Glottal Source and Prosodic Prominence Modelling in HMM-based Speech Synthesis for the Blizzard Challenge 2009

J. Sebastian Andersson, Joao P. Cabral, Leonardo Badino, Junichi Yamagishi, Robert A.J. Clark

The Centre for Speech Technology Research
University of Edinburgh, Edinburgh, UK
J.S.Andersson@sms.ed.ac.uk, jscabral@inf.ed.ac.uk, l.badino@sms.ed.ac.uk

Abstract

This paper describes the CSTR entry for the Blizzard Challenge 2009. The work focused on modifying two parts of the Nitech 2005 HTS speech synthesis system to improve naturalness and contextual appropriateness. The first part incorporated an implementation of the Linjencrants-Fant (LF) glottal source model. The second part focused on improving synthesis of prosodic prominence including emphasis through context dependent phonemes. Emphasis was assigned to the synthesised test sentences based on a handful of theory based rules. The two parts (LF-model and prosodic prominence) were not combined and hence evaluated separately. The results on naturalness for the LF-model showed that it is not yet perceived as natural as the Benchmark HTS system for neutral speech. The results for the prosodic prominence modelling showed that it was perceived as contextually appropriate as the Benchmark HTS system, despite a low naturalness score. The Blizzard challenge evaluation has provided valuable information on the status of our work and continued work will begin with analysing why our modifications resulted in reduced naturalness compared to the Benchmark HTS system.
Index Terms:Speech Synthesis, HMM, LF-Model, prosodic prominence, emphasis

1. Introduction

The quality of HMM-based speech synthesisers has been improving in the recent years. However, speech synthesised with this type of synthesisers does not sound as natural as the speech produced by the best unit-selection systems [1].

The main problems with HMM-based synthesisers is that the synthetic speech sounds "buzzy" and "muffled". These characteristics are related with the limitations of the parametric model of speech used by this type of synthesisers and the over-smoothing of the parameter trajectories due to the statistical modelling, respectively.

1.1. Blizzard Challenge 2009

The speech databases released for the Blizzard Challenge 2009 included the same UK English and Mandarin Chinese databases which were released for the Blizzard 2008 [1].

The participants were able to use the speech databases to build voices for several different evaluation tasks. We focused our work on two of these available tasks and our entry consisted of two independent modifications to the Nitech HTS 2005 system [2]:

- Task EH2: A voice built with the *Arctic* subset of the speech database where we aimed to reduce the buzziness of the synthetic speech by integrating an acoustic glottal source model (described in section 3).
- Task ES3: A voice that sound natural and appropriate within a dialogue context where we aimed to improve prosodic expressiveness and realization of context-dependent prosody in TTS synthesis by identifying contrast and new information and prosodically mark it with emphatic pitch accents (described in section 4).

2. Nitech-HTS 2005

The following sections describes the Nitech HTS 2005 [2] which we refer to as the standard Nitech HTS 2005. This system was also used as the HTS Benchmark system in the Blizzard Challenge 2009 evaluation.

2.1. Overview

We have modified the speaker-dependent HMM-based speech synthesizer Nitech-HTS 2005 [2]. The general architechture of the Nitech-HTS 2005 is shown in figure 1.

In the analysis part, Nitech-HTS 2005 uses the STRAIGHT method [3] to calculate the spectral envelope of the speech signal and the aperiodicity measurements. F_0 can also be calculated with STRAIGHT, using fixed point analysis [4], or with the F_0 detector from the Entropic Signal Processing System (ESPS), [5] and [6].

In the training, the HMM parameters (including state duration densities) are estimated automatically using maximum likelihood estimation. After the training, the speech parameters can be calculated from the input text labels using the parameter generation algorithm.

The system synthesises the speech by shaping the speech spectrum on the mixed excitation signal. This excitation is modelled by the weighted sum of a delta pulse train with multi-band Gaussian noise calculated from the aperiodicity features.

2.2. Acoustic features

The spectral features are the mel-cepstral coefficients, which are calculated from the speech spectrum. The excitation features are the energy of noise in five different frequency bands and the fundamental frequency, F_0.

2.3. Context-Dependent Labels/Phonemes

The text of the training and test sentences are analysed with regards to phonetic, linguistic and prosodic information and converted into context dependent phoneme models (triphone or quinphone plus linguistic and prosodic information).

