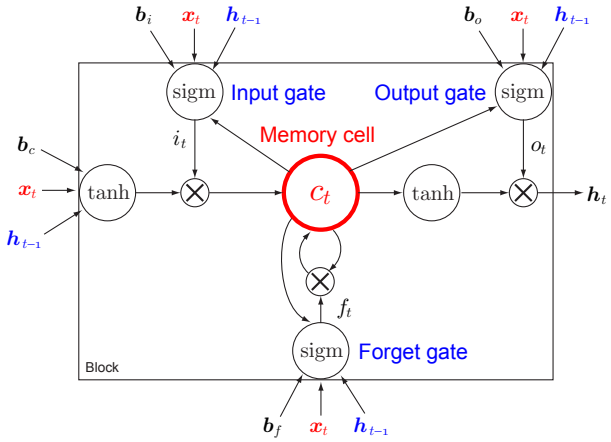


Figure 5.4 The memory block in a LSTM-RNN (Zen, 2015).

In LSTM-RNN, the neurons on the hidden layers of DNN are replaced with *memory blocks*. A typical memory block (Figure 5.4) is composed of one *memory cell* to store the temporal state of the network and three *gates* ('input', 'output' and 'forget') to control the information flow. The information flow is typically unidirectional (left to right), but can also be modified to be bidirectional so that both past and future features can be accessed by each memory block for prediction (Schuster, 1999; Liwicki, Graves, Bunke & Schmidhuber, 2007). Figure 5.5 displays a schematic comparison of dependency structures between a typical DNN and a unidirectional RNN. It can be seen clearly that the neurons on each layer of the RNN are unidirectionally connected.

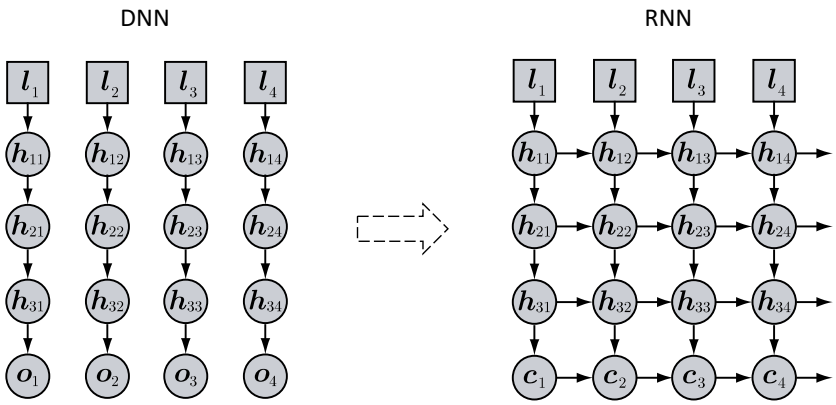


Figure 5.5 Different dependency structures of DNN and RNN (Zen, 2015).

Applications in SPSS show that the quality of synthetic speech (including F0 contours) can be significantly improved by using either bidirectional LSTM-RNNs (Fan et al., 2014; Fernandez et al., 2014) or unidirectional LSTM-RNN (Zen & Sak, 2015). Moreover, some latest studies found that the structure of memory block can be simplified by disabling some gates, which can significantly reduce the complexity of neural network without degrading performance of some signal processing (Chung, Gülçehre, Cho & Bengio, 2014) and speech synthesis (Wu & King, 2016b) tasks. So far, variations of the RNN-based approach are still springing up and no implementation can actually be considered as a widely-recognised standard. In this chapter, we will also experiment with an RNN but with a simplified architecture according the latest development.

5.2 Introduction

Given the TA model as a valid 'motor-to-acoustic' model, in order to produce meaningful F0 contours, the TA model needs to be driven by a 'linguistic-to-motor' mapping. The mapping can be learned by associating the TA motor parameters to *syllable*-level contextual linguistic features via various learning methods (e.g. decision tree, DNN or RNN).

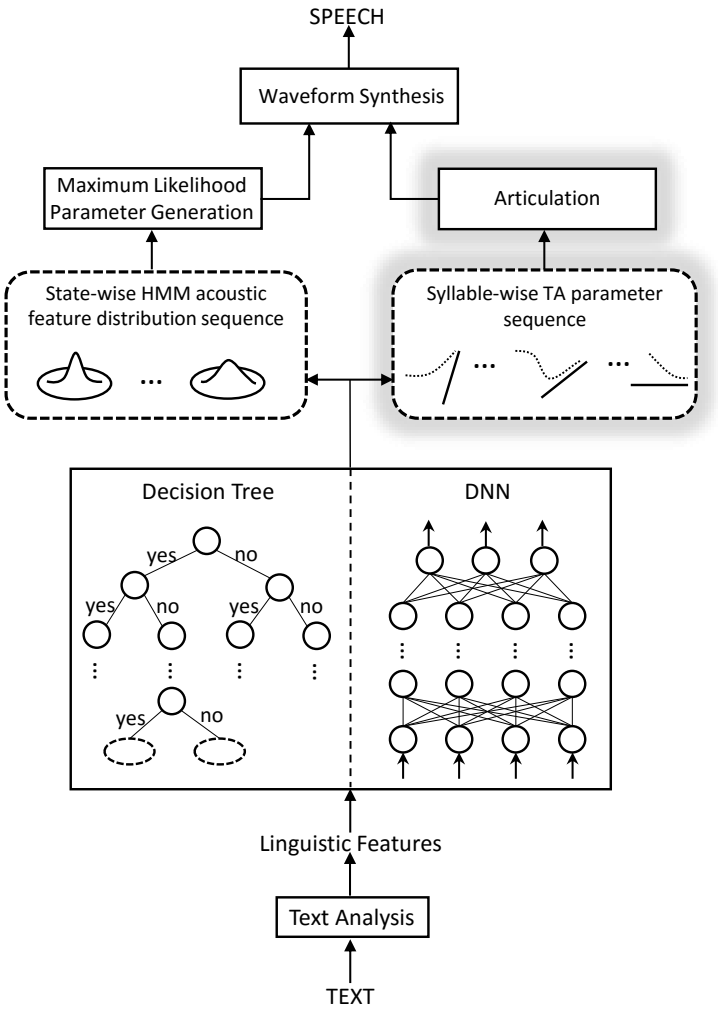


Figure 5.6 Comparative overview between SPSS and our TA-based approach.

Figure 5.6 is a comparative diagram showing the differences between the decision tree or DNN based SPSS approach and the proposed TA-based approach. Basically, state-level acoustic modelling with the HMM model and MLPG algorithm are replaced by syllable-level acoustic modelling with the TA model (TA-based F0 parameter generation). The proposed approach addresses limitations in the SPSS approach by bringing in four major improvements:

- Syllable-level modelling. As discussed above, the simulation of the dynamical process of articulation provided by the TA model helps to uniformly model independent F0 samples at the syllable level and thus resolves the

problem of temporal dependency which is currently coped with by the HMM model with MLPG algorithm or tuning computationally expensive RNNs.

- Economic representation. In contrast to other SPSS approaches, the TA-based approach does not rely on static and Gaussian distributed features for acoustic representation in training and synthesis. Instead, it uses only three articulatory control parameters as representation of F0 contours at the syllable level. This reduces hundreds or even thousands of F0 features per utterance to only tens of parameters.

- Small footprint. Small number of syllable-level linguistic features with economic acoustic representation lowers model complexity, which in turn leads to efficient training and synthesis.

- Fast synthesis. As summarised by Zen (2015), the typical DNN-based approach is slow because matrix multiplication operations need to apply for every *frame* to predict acoustic features. With TA as the link to a DNN, prediction is needed only for each syllable instead of each frame. Thus a faster synthesis process can be achieved. However, it is not easy to directly compare TA with the HMM-based approach in terms of computational cost, which involves a different operation (tree traversing) that needs to be done for every *state*. Even so, the TA-based approach wins in terms of number of predictions.

In short, current SPSS approaches jumps over the physical process of human speech production and sets up a frame-level direct 'linguistic-to-acoustic' mapping, whereas the TA-based approach models a relatively complete human speech production pipeline with a two-staged 'linguistic-motor-acoustic' mapping.

5.3 DNN-based F0 Modelling with TA

DNN-based F0 modelling with the TA model can be done by following the workflow shown in Figure 5.6. As a two-stage 'linguistic-motor-acoustic' production process (Figure 1.6), each stage needs to be optimised separately to achieve an optimal end-to-end mapping.

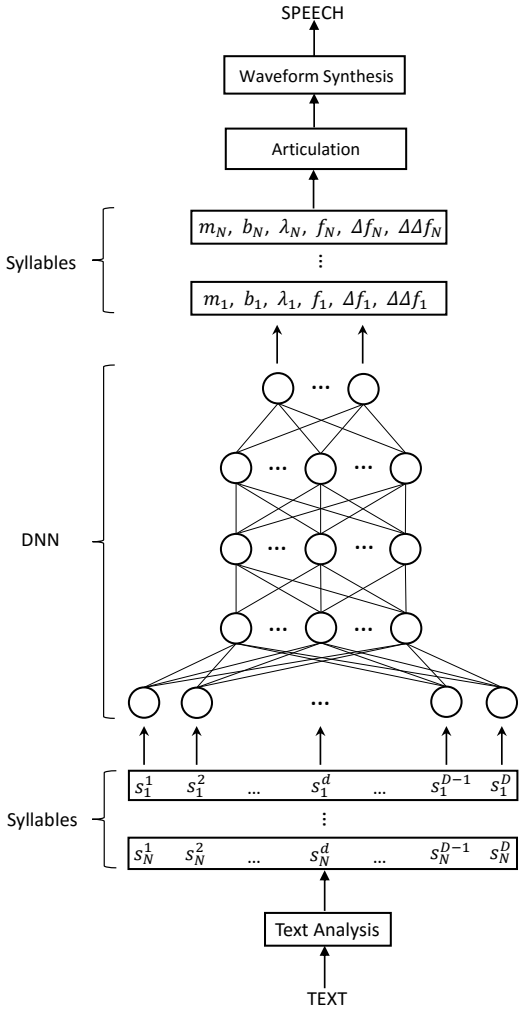


Figure 5.7 DNN-TA system associates linguistic features with TA parameters for each syllable.

As described above, the TA process is implemented by a dynamical system, so the Levenberg-Marquardt nonlinear least-squares method (Moré, 1978) can be applied to find locally optimal TA parameters of each syllable. Therefore,

segmentation needs to apply first to obtain syllabic F0 contours. For each syllable, its optimal TA parameters together with the onset F0 state form a vector $\boldsymbol{v}_n = [m_n, b_n, \lambda_n, f_n, \Delta f_n, \Delta\Delta f_n]^\top$. Although it is increasingly popular to do interpolation on unvoiced parts of F0 in order to obtain overall continuous contours for universal modelling, and it is also reasonable to do so with the hypothesis that articulatory movements are continuous even during unvoiced sessions (Xu & Prom-on, 2014), here in our local optimum searching task unstable TA parameters may be found due to the undesirable errors introduced by the pitch tracking and interpolation. Therefore, simple heuristics need to be developed to skip initial unvoiced parts in syllables with voiceless consonants, and optimal TA parameters were obtained based only on the voiced parts. The resulting TA parameter vector can then be associated to corresponding *syllable*-level contextual linguistic feature vector $\boldsymbol{s}_n$ (e.g. syllable identity, syllable position, etc.) to train a DNN (Figure 5.7). Note that the dimension D of this vector is greatly smaller than I (the input dimension in the typical DNN-based approach), because all the phone-level and frame-level features are removed here.

At the synthesis stage, the trained DNN can be used for syllable-wise TA parameter prediction. If a syllable is voiced and it is not the first one of the utterance, the offset F0 state of its preceding syllable will be transferred as the onset state of the current syllable for TA-based F0 generation, so that the predicted onset state parameters will not be used. Otherwise, all the six predicted parameters will be used for F0 generation. MLPG-like smoothing process is not necessary, since the TA model directly generates vocoder-ready F0 parameters with in-syllable and cross-syllable dynamics.

