Doctoral Thesis reviewed
by Ritsumeikan University

Studies on Dynamic Feature Modeling and Fundamental Frequency Extraction in HMM-based Speech Synthesis

(HMMに基づいた音声合成における動的特徴のモデル化および基本周波数の抽出に関する研究)

March 2016

2016年 3月

Doctoral Program in Integrated Science and Engineering

Graduate School of Science and Engineering

Ritsumeikan University

立命館大学大学院理工学研究科

総合理工学専攻博士課程後期課程

NINH Khanh Duy

ニン　カン　ズイ

Supervisor : Professor YAMASHITA Yoichi

研究指導教員：山下 洋一教授

Abstract

This thesis aims to improve the speech quality of HMM-based speech synthesis systems by considering two issues: the modeling of the dynamic features of speech parameters, and the extraction of the fundamental frequency (or F0) parameter in glottalized regions of speech signals. The dynamic features capture dynamic properties of speech parameter trajectories, thus containing important information about speech dynamics such as spectral transition. Meanwhile, the F0 parameter conveys the intonation of speech, however, difficult to accurately estimate in speech affected by glottalization. Therefore, accurate modeling of the dynamic features and accurate extraction of the F0 in glottalized speech can help enhance the naturalness and expressiveness of speech synthesized from HMMs.

First, the author improves the modeling accuracy for the dynamic features by incorporating the generation error of dynamic features into the generation error function of the Minimum Generation Error (MGE) criterion, a state-of-the-art HMM training framework for speech synthesis. The author also proposes a method for adaptively changing the weight associated with the newly added error component based on the dynamicity degree of portions of the speech signal. As a result, the proposed technique improves the capability of HMMs in capturing dynamic properties of speech while maintaining a computational complexity similar to that of the conventional MGE criterion.

Second, the author tackles the problem of F0 extraction in glottalized speech signals by examining a language possessing a heavy glottalization feature, (Hanoi) Vietnamese. As a tonal language with several glottalized tones in its tone set, the inaccurate F0 estimation has severe effects on the F0 modeling, thus degrading the tone naturalness and causing the hoarseness in synthesized speech. The author proposes an F0 parameterization scheme for the Vietnamese glottalized tones by using a pitch mark propagation algorithm in combination with a conventional F0 extractor. The proposed scheme is capable of deriving more complete and accurate F0 contours representing the tones compared to the simple use of the F0 extractor, thereby significantly alleviating the hoarseness and slightly improving the tone naturalness of synthetic speech.

Abstract in Japanese

本論文は，HMM に基づいた音声合成システムにおける品質向上を目的としている。音声パラメータの動的特徴のモデル化，および音声信号において声帯振動が不規則となっている区間での基本周波数（F0）抽出の 2 つの問題を対象としている。動的特徴は音声パラメータ変化の動的な性質を表現しており，スペクトル遷移などの音声の変化特性に関する重要な情報を含んでいる。一方で，F0 パラメータは，音声のイントネーションを表現しており，声帯振動が不規則となっている音声においては正確に推定することが難しい。動的特徴の正確なモデル化および声帯振動が不規則となっている音声における正確な F0 推定は，HMM から合成された音声における自然性や感情表現を高めることになる。

第 1 に，音声合成において広く用いられている HMM 学習の枠組みにおいて，生成誤差最小（MGE）基準の評価関数に動的特徴の生成誤差を導入することによって動的特徴のモデル化精度を向上させた。また，新たに導入した誤差項の重みを，音声の動的変化の強度に応じて適応的に変化させる手法を提案している。結果として，提案手法は，計算量を従来の MGE 基準の手法と同等に保ちながら，音声の動的特徴において HMM の表現能力を向上させている。

第 2 に，声帯の不規則振動が頻繁に出現する言語であるベトナム語（のハノイ方言）を対象として，声帯振動が不規則となっている音声の F0 抽出の問題に取り組んでいる。声帯の不規則振動が頻繁に発生する声調を持っている声調言語においては，不正確な F0 推定が F0 のモデル化に悪影響をもたらし，合成音声において声調の不自然さやしわがれ声をもたらす。従来の F0 抽出にピッチマークの伝搬アルゴリズムを組み込むことによって，ベトナム語の声帯振規則となる声調に対する F0 分析の枠組みを提案している。提案手法は，従来の F0 抽出法に比べて，抽出誤差なく正確に声調を表現する F0 系列を抽出できており，合成音声のしわがれ声らしさを軽減し，声調の自然さを向上している。

Acknowledgments

The present thesis was made possible by a PhD scholarship from the 322 Project, the Ministry of Education and Training of Vietnam.

I would like to express my deep gratitude to my supervisor, Prof. Yoichi Yamashita for his kindness, his insightful guidance, his availability, and his support throughout my doctoral study.

I am also thankful to Assistant Professors Masanori Morise and Masahiro Niitsuma, and Doctor Kook Cho for the fruitful collaborations during my research work.

I would also like to thank Professor Takanobu Nishiura for the permission to use his laboratory's recording room, and Assistant Professor Takahiro Fukumori for his help during speech recording sessions.

I would like to thank all Japanese and Vietnamese students studying at Ritsumeikan University who have participated in the listening experiments, an indispensable part of my research.

For their availability and their judicious comments on my thesis, I am thankful to all my thesis's screening committee members: Prof. Akira Hirabayashi, Prof. Takanobu Nishiura, and Prof. Yoichi Yamashita.

Finally, I would like to express my deepest gratitude to my family for their kind support and encouragement along this journey.

Table of contents

List of Figures

List of Tables

Chapter 1 Introduction

Speech is one of the most important means of communication for human beings. Consequently, there have been enormous efforts to facilitate human-machine interaction using speech. Key technologies in speech processing such as speech recognition and understanding, and speech synthesis have recently been made viable thanks to the availability of powerful microprocessors, large datasets, and advanced machine learning algorithms. The integration of these technologies also supports human-to-human communication. Typical applications of speech processing include speech-to-speech translation, hands-free and/or eyes-free communication and control for people with disabilities, virtual assistant on hand-held devices, etc.

Text-to-speech (TTS) conversion is a technique for making a computer to produce human-like speech from any input text. To fully transmit the information conveyed in synthesized speech signals, it is often desired that speech synthesis systems are able to generate natural sounding and easily understandable speech with arbitrary speaker's voice characteristics and various speaking styles and/or emotional expressions. In other words, the quality of synthetic speech is characterized by its naturalness, its intelligibility and its expressivity. Meanwhile, a practical concern is the amount of resources required for building a speech synthesizer. It is always expected that the amount of recorded speech data, and the amount of time and labor needed for processing them afterwards (i.e., segmentation, annotation, etc.), are as small as possible to attain a given level of synthetic speech quality.

Modern speech synthesis is being dominated by two techniques: unit selection [1] and statistical parametric speech synthesis [2]. In unit selection approach, appropriate sub-word units are automatically chosen from a speech database based on two cost functions: the target cost, which represents how well the selected unit matches the target, and the concatenation cost, which represents how well two selected units combine. At synthesis time, the goal is to minimize the overall cost of a label sequence (derived from the input text) to be synthesized, which is equal to the sum of the target and concatenation costs.

This method allows the generation of high-quality speech since speech units are directly selected from a database of actual human speech, then concatenated together without any modification. However, it requires a huge amount of resources since a larger database implies a better unit coverage in terms of phonetic and prosodic contexts. As speech units cannot be easily and systematically modified (i.e., simultaneous changes in spectrum, fundamental frequency, and phone duration), a database much be separately recorded for each voice or speaking style/emotion expression. Unfortunately, recording large databases with variations is very difficult and costly [3], thus limiting the flexibility of the unit selection approach. Most of current commercial systems use this synthesis technique, and an extensive literature review can be found in [4]

In direct contrast with the selection of original instances of speech units from a database, statistical parametric speech synthesis can be described as generating the average of some sets of similarly sounding speech segments [2]. In this approach, sequences of speech parameters are firstly extracted from a collection of speech signals. The parameters of statistical models are then learned from this speech corpus in a process referred to as model parameter estimation (or model training). After that, the trained statistical models are used to synthesize speech parameter sequences for any input text in a process referred to as speech parameter generation. The speech signal corresponding to this text is finally reconstructed from some parametric representation of speech. This method produces reasonably high-quality speech, but slightly degraded by its buzziness characteristic compared to what is generated by the unit selection approach. However, it has the significant advantage of high flexibility thanks to the possibility of using voice conversion or speaker adaptation techniques for controlling speaker's voice characteristics, speaking styles or emotions. In addition, it results in systems with much smaller memory footprint (typically several hundreds kilobytes) than ones produced by the unit selection approach (typically several hundreds megabytes) because only the parameters of the statistical models have to be stored rather than a large speech database. This makes the statistical parametric approach highly suitable for embedded devices, which often have limited memory resources. In recent years, statistical parametric speech synthesis has been widely adopted for building TTS systems and gradually become a mainstream in research and development in speech synthesis. The field has largely converged on a standard approach to statistical parametric speech synthesis based on Hidden Markov Models (HMMs) [5], and the author refers to this as HMM-based speech synthesis framework in the present thesis.

Despite its success, there are many potential areas for improving the quality of speech synthesized from HMMs, and the author considers two of these in detail in the thesis: the

modeling of the dynamic features of speech parameters, and the extraction of the fundamental frequency parameter from speech signals. The first issue the thesis considers is the drawbacks of conventional training criteria, the maximum likelihood (ML) and the minimum generation error (MGE), in modeling the dynamic features. In the ML training framework [6], the constraints between static and dynamic features are ignored in the training stage, yet used in the synthesis stage. In order to resolve this inconsistency of the ML-based approach, the MGE criterion has been proposed for training the HMMs [7]. By incorporating the speech parameter generation algorithm into the training stage, the HMMs are estimated to minimize the deviation between the original and generated speech data, referred to as generation error. Although the conventional MGE criterion effectively addresses the inconsistency of the ML criterion and improves synthetic speech quality, it stills has a weakness, which is the ignorance of the dynamic features in the generation error definition. The dynamic features capture dynamic properties of speech parameter trajectories, thus containing important information about speech dynamics such as spectral transition, an important acoustic cue in speech perception. Therefore, the absence of the dynamic features in the generation error definition may have some negative effect on the quality of an HMM-based synthesis system. To address this issue, the thesis incorporates dynamic properties of speech parameters into MGE training by introducing the error component of dynamic features into the generation error definition. The author proposes two methods for setting the weight associated with the additional error component. In the fixed weighting approach, this weight is kept constant over the course of speech. In the adaptive weighting approach, it is adjusted according to the degree of dynamicity of speech segments. The newly derived MGE criterion improves the capability of HMMs in capturing dynamic properties of speech without increasing the computational complexity of the training process compared with the conventional MGE criterion.

The second issue the thesis considers is the extraction of the fundamental frequency (or F0) in glottalized regions of a speech signal within the framework of an HMM-based TTS system. Specifically, the author investigates the problem of F0 parameterization of glottalized tones in HMM-based TTS for Vietnamese (Hanoi dialect in particular), a tonal language possessing a heavy glottalization feature. The term “glottalization” is used to cover any non-modal phonation that occurs during tone production and leads to speech waveform with irregular pitch periods, often accompanied by very low F0 values or voicelessness. The F0 contour extracted from the speech signal of an utterance represents its intonation (or melody). For a tonal language like Vietnamese, each syllabic tone can be represented by an F0 contour covering a whole syllable. Since each tone has an essential role in the meaning of the associated syllable, accurate analysis, modeling and synthesis of

the tones, thus the F0 contour of a speech signal, is crucial for an HMM-based Vietnamese TTS system. Modern F0 extractors often assign erroneous F0 values or fail to detect any F0 in glottalized waveforms, which result from the production of the Vietnamese glottalized tones, since these waveforms often exhibit rather weak periodicity. As a result, subsequent F0 modeling and generation suffer. Standard HMM-based TTS uses multi-space distribution (MSD) to model and generate discontinuous F0 trajectories [8]. Faulty voicing decisions resulting from the F0 extraction phase will cause the deteriorately trained MSD-HMMs to synthesize voiced frames as unvoiced, resulting in hoarse speech, or to synthesize unvoiced frames as voiced, resulting in buzzy speech. When listening to the output of the baseline HMM-based Vietnamese TTS system, the author perceived that although the synthetic speech is highly intelligible, its overall quality is greatly degraded by the hoarseness frequently occurring at syllables bearing a glottalized tone. This is due to voiced (or F0) missing errors (i.e., voiced frames classified as unvoiced) caused by F0 trackers when dealing with glottalized waveforms. To alleviate the perceived hoarseness, popular solutions include the use of continuous F0 contours either for modeling [9-11] or for synthesis [12]. In this thesis, the author directly tackles the problem of F0 estimation in glottalized regions of a speech signal. The research of Ishi et al. [13] suggests that a cross-correlation-based measure could help for the detection of similar successive glottal pulses in glottalized speech segments. Based upon this work, the author proposes a pitch marking algorithm, where the pitch marks are propagated from regularly spaced pitch periods to irregularly spaced ones, from which the refined F0 contour of a glottalized tone is derived. This F0 parameterization method reduces the hoarseness whilst improving the tone naturalness of synthetic speech. The pitch marking procedure works as a refinement step based on the results of a conventional F0 extractor. Thus it can be combined with any F0 extractor.

The remainder of this thesis is organized as follows. Chapter 2 reviews the fundamentals of statistical parametric speech synthesis based on HMMs. Chapter 3 presents the MGE training criterion incorporating the dynamic features of speech parameters into the generation error definition, and compares it to the standard MGE criterion in subjective and objective evaluations. Chapter 4 describes an F0 parameterization method for the glottalized tones in HMM-based Vietnamese TTS, and compares it to the standard use of a conventional F0 extractor in subjective and objective evaluations. Chapter 5 gives some perspectives and concludes the thesis.

Chapter 2 HMM-based Text-to-Speech Synthesis

2.1. Introduction

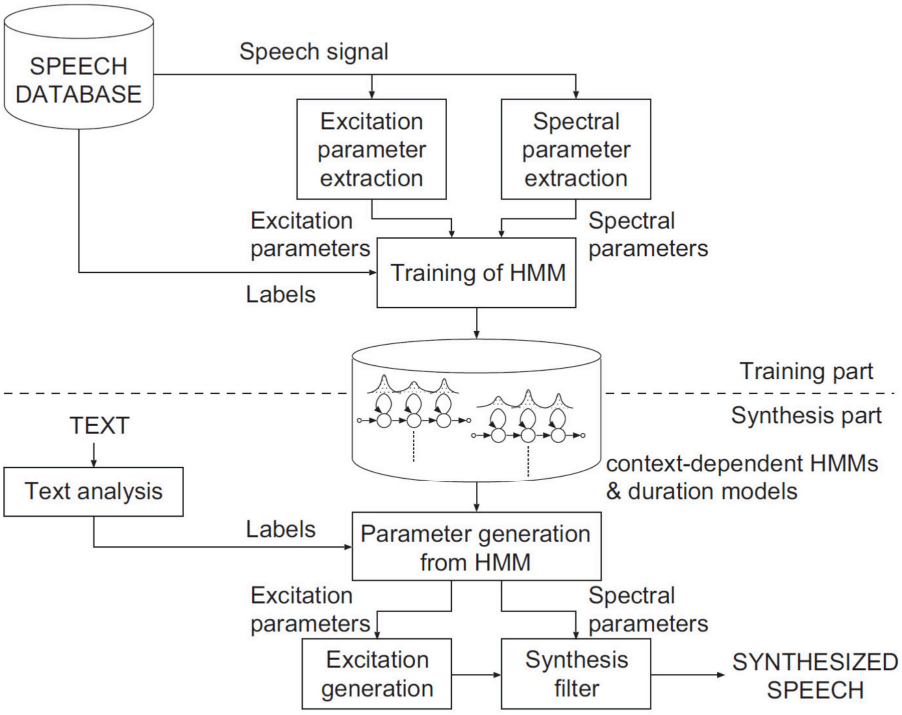


Figure 2.1: The scheme of a typical HMM-based TTS system [2].

This chapter gives an overview of HMM-based TTS. A typical HMM-based TTS system consists of the training and synthesis parts, as shown in Figure 2.1. In the training part, spectral parameters (e.g., mel-cepstral coefficients) and excitation parameters (e.g., fundamental frequency) are firstly extracted from a speech database. The extracted parameters are then modeled by context-dependent HMMs. Context-dependent duration models are also estimated during this phase. In the synthesis part, a context-dependent

label sequence is firstly derived from the input text by a text analyzer. Then, a sentence HMM is composed by concatenating context-dependent HMMs according to the label sequence. After that, spectral and excitation parameters are generated from the sentence HMM by using the parameter generation algorithm. Finally, speech is synthesized from the generated spectral and excitation parameters by using a synthesis filter.

This chapter is organized as follows. Section 2.2 introduces a source-filter model of speech production, which lays the foundation for speech analysis (or feature extraction) and synthesis steps in HMM-based TTS. The training and synthesis parts are then covered in Sections 2.3 and 2.4, respectively.

2.2. Speech analysis/synthesis framework

For all the work in this thesis, the speech analysis/synthesis framework is based on a traditional source-filter model of speech production [14] (Figure 2.2). In this model, the speech signal is assumed as the output of a linear time-invariant system (i.e., the filter) excited by an excitation signal (i.e., the source) that switches between periodic pulse train for voiced speech and white noise for unvoiced speech. The excitation signal $e(n)$ reflects the pulse/noise characteristic of the air flow at the vocal folds, while the filter $h(n)$ simulates the resonance effect of the vocal tract during the speech production of a human. To produce a speech-like signal, the excitation mode and the properties of the vocal tract filter must change with time.

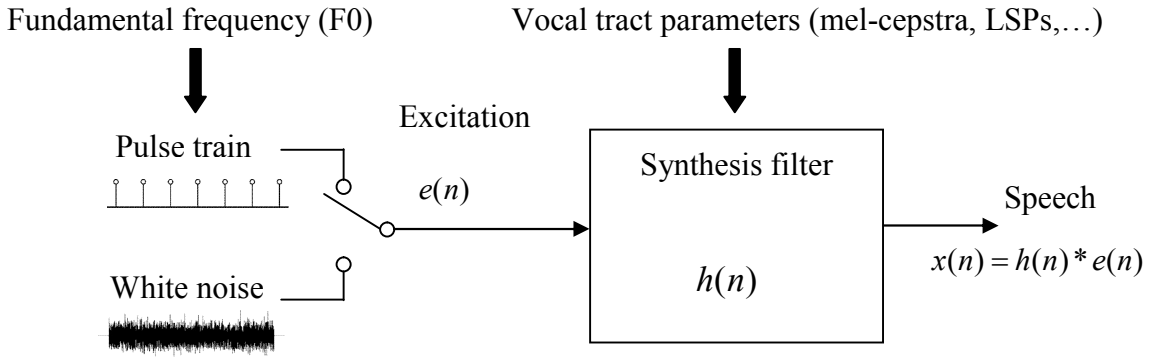


Figure 2.2: The source-filter model of speech production.

Speech analysis must consider the fact that the properties of speech signals change relatively slowly with time. Therefore, it is often assumed that the general characteristics of the excitation and the vocal tract remain unchanged for periods of 10–40 ms. In practice,

analysis frames of 25 ms with 5-ms shifts are used to extract smooth representations of the speech signal, including excitation and spectral parameters. Excitation parameters are the voiced/unvoiced classification and the fundamental frequency (or F0) for voiced speech. The problem of F0 extraction will be examined in details in Chapter 4 when speech signals having some abnormal periodicity are dealt with. Spectral parameters describe the vocal tract filter's frequency response, which are usually either mel-cepstral or line spectral pair (LSP) coefficients. In this thesis, mel-cepstral coefficients obtained by a mel-cepstral analysis technique [15] are used as the spectral parameters.

To synthesize speech from given excitation and spectral parameters, a synthesis filter must be realized based on the mel-cepstral coefficients. Here, the Mel Log Spectrum Approximation (MLSA) filter [16] is used to synthesize a speech waveform from the obtained mel-cepstra.

Although more complex excitation models (e.g., mixed excitation [17]) and analysis/synthesis framework (e.g., STRAIGHT [18]) have shown effectiveness in HMM-based TTS (e.g., [19], [20]), this thesis uses the pulse/noise excitation and analysis/synthesis framework described above for simplicity.

2.3. Training stage

This section describes how the spectrum, F0 and duration are simultaneously modeled in a unified framework of HMMs under the maximum likelihood (ML) criterion.

2.3.1. Spectral modeling with continuous distribution HMMs

2.3.1(1) Continuous distribution HMM

In HMM-based speech synthesis, spectral parameters are modeled using HMMs in the same way as speech recognition systems [21]. An HMM is a finite state machine that generates a sequence of observations, however, its states are hidden (i.e., unobservable). Mathematically speaking, an HMM is a doubly embedded stochastic process in which the state changes at each time unit according to the state transition probabilities, then the observations are generated through the output probability distribution associated with each state.

An N-state HMM λ is defined by a set of model parameters including:

- the initial state probabilities $\boldsymbol{\pi} = \{\pi_i\}_{i=1}^{N}$, where

$$\pi_i = P(q_1 = i) \tag{2.1}$$

is the probability of being state i at the first time unit. These probabilities are subject to the following constraint

$$\sum_{i=1}^{N} \pi_i = 1. \tag{2.2}$$

- the state transition probabilities $\boldsymbol{A} = \{a_{ij}\}_{i,j=1}^{N}$, where

$$a_{ij} = P\left(q_{t+1} = j \mid q_t = i\right) \tag{2.3}$$

is the probability of changing from state i to state j, under the hypotheses that the transition probabilities obey a first-order Markov process (i.e., the probability of being a state at next time unit depends only on the current state, not on the previous ones) and are time-independent. These probabilities are subject to the following constraint

$$\sum_{j=1}^{N} a_{ij} = 1, \quad 1 \le i \le N. \tag{2.4}$$

- the state output probabilities $\boldsymbol{B} = \{b_j(\boldsymbol{o}_t)\}_{j=1}^{N}$, where

$$b_j(\boldsymbol{o}_t) = P(\boldsymbol{o}_t \mid q_t = j) \tag{2.5}$$

is the probability of generating the observation $\boldsymbol{o}_t$ when being in state j at time t. The output probability distribution $b_j(\boldsymbol{o}_t)$ can be discrete or continuous depending on whether the observations are discrete or continuous-valued, respectively.

