PREDICTION OF CREAKY VOICE FROM CONTEXTUAL FACTORS

[1]*Thomas Drugman,* [2]*John Kane,* [3]*Tuomo Raitio,* [2]*Christer Gobl*

[1]TCTS Lab - University of Mons, Belgium
[2]Phonetics and Speech Laboratory, School of Linguistic, Speech and Communication Sciences, Trinity College Dublin, Ireland
[3]Aalto University, Department of Signal Processing and Acoustics, Espoo, Finland

ABSTRACT

Creaky voice, also referred to as vocal fry, is a voice quality frequently produced in many languages, in both read and conversational speech. In order to enhance the naturalness of speech synthesisers, these latter should be able to generate speech in all its expressive diversity. This includes a proper use of creaky voice. The goal of this paper is two-fold. Firstly we analyse how contextual factors can be informative for the prediction of creaky use. It is observed that a few contextual factors related to speech production preceding a silence or a pause are of particular interest. This study validates that creaky voice plays a crucial syntactic role, allowing for a better structuring of phrases. In a second experiment, we investigate the prediction of creakiness from contextual factors based on HMMs. Four methods are compared on a US English and a Finnish speaker. It is shown that the best prediction technique achieves a promising performance comparable to what is carried out with the creaky detection algorithm on which HMMs were trained.

Index Terms— Creaky voice, Speech Synthesis, Expressive Speech, Contextual Factors

1. INTRODUCTION

Creaky voice, also called vocal fry or laryngealisation, is a voice quality brought about by a distinctive non-modal phonation type involving low-frequency vocal fold vibration. The temporal periodicity of this is often highly irregular and secondary laryngeal excitations are also common. The perceptual consequence can be described as *"a rough quality with the sensation of additional impulses"* [1]. For a description of the physiological and acoustic characteristics of creaky voice the reader can refer to [1]-[5]. Although produced by speakers involuntarily, various systematic usages of creaky voice have been reported. For instance, creaky voice has been observed as a phrase boundary marker in American English [6]. Another study investigated the use of creaky voice as a turn-yielding mechanism in Finnish [7]. The relevance of creaky voice for hesitations has been examined [8] as well its usage in portraying social status [9]. Creaky voice is also known to be important for communicating attitude and affective states [10].

Some of our previous work on creaky voice involved developing methods for automatic detection [11, 5]. Further work by the present authors was concerned with developing an excitation model of creaky production capable of providing a natural rendering of the voice quality [12]. One major application of this line of research is incorporating creaky voice in a statistical speech synthesis system. There are several reasons why this is desirable. Firstly, speakers often use creaky voice sentence finally, as well as in other positions, in the read speech used for developing text-to-speech (TTS) systems. For such speakers, providing the proper mechanisms for modelling creaky voice will inevitably improve the naturalness of the synthesis [14]. Furthermore, as creaky voice is frequently adopted in lively story-telling and natural interactive conversation, incorporating it may contribute significantly to the development of expressive synthesis. A major hurdle in the direction of this goal is determining where in the synthesised speech the creaky voice part should lie.

To address this, the present paper looks to investigate whether the creaky voice regions can be predicted solely from contextual factors (e.g., phoneme identity, position of current syllable, etc.). More specifically, we investigate whether these contextual factors can be applied within the Hidden Markov Model (HMM) framework used for statistical speech synthesis, in predicting the location of creaky regions. The benefit of such a prediction would be that creaky usage could be automatically determined from the input text, thereby allowing the use of a separate modelling for these regions [12]. Although previous work has been carried out on detecting creaky voice [5, 1], to the best of our knowledge previous research to date has yet to focus on the prediction of voice quality from contextual factors. However, somewhat related research was carried out in [13] where the authors investigated correlates of creaky voice in conversational speech, using latent semantic analysis. Note that the present initial study is limited to the prediction of creaky voice regions for read speech used for TTS development. The prediction of creaky voice regions from expressive speech and natural conversation will be the subject of a future study.

The paper is organised as follows. The speech data we used is introduced in Section 2. Section 3 describes the method for automatically detecting creaky voice from the speech waveform. In Section 4 we present the contextual information used in the prediction method and inspect which factors are most relevant for the prediction method, and to which extent. The prediction of creaky voice using this contextual information is described in Section 5, including the various approaches investigated as well as the assessment of the prediction performance. Finally, Section 6 concludes the paper.

