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

Abstract ... i

Abstract in Japanese ... ii

Acknowledgments ... iii

Table of contents ... iv

List of Figures ... vi

List of Tables ... viii

Chapter 1 Introduction ... 1

Chapter 2 HMM-based Text-to-Speech Synthesis ... 5

2.1. Introduction ... 5

2.2. Speech analysis/synthesis framework ... 6

2.3. Training stage ... 7

2.3.1. Spectral modeling with continuous distribution HMMs ... 7

2.3.2. F0 modeling with multi-space distribution HMMs ... 11

2.3.3. Dynamic feature calculation ... 13

2.3.4. Duration modeling ... 14

2.3.5. Context-dependent modeling and context clustering ... 15

2.4. Synthesis stage ... 17

2.4.1. Text analysis ... 17

2.4.2. State duration determination ... 18

2.4.3. Effect of dynamic features in speech parameter generation ... 19

Chapter 3 Minimum Generation Error Training Considering Dynamic Features 22

3.1. Introduction ... 22

3.2. Speech parameter generation algorithms ... 23

3.2.1. Original parameter generation algorithm ... 24

3.2.2. Parameter generation algorithm considering global variance ... 25

3.3. Minimum generation error training criterion ... 26

3.4. Proposed MGE criterion considering dynamic features ... 28

3.4.1. Fixed weighting approach to MGE-dynamics ... 29

3.4.2. Adaptive weighting approach to MGE-dynamics ... 29

3.5. Experiments....31
3.5.1. Experimental conditions.... 32
3.5.2. Evaluation with the original parameter generation algorithm.... 33
3.5.3. Evaluation with the parameter generation algorithm considering GV.... 38
3.6. Discussions....39
3.7. Conclusion....40
Chapter 4 F0 Parameterization of Glottalized Tones in HMM-based TTS for Hanoi Vietnamese....41
4.1. Introduction....41
4.2. Vietnamese glottalized tones....42
4.3. Problems with F0 extraction for glottalized tones....44
4.3.1. Popular F0 extraction errors of RAPT.... 45
4.3.2. Effects of F0 extraction errors on MSD-HMM modeling and generation.... 47
4.4. Proposed F0 parameterization of glottalized tones....47
4.4.1. Pre-processing of F0s.... 48
4.4.2. Detection of pitch marks in regular region.... 48
4.4.3. Detection of pitch marks in glottalized region.... 49
4.4.4. Derivation of F0 estimates from pitch marks.... 54
4.5. Experimental evaluations....54
4.5.1. Common system setups.... 54
4.5.2. Parameter tuning for the proposed method.... 55
4.5.3. Remark on the resulting F0 model sizes.... 58
4.5.4. Objective evaluations.... 59
4.5.5. Perceptual evaluations.... 60
4.6. Discussions....61
4.7. Conclusion....63
Chapter 5 Conclusions and Future Work....64
5.1. Conclusions....64
5.2. Future work....65
Bibliography....67
List of Publications....72

List of Figures

Figure 2.1: The scheme of a typical HMM-based TTS system [2]. 5

Figure 2.2: The source-filter model of speech production. 6

Figure 2.3: A 3-state no-skip left-to-right HMM generates an observation sequence (adapted from [22]). 8

Figure 2.4: F0 pattern modeling on two spaces [25]. 12

Figure 2.5: MSD-HMM for F0 modeling [22]. 13

Figure 2.6: Feature vector of a speech frame [26]. 14

Figure 2.7: An example of decision tree based context clustering [26]............................. 16

Figure 2.8: Block diagram of the synthesis stage [33].. 18

Figure 2.9: State duration generation [26]. 19

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. 20

Figure 3.1: Trajectories of second mel-cepstral coefficients extracted from natural speech and that generated from HMM [2]. 26

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. 30

Figure 3.3: Evolution of generation error of the 2nd mel-cepstral coefficient on test data. 34

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. 35

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. 36

Figure 3.6: Natural and generated trajectories of the 2nd mel-cepstral coefficient for an utterance included in the training data. 37

Figure 4.1: Waveforms of the syllable /ɗa/ exhibit different periodicity characteristics when accompanied by (a) a non-glottalized tone, and (b) a glottalized tone. 43

Figure 4.2: Number of occurrences of each tone in the database used in this research. 44

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. 46

Figure 4.4: Proposed F0 parameterization method for a glottalized syllable. 48

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 4.2a was used. 49

Figure 4.6: Example of pitch mark searching forward (with *MaxPR* = 2.7). 51

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 4.2a was used. 52

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. 53

Figure 4.9: Results of the forward and backward pitch mark combination for the example in Figure 4.6 (top plot) and F0 contours estimated by RAPT and the proposed method (bottom plot). The same settings of Figure 4.2a were used in both plots..53

Figure 4.10: Performance of the proposed F0 refinement method on the development set when varying the *MaxPR* and *CorrThs* parameters. 57

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. 60

List of Tables

Table 3.1: Mean preference score (with 95% confidence interval) in evaluation with the original parameter generation algorithm. 38

Table 3.2: Mean preference score (with 95% confidence interval) in evaluation with the parameter generation algorithm considering GV. 38

Table 4.1: Characteristics of the glottalized tones. 44

Table 4.2: Performances of RAPT and the proposed F0 refinement method on glottalized syllables in the development set. 57

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. 59

Table 4.4: Performances of the conventional and proposed systems on the test set. 60

Table 4.5: Results of the two paired preference tests (Test 1: only F0 was generated, Test 2: all features were generated). 61

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. 63