5.4 RNN-based F0 Modelling with TA

Fundamentally, both conventional HMM-based and the typical feedforward DNN-based approaches are state/frame independent. Although there is some positional information in the contextual linguistic features, their contribution to acoustic variations at the frame level is small. When not applying the MLPG algorithm, correlations between adjacent frames are hardly considered. Thus the sequentially changing nature of acoustic features is largely ignored.

With this in mind, RNN as a sequence-to-sequence mapping method which is capable of capturing the long-term temporal dependency of sequential data are considered by some reserachers (Fan et al., 2014; Fernandez et al., 2014; Zen & Sak, 2015). Systems based on RNN with bidirectional or unidirectional LSTM can directly output vocoder-ready acoustic parameters for synthesis and the subjective scores are significantly improved. Particularly, an investigation by Wu & King (2016b) shows that the complexity of LSTM can be significantly reduced with only 'forget gate' left in use without degrading perceptual quality in synthesis. And this simplified architecture is similar to the gated recurrent unit (GRU) based RNN (Cho, van Merriënboer, Gülçehre, Bahdanau, Bougares, Schwenk & Bengio, 2014).

While the in-syllable and cross-syllable dynamics of F0 production has been simulated by the TA model, there remains a question as to whether longer-term suprasegmental variations (e.g. phrase or utterance level) still need to be handled. The DNN-TA approach can deal with this to a certain extent since some positional information has been included in the linguistic features and considered during training. However, how much the TA-model can benefit from sequence-to-sequence mapping provided by an RNN is unknown. To this end, we propose a GRU based RNN architecture to assist the TA model to accommodate the dynamics outside the temporal scope of TA.

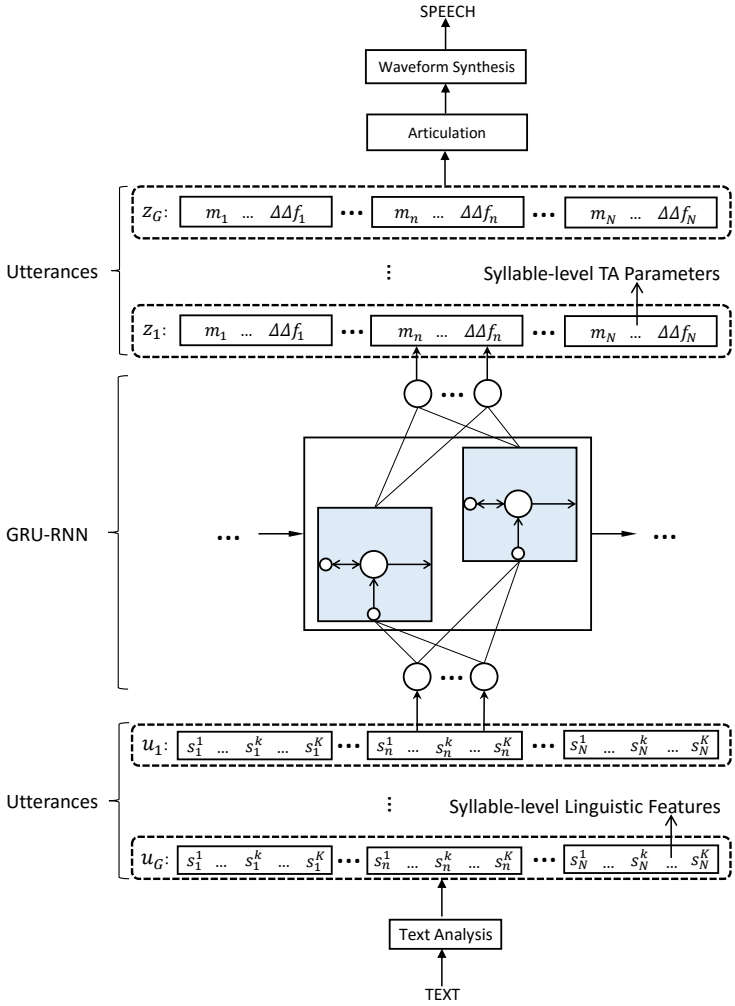


Figure 5.8 RNN-TA system associates syllable-wise linguistic features with corresponding TA parameters for each utterance.

Figure 5.8 shows a diagram of the proposed RNN-TA approach. As we can see, the mapping is organised at the utterance level. Namely, the originally separated syllables of an utterance are now grouped together and embedded in a vector sequence, i.e. $\boldsymbol{u}_g = \left[\boldsymbol{s}_1^\top, \boldsymbol{s}_2^\top, ..., \boldsymbol{s}_N^\top\right]^\top$ for input linguistic features and $\boldsymbol{z}_g = \left[\boldsymbol{v}_1^\top, \boldsymbol{v}_2^\top, ..., \boldsymbol{v}_N^\top\right]^\top$ for output TA parameters.

At the training stage, the GRU-RNN will first unfold itself to align with each syllable and then be trained with the backpropagation through time (BPTT) algorithm (Williams & Peng, 1990). At the synthesis stage, when the syllable sequence of an utterance is given, the trained GRU-RNN will read the entire sequence of linguistic

feature vectors and predict a sequence of TA parameter vectors. Note that here the GRU-RNN still needs to do syllable-by-syllable predictions internally, by accepting the whole sentence as an input and processing the syllables inside it sequentially. Modifications can be done to some deep learning frameworks to make it do stateful predictions (i.e. one syllable at a time), with time dependencies still considered. Therefore, RNN-TA will not be faster than DNN-TA in terms of number of predictions. Nevertheless, the computational cost of RNN-TA may still be lower than DNN-TA since the network size of an RNN can be smaller than a DNN but still achieve similar performance (Zen, 2015). Following the same parameter generation process as in the DNN-TA approach, an F0 contour of the whole utterance can be obtained.

5.5 Experiments

5.5.1 Dataset

A Mandarin Chinese spontaneous speech dataset was used in this experiment, which consisted of 6233 phonetically and prosodically balanced utterances (around 5 hours) as the training set and 60 extra utterances as the test set. Of the 6233 utterances, 701 were questions and the rest were statements. The test set was evenly divided into statements and questions. The dataset was recorded from a female speaker at 22.5kHz/16bit. Spectral analysis was performed with a 25-ms Hamming window shifted every 5 ms. Extracted acoustic features include logarithmic F0 (by the RAPT algorithm (Talkin, 1995)), 31-order Mel-generalised cepstrum (MGC) coefficients as well as their delta and delta-delta. Phone durations were obtained through forced alignment, and the contextual linguistic features include triphone, phone position in word and in phrase, syllable and its position in word and in

phrase, word/phrase length, sentence length, sentence type, phone/syllable stress, prominence, part-of-speech (POS), etc.

5.5.2 System configurations

To test the proposed TA-based approaches, we built five systems for comparison. They were trained with different levels of linguistic features as shown in Table 5.1.

The HMM-baseline system is typical as used in other studies with five-state left-to-right-with-no-skip HMM contextual phone models, and each HMM state is modelled by a single Gaussian output distribution with diagonal covariance. In particular, the log F0s with voiced/unvoiced observations were modelled by multi-space probability distributions (MSD) (Tokuda et al., 1999). The HMM-based system used all the contextual linguistic features. A total number of 5268 questions were developed based on these linguistic features for decision tree-based state clustering. The minimum description length (MDL) criterion factor α was set to 1 (Shinoda & Watanabe, 1997).

The DNN-baseline system is a typical feed-forward neural network with 3 hidden layers and 1500 nodes on each layer. Two extra frame level linguistic features were added to indicate the frame index in the phone and the total number of frames of the phone. These linguistic features were encoded as 485-dimensional input vectors with 57 binary features one-hot encoded and 25 numeric features normalised to zero-mean unit-variance. The output vectors are 4-dimensional with F0 level, velocity, acceleration and voicing flag embedded, which are normalised to $[0.01, 0.99]$ based on their minimum and maximum values extracted from the dataset. The missing F0 values of unvoiced frames were linearly interpolated and they were all used for training. The activation functions used in the network were hyperbolic tangent for the hidden layers and linear for the output layer. The network was trained with the backpropagation algorithm using mini-batch stochastic gradient

Table 5.1 F0 modelling systems with their levels and amount of linguistic feature used as well as input and output dimension of each neural network.

System	Linguistic Feature Level	Linguistic Feature Amount	Input Dimension	Output Dimension
HMM-baseline	all	80	—	—
DNN-baseline	all + frame	82	485	4
DNN-TA	above syllable	58	287	6
RNN-TA (*full*)	above syllable	58	287	6
RNN-TA (*lite*)	above syllable − positional ones	42	190	6

descent (SGD) as the optimiser.[2] Following Zen et al. (2013), the MLPG algorithm were used with dynamic features considered to generate smooth acoustic feature sequences.

Similar to the DNN-baseline, the DNN-TA system is also a three-hidden-layer feed-forward neural network. The difference is that, by using the TA model with fewer linguistic features (only those above the syllable level, see Table 5.1), the burden of the network was significantly alleviated. As a result, the nodes on each layer were reduced to 1024 in the experiment (further reduction of nodes is also possible).

The number of input dimensions of the DNN-TA were 287 formed by 35 binary features with one-hot encoding and 23 numeric features with zero-mean unit-variance normalisation. The outputs were 6-dimensional, including the three TA motor parameters as well as dynamic onset state of the syllable (onset F0 level, velocity and acceleration), which were normalised to $[0.01, 0.99]$. With the help of the LMFIT Python package (Newville et al., 2014), optimal TA parameters could be found for each syllable and used in the outputs. Unvoiced parts are not considered in the experiment except for those missed by the pitch tracker which were recognised by some heuristics and then linearly interpolated. Based on previous studies (Prom-on et al., 2009; Xu & Prom-on, 2014), certain ranges were applied to limit the search range of TA parameters: $m \in [-100, 100]$, $b \in [-30, 30]$ and $\lambda \in [1, 80]$ [3]. To ensure that global optima are found, hundreds of points uniformly distributed in the above ranges were set as the initial conditions for the least-squares algorithm. Figure 5.9 illustrates the performance of the TA model when local optimal parameters were found (m and b are plotted as underlying pitch targets defined in TA, λ is not presented). It needs to be mentioned that the

[2]The deep learning library Keras (Chollet, 2015) with Theano as the backend (Theano Development Team, 2016) was used for building all the neural networks in this experiment.

[3]Some preliminary experiments also found that fixing λ won't degrade natualness very much.

syllabic F0 contours were never length-normalised while searching for the optimal TA parameters. This is different from the work done by Gao et al. (2014). This system followed the same training method as used in the DNN-baseline.

Two versions of RNN-TA system were built with respect to the levels of linguistic features they used. One is the *full* version using the same input linguistic features as in the DNN-TA system, and the other is the *lite* version which also used linguistic features above the syllable level but with all the positional information removed (Table 5.1). The inputs of the RNN-TA (*lite*) system are 190-dimensional converting from 19 binary features and 23 numeric features. Except that, the two versions of RNN-TA shared the same architecture. There was only 1 forward-directed hidden layer with 256 GRU units in the RNN. A feedforward output layer instead of a recurrent one was used in this RNN, which is different from the pioneer study by Zen & Sak (2015). The motivation is that unlike those frame-level acoustic features which need a recurrent output layer for further smoothing, the output TA parameters in our experiment are mostly not static acoustic features and the modelling is conducted at the syllable level so that the benefit from such extra process will be minimal. Mini-batch RMSprop (Tieleman & Hinton, 2012) based backpropagation through time (BPTT) algorithm was used to train both systems.