2. SPEECH DATA

The speech data we use in the present study are two databases recorded for the purpose of developing text-to-speech (TTS) synthe-

This research was supported by FNRS, the Science Foundation Ireland Grant 09/IN.1/I2631 (FASTNET), the Irish Department of Arts, Heritage and the Gaeltacht (ABAIR project), the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n 287678, Academy of Finland (projects 135003, 256961), and MIDE UI-ART.The authors would like to thank Antti Suni from University of Helsinki for his help.

sis. The first is 1131 sentences produced by an American English male (labelled BDL) recorded for the ARCTIC database [15]. The second is 692 sentences read by a Finnish male (labelled MV) [16]. 100 sentences from each corpus are held out to be used solely for testing. In order to have reference data to evaluate the proposed prediction methods, human annotation of these test sentences was carried out by two of the present authors. In order to be consistent with previous studies we adopt the same annotation approach used in [1]. This involves basing the binary annotation of creaky voice on the auditory criterion: "a rough quality with the sensation of additional impulses", and allowing the annotation to be also visually guided through the use of speech waveforms, spectrograms and fundamental frequency (f_0) contours.

3. CREAKY VOICE DETECTION

Since the experiments conducted in this paper are based on statistical methods, a sufficiently large amount of annotated data is required. As manually labelling entire corpora (or even a sufficient proportion) is practically infeasible, a technique for the automatic detection of creaky voice is needed. This is carried out in the present study using the algorithm described in [5] which was shown to clearly outperform state-of-the-art approaches. The algorithm involves the use of two acoustic features which characterise two different aspects of the creaky excitation. These features are used as part of a decision tree classifier and a brief outline of the algorithm is now given.

The first acoustic feature (originally presented in [11]) involves the use of two resonators both with a centre frequency set according to the speakers mean f_0. The input to both resonators is the Linear Prediction (LP) residual. One resonator is set with a large bandwidth in order to provide reasonable f_0 values in creaky regions. The second resonator is set with a narrow bandwidth. A combination of the richer harmonicity brought about by the characteristic secondary excitations combined with the considerably lower f_0 in creaky regions results in a stronger second harmonic (H2) compared to the first (H1). Note that the harmonics are located using the f_0 value derived from the output of the first resonator. The first acoustic feature is, hence, H2-H1 calculated for the second resonator output. The second feature is calculated using a rectangular window centred on peaks in the first resonator output. Using a window twice the length of the shortest expected creaky pulse (2×15 ms, as determined following previous psycho-acoustic experiments) the prominence of the centre peak is measured and compared to other peaks within the window. The prominent residual peaks in creaky regions combined with long glottal pulse lengths ensure this *residual peak prominence* feature is high for creaky regions and low for other speech regions.

These two acoustic parameters are used as input features to a decision tree classifier which was trained according to the description given in [5]. Binary decisions of creaky use are derived by applying a threshold on the posterior probability provided by the decision tree. For the post-processing, zero-crossing rate is used to help remove misdetections in silent and unvoiced regions. Also, overly short regions are removed and adjacent detected regions are merged. One can refer to the original publication [5] for a comprehensive description.

4. ANALYSIS OF CONTEXTUAL INFORMATION RELATED TO CREAKY VOICE

The goal of this section is to investigate which contextual factors are the most relevant to predict creaky usage, and to quantify to what extent they can be useful for this purpose. If successful, this analysis would imply that contextual factors, on the present data, carry enough information to allow effective prediction. In order to achieve this, the predictability power of each contextual factor is assessed based on its Mutual Information (MI) with the creaky use decisions. Let us denote F_i the random variable corresponding to a particular contextual factor i with possible discrete values f_i, and C the binary decision indicating whether creak is used ($c = 1$) or not ($c = 0$). The mutual information $I(F_i; C)$ between F_i and C is defined as [17]:

$$I(F_i; C) = \sum_{f_i} \sum_{c} p(f_i, c) \log_2 \frac{p(f_i, c)}{p(f_i)p(c)}, \quad (1)$$

and is the amount of information (in bits) that F_i conveys about C. The amount of uncertainty on the determination of C is its entropy:

$$H(C) = -\sum_{c} p(c) \log_2 p(c). \quad (2)$$

Using these measures, the predictability power of a given contextual factor can be estimated using the normalised mutual information $\frac{I(F_i;C)}{H(C)}$, i.e. the proportion of relevant information it conveys over the creaky usage. Since MI computation involves the estimation of probability density functions, a sufficient amount of data is required. Therefore the assessment of the relevance of contextual factors is achieved on the whole datasets considering the automatic detections (as explained in Section 3) as the creaky use decisions C, rather than using only the 100 manually-annotated files. Since the creaky detection performance was shown in [5] to be relatively high (and particularly for the two speakers considered in this study), making such an approximation is acceptable and will not affect the validity of our findings.