10.21437/Blizzard.2009-5

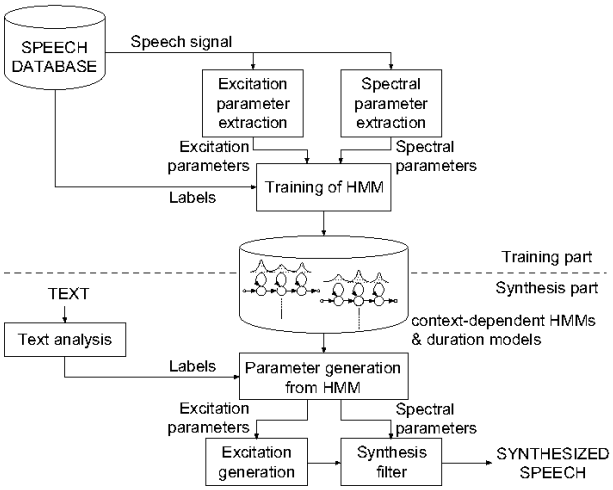

Figure 1: *Overview of the Nitech 2005 HMM-based speech synthesis system.*

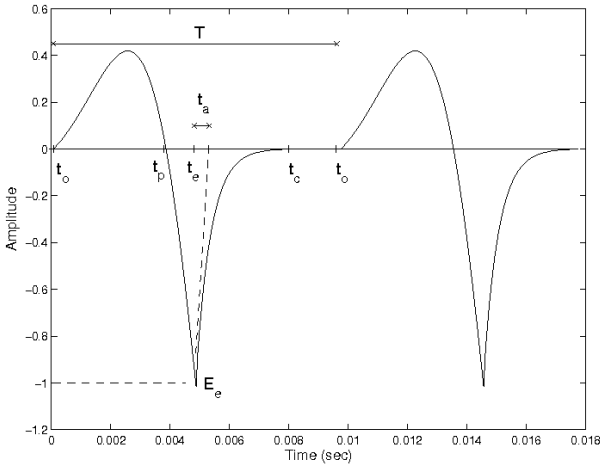

Figure 2: *Segment of the LF-model waveform with the representation of the glottal parameters during one period.*

The specifications of context dependent models for neutral English is generally very similar to [7] or its more recent variant[8], where most contexts are related to counts, positions and distances of stressed or accented syllables, and stretches from phoneme to utterance level contexts, e.g.:

- {preceding, current, succeeding} phoneme
- position of current phoneme in current syllable
- number of phonemes in {preceding, current, succeeding} syllable
- accent of {preceding, current, succeeding} syllable
- number of {preceding, succeeding} accented syllables in current phrase
- position of current word in current phrase
- number of syllables in current utterance

The large number of context gives rise to a very large number of context dependent phonemes and to be able to train and generate from these models they are clustered based on these contexts with tree-based clustering to share parameters and reduce the number of models.

In addition to the contexts described above the Benchmark HTS system in the Blizzard Challenge 2009 also clusters based on articulatory features (e.g. front vowels, plosives, fricatives, etc.) [9]

2.4. Statistical Modelling

The statistical model is a five-state left-to-right HMM. Each state output probability density function consists of five streams: spectral parameters, noise parameters, log$F0$, Δ of logF_0 and Δ^2 of logF_0. The spectrum and aperiodicity parameters are modeled by continuous HMMs while the last three streams are modelled by HMMs based on multi-space probability distributions (MSD-HMMs) [10] because F_0 is undefined in unvoiced regions.

The time structure of speech is also modelled by continuous probability functions for state durations.

The context-dependent HMMs take into account the contextual factors described in [8]. However, they are also clustered using decision trees [11] to better model the contextual factors (phonetic, prosodic and linguistic). Since the spectral, F_0 and duration parameters have their own contextual factors, they are clustered independently.