For notational simplicity, the model parameters of the HMM λ are denoted as

$$\lambda = \left(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}\right). \tag{2.6}$$

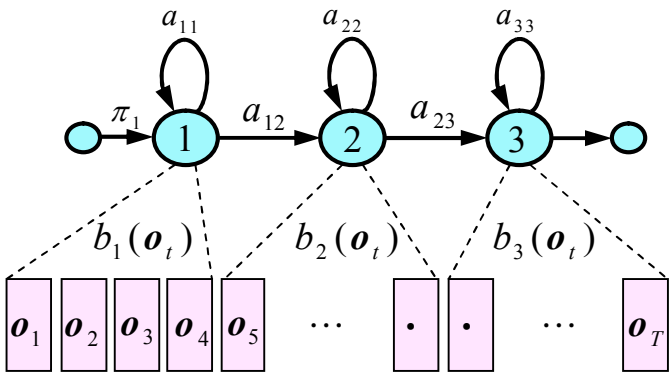


Figure 2.3: A 3-state no-skip left-to-right HMM generates an observation sequence (adapted from [22]).

Figure 2.3 shows an N-state left-to-right HMM with no skip (N = 3), an HMM topology widely used as a speech unit (e.g., phoneme) for modeling speech parameter sequences due to the fact that the speech signal has properties varying successively with

time. In this HMM structure, the state index either increases by one or stays the same when the time index increases. In the figure, the HMM is assumed to generate the observation sequence $\boldsymbol{O} = (\boldsymbol{o}_1, \boldsymbol{o}_2, ..., \boldsymbol{o}_T)$ in which the first four observations are generated from the first state. An observation $\boldsymbol{o}_t$ is typically a *D*-dimensional speech parameter vector obtained after the parameterization of the analysis frame at time index *t*. For modeling such multi-dimensional continuous observational data, we use *continuous distribution HMM* (CD-HMM) in which the output probability of a state is usually modeled by a mixture of multivariate Gaussian distributions as follows

$$b_j(\boldsymbol{o}_t) = \sum_{m=1}^{M} w_{jm} \mathcal{N}\left(\boldsymbol{o}_t; \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm}\right), \quad 1 \le j \le N, \tag{2.7}$$

where *M* is the number of Gaussian components in the mixture; w_{jm}, $\boldsymbol{\mu}_{jm}$, and $\boldsymbol{\Sigma}_{jm}$ are the weight, the *D*-dimensional mean vector, and the $D \times D$ covariance matrix of Gaussian component *m* of state *j*, respectively. The Gaussian probability distribution function (PDF) $\mathcal{N}\left(\boldsymbol{o}_t; \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm}\right)$ is defined as

$$\mathcal{N}\left(\boldsymbol{o}_t; \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm}\right) = \frac{1}{(2\pi)^{D/2} \left|\boldsymbol{\Sigma}_{jm}\right|^{1/2}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{o}_t - \boldsymbol{\mu}_{jm}\right)^T \boldsymbol{\Sigma}_{jm}^{-1} \left(\boldsymbol{o}_t - \boldsymbol{\mu}_{jm}\right)\right\}. \tag{2.8}$$

When the components of the *D*-dimensional feature vector are assumed to be not correlated with each other (an assumption is of popular use in HMM-based TTS), $\boldsymbol{\Sigma}_{jm}$ becomes a diagonal covariance matrix and the above equation is reduced to

$$\mathcal{N}\left(\boldsymbol{o}_t; \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm}\right) = \prod_{i=1}^{D} \frac{1}{\sqrt{2\pi\Sigma_{jmi}^2}} \exp\left(-\frac{1}{2}\frac{\left(o_{ti} - \mu_{jmi}\right)^2}{\Sigma_{jmi}^2}\right), \tag{2.9}$$

where o_{ti} is the i^{th} component of $\boldsymbol{o}_t$, μ_{jmi} is the i^{th} component of $\boldsymbol{\mu}_{jm}$, and Σ_{jmi}^2 is the i^{th} diagonal element of $\boldsymbol{\Sigma}_{jm}$.

For the use of HMM in modeling real-world phenomena, it is essential to solve efficiently three following problems, whose mathematical solutions are detailed in [21]:

- problem #1 (*probability evaluation*): given an HMM model λ, how to compute the probability $P(\boldsymbol{O} \mid \lambda)$ of the observation sequence $\boldsymbol{O} = (\boldsymbol{o}_1, \boldsymbol{o}_2, ..., \boldsymbol{o}_T)$?
- problem #2 (*optimal state sequence determination*): given an HMM model λ, how to determine the state sequence $\boldsymbol{q} = (q_1, q_2, ..., q_T)$ that best explains the observation sequence $\boldsymbol{O} = (\boldsymbol{o}_1, \boldsymbol{o}_2, ..., \boldsymbol{o}_T)$?
- problem #3 (*model parameter estimation*): given the observation sequence $\boldsymbol{O} = (\boldsymbol{o}_1, \boldsymbol{o}_2, ..., \boldsymbol{o}_T)$, how to adjust the model parameters $\lambda = \left(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}\right)$ to maximize $P(\boldsymbol{O} \mid \lambda)$?

For the sake of compactness, this thesis only describes the solution of problem #3, which is also referred to as the *model training* problem, in the next sub-section.

2.3.1(2) HMM training under maximum likelihood criterion

There is no known analytical solution for finding the model parameter set $\lambda = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$ that globally maximize the likelihood of a given observation sequence $\boldsymbol{O} = (\boldsymbol{o}_1, \boldsymbol{o}_2, ..., \boldsymbol{o}_T)$:

$$\hat{\lambda} = \arg\max_{\lambda} P(\boldsymbol{O} \mid \lambda) = \arg\max_{\lambda} \sum_{all\ \boldsymbol{q}} P(\boldsymbol{O}, \boldsymbol{q} \mid \lambda), \tag{2.10}$$

where $\boldsymbol{q}$ denotes a possible state sequence, which is a hidden or latent variable. However, a model parameter set λ which locally maximizes the likelihood $P(\boldsymbol{O} \mid \lambda)$ can be obtained using an iterative procedure such as the Expectation-Maximization (EM) algorithm [23], a general technique for finding the maximum likelihood estimators of models including hidden variables such as HMM states.

In the following, the EM algorithm for CD-HMM with single Gaussian distribution, which is widely used in HMM-based TTS, is briefly described. Corresponding formulae for mixture Gaussian or discrete output distributions can be derived similarly.

This algorithm is based on the forward and backward probabilities defined as:

- $\alpha_t(i) = P(\boldsymbol{o}_1, \boldsymbol{o}_2, ..., \boldsymbol{o}_t, q_t = i \mid \lambda')$, the probability of having the partial observation sequence $(\boldsymbol{o}_1, \boldsymbol{o}_2, ..., \boldsymbol{o}_t)$ and being in state i at time t, given the model λ'.
- $\beta_t(i) = P(\boldsymbol{o}_{t+1}, \boldsymbol{o}_{t+2}, ..., \boldsymbol{o}_T \mid q_t = i, \lambda')$, the probability of having the partial observation sequence $(\boldsymbol{o}_{t+1}, \boldsymbol{o}_{t+2}, ..., \boldsymbol{o}_T)$, given state i at time t and the model λ'.

These probabilities are efficiently calculated by the Forward-Backward algorithm [21].

In the *Expectation step*, the current model parameter set λ' is used to compute the following posterior probabilities:

- the state occupancy probability $\gamma_t(i)$ is the probability of being in state i at time t given the model λ' and the observation sequence $\boldsymbol{O}$, i.e.,

  $$\gamma_t(i) = P(q_t = i \mid \boldsymbol{O}, \lambda'). \tag{2.11}$$

  It follows that

  $$\gamma_t(i) = \frac{P(q_t = i, \boldsymbol{O} \mid \lambda')}{P(\boldsymbol{O} \mid \lambda')} = \frac{\alpha_t(i)\beta_t(j)}{\sum_{k=1}^{N} \alpha_t(k)\beta_t(k)}. \tag{2.12}$$

- the state transition probability $\xi_t(i, j)$ is the probability of being in state i at time t and in state j at time $t+1$ given the model λ' and the observation sequence $\boldsymbol{O}$, i.e.,

  $$\xi_t(i, j) = P(q_t = i, q_{t+1} = j \mid \boldsymbol{O}, \lambda'). \tag{2.13}$$

  It follows that

$$\xi_t(i,j) = \frac{P(q_t = i, q_{t+1} = j, \boldsymbol{O} \mid \lambda')}{P(\boldsymbol{O} \mid \lambda')} = \frac{\alpha_t(i) a_{ij} b_j\left(o_{t+1}\right) \beta_{t+1}(j)}{\sum_{k=1}^{N} \sum_{l=1}^{N} \alpha_t(k) a_{kl} b_l\left(o_{t+1}\right) \beta_{t+1}(l)}. \tag{2.14}$$

In the *Maximization step*, the current model parameter set λ' is replaced by the new parameter set λ that maximizes the auxiliary function

$$Q(\lambda, \lambda') = \sum_{all\ \boldsymbol{q}} P(\boldsymbol{q} \mid \boldsymbol{O}, \lambda') \ln P(\boldsymbol{q} \mid \boldsymbol{O}, \lambda'). \tag{2.15}$$

When applied iteratively, this procedure can be proved to increase the likelihood $P(\boldsymbol{O} \mid \lambda)$ monotonically and converge to a certain critical point. The maximization of the auxiliary function $Q(\lambda, \lambda')$ over λ, subject to the constraints of Eqs. (2.4) and (2.5) leads to the following re-estimation formulae for state i, $1 \le i \le N$:

$$\pi_i = \gamma_1(i), \tag{2.16}$$

$$a_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}, \tag{2.17}$$

$$\boldsymbol{\mu}_i = \frac{\sum_{t=1}^{T} \gamma_t(i) \cdot \boldsymbol{o}_t}{\sum_{t=1}^{T} \gamma_t(i)}, \tag{2.18}$$

$$\boldsymbol{\Sigma}_i = \frac{\sum_{t=1}^{T} \gamma_t(i) \cdot \left(\boldsymbol{o}_t - \boldsymbol{\mu}_i\right)\left(\boldsymbol{o}_t - \boldsymbol{\mu}_i\right)^{\mathrm{T}}}{\sum_{t=1}^{T} \gamma_t(i)}. \tag{2.19}$$

2.3.2. F0 modeling with multi-space distribution HMMs

In the previous section, we have seen how a fixed-dimensional speech parameter such as the spectrum can be modeled by continuous Gaussian distribution. However, it is difficult to use either discrete or continuous distribution to model a variable-dimensional parameter such as the fundamental frequency (F0). The F0 contour extracted from a speech signal is composed of observations that are either real-valued for voiced regions or undefined for unvoiced regions. In other words, an F0 observation sequence includes both of one-dimensional continuous values, which represent voiced speech, and a discrete (zero-dimensional) symbol, which represent unvoiced speech. For modeling such variable-

dimensional observation sequence, multi-space probability distribution (MSD) based HMM has been proposed and applied to F0 pattern modeling in HMM-based TTS [8, 24].

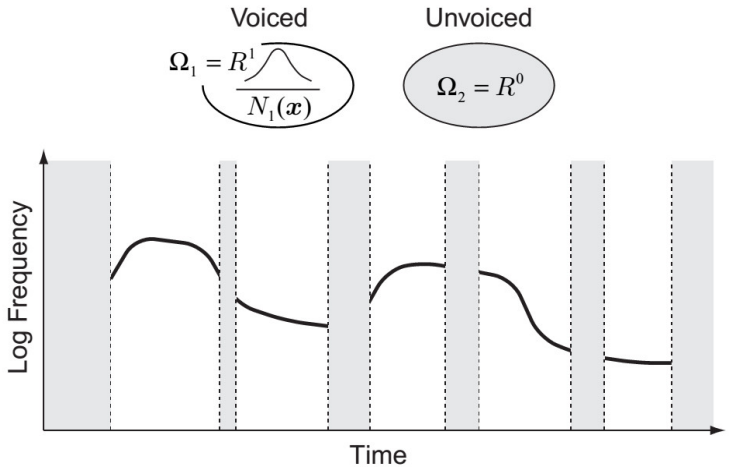


Figure 2.4: F0 pattern modeling on two spaces [25].

Figure 2.4 illustrates F0 pattern modeling using MSD in which an F0 observation is assumed to be outputted from either one-dimensional space Ω_1 for voiced segments or zero-dimensional space Ω_2 for unvoiced segments. Each space Ω_g has its own space weight w_g (referred to as MSD-weight) and satisfies the probabilistic constraint

$$\sum_{g=1}^{2} w_g = 1. \tag{2.20}$$

The space Ω_1 has a one-dimensional Gaussian probability density function $\mathcal{N}_1(\boldsymbol{x})$, while the space Ω_2 has only one sample point. An F0 observation $\boldsymbol{o}$ consists of a continuous random variable $\boldsymbol{x}$ and a set of space indices X, i.e.,

$$\boldsymbol{o} = (X, \boldsymbol{x}), \tag{2.21}$$

where $X = \{1\}$ for voiced region and $X = \{0\}$ for unvoiced region. The probability of observation $\boldsymbol{o}$ is defined by

$$b(\boldsymbol{o}) = \sum_{g \in S(\boldsymbol{o})} w_g \mathcal{N}_g\left(V(\boldsymbol{o})\right), \tag{2.22}$$

where $V(\boldsymbol{o}) = \boldsymbol{x}$ and $S(\boldsymbol{o}) = X$. It should be noted that $\mathcal{N}_2(\boldsymbol{x}) \equiv 1$ for notational simplicity.

By using an HMM in which the state-output probability distribution is an MSD as specified in Eq. (2.22), hereinafter referred to as MSD-HMM, F0 observations for voiced and unvoiced regions can be modeled in a unified model without any heuristic assumption [24]. Figure 2.5 shows the structure of MSD-HMM specialized for F0 modeling. Each

state has MSD-weights (or voiced/unvoiced weights) which represent the probabilities of voiced and unvoiced, and a continuous distribution for voiced observations. The training of MSD-HMM under the ML criterion using the EM algorithm is similar to that of CD-HMM (Section 2.3.1(2)), and re-estimation formulae can be derived in a similar manner as detailed in [24].

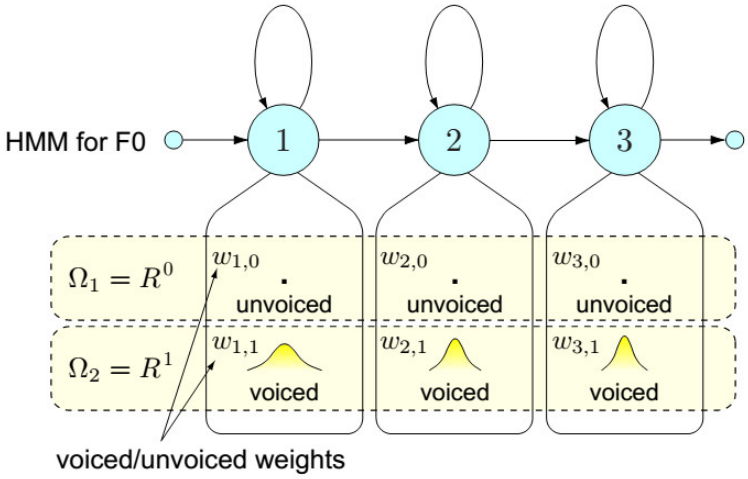


Figure 2.5: MSD-HMM for F0 modeling [22].

However, if spectrum and F0 are modeled separately, speech segmentations may be discrepant between them. To avoid this problem, they are jointly modeled by multi-stream MSD-HMM, in which the spectral part is modeled by continuous distribution and the F0 part is modeled by MSD (Figure 2.6). In the figure, $\boldsymbol{c}_t$, X_t^p, and $\boldsymbol{x}_t^p$ represent the spectral parameter vector, a set of space indices of F0, and F0 parameter at time t, respectively; Δ and Δ^2 represent the delta and delta-delta parameters, respectively.

2.3.3. Dynamic feature calculation

In HMM-based TTS, not only spectral and F0 parameters (called as static features) but also their delta and delta-delta counterparts (referred to as dynamic features) are modeled by HMMs. These dynamic features capture dynamic properties of speech parameter trajectories. Therefore, the integration of dynamic features into the feature vector of a speech frame is essential for the modeling and generation of parameter trajectories. In this thesis, spectral dynamic features are calculated as follows

$$\Delta \boldsymbol{c}_t = 0.5\left(\boldsymbol{c}_{t+1} - \boldsymbol{c}_{t-1}\right), \tag{2.23}$$

$$\Delta^2 \boldsymbol{c}_t = \boldsymbol{c}_{t+1} - 2\boldsymbol{c}_t + \boldsymbol{c}_{t-1}. \tag{2.24}$$

Similarly, F0 dynamic features are given by

$$\boldsymbol{x}_t^{\Delta p} = 0.5\left(\boldsymbol{x}_{t+1}^{p} - \boldsymbol{x}_{t-1}^{p}\right), \tag{2.25}$$

$$\boldsymbol{x}_t^{\Delta^2 p} = \boldsymbol{x}_{t+1}^{p} - 2\boldsymbol{x}_t^{p} + \boldsymbol{x}_{t-1}^{p}. \tag{2.26}$$

In unvoiced regions, $\boldsymbol{x}_t^{p}$, $\boldsymbol{x}_t^{\Delta p}$, and $\boldsymbol{x}_t^{\Delta^2 p}$ are defined as the discrete symbol. For the frames at the boundary between voiced and unvoiced regions where the F0 dynamic features cannot be calculated, they are also defined as the discrete symbol.

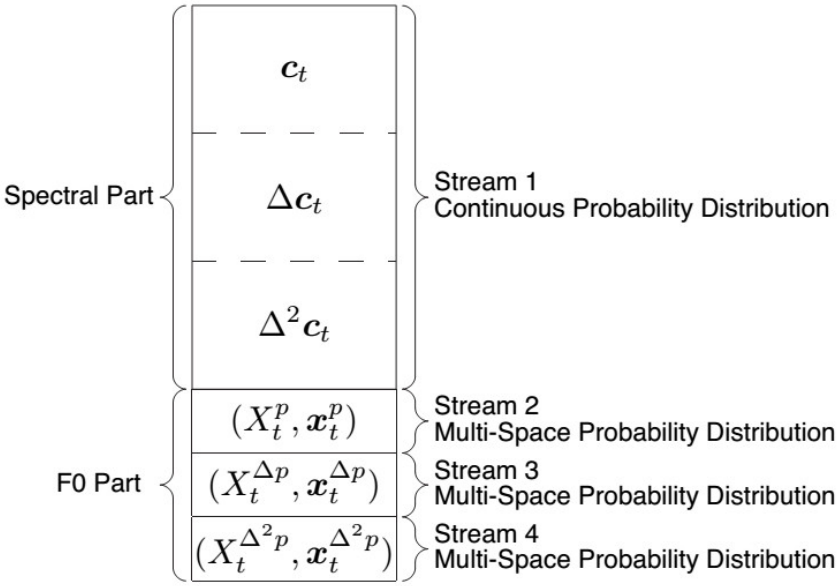


Figure 2.6: Feature vector of a speech frame [26].

2.3.4. Duration modeling

In HMM-based TTS, rhythm and tempo of synthesized speech is controlled by the duration of HMM states, i.e., how many consecutive speech frames that an HMM state spans. Since a left-to-right HMM with no skip is usually used as the model of a phoneme, a state can be regarded as the model of a sub-phoneme unit. To flexibly control synthesized phoneme durations, state durations are modeled by Gaussian distributions [27]. Specifically, the set of state durations of each phoneme HMM is modeled by a multi-dimensional Gaussian PDF. The dimension of this Gaussian distribution is equal to the number of states in the phoneme HMM, in which its n^{th} dimension corresponds to the duration PDF of the n^{th} state. In the training framework of conventional HMMs, these state duration distributions are not incorporated into the expectation step of the EM algorithm. Their parameters are estimated from statistical variables obtained in the last iteration of the EM algorithm, yet not re-estimated in the EM iteration [27]. However, the state duration PDFs are explicitly used in

the speech parameter generation process of the synthesis part, as described in Section 2.4. This inconsistency can make the synthesized speech sound less natural.

To resolve this inconsistency, the Hidden Semi-Markov Model (HSMM) is introduced into statistical speech synthesis in replacement of the conventional HMM [28]. An HSMM can be viewed as an HMM with explicit state duration PDFs. In other words, an HSMM is an HMM in which state transition probabilities are replaced with state duration distributions. The use of HSMMs can solve the above inconsistency because the state duration PDFs can be explicitly incorporated into both the training and synthesis parts of the system. The HSMM training is performed according to the generalized forward-backward algorithm (expectation step) and parameter re-estimation formulae (maximization step) are detailed in [28]. Nowadays, most of statistical speech synthesis systems, including the ones used for the experiments in this thesis, use HSMMs as the acoustic models. However, the field is still named HMM-based speech synthesis as a convention, and the term "HMM" is often used instead of "HSMM".

2.3.5. Context-dependent modeling and context clustering

The realization of acoustic parameters such as spectrum, excitation, and duration in natural speech is affected by different phonetic, prosodic, and linguistic factors. The factors which can affect the acoustic realization of a phone are referred to as its contexts. To achieve high-quality synthesized speech, a large set of contexts is required to be represented. Widely used contexts for speech synthesis include [29]:

- Identity of neighboring phones to the central phone. Normally, two phones to the left and the right of the central phone are considered as phonetic contexts (the so-called quinphone context).
- Phonological categories (e.g., consonant/vowel/fricative, voiced/unvoiced).
- Position of phones, syllables, words, phrases with respect to higher level units.
- Number of phones, syllables, words, phrases with respect to higher level units.
- Syllable stress for stressed languages (e.g. English), accent status for accentual languages (e.g., Japanese), or tone identity for tonal languages (e.g., Vietnamese).
- Linguistic roles (e.g., part-of-speech tag).

Each phoneme is associated with a label that integrates all of the contextual information related to it, often referred to as *full-context label*. In order to handle contextual complexity, a distinct HMM should be used for each combination of possible contexts, referred to as a *context-dependent HMM*. However, the total number of possible combinations of these factors exponentially increases with the number of contexts taken

into account, which can be around 50. The amount of available training data is normally inadequate for robustly estimating all context-dependent HMMs since there is rarely sufficient data to cover all of the contextual combinations. Besides, there is great variation in the appearance frequency of each context-dependent unit. To alleviate these problems, top-down decision tree based context clustering is widely used to cluster HMM states and share model parameters among states in each cluster [30, 31].

An example of context clustering based on decision tree is shown in Figure 2.7. The decision tree is a binary tree. Each node (except for leaf nodes) has a context-related question, such as R-silence? ("Is the current phoneme on the right of a silence?") or L-vowel? ("Is the current phoneme on the left of a vowel?"), and two child nodes representing "Yes" and "No" answers to the question. Leaf nodes have state output distributions. Using the decision tree based context clustering, model parameters of the speech units for unseen contexts can be obtained because any context reaches one of the leaf nodes by going down the tree, starting from the root node then selecting the next node depending on the answer about the current context.

