HMM-based synthesis of creaky voice

Raitio, T., Kane, J., Drugman, T., & Gobl, C. (2013). HMM-based synthesis of creaky voice. In INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association. (pp. 2316-2320). ISCA-INST SPEECH COMMUNICATION ASSOC.

HMM-based synthesis of creaky voice

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

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

tuomo.raitio@aalto.fi, kanejo@tcd.ie, thomas.drugman@umons.ac.be, cegobl@tcd.ie

Abstract

Creaky voice, also referred to as vocal fry, is a voice quality frequently produced in many languages, in both read and conversational speech. To enhance the naturalness of speech synthesis, these latter should be able to generate speech in all its expressive diversity, including creaky voice. The present study looks to exploit our recent developments, including creaky voice detection, prediction of creaky voice from context, and rendering of the creaky excitation, into a fully functioning and automatic HMM-based synthesis system. HMM-based synthetic creaky voices are built and evaluated in subjective listening tests, which show that the best synthetic creaky voices are rated more natural and more creaky compared to a conventional voice. A non-creaky voice is also successfully transformed to use creak by modifying the F0 contour and excitation of the predicted creaky parts. The transformed voice is rated equal in terms of naturalness and clearly more creaky compared to the original voice.
Index Terms: speech synthesis, creaky voice, contextual factors, F0 estimation, excitation modeling

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 creak is often highly irregular and secondary laryngeal excitations are also common. The perceptual consequence of this 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 can be found e.g. in [1]–[5]. Although creak is 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 of creak [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]. Also the prediction of creaky voice from contextual factors was investigated in [13], which enables automatic determination of the creaky usage from the input text. One obvious application of this line of research is incorporating creaky voice in a statistical parametric speech synthesis system. There are several reasons why this is desirable. Firstly, many speakers use creaky voice 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 creak will also contribute significantly to the development of expressive speech synthesis.

In this paper, statistical parametric speech synthesis of creaky voice is investigated. First, Sec. 2 describes speech data used in the study and Sec. 3 describes methods required for successful analysis of creaky voice: creaky voice detection and fundamental frequency (F0) estimation. In Sec. 4, hidden Markov model (HMM) based synthesis of creaky voice is experimented: synthetic creaky voices are built and evaluated in subjective listening tests in terms of naturalness and creakiness. Adding creaky voice to a non-creaky speaker is experimented in Sec. 5, and finally Sec. 6 summarises the current findings.

2. Speech data

The speech data used in the present study consist of three databases recorded for the purpose of developing TTS synthesis. 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]. The first two speakers use creaky voice in the recordings. The third corpus contains 1138 utterances spoken by a Scottish English male (labelled AWB) who does not generally exhibit creaky voice. This corpus is thus used in experiments of adding a creaky voice for a non-creaky speaker.

Additionally, conversational speech data is used for assessing the performance of F0 and voicing estimation algorithms in creaky voice regions. This consists of conversational speech data recorded from 7 speakers in a range of conditions, and covering a set of languages (English, Japanese and Swedish). A full description of these conversational speech databases is given in [5]. Note that an additional TTS database of a female Finnish speaker was included to evaluate the F0 algorithms (see [14]).

All of the conversational data, as well as 100 sentences from the TTS databases (which was used as test data), were hand-labelled for creaky voice (the annotation procedure is outlined in [5]). Note that it is not generally possible to obtain objective reference annotation for creaky voice.

3. Analysis of creaky voice

During the production of creaky voice, the glottal behaviour is dramatically modified. The physiological settings [17] bring about acoustic characteristics which are quite distinct from

modal voice. As a result, proper automatic analysis of creaky voice then requires specific tools for i) the accurate detection of creaky voice parts and ii) the accurate F0 estimation in difficult creaky voice parts. In the following, these methods for creaky voice analysis are described.

3.1. Creaky voice detection

In order to have proper treatment of the distinctive acoustic characteristics of creaky voice in a speech synthesis system it is essential to have annotation of creaky voice regions in a given corpus. Hand-annotation of large corpora is, of course, extremely time-consuming and besides, in order to have a fully automatic and reproducible synthesis development method, automatic detection of creaky voice is required. In this study we utilise a recently developed detection algorithm [5] (which built on initial work in [11]). The algorithm involves the use of two features derived from the Linear Prediction (LP) residual which have been tailor-designed to characterise aspects of the creaky excitation. These features are used as inputs to a binary decision tree classifier, which outputs a posterior probability of the occurrence of creaky voice. This contour can be thresholded to obtain a binary creaky decision. The detection method was trained on a range of speech data including read speech recorded for TTS development as well as a range of conversational speech databases recorded under different conditions.