This experiment is run on both BDL and MV. For BDL (US English), we consider the standard complete list of 53 contextual factors used in the HTS implementation [18, 19]. Among these, only 13 are found to have interesting normalised MI values higher than 15%, while all others do not exceed 7%. Inspecting these 13 contextual factors reveals that they are closely related with creaky use at the end of a sentence or a word group. For example, the most relevant factor turns out to be the position of the current syllable in the current phrase (counting backward), which reach a MI of 23.3%. The histogram of creaky usage for this contextual factor on BDL is displayed in Figure 1. It can be observed that creak is used for around 40% of the last syllables in a phrase. Further analysis highlights that BDL produced creaky voice for about 60% of the two phones preceding a pause.

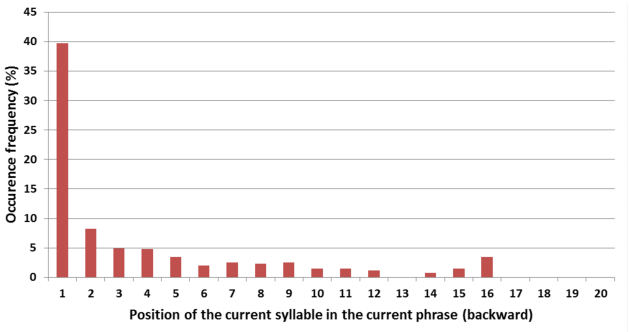

Fig. 1. *Conditional probability, for speaker BDL, of creaky use knowing the position of the current syllable in the phrase (counting backward).*

For the Finnish speaker, MV, we use a total of 66 contextual factors, as described in [20]. The trends seen for BDL appear even more pronounced for this speaker. More precisely, only 6 factors clearly emerge with values of normalised MI ranging from 22.9% to 32.1%, while others barely reach 10%. As it is the case for BDL, the most informative factor is the position of current syllable in the phrase counting backward. Again, the most discriminative contextual factors are noticed to be linked with a creaky usage at the end of a group of words, preceding a pause. It can therefore be reasonably hypothesised that the production of creaky voice, in the present data, has a crucial syntactic role, allowing a better structuring of phrases which consequently facilitates the listener's understanding. This perhaps corroborates previous findings on the segmentation of speech using creaky voice [6]. As a conclusion, contextual factors convey, on our two TTS speakers, a relatively high amount of relevant information to predict creaky usage. However, observation of these trends alone are not sufficient to provide robust prediction due to the complex interactions of contextual factors relevant to creaky voice.

5. CREAKY VOICE PREDICTION

This section describes the prediction of creaky voice using the HMM-based approach. This idea stems from the framework of statistical parametric speech synthesis [21] in which sequential speech data is modelled with context-dependent HMMs. In the following, HMM-based parameter training and generation is first illustrated after which the creaky voice prediction based on the approach is described.

5.1. HMM-based parameter training and generation

In statistical speech synthesis [21], speech is modelled in a parameterised form, consisting of features such as f_0, energy, and spectrum. Context dependent HMMs are used as acoustic units for modeling the sequential speech features. In each HMM state, each parameter is modelled by a Gaussian distribution and a diagonal covariance matrix. The duration of the state is modelled by a Gaussian distribution with scalar variance. Speech features are individually modelled in continuous probability distribution (CD) streams except f_0, which is modelled in a multi-space probability distribution (MSD) stream [24] in order to model a mixture sequence of continuous real numbers for voiced regions and symbol strings for unvoiced regions.

The training of the system begins with estimating the parameters statistics for monophone HMMs based on phonetic labels having time information. Monophone HMMs are then converted into context-dependent HMMs describing the characteristics of the phonemes in a specific context. The model parameters are then re-estimated, and decision-tree-based context clustering is applied in order to tie the model parameters of the HMMs at each leaf node of the decision trees. Finally, the clustered context dependent HMMs are re-estimated once more.