The basic heuristics that we used to determine these neural network architectures was not complex. The latest standard as reported in recent pioneering studies (e.g. Qian et al. (2014) and Zen & Sak (2015)) was followed in the beginning. The number of neurons on each hidden layer was increased first to see if any performance gain is possible. If not, the number of hidden layers was then increased. We found that the three-hidden-layer DNN was already sufficient for the purpose of F0 modelling and any RNN architecture with more than one hidden layer turned out to be redundant. Also, we would like to emphasise that choosing similar architectures as used in other studies is more useful to testify the effectiveness of our F0 modelling

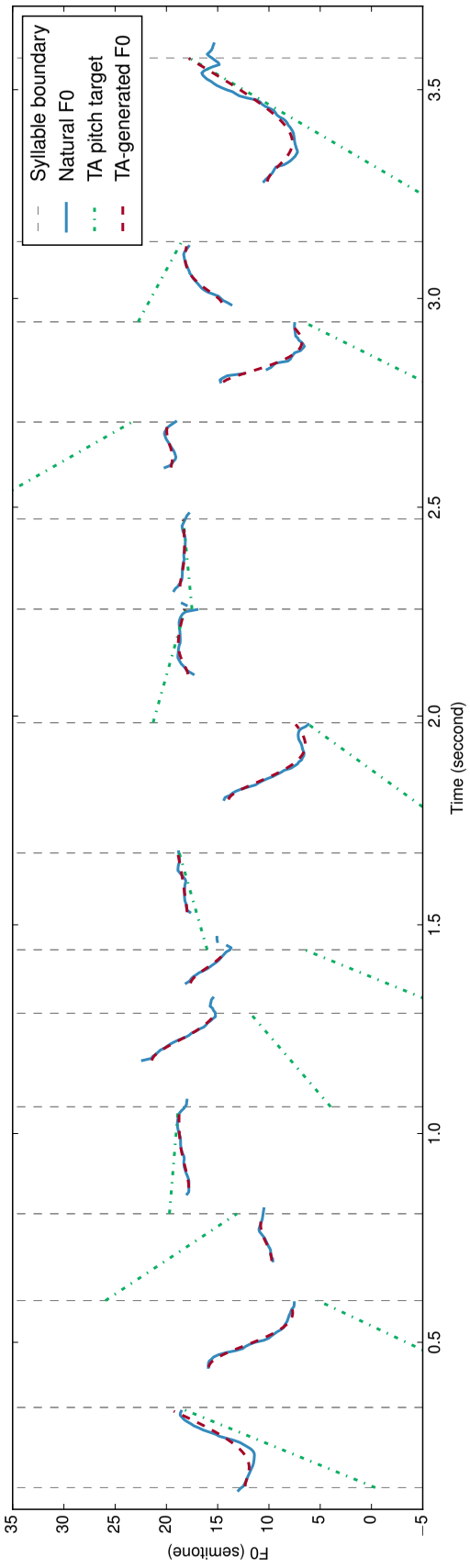


Figure 5.9 Syllabified natural F0 contours and those generated by the TA model with optimal motor parameters (a training set statement sentence).

approach. In other words, if we can achieve better evaluation scores than others with similar or even simpler DNN or RNN architectures, the proposed F0 modelling approach can be seen as useful. Therefore, not a lot effort was put on exploring more powerful neural network architectures in this study.

5.6 Evaluations

As mentioned earlier, the test set consisted of 30 statements and 30 questions. The evaluations were therefore run separately for them to show differences in performance. Both objective and subjective evaluations were conducted. During synthesis, durations obtained through forced alignment on the test set were used for all the systems. The TA-based systems generally followed the voicing decisions predicted by the HMM-baseline system.

5.6.1 Objective evaluation

For the objective test, F0 discrepancies between natural and synthetic speech were measured with root mean square errors (RMSE) in Hz as well as correlation scores for each system in each task.

Table 5.2 Objective scores of each system on different sentence types. Lowest RMSE scores are in bold.

System	Statement		Question	
	RMSE	Corr.	RMSE	Corr.
HMM-baseline	22.32	0.91	33.67	0.85
DNN-baseline	22.17	0.92	32.19	0.85
DNN-TA	**21.10**	0.91	33.20	0.86
RNN-TA (*full*)	23.58	0.90	**29.95**	0.87
RNN-TA (*lite*)	23.18	0.89	30.69	0.87

As shown in Table 5.2, the systems generally performed better in the statement task than in the question task (lower RMSE and higher correlation). This is reasonable due to the fact that the number of question sentences in the training data is relatively limited, and more importantly F0 contours tend to be more stable in statements than in questions. For this particular test set, the F0 contours of questions were very different from those in the training set that the systems were trained on. Looking closely, in the statement task, the DNN-TA system achieved the lowest RMSE score followed by the two baseline systems, whereas the scores for the two RNN-TA systems are relatively higher. However, the situation is reversed in the question task, where the RNN-TA systems achieved significantly lower RMSE scores than the rest systems. While the exact reason for this is still unclear, it is noteworthy that the objective test scores are not highly indicative of system performance in terms of perceptual quality. It has been found in other studies (Zen et al., 2013; Yin et al., 2016) that, some systems using deep neural networks outperformed baseline systems in the subjective test but achieved higher RMSE scores in the objective test. Similar situations can also be found in the following subjective tests.

5.6.2 Subjective evaluation

Subjective tests were focused on comparing naturalness of sentence prosody only. Similar to the objective test, statements and questions were tested separately. Subjects were asked to do A/B preference test based on the synthetic sentence pairs that they heard. The outcome of the test is computed in terms of percentages of preference expressed for each system (A, B or 'no preference'). A one-tailed binomial test with an expected 50% split was applied to check if one system is significantly preferred than the other. Note that in order to enable the test, the 'no preference' votes need to be split equally over the two systems to simulate a forced

choice situation in which people who really have no preference can be expected to vote for A as often as for B (Eskenazi, Levow, Meng, Parent & Suendermann, 2013). Fifteen sentence pairs for each sentence type were randomly selected from the test set. Twenty native speakers participated in the evaluation.

Table 5.3 Subjective preference scores (%) of each system on **statements**. In each paired test, the system achieved significantly better preference than the other ($p < 0.01$) is in bold. N/P stands for 'no preference'.

			RNN-TA			
HMM	DNN	DNN-TA	*full*	*lite*	N/P	p-value
18.4	**30.0**	—	—	—	51.6	2.98×10^{-5}
15.8	—	**32.0**	—	—	52.2	1.22×10^{-5}
—	21.4	**30.3**	—	—	48.3	1.71×10^{-3}
—	—	12.1	**23.6**	—	64.3	4.03×10^{-6}
—	—	—	17.6	11.9	70.5	0.38

Table 5.4 Subjective preference scores (%) of each system on **questions**. In each paired test, the system achieved significantly better preference than the other ($p < 0.01$) is in bold. N/P stands for 'no preference'.

			RNN-TA			
HMM	DNN	DNN-TA	*full*	*lite*	N/P	p-value
22.0	**35.2**	—	—	—	42.8	4.5×10^{-6}
9.5	—	**67.0**	—	—	23.5	6.74×10^{-14}
—	17.4	**56.1**	—	—	26.5	1.42×10^{-12}
—	—	21.8	**32.9**	—	45.3	4.32×10^{-3}
—	—	—	27.1	21.0	51.9	0.67

The preference scores are shown in Table 5.3 and Table 5.4 with p values obtained from the binomial tests. It can be seen that the situations are quite similar between the statement and question tasks. That is, while the DNN-baseline system is perceptually better than the HMM-baseline, the DNN-TA system outperformed both of them. In particular, the DNN-TA system almost doubled its preference scores

in the question task from the statement task. More importantly, both tasks suggest that the RNN-TA system with full set of syllable-level linguistic features performed significantly better than the DNN-TA system. This indicates that the GRU-RNN architecture did help the TA model to simulate some long-term suprasegmental dynamics of F0 production. Another interesting finding is that the preference scores of the RNN-TA (*lite*) system for the two tasks are not significantly different from the RNN-TA (*full*) system. This indicates that the GRU-RNN is capable of capturing long-term dependencies between syllables without relying on any positional information in the linguistic features. Although this actually reflects the sequence-to-sequence mapping nature of RNN, to our knowledge, so far full linguistic features are still used in other RNN-based TTS studies. The reason might be that those studies are still focusing on frame-level modelling and the coverage of the RNN is therefore limited to phone level only (Zen & Sak, 2015), which does not eliminate the need for positional information in the linguistic features. In contrast, the combination of RNN and TA enables an easy utterance-level modelling.

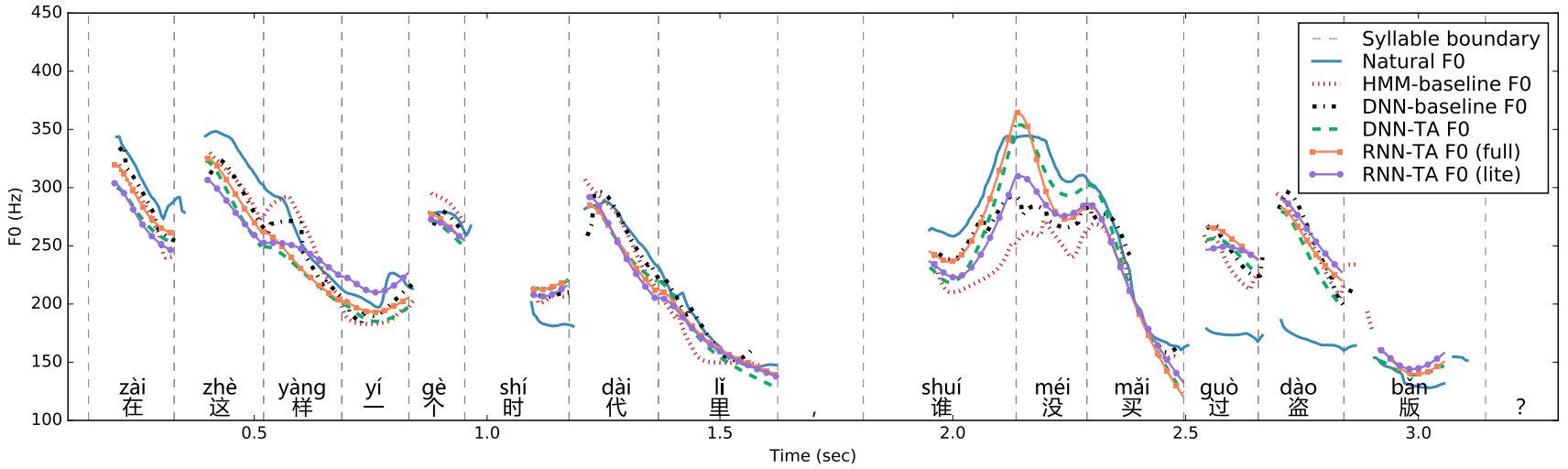


Figure 5.10 Syllabified natural F0 contours together with those generated by the five systems in the experiment (a test set question sentence).

As an illustration, a comparison between F0 contours generated by these systems is shown in Figure 5.10. The sentence is a question selected from the test set. It can be seen that the F0 contours generated by the TA-based systems generally show much closer resemblance to the natural ones than those generated by the baseline systems. Also the TA-based trajectories in each syllable are stable and smooth, which are different from those generated by the baseline systems where unreasonable perturbations are often seen due to their independent state/frame modelling nature.

More importantly, Figure 5.10 also reveals an interesting issue. As we can see that, near the end of the sentence, all the systems predicted much higher than natural F0 values for the words '过' and '盗'. Looking closely, we should notice that they are preceded (though not immediately) by the question word '谁'. From the functional perceptive, question words are usually 'focused' (Cooper, Eady & Mueller, 1985; Xu, 1999; Pell, 2001). And there are studies showing that prosodic focus is realised not only by increasing F0, duration, intensity and upper spectral energy on the focused component itself, but also by compressing the pitch range and intensity of the post-focus components (Cooper et al., 1985; Xu, 1999; Pell, 2001; Xu, 2005; Xu & Xu, 2005). This post-focus effect is called post-focus compression (PFC) (Xu, 2011). In the example displayed here, the question word '谁' (or possibly together with the following negative adverb '没') may be focused. Therefore, a possible account can be made here that during natural production '过' and '盗' were affected by PFC so that their pitch levels were lowered and pitch ranges were compressed (as shown in the figure). However, the absence of functional labelling of 'focus' in the dataset made no system predict correct values for these post-focus components. This finding suggests that functional (or pragmatic) labelling is really in need for better TTS.