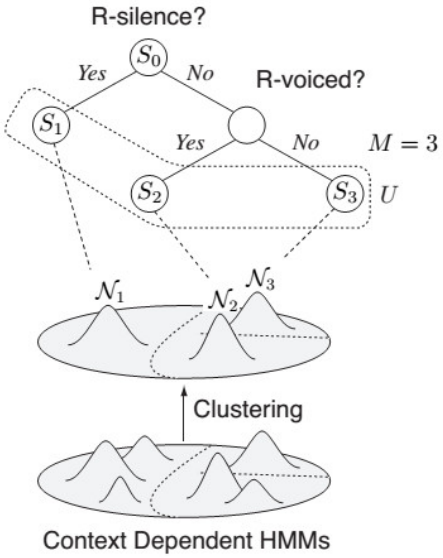


Figure 2.7: An example of decision tree based context clustering [26].

The tree growing process using the minimum description length (MDL) criterion [32], which exhibits the trade-off between model complexity and likelihood increases, is summarized as follows:

1. A set of context-dependent HMMs with single Gaussian output distribution per state is trained under the ML criterion as described in Section 2.3.1(2).

2. The estimated distributions of all states to be clustered are gathered and placed in the root node of the tree, and the likelihood of the training data is calculated with an assumption that all of the states are tied (i.e., model parameters are shared among the states).
3. For each leaf node, a question that gives maximum increase in the likelihood of the training data when the leaf node is split into two using the question is found.
4. Among all leaf nodes, the node that minimizes the description length of the model set when split into two using the question found in step 3 is chosen.
5. The selected node is split into two if the description length of the after-splitting model set is smaller than that of the before-splitting one. Otherwise, the tree growing procedure is stopped.
6. Steps 3, 4, and 5 are repeated until the tree growing procedure is stopped.

It should be noted that, one tree is built for each state index to construct the parameter sharing structure. Furthermore, separate trees are built for spectrum, excitation, and duration because each of them has its own context dependency. Although decision tree based context clustering technique was originally derived for HMM, it can be applied to HSMM with no modification. It also has been extended for MSD-HMMs in [33].

2.4. Synthesis stage

In the synthesis stage, the text to be synthesized is firstly converted into a full-context label sequence by a text analyzer. Based on this label sequence, a sentence HMM is built by concatenating the corresponding context-dependent phoneme HMMs. After that, state durations of the sentence HMM are calculated so as to maximize the state duration probability of the state sequence [27]. Based on the obtained state durations, sequences of mel-cepstral coefficients and F0 values are then generated so as to maximize their output probability given the sentence HMM [34]. Finally, the MLSA filter [16] is used to synthesize a speech waveform from the obtained mel-cepstral and F0 sequences. The whole synthesis process is illustrated in Figure 2.8.

2.4.1. Text analysis

The task of a text analyzer is to extract contextual information described in Section 2.3.5, and convert them into a full-context label sequence from an input text. The experiments in this thesis use available contextual labels provided by the HMM-based speech synthesis

system (HTS) toolkit [35] for the Japanese TTS system (Chapter 3), and the freely available text analysis tool JVnTextPro [36] for the Vietnamese TTS system (Chapter 4).

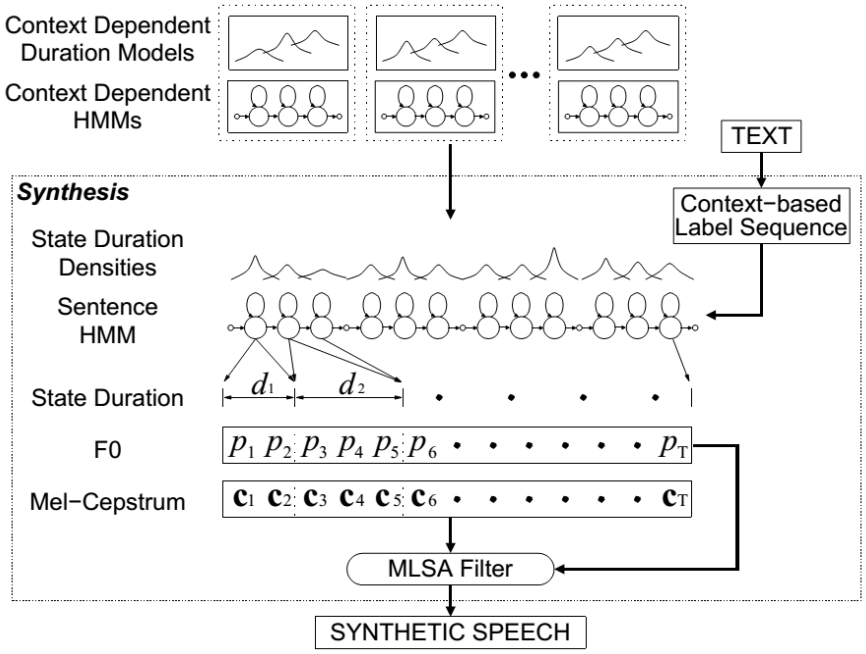


Figure 2.8: Block diagram of the synthesis stage [33].

2.4.2. State duration determination

Given the contextual label sequence W, the estimated sentence HMM $\hat{\lambda}$ (a left-to-right with no skip structure is assumed), and the desired length (in frames) of the synthesized speech T, the probability of a state sequence $\boldsymbol{q} = (q_1, q_2, ..., q_T)$ is given as [27]

$$P(\boldsymbol{q} \mid W, \hat{\lambda}) = \prod_{k=1}^{K} p_k(d_k), \tag{2.27}$$

where $p_k(d_k)$ is the probability of being in state k for d_k frames, K is the number of states in the HMM $\hat{\lambda}$, and

$$\sum_{k=1}^{K} d_k = T. \tag{2.28}$$

When each state duration distribution is modeled by a single Gaussian PDF

$$p_k(d_k) = \mathcal{N}\left(d_k; \mu_k, \sigma_k^2\right) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{\left(d_k - \mu_k\right)^2}{2\sigma_k^2}\right), \tag{2.29}$$

the state durations $\{d_k\}_{k=1}^{K}$ that maximize Eq. (2.27) under the constraint of Eq. (2.28) are given by:

$$d_k = \mu_k + \rho \cdot \sigma_k^2, \quad 1 \le k \le K, \tag{2.30}$$

$$\rho = \left(T - \sum_{k=1}^{K} \mu_k \right) / \sum_{k=1}^{K} \sigma_k^2, \tag{2.31}$$

where μ_k and σ_k^2 are the mean and variance of the duration density of state k, respectively (Figure 2.9).

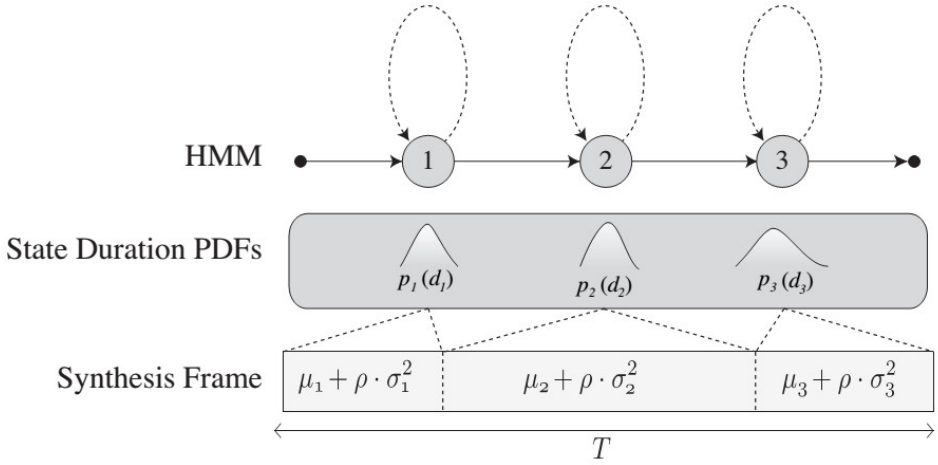


Figure 2.9: State duration generation [26].

It is noted that the speaking rate can be controlled by ρ instead of T since they are related each other by Eq. (2.31). To synthesize speech with average speaking rate, ρ should be set to 0, i.e.,

$$T = \sum_{k=1}^{K} \mu_k \,. \tag{2.32}$$

In order to increase or decrease the speaking rate, ρ is set to a positive or negative value, respectively.

2.4.3. Effect of dynamic features in speech parameter generation

The state sequence $\hat{\boldsymbol{q}} = (q_1, q_2, ..., q_T)$ used for the synthesis can be trivially inferred from the synthesized state durations $\{d_k\}_{k=1}^{K}$, from which a feature vector sequence $\boldsymbol{o} = \left(\boldsymbol{o}_1^{\mathrm{T}}, \boldsymbol{o}_2^{\mathrm{T}}, \ldots, \boldsymbol{o}_T^{\mathrm{T}}\right)^{\mathrm{T}}$ (T denotes the matrix transpose operation) is generated so as to maximize its output probability, given the estimated model set $\hat{\lambda}$, that is [34],

$$\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{o}} P(\boldsymbol{o} \mid \hat{\boldsymbol{q}}, \hat{\lambda}) \,. \tag{2.33}$$

The solution for Eq. (2.33) will be presented in details in Section 3.2.1. In the following, the effect of dynamic features in the speech parameter generation is briefly considered.

To simplify the notation, it is assumed that each state output distribution is a single multivariate Gaussian PDF, i.e.,

$$b_k(\boldsymbol{o}_t) = \mathcal{N}\left(\boldsymbol{o}_t; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right), \tag{2.34}$$

where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the mean vector and covariance matrix of the k^{th} state, respectively. From Eqs. (2.33) and (2.34), we have

$$\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{o}} \mathcal{N}\left(\boldsymbol{o}; \boldsymbol{\mu}_{\hat{q}}, \boldsymbol{\Sigma}_{\hat{q}}\right), \tag{2.35}$$

where $\boldsymbol{\mu}_{\hat{q}} = \left(\boldsymbol{\mu}_{q_1}^{\mathrm{T}}, \boldsymbol{\mu}_{q_2}^{\mathrm{T}}, \ldots, \boldsymbol{\mu}_{q_T}^{\mathrm{T}}\right)^{\mathrm{T}}$ and $\boldsymbol{\Sigma}_{\hat{q}} = diag\left(\boldsymbol{\Sigma}_{q_1}, \boldsymbol{\Sigma}_{q_2}, \ldots, \boldsymbol{\Sigma}_{q_T}\right)$ are respectively the mean vector and covariance matrix related to the state sequence $\hat{\boldsymbol{q}}$.

If the feature vector at time t only included the static parameters, i.e., $\boldsymbol{o}_t = \boldsymbol{c}_t$, the generated feature vector sequence $\hat{\boldsymbol{o}}$ according to Eq. (2.35) would be the mean vector sequence $\boldsymbol{\mu}_{\hat{q}}$ due to the Gaussian PDF assumption (the red horizontal lines in Figure 2.10). Such step-wise parameter trajectories are poor representation of natural speech. They would severely degrade the synthesized speech quality due to the discontinuities occurring at state boundaries.

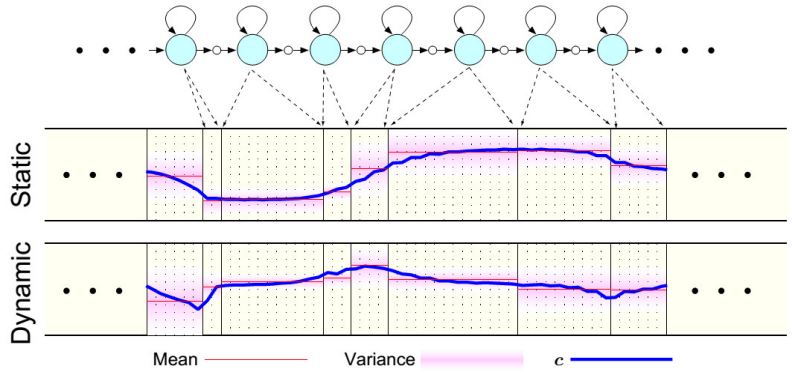


Figure 2.10: Generated speech parameter trajectory [22] (showing only one dimension of the feature vector). Delta parameters are shown as representative for dynamic features.

To generate more speech-like parameter trajectories, relationships between static and dynamic features are introduced as the constraints for the maximization problem of Eq. (2.35) by considering the fact that the feature vector $\boldsymbol{o}_t$ consists of not only the static feature vector $\boldsymbol{c}_t$ but also the first and second order dynamic feature vectors $\Delta\boldsymbol{c}_t$ and $\Delta^2\boldsymbol{c}_t$, that is,

$$\boldsymbol{o}_t = \left(\boldsymbol{c}_t^{\mathrm{T}}, \Delta\boldsymbol{c}_t^{\mathrm{T}}, \Delta^2\boldsymbol{c}_t^{\mathrm{T}}\right)^{\mathrm{T}}, \tag{2.36}$$

where the dynamic features can be calculated, for example, as given in Section 2.3.3. By solving this constrained maximization problem (see Section 3.2.1), generated parameter trajectories are no longer piece-wise stationary since associated dynamic features also contribute to the likelihood. The inclusion of the dynamic features into the feature vector makes generated parameter trajectories become smooth and realistic similar to the solid blue lines in Figure 2.10.

Chapter 3 Minimum Generation Error Training Considering Dynamic Features

3.1. Introduction

In HMM-based TTS, the use of dynamic features (e.g., delta and delta-delta cepstral coefficients) is crucial for generating smoothly varying parameter trajectories (see Section 2.4.3). In the training phase, dynamic features are modeled independently with static feature. In the conventional system described in Chapter 2, HMM parameters are trained under the maximum likelihood (ML) criterion. In the synthesis phase, the most probable parameter sequence is generated given the distributions of static and dynamic features using the speech parameter generation algorithm [34]. Here, the constraints between static and dynamic features are taken into account to generate smooth, realistic feature trajectories. Although dynamic features of both spectral and F0 parameters are used in HMM-based TTS, this chapter only focuses on the modeling of dynamic spectral features.

Considering the ignorance of the constraints between static and dynamic features in the training phase and the mismatch between the ML-based training criterion and the objective of speech synthesis, the minimum generation error (MGE) criterion [7] has been proposed for training HMMs. By incorporating the parameter generation process into the HMM training, the error between the original and generated data for a training sentence can be calculated as a function of HMM parameters, which is called the generation error function (GEF). HMM parameters are then re-estimated to minimize the total generation error for all training sentences. As a result, the above two issues of the ML-training-based conventional system can be solved effectively, and improved synthetic speech quality has been reported [7].

A key point in MGE training is the definition of the GEF. In the baseline MGE criterion [7], the GEF is defined as the Euclidean distance between the natural and generated static feature vector sequences. Under the viewpoint of parameter trajectory modeling, this has a drawback, which is the ignorance of dynamic properties of parameter

trajectories in the generation error definition. These dynamic properties are captured, to a certain extent, by the dynamic features of speech parameters since the dynamic features of a speech frame are generally calculated as regression coefficients from the static features of neighboring frames [37]. Moreover, dynamic features convey spectral transition information, which is believed to be an important acoustic cue in speech perception [38]. Therefore, it is expected that the introduction of the error component of dynamic features into the GEF could be beneficial.

In this chapter, the generation error of dynamic features is firstly defined and incorporated into the GEF. Then the newly derived GEF is minimized under the MGE criterion. It is worth pointing out that the objective of this research is different from those of recent improvements of the baseline MGE criterion [39-41]. In [39], the error component of the global/local variance of feature trajectories was introduced into the GEF to obtain an over-smoothing alleviation effect similar to the parameter generation algorithm considering global variance (GV) [42] without introducing any extra computational cost during synthesis. In [40] and [41], two perceptually motivated distance metrics on line spectral pairs (LSPs), log spectral distortion and weighted Euclidean distance, respectively, were proposed to enhance the correlation between the objective GEF and the subjective perception of spectral distortion. In contrast, this research aims to investigate the effect of more accurately modeling the dynamic properties of speech parameters under the MGE training framework.

The chapter is organized as follows. Sections 3.2 and 3.3 review the speech parameter generation algorithms and the baseline MGE criterion, respectively. The proposed MGE criterion considering dynamic properties of speech parameters is described in Section 3.4. Section 3.5 presents experimental results. Section 3.6 comprises several discussions. Finally, conclusions are given in Section 3.7.

3.2. Speech parameter generation algorithms

This section provides brief reviews of the speech parameter generation algorithms used in the experiments. First, the original parameter generation algorithm is described. Then, the parameter generation algorithm considering GV, which alleviates the over-smoothing effect of the original algorithm, is presented.

3.2.1. Original parameter generation algorithm

For a given HMM λ and a state sequence $\boldsymbol{q}$, the speech parameter generation algorithm aims to generate the parameter vector sequence $\boldsymbol{o} = \left[\boldsymbol{o}_1^{\mathrm{T}}, \boldsymbol{o}_2^{\mathrm{T}}, ..., \boldsymbol{o}_T^{\mathrm{T}}\right]^{\mathrm{T}}$ by maximizing $P\left(\boldsymbol{o}\middle|\lambda, \boldsymbol{q}\right)$ with respect to $\boldsymbol{o}$ [34] (Case 1), where T is the number of frames in $\boldsymbol{o}$ and T denotes the matrix transpose operation. The t^{th} frame's parameter vector $\boldsymbol{o}_t$ includes the M-dimensional static feature vector $\boldsymbol{c}_t$ and dynamic feature vectors $\Delta^{(1)}\boldsymbol{c}_t$ and $\Delta^{(2)}\boldsymbol{c}_t$, and can be written as $\boldsymbol{o}_t = \left[\boldsymbol{c}_t^{\mathrm{T}}, \Delta^{(1)}\boldsymbol{c}_t^{\mathrm{T}}, \Delta^{(2)}\boldsymbol{c}_t^{\mathrm{T}}\right]^{\mathrm{T}}$, where the d^{th}-order dynamic feature vector is calculated as

$$\Delta^{(d)}\boldsymbol{c}_t = \sum_{\tau=-L_-^{(d)}}^{L_+^{(d)}} w^{(d)}(\tau)\boldsymbol{c}_{t+\tau}. \tag{3.1}$$

Here, $w^{(d)}(\cdot)$ are window coefficients; $L_-^{(d)}$ and $L_+^{(d)}$ are the numbers of frames preceding and succeeding frame t involved in the calculation of $\Delta^{(d)}\boldsymbol{c}_t$ $\left(d = 1, 2\right)$, respectively.

From Eq. (3.1), the constraints between static and dynamic features can be expressed as $\boldsymbol{o} = \boldsymbol{W}\boldsymbol{c}$, where $\boldsymbol{c} = \left[\boldsymbol{c}_1^{\mathrm{T}}, \boldsymbol{c}_2^{\mathrm{T}}, ..., \boldsymbol{c}_T^{\mathrm{T}}\right]^{\mathrm{T}}$ and $\boldsymbol{W}$ is a $3MT \times MT$ window matrix defined as

$$\boldsymbol{W} = [\boldsymbol{W}_1, \boldsymbol{W}_2, \ldots, \boldsymbol{W}_T]^{\mathrm{T}} \otimes \boldsymbol{I}_{M\times M}, \tag{3.2}$$

$$\boldsymbol{W}_t = [\boldsymbol{w}_t^{(0)}, \boldsymbol{w}_t^{(1)}, \boldsymbol{w}_t^{(2)}], \tag{3.3}$$

$$\boldsymbol{w}_t^{(0)} = [\underbrace{0, \ldots, 0}_{t-1}, 1, \underbrace{0, \ldots, 0}_{T-t}]^{\mathrm{T}}, \tag{3.4}$$

$$\boldsymbol{w}_t^{(d)} = \left[\underbrace{0, \ldots, 0}_{t-L_-^{(d)}-1}, w^{(d)}(-L_-^{(d)}), \ldots, w^{(d)}(0), \ldots, w^{(d)}(L_+^{(d)}), \underbrace{0, \ldots, 0}_{T-(t+L_+^{(d)})}\right]^{\mathrm{T}}, \tag{3.5}$$

where $\otimes$ denotes the Kronecker product operation and $\boldsymbol{I}$ is the identity matrix.

For the particular formulae of dynamic feature calculation specified in Section 2.3.3, $\boldsymbol{W}$ is given as

$$\boldsymbol{W} = \begin{bmatrix} \cdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots \\ \cdots & \boldsymbol{0} & \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \cdots \\ \cdots & -1/2\boldsymbol{I} & \boldsymbol{0} & 1/2\boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} & \cdots \\ \cdots & \boldsymbol{I} & -2\boldsymbol{I} & \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} & \cdots \\ \cdots & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} & \cdots \\ \cdots & \boldsymbol{0} & -1/2\boldsymbol{I} & \boldsymbol{0} & 1/2\boldsymbol{I} & \boldsymbol{0} & \cdots \\ \cdots & \boldsymbol{0} & \boldsymbol{I} & -2\boldsymbol{I} & \boldsymbol{I} & \boldsymbol{0} & \cdots \\ \cdots & & & & & & \end{bmatrix}, \tag{3.6}$$

where each zero or identity matrix in $\boldsymbol{W}$ is a $M \times M$ matrix.

Under these constraints, finding $\boldsymbol{o}$ that maximizes $P(\boldsymbol{o}|\lambda,\boldsymbol{q})$ is equivalent to finding $\boldsymbol{c}$ that maximizes $P(\boldsymbol{o}|\lambda,\boldsymbol{q})$. By solving $\partial P(\boldsymbol{o}|\lambda,\boldsymbol{q})/\partial \boldsymbol{c} = 0$, the generated static feature vector sequence is obtained as

$$\overline{\boldsymbol{c}}_{\boldsymbol{q}} = \boldsymbol{R}_{\boldsymbol{q}}^{-1}\boldsymbol{r}_{\boldsymbol{q}}, \tag{3.7}$$

where

$$\boldsymbol{R}_{\boldsymbol{q}} = \boldsymbol{W}^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1}\boldsymbol{W}, \tag{3.8}$$

$$\boldsymbol{r}_{\boldsymbol{q}} = \boldsymbol{W}^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1}\boldsymbol{\mu}_{\boldsymbol{q}}, \tag{3.9}$$

and $\boldsymbol{\mu}_{\boldsymbol{q}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{q}}$ are the mean vector and covariance matrix related to $\boldsymbol{q}$ as specified in Section 2.4.3, respectively [34].

3.2.2. Parameter generation algorithm considering global variance

With the original parameter generation algorithm, intelligible and smooth speech can be synthesized. However, synthetic speech is evidently muffled compared with natural speech because the generated speech parameter trajectories are often over-smoothed. It is due to the fact that detailed characteristics of speech parameters are removed in the statistical averaging process in the modeling stage and cannot be recovered in the generation stage. Figure 3.1 shows the trajectories of second mel-cepstral coefficients extracted from natural speech and those generated from an HMM. We can see that the dynamic range of the generated mel-cepstral coefficients is smaller than that of the natural ones. The speech parameter generation algorithm considering GV [42] tries to recover the dynamic range of generated trajectories close to that of natural ones. A GV, $\boldsymbol{v}(\boldsymbol{c})$, is defined as an intra-utterance variance of a M-dimensional T-frames parameter trajectory $\boldsymbol{c} = [\boldsymbol{c}_1^{\mathrm{T}}, \boldsymbol{c}_2^{\mathrm{T}}, ..., \boldsymbol{c}_T^{\mathrm{T}}]^{\mathrm{T}}$, that is,

$$\boldsymbol{v}(\boldsymbol{c}) = [v(1), v(2), ..., v(M)]^{\mathrm{T}}, \tag{3.10}$$

where each component of the M-dimensional GV vector is calculated as

$$v(m) = \frac{1}{T}\sum_{t=1}^{T}\{c_t(m) - \mu(m)\}^2, \tag{3.11}$$

$$\mu(m) = \frac{1}{T}\sum_{t=1}^{T} c_t(m). \tag{3.12}$$

In the above equations, $c_t(m)$ denotes the m^{th} component of the static feature vector $\boldsymbol{c}_t$.