3. Nitech-HTS 2005 with LF-Model

The major factor which causes the buzziness is the use of a simple delta pulse signal to generate the excitation of voiced speech. To reduce this effect, the standard Nitech HTS 2005 uses a multi-band mixed excitation. Recently, other types of excitation models have been used in HMM-based synthesisers to improve the speech quality. For example, an excitation model based on state-dependent filters and pulse-trains mixed with noise has been proposed [12]. Another method is to manipulate a glottal flow pulse extracted from real speech using the glottal source parameters modelled and generated by HMMs [13]. An acoustic glottal source model, the Liljencrants-Fant (LF) Model [14] has also been integrated in the Nitech HTS 2005, [15] and [16]. In [15], the delta pulse was replaced by a post-filtered LF-model signal, but the glottal parameters were not modelled within the statistical framework of the synthesiser. We have used the second approach, [16], which models the LF-model parameters and uses them to synthesise the excitation signal, without applying any filtering operation.

3.1. Liljencrants-Fant Model

The Liljencrants-Fant model (LF-model) is a popular acoustic glottal source model [14]. The model is divided into three parts and is given by the following equation:

$$e_{LF}(t) = \begin{cases} E_0 e^{\alpha t} \sin(w_g t), & 0 \le t \le t_e \\ -\frac{E_e}{\epsilon T_a}[e^{-\epsilon(t-t_e)} - e^{-\epsilon(t_c-t_e)}], & t_e < t \le t_c \\ 0, & t_c < t \le T \end{cases} \quad (1)$$

where $w_g = \pi / t_p$. The parameters E_0, ϵ and α can be calculated from equations 1. Figure 2 shows a segment of the LF-model and the five glottal parameters: t_p, t_e, t_c, T_a, E_e, and the pitch period T. To reduce the number of parameters, we set t_c equal to the fundamental period ($t_c = T$).

3.2. Analysis

In spectrum and glottal source parameters are calculated as described in [16].

3.2.1. Glottal Source Parameters

The LF-Model is estimated from the linear prediction (LP) residual, as described in [16]. The LP coefficients are calculated pitch-synchronously using Hanning windows with length 20 ms and centered in the glottal epochs. Then, the five LF-model parameters are calculated for each pitch cycle of the residual, which starts at each glottal epoch and has the duration of the pitch period. The pitch period ($T = 1/F_0$) and the glottal epochs are calculated using the Entropic Signal Processing System (ESPS) tools. The other LF-parameters are obtained by fitting the LF-model to a low-pass filtered version of the LPC residual signal using a non-linear optimisation algorithm.

The initial estimates of the parameters for the optimization algorithm are calculated using the direct methods in [15]. Then, the values of the parameters are varied for a maximum number of iterations to minimize the mean-squared error between the LF-model and the short-time signal using the Levenberg-Marquardt algorithm [17].

3.2.2. Spectral Parametets

The spectral envelope estimated by STRAIGHT cannot be used in conjunction with the LF-model, because it too contains information about the source, e.g. the spectral tilt. We overcome this problem by using the Glottal Separation Method [16] which removes pitch-synchronously the spectral effects of the LF-Model from the speech spectrum before calculating the spectral envelope with STRAIGHT. Basically, this method consists in dividing the speech spectrum by the amplitude spectrum of the estimated LF-Model for each short-time signal (pitch cycle delimited by contiguous epochs).

3.3. Acoustic Modelling

The five-state HMM structure was not modified. However, the dimension of the F_0 streams was set to 5 to take into account the new glottal features: logarithm values of the LF-parameters, Δ and Δ^2 of the logarithm LF-parameters values. 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. This simplification is based on the assumption that the LF-Model parameters are correlated with F_0.

3.4. 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. Then, the short-time signals are pitch-synchronously overlapped-and-added (PSOLA) to obtain smooth transitions. This method is described in more detail in [16].

4. Nitech-HTS 2005 with Emphasis Realisation

The main goal of task ES3 was that of evaluating the appropriateness of synthetic speech when generated in context. In that respect prosodic prominence is the dimension of speech that play the most important role. This is especially true for languages like English where the most informative and salient words tend to be prosodically more prominent than remaining words. If we assume that prosodic prominence is somehow categorizable, the task of identifying prosodic prominence patterns can be reformulated as the task of identifying sequences of pitch accents.

In order to generate contextually appropriate prosodic prominence we distinguished between "standard" and "emphatic" (i.e., more prominent than standard accents) accents and included them in the set of training features.

Emphatic accents were used to mark both new informative words and contrastive words (although it still is under debate whether the two discourse related effects are marked by different accents or not, see [18]).

