HMM-based Speech Synthesis with an Acoustic Glottal Source Model

Article · January 2009

CITATIONS
5

READS
57

4 authors, including:

Korin Richmond
The University of Edinburgh
84 PUBLICATIONS **1,762** CITATIONS
SEE PROFILE

Junichi Yamagishi
National Institute of Informatics, Japan
353 PUBLICATIONS **6,836** CITATIONS
SEE PROFILE

HMM-based Speech Synthesis with an Acoustic Glottal Source Model

João P. Cabral, Steve Renals, Korin Richmond and Junichi Yamagishi

The Centre for Speech Technology Research
University of Edinburgh,UK
jscabral@inf.ed.ac.uk, s.renals@ed.ac.uk, korin@cstr.ed.ac.uk, jyamagis@inf.ed.ac.uk

Abstract

A major cause of degradation of speech quality in HMM-based speech synthesis is the use of a simple delta pulse signal to generate the excitation of voiced speech.

This paper describes a new approach to using an acoustic glottal source model in HMM-based synthesisers. The goal is to improve speech quality and parametric flexibility to better model and transform voice characteristics.

Index Terms: HMM-based Speech Synthesis, LF-Model, Glottal Spectral Separation

1. Introduction

Standard HMM-based speech synthesisers produce speech by shaping a multi-band mixed excitation with the spectral envelope e.g. [1]. The periodic component of the excitation is modelled by a delta pulse, which produces a buzzy speech quality.

Recently, different types of excitation signals have been used in HMM-based synthesisers to reduce the buzzy effect. For example, a multipulse-based excitation [3] and a glottal flow pulse parameterised with the linear prediction (LP) coefficients [4]. In previous work, we have similarly tried to reduce buziness by using a post-filtered LF-Model signal [2]. In that work, the parameters were not modelled within the statistical framework of the synthesiser. In our current work, we extend that approach by modelling the LF-model parameters and using them in the synthesis, without applying any post-filter.

2. System

Our system is a modified version of the speaker-dependent HMM-based speech synthesiser called Nitech-HTS 2005 [1]. Table 1 shows the main differences between the two systems.

2.1. Analysis

In our system, the LF-Model parameters are calculated from the LP residual signal as described in [5]. One of the parameters is $F_0 = 1/T$, where T is the duration of the LF-Model. The other parameters are obtained by fitting the LF-model to the residual pitch-synchronously, using a non-linear optimisation algorithm.

In Nitech-HTS 2005 the spectrum features are obtained from the spectral envelope of the speech signal, which is calculated with the STRAIGHT method. This spectrum cannot be used to synthesise speech with the LF-model, because it too contains information about the source. We use the Glottal Spectral Separation (GSS) method [5] to remove the spectral effects of the LF-model from the speech signal prior to estimating the spectral envelope with STRAIGHT.

Supported by Marie Curie Early Stage Training Site EdSST (MEST-CT-2005-020568)

Table 1: *Main aspects of the Nitech-HTS 2005 with LF-Model.*

	Nitech-HTS 2005	HTS with LF-Model
Analysis	STRAIGHT	LF-model Extract., GSS
HMM Feature Vectors	39 mel-sp.c., Δ, Δ^2 5 aperiod., Δ, Δ^2 log F_0, Δ, Δ^2	39 mel-sp.c., Δ, Δ^2 5 aperiod., Δ, Δ^2 5 log LF-param., Δ, Δ^2
Synthesis	MLSA filter & OLA	FFT proc. & OLA

2.2. Acoustic Modelling

The five-state HMM structure was not modified. However, the HMMs settings were adjusted to take into account the new glottal features , e.g. the dimension of the LF-model, Δ and Δ^2 streams was set to 5. These streams are modelled by HMMs based on multi-space probability distributions (MSD-HMMs), because the LF-parameters are undefined in unvoiced regions. Also, the decision tree used for HMM clustering with the LF-parameters is the same as the one used with F_0 in Nitech 2005.

2.3. Synthesis

Speech is synthesised with the LF-model by multiplying the FFT of the mixed excitation by the amplitude spectrum obtained from the spectral parameters, as in [5]. Then, the short-time signals are overlapped-and-added (OLA).

We have found the HTS with LF-model produced improvements in speech quality (http://homepages.inf.ed.ac.uk/jscabral /HTS.html). Also, formal evaluations are being conducted.

3. Conclusions

A glottal source model has been successfully integrated into a standard HMM-based speech synthesiser by using the LF-Model, a new spectrum and HMM modelling adjustments.

4. References

[1] Zen, H., Toda, T., Nakamura, M. and Tokuda, K., "Details of Nitech HMM-based Speech Synthesis System for the Blizzard Challenge 2005", IEICE Tran., V.E90-D, No1, pp. 325–333, 2007.

[2] Cabral, J., Renals, S., Richmond, K. and Yamagishi, J., "Towards an improved modeling of the glottal source in statistical parametric speech synthesis", In SSW6-2007, Germany, 2007.

[3] Maia, R., Toda, T., Zen, H., Nankaku, Y. and Tokuda, K., "An excitation model for HMM-based speech synthesis based on residual modeling", In SSW6-2007, pp. 131–136, 2007.

[4] Raitio, T., Suni, A., Pulakka, H., Vainio, M. and Alku, P, "HMM-Based Finnish Text-to-Speech System Utilizing Glottal Inverse Filtering", In Proc. of the Interspeech, Australia, 2008.

[5] Cabral, J., Renals, S., Richmond, K. and Yamagishi, J., "Glottal Spectral Separation for Parametric Speech Synthesis", In Proc. of the Interspeech, Australia, 2008.