GVs for all training utterances are calculated and modeled by using a single multivariate Gaussian distribution as

$$P(\boldsymbol{v}(\boldsymbol{c}) \mid \lambda_{GV}) = \frac{1}{(2\pi)^{M/2}\left|\boldsymbol{\Sigma}_{GV}\right|^{1/2}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{v}(\boldsymbol{c}) - \boldsymbol{\mu}_{GV}\right)^{T} \boldsymbol{\Sigma}_{GV}^{-1}\left(\boldsymbol{v}(\boldsymbol{c}) - \boldsymbol{\mu}_{GV}\right)\right\}, \tag{3.13}$$

where $\boldsymbol{\mu}_{GV}$ and $\boldsymbol{\Sigma}_{GV}$ are the mean vector and covariance matrix of the Gaussian PDF.

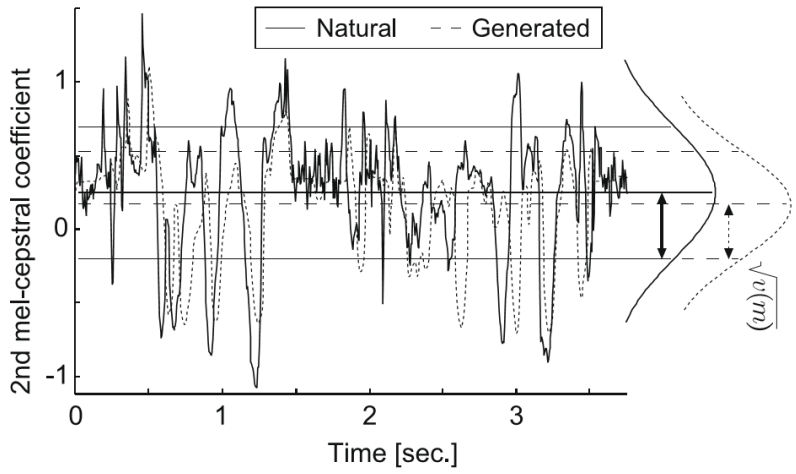


Figure 3.1: Trajectories of second mel-cepstral coefficients extracted from natural speech and that generated from HMM [2].

Instead of maximizing the probability $P(\boldsymbol{o}|\lambda,\boldsymbol{q})$ as in the original algorithm, the parameter generation algorithm considering GV maximizes the following probability represented as the product of the two probabilities,

$$P(\boldsymbol{o}\,|\,\lambda,\lambda_{GV},\boldsymbol{q}) = P\left(\boldsymbol{o}\middle|\lambda,\boldsymbol{q}\right)^{\omega} P\left(\boldsymbol{v}(\boldsymbol{c})\,|\,\lambda_{GV}\right), \tag{3.14}$$

where the constant ω is a weight to balance the HMM and GV probabilities. The second term in Eq. (3.14) can be viewed as a penalty to prevent over-smoothing because it works to retain the dynamic range of the generated trajectory close to that of the training data. This method can be viewed as a statistical post-filtering technique to a certain extent. Experiments showed that, by considering GV, the spectral structure becomes clearer and the synthetic speech quality is dramatically improved [42]. However, its effect on the F0 feature is unclear because the GV model seems too simple to capture variance characteristics of a natural F0 contour.

3.3. Minimum generation error training criterion

This section gives a review of the baseline MGE criterion [7]. In HMM-based TTS, the original speech parameter generation algorithm is used to generate the most probable feature vector sequence. Then HMM parameters are optimized to minimize the total generation error of all training data under the MGE criterion. The following subsections follow the notations and formulations in [39] for the sake of coherence and compactness.

In the baseline MGE criterion, the Euclidean distance is used to measure the error between the original and generated static feature vector sequences, which is

$$D(\boldsymbol{c}, \bar{\boldsymbol{c}}_q) = \left\| \boldsymbol{c} - \bar{\boldsymbol{c}}_q \right\|^2. \tag{3.15}$$

Theoretically, all possible state sequences underlying the original parameter vector sequence $\boldsymbol{o}$ could be involved in the calculation of the generation error, where $P(\boldsymbol{q}|\lambda, \boldsymbol{o})$ can be used to weight the error corresponding to the state sequence $\boldsymbol{q}$. However, this is computationally expensive. In practice, only the most probable state sequence $\hat{\boldsymbol{q}}$ for $\boldsymbol{o}$ is used and the GEF is defined as

$$e(\boldsymbol{c}, \lambda) = D(\boldsymbol{c}, \bar{\boldsymbol{c}}_{\hat{q}}). \tag{3.16}$$

For notational convenience, $\boldsymbol{q}$ is used to denote $\hat{\boldsymbol{q}}$ in the rest of the chapter.

The MGE criterion aims to minimize the total generation error for all training sentences,

$$\hat{\lambda} = \arg\min_{\lambda} \sum_{n} e(\boldsymbol{c}_n, \lambda) \tag{3.17}$$

with respect to

$$\boldsymbol{\mu} = [\boldsymbol{\mu}_1^{\mathrm{T}}, \boldsymbol{\mu}_2^{\mathrm{T}}, ..., \boldsymbol{\mu}_K^{\mathrm{T}}]^{\mathrm{T}}, \tag{3.18}$$

$$\boldsymbol{U} = [\boldsymbol{\Sigma}_1^{-1}, \boldsymbol{\Sigma}_2^{-1}, ..., \boldsymbol{\Sigma}_K^{-1}]^{\mathrm{T}}, \tag{3.19}$$

where $\boldsymbol{c}_n$ is the static feature vector sequence of the n^{th} training sentence, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the mean vector and covariance matrix of the k^{th} unique Gaussian component, respectively, and K is the total number of Gaussian components in the model set λ.

For each training data $\boldsymbol{c}_n$, the model parameters are updated using the probabilistic descent method [43] as

$$\lambda(n+1) = \lambda(n) - \varepsilon_n \nabla e(\boldsymbol{c}_n, \lambda)\big|_{\lambda=\lambda(n)}, \tag{3.20}$$

where ε_n is the learning rate, which decreases when the sentence index n increases.

The derivatives of the GEF with respect to the model set's mean and variance parameters can be derived as [39]

$$\frac{\partial e(\boldsymbol{c}_n, \lambda)}{\partial \boldsymbol{\mu}} = 2\boldsymbol{S}_q^{\mathrm{T}} \boldsymbol{\Sigma}_q^{-1} \boldsymbol{W} \boldsymbol{R}_q^{-1} \boldsymbol{\varsigma}, \tag{3.21}$$

$$\frac{\partial e(\boldsymbol{c}_n, \lambda)}{\partial \boldsymbol{U}} = 2\boldsymbol{S}_q^{\mathrm{T}} \mathrm{diag}^{-1}\left(\boldsymbol{W} \boldsymbol{R}_q^{-1} \boldsymbol{\varsigma} (\boldsymbol{\mu}_q - \boldsymbol{W} \bar{\boldsymbol{c}}_q)^{\mathrm{T}} \right), \tag{3.22}$$

where

$$\boldsymbol{\Sigma}_q^{-1} = \mathrm{diag}(\boldsymbol{S}_q \boldsymbol{U}), \tag{3.23}$$

$$\boldsymbol{\mu}_q = \boldsymbol{S}_q \boldsymbol{\mu}, \tag{3.24}$$

$$\boldsymbol{\varsigma} = \bar{\boldsymbol{c}}_q - \boldsymbol{c}_n. \tag{3.25}$$

In the above equations, $\boldsymbol{S}_q$ is a $3MT \times 3MK$ matrix whose elements are 0 or 1, determined according to the optimal state sequence $\boldsymbol{q}$ for $\boldsymbol{c}_n$, $\mathrm{diag}(.)$ is the operation to convert a $3MT \times 3M$ matrix to a $3MT \times 3MT$ block-diagonal matrix with a block size of $3M$, and $\mathrm{diag}^{-1}(.)$ is the inverse operation of $\mathrm{diag}(.)$.

It should be noted that the highest computational cost of MGE training is related to the calculation of $\boldsymbol{R}_q^{-1}$ in Eqs. (3.21) and (3.22). This cost can be markedly reduced by using an approximation of $\boldsymbol{R}_q^{-1}$ taking into account the special structure of the matrix $\boldsymbol{R}_q$ [7].

3.4. Proposed MGE criterion considering dynamic features

The baseline MGE criterion described in the previous section has a drawback, which is the ignorance of dynamic properties of speech parameters in the generation error definition. This section incorporates these dynamic properties into MGE training by defining the generation error of dynamic features and introducing this new error component into the GEF. Controlling the weight associated with the newly added error component is of essential importance in balancing the performance of the two error components comprising the newly derived GEF. Two methods are proposed for setting this weight: fixed and adaptive weighting. The effects of these two weighting methods are discussed in the next section.

The generation error of the d^{th}-order dynamic feature is defined as the Euclidean distance between the d^{th}-order dynamic feature vector sequences derived from the corresponding original and generated static feature vector sequences, that is,

$$D\left(\Delta^{(d)}\boldsymbol{c}, \Delta^{(d)}\overline{\boldsymbol{c}}_{\boldsymbol{q}}\right) = \left\|\Delta^{(d)}\boldsymbol{c} - \Delta^{(d)}\overline{\boldsymbol{c}}_{\boldsymbol{q}}\right\|^2, \tag{3.26}$$

where

$$\Delta^{(d)}\boldsymbol{c} = \boldsymbol{A}_d\boldsymbol{c}, \tag{3.27}$$

$$\Delta^{(d)}\overline{\boldsymbol{c}}_{\boldsymbol{q}} = \boldsymbol{A}_d\overline{\boldsymbol{c}}_{\boldsymbol{q}}. \tag{3.28}$$

Here, $\boldsymbol{A}_d$ is an $MT \times MT$ window matrix defined as

$$\boldsymbol{A}_d = [\boldsymbol{w}_1^{(d)}, \boldsymbol{w}_2^{(d)}, \ldots, \boldsymbol{w}_T^{(d)}]^{\mathrm{T}} \otimes \boldsymbol{I}_{M\times M}, \tag{3.29}$$

where $\boldsymbol{w}_t^{(d)}$ is given by Eq. (3.5).

The new GEF incorporating the original error component of the static feature and that of the delta feature (i.e., the first-order dynamic feature) is defined as

$$e'\left(\boldsymbol{c}, \lambda\right) = D\left(\boldsymbol{c}, \overline{\boldsymbol{c}}_{\boldsymbol{q}}\right) + aD\left(\Delta^{(1)}\boldsymbol{c}, \Delta^{(1)}\overline{\boldsymbol{c}}_{\boldsymbol{q}}\right), \tag{3.30}$$

where a is the weight associated with the error component of the delta feature, which is used to control the balance between the two error components. The effects of different values of a are discussed in the next section.

It should be noted that the GEF given by Eq. (3.30) can be extended to higher-order dynamic feature(s) in a straightforward way. The MGE criterion with the new GEF incorporating dynamic feature(s) is referred to as *MGE-dynamics* for brevity.

3.4.1. Fixed weighting approach to MGE-dynamics

In this weighting approach, the delta weight a is kept unchanged over all training data. By substituting Eqs. (3.27) and (3.28) into Eq. (3.26), the derivatives of the new GEF with respect to the mean and variance parameters can be obtained as

$$\frac{\partial e'(\boldsymbol{c}_n, \lambda)}{\partial \boldsymbol{\mu}} = 2\boldsymbol{S}_q^{\mathrm{T}} \boldsymbol{\Sigma}_q^{-1} \boldsymbol{W} \boldsymbol{R}_q^{-1} \boldsymbol{P} \boldsymbol{\varsigma}, \tag{3.31}$$

$$\frac{\partial e'(\boldsymbol{c}_n, \lambda)}{\partial \boldsymbol{U}} = 2\boldsymbol{S}_q^{\mathrm{T}} \mathrm{diag}^{-1}\left(\boldsymbol{W} \boldsymbol{R}_q^{-1} \boldsymbol{P} \boldsymbol{\varsigma} \left(\boldsymbol{\mu}_q - \boldsymbol{W} \bar{\boldsymbol{c}}_q \right)^{\mathrm{T}} \right), \tag{3.32}$$

where

$$\boldsymbol{P} = \boldsymbol{I} + a \boldsymbol{A}_1^{\mathrm{T}} \boldsymbol{A}_1 . \tag{3.33}$$

Here, $\boldsymbol{A}_1$ is given by Eq. (3.29) when d is equal to one.

Comparing the newly formulated updating rules (Eqs. (3.31) and (3.32)) with the original ones (Eqs. (3.21) and (3.22)), we can see that a matrix factor $\boldsymbol{P}$ is added as a consequence of the introduction of the delta feature error component into the GEF. It can be seen that $\boldsymbol{P}$ is a constant matrix for a given number of frames T of a training sentence and the delta weight a. Hence, the MGE-dynamics criterion with the fixed weighting approach gives rise to no additional computational complexity compared with the baseline MGE criterion.

3.4.2. Adaptive weighting approach to MGE-dynamics

The above fixed weighting approach has the effect of assigning an equal delta weight to every time sample of an utterance. However, speech consists of stationary and transitional parts, and transitional parts possess higher dynamicity than stationary ones. This suggests that the error component of the delta feature corresponding to transitional parts should be given more emphasis than that corresponding to stationary ones. Therefore, this subsection proposes an adaptive weighting approach where the delta weight is adjusted according to the degree of dynamicity of speech segments. The effects of emphasizing transitional or stationary parts of speech have also been reported in speech recognition [44].

To characterize the degree of dynamicity of speech segments, the thesis proposes to divide speech signals into portions corresponding to the state boundaries of phoneme HMMs found by Viterbi decoding [21]. The reason for this is twofold. Firstly, the states of phoneme HMMs provide natural sub-phonetic boundaries and may contain dynamic information of speech segments. Secondly, the updating rules of MGE training (i.e., Eqs.

(3.21) and (3.22) as well as Eqs. (3.31) and (3.32)) are basically performed on a state-wise basis, which means that the original and generated feature vectors of frames belonging to an HMM state are used to re-estimate the model parameters of that state HMM.

The formulation proposed in [44] is used to estimate the degree of dynamicity of a frame t, that is,

$$DF_t = \sum_{p=1}^{M-1} \left| \Delta^{(1)} c_t(p) \right|, \tag{3.34}$$

where $\Delta^{(1)} c_t(p)$ is the p^{th} component of the M-dimensional delta feature vector of frame t. Note that the zero$^{\text{th}}$ component of this vector, which is related to the log-energy of the speech signal, is excluded from the sum.

The thesis proposes to calculate the degree of dynamicity of an HMM state s as the average of the degree of dynamicity of all frames belonging to that state, i.e.,

$$DS_s = \frac{1}{T_s} \sum_{t=t_s}^{t_s+T_s-1} DF_t, \tag{3.35}$$

where T_s is the number of consecutive frames, starting from frame t_s, belonging to state s.

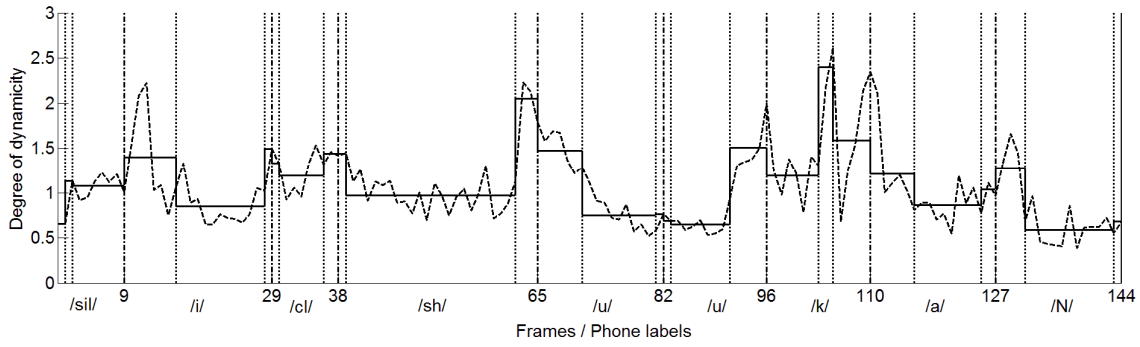


Figure 3.2: Degree of dynamicity of frames (ragged dashed line) and that of HMM states (stepwise solid line) for the Japanese utterance /sil-i-cl-sh-u-u-k-a-N/ (“one week” in English). Vertical dotted lines show HMM state boundaries and vertical dash-dotted lines associated with frame numbers show HMM phoneme boundaries.

Equation (3.35) gives a rough estimate of the degree of dynamicity of an HMM state considered as a speech segment. Figure 3.2 shows an example where the degree of dynamicity of frames and that of HMM states for an utterance are illustrated on the same plot. Here, a three-state left-to-right no-skip HMM structure was used. This example indicates that the formulation in Eqs. (3.34) and (3.35) can capture, to a certain extent, the degree of dynamicity of speech segments, even for portions where a sudden change in spectral dynamics occurs (e.g., the middle part of the stop consonant /k/).

Then, the delta weight for all frames belonging to an HMM state s is set according to the degree of dynamicity of that state as

$$a_s = a_{\max} \frac{DS_s}{DS_{\max}}, \tag{3.36}$$

where $a_{\max}$ is the maximum delta weight, which is assigned to the state possessing the maximum degree of dynamicity $DS_{\max}$ of all HMM states in the training data.

Finally, the same updating rules as those in the fixed weighting approach can be reused, although matrix $\boldsymbol{P}$ in Eq. (3.33) should be reformulated appropriately since the delta weight is adjusted state-by-state according to Eq. (3.36). Specifically, it is assumed that the n^{th} training sentence $\boldsymbol{c}_n$ has T frames belonging to N HMM states, where state i has state duration T_i, i.e., $T = \sum_{i=1}^{N} T_i$. The $MT \times MT$ banded matrix $\boldsymbol{A}_1$ given in Eq. (3.29) can be approximated as a block-diagonal matrix having block sizes that change according to the durations of HMM states, that is,

$$\boldsymbol{A}_1 \simeq \boldsymbol{A}_{1,1} \oplus \boldsymbol{A}_{1,2} \oplus \cdots \oplus \boldsymbol{A}_{1,N}, \tag{3.37}$$

where $\oplus$ denotes the direct sum operation, and $\boldsymbol{A}_{1,i}$ is an $MT_i \times MT_i$ matrix having a similar composition to $\boldsymbol{A}_1$.

Similarly, $\boldsymbol{A}_1^{\mathrm{T}}$ can be approximated as

$$\boldsymbol{A}_1^{\mathrm{T}} \simeq \boldsymbol{A}_{1,1}^{\mathrm{T}} \oplus \boldsymbol{A}_{1,2}^{\mathrm{T}} \oplus \cdots \oplus \boldsymbol{A}_{1,N}^{\mathrm{T}}. \tag{3.38}$$

The delta weight a in Eq. (3.33) is now adjusted state-by-state in the adaptive weighting approach. Thus, the matrix $a\boldsymbol{I}_{MT \times MT}$ appearing implicitly in Eq. (3.33) can be rewritten as

$$a\boldsymbol{I}_{MT \times MT} = a_1 \boldsymbol{I}_{MT_1 \times MT_1} \oplus a_2 \boldsymbol{I}_{MT_2 \times MT_2} \oplus \cdots \oplus a_N \boldsymbol{I}_{MT_N \times MT_N}, \tag{3.39}$$

where a_i is the delta weight for HMM state i, which is determined by Eq. (3.36).

From Eqs. (3.37)–(3.39), the matrix $\boldsymbol{P}$ in Eq. (3.33) can be reformulated as

$$\boldsymbol{P} \simeq \boldsymbol{I} + (a_1 \boldsymbol{A}_{1,1}^{\mathrm{T}} \boldsymbol{A}_{1,1} \oplus a_2 \boldsymbol{A}_{1,2}^{\mathrm{T}} \boldsymbol{A}_{1,2} \oplus \cdots \oplus a_N \boldsymbol{A}_{1,N}^{\mathrm{T}} \boldsymbol{A}_{1,N}). \tag{3.40}$$

Equation (3.40) allows us to compute $\boldsymbol{P}$ for a given training sentence in a straightforward manner if the delta weight for each HMM state has already been obtained following Eq. (3.36). It can be concluded that the MGE-dynamics criterion with adaptive weighting has similar computational complexity to that with fixed weighting and the baseline MGE criterion, since the highest computational cost of the training process is still related to the calculation of $\boldsymbol{R}_{\boldsymbol{q}}^{-1}$.

3.5. Experiments

To evaluate the effectiveness of the two proposed methods, i.e., the MGE-dynamics criterion with fixed weighting (*MGE-dynamics-FW*) and the MGE-dynamics criterion with

adaptive weighting (*MGE-dynamics-AW*), evaluation experiments were carried out to compare the performance of HMMs trained by the proposed techniques with that of HMMs trained by the baseline MGE technique.

3.5.1. Experimental conditions

503 phonetically balanced sentences uttered by the male speaker MHT from the ATR Japanese speech database (B-set) [45] were used in the experiments. The first 450 sentences were used for training, and the remaining 53 sentences were used for testing. Speech signals were sampled at 16 kHz and windowed by a 25 ms Hamming window with a 5 ms shift. The feature vector consists of static feature, including the 0^{th} through 24^{th} mel-cepstral coefficients obtained by a mel-cepstral analysis technique [15] and the logarithm of F0, delta and delta-delta features. A three-state left-to-right no-skip HMM structure was used. Each state output distribution was composed of spectrum and F0 streams. The spectrum stream was modeled by single multivariate Gaussian distributions with diagonal covariance matrices. The F0 stream was modeled by MSDs [8]. In the synthesis part, the MLSA filter [16] was used to synthesize the speech waveform from the generated mel-cepstral coefficients and F0 values. The experiments were based on the HTS toolkit [35], where the dynamic feature vectors are calculated as given in Section 2.3.3.