New informative words are the "answers within the system's answer", i.e. the words that carry relevant information and have not been previously mentioned (neither in the answer nor in the user's question). For example, given the question-answer pair:

Q - I'm looking for a French restaurant in the Old Town

A - Pascal's and the Old Town Bistro are both French restaurants.

Pascal's and the *Old Town Bistro* are the new informative words of the answer.

Contrastive words are words that explicitly contrast each other, as in the sentence "Pascal's is expensive. On the other hand the Old Town Bistro is cheap" where *Pascal's - Old Town Bistro* and *expensive - cheap* are two contrastive pairs.

Emphatic accents were available and already annotated (as capitalised words) at the "emphasis" section of the Blizzard training data. Emphatic words in the Blizzard test sentences were manually determined following some simple rules (see 4.1.2).

Standard accents were used to mark all other words that are neither new nor contrastive but are usually accented, independently of the discourse context. Standard accents were predicted using a pitch accent predictor using discourse-context independent features (see below). The accent predictor was used to predict accents on both training data and Blizzard test sentences.

4.1. Identifying Prominent Words

The degree of informativeness and salience of a word (which in turn determines its prosodic prominence) depends on intrinsic properties of the word (e.g. the indefinite article "a" is usually much less informative then word "Porsche"), discourse context (e.g. "I said **that** Porsche not **a** Porsche") and speaker's intention, i.e. speakers tend to mark those words that convey the message they want the hearers to receive. In general the last two factors are very hard to compute. However in short dialogues, consisting of single question-answer pairs as those presented in task ES3, discourse-context factors like new/given and contrast/no-contrast dichotomies, and speaker's intentions (simply determined by the application's ultimate goal, i.e. "inform the customer about the restaurants that match her request") are certainly much easier to identify and compute, and so can be used to model prosodic prominence in context.

The pitch accent predictor we used to predict standard pitch accents was intended to identify words that are prosodically prominent because of their intrinsic informativeness and of a

very limited local context, while emphatic accents were (manually) assigned looking at new/given and contrast/no-contrast distinction.

4.1.1. "Standard" Pitch Accent Prediction

Our pitch accent predictor is a CART based predictor using a five-word observation window on three training features: logarithm of the probability of unigram and of bigram, and Part-of-Speech. These features have been proved to be highly correlated to pitch accenting and more correlated than more complex semantic and syntactic features [19]. The predictor was trained and tested on the f2b voice of the Boston University Radio Corpus. Tested on a 10-fold cross validation the predictor achieves a 84.8% accuracy which is comparable with accuracy of predictors reported in the recent literature (see [20]for example).

In an attempt to improve the predictor accuracy and model a possible dependency of pitch accent placement on previous accent placements we also trained and tested predictors based on HMM and on Conditional Random Fields models, but they did not compare favourably with the CART based predictor.

4.1.2. Identification of emphatic words

Emphatic accents were assigned to informative new words and contrastive words that did not fall into the following two cases:

1. the word was the last or the first of a clause
2. the word was not the most informative part of a proper name (e.g. New in "New Town") or (in case we reckoned all words to be equally informative) the last part of a proper name (e.g. the first Chop of "Chop Chop")

Words following into the two exceptions above were accented with a standard accent. We had to apply exception 1 since clause initial and clause final emphatic words turned out to be exorbitantly emphatic.

In order to investigate whether a task like ES3 is entirely automatizable, we also investigated the feasibility of the automatic identification of new and contrastive words on the 45 sentences shown on the Blizzard website as examples of the ES3 task. While the identification of new informative words is quite straightforward (they are all cardinal numbers or proper nouns or adjectives not previously mentioned), identification of contrastive word pairs is more difficult. Using the contrast tagger proposed in [21], which combines lexical, syntactic and semantic information, we achieved a 78% recall and 90% precision on stratified leave-one-out accuracy estimation.

4.2. Selected Speech Data

The provided Blizzard Speech data have previously been used to synthesise emphatic speech accents with unit selection and had existing mark-up of emphasis [22].

For the purpose of synthesising contextually appropriate speech we selected:

- *Arctic* containing 1132 utterances for general phonetic coverage
- *Emphasis* containing 1683 carrier sentences with more than 1100 emphasised names in the following template format:

  "It was JAMES who did it."