5.7 Discussion

The comparison of the five F0 generation systems has shown that the TA-based systems achieved significantly better performance than the baseline systems. This provides support for our hypothesis that data-driven computer speech synthesis can be simulated as a two-stage 'linguistic-motor-acoustic' process. It also demonstrates that F0 modelling can be done entirely at the syllable level. More notably, the TA-based systems were found to be much more efficient than the baseline systems. Given its significance, this enhanced efficiency deserves more in-depth discussion.

The efficiency enhancement of the TA model is twofold. The first is the substantial reduction of training data size. In a typical DNN-based approach, linguistic and acoustic features need to be collected for every frame, resulting in a massive training data set. The TA-based approach, in contrast, requires only a subset of the linguistic features, and only three TA parameters for each syllable. The training dataset used in this experiment for the DNN-baseline system was 10.4 GB ($3,531,502$ entries), but the one for the DNN-TA system was merely 221 MB ($79,691$ entries). Smaller training data naturally leads to a faster loop for each training epoch.

Table 5.5 Model complexities of the systems for F0 modelling. (M stands for million.)

System	Parameter Amount
HMM-baseline	0.34M
DNN-baseline	5.23M
DNN-TA	2.40M
RNN-TA (*full*)	0.42M
RNN-TA (*lite*)	0.35M

The second efficiency enhancement is the reduction of model complexity. As can be seen in Table 5.5, within the neural network family the typical frame-based DNN-baseline system is powered by a fairly large network and therefore is the most

computationally expensive during both training and synthesis. The syllable-based DNN-TA system, on the other hand, is capable of producing better results with only half of the parameters. The two RNN-TA systems are the smallest in network size, comparable even to the HMM-baseline. Note, however, because only F0 modelling is considered here, the number of parameters of the HMM-baseline system is very small, with only 9 F0-related parameters in each decision tree node (i.e. means, variances and MSD weights).

Lastly, as may be noticed from the evaluations, the further improvement introduced by the RNN architecture relative to the DNN-TA system is not as sizeable as in other studies (Fan et al., 2014; Zen & Sak, 2015). A possible reason is that the RNN model is used in those studies to capture some in-syllable dynamics of F0 production, but that is already largely simulated by the TA model, as explained in Chapter 3. What the RNN has captured in this study is only those contextual variations that are not due to articulatory dynamics. The nature of these inter-syllable non-articulatory variations can be further examined in future research.

5.8 Summary

In this chapter, we attempted to tackle the issues in speech synthesis from an articulatory perspective and proposed a 'linguistic-motor-acoustic' two-stage synthesis approach. The model is the target approximation (TA) model, which serves as the link between linguistic features and surface acoustics. With TA, syllable is the basic prosody modelling unit instead of frame, which greatly increases processing efficiency and addresses the notorious frame-by-frame independence issue within the current typical DNN-based synthesis frameworks. The TA model is first driven by a DNN to test its efficacy. A GRU-RNN architecture is then experimentally adopted in order to capture the dynamics of F0 production beyond the syllable

level. The number of linguistic features can be significantly reduced with the proposed approach. Thanks to the sequence-to-sequence mapping power of RNN, the evaluation results were further improved.

Chapter 6

Conclusions

6.1 Summary

One of the major issues in current statistical parametric speech synthesis (SPSS) approaches is that they typically aim at state/frame-level acoustic modelling so that the dependencies between neighbouring acoustic frames, which are so important in speech, are largely ignored. Although the MLPG algorithm is able to generate smooth trajectories by considering dynamic features, these features themselves could be inconsistently predicted. Besides that, the SPSS approaches generally set up a direct mapping from linguistic to acoustic features and are purely statistical data-driven without exploiting any mechanisms found in speech production research. Therefore, they are not flawless in terms of biological plausibility.

To address these issues, this thesis proposed and tested a more human-like F0 modelling paradigm by integrating dynamic mechanisms of human speech production as a core component of F0 generation. The proposed F0 modelling paradigm operates on the syllable level with an articulatory model, target approximation (TA), which largely simulates the dynamic process of F0 production. The proposed

paradigm is two-staged: the linguistic-to-motor stage links linguistic features to TA motor parameters, and the motor-to-acoustic stage is the TA model.

On the motor-to-acoustic stage, a simulation experiment was conducted, which for the first time successfully replicated the pitch-level online auditory feedback compensation behaviour of human with systematic TA-based pitch control. This experiment supports the idea that human speech movement can be considered as a dynamic process of target approximation and the TA model is a valid F0 generation model that can be used for speech synthesis. Moreover, this study also opens the possibility of further research on pitch-related neural control mechanisms in the brain through TA-based computational modelling.

On the linguistic-to-motor stage, the extracted TA parameters from a spontaneous speech dataset were associated to syllable-level linguistic features by training deep or recurrent neural networks (DNN/RNN), which completed the proposed F0 modelling pipeline. We trained five systems on a Mandarin Chinese dataset consisting of both statements and questions. The five systems are: HMM-based baseline, DNN-based baseline, DNN-TA, RNN-TA and RNN-TA without positional features. The TA-based systems generally outperformed the baseline systems in both objective and subjective evaluations, and more importantly we showed that the RNN-TA system was able to abandon all the positional features without degrading synthesis quality. In general, the TA-based F0 modelling approach used fewer linguistic features than existing SPSS approaches, which led to less training data and lower model complexity and in turn led to shortened training time and a faster synthesis process.

As discussed in the thesis, the TA model is quantitatively implemented as a dynamical system which makes it capable of generating trajectories that are close to the reality. Once the TA parameters are learned, the acoustic contexts are largely preserved, which is the major benefit of using the TA model. Compared with other

studies addressing the same frame-by-frame independence issue of SPSS, there are at least two advantages of the TA-based approach. The first is that the procedure of TA-based modelling is simple, which only requires an extra step of extracting TA parameters from the dataset. In contrast, the studies relying on DCT transformation to capture inter-frame dependencies not only need to extract multi-level DCT parameters in advance, but also require training multiple models to hierarchically predict the frame-level residuals and the high-level DCT parameters (Yin et al., 2014, 2016). The second is that the TA-based modelling is computationally less expensive since the model is trained at the syllable level with much less data than at the frame level. In contrast, the studies that apply the minimum generation error (MGE) training criterion (Fan et al., 2015; Wu & King, 2015, 2016a) are primarily still based on frame-level modelling, the only difference is that they minimise utterance-level vocoder parameter trajectory errors rather than frame-level acoustic feature errors. However, the utterance-level trajectories are actually the output of the MLPG algorithm, which means that the algorithm needs to be applied iteratively during training.

Our use of the GRU-RNN is still experimental. However, the results are very informative which suggest that with this architecture the positional information included in the input linguistic features is somehow redundant. This finding implies that long-term variations mostly come from functional contrasts, which are the truly important properties that should be controlled by input features.

6.2 Limitations and Future Directions

Syllable segmentation. The accuracy of syllable segmentation affects synthesis quality of the TA-based F0 modelling approach proposed in this thesis. While the syllables were identified automatically with manual corrections in the experiments,

the phone boundaries were determined simply by forced alignment. Those misaligned boundaries directly lead to bad individual local fitting results during TA parameter extraction, and may also be harmful to the final learning outcome if such errors accumulate. Besides that, some very short syllables were inevitably wasted. It can be easily imagined that when a very short syllable is encountered (e.g. fewer than 5 samples given by the pitch tracker), finding reliable TA parameters for it would be rather difficult. As a consequence, these short syllables were usually dropped in the experiments. Although abandoning these short syllables is usually not that harmful to the final learning outcome, frame-level modelling approach may escape from this situation.

Chain effect. As introduced in Section 3.3, the TA model transfers the offset state of the preceding syllable to the current one. As a consequence, if the preceding syllable is badly predicted during synthesis, its improper offset will be transferred to the current syllable and deviate the current syllable to some extent. Similarly, for syllables starting with voiceless consonant, the predicted onset states play a critical role in shaping their F0 contours during synthesis. A badly predicted onset state will lead to a less satisfactory F0 contour of the syllable. And it is not easy to recover from such errors.

Functional labelling. As we have discussed, functional labelling is critical to speech synthesis. However, only a very limited number of functional labels were used in this thesis. While those non-functional (e.g. contextual) labels consumed a large proportion of computation power, their contribution can actually be taken over by the TA model and RNN.

There are two possible directions of future work. The first is to extend the use of TA to spectrum modelling. Given the success of the pilot study in synthesising basic CV utterances (Prom-on et al., 2013), it is reasonable to anticipate the efficacy of TA in directly reproducing some of the spectral trajectories. The second is to

explore ways to improve current frontends with more focus on natural language understanding in order to make them capable of providing more functional labels for TTS. Along this goal, training data also need to be improved. Instead of only including isolated statements and questions, more engaged and sophisticated discourse utterances are needed.

Bibliography

Agarwal, S., Mierle, K. et al. (2010). Ceres Solver. http://ceres-solver.org.

Allen, J., Hunnicutt, M. S., Klatt, D. H., Armstrong, R. C. & Pisoni, D. B. (1987). *From text to speech: The MITalk system*. Cambridge University Press.

Armson, J., Foote, S., Witt, C., Kalinowski, J. & Stuart, A. (1997). Effect of frequency altered feedback and audience size on stuttering. *Eur. J. Disord. Commun. 32*(3), 359–366.

Arvaniti, A. & Ladd, D. R. (2009). Greek wh-questions and the phonology of intonation. *Phonology*, *26*(1), 43–74.

Arvaniti, A. & Ladd, D. R. (2015). Underspecification in intonation revisited: A reply to Xu, Lee, Prom-on and Liu. *Phonology*, *32*(3), 537–541.

Bailly, G. & Gorisch, I. (2006). Generating German intonation with a trainable prosodic model. In *Proc. Interspeech* (pp. 2366–2369).

Bailly, G. & Holm, B. (2005). SFC: a trainable prosodic model. *Speech Commun. 46*(3), 348–364.

Bauer, J. J. & Larson, C. R. (2003). Audio-vocal responses to repetitive pitch-shift stimulation during a sustained vocalization: Improvements in methodology for the pitch-shifting technique. *J. Acoust. Soc. Am. 114*, 1048–1054.

Beckman, M. E. & Pierrehumbert, J. B. (1986). Intonational structure in Japanese and English. *Phonology*, *3*(1), 255–309.

Birkholz, P. (2013). Modeling consonant-vowel coarticulation for articulatory speech synthesis. *PLoS ONE*, *8*(4), 1–17.

Birkholz, P., Jackèl, D. & Kröger, B. J. (2006). Construction and control of a three-dimensional vocal tract model. In *Proc. ICASSP* (Vol. 1, pp. 873–876).

Birkholz, P., Kröger, B. J. & Neuschaefer-Rub, C. (2011). Model-based reproduction of articulatory trajectories for consonant-vowel sequences. *IEEE Trans. Audio, Speech and Lang. Process*, *19*(5), 1422–1433.

Black, A. W. & Hunt, A. J. (1996). Generating F0 contours from ToBI labels using linear regression. In *Proc. ICSLP* (Vol. 3, pp. 1385–1388).

Black, J. W. (1951). The effect of delayed side-tone upon vocal rate and intensity. *J. Speech Hear. Disord. 16*, 56–60.

Boersma, P. (2002). Praat, a system for doing phonetics by computer. *Glot International*, *5*(9/10), 341–345.

Borden, G. J. (1979). An interpretation of research on feedback interruption in speech. *Brain Lang. 7*(3), 307–319.

Brainard, M. S. & Doupe, A. J. (2000). Auditory feedback in learning and maintenance of vocal behaviour. *Nature Rev. Neurosci. 1*(1), 31–40.

Burnett, T. A., Freedland, M. B., Larson, C. R. & Hain, T. C. (1998). Voice F0 responses to manipulations in pitch feedback. *J. Acoust. Soc. Am. 103*(6), 3153–3161.

Burnett, T. A. & Larson, C. R. (2002). Early pitch-shift response is active in both steady and dynamic voice pitch control. *J. Acoust. Soc. Am. 112*, 1058–1063.

Burnett, T. A., Senner, J. E. & Larson, C. R. (1997). Voice F0 responses to pitch-shifted auditory feedback: A preliminary study. *J. Voice*, *11*(2), 202–211.

