博士論文要旨

論文題名：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

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

ニン　カン　ズイ
NINH Khanh Duy

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。

Abstract of Doctoral Thesis

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

Doctoral Program in Integrated Science and Engineering
Graduate School of Science and Engineering
Ritsumeikan University

ニン　カン　ズイ
NINH Khanh Duy

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 contain 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.