In parameter generation, an input text is first transformed into a sequence of context-dependent phoneme labels, and a sentence HMM corresponding to the label sequence is constructed by concatenating the context-dependent HMMs. Then, speech feature trajectories are statistically generated from the sentence HMMs. Trajectories are generated considering the global variance (GV) of the original training data [23]. The most common platform for statistical parametric speech synthesis is the freely available tool HTS [18, 19].

5.2. HMM-based prediction methods

In this study, the HMM-based approach, described in Section 5.1, is used for predicting creaky voice from contextual information. Creaky features, described in Section 3, are first trained as normal speech features in HTS. In addition to the creaky features, conventional features, namely f_0 and spectral coefficients, extracted with the GlottHMM synthesiser [22], are used for enabling proper alignment for the training. After the training, the creaky features are generated from labels aligned with the original speech samples, and finally the prediction of creaky use is carried out based on the detection method explained in Section 3. The process is fully automatic and reproducible with any other new voice provided that there is a sufficiently large amount of data. Although f_0 and spectrum are used in the parameter training for alignment, the prediction of creaky voice is made independently from those parameters (except the method using the MSD stream).

Four different prediction approaches based on the detection workflow depicted in Section 3 are experimented with:

- **PredictedFeat**: Modelling of the two creaky features in individual CD streams, after which the prediction is drawn from the decision tree used in Section 3
- **Binary**: Modelling of the binary creaky decision in a CD stream
- **Binary MSD**: Modelling of the binary creaky decision in a MSD stream that is aligned with f_0
- **PosteriorProb**: CD stream modelling of the posterior probability given by the detection algorithm (see Section 3)

The *PredictedFeat* method is an obvious choice since it directly models the speech features that are shown to be highly correlated with creaky voice. After the generation of the two creaky features, the creaky boundary determination is made according to the binary decision tree as detailed in Section 3. The methods *Binary* and *Binary MSD* both model the binary creaky decision that is the output of the detection algorithm. In the *Binary* technique, the binary decision is modelled in a CD stream, whereas in *Binary MSD* the unvoiced ($f_0 = 0$) segments are left undefined and voiced sections are defined by the given binary decision from the detection algorithm. The MSD-based method ensures that creaky voice cannot occur in unvoiced sections. With the assumption that f_0 is correctly modelled, the MSD-based method provides no false alarms in unvoiced regions. In the *PosteriorProb* approach, the posterior probability of creaky use provided by the binary decision tree (as explained in Section 3) without thresholding is modelled in a CD stream.

5.3. Assessing prediction performance

For each speaker and method, training is carried out on the whole database except the 100 manually-annotated sentences that are used solely for testing. To assess the performance of the prediction methods, we calculate evaluation metrics at both the event level and the frame level. For the event level, we used the number of hits (i.e. some part of a reference creak region was correctly detected), misses (i.e. for a reference creak region no positive detection was made) and false alarms (i.e. within a detected creak region there was no reference creak). At the frame level we use the metrics True Positive Rate (TPR, as known as recall) and False Positive Rate (FPR). We also use the F1 score which combines true positives (TPs), false positives (FPs) and false negatives (FNs) into one single metric. This metric is particularly useful when analysing skewed datasets where the feature being considered has a rather sparse occurrence. The metric is bound between 0 and 1, with 1 indicating perfect detection:

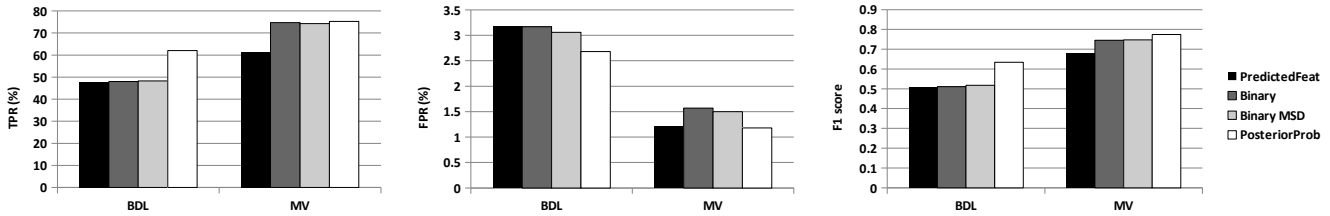

Fig. 2. *Results of creaky prediction at the frame level for the four compared techniques.*

$$F1 = \frac{2 \cdot \text{TPs}}{2 \cdot \text{TPs} + \text{FPs} + \text{FNs}} \in [0, 1] \quad (3)$$