The HMM training procedure was conducted as follows. Firstly, HMMs were trained based on the conventional ML criterion [6]. Then, the resulting HMMs were used as the initial models for MGE-based training techniques. These ML-trained HMMs were also utilized to obtain the optimal state sequences for all training sentences with the Viterbi algorithm. Finally, each MGE training technique was performed iteratively until the total generation error of static and delta features for all training data converged. The online probabilistic descent updating strategy on a sentence-by-sentence basis, as described in Section 3.3, was adopted for its insensitivity to the learning rate and ease of implementation [46]. The learning rate ε_n in Eq. (3.20) was empirically set as

$$\varepsilon_n = \frac{1}{100+n}. \tag{3.41}$$

For each training sentence, HMM parameters related to the optimal state sequence were re-estimated state-by-state using the updating rules. In the experiments, only spectral model parameters were updated as the effect of dynamic spectral features is of interest in this research.

First, the three MGE training techniques were objectively and subjectively evaluated with the original parameter generation algorithm [34], which was used in the MGE training

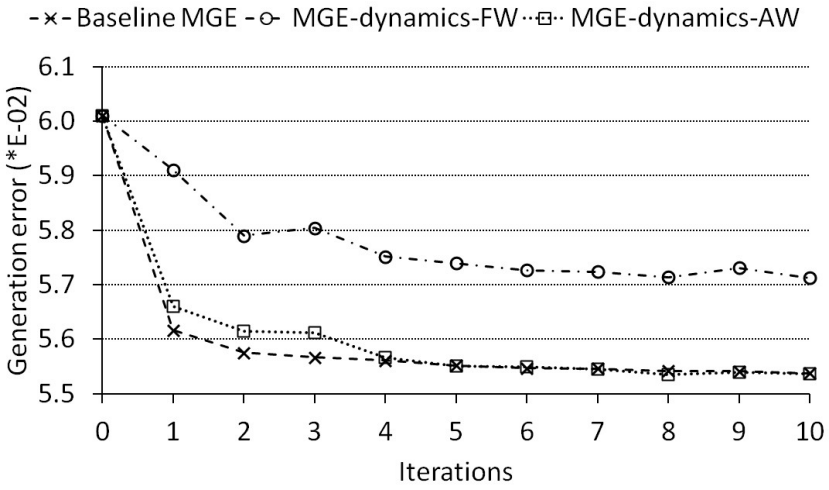
framework. Then they were subjectively evaluated with the parameter generation algorithm considering GV [42] to show their effectiveness under this highly effective synthesis scheme. No-silence GV modeling in which the GV weight was set to 1.0 was used. In all evaluation experiments, the optimal state sequences obtained by the forced alignment of the original speech features with the ML-trained HMMs were used for synthesis.

3.5.2. Evaluation with the original parameter generation algorithm

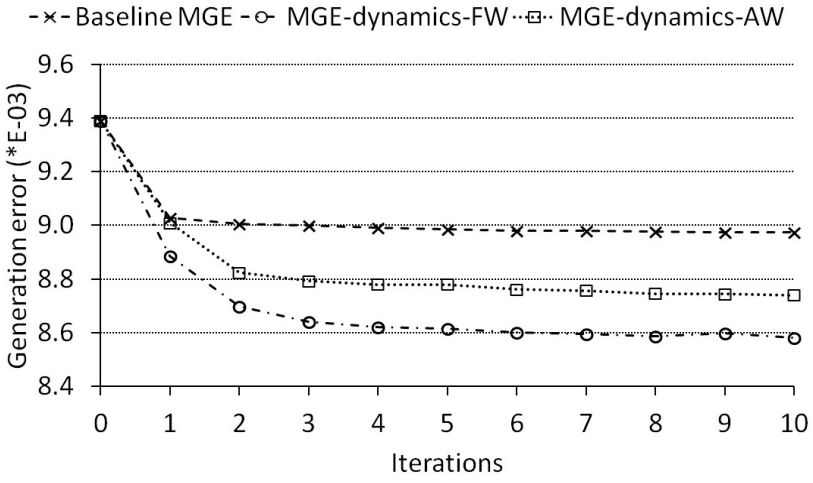
3.5.2(1) Objective evaluation

Since multivariate Gaussian distributions with diagonal covariance matrices were used for spectral parameter modeling, HMM parameters were optimized and the generation error was calculated independently for each dimension of mel-cepstral coefficients. With the above setting of the learning rate, it was observed that all MGE training techniques under investigation converged within 10 iterations. Figure 3.3 shows plots of the evolution of the generation error of the 2nd mel-cepstral coefficient on the test data as an example, where the delta weight for MGE-dynamics-FW and the maximum delta weight for MGE-dynamics-AW were both set to 100. Similar evolutions were also observed for other dimensions, on the training data, and for other settings related to the delta weight.

Figure 3.4 shows the relative error reduction (the performance of ML training was used as the reference) of static and delta features on the test data for several representative mel-cepstrum orders after MGE-dynamics-FW training with various settings of the delta weight. When the delta weight was set to zero, we obtained the result of baseline MGE training. It can be seen that baseline MGE training reduces the generation error of the delta feature as a side effect, although its GEF does not incorporate the delta feature. When the delta weight is increased, the relative error reduction of the static feature exhibits a steady downward trend while that of the delta feature increases and saturates as the delta weight approaches 100. This trade-off between the performances of error components included in the GEF was also observed in MGE-dynamics-FW training when the delta-delta feature was incorporated into its GEF in a similar manner to that described in Section 3.4. Considering the lower significance of the delta-delta feature compared with its static and delta counterparts, it is sufficient to investigate the effect of introducing the delta feature, without considering higher-order dynamic features, into MGE training in this thesis.

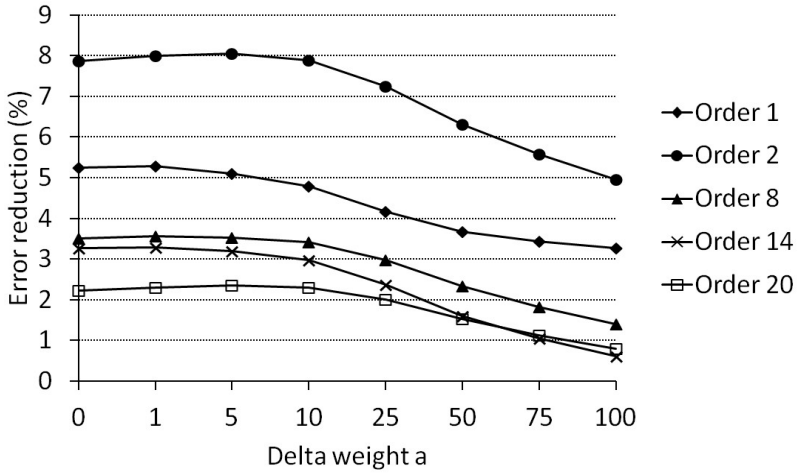


(a) Generation error of static feature.

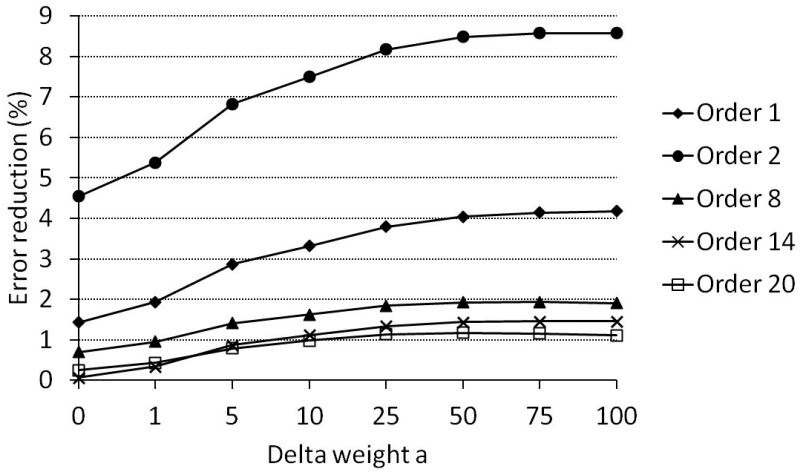


(b) Generation error of delta feature.

Figure 3.3: Evolution of generation error of the 2nd mel-cepstral coefficient on test data.



(a) Relative error reduction of static feature.

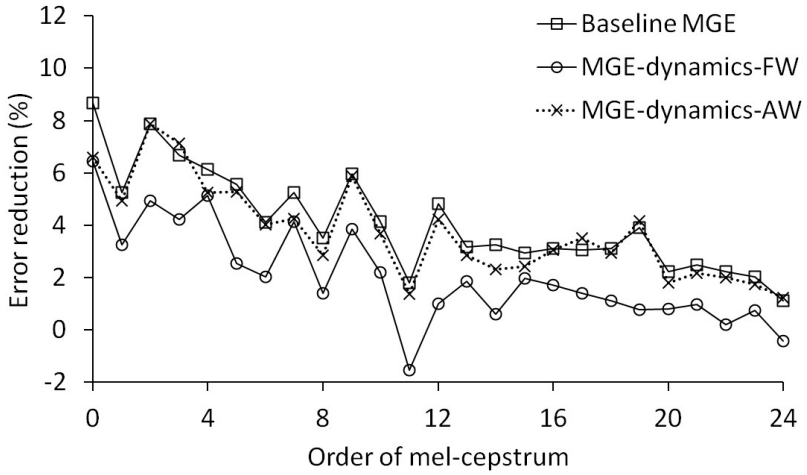


(b) Relative error reduction of delta feature.

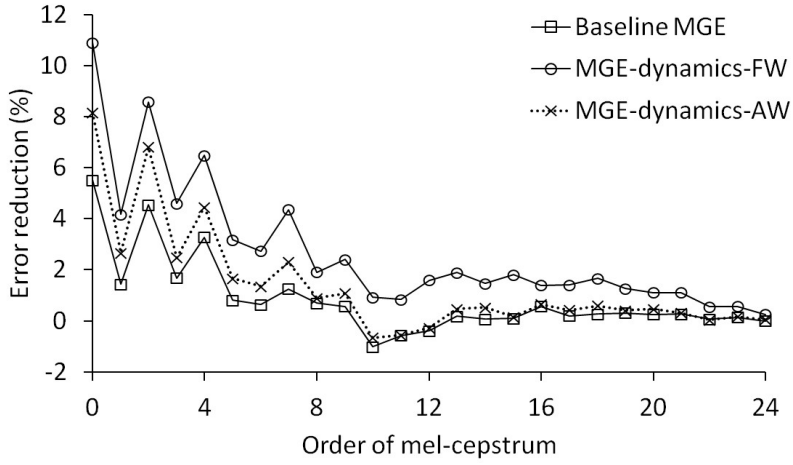
Figure 3.4: Performance of MGE-dynamics-FW with different delta weights on test data for several mel-cepstrum orders. Other orders shows similar trends but are not plotted here for readability.

To compare the performance of MGE-dynamics-AW with those of MGE-dynamics-FW and baseline MGE, the thesis performed training experiments in which the delta weight for MGE-dynamics-FW and the maximum delta weight for MGE-dynamics-AW were both set to 100. Figure 3.5 shows the relative error reduction of the static and delta features on the test data for these MGE training techniques. Compared with baseline MGE, MGE-dynamics-AW has a comparable relative error reduction of the static feature and a larger relative error reduction of the delta feature. Compared with MGE-dynamics-FW,

MGE-dynamics-AW has a larger relative error reduction of the static feature but a smaller relative error reduction of the delta feature. It can be seen that MGE-dynamics-AW alleviates the trade-off effect observed in MGE-dynamics-FW. Since both MGE-dynamics-AW and MGE-dynamics-FW obtain better performance on the delta feature than baseline MGE, it can be concluded that the MGE-dynamics criterion improves the capability of HMMs in capturing dynamic properties of speech over the baseline MGE criterion.



(a) Relative error reduction of static feature.



(b) Relative error reduction of delta feature.

Figure 3.5: Performances of three MGE training techniques on test data. The delta weight for MGE-dynamics-FW and the maximum delta weight for MGE-dynamics-AW were both set to 100.

Figure 3.6 illustrates an example of the trajectory of the 2nd mel-cepstral coefficient of natural speech included in the training data and those generated from HMMs trained by the baseline and proposed MGE training techniques. It can be seen that the generated trajectories from the proposed techniques almost always have more similar dynamics to the natural trajectory than that from the baseline MGE (the clearest improvements in the dynamics can be observed around the 45th and 305th frames in the figure).

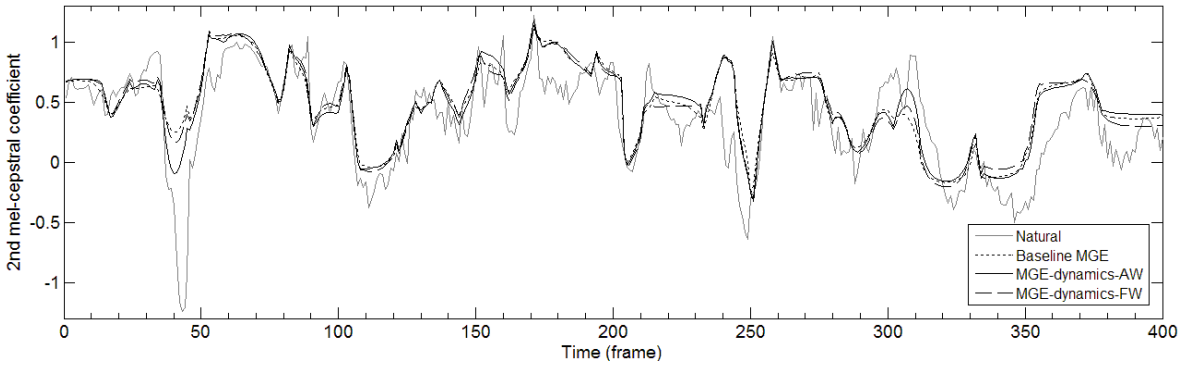


Figure 3.6: Natural and generated trajectories of the 2nd mel-cepstral coefficient for an utterance included in the training data.

3.5.2(2) Subjective evaluation

The thesis also carried out subjective listening tests to evaluate the effectiveness of the two proposed techniques. Two preference tests were conducted. In the first test, MGE-dynamics-FW was compared with baseline MGE. In the second test, MGE-dynamics-AW was compared with baseline MGE. The settings related to the delta weight for MGE-dynamics-FW and MGE-dynamics-AW were the same as those in the previous experiment. Twelve Japanese listeners participated in the tests. They were presented with pairs of synthesized speech in random order, and asked to choose which one sounded better or to give an answer of "No preference" if the stimuli sounded the same. For each listener, 20 test sentences were randomly selected from the evaluation set consisting of 53 sentences.

Table 3.1 shows the results of the two preference tests. It can be seen that the difference in preference between MGE-dynamics-FW and baseline MGE is insignificant (28.3% vs 29.6%), whereas MGE-dynamics-AW has a higher preference score than baseline MGE (32.9% vs 27.1%). Furthermore, the result of a paired one-tailed *t*-test [47] indicates that the mean preference score of MGE-dynamics-AW is statistically significantly greater than that of baseline MGE at a 5% significance level (*p*-value = 0.031). For the second listening test, informal feedback from the test subjects suggested that the

utterances synthesized using the models trained by MGE-dynamics-AW and baseline MGE had almost the same naturalness; however, the clearness of some parts of utterances belonging to the MGE-dynamics-AW category was improved. The higher preference score of MGE-dynamics-AW than baseline MGE could be interpreted as the consequence of improved objective performance of the delta feature while maintaining an objective performance of the static feature comparable with that of baseline MGE.

Table 3.1: Mean preference score (with 95% confidence interval) in evaluation with the original parameter generation algorithm.

	No preference (%)	Baseline (%)	Proposed (%)
Baseline vs dynamics-FW	42.1 ± 16.4	29.6 ± 9.7	28.3 ± 10.3
Baseline vs dynamics-AW	40.0 ± 12.6	27.1 ± 6.1	32.9 ± 7.8

Table 3.2: Mean preference score (with 95% confidence interval) in evaluation with the parameter generation algorithm considering GV.

	No preference (%)	Baseline (%)	Proposed (%)
Baseline vs dynamics-FW	58.7 ± 15.7	17.1 ± 8.0	24.2 ± 10.9
Baseline vs dynamics-AW	55.0 ± 14.1	17.9 ± 6.1	27.1 ± 9.6

3.5.3. Evaluation with the parameter generation algorithm considering GV

Two preference tests similar to those described in the preceding subsection were additionally conducted. The only difference from the previous tests is that GV was considered in the synthesis process. Table 3.2 shows the results of the preference tests with the GV technique. Although the "No preference" rates increase by around 15% compared with those in the previous tests, both of the proposed techniques have a higher preference score than the baseline MGE. The results of paired one-tailed *t*-tests indicate that while the preference of MGE-dynamics-AW over baseline MGE is again significant at a 5%

significance level (p-value = 0.012), the preference of MGE-dynamics-FW over baseline MGE is much less significant (p-value = 0.090). The visual inspection of several samples revealed that while enhancing the dynamic range of the generated parameter trajectory, the GV technique seems to keep the trajectory dynamics similar to that in the case without considering GV.

3.6. Discussions

Although a preference test between MGE-dynamics-AW and MGE-dynamics-FW was not conducted, it is necessary to point out the merit of the former over the latter. MGE-dynamics-AW provides a data-driven technique of determining the delta weight for each portion of speech, provided that the maximum delta weight is set appropriately. In contrast, one must manually tune the delta weight to obtain the best performance for MGE-dynamics-FW. Moreover, considering the fact that HMM parameters are optimized independently for each dimension owing to the use of diagonal covariance matrices in spectral parameter modeling, MGE-dynamics-AW is likely to result in improved dynamics occurring synchronously among the dimensions since the delta weight is adjusted segment-by-segment over the course of an utterance. This synchronously improved dynamic property is less likely to occur in MGE-dynamics-FW because the delta weight is kept constant over the entire speech. Whether synchronously adaptive control among the dimensions is effective for spectral dynamics representation is still unclear. More work is needed to confirm this.

This work also exhibits some limitations. First, the high-quality vocoder STRAIGHT [18] was not used since the magnitude of several speech samples synthesized by the STRAIGHT filter exceeded the data range specified for a 16-bit WAV sound file. Compared with mel-cepstral vocoding, the relative error reduction of the MGE training techniques had similar trends but slightly lower magnitudes when STRAIGHT was used. Second, the more popular five-state HMM structure was not employed in the experiments because it was found that the use of the three-state one to be sufficient to capture the degree of dynamicity of speech segments. The use of five-state phoneme HMMs resulted in over-detailed representation of the proposed degree of dynamicity of HMM-state-sized speech segments, causing an over-fitting effect when MGE-dynamics-AW training was performed with this model setting. From this viewpoint, MGE-dynamics-FW has the flexibility to work well irrespective of whether a three-state or five-state structure is used.

3.7. Conclusion

In this chapter, dynamic properties of speech parameters are incorporated into the MGE criterion by defining the generation error of dynamic features and introducing this error component into the GEF, resulting in the so-called MGE-dynamics criterion. Two methods for setting the weight associated with this newly added error component are also proposed, which are fixed weighting (FW) and adaptive weighting (AW). An objective evaluation shows that MGE-dynamics-AW obtains comparable performance for the static feature and better performance for the delta feature compared with baseline MGE training. Subjective listening tests indicate that a small but statistically significant improvement in the quality of synthesized speech was perceived in the case of MGE-dynamics-AW training. The newly derived MGE-dynamics criterion improves the capability of HMMs in capturing dynamic properties of speech while maintaining a computational complexity similar to that of the baseline MGE criterion. Future work should target the investigation of the effect of the window length used in dynamic feature calculation on MGE-dynamics training and the effect of MGE-dynamics training on F0 modeling.

Chapter 4 F0 Parameterization of Glottalized Tones in HMM-based TTS for Hanoi Vietnamese

4.1. Introduction

A couple of HMM-based TTS systems for Hanoi Vietnamese, the standard dialect of Vietnamese [48], have been introduced recently [49, 50]. Latest refinements being made to these systems involved in the integration of syntactic information and intonational tags to improve the overall naturalness of generated prosody [51-53]. Although the obtained results are promising, none of the above improvements are directly related to the perception of tones, the most dominant factor in the prosody of a tonal language like Vietnamese.

Vietnamese has a complex tone system where each syllabic tone can be represented by an F0 contour in its isolated mode. The realization of the F0 contour of a Vietnamese utterance is mainly the result of the interaction between the contributing tones, called tone coarticulation [54]. Not only for making the utterance's intonation, each tone has an essential role in the meaning of the associated syllable. Accurate modeling and synthesis of the tones, thus the F0 contour of a speech signal, is crucial for an HMM-based Vietnamese TTS system.

A distinctive feature of Hanoi Vietnamese that makes the above task more challenging is that its tones are realized by a combination of pitch and voice quality (phonation type) [55, 56]. Specifically, 3 out of 8 tones are often produced with some degree of glottalization, called glottalized tones. In one work [56], "glottalization" is used as the cover term for glottal constriction and laryngealization (or creaky voice/vocal fry [13], [57]). This thesis uses "glottalization" to cover any non-modal phonation [58] that occurs during tone production and leads to speech waveform with irregular pitch periods, often accompanied by very low F0 values or voicelessness. Note that this definition of glottalization excludes glottalized events not related to the tone production [59].

In HMM-based TTS, speech parameters, including F0, have to be extracted from recorded waveforms prior to HMM modeling and generation. Modern F0 extractors often assign erroneous F0 values or fail to detect any F0 (or periodicity) in glottalized waveform [9, 57, 60]. As a result, subsequent F0 modeling and generation suffer. Standard HMM-based TTS uses multi-space distribution (MSD) to model and generate discontinuous F0 trajectories [61] as mentioned in Section 2.3.2. Faulty voicing decisions resulting from the F0 extraction phase will cause the deteriorately trained MSD-HMMs to synthesize voiced frames as unvoiced, resulting in hoarse speech, or to synthesize unvoiced frames as voiced, resulting in buzzy speech [12]. When listening to the output of the baseline HMM-based Vietnamese TTS system, it can be perceived that although the synthetic speech is highly intelligible, its overall quality is greatly degraded by the hoarseness frequently occurring at syllables bearing a glottalized tone. This is due to voiced (or F0) missing errors (i.e., voiced frames classified as unvoiced) caused by F0 trackers when dealing with glottalized waveform.