Cai, S. (2012). *Online control of articulation based on auditory feedback in normal speech and stuttering: behavioral and modeling studies* (Doctoral dissertation, Massachusetts Institute of Technology).

Cai, S., Ghosh, S. S., Guenther, F. H. & Perkell, J. S. (2011). Focal manipulations of formant trajectories reveal a role of auditory feedback in the online control of both within-syllable and between-syllable speech timing. *J. Neurosci. 31*, 16483–16490.

Chang, E. F., Niziolek, C. A., Knight, R. T., Nagarajan, S. S. & Houde, J. F. (2013). Human cortical sensorimotor network underlying feedback control of vocal pitch. *Proc. Natl. Acad. Sci. 110*(7), 2653–2658.

Chase, R. A., Sutton, S. & Rapin, I. (1961). Sensory feedback influences on motor performance. *J. Aud. Res. 1*, 212–223.

Chen, G.-P., Bailly, G., Liu, Q.-F. & Wang, R.-H. (2004). A superposed prosodic model for Chinese text-to-speech synthesis. In *Proc. ISCSLP* (pp. 177–180).

Chen, S. H., Liu, H., Xu, Y. & Larson, C. R. (2007). Voice F0 responses to pitch-shifted voice feedback during English speech. *J. Acoust. Soc. Am. 121*, 1157–1163.

Chen, Y. & Xu, Y. (2006). Production of weak elements in speech – Evidence from F0 patterns of neutral tone in standard Chinese. *Phonetica*, *63*(1), 47–75.

Chen, Z., Liu, P., Jones, J. A., Huang, D. & Liu, H. (2010). Sex-related differences in vocal responses to pitch feedback perturbations during sustained vocalization. *J. Acoust. Soc. Am. 128*(6), 355–360.

Cherry, C. & Sayers, B. M. (1956). Experiments upon the total inhibition of stammering by external control, and some clinical results. *J. Psychosom. Res. 1*(4), 233–246.

Cho, K., van Merriënboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H. & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proc. EMNLP* (pp. 1724–1734).

Chollet, F. (2015). Keras. https://github.com/fchollet/keras. GitHub.

Chung, J., Gülçehre, Ç., Cho, K. & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *CORR*. Retrieved from http://arxiv.org/abs/1412.3555

Clark, R. A. J., Richmond, K. & King, S. (2007). Multisyn: Open-domain unit selection for the Festival speech synthesis system. *Speech Commun. 49*(4), 317–330.

Cooper, W. E., Eady, S. J. & Mueller, P. R. (1985). Acoustical aspects of contrastive stress in question–answer contexts. *J. Acoust. Soc. Am. 77*(6), 2142–2156.

Cooper, W. E. & Sorensen, J. M. (2012). *Fundamental Frequency in Sentence Production*. Springer Science & Business Media.

Cowie, R., Douglas-Cowie, E. & Kerr, A. (1982). A study of speech deterioration in post-lingually deafened adults. *J. Laryngol. Otol. 96*(2), 101–112.

Curlee, R. F. & Perkins, W. H. (1969). Conversational rate control therapy for stuttering. *J. Speech Hear. Disord. 34*, 245–250.

Davis, B. L. & MacNeilage, P. F. (1995). The articulatory basis of babbling. *J. Speech Lang. Hear. Res. 38*(6), 1199–1211.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Series B Stat. Methodol. 39*(1), 1–38.

d'Imperio, M. & House, D. (1997). Perception of questions and statements in Neapolitan Italian. In *Proc. Eurospeech* (pp. 251–254).

Donath, T. M., Natke, U. & Kalveram, K. T. (2002). Effects of frequency-shifted auditory feedback on voice F0 contours in syllables. *J. Acoust. Soc. Am. 111*, 357–366.

Dusterhoff, K. & Black, A. W. (1997). Generating F0 contours for speech synthesis using the Tilt intonation theory. In *Proc. INT* (pp. 107–110).

Elman, J. L. (1981). Effects of frequency-shifted feedback on the pitch of vocal productions. *J. Acoust. Soc. Am. 70*(1), 45–50.

Eskenazi, M., Levow, G.-A., Meng, H., Parent, G. & Suendermann, D. (2013). *Crowdsourcing for speech processing: Applications to data collection, transcription and assessment*. John Wiley & Sons.

Fan, Y., Qian, Y., Soong, F. K. & He, L. (2015). Sequence generation error (SGE) minimization based deep neural networks training for text-to-speech synthesis. In *Proc. Interspeech* (pp. 864–868).

Fan, Y., Qian, Y., Xie, F.-L. & Soong, F. K. (2014). TTS synthesis with bidirectional LSTM based recurrent neural networks. In *Proc. Interspeech* (pp. 1964–1968).

Fechner, G. T. (1966). *Elements of Psychophysics. [Translated by Helmut E. Adler]* (D. H. Howes & E. G. Boring, Eds.). New York: Holt, Rinehart and Winston.

Feder, M. & Weinstein, E. (1988). Parameter estimation of superimposed signals using the EM algorithm. *IEEE Trans. Audio, Speech and Lang. Process. 36*(4), 477–489.

Fernandez, R., Rendel, A., Ramabhadran, B. & Hoory, R. (2014). Prosody contour prediction with long short-term memory, bi-directional, deep recurrent neural networks. In *Proc. Interspeech* (pp. 2268–2272).

Forney, G. D. (1973). The Viterbi algorithm. *Proc. IEEE*, *61*(3), 268–278.

Fujisaki, H. (1983). Dynamic characteristics of voice fundamental frequency in speech and singing. In P. F. MacNeilage (Ed.), *The Production of Speech* (pp. 39–55). Springer.

Fujisaki, H. (2004). Information, prosody, and modeling - with emphasis on tonal features of speech. In *Proc. Speech Prosody* (pp. 1–10).

Fujisaki, H. & Hirose, K. (1982). Modelling the dynamic characteristics of voice fundamental frequency with application to analysis and synthesis of intonation. In *Proc. ICL* (pp. 57–70).

Fujisaki, H. & Hirose, K. (1984). Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *J. Acoust. Soc. Jp. (E)*, *5*(4), 233–242.

Fujisaki, H. & Kawai, H. (1988). Realization of linguistic information in the voice fundamental frequency contour of the spoken Japanese. In *Proc. ICASSP* (pp. 663–666).

Gandour, J., Potisuk, S. & Dechongkit, S. (1994). Tonal coarticulation in Thai. *J. Phonet. 22*(4), 477–492.

Gao, L., Ling, Z.-H., Chen, L.-H. & Dai, L.-R. (2014). Improving F0 prediction using bidirectional associative memories and syllable-level F0 features for HMM-based Mandarin speech synthesis. In *Proc. ISCSLP* (pp. 275–279).

Goldiamond, I. (1965). Stuttering and fluency as manipulatable operant response classes. In L. Krasner & L. Ullman (Eds.), *Research in Behavior Modification* (pp. 106–156). New York: Holt, Rhinehart and Winston.

Goldsmith, J. A. (1990). *Autosegmental and Metrical Phonology*. Basil Blackwell.

Gracco, V. L., Ross, D., Kalinowski, J. & Stuart, A. (1994). Articulatory changes following spectral and temporal modifications in auditory feedback. *J. Acoust. Soc. Am. 95*(5), 2821–2821.

Guenther, F. H. (1995). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychol. Rev. 102*(3), 594–621.

Guenther, F. H., Ghosh, S. S. & Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain Lang. 96*(3), 280–301.

Guenther, F. H. & Perkell, J. S. (2004). A neural model of speech production and its application to studies of the role of auditory feedback in speech. In B. Maassen, R. D. Kent, H. F. M. Peters, P. H. H. M. van Lieshout & W. Hulstijn (Eds.), *Speech Motor Control in Normal and Disordered Speech* (pp. 29–49). Oxford University Press.

Guenther, F. H. & Vladusich, T. (2012). A neural theory of speech acquisition and production. *Journal of Neurolinguist. 25*(5), 408–422.

Hain, T. C., Burnett, T. A., Kiran, S., Larson, C. R., Singh, S. & Kenney, M. K. (2000). Instructing subjects to make a voluntary response reveals the presence of two components to the audio-vocal reflex. *Exp. Brain Res. 130*, 133–141.

Ham, R. & Steer, M. (1967). Certain effects of alterations in auditory feedback. *Folia Phoniatrica et Logopaedica*, *19*(1), 53–62.

Hirose, K., Sato, K., Asano, Y. & Minematsu, N. (2005). Synthesis of F0 contours using generation process model parameters predicted from unlabeled corpora: Application to emotional speech synthesis. *Speech Commun. 46*(3), 385–404.

Hirst, D. J. (2005). Form and function in the representation of speech prosody. *Speech Commun. 46*(3), 334–347.

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural Comput. 9*(8), 1735–1780.

Holmes, J. N., Mattingly, I. G. & Shearme, J. N. (1964). Speech synthesis by rule. *Lang. Speech*, *7*(3), 127–143.

Houde, J. F. & Jordan, M. I. (1998). Sensorimotor adaptation in speech production. *Science*, *279*, 1213–1216.

Houde, J. F. & Jordan, M. I. (2002). Sensorimotor Adaptation of Speech I: Compensation and Adaptation. *J. Speech Lang. Hear. Res. 45*, 295–310.

Houde, J. F. & Nagarajan, S. S. (2011). Speech production as state feedback control. *Front. Hum. Neurosci. 5*(82).

Howard, I. S. & Messum, P. (2011). Modeling the development of pronunciation in infant speech acquisition. *Motor Control*, *15*(1), 85–117.

Howell, P. (2002). The EXPLAN theory of fluency control applied to the treatment of stuttering. In E. Fava (Ed.), *Current Issues in Linguistic Theory Series: Pathology and Therapy of Speech Disorders* (pp. 95–118). Amsterdam: John Benjamins.

Howell, P. (2004). Effects of delayed auditory feedback and frequency-shifted feedback on speech control and some potentials for future development of prosthetic aids for stammering. *Stammering Research*, *1*(1), 31–46.

Howell, P. & Archer, A. (1984). Susceptibility to the effects of delayed auditory feedback. *Percept. Psychophys. 36*(3), 296–302.

Howell, P. & Powell, D. J. (1984). Hearing your voice through bone and air: Implications for explanations of stuttering behavior from studies of normal speakers. *J. Fluency Disord. 9*(4), 247–263.

Howell, P., Powell, D. J. & Khan, I. (1983). Amplitude contour of the delayed signal and interference in delayed auditory feedback tasks. *J. Exp. Psychol.–Hum. Percept. Perform. 9*(5), 772.

Howell, P. & Sackin, S. (2002). Timing interference to speech in altered listening conditions. *J. Acoust. Soc. Am. 111*(6), 2842–2852.

Howell, P., El-Yaniv, N. & Powell, D. J. (1987). Factors affecting fluency in stutterers when speaking under altered auditory feedback. In H. F. M. Peters & W. Hulstijn (Eds.), *Speech Motor Dynamics in Stuttering* (pp. 361–369). Springer.

Howell, P. & Au-Yeung, J. (2002). The EXPLAN theory of fluency control applied to the diagnosis of stuttering. In E. Fava (Ed.), *Current Issues in Linguistic Theory Series: Pathology and Therapy of Speech Disorders* (pp. 75–94). Amsterdam: John Benjamins.

Huckvale, M. (2002). Speech synthesis, speech simulation and speech science. In *Proc. ICSLP* (pp. 1261–1264).

Huckvale, M. (2011). Recording caregiver interactions for machine acquisition of spoken language using the KLAIR virtual infant. In *Proc. Interspeech* (pp. 3277–3280).

Huckvale, M., Howard, I. S. & Fagel, S. (2009). KLAIR: a virtual infant for spoken language acquisition research. In *Proc. Interspeech* (pp. 696–699).

Hunt, A. J. & Black, A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *Proc. ICASSP* (pp. 373–376).

Jones, J. A. & Munhall, K. G. (2002). The role of auditory feedback during phonation: studies of Mandarin tone production. *J. Phonet. 30*, 303–320.