  "No, it was JOHN who did it."

  "It was JOHN, not JAMES!"

4.3. Prosodic Prominence as Context Dependent Phonemes

The context dependent phonemes in HMM-based speech synthesis determines the phonetic, linguistic and prosodic categories for training as well as generation. In the Nitech HTS 2005 [2], as well as the Blizzard Challenge Benchmark HTS, prosodic prominence categories are restricted to lexical stress and pitch accents. To be able to synthesise more contextually appropriate speech we included emphasis in addition to lexical stress and pitch accents (see section 4 for motivation of emphasis).

In the speech synthesis Blizzard Challenge 2008 [1] some teams [23] [24] included emphasis contexts in HMM-based speech synthesis systems, but no results were reported.

As part of an ongoing investigation into prosodic modelling through context dependent phonemes in HMM-based speech synthesis we selected a different set of contexts than [7] on the basis that there were potentially important information missing, and some contexts had rather opaque prosodic relevance.

Instead we selected a small set of concrete (as opposed to counts) contexts within a more controllable prosodic window of at most preceding, current and succeeding word:

- *which* {preceding, current, succeeding} phoneme (e.g. uh1)
- *which* {preceding, current, succeeding} syllable (e.g. b_uh1_t)
- *which* {preceding, current, succeeding} word (e.g. but)

The phoneme and syllable names both included lexical stress (0,1,2). Phonemes were clustered both on articulatory features [9] and stress level. word context, clustering was only applied to words with frequency above 20 in the training data, which limit the word context to mainly closed class words, and thereby separated function from content words. A distinction was made between utterance internal and beginning/final silences.

The contexts for pitch accent and emphasis were binary values set for pitch accents on:

- *which* current syllable nucleus
- {preceding,current,succeeding} syllable
- {preceding,current,succeeding} word

And for emphasis on:

- *which* {preceding, current, succeeding} phoneme
- *which* current syllable nucleus
- {preceding, current, succeeding} syllable

That pitch accents and emphasis did not have the same context specifications is partly motivated by that emphasis is stronger than pitch accents and affect nearby phones more, and partly that all our emphasis is in carrier sentences (see section 4.2) and a larger prosodic window might have resulted in modelling artifacts.

Pitch accents were automatically predicted for the training data using the pitch accent predictor described in section 4.1.1.

4.4. Resulting Voice

Informal listening tests of the resulting voice suggests that the general quality was reasonably good, that pitch accents made a positive impact on the quality and that emphasis could be realized.

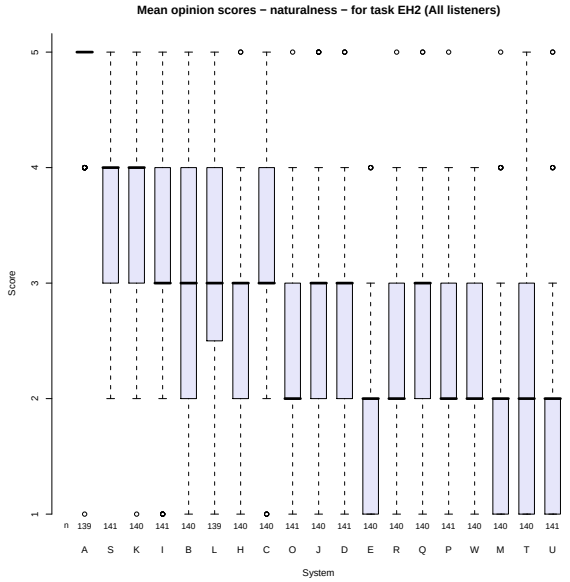

Figure 3: *Mean opinion scores of* naturalness *for all systems and listeners. Our entry is represented by letter U, the Benchmark HTS (Nitech HTS 2005) by letter C and natural speech by letter A*

