論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Statistical Parametric Speech Synthesis Using Local Variance and Quantized F0 Context
著者(和文)	CHUNWIJITRAVATAYA
Author(English)	Vataya Chunwijitra
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第9322号, 授与年月日:2013年9月25日, 学位の種別:課程博士, 審査員:小林　隆夫,羽鳥　好律,小池　康晴,杉野　暢彦,篠崎　隆宏
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第9322号, Conferred date:2013/9/25, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

専攻： Department of	物理情報システム　専攻	申請学位（専攻分野）： Academic Degree Requested	博士 Doctor of　（　工学　）
学籍番号： Student ID Number		指導教員（主）： Academic Advisor(main)	小林　隆夫
学生氏名： Student's Name	CHUNWIJITRA Vataya	指導教員（副）： Academic Advisor(sub)	

要旨（英文 800 語程度）
Thesis Summary（approx.800 English Words ）

This thesis describes novel tone-modeling and parameter generation approaches to generate more natural-sounding speech in statistical parametric speech synthesis. In this work, we concentrate on a major technique of statistical parametric speech synthesis approach, hidden Markov model (HMM)-based speech synthesis. In this technique, there are some issues which should be investigated and improved. Therefore, in this thesis, the work is divided into two phases including the first phase of the tone correctness improvement in tonal language and the second phase of the spectral reproducibility improvement.

In the first phase, we describe a novel approach to improving the tone correctness in speech synthesis of a tonal language based on an average-voice model trained with a corpus from nonprofessional speakers' speech. The intelligibility and naturalness of synthetic speech are degraded in HMM-based speech synthesis when using a small amount of speech data from nonprofessional speakers. There are tone disagreements between the tonal labels and the recorded speech samples when the labels are automatically generated from transcriptions. The problem with inconsistent labeling in tonal languages is crucial because incorrect tone labels affect the tone correctness of synthesized speech. Moreover, it is not an easy task to manually modify incorrect tone labels inexpensively. Therefore, we focused on reducing tone disagreements in speech data acquired from nonprofessional speakers without manually modifying the labels. To reduce the distortion in tone caused by inconsistent tonal labeling, quantized F0 symbols were utilized as the tone context to obtain an appropriate F0 model. With this technique, the tonal context label can be directly extracted from the original speech and this prevents inconsistency between speech data and F0 labels generated from transcriptions, which affect naturalness and the tone correctness in synthetic speech. We examined two types of labeling for the tonal context using phone-based and sub-phone-based quantized F0 symbols. Subjective and objective evaluations of the synthetic voice were carried out

in terms of the intelligibility of tone and its naturalness. The experimental results from both the objective and subjective tests revealed that the proposed technique could improve not only naturalness but also the tone correctness of synthetic speech under conditions where a small amount of speech data from nonprofessional target speakers was used.

In the second phase, we describe a novel approach to improving the spectral reproducibility by reducing the over-smoothing problem. In the conventional parameter generation algorithm, the resultant spectral trajectory is often excessively smoothed by the parameter tying in the model training. This causes the degradation of perceptual quality and makes the synthetic speech sounding buzzy and muffled. To alleviate the over-smoothing effect, we propose a parameter generation algorithm using a local variance (LV) model in HMM-based speech synthesis. In the proposed technique, we define LV as a feature that represents the local variation of a spectral parameter sequence and model LVs using HMMs. Context-dependent HMMs are used to capture the dependence of LV trajectories on phonetic and prosodic contexts. In addition, the dynamic features of LVs are taken into account as well as the static one to appropriately model the dynamic characteristics of LV trajectories. By introducing the LV model into the speech parameter generation process, the proposed technique can impose a more precise variance constraint for each frame than the conventional technique with a global variance (GV) model. Consequently, the proposed technique alleviates the excessive spectral peak enhancement that often occurs in GV-based parameter generation. Objective evaluation results showed that the proposed technique can generate better spectral parameter trajectories than the GV-based technique in terms of spectral and LV distortion. Moreover, the results of subjective evaluation demonstrated that the proposed technique can generate synthetic speech significantly closer to the original one than the conventional technique while maintaining speech naturalness.