5.4. Results

Results of creaky prediction ability at the frame level obtained with the four compared methods described in Section 5.2 are displayed in Figure 2. Across both speakers and all metrics, it turns out that the *PosteriorProb* method gives the best performance. This is especially striking for speaker BDL for whom the performance of the four techniques is lower compared to what is achieved for MV. For the BDL voice, the first three methods provide sensibly comparable results with a F1 score around 0.51, while *PosteriorProb* clearly emerges with a value of 0.635 for this measure. Although to a lesser extent, the superiority of *PosteriorProb* is also observed for speaker MV. In that case, using the predicted features (*PredictedFeat*) is noticed to be the worst choice as it leads to a much lower TPR. Comparing the utility of opting for a MSD stream instead of a CD one (and vice versa) in order to model the binary creaky decisions, it is observed on our two speakers that both approaches carry out very similar results and therefore no preference can be given of favour of one of them. This also validates that the voicing decision extracted in GlottHMM [22] provides reasonable estimates in creaky regions for these two speakers. It is nonetheless worth emphasising that our previous studies [11, 5] reported that some other speakers produce creakiness with much less periodic patterns. Thus it is important to ensure that the used f_0 estimation method gives reliable f_0 estimates in creaky regions.

Results at the event level are shown in Table 1. Conclusions drawn from Figure 2 about the comparative performance across the four methods remain valid here. It is seen that *PosteriorProb* produces the lowest number of misses (and consequently the best hit rate) compared to the 3 other approaches. This is at the expense of an increase of the amount of false alarms. Contrastingly, *Binary MSD* leads to the fewest false alarms, closely followed by *Binary*.

Table 1. *Summary of the performance at the event level across the 2 speakers for the four prediction methods. Best results are highlighted in bold.*

Database	Method	Misses	FAs	Hits
BDL	PredictedFeat	66	24	98
	Binary	68	19	96
	Binary MSD	68	**17**	96
	PosteriorProb	**40**	37	**124**
MV	PredictedFeat	54	29	109
	Binary	46	25	117
	Binary MSD	47	**24**	116
	PosteriorProb	**28**	39	**135**

Overall, *PosteriorProb* generates the lowest number of errors with a good trade-off between misses and false alarms.

The best results achieved by prediction from contextual factors can be compared with those obtained by the creaky detection technique which was used to train the prediction techniques. This latter method was shown in [5] to reach F1 scores of 0.724 for BDL and 0.81 for MV, against respectively 0.634 and 0.77 after prediction. It can then be concluded from our experiment that contextual factors are sufficiently informative to predict the creaky use with a performance comparable (albeit slightly worse) to the detection technique with which the prediction method was trained. This is achieved thanks to the ability of the HMM-based system to properly model the trajectory of the creaky posterior probability.

6. CONCLUSION

This paper investigates the possibility of using contextual factors to predict the production of creaky voice. Since such a statistical study requires analysis of a sufficiently large amount of data, our approach is based on an efficient creaky detection technique we proposed in a previous work. Therefore the whole process is fully automatic. Experiments are divided into two parts and are carried out on both a US English and a Finnish speaker. Firstly we investigate how contextual information is related to the use of creaky voice. A set of contextual factors linked to speech production preceding a silence or a pause appears to be highly relevant, leading to normalised mutual information values up to 32%. This suggests that, in the present data, vocal fry has among others a syntactic role by making a better delimitation of groups of words and by consequently making phrase segmentation easier. In the second experiment, four methods are proposed to predict the use of creaky voice based on Hidden Markov Models. It is shown that modelling the posterior probability given by the detection algorithm in a continuous stream leads to the best results across all metrics and for both speakers. This technique achieves performance scores comparable to the determination rates obtained by the detection method on which it is trained. This therefore confirms the relatively successful ability of HMMs to predict the creaky usage only from contextual factors. This finding is, of course, limited to the speakers and languages in the present study. Other speakers and languages may use creaky voice in different ways to the observations in the present data. Nevertheless, provided there are the necessary contextual factors to characterise the usage, we believe that the present method continue to be effective. Ultimately, the suitability of this method for speech synthesis requires a formal perceptual evaluation and this is an immediate objective of our future work. We also intend to investigate whether a similar approach could be applied for predicting creaky voice in expressive speech and natural conversation. For this to be effective a richer set of contextual factors will be required which cover issues such as discourse functions, affective states etc.