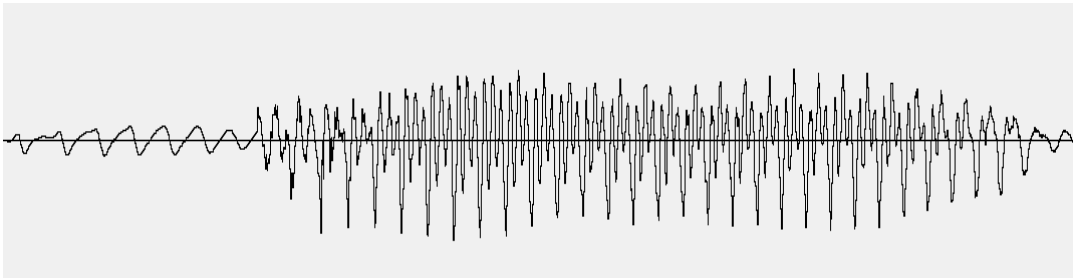
To alleviate the perceived hoarseness, popular solutions include the use of continuous F0 contours either for modeling [9-11], or for synthesis [12]. This study directly tackles the problem of F0 estimation in glottalized regions of a speech signal. The research of Ishi et al. [13] suggests that a cross-correlation-based measure could help for the detection of similar successive glottal pulses in glottalized speech segments, where the waveforms often exhibit rather weak periodicity. Based upon this work, this thesis proposes a pitch marking algorithm, where the pitch marks [62] are propagated from regularly spaced pitch periods to irregularly spaced ones, from which the refined F0 contour of a glottalized tone is derived. This F0 parameterization method can be expected to reduce the hoarseness whilst improving the tone naturalness of synthetic speech. While the pitch marking procedure works as a refinement step based on the results of an F0 extractor, it is independent from the F0 extraction step. Therefore it can be combined with any F0 extractor.

This chapter is organized as follows. Section 4.2 gives an overview of Vietnamese glottalized tones. Section 0 explains the problems with F0 extraction for the glottalized tones. The proposed F0 parameterization method is described in Section 4.4, and the results of the objective and perceptual evaluations are reported in Section 4.5. Section 4.6 presents some discussions, and Section 4.7 concludes the chapter.

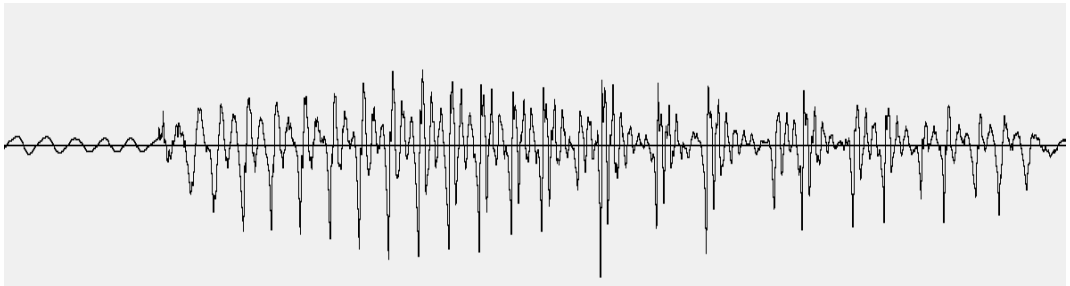
4.2. Vietnamese glottalized tones

Vietnamese is a tonal language where each syllable carries a tone. The tone contributes not only to the syllable's meaning but also to the utterance's intonation. In terms of acoustic

feature, a tone is represented by a contiguous F0 contour covering the whole syllable. A full description of the tone system of Hanoi Vietnamese can be found in several works such as [56] (pp. 120–123) or [48] (pp. 385–388). Among the total of 8 tones, there are 3 tones often accompanied by glottalization, called glottalized tones. In one study [56], glottalization is described as those abnormal activities of the glottis that lead to either "a tense gesture of adduction of the vocal folds that extends over the whole of a syllable rhyme" (referred to as glottal constriction or glottal interrupt) or "irregular vocal fold vibration" (referred to as laryngealization or creaky voice). According to the source-filter model of speech production, these irregular activities of the glottal source result in abnormal speech waveforms, causing troubles to F0 estimators. Figure 4.1 shows an example of waveforms of a syllable when produced with a non-glottalized tone and with a glottalized one. While the former exhibits regular periodicity, the latter indicates irregular periodicity in the second half due to the glottalization. Table 4.1 summarizes some characteristics of the glottalized tones. This thesis uses the same tone identifiers (IDs) as in the research of Vu et al. [49] for brevity.



(a) Waveform with regular periodicity.



(b) Waveform with irregular periodicity in the second half due to glottalization.

Figure 4.1: Waveforms of the syllable /dȁ/ exhibit different periodicity characteristics when accompanied by (a) a non-glottalized tone, and (b) a glottalized tone.

Table 4.1: Characteristics of the glottalized tones.

Tone ID in [49]	Tone ID in [56], [48]	F0 contour	Position of glottalization
3	C2	Falling-Rising	Middle of syllable
4	C1	Falling	End of syllable
6	B2	Falling	End of syllable

Figure 4.2 shows the occurring frequency of 8 tones in the Vietnamese speech database used in this research. The database consists of 1107 recorded sentences. On average, there are 2.7 glottalized tones among the total 11.4 syllables per sentence.

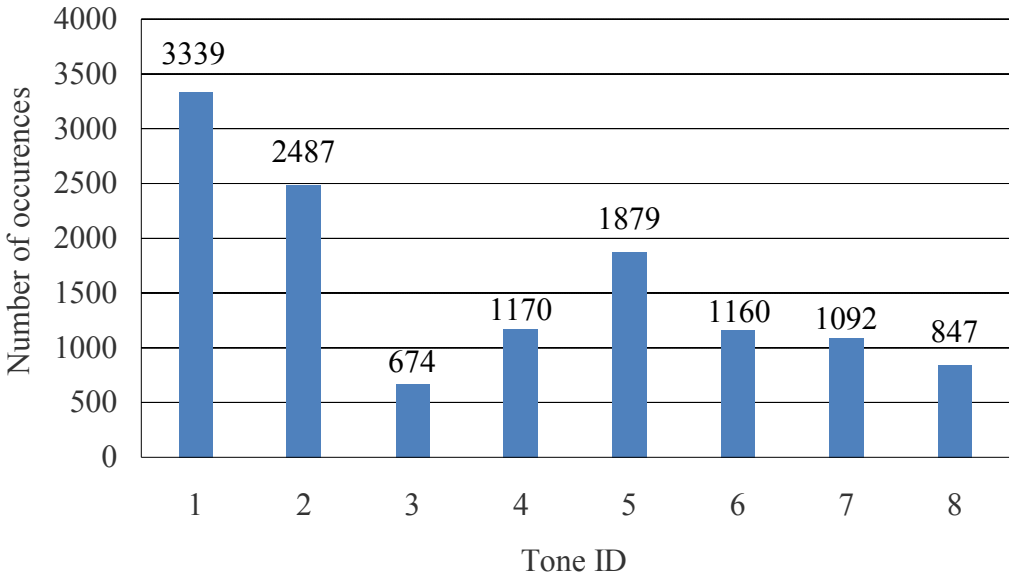


Figure 4.2: Number of occurrences of each tone in the database used in this research.

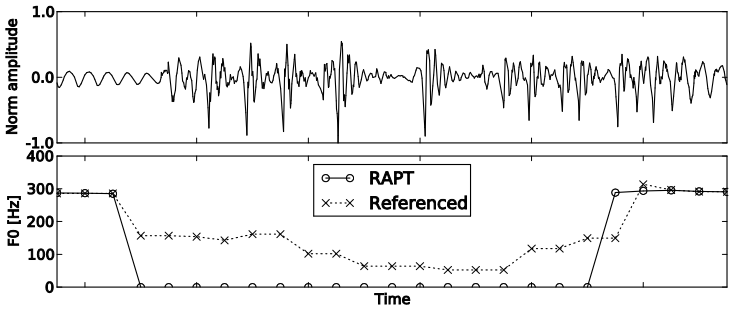
4.3. Problems with F0 extraction for glottalized tones

In an irregular phonation like glottalization, even sounds regarded as voiced (e.g., vowels and nasals) may show no obvious regularity in speech waveform. This makes the task of obtaining a useful estimate of F0 during glottalization rather difficult. The current study chooses the robust F0 tracker RAPT [63], which is included in the standard HTS toolkit

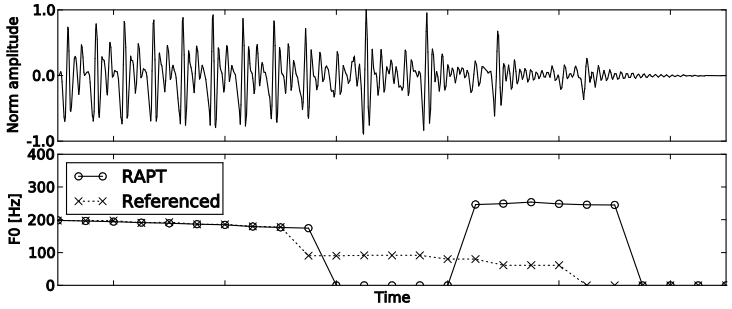
[35] and was reported to obtain a fairly good performance in a voicing classification task for creaky voice [60], as the conventional F0 extractor for comparison with the proposed one.

4.3.1. Popular F0 extraction errors of RAPT

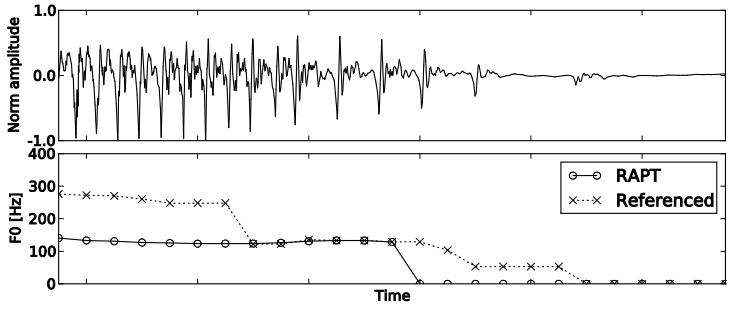
For the Hanoian female voice used in this study, it is observed that F0 extraction errors regularly occur when RAPT encounters syllables carrying a glottalized tone (hereinafter referred to as glottalized syllables). Figure 4.3 shows examples of waveform segments of voiced phones in glottalized syllables and F0 extraction results of RAPT together with referenced F0 data. As can be seen, the sudden increase of pitch period (or the sharp drop of F0), typically from double to triple, during glottalization is troublesome for an F0 detector like RAPT. In such cases, RAPT often fails to detect any F0 even if the lowest F0 for detection is set to a very low value, resulting in voiced missing errors (VMEs). Furthermore, gross F0 errors may appear when a glottalized waveform exhibits spurious periodicity or a period-doubled/diplophonic phonation [13], [58] happens (which causes pitch halving/doubling errors). These errors occur quite often with glottalized syllables, which are prevalent in the training corpus, thus causing considerable effects on the performance of the baseline HMM-based Vietnamese TTS system. Although voiced insertion errors (VIEs) (i.e., unvoiced frames classified as voiced) also appear (for example, in the last F0 samples in Figure 4.3b), its effect is negligible compared to VMEs in both objective evaluation (Section 4.5.4) and perceptual tests [12].



(a) VMEs in the middle of Tone 3, syllable /ɗa/.



(b) VMEs and gross F0 errors at the end of Tone 4, syllable /bɔ/.



(c) VMEs at the 2nd half and pitch halving errors at the 1st quarter of Tone 6, syllable /ʈɔŋ/.

Figure 4.3: Typical errors occur when the F0 extractor RAPT copes with glottalized syllables. Only the final segment of 120 ms of a syllable is shown in each top plot. F0 values were extracted at every 5 ms. F0 range for the extraction was set to 40–400 Hz. An F0 of zero means that the associated speech frame is considered as unvoiced. Referenced F0 contours were produced by a method described in Section 4.5.2.

4.3.2. Effects of F0 extraction errors on MSD-HMM modeling and generation

In the MSD-HMM training, gross F0 errors, including pitch halving/doubling ones, decrease the accuracy of continuous Gaussian distributions estimated from erroneous voiced F0 observations. Meanwhile, VMEs not only deteriorate the estimation of continuous Gaussian distributions due to the lack of meaningful voiced F0 observations, but also ruin the estimation of MSD-weights, which represent the probability of being voiced or unvoiced of HMM states (see Section 2.3.2). Specifically, the MSD of a state in a context-dependent phoneme HMM is estimated from F0 observations belonging to a leaf node of the decision tree resulted from context clustering process. Due to VMEs, there is high probability that leaf-node state models matched with a combinatorial context such as "current phone is a vowel and current syllable bears Tone 3" and the like are estimated with more unvoiced F0 observations than voiced ones. Consequently, the MSD-weights of these state models will be biased toward the unvoiced component, which is unexpected.

In the MSD-HMM synthesis, while the weakly trained continuous Gaussian distributions contribute to the generation of distorted F0 contours for glottalized tones, the biasedly estimated MSD-weights cause the system to generate voiced phones (e.g., vowel) in glottalized syllables with unvoiced excitation (i.e., noise source). In terms of perception, the former leads to tone distortion, whilst the latter results in hoarseness (i.e., noise).

4.4. Proposed F0 parameterization of glottalized tones

As directly estimating F0 from irregularly periodic waveforms due to glottalization is often problematic for a conventional F0 extractor, this section proposes a pitch marking algorithm to indirectly derive F0 estimates from the pitch marks. Pitch marks are the locations marking signal periods in a voiced speech segment, hence the distance between them inversely correlates with F0 values of the segment. The detection of pitch marks has drawn much attention in concatenative TTS using TD-PSOLA because its accuracy has direct influence on the quality of prosodic modification at waveform level [62]. In this paper, the pitch marks are means of deriving F0 parameters for statistical modeling, thus the precision of their locations is of lesser importance, making simpler marking algorithm be of potential use.

The principle of the marking algorithm is to propagate pitch marks period-by-period from region with regularly spaced pitch periods (called regular region) to region with irregularly spaced ones (called glottalized region). Normally, a glottalized syllable's waveform consists of both regular and glottalized regions, and their relative positions

depend on where the glottalization occurs in the corresponding glottalized tone's production. Figure 4.4 shows the flowchart of the proposed method, which is only applied to the glottalized syllables in an utterance. The proposed method can be seen as an F0 refinement scheme, thus can be combined with any F0 extractor. Detailed description of every processing stage is given in the following subsections.

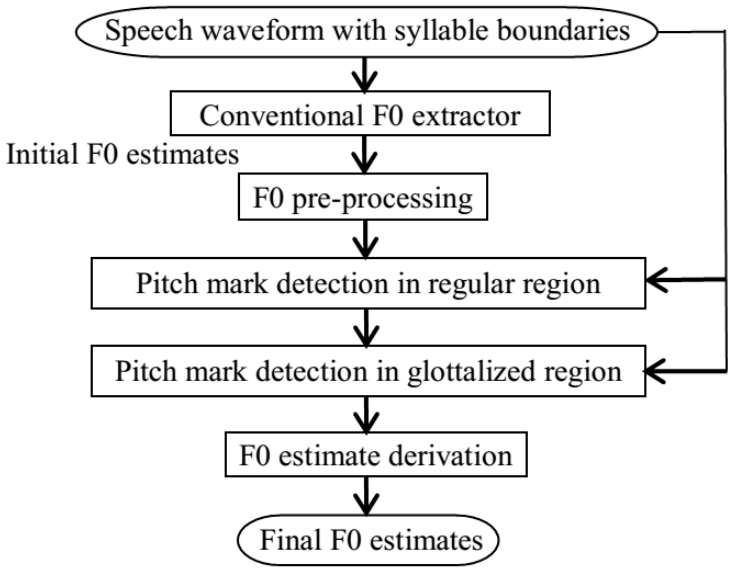


Figure 4.4: Proposed F0 parameterization method for a glottalized syllable.

4.4.1. Pre-processing of F0s

The objective of this stage is to fix some of the gross F0 errors caused by glottalization. Specifically, voiced segments with too short duration (e.g., less than 20 ms) are removed. Besides, a voiced segment having an average F0 larger than double (or smaller than half) of those of the preceding and succeeding voiced segments is considered as a pitch doubling (or halving) error. F0 values of a voiced segment like this will be replaced with the ones obtained by linearly interpolating F0 values of the two neighboring voiced segments.

4.4.2. Detection of pitch marks in regular region

Given the above F0 estimates, corresponding pitch marks can be determined by any of available pitch marking algorithms. Current work used a simple one presented in [64] to propagate the pitch marks from the middle to the both ends of each voiced isle on a pitch-synchronous basis as follows. For a voiced isle, the 0^{th} pitch mark locates at the time instant t_0 that has the minimum waveform amplitude in the interval $[t_{mid} - T_0/2, t_{mid} + T_0/2]$,

where t_{mid} is the middle instant of the voiced isle, and T_0 is the pitch period at t_{mid}, computed as the reciprocal of the nearest F0 estimate around t_{mid}. Then the i^{th} pitch mark ($i>0$) toward the left boundary of the voiced isle can be recursively determined as the instant t_i having the minimum amplitude in the interval $[t_{i-1} - 1.2T_i, t_{i-1} - 0.8T_i]$, where t_{i-1} is the location of the $(i-1)^{th}$ pitch mark, and T_i is the pitch period at t_{i-1}, computed as the reciprocal of the nearest F0 estimate around t_{i-1}. A similar process is also carried out toward the right boundary of the voiced isle. Two right-most marks and two left-most marks of each voiced isle will be used as anchor ones to detect potential marks in glottalized region. Figure 4.5 illustrates pitch marks detected in two regular regions (vertical dashed lines) for a syllable with Tone 3.

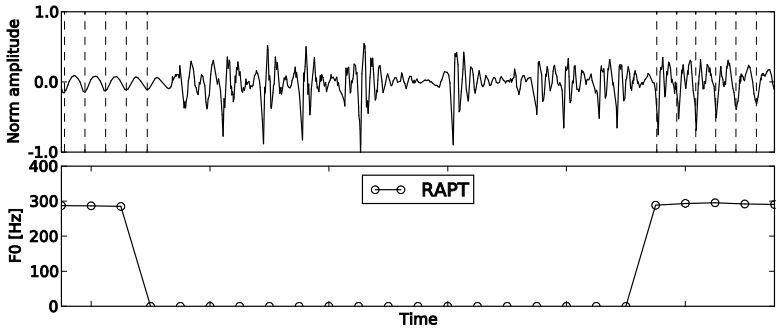


Figure 4.5: Pitch marks in regular regions of speech (vertical dashed lines in top plot) were detected from F0 estimates extracted by RAPT (bottom plot). The same example in Figure 4.3a was used.

4.4.3. Detection of pitch marks in glottalized region

For those glottalized syllables having undefined F0s between the voiced isles or at the syllable's ending part, which are likely due to glottalized waveform, a pitch mark searching process need to be carried out to infer these F0 estimates. Here, prior knowledge about the glottalized tones is used to specify the search direction, given that the anchor pitch marks detected in regular regions are available. Tone 4 and Tone 6 have glottalization at the end, thus the search is only performed forward, starting from the right-most anchor marks of a voiced isle toward the syllable's end; however, Tone 3 normally has glottalization in the middle, thus the search is performed both forward, starting from the right-most anchor marks of a voiced isle toward the syllable's end, and backward, starting from the left-most ones of a voiced isle toward the syllable's start. Since the procedures for searching forward and backward are identical, only the searching forward algorithm is presented next. For undefined F0 regions in syllables bearing Tone 3 where there exists both forward and backward pitch marks, an additional procedure need to be

carried out to combine these marks into a unique mark sequence, described in Section 4.4.3(2).

4.4.3(1) Searching forward for pitch marks

The searching forward problem can be formulated as follows: given the current pitch mark at time t and its pitch period P (computed as the time distance between the current mark and its preceding one), find the next pitch mark (if any) at time $t' = t + P + S$, where $P + S$ and S are the pitch period and period shift of the next mark, respectively. The search range for S can be found based on the variation range of consecutive pitch periods in glottalized region. Empirically, the next pitch period may range from half to triple of the current one. Let *MinPR* and *MaxPR*, respectively, the minimum and maximum ratios of the next period to the current one. While *MinPR* can be reasonably set to a value around 0.5 to account for sudden period contractions, *MaxPR* represents the degree to which the next period expands compared to the current one, thus reflecting the degree of glottalization of a particular speaker. Here, *MinPR* was set to 0.4, whereas *MaxPR* was tuned on a development set, detailed in Section 4.5.2. Then the search range for S can be inferred as [*(MinPR – 1)P*, *(MaxPR – 1)P*].

In glottalized regions, the signal still exhibits its periodicity, though much weaker than regular regions. Pitch marks are expected to locate the signal periods, thus the signals around two consecutive marks should show some similarity. Hence, the strength of a next mark candidate can be measured by the cross-correlation coefficient between the signal around the current mark and the one around the next mark candidate as

$$XCorr\left(X_t, Y_{t'}\right) = \frac{L\sum_{i=-\frac{L}{2}}^{\frac{L}{2}-1} x_i y_i - \sum_{i=-\frac{L}{2}}^{\frac{L}{2}-1} x_i \sum_{i=-\frac{L}{2}}^{\frac{L}{2}-1} y_i}{\sqrt{L\sum_{i=-\frac{L}{2}}^{\frac{L}{2}-1} x_i^2 - \left(\sum_{i=-\frac{L}{2}}^{\frac{L}{2}-1} x_i\right)^2}\sqrt{L\sum_{i=-\frac{L}{2}}^{\frac{L}{2}-1} y_i^2 - \left(\sum_{i=-\frac{L}{2}}^{\frac{L}{2}-1} y_i\right)^2}}, \tag{4.1}$$

where X_t and $Y_{t'}$ are the windowed signals centered at the current mark t and the next mark candidate t', respectively, and L is the window length for the cross-correlation calculation. Specifically, L was set to the current pitch period P, yet limited to 15 ms in order to avoid the effect of irregularly spaced glottal pulses in the strength measure [13], and

$$X_t = \left\{x_i, -L/2 \le i < L/2 \middle| x_i = a\left(i + t\right)\right\}, \tag{4.2}$$

$$Y_{t'} = \left\{y_i, -L/2 \le i < L/2 \middle| y_i = a\left(i + t'\right)\right\}, \tag{4.3}$$

where *a(i)* is the signal amplitude at the i^{th} sample. Then the strongest pitch mark candidate can be inferred from the optimal period shift, which is defined as

$$S_{opt} = \underset{S \in [(MinPR-1)P,(MaxPR-1)P]}{\arg\max} XCorr\left(X_t, Y_{t'}\right). \tag{4.4}$$

Figure 4.6 depicts an example of pitch mark searching forward. The top plot shows previously defined and current marks in vertical solid lines and a candidate for the next mark (when the period shift is equal to zero) in vertical dash line. The bottom plot shows the cross-correlation variation over the search range for the period shift. The optimal shift was found at 5.5 ms, which means that the strongest mark candidate was on the right of the above candidate at a time distance of 5.5 ms.