Jones, J. A. & Munhall, K. G. (2005). Remapping auditory-motor representations in voice production. *Curr. Biol. 15*(19), 1768–1772.

Jones, J. A. & Munhall, K. G. (2000). Perceptual calibration of F0 production: Evidence from feedback perturbation. *J. Acoust. Soc. Am. 108*(3), 1246–1251.

Kalinowski, J., Armson, J., Stuart, A. & Gracco, V. L. (1993). Effects of alterations in auditory feedback and speech rate on stuttering frequency. *Lang. Speech*, *36*(1), 1–16.

Kameoka, H., Yoshizato, K., Ishihara, T., Kadowaki, K., Ohishi, Y. & Kashino, K. (2015). Generative modeling of voice fundamental frequency contours. *IEEE/ACM Trans. Audio Speech Lang. Process. 23*(6), 1042–1053.

Karaali, O., Corrigan, G., Gerson, I. & Massey, N. (1997). Text-to-speech conversion with neural networks: A recurrent TDNN approach. In *Proc. Eurospeech* (pp. 561–564).

Katseff, S., Houde, J. & Johnson, K. (2012). Partial compensation for altered auditory feedback: A tradeoff with somatosensory feedback? *Lang. Speech*, *55*(2), 295–308.

Kawahara, H. (1993). Transformed auditory feedback: Effects of fundamental frequency perturbation. *J. Acoust. Soc. Am. 94*(3), 1883–1884.

Kawahara, H., Masuda-Katsuse, I. & de Cheveigne, 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 Commun. 27*(3), 187–207.

Kochanski, G. & Shih, C. (2000). Stem-ML: Language-independent prosody description. In *Proc. ICSLP* (Vol. 3, pp. 239–242).

Kochanski, G. & Shih, C. (2001). Automated modeling of Chinese intonation in continuous speech. In *Proc. Eurospeech* (pp. 911–914).

Kochanski, G. & Shih, C. (2003). Prosody modeling with soft templates. *Speech Commun. 39*(3), 311–352.

Kohler, K. J. (1990). Macro and micro F0 in the synthesis of intonation. In J. Kingston & M. E. Beckman (Eds.), *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech* (pp. 115–138). Cambridge University Press.

Koopmans-van Beinum, F. J., Clement, C. J., Den Dikkenberg-Pot, V. et al. (2001). Babbling and the lack of auditory speech perception: a matter of coordination? *Developmental Science*, *4*(1), 61–70.

Kröger, B. J., Kannampuzha, J. & Neuschaefer-Rube, C. (2009). Towards a neuro-computational model of speech production and perception. *Speech Commun. 51*(9), 793–809.

Lackner, J. R. & DiZio, P. (2005). Motor control and learning in altered dynamic environments. *Curr. Opin. Neurobiol.y*, *15*(6), 653–659.

Ladd, D. R. (1984). Declination: A review and some hypotheses. *Phonology Yearbook*, *1*, 53–74.

Ladd, D. R. (2008). *Intonational Phonology*. Cambridge University Press.

Lane, H. & Tranel, B. (1971). The Lombard sign and the role of hearing in speech. *J. Speech Lang. Hear. Res. 14*(4), 677–709.

Lane, H. & Webster, J. W. (1991). Speech deterioration in postlingually deafened adults. *J. Acoust. Soc. Am. 89*(2), 859–866.

Larson, C. R. (1998). Cross-modality influences in speech motor control: The use of pitch shifting for the study of F0 control. *J. Commun. Disord. 31*(6), 489–503.

Larson, C. R., Burnett, T. A., Bauer, J. J., Kiran, S. & Hain, T. C. (2001). Comparison of voice F0 responses to pitch-shift onset and offset conditions. *J. Acoust. Soc. Am. 110*, 2845–2848.

LeCun, Y., Bengio, Y. & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lee, B. S. (1951). Artificial stutter. *J. Speech Hear. Disord. 15*, 53–55.

Lee, T., Kochanski, G., Shih, C. & Li, Y. (2002). Modeling tones in continuous Cantonese speech. In *Proc. ICSLP* (pp. 2401–2404).

Lei, M., Wu, Y.-J., Soong, F. K., Ling, Z.-H. & Dai, L.-R. (2010). A hierarchical F0 modeling method for HMM-based speech synthesis. In *Proc. Interspeech* (pp. 2170–2173).

Levelt, W. J., Roelofs, A. & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behav. Brain Sci. 22*(1), 1–38.

Liberman, M. Y. (1975). *The intonational system of English.* (Doctoral dissertation, Massachusetts Institute of Technology).

Liberman, M. & Pierrehumbert, J. (1984). Intonational invariance under changes in pitch range and length. In M. Aronoff & R. Oerhle (Eds.), *Language Sound Structure* (pp. 157–233). MIT Press.

Liberman, M. & Prince, A. (1977). On stress and linguistic rhythm. *Linguist. Inq. 8*(2), 249–336.

Ling, Z.-H., Deng, L. & Yu, D. (2013a). Modeling spectral envelopes using restricted Boltzmann machines and deep belief networks for statistical parametric speech synthesis. *IEEE Trans. Audio, Speech and Lang. Process. 21*(10), 2129–2139.

Ling, Z.-H., Kang, S.-Y., Zen, H., Senior, A., Schuster, M., Qian, X.-J., Meng, H. M. & Deng, L. (2015). Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends. *IEEE Signal Process. Mag. 32*(3), 35–52.

Ling, Z.-H., Richmond, K. & Yamagishi, J. (2012). Vowel creation by articulatory control in HMM-based parametric speech synthesis. In *Proc. Interspeech* (pp. 991–994).

Ling, Z.-H., Richmond, K. & Yamagishi, J. (2013b). Articulatory control of HMM-based parametric speech synthesis using feature-space-switched multiple regression. *IEEE Trans. Audio, Speech and Lang. Process. 21*(1), 207–219.

Ling, Z.-H., Richmond, K., Yamagishi, J. & Wang, R.-H. (2008). Articulatory control of HMM-based parametric speech synthesis driven by phonetic knowledge. In *Proc. Interspeech* (pp. 573–576).

Ling, Z.-H., Richmond, K., Yamagishi, J. & Wang, R.-H. (2009). Integrating articulatory features into HMM-based parametric speech synthesis. *IEEE Trans. Audio Speech Lang. Process. 17*(6), 1171–1185.

Liu, F. & Xu, Y. (2006). Parallel encoding of focus and interrogative meaning in Mandarin intonation. *Phonetica*, *62*(2-4), 70–87.

Liu, H. & Larson, C. (2007). Effects of perturbation magnitude and voice F0 level on the pitch-shift reflex. *J. Acoust. Soc. Am. 122*, 3671–3677.

Liu, H., Russo, N. M. & Larson, C. R. (2010). Age-related differences in vocal responses to pitch feedback perturbations: A preliminary study. *J. Acoust. Soc. Am. 127*(2), 1042–1046.

Liu, H. & Xu, Y. (2014). A simplified method of learning underlying articulatory pitch target. In *Proc. Speech Prosody* (pp. 1017–1021).

Liu, H. & Xu, Y. (2015). Simulating online compensation for pitch-shifted auditory feedback with the target approximation model. In *Proc. ICPhS* (Paper no. 0437).

Liu, P., Chen, Z., Jones, J. A., Huang, D. & Liu, H. (2011). Auditory feedback control of vocal pitch during sustained vocalization: A cross-sectional study of adult aging. *PLoS ONE*, *6*(7), 1–8.

Liwicki, M., Graves, A., Bunke, H. & Schmidhuber, J. (2007). A novel approach to on-line handwriting recognition based on bidirectional long short-term memory networks. In *Proc. ICDAR* (Vol. 1, pp. 367–371).

Lombard, E. (1911). Le signe de l' elevation de la voix. *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, *37*(101-119), 25.

Lu, H., King, S. & Watts, O. (2013). Combining a vector space representation of linguistic context with a deep neural network for text-to-speech synthesis. *Proc. SSW-8*, 281–285.

MacDonald, E. N., Goldberg, R. & Munhall, K. G. (2010). Compensations in response to real-time formant perturbations of different magnitudes. *J. Acoust. Soc. Am. 127*, 1059–1068.

MacDonald, E. N., Johnson, E. K., Forsythe, J., Plante, P. & Munhall, K. G. (2012). Children's development of self-regulation in speech production. *Curr. Biol. 22*(2), 113–117.

MacNeilage, P. F. (1970). Motor control of serial ordering of speech. *Psychol. Rev. 77*(3), 182.

MacNeilage, P. F. (1998). The frame/content theory of evolution of speech production. *Behav. Brain. Sci. 21*(04), 499–511.

MacNeilage, P. F. & Davis, B. L. (1993). Motor explanations of babbling and early speech patterns. In B. Boysson-Bardies, S. de Schoen, P. Jusczyk, P. MacNeilage & J. Morton (Eds.), *Developmental neurocognition: Speech and face processing in the first year of life* (pp. 341–352). Springer.

Mann, H. B. & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* 50–60.

MathWorks. (2015). *MATLAB version 8.5.0.197613 (R2015a)*. The Mathworks, Inc. Natick, Massachusetts.

Mixdorff, H. (2000). A novel approach to the fully automatic extraction of Fujisaki model parameters. In *Proc. ICASSP* (pp. 1281–1284).

Mixdorff, H. (2004). Quantitative tone and intonation modeling across languages. In *Proc. TAL* (pp. 137–142).

Moré, J. J. (1978). The Levenberg-Marquardt algorithm: Implementation and theory. In G. A. Watson (Ed.), *Numerical Analysis* (pp. 105–116). Springer.

Moulines, E. & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Commun. 9*, 453–467.

Munhall, K. G., MacDonald, E. N., Byrne, S. K. & Johnsrude, I. (2009). Talkers alter vowel production in response to real-time formant perturbation even when instructed not to compensate. *J. Acoust. Soc. Am. 125*, 384–390.

Na, X. & Garner, P. N. (2013). *Convolutional pitch target approximation model for speech synthesis* (tech. rep. No. EPFL-REPORT-192548). Idiap.

Natke, U., Donath, T. M. & Kalveram, K. T. (2003). Control of voice fundamental frequency in speaking versus singing. *J. Acoust. Soc. Am. 113*, 1587–1593.

Natke, U., Grosser, J. & Kalveram, K. T. (2001). Fluency, fundamental frequency, and speech rate under frequency-shifted auditory feedback in stuttering and nonstuttering persons. *J. Fluency Disord. 26*(3), 227–241.

Natke, U. & Kalveram, K. T. (2001). Effects of frequency-shifted auditory feedback on fundamental frequency of long stressed and unstressed syllables. *J. Speech Lang. Hear. Res. 44*, 577–584.

Neelley, J. N. (1961). A study of the speech behavior of stutterers and nonstutterers under normal and delayed auditory feedback. *J. Speech Hear. Disord. 7*, 63–82.

Newville, M., Stensitzki, T., Allen, D. B. & Ingargiola, A. (2014). LMFIT: Non-Linear Least-Square Minimization and Curve-Fitting for Python. doi:10.5281/zenodo.11813

Niziolek, C. A., Nagarajan, S. S. & Houde, J. F. (2013). Feedback-driven corrective movements in speech in the absence of altered feedback. *J. Acoust. Soc. Am. 134*(5), 4167–4167.

Oller, D. K. & Eilers, R. E. (1988). The role of audition in infant babbling. *Child Dev.* 441–449.

Osberger, M. J. & McGarr, N. S. (1982). Speech production characteristics of the hearing impaired. In N. J. Lass (Ed.), *Speech and Language: Advances in Basic Research and Practice* (Vol. 8, pp. 221–284). New York: Academic Press.

Paine, T. L., Khorrami, P., Chang, S., Zhang, Y., Ramachandran, P., Hasegawa-Johnson, M. A. & Huang, T. S. (2016). Fast Wavenet Generation Algorithm. *CoRR*. Retrieved from http://arxiv.org/abs/1611.09482

Pang, H., Wu, Z. & Cai, L. (2012). Modeling pitch contour of Chinese Mandarin sentences with the PENTA model. *Tsinghua Sci. Technol. 17*(2), 218–224.