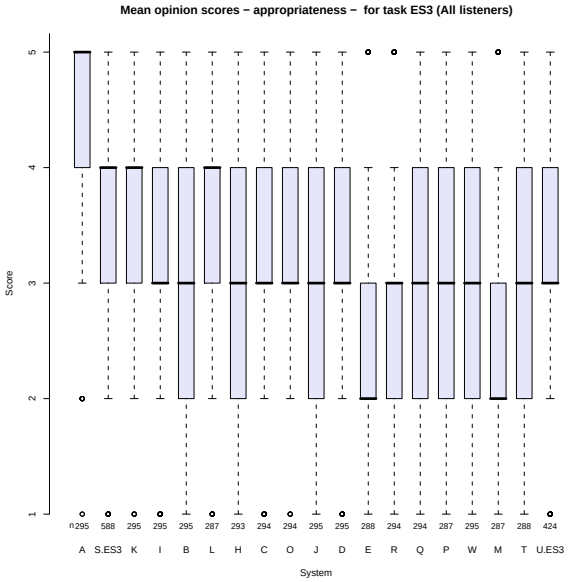

Figure 4: *Mean opinion scores of* appropriateness *for all systems and listeners. Our entry is represented by letter U, the Benchmark HTS (Nitech HTS 2005) by letter C and natural speech by letter A*

All our context categories were used substantially in the tree based clustering, but more analysis is required before making claims about the usefulness of specific categories.

5. Blizzard 2009 Evaluation

The submitted voices were evaluated in the large scale evaluation conducted by the Blizzard Challenge comittee.

- Task EH2 (voice built with the Arctic speech database) was evaluated with mean opinion scores (MOS) of naturalness and similarity to original speaker. Intelligibility was evaluated with semantically unpredictable sentences.
- Task ES3 was evalauted with mean opinion scores of naturalness and appropriateness within a dialogue context.

5.1. Results

5.1.1. Naturalness (EH2)

The naturallness mean opinion scores (MOS) for all systems and all listeners (listeners who participated in the lab and web-based evaluations) are shown in Figure 3. In the figure, our entry in the Blizzard listening test is identified by the letter U. The Benchmark HTS is represented by the letter C in the figure. Letter A represents the natural speech, which obtains the highest score as expected.

In general, the synthetic speech produced with the Nitech-HTS 2005 with LF-Model does not sound as natural as the speech produced with the standard Nitech-HTS 2005. Our system also performed worse than the standard HTS in the similarity and intelligibility evaluations.

We expected the HTS-system with LF-model to perform at least as good as the standard Nitech HTS 2005, which does not have an acoustic glottal source model. However, this new HTS system is at an early stage and it is not performing at its best yet. For example, we have already improved the robustness of the glottal features extraction.

5.1.2. Contextually Appropriate Synthesis (ES3)

The contextual *appropriateness* mean opinion scores (MOS) for all systems and listeners are shown in figure 4. In the figure, our entry is identified by the letter U, the Benchmark HTS (Nitech HTS 2005) is represented by the letter C and letter A represents natural speech.

The evaluation showed that our system was perceived as contextually appropriate as the standard Nitech HTS 2005, despite that it was not perceived as very natural (median 2, not shown in any figure). Given that we manually selected appropriate *places* for emphasis, the low naturalness score suggests that it is the *manner* or realisation of in particular emphasis that needs to be improved.

To investigate whether our method of context dependent models is at least as natural as the Benchmark HTS, we should also compare the two methods on the same dataset, for training and synthesis, with the same categories of prosodic prominence (lexical stress and pitch accents only) to identify which important contexts we potentially lack.

6. Conclusions

Our entry to the Blizzard Challenge 2009 incorporated two independent modifications to HMM-based speech synthesis:

- An implementation of Liljencrants-Fant (LF) glottal source model
- A label method for generating context-dependent phonemes capable of realising prosodic prominence including emphasis.

Our submitted systems are in the initial stages of development and participating in the Blizzard Challenge gave us valuable information how they compared with the Benchmark and best speech synthesisers and will guide further developments of our systems.

7. Acknowledgements

1st and 2nd author is supported by Marie Curie Early Stage Training Site EdSST (MEST-CT-2005-020568).

This work has made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF). (http://www.ecdf.ed.ac.uk/).

8. References

[1] V. Karaiskos, S. King, R. Clark, and C. Mayo, "The blizzard challenge 2008," in *The Blizzard Challenge*, Brisbane, Australia, 2008.

[2] 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, 2007.

[3] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of repetitive structure in sounds," *Speech Communication*, vol. 27, pp. 187–207, 1999.