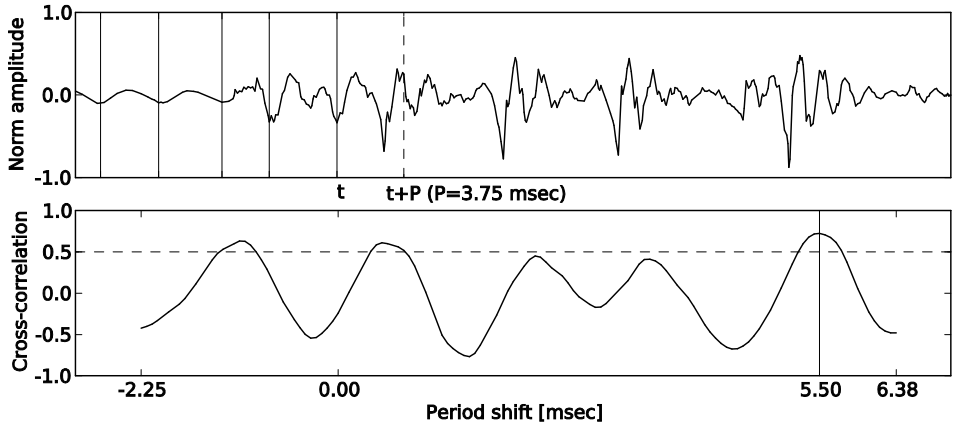


Figure 4.6: Example of pitch mark searching forward (with *MaxPR* = 2.7).

Finally, the strongest candidate is accepted as the next pitch mark only when its cross-correlation coefficient is larger than a pre-defined threshold. Setting this threshold is important. A threshold that is too low (or high) will leads to an increase of pitch mark insertion (or missing) errors and thus VIEs (or VMEs). The cross-correlation threshold (*CorrThs*) was also tuned on the same development set used for tuning *MaxPR*.

The above pitch mark searching process is carried out repeatedly until an anchor mark or a boundary of the syllable is encountered. It may also stop earlier if the cross-correlation coefficient of the strongest mark candidate is below the preset *CorrThs*, which means that the waveform around there is too irregular to proceed the pitch mark propagation. Figure 4.7 illustrates both of the above stopping conditions, where the forward search was stopped early and the backward search was stopped when an anchor mark was met. The searching forward algorithm can be summarized as follows:

Step 1: Initialization

ForwardMarks = empty list

t = right-most mark of the preceding voiced isle

P = pitch period associated with the right-most mark

Step 2: Iteration

Repeat

For S in range [$(MinPR - 1)P$, $(MaxPR - 1)P$] {

$t' = t + P + S$

compute $XCorr(X_t, Y_{t'})$ based on Eqs. (4.1)–(4.3)

}

determine S_{opt} using Eq. (4.4)

$t' = t + P + S_{opt}$

$XCorr_{opt} = XCorr(X_t, Y_{t'})$

append t' into *ForwardMarks*

$t = t'$

$P = P + S_{opt}$

Until $XCorr_{opt} < CorrThs$ OR t exceeds the left-most mark of the following voiced isle or the syllable's end.

Step 3: Termination

Withdraw the lastly appended mark from *ForwardMarks* and stop.

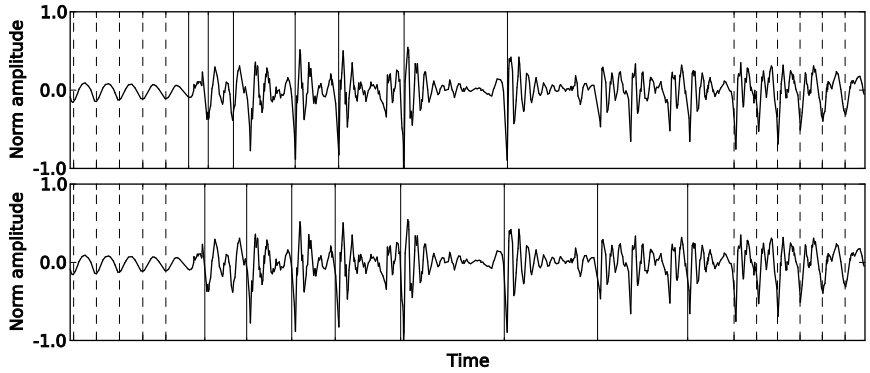


Figure 4.7: Results of pitch mark searching in glottalized region (with *MaxPR* = 2.7, *CorrThs* = 0.4). Forward (top) and backward (bottom) results are shown in vertical solid lines. Vertical dashed lines are the marks detected in regular regions. The example in Figure 4.3a was used.

4.4.3(2) Combination of forward and backward results

A consistent pitch mark arrangement can be obtained from the forward and backward marks based on their comparative reliability. It is noticed that the marks were sequentially propagated, thus an earlier founded mark is considered more reliable than lately founded ones. Hence, N_L left-most forward and N_R right-most backward marks are to be combined into a single mark sequence. To define N_L and N_R so that the combination process can be

done with ease, the centering point of the most coincident forward-and-backward mark pair is chosen to make a common split on the two mark sequences. Then N_L and N_R are set to the number of marks in the left-sided split forward sequence and in the right-sided split backward sequence, respectively. In the combined mark sequence, if the interval d between the most coincident forward-and-backward mark pair is too small compared to the neighboring pitch periods p_L and p_R (i.e., violating the constraint set by *MinPR*), this pair will be merged into one mark located at its centering point. Figure 4.8 illustrates the split-combine-merge process for a self-designed example, where $N_L = 2$ and $N_R = 4$.

The top plot in Figure 4.9 displays the results of the forward and backward mark combination process for the example in Figure 4.7 (here, $N_L = 7$ and $N_R = 3$). As can be seen, the final pitch mark arrangement is quite consistent thanks to the reuse of mark orders available in the forward and backward results.

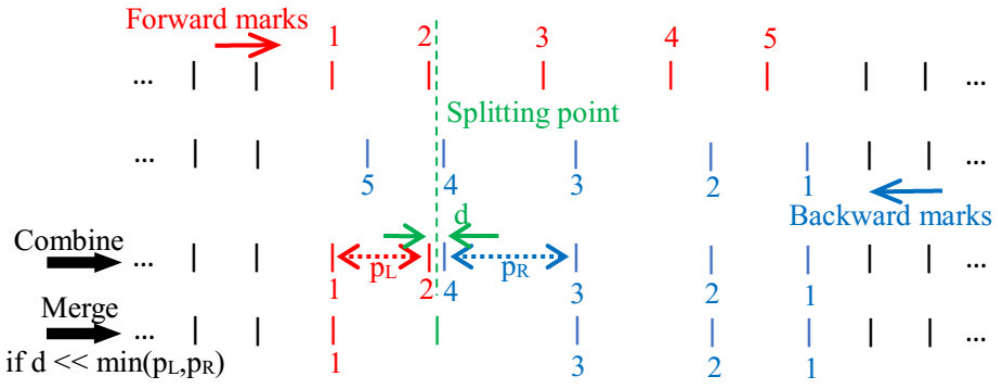


Figure 4.8: Illustration of the split-combine-merge process. Pitch marks are denoted by vertical solid lines with different colors depending on their types: anchor (black), forward (red), backward (blue), merged (green). Numbers indicate mark orders in forward and backward results.

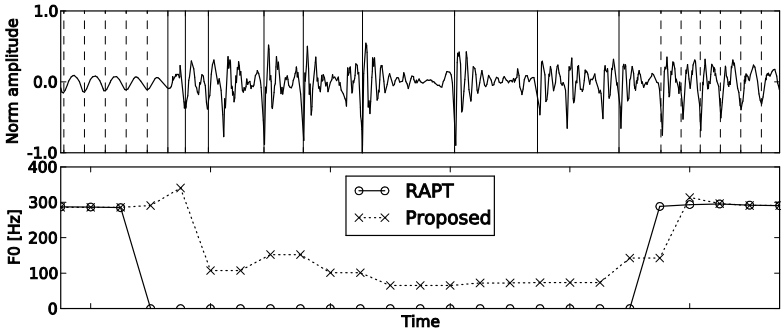


Figure 4.9: Results of the forward and backward pitch mark combination for the example in Figure 4.7 (top plot) and F0 contours estimated by RAPT and the proposed method (bottom plot). The same settings of Figure 4.3a were used in both plots.

4.4.4. Derivation of F0 estimates from pitch marks

Any F0 instant between two consecutive marks is assigned a value equal to the reciprocal of the time distance between them. However, exception will be made to those marks that are too far away each other to take into account abnormally spaced glottal pulses and possible mark missing errors. As observed from glottalized waveforms of the recorded voice, segments containing interpulse intervals longer than 25 ms usually exhibit almost no periodicity due to the abnormal pulse shape, often associated with only one or two pulses. Thus 25-ms was chosen as the maximum period between two pitch marks for deriving voiced F0 values (i.e., the minimum detectable F0 is 40 Hz). The bottom plot of Figure 4.9 shows the F0 contour estimated from the pitch marks detected by the proposed method for the same example as in Figure 4.3a. Although the F0 contour obtained by the proposed method is deviated from the referenced one in some parts, the global shape Falling-Rising of a Tone 3 is preserved. Compared to the F0 contour extracted by RAPT, the proposed one is a much more complete and accurate representation of the tone.

4.5. Experimental evaluations

This section evaluates the proposed F0 parameterization method in combination with a version of the F0 extractor RAPT implemented in the Snack sound library, which is used in the HTS toolkit version 2.1.1 [35]. Two HMM-based synthesis systems were built as follows:

- Conventional system used the original F0 estimates extracted by RAPT for training F0 models. The extraction range was set to 40–400 Hz.
- Proposed system used the F0 estimates resulting from the refinement of RAPT's extraction results with the proposed method for training F0 models. Only the F0 estimates of glottalized syllables were refined.

4.5.1. Common system setups

A phonetically balanced Vietnamese corpus, uttered by a Hanoian female speaker, was used for the experiments. The corpus consists of 1007 training, 50 development, and 50 test sentences, with the total length of about 50 minutes of speech. Speech data was manually labeled at phoneme level. On average, there are 2.7 glottalized tones among the total 11.4 syllables per sentence. To prepare the context-dependent labels, similar phonetic and linguistic contexts of the English system [2] were used, excluding those related to stress, accent and intonational tag. Instead, information on tone identity of the current,

preceding and succeeding two syllables was added to capture the rich tonal features of Vietnamese.

Speech signals were sampled at 16 kHz and windowed by a 25-ms Hamming window with a 5-ms shift. Spectral parameters were the 0^{th} through 24^{th} mel-cepstral coefficients obtained by a mel-cepstral analysis technique [15]. Excitation parameter was the logarithm of F0. Each of the spectral and F0 parameter vectors included the static feature and their delta and delta-delta features.

A five-state left-to-right no-skip HMM structure was used. Each state output distribution was composed of spectrum and F0 streams. The spectrum stream was modeled by single Gaussian distributions with diagonal covariance matrices. The F0 streams were modeled by MSDs. Context-dependent HMMs for each of the spectral and F0 parts were constructed with a decision tree based context clustering technique based on the minimum description length (MDL) criterion [32].

Given contextual labels of an input sentence, mel-cepstral coefficients were generated from a sequence of Gaussian distributions of the context-dependent HMMs using the original parameter generation algorithm, and emphasized with post-filtering [33]. F0 parameter sequences were generated from sequences of contiguous voiced states, determined based on the MSD-weights of the MSD-HMMs. The generated F0 parameters were used to drive a simple pulse train/white noise excitation. Then the MLSA filter [16] was used to synthesize a speech waveform based on the generated mel-cepstra.

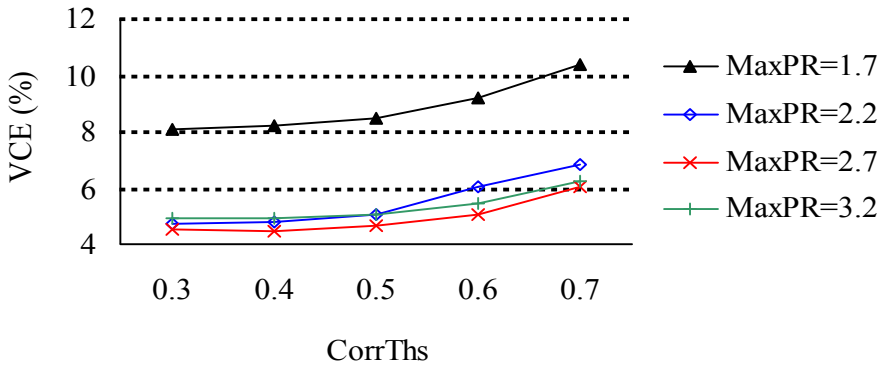
4.5.2. Parameter tuning for the proposed method

The development set was used to find the optimal values of the *MaxPR* and *CorrThs* parameters for the proposed F0 refinement method. Objective measures were the voicing classification error (VCE), including VME and VIE, and the root mean square error (RMSE) of the refined F0 contours. For comparison with the initial F0 contours extracted by RAPT, the RMSE was only calculated for frames that were simultaneously voiced in the initial, refined, and referenced F0 contours. For parameter tuning, only the glottalized syllables in a sentence were of interest. The referenced F0 data for the glottalized syllables were created as follows:

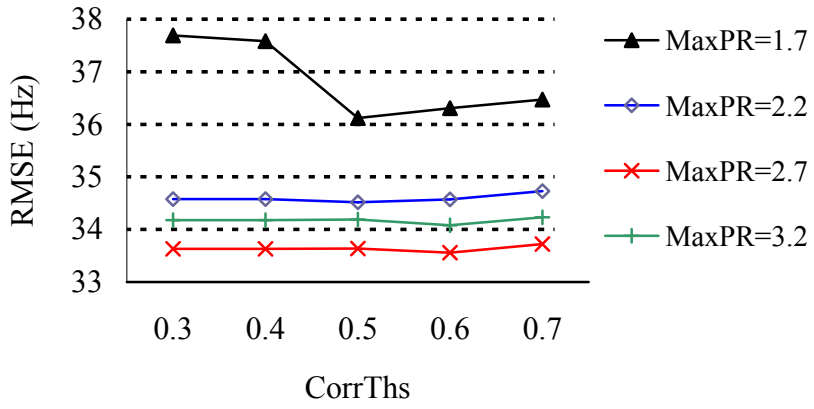
- Firstly, pitch marks were automatically detected from RAPT's F0 estimates by using the proposed pitch marking method with some sub-optimal settings (e.g., *MaxPR* = 2.2, *CorrThs* = 0.5).
- Then, these pitch marks were manually corrected based on the speech waveform.

- The corrected pitch marks then were used to automatically derive F0 estimates as described in Section 4.4.4.
- Finally, these F0 values were manually corrected by inspecting the speech waveform in order to resolve unexpected situations (e.g., unvoiced speech may appear between two pitch marks even if their time distance is less than 25 ms).

Figure 4.10 shows the performance of the proposed F0 refinement method when varying *MaxPR* and *CorrThs*. *MaxPR* showed consistent trends on the two measures with the turning point at 2.7. As *MaxPR* increased from 1.7 to 2.7, the VCE (dominant by the VME) and the RMSE decreased. However, these trends were reversed when *MaxPR* was further increased. Hence the optimal value for *MaxPR* was fixed at 2.7. Then the optimal value for *CorrThs* was set to 0.4 because it resulted in the lowest VCE (*CorrThs* has almost no effect on the RMSE while *MaxPR* equals 2.7). Table 4.2 compares RAPT to the proposed method with the optimal settings. The proposed method remarkably reduces the VME, and slightly improves the VIE and the RMSE.



(a) Voicing classification error rate.



(b) Root mean square error.

Figure 4.10: Performance of the proposed F0 refinement method on the development set when varying the *MaxPR* and *CorrThs* parameters.

Table 4.2: Performances of RAPT and the proposed F0 refinement method on glottalized syllables in the development set.

Method	VME (%)	VIE (%)	RMSE (Hz)
RAPT	12.76	2.47	38.70
Proposed	3.02	1.46	33.63

4.5.3. Remark on the resulting F0 model sizes

As foreseeable, the proposed system was trained with (2.8%) more voiced frames than the conventional system due to the improved VME of the F0 refinement method. However, the former had unexpectedly much smaller (19.8%) F0 model size (i.e., number of clustered states in the resulting F0 models) compared to the latter, though the tree growing processes of both systems were controlled with the same standard MDL factor of 1.0. If a voiced frame was regarded as a unit of training data for estimating the F0 models as usual, it would mean that more training data leads to less complex model in this particular case.

The author tries to explain this contradiction by considering the fact that F0 is a supra-segmental feature, thus a unit of training data for F0 modeling should be longer than a frame. To confirm this, the author made a rough hypothesis that a voiced isle is a unit of F0 data, and examined the correlation between amount of training data in terms of the hypothesized unit and resulting F0 model complexity among the systems.

In addition to the conventional and proposed systems, the author created a dummy system, whose size of training F0 data lies somewhere in between those of the previously built systems. It is noted that (i) in the conventional system, an F0 contour representing Tone 3 is more likely to be broken into multiple voiced isles than Tone 4 and Tone 6 (ideally, there should be only one voiced isle for a tone) due to their glottalization positions (Table 4.1), and (ii) the proposed F0 refinement technique works as a signal-based F0 interpolator for all three glottalized tones (see Figure 4.9 for an example of Tone 3). Consequently, a simple way to build F0 data of the dummy system that meets the above data-size requirement is to reuse that of the conventional system and linearly interpolate over undefined F0 regions separating those voiced isles belonging to a single Tone 3. By using this trick, the dummy system would have the number of voiced isles smaller than the conventional system but larger than the proposed one because Tone 4 and Tone 6 are not interpolated. In terms of the number of voiced frames, a reverse trend would be observed among the systems.

Table 4.3 shows some statistics of the three systems. The resulting number of clustered F0 states of the dummy system also lies somewhere in between those of the other two systems, which is an expected result. Of the three systems, the F0 model size is positively correlated with the number of voiced isles, yet negatively correlated with the number of voiced frames (correlation coefficients are 0.85 and -0.97, respectively) used in the training F0 data. These results resolve, to some extend, the above contradiction on the reverse correlation between training data size and model complexity when a voiced frame

was taken as a data unit in F0 modeling, and support the notion that a unit of F0 data should span longer than a frame.

Table 4.3: Number of voiced units in training data and of clustered F0 states for systems trained with different F0 parameterization methods. The same MDL factor of 1.0 was used for the three systems.

System	# voiced frames	# voiced isles	# clustered F0 states
Conventional	357810	8434	3060
Dummy	361551	8071	2972
Proposed	367660	7865	2455

4.5.4. Objective evaluations

The test set was used to compare the performances of the two synthesis systems. The same objective measures as those in Section 4.5.2 were used for the evaluations. Ideal state durations were firstly obtained by forced-aligning natural speech of the test sentences. F0 contours were then generated from these same state alignments for the conventional and proposed systems. Both glottalized and non-glottalized tones were of consideration. Referenced F0 data for the glottalized tones were created as described in Section 4.5.2, whereas those for the non-glottalized tones were the original F0 estimates extracted from natural speech by RAPT without any correction.

Table 4.4 shows the performances of the conventional and proposed systems on the test set. As for the non-glottalized tones, the two systems obtain similar results, which means that the proposed F0 parameterization method for the glottalized tones has almost no effect on the synthesis performance of the non-glottalized tones although all the tones were jointly trained. As for the glottalized tones, the proposed system yields a significant reduction in the VME and a slight decrease in the RMSE compared to the conventional system. These are the results of the improvements made by the proposed F0 refinement method reported in Table 4.2. However, the proposed system has a slightly higher VIE than the conventional system partly because the F0 refinement method is not able to detect unvoiced speech between closed-distant (less than 25 ms) pitch marks as noted earlier.

Figure 4.11 shows an example of F0 trajectories generated by the two systems compared to the referenced one for a phrase consisting of three glottalized tones. Compared to the conventional system, the proposed system is capable of synthesizing the

F0 contours representing the tones that are not only much more complete (due to the significantly lower VME) but also more closely matched to the referenced ones (due to the lower RMSE). Consequently, the proposed system reproduces the tone shape better than the conventional system. By listening to the two outputs, the author perceived that the proposed system well generates the tones with clear speech, whilst the conventional system produces heavily hoarse speech with distorted tones. The effect of VIE was not perceptible.

Table 4.4: Performances of the conventional and proposed systems on the test set.

Tone type	System	VME (%)	VIE (%)	RMSE (Hz)
Glottalized	Conventional	14.51	2.44	44.00
	Proposed	**6.81**	3.39	**40.75**
Non-glottalized	Conventional	3.09	1.66	19.25
	Proposed	3.22	1.78	20.07

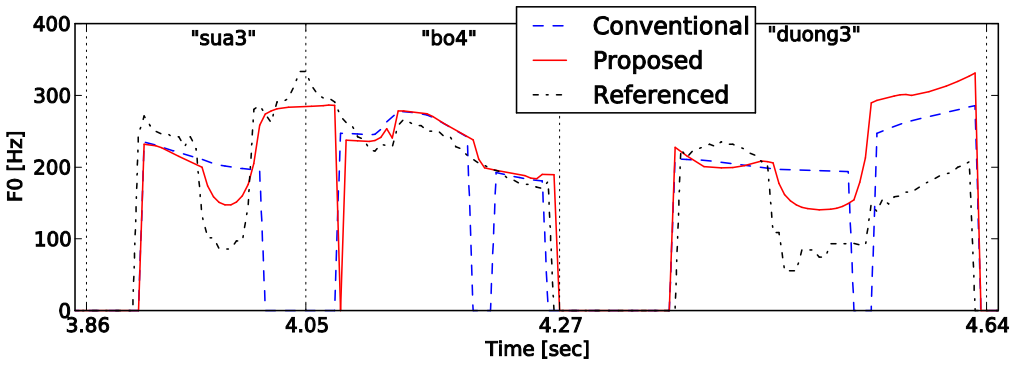


Figure 4.11: Example F0 trajectories generated by two systems compared to the referenced one. Vertical dotted lines show syllable boundaries. The phrase /sɯə3 ɓo4 zɯəŋ3/ was taken from a sentence in test set.

4.5.5. Perceptual evaluations

Two paired preference tests were conducted to evaluate the perceptual effectiveness of the proposed F0 refinement method on HMM synthesis. In the first test, speech was synthesized from ideal spectral parameters (extracted from natural speech) and state durations (obtained from forced alignment of natural speech). This is to test the effect of

generated F0 contours only. In the second test, speech was synthesized from generated spectral, F0, and durational features in order to test the whole synthesis systems.

Ten native Vietnamese listeners participated in both listening experiments. Each subject was given the texts of 15 sentences randomly selected from the test set, and asked to listen to the corresponding pairs of waveforms synthesized by the two systems in randomized order. With each pair, subjects were required to listen to the whole sentences and give their preference on speech naturalness based on two criteria: tone distortion and hoarseness. They were also asked to mark the syllables that sounded better in the pair and write down the reasons for their judgments. The test results in Table 4.5 indicate a clear preference toward the proposed system in both tests. The listeners' comments tell us that most of their preferences were based on the hoarseness, while a few of them were based on the tone distortion, of glottalized syllables. This means that although hoarseness and tone distortion often co-occur at the glottalized syllables in a synthesized utterance, the perception of the former is dominant over the latter. It is perhaps due to some masking effect of the noise introduced by a hoarse speech and/or the fact that native speakers have some ability to infer the tones and word meanings in continuous speech, making a hoarse speech easier to notice than a distorted tone. Very few comments mentioned non-glottalized syllables. The listeners' comments are consistent with the significant reduction of the VME for glottalized tones of the proposed system compared to the conventional one, and the similar performances of the two systems on non-glottalized tones, reported in Table 4.4.