Patel, R., Niziolek, C., Reilly, K. & Guenther, F. H. (2011). Prosodic adaptations to pitch perturbation in running speech. *J. Speech Lang. Hear. Res. 54*(4), 1051–1059.

Pell, M. D. (2001). Influence of emotion and focus location on prosody in matched statements and questions. *J. Acoust. Soc. Am. 109*(4), 1668–1680.

Perkell, J. S. (2012). Movement goals and feedback and feedforward control mechanisms in speech production. *J. Neurolinguist. 25*, 382–407.

Pierrehumbert, J. (1981). Synthesizing intonation. *J. Acoust. Soc. Am. 70*(4), 985–995.

Pierrehumbert, J. B. (1980). *The phonology and phonetics of English intonation* (Doctoral dissertation, Massachusetts Institute of Technology).

Pierrehumbert, J. & Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In P. Cohen, J. Morgan & M. Pollack (Eds.), *Intentions in Communication* (pp. 271–311). Cambridge MA.: MIT Press.

Pile, E. J. S., Dajani, H. R., Purcell, D. W. & Munhall, K. G. (2007). Talking under conditions of altered auditory feedback: Does adaptation of one vowel generalize to other vowels. In *ICPhS* (pp. 645–648).

Potisuk, S., Gandour, J. & Harper, M. P. (1997). Contextual variations in trisyllabic sequences of Thai tones. *Phonetica*, *54*(1), 22–42.

Prom-on, S., Birkholz, P. & Xu, Y. (2013). Training an articulatory synthesizer with continuous acoustic data. In *Proc. Interspeech* (pp. 349–353).

Prom-on, S., Xu, Y. & Thipakorn, B. (2009). Modeling tone and intonation in Mandarin and English as a process of target approximation. *J. Acoust. Soc. Am. 125*(1), 405–424.

Purcell, D. W. & Munhall, K. G. (2006a). Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation. *J. Acoust. Soc. Am. 120*, 966–977.

Purcell, D. W. & Munhall, K. G. (2006b). Compensation following real-time manipulation of formants in isolated vowels. *J. Acoust. Soc. Am. 119*, 2288–2297.

Qian, Y., Wu, Z., Gao, B. & Soong, F. K. (2011). Improved prosody generation by maximizing joint probability of state and longer units. *IEEE Trans. Audio, Speech and Lang. Process. 19*(6), 1702–1710.

Qian, Y., Fan, Y., Hu, W. & Soong, F. K. (2014). On the training aspects of Deep Neural Network (DNN) for parametric TTS synthesis. In *Proc. ICASSP* (pp. 3829–3833).

Qian, Y., Liang, H. & Soong, F. K. (2008). Generating natural F0 trajectory with additive trees. In *Proc. Interspeech* (pp. 2126–2129).

Rabiner, L. (1977). On the use of autocorrelation analysis for pitch detection. *IEEE Trans. Audio, Speech and Lang. Process. 25*(1), 24–33.

Raidt, S., Bailly, G., Holm, B. & Mixdorff, H. (2004). Automatic generation of prosody: Comparing two superpositional systems. In *Proc. Speech Prosody* (pp. 417–420).

Raphael, L., Borden, G. & Harris, K. (2011). *Speech Science Primer: Physiology, Acoustics, and Perception of Speech*. Lippincott Williams & Wilkins.

Reddy, V. R. & Rao, K. S. (2013). Two-stage intonation modeling using feedforward neural networks for syllable based text-to-speech synthesis. *Comput. Speech Lang. 27*(5), 1105–1126.

Reddy, V. R. & Rao, K. S. (2016). Prosody modeling for syllable based text-to-speech synthesis using feedforward neural networks. *Neurocomputing*, *171*, 1323–1334.

Rosenberg, A. (2010). AuToBI - a tool for automatic ToBI annotation. In *Proc. Interspeech* (pp. 146–149).

Ross, K. N. & Ostendorf, M. (1999). A dynamical system model for generating fundamental frequency for speech synthesis. *IEEE Trans. Audio, Speech and Lang. Process. 7*(3), 295–309.

Ross, M. & Giolas, T. G. (1978). *Auditory Management of Hearing-impaired Children: Principles and Prerequisites for Intervention*. Baltimore: University Park Press.

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*, 533–536.

Ryan, B. (1974). Programmed stuttering therapy for children and adults. *Springfield: CC Thomas*.

Saltzman, E. L. & Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecol. Psychol. 1*(4), 333–382.

Schuster, M. (1999). *On supervised learning from sequential data with applications for speech recognition* (Doctoral dissertation, Nara Institute of Science and Technology).

Shen, X.-N. S. (1990). *The Prosody of Mandarin Chinese*. University of California Press.

Shih, C. & Kochanski, G. (2003). Modeling intonation: Asking for confirmation in English. In *Proc. ICPhS* (pp. 551–554).

Shinoda, K. & Watanabe, T. (1997). Acoustic modeling based on the MDL criterion for speech recognition. In *Proc. Eurospeech* (pp. 99–102).

Silverman, K. (1986). F0 segmental cues depend on intonation: The case of the rise after voiced stops. *Phonetica*, *43*(1-3), 76–91.

Silverman, K. E., Beckman, M. E., Pitrelli, J. F., Ostendorf, M., Wightman, C. W., Price, P., Pierrehumbert, J. B. & Hirschberg, J. (1992). TOBI: a standard for labeling English prosody. In *Proc. ICSLP* (pp. 867–870).

Sivasankar, M., Bauer, J. J., Babu, T. & Larson, C. R. (2005). Voice responses to changes in pitch of voice or tone auditory feedback. *J. Acoust. Soc. Am. 117*, 850–857.

Smith, C. R. (1975). Residual hearing and speech production in deaf children. *J. Speech Lang. Hear. Res. 18*(4), 795–811.

Soderberg, G. A. (1960). *A study of the effects of delayed side-tone on four aspects of stutterers' speech during oral reading and spontaneous speech* (Doctoral dissertation, Ohio State University).

Sun, X. (2002). *The determination, analysis, and synthesis of fundamental frequency* (Doctoral dissertation, Northwestern University).

Syrdal, A. K. & McGory, J. T. (2000). Inter-transcriber reliability of ToBI prosodic labeling. In *Proc. Interspeech* (pp. 235–238).

't Hart, J. & Cohen, A. (1973). Intonation by rule: A perceptual quest. *J. Phonet. 1*, 309–327.

't Hart, J. & Collier, R. (1975). Integrating different levels of intonation analysis. *J. Phonet. 3*(4), 235–255.

't Hart, J., Collier, R. & Cohen, A. (1990). *A perceptual study of intonation: an experimental-phonetic approach to speech melody*. Cambridge University Press.

Takamichi, S., Toda, T., Neubig, G., Sakti, S. & Nakamura, S. (2014). A postfilter to modify the modulation spectrum in HMM-based speech synthesis. In *Proc. ICASSP* (pp. 290–294).

Talkin, D. (1995). A robust algorithm for pitch tracking (RAPT). *Speech Coding and Synthesis*, 495–518.

Taylor, P. (1998). The Tilt intonation model. In *Proc. ICSLP* (pp. 1383–1386).

Taylor, P. (2000). Analysis and synthesis of intonation using the Tilt model. *J. Acoust. Soc. Am. 107*(3), 1697–1714.

Taylor, P. (2009). *Text-to-Speech Synthesis*. Cambridge University Press.

Taylor, P. A. (1992). *A phonetic model of English intonation* (Doctoral dissertation, The University of Edinburgh).

Taylor, P. A. & Black, A. W. (1994). Synthesizing conversational intonation from a linguistically rich input. *Proc. SSW-2*, 175–178.

Taylor, P., Black, A. W. & Caley, R. (1998). The architecture of the Festival speech synthesis system. *Proc. SSW-3*, 147–151.

Terband, H. & van Brenk, F. (2015). Compensatory and adaptive responses to real-time formant shifts in adults and children. In *Proc. ICPhS* (Paper no. 1017).

Terband, H., van Brenk, F. & van Doornik-van der Zee, A. (2014). Auditory feedback perturbation in children with developmental speech sound disorders. *J. Commun. Disord. 51*, 64–77.

Teutenberg, J., Watson, C. & Riddle, P. (2008). Modelling and synthesising F0 contours with the discrete cosine transform. In *Proc. ICASSP* (pp. 3973–3976).

Theano Development Team. (2016). Theano: A Python framework for fast computation of mathematical expressions. *CoRR*. Retrieved from http://arxiv.org/abs/1605.02688

Tieleman, T. & Hinton, G. (2012). *Lecture 6.5 - RMSProp*. COURSERA: Neural Networks for Machine Learning.

Titze, I. R. (1989). Physiologic and acoustic differences between male and female voices. *J. Acoust. Soc. Am. 85*(4), 1699–1707.

Toda, T., Black, A. W. & Tokuda, K. (2004). Mapping from articulatory movements to vocal tract spectrum with Gaussian mixture model for articulatory speech synthesis. In *Proc. SSW-5* (pp. 31–36).

Toda, T. & Tokuda, K. (2007). A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. *IEICE Trans. Inf. Syst. 90*(5), 816–824.

Tokuda, K., Masuko, T., Miyazaki, N. & Kobayashi, T. (1999). Hidden Markov models based on multi-space probability distribution for pitch pattern modeling. In *Proc. ICASSP* (Vol. 1, pp. 229–232).

Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J. & Oura, K. (2013). Speech synthesis based on hidden Markov models. *Proc. IEEE*, *101*(5), 1234–1252.

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T. & Kitamura, T. (2000). Speech parameter generation algorithms for HMM-based speech synthesis. In *Proc. ICASSP* (Vol. 3, pp. 1315–1318).

Tokuda, K. & Zen, H. (2015). Directly modeling speech waveforms by neural networks for statistical parametric speech synthesis. In *Proc. ICASSP* (pp. 4215–4219).

Tokuda, K. & Zen, H. (2016). Directly modeling voiced and unvoiced components in speech waveforms by neural networks. In *Proc. ICASSP* (pp. 5640–5644).

Torres, H. & Gurlekian, J. (2016). Novel estimation method for the superpositional intonation model. *IEEE/ACM Trans. Audio Speech Lang. Process. 24*(1), 151–160.

Tourville, J. A. & Guenther, F. H. (2011). The DIVA model: A neural theory of speech acquisition and production. *Lang. Cognitive Proc. 26*(7), 952–981.

Tourville, J. A., Reilly, K. J. & Guenther, F. H. (2008). Neural mechanisms underlying auditory feedback control of speech. *Neuroimage*, *39*, 1429–1443.

Tremblay, S., Shiller, D. M. & Ostry, D. J. (2003). Somatosensory basis of speech production. *Nature*, *423*(6942), 866–869.

Tuerk, C. & Robinson, T. (1993). Speech synthesis using artificial neural networks trained on cepstral coefficients. In *Proc. Eurospeech* (pp. 1713–1716).

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. W. & Kavukcuoglu, K. (2016). WaveNet: A Generative Model for Raw Audio. *CoRR*. Retrieved from http://arxiv.org/abs/1609.03499

van Brenk, F., Terband, H. & Cai, S. (2014). Auditory feedback perturbation in adults and children. In *Proc. Motor Speech*. Poster presentation.

van Santen, J., Mishra, T. & Klabbers, E. (2008). Prosodic processing. In J. Benesty, M. M. Sondhi & Y. Huang (Eds.), *Springer Handbook of Speech Processing* (pp. 471–488). Springer.

Verhelst, W. & Roelands, M. (1993). An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech. In *Proc. of ICASSP* (pp. 554–557).

Villacorta, V. M., Perkell, J. S. & Guenther, F. H. (2007). Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. *J. Acoust. Soc. Am. 122*, 2306–2319.

von Békésy, G. (1960). *Experiments in Hearing*. New York: McGraw-Hill.

Wang, C.-C., Ling, Z.-H., Zhang, B.-F. & Dai, L.-R. (2008). Multi-layer F0 modeling for HMM-based speech synthesis. In *Proc. ISCSLP* (pp. 1–4).