7. REFERENCES

[1] Ishi, C., Sakakibara, K., Ishiguro, H., Hagita, N., "A method for automatic detection of vocal fry", IEEE Transactions on Audio, Speech and Language Processing, 16 (1), pp. 47-56, 2008.

[2] Laver, J. "The Phonetic Description of Voice Quality", Cambridge University Press, 1980.

[3] Gobl, C., Ní Chasaide, A., "Acoustic characteristics of voice quality", Speech Communication, 11, pp. 481-490, 1992.

[4] Blomgren, M., Chen, Y., Ng, M., Gilbert, H. "Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers", J. Acoust. Soc. Am., 103(5), pp. 2649-2658, 1998.

[5] Kane, J., Drugman, T., Gobl, C., "Improved automatic detection of creak", Computer Speech and Language [Accepted].

[6] Surana, K., Slifka, J. "Is irregular phonation a reliable cue towards the segmentation of continuous speech in American English", Proceedings of Speech Prosody, Dresden, Germany, Paper 177, 2006.

[7] Ogden, R., "Turn transition, creak and glottal stop in Finnish talk-in-interaction", Journal of the Internation Phonetic Association, 31 (1), pp. 139-152, 2001.

[8] Carlson, R., Gustafson, K., Strangert, E., "Prosodic Cues for Hesitation," in Proceedings of Fonetik 2006 , pp. 2124, 2006.

[9] Yuasa, I. K., Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile American women?, American Speech 85(3), 315337, 2010.

[10] Yanushevskaya, I., Gobl, C., Ní Chasaide, A., "Voice parameter dynamics in portrayed emotions", Proceedings of Maveba, Florence, 2124, 2009.

[11] Drugman, T., Kane, J., Gobl, C. "Resonator-based Creaky Voice Detection", Proceedings of Interspeech, Portland, Oregon, USA, 2012.

[12] Drugman, T., Kane, J., Gobl, C. "Modeling the creaky excitation for parametric speech synthesis", Proceedings of Interspeech, Portland, Oregon, USA, 2012.

[13] Zhuang, X., Hasegawa-Johnson, M., "Towards interpretation of creakiness in switchboard", in Proceedings of Speech Prosody, 2008.

[14] Silén, H., Helander, E., Nurminen, J., Gabbouj, M., "Parameterization of vocal fry in HMM-based speech synthesis", in Proceedings of Interspeech, Brighton, UK, pp. 1775-1778, 2009.

[15] [Online], "CMU ARCTIC speech synthesis databases", *http://festvox.org/cmu arctic/*.

[16] Vainio, M., "Artificial neural network based prosody models for Finnish text-to-speech synthesis," Ph.D. dissertation, University of Helsinki, Finland, 2001.

[17] Cover, T., Thomas, J., "Elements of Information Theory", Wiley Series in Telecommunications, New York, 1991.

[18] Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. and Tokuda, K. "The HMM-based speech synthesis system (HTS) version 2.0", in Sixth ISCA Workshop on Speech Synthesis, Aug. 2007, pp. 294–299.

[19] [Online] "HMM-based speech synthesis system", *http://hts.sp.nitech.ac.jp*.

[20] Vainio, M., Suni, A. and Sirjola, P. "Accent and prominence in Finnish speech synthesis", in Proceedings of the 10th International Conference on Speech and Computer (Specom 2005), G. Kokkinakis, N. Fakotakis, E. Dermatos, and R. Potapova, Eds. University of Patras, Greece, Oct. 2005, pp. 309–312.

[21] Zen, H.,Tokuda, K. and Black, A. "Statistical parametric speech synthesis", Speech Communication, vol. 51, no. 11, pp. 1039–1064, 2009.

[22] Raitio, T., Suni, A., Yamagishi, J., Pulakka, H., Nurminen, J., Vainio, M. and Alku, P. "HMM-Based Speech Synthesis Utilizing Glottal Inverse Filtering", in IEEE Trans. on Audio, Speech, and Language Processing vol. 19, no. 1, pp. 153–165, 2011.

[23] Toda, T. and Tokuda, K. "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis", IEICE Trans. Inf. & Syst., vol. E90-D, no. 5, pp. 816–824, May 2007.

[24] Tokuda, K., Masuko, T., Miyazaki, N. and Kobayashi, T. "Multi-space probability distribution HMM", IEICE Trans. Inf. & Syst., vol. E85-D, no. 3, pp. 455–464, Mar. 2002.