Table 4.5: Results of the two paired preference tests (Test 1: only F0 was generated, Test 2: all features were generated).

Test	Conventional	No Preference	Proposed
Test 1	17.3 %	18.0 %	64.7 %
Test 2	16.0 %	10.7 %	73.3 %

4.6. Discussions

The aim of this research is to propose an F0 refinement technique that follows the out-of-the-box use (i.e., without manual intervention) of a conventional F0 extractor for glottalized syllables. Nevertheless, it is still worth considering the behavior of the conventional extractor (in this study, RAPT) when its parameters are tuned for potentially

better performance on those syllables, and compare with the proposed F0 parameterization scheme. The operation of RAPT is controlled by a set of 14 parameters [63] (pp. 509), among which $F0_{min}$, $F0_{max}$, and t were respectively fixed to 40 Hz, 400 Hz, and 5 ms as in the above experiments, while *A_FACT* is unused in the Snack's implementation. In each iteration of parameter tuning, each of the remaining 10 parameters was tuned separately to avoid their co-effects on estimation results, and the one yielded the highest performance improvement had its value fixed for the succeeding iterations. The tuning process was performed iteratively until a stopping condition was met.

With the same evaluation framework as that in Section 4.5.2, the author have observed that the tuned RAPT consistently showed a trade-off relationship between the VME and, on the other side, the VIE and RMSE. This makes it difficult to determine the highest performance improvement for each tuning iteration as well as the stopping condition. Considering the remarkable inferior of the default RAPT compared to the proposed scheme in terms of the VME (Table 4.2), the author set out to employ a tuning strategy for each iteration so that RAPT maximizes its improvement on the VME, compared to the preceding iteration, while maintaining acceptable levels of the VIE and RMSE, compared to the proposed scheme. Specifically, the differences of 5% for the VIE and 20 Hz for the RMSE were used as the tolerable thresholds since these cause remarkably more perceptible buzziness and tone distortion, respectively. When no VME decrease can be further made while the other two measures are kept acceptable, the tuning process is stopped.

Table 4.6 shows RAPT's performance on glottalized syllables in the development set after two iterations of parameter adjustment. The iteration #0 denotes RAPT with default parameters, the same as in Table 4.2. In the iteration #1, the weight given to F0 trajectory smoothness *FREQ_WT* was decreased from 0.02 to zero. In the iteration #2, the correlation window size w was decreased from 7.5 to 5 ms. It can be seen that the VME reduction of RAPT was always achieved at the expense of its increases on the VIE and RMSE. After two iterations, no further VME improvement was attainable without making the other two measures exceed the preset thresholds, thus the tuning was ceased. The proposed F0 refinement method had superior performance to RAPT regardless of the adjusted parameters. Even when RAPT obtained a slightly lower VME after the tuning iteration #2, the proposed method had considerably lower VIE and RMSE. Moreover, the proposed method let the parameter adjustment process much more controllable thanks to the fewer parameters for varying (only two, *MaxPR* and *CorrThs*) and the less sensitivity of the estimator's performance to different parameter values compared to RAPT. It should be noted that, in Table 4.6, the RMSE of the proposed method changes after each iteration of RAPT tuning because the number of frames that are simultaneously voiced in three F0

contours (the one estimated by RAPT, the one estimated by the proposed method, and the referenced one) also varies during the tuning process.

Table 4.6: Performance of RAPT on glottalized syllables in the development set after two tuning iterations. That of the proposed F0 refinement method is also provided for comparison.

Iteration	RAPT's parameters	Method	VME (%)	VIE (%)	RMSE (Hz)
#0	$FREQ\_WT = 0.02$, $w = 0.0075$	RAPT	12.76	2.47	38.70
		Proposed	3.02	1.46	33.63
#1	$FREQ\_WT = 0$, $w = 0.0075$	RAPT	3.84	5.23	54.93
		Proposed	3.02	1.46	37.41
#2	$FREQ\_WT = 0$, $w = 0.005$	RAPT	1.57	6.35	56.88
		Proposed	3.02	1.46	37.27

4.7. Conclusion

This chapter presents an F0 parameterization scheme for the Vietnamese glottalized tones by using a pitch mark propagation algorithm in combination with an F0 extractor. The proposed scheme is capable of deriving more complete and accurate F0 contours representing the tones compared to the simple use of an F0 extractor, thereby significantly alleviating the hoarseness and slightly improving the tone naturalness of a standard HMM-based speech synthesis system. It is expected that the proposed technique is also useful for other tonal languages having glottalization feature such as Chinese [65]. Future work includes extending the proposed technique to other glottalized events [59] and finding a proper way to limit voiced insertion errors. Experiments on the data of other speakers will also be conducted to confirm the effectiveness of the proposed method.

Chapter 5 Conclusions and Future Work

5.1. Conclusions

This thesis aims to improve the synthesized speech quality of current HMM-based TTS systems by considering two issues: the modeling of the dynamic features of speech parameters, and the extraction of the fundamental frequency (or F0) parameter in glottalized regions of speech signals. The dynamic features capture dynamic properties of speech parameter trajectories, thus containing important information about speech dynamics such as spectral transition. Meanwhile, the F0 parameter conveys the intonation of speech, however, is difficult to extract in speech affected by glottalization. Moreover, F0 is one of the parameters used to characterize the glottalization (or vocal fry/creaky voice) [57, 60], a non-modal phonation popular among languages and speakers [59]. Therefore, accurate modeling of the dynamic features and accurate extraction of the F0 in glottalized speech can help enhance the naturalness and expressiveness of speech synthesized from HMMs.

In Chapter 3, the modeling accuracy for the dynamic features was improved thanks to the incorporation of the generation error of dynamic features into the generation error function of the MGE criterion, a state-of-the-art HMM training framework for speech synthesis. A method for adaptively changing the weight associated with the newly added error component based on the dynamicity degree of portions of the speech signal, named MGE-dynamics-AW, was also proposed. Listening tests show that the MGE-dynamics-AW criterion obtains higher mean preference scores than the baseline one, 32.9% vs. 27.1% in the evaluation with the original parameter generation algorithm (Table 3.1) and 27.1% vs. 17.9% in the evaluation with the parameter generation algorithm considering GV (Table 3.2). The newly derived criterion improves the capability of HMMs in capturing dynamic properties of speech while maintaining a computational complexity similar to that of the baseline MGE criterion.

In Chapter 4, the problem of F0 extraction in glottalized speech signals was tackled by examining a language possessing a heavy glottalization feature, (Hanoi) Vietnamese. As a

tonal language with several glottalized tones in its tone set, the inaccurate F0 estimation has severe effects on the F0 modeling, thus degrading the tone naturalness and causing the hoarseness in synthesized Vietnamese speech. An F0 parameterization scheme for the Vietnamese glottalized tones by using a pitch mark propagation algorithm in combination with the conventional F0 extractor RAPT was presented. The proposed scheme is capable of deriving more complete and accurate F0 contours representing the glottalized tones compared to the simple use of the F0 extractor. Specifically, the voiced missing error rate was decreased from 14.51% to 6.81%, and the root mean square error was reduced from 44.00 Hz to 40.75 Hz (Table 4.4). Therefore, significantly alleviated hoarseness and slightly improved tone naturalness were perceived in the output of a standard HMM-based TTS system. The perceptual tests show remarkably higher preference scores of the proposed method over the conventional one with the differences in mean scores around 50% (Table 4.5). It is expected that the proposed technique is also useful for other tonal languages having glottalization feature such as Chinese.

5.2. Future work

The performance of MGE training considering dynamic features of speech parameters (i.e., MGE-dynamics) may be dependent on how the dynamic features are calculated. The experiments in Chapter 3 only used the standard formulae presented in Section 2.3.3 for this calculation, in which the dynamic features of a frame are estimated based on the static features of the current frame and the two frames on the left and the right. The effect of broader window lengths for dynamic feature calculation on MGE-dynamics training should be examined in the future. Besides, the thesis only investigates the effect of MGE-dynamics training on spectral parameters. A follow-up study should consider the effectiveness of the proposed method on F0 parameter with the note that the F0 is a supra-segmental feature, thus broader window lengths for dynamic feature calculation should be employed.

The research presented in Chapter 4 also exhibits several problems for further study. Firstly, since the glottalization is a non-modal phonation popular among languages [59], the proposed F0 parameterization method should be extended to cope with not only tone-related event but also other glottalized events, and not only tonal languages but also other non-tonal languages. In Section 4.4.3, tone type and syllable boundaries, which are information limited to tonal event in a tonal language, are used to specify the search direction and search boundaries for the pitch mark propagation algorithm. This algorithm needs to be improved at this point to make it become useful to other glottalized events as

well as other non-tonal languages. Secondly, experiments on the data of other speakers need to be conducted to confirm the sensitiveness of the parameter tuning process and the effectiveness of the proposed method across speakers. Thirdly, a proper way to limit voiced insertion errors when deriving F0 estimates from the pitch marks need to be developed. Finally, to fully synthesize the glottalization (or vocal fry/creaky voice) feature, not only the F0 but also other speech parameters such as the spectrum and aperiodicity need to be parameterize from speech signals [57, 60]. Future studies on characterizing this phonation should use more complex analysis/synthesis framework (e.g., STRAIGHT [18]) to extract the aperiodicity component. It is noted that this component can be used as a voicing measure to detect voiced/unvoiced, which might be a helpful solution to the third problem mentioned above.

Bibliography

[1] A. J. Hunt and A. W. Black, *Unit selection in a concatenative speech synthesis system using a large speech database*, Proc. ICASSP, pp.373-376, 1996.

[2] H. Zen, K. Tokuda, and A. W. Black, *Statistical parametric speech synthesis*, Speech Communication, vol.51, no.11, pp.1039-1064, 2009.

[3] Alan W. Black, *Unit Selection and Emotional Speech*, Proc. EUROSPEECH, pp.1649-1952, 2003.

[4] P. Taylor, *Text-to-speech synthesis*, Cambridge University Press, 2009.

[5] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, *Speech Synthesis Based on Hidden Markov Models*, Proceedings of the IEEE, vol.101, no.5, pp.1234-1252, 2013.

[6] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, *Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis*, Proc. EUROSPEECH, pp.2347–2350, 1999.

[7] Y.-J. Wu and R. H. Wang, *Minimum generation error training for HMM-based speech synthesis*, Proc. ICASSP, pp.889–892, 2006.

[8] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, *Hidden Markov models based on multi-space probability distribution for pitch pattern modeling*, Proc. ICASSP, pp.229-232, 1999.

[9] M. Wang, M. Wen, D. Saito, K. Hirose, and N. Minematsu, *Improved Generation of Prosodic Features in HMM-based Mandarin Speech Synthesis*, Proc. 7th ISCA Workshop on Speech Synthesis (SSW7), pp.359-364, 2010.

[10] Q. Zhang, F. Soong, Y. Qian, Z. Yan, J. Pan, and Y. Yan, *Improved modeling for F0 generation and V/U decision in HMM-based TTS*, Proc. ICASSP, pp.4606-4609, 2010.

[11] K. Yu and S. Young, *Continuous F0 Modeling for HMM Based Statistical Parametric Speech Synthesis*, IEEE Transactions on Audio, Speech, and Language Processing, vol.19, no.5, pp.1071-1079, 2011.

[12] J. Latorre, M. J. F. Gales, S. Buchholz, K. Knill, M. Tamurd, Y. Ohtani, and M. Akamine, *Continuous F0 in the source-excitation generation for HMM-based TTS: Do we need voiced/unvoiced classification?*, Proc. ICASSP, pp.4724-4727, 2011.

[13] C. T. Ishi, K. I. Sakakibara, H. Ishiguro, and N. Hagita, *A Method for Automatic Detection of Vocal Fry*, IEEE Transactions on Audio, Speech, and Language Processing, vol.16, no.1, pp.47-56, 2008.

[14] L. Rabiner and R. Schafer, *Theory and Applications of Digital Speech Processing*, Prentice Hall, 2011.

[15] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, *An adaptive algorithm for mel-cepstral analysis of speech*, Proc. ICASSP, pp.137–140, 1992.

[16] S. Imai, *Cepstral analysis synthesis on the mel frequency scale*, Proc. ICASSP, pp.93-96, 1983.

[17] A. V. McCree and T. P. Barnwell, III, *A mixed excitation LPC vocoder model for low bit rate speech coding*, IEEE Transactions on Speech and Audio Processing, vol.3, no.4, pp.242-250, 1995.

[18] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, *Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds*, Speech Communication, vol.27, no.3–4, pp.187-207, 1999.

[19] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, *Mixed excitation for HMM-based speech synthesis*, Proc. EUROSPEECH, pp.2263–2266, 2001.

[20] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, *Details of the Nitech HMM-Based Speech Synthesis System for the Blizzard Challenge 2005*, IEICE transactions on Information and Systems, vol.E90-D, no.1, pp.325-333, 2007.

[21] L. R. Rabiner, *A tutorial on hidden Markov models and selected applications in speech recognition*, Proceedings of the IEEE, vol.77, no.2, pp.257-286, 1989.

[22] K. Tokuda and H. Zen, *Fundamentals and recent advances in HMM-based speech synthesis*, Proc. Tutorial of INTERSPEECH, 2009.

[23] A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum Likelihood from Incomplete Data via the EM Algorithm*, Journal of the Royal Statistical Society, vol.39, no.1, pp.1-38, 1977.

[24] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, *Multi-Space Probability Distribution HMM*, IEICE Transactions on Information and Systems, vol.85, no.3, pp.455-464, 2002.

[25] T. Masuko, *HMM-Based Speech Synthesis and Its Applications*, PhD thesis, Tokyo Institute of Technology, 2002.

[26] Junichi Yamagishi, *Average-Voice-Based Speech Synthesis*, PhD thesis, Tokyo Institute of Technology, 2006.

[27] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, *Duration Modeling For HMM-Based Speech Synthesis*, Proc. ICSLP, pp.29-32, 1998.

[28] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, *A Hidden Semi-Markov Model-Based Speech Synthesis System*, IEICE Transactions on Information and Systems, vol.E90-D, no.5, pp.825-834, 2007.

[29] K. Yu, H. Zen, F. Mairesse, and S. Young, *Context adaptive training with factorized decision trees for HMM-based statistical parametric speech synthesis*, Speech Communication, vol.53, no.6, pp.914-923, 2011.

[30] J. J. Odell, *The use of context in large vocabulary speech recognition*, PhD Thesis, University of Cambridge, 1995.

[31] S. Young, J. Odell, and P. Woodland, *Tree-based state tying for high accuracy acoustic modelling*, Proc. ARPA Workshop on Human Language Technology, pp.307-312, 1994.

[32] K. Shinoda and T. Watanabe, *MDL-based context-dependent subword modeling for speech recognition*, Journal of the Acoustical Society of Japan (E), vol.21, no.2, pp.79-86, 2000.

[33] T. Yoshimura, *Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for HMM-based text-to-speech systems*, PhD thesis, Nagoya Institute of Technology, 2002.

[34] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, *Speech parameter generation algorithms for HMM-based speech synthesis*, Proc. ICASSP, pp.1315–1318, 2000.

[35] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, *The HMM-based speech synthesis system version 2.0*, Proc. 6th ISCA Workshop on Speech Synthesis (SSW6), pp.294–299, 2007.

[36] C.-T. Nguyen, X.-H. Phan, and T.-T. Nguyen, *JVnTextPro: A Java-based Vietnamese Text Processing Tool*, http://jvntextpro.sourceforge.net/, 2010.

[37] S. Furui, *Speaker-independent isolated word recognition using dynamic features of speech spectrum*, IEEE Transactions on Acoustics, Speech and Signal Processing, vol.34, no.1, pp.52-59, 1986.

[38] S. Furui, *On the role of spectral transition for speech perception*, Journal of the Acoustical Society of America, vol.80, no.4, pp.1016-1025, 1986.

[39] Y.-J. Wu, H. Zen, Y. Nankaku, and K. Tokuda, *Minimum generation error criterion considering global/local variance for HMM-based speech synthesis*, Proc. ICASSP, pp.4621-4624, 2008.

[40] Y.-J. Wu and K. Tokuda, *Minimum generation error training with direct log spectral distortion on LSPs for HMM-based speech synthesis*, Proc. INTERSPEECH, pp.577-580, 2008.

[41] M. Lei, Z.-H. Ling, and L.-R. Dai, *Minimum generation error training with weighted Euclidean distance on LSP for HMM-based speech synthesis*, Proc. ICASSP, pp.4230-4233, 2010.

[42] T. Toda and K. Tokuda, *A Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis*, IEICE Transactions on Information and Systems, vol.E90-D, no.5, pp.816-824, 2007.

[43] S. Amari, *A theory of adaptive pattern classifiers*, IEEE Transactions on Electronic Computers, vol.16, no.3, pp.299-307, 1967.

[44] K. Elenius and M. Blomberg, *Effects of emphasizing transitional or stationary parts of the speech signal in a discrete utterance recognition system*, Proc. ICASSP, pp.535-538, 1982.

[45] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, *ATR Japanese speech database as a tool of speech recognition and synthesis*, Speech Communication, vol.9, no.4, pp.357-363, 1990.

[46] Y.-J. Wu, H. Zen, Y. Nankaku, and K. Tokuda, *Evaluation of parameter optimization methods for minimum generation error based HMM training*, Proc. Autumn Meeting of ASJ, pp.370-371, 2007.

[47] L. R. Ott and M. T. Longnecker, *An Introduction to Statistical Methods and Data Analysis*, Brooks/Cole, 2010.

[48] J. P. Kirby, *Vietnamese (Hanoi Vietnamese)*, Journal of the International Phonetic Association, vol.41, no.03, pp.381–392, 2011.

[49] T. T. Vu, M. C. Luong, and S. Nakamura, *An HMM-based Vietnamese speech synthesis system*, Proc. Oriental COCOSDA, pp.116-121, 2009.

[50] T. T. T. Nguyen, C. Alessandro, A. Rilliard, and D. D. Tran, *HMM-based TTS for Hanoi Vietnamese: Issues in design and evaluation*, Proc. INTERSPEECH, pp.2311-2315, 2013.

[51] A. T. Dinh, T. S. Phan, T. T. Vu, and C. M. Luong, *Vietnamese HMM-based speech synthesis with prosody information*, Proc. 8th ISCA Workshop on Speech Synthesis, pp.55–59, 2013.

[52] T. T. T. Nguyen, A. Rilliard, D. D. Tran, and C. Alessandro, *Prosodic phrasing modeling for Vietnamese TTS using syntactic information*, Proc. INTERSPEECH, pp.2332-2336, 2014.

[53] T. T. T. Nguyen, D. D. Tran, A. Rilliard, C. Alessandro, and T. N. Y. Pham, *Intonation issues in HMM-based speech synthesis for Vietnamese*, Proc. 4th International Workshop on Spoken Language Technologies for Under-resourced Languages, pp.99-104, 2014.

[54] M. Brunelle, *Northern and Southern Vietnamese tone coarticulation: a comparative case study*, Journal of the Southeast Asian Linguistics Society, vol.1, pp.49–62, 2009.

[55] V. L. Nguyen and J. A. Edmondson, *Tones and voice quality in modern Northern Vietnamese: instrumental case studies*, Mon-Khmer Studies Journal, vol.28, pp.1-18, 1998.

[56] A. Michaud, *Final consonants and glottalization: new perspectives from Hanoi Vietnamese*, Phonetica, vol.61, no.2–3, pp.119–146, 2004.

[57] H. Silén, E. Helander, J. Nurminen, and M. Gabbouj, *Parameterization of vocal fry in HMM-based speech synthesis*, Proc. INTERSPEECH, pp.1775-1778, 2009.

[58] B. R. Gerratt and J. Kreiman, *Toward a taxonomy of nonmodal phonation*, Journal of Phonetics, vol.29, no.4, pp.365-381, 2001.

[59] H. Ding, O. Jokisch, and R. Hoffmann, *The Effect of Glottalization on Voice Preference*, Proc. SPEECH PROSODY, pp.851-854, 2006.

[60] T. Raitio, J. Kane, T. Drugman, and C. Gobl, *HMM-based synthesis of creaky voice*, Proc. INTERSPEECH, pp.2316-2320, 2013.

[61] T. Masuko, K. Tokuda, N. Miyazaki, and T. Kobayashi, *Pitch pattern generation using multi-space probability distribution HMM*, IEICE Transactions, vol.J83-D-II, no.7, pp.1600–1609, 2000.

[62] V. Colotte and Y. Laprie, *Higher precision pitch marking for TD-PSOLA*, Proc. EUSIPCO, pp.419–422, 2002.

[63] D. Talkin, *A robust algorithm for pitch tracking (RAPT)*, in *Speech Coding and Synthesis*, editors W.B. Kleijn and K.K. Paliwal, Elsevier: New York, pp.495-518, 1995.

[64] D. Stadniczuk, G. Bauckmann, and D. Suendermann-Oeft, *An open-source Octave toolbox for VTLN-based voice conversion*, Proc. International Conference of the German Society for Computational Linguistics and Language Technology, 2013.

[65] K. M. Yu, *Laryngealization and features for Chinese tonal recognition*, Proc. INTERSPEECH, pp.1529-1532, 2010.

List of Publications

Journal Papers

D. K. Ninh, M. Morise, and Y. Yamashita, *A generation error function considering dynamic properties of speech parameters for minimum generation error training for hidden Markov model-based speech synthesis*, Acoustical Science and Technology, vol.34, no.2, pp.123–132, 2013.

D. K. Ninh and Y. Yamashita, *F0 parameterization of glottalized tones in HMM-based speech synthesis for Hanoi Vietnamese*, IEICE Transactions on Information and Systems, vol.E98-D, no.12, pp.2280–2289, 2015.

International Conference Proceedings

D. K. Ninh, M. Morise, and Y. Yamashita, *Incorporating dynamic features into minimum generation error training for HMM-based speech synthesis*, Proc. 8th International Symposium on Chinese Spoken Language Processing (ISCSLP), pp.55–59, 2012.

D. K. Ninh and Y. Yamashita, *F0 parameterization of glottalized tones for HMM-based Vietnamese TTS*, Proc. INTERSPEECH, pp.2202–2206, 2015.

Domestic Conference Proceedings

D. K. Ninh, K. Cho, and Y. Yamashita, *Introduction of duration models and dynamic features in MGE training for HSMM-based speech synthesis*, Proc. Spring Meeting of ASJ, pp.431–434, 2012.

D. K. Ninh, M. Morise, and Y. Yamashita, *An adaptive weighting approach for minimum generation error training considering dynamic features in HMM-based speech synthesis*, Proc. Autumn Meeting of ASJ, pp.383–386, 2012.