Watts, O., Henter, G. E., Merritt, T., Wu, Z. & King, S. (2016). From HMMs to DNNs: where do the improvements come from? In *Proc. ICASSP* (pp. 5505–5509).

Willems, N. J. (1983). STEP: A model of standard English intonation patterns. *IPO Annual Progress Report*, *18*, 37–42.

Williams, R. J. & Peng, J. (1990). An efficient gradient-based algorithm for on-line training of recurrent network trajectories. *Neural Comput. 2*(4), 490–501.

Wu, Y.-J. & Wang, R.-H. (2006). Minimum generation error training for HMM-based speech synthesis. In *Proc. ICASSP* (pp. 89–92).

Wu, Y. J. & Soong, F. (2012). Modeling pitch trajectory by hierarchical HMM with minimum generation error training. In *Proc. ICASSP* (pp. 4017–4020).

Wu, Z., Qian, Y., Soong, F. K. & Zhang, B. (2008). Modeling and generating tone contour with phrase intonation for Mandarin Chinese speech. In *Proc. ISCSLP* (pp. 1–4).

Wu, Z. & King, S. (2015). Minimum trajectory error training for deep neural networks, combined with stacked bottleneck features. In *Proc. Interspeech* (pp. 309–313).

Wu, Z. & King, S. (2016a). Improving Trajectory Modelling for DNN-Based Speech Synthesis by Using Stacked Bottleneck Features and Minimum Generation Error Training. *IEEE/ACM Trans. Audio Speech Lang. Process. 24*(7), 1255–1265.

Wu, Z. & King, S. (2016b). Investigating gated recurrent neural networks for speech synthesis. *Proc. ICASSP*, 5140–5144.

Xu, Y. & Liu, F. (2006). Tonal alignment, syllable structure, and coarticulation: Toward an integrated model. *Italian Journal of Linguistics*, *18*, 125–159.

Xu, Y. & Prom-on, S. (2010–2012). PENTAtrainer1. Retrieved from http://www.homepages.ucl.ac.uk/~uclyyix/PENTAtrainer1

Xu, Y. (1997). Contextual tonal variations in Mandarin. *J. Phonet. 25*(1), 61–83.

Xu, Y. (1999). Effects of tone and focus on the formation and alignment of F0 contours. *J. Phonet. 27*(1), 55–105.

Xu, Y. (2005). Speech melody as articulatorily implemented communicative functions. *Speech Commun. 46*(3–4), 220–251.

Xu, Y. (2007). Speech as articulatory encoding of communicative functions. In *Proc. ICPhS* (pp. 25–30).

Xu, Y. (2011). Post-focus compression: Cross-linguistic distribution and historical origin. In *Proc. ICPhS* (pp. 152–155).

Xu, Y. (2015). Speech prosody - Theories, models and analysis. In A. R. Meireles (Ed.), *Courses on Speech Prosody* (pp. 171–204). Cambridge Scholars Publishing.

Xu, Y., Larson, C. R., Bauer, J. J. & Hain, T. C. (2004). Compensation for pitch-shifted auditory feedback during the production of Mandarin tone sequences. *J. Acoust. Soc. Am. 116*, 1168–1178.

Xu, Y., Lee, A., Prom-on, S. & Liu, F. (2015). Explaining the PENTA model: A reply to Arvaniti and Ladd. *Phonology*, *32*, 505–535.

Xu, Y. & Prom-on, S. (2014). Toward invariant functional representations of variable surface fundamental frequency contours: Synthesizing speech melody via model-based stochastic learning. *Speech Commun. 57*, 181–208.

Xu, Y. & Prom-on, S. (2015). Degrees of freedom in prosody modeling. In K. Hirose & J. Tao (Eds.), *Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis* (pp. 19–34). Springer.

Xu, Y. & Sun, X. (2002). Maximum speed of pitch change and how it may relate to speech. *J. Acoust. Soc. Am. 111*(3), 1399–1413.

Xu, Y. & Wang, Q. E. (2001). Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Commun. 33*(4), 319–337.

Xu, Y. & Xu, C. X. (2005). Phonetic realization of focus in English declarative intonation. *J. Phonet. 33*(2), 159–197.

Yin, X., Lei, M., Qian, Y., Soong, F. K., He, L., Ling, Z.-H. & Dai, L.-R. (2014). Modeling DCT parameterized F0 trajectory at intonation phrase level with DNN or decision tree. In *Proc. Interspeech* (pp. 2273–2277).

Yin, X., Lei, M., Qian, Y., Soong, F. K., He, L., Ling, Z.-H. & Dai, L.-R. (2016). Modeling F0 trajectories in hierarchically structured deep neural networks. *Speech Commun. 76*, 82–92.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. & Kitamura, T. (1999). Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *Proc. Eurospeech* (pp. 2347–2350).

Yu, K. & Young, S. (2011). Continuous F0 modeling for HMM based statistical parametric speech synthesis. *IEEE Trans. Audio Speech Lang. Process. 19*(5), 1071–1079.

Yuan, J. (2006). Mechanisms of question intonation in Mandarin. In Q. Huo, B. Ma, E.-S. Chng & H. Li (Eds.), *Chinese Spoken Language Processing* (pp. 19–30). Springer.

Zen, H., Senior, A. & Schuster, M. (2013). Statistical parametric speech synthesis using deep neural networks. In *Proc. ICASSP* (pp. 7962–7966).

Zen, H. (2015). Acoustic modeling in statistical parametric speech synthesis - From HMM to LSTM-RNN. In *Proc. MLSLP*. Invited paper.

Zen, H. & Braunschweiler, N. (2009). Context-dependent additive log F0 model for HMM-based speech synthesis. In *Proc. Interspeech* (pp. 2091–2094).

Zen, H. & Sak, H. (2015). Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis. In *Proc. ICASSP* (pp. 4470–4474).

Zen, H. & Senior, A. (2014). Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis. In *Proc. ICASSP* (pp. 3844–3848).

Zen, H., Tokuda, K. & Black, A. W. (2009). Statistical parametric speech synthesis. *Speech Commun. 51*(11), 1039–1064.

Zen, H., Tokuda, K. & Kitamura, T. (2007). Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences. *Comput. Speech Lang. 21*(1), 153–173.

Zhang, Z., Wang, X., Yu, Y. & Wu, X. (2010). Hierarchical pitch target model for Mandarin speech. In *Proc. ISCSLP* (pp. 378–382).

Zimmerman, S., Kalinowski, J., Stuart, A. & Rastatter, M. (1997). Effect of altered auditory feedback on people who stutter during scripted telephone conversations. *J. Speech Lang. Hear. Res. 40*(5), 1130–1134.

Appendix A

Python Implementation of the Target Approximation Model

Requirements:

Python 2.7.x, 3.4.x or higher (http://www.python.org)
numpy 1.9.x or higher (http://www.numpy.org)
matplotlib 1.4.x or higher (http://matplotlib.org)

```
#==================================
# qta.py
# by Hao Liu <h.liu.12@ucl.ac.uk>
# University College London
#==================================


import math
import numpy as np


class QTA(object):
    '''
    The quantitative implementation of
    the Target Approximation (qTA) model
    for dynamic F0 generation
    '''
    def __init__(self, target, onset):
        ''' Initialize a qTA instance '''
        if not isinstance(target, tuple) or len(target) != 3:
            raise TypeError('"target" should be a 3-item tuple')
        if not isinstance(onset, tuple) or len(onset) != 3:
            raise TypeError('"onset" should be a 3-item tuple')

        self.m = target[0]
        self.b = target[1]
        self.r = target[2]

        self.onset_f0 = onset[0]
        self.onset_vel = onset[1]
```

```
        self.onset_acc = onset[2]

    def generate(self, t):
        '''
        Generate f0 at time t or along the given time series
        (t should be a relative time or time series
        to the syllable onset)
        '''
        c1 = self.onset_f0 - self.b
        c2 = self.onset_vel + c1*self.r - self.m
        c3 = 0.5*(self.onset_acc + 2*c2*self.r - c1*self.r*self.r)

        if isinstance(t, float):
            f0 = self.m*t + self.b +
                    (c1 + c2*t + c3*t*t) * math.exp(-self.r*t)
        elif isinstance(t, np.ndarray):
            f0 = self.m*t + self.b +
                    (c1 + c2*t + c3*t*t) * np.exp(-self.r*t)
        else:
            raise TypeError('The argument t should be' \
                            'either a float or a numpy.ndarray.')
        return f0
```

```
#==================================
# ta_process.py
# by Hao Liu <h.liu.12@ucl.ac.uk>
# University College London
#==================================


from qta import QTA
import numpy as np
from matplotlib.mlab import frange


class TAProcess(object):
    '''
    The process of target approximation, which may form a full
    syllable of part of a syllable

    a. produce an F0 trajectory with a properly initialized QTA model
    b. handle other important attributes for the TA process
    '''

    def __init__(self, t_series, target, onset=(0., 0., 0.)):
        ''' Initialize a qTA instance '''
        if not isinstance(t_series, np.ndarray):
            raise TypeError('"t_series" should be a 1D numpy array')
        if not isinstance(target, tuple) or len(target) != 3:
            raise TypeError('"target" should be a 3-item tuple')
        if not isinstance(onset, tuple) or len(onset) != 3:
            raise TypeError('"onset" should be a 3-item tuple')

```

```
31          self.t_series = t_series
32          self._duration = t_series[-1] - t_series[0]
33
34          # QTA actually uses onset b value,
35          # but the target defined for a TA process is the offset.
36          # So convert offset b to onset b!
37          self._model = QTA((target[0],
38                             target[1]-self._duration*target[0],
39                             target[2]), onset)
40
41          self.onset = onset
42          self.f0 = None
43          self.offset = None
44
45      def produce(self):
46          ''' Produce a F0 series of the defined TA process '''
47          self.f0 = self._model.generate(self.t_series)
48          return self.f0
49
50      def _get_offset_f0(self):
51          ''' Get offset f0 (semitone) '''
52          return self.f0[-1]
53
54      def _get_offset_vel(self):
55          ''' Get offset f0 velocity (semitone/sec) '''
56          return (self.f0[-1] - self.f0[-2]) /
57                     (self.t_series[-1] - self.t_series[-2])
58
59      def _get_offset_acc(self):
60          ''' Get offset f0 acceleration (semitone/sec^2) '''
61          dur1 = self.t_series[-1] - self.t_series[-2]
62          dur2 = self.t_series[-2] - self.t_series[-3]
63          vel1 = (self.f0[-1] - self.f0[-2]) / dur1
64          vel2 = (self.f0[-2] - self.f0[-3]) / dur2
65          return (vel1 - vel2) / dur1
66
67      def get_offset(self):
68          ''' Get the offset state '''
69          if self.f0 is None:
70              raise ValueError('Empty F0.'\
71                               'Run TAProcess.produce() first.')
72          elif len(self.f0) >= 3:
73              self.offset = (self._get_offset_f0(),
74                             self._get_offset_vel(),
75                             self._get_offset_acc())
76          else:
77              # Too short production! Use offset F0 level,
78              # onset vel & acc as an estimate of
79              # the offset state
80              self.offset = (self.f0[-1],
81                             self.onset[1], self.onset[2])
82          return self.offset
83
84
```

```
85 def multi_TAProcess(targets, durs, t_step,
86                                 initial_onset=(0., 0., 0.)):
87     ''' Dynamically stringing multiple TA processes '''
88     if len(targets) != len(durs):
89         raise ValueError('Dimension mismatch:'\
90                          'len(targets) != len(durs)')
91 
92     f0 = None
93     onset = None
94     for i in range(len(targets)):
95         if i == 0:
96             tap = TAProcess(frange(0., durs[i], t_step, closed=1),
97                                           targets[i], initial_onset)
98             f0 = tap.produce()
99             onset = tap.get_offset()
100        else:
101            tap = TAProcess(frange(t_step, durs[i], t_step,
102                                       closed=1), targets[i], onset)
103            f0 = np.concatenate((f0, tap.produce()))
104            onset = tap.get_offset()
105    return f0
```