[4] H. Kawahara, H. Katayose, and R. Cheveigné, Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of f0 and periodicity," in *Proc. Eurospeech*, 1999, pp. 2781–2784.

[5] D. Talkin, "Voicing epoch determination with dynamic programming," *J. of the Acoustical Society of America*, vol. 85, supplement 1, 1989.

[6] D. Talkin and J. Rowley, "Pitch synchronous analysis and synthesis for tts systems," in *Proc. of the ESCA Workshop on Speech synthesis*, C. ed. Benoit, Ed., Gieres, France, 1990.

[7] K. Tokuda, H. Zen, and A. Black, "An HMM-based speech synthesis system applied to english," in *Proc. of 2002 IEEE SSW*, 2002.

[8] H. Zen, K. Tokuda, and T. Kitamura, "An introduction of trajectory model into HMM-based speech synthesis," in *Proc. of 5th ISCA Speech Synthesis Workshop*, 2004.

[9] J. Yamagishi, H. Zen, T. Toda, and K. Tokuda, "Speaker-independent hmm-based speech synthesis system – hts-2007 system for the blizzard challenge 2007," in *BLZ3-2007*, Bonn, Germany, 2007.

[10] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden markov models based on multi-space probability distribution for pitch pattern modelling," in *Proc. ICASSP*, 1999, pp. 229–231.

[11] J. Odell, "The use of context in large vocabulary speech recognition," Ph.D. dissertation, Cambridge University, 1995, phD Thesis.

[12] R. Maia, T. Toda, H. Zen, Y. Nankaku, and K. Tokuda, "A trainable excitation model for hmm-based speech synthesis," in *8th Annual Conference of the ISCA (Interspeech 2007)*, Antwerp, Belgium, Sep. 2007.

[13] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, "Hmm-based finnish text-to-speech system utilizing glottal inverse filtering," in *Proc. Interspeech*, Brisbane, Australia, 2008.

[14] G. Fant, "The voice source in connected speech," *Speech Communication*, vol. 22, pp. 125–139, 1997.

[15] J. Cabral, S. Renals, K. Richmond, and J. Yamagishi, "Towards an improved modelling of the glottal source in statistical parametric speech synthesis," in *Proc. of the 6th ISCA Workshop on Speech Synthesis*, Bonn, Germany, 2007.

[16] ——, "Hmm-based speech synthesis with an acoustic glottal source model," in *1st Young Researchers Workshop on Speech Synthesis*, Dublin, Ireland, 2008.

[17] D. Marquardt, "An algorithm for least squares estimation of nonlinear parameters," *SIAM Journal on Applied Mathematics*, vol. 11, pp. 431/226–441, 1963.

[18] E. Krahmer and M. Swerts, "On the alleged existence of contrastive accents," *Speech Communication*, vol. 34, no. 4, 2001.

[19] S. Pan, K. McKeown, and J. Hirschberg, "Exploring features from natural language generation for prosody modelling," *Computer, Speech and Language*, vol. 16, pp. 457–490, 2002.

[20] J. Yuan, J. Brenier, and D. Jurafsky, "Pitch accent prediction: Effects of genre and speaker," in *Proc. Interspeech*, Lisbon, Portugal, 2005.

[21] L. Badino, J. Andersson, J. Yamagishi, and R. Clark, "Identification of contrast and its emphatic realisation in hmm based speech synthesis," in *Proc. Interspeech*, Brighton, UK, 2009.

[22] V. Strom, A. Nenkova, R. Clark, Y. Vazquez-Alvarez, J. Brenier, S. King, and D. Jurafsky, "Modelling prominence and emphasis improves unit-selection synthesis," in *Interspeech*, Antwerp, Belgium, 2007, pp. 1282–1285.

[23] P. Scholtz, A. Visagie, and J. du Preez, "Statistical speech synthesis for the blizzard challenge 2008," in *Proc. Blizzard Challenge 2008*, Brisbane, Australia, 2008.

[24] R. Maia, J. Ni, S. Sakai, T. Toda, K. Tokuda, T. Shimizu, and S. Nakamura, "The NICT/ATR speech synthesis system for the blizzard challenge 2008," in *Proc. Blizzard Challenge 2008*, 2008.