3.2. F0 estimation

To develop a synthesis system with effective rendering of creaky voice, one must use an F0 tracker capable of outputting meaningful values in these regions. However, due to the very low F0 and often highly irregular periodicity of creaky voice many F0 trackers either output spurious values or incorrectly determine the region to be unvoiced. To decide on an appropriate F0 tracker for our synthesis approach, we first evaluated a range of state-of-the-art F0 algorithms:

- GlottHMM [18]
- SWIPE [19]
- RAPT [20]
- SPTK 3.1 cepstrum based pitch function [21]
- STRAIGHT TEMPO [22, 23])

These methods are assessed in terms of the extent to which they incorrectly determine creaky voice regions to be unvoiced. The methods were mostly used with their default settings, except that the frame length was set to 45 ms whenever possible to assist the F0 detection in low-pitch creaky sections. A range of speech data, previously hand-labelled for creaky voice (see Section 2), including 3 databases of read speech for TTS synthesis development as well as conversational speech data from 7 other speakers was used. For the TTS data (Figure 1, left panel) the GlottHMM method performs best in terms of not incorrectly determining creaky voice regions to be unvoiced, with SPTK also performing well. In general for the conversational data (Figure 1, right panel) performance is degraded somewhat. This is to a large extent due to lower quality recording conditions. Here SPTK performs best with GlottHMM the next best. Considering these findings we opted to use the GlottHMM F0 and voicing decision algorithm for our synthesis approach.

4. Synthesis of creaky voice

Synthesis of voice with creak requires i) the prediction of creaky parts from context and ii) the ability to render creaky excitation.

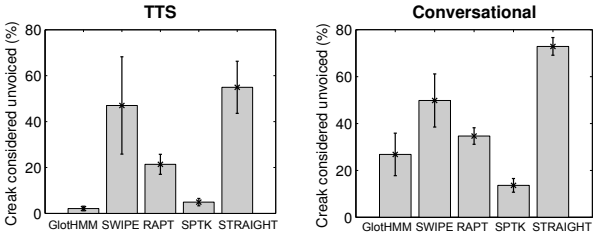

Figure 1: *Percentage of hand-labelled creaky voice regions incorrectly determined as unvoiced using 5 F0 tracking algorithms for TTS (left panel) and conversational (right panel) speech data. Data is displayed as means and standard errors.*

In our previous work, we have developed methods for creaky voice prediction from context [13] and rendering of creaky excitation [12]. However, these methods have not been utilised in a full TTS voice before. In the following, HMM-based synthetic voices with the ability to exhibit creak are created.

4.1. Prediction of creaky voice from context

To render creaky voice in appropriate parts in a sentence, creak must be predicted from input text. Although it is possible to have an external control over a creaky voice in speech synthesis, in a pure TTS application, creaky parts must be predicted from the only available source of information, the context of the input text. The process of the prediction begins with first detecting the existing creaky parts in the training corpus by a creaky voice detection algorithm (see Section 3.1). In this study, the algorithm in [5] is used, which provides a frame-wise probability of creak. This parameter is used as a feature in the HMM-training for determining if a segment is creaky or not [13]. More specifically, the parameter indicating the probability of creak is trained as an additional 1-dimensional feature along with other speech features, such as F0 and spectrum.

The contextual features according to which the creaky probability parameter is trained, are defined by the list of phonetic and linguistic information that is used for training a HMM-based synthetic voice. For BDL voice, the standard list of 53 contextual factors in the HTS implementation [24, 25] is used. For the Finnish speaker, MV, a total of 66 contextual factors are used, described in [26]. According to the study in [13], only a few of the contextual factors are useful in predicting creaky voice, and the useful factors are closely related with creaky use at the end of a sentence or a word group.

After the training, a statistical model (i.e. HMM system) is created that links the creaky probability with the contextual factors. In synthesis, the input text is fed into a front-end that extracts the contextual information according to the list of contextual features. This information is then used to generate a creaky probability trajectory from the trained statistical model. Investigations on this procedure in [13] indicate that the accuracy of the prediction of creaky voice from context is comparable to the accuracy of the creaky detection algorithm on which the HMM system was trained.

4.2. Rendering of creaky voice

As described in Section 1, the creaky excitation is dramatically different from the excitation of modal speech arising from certain distinctive physiological factors [17]. More precisely, the creaky excitation signal not only exhibits discontinuities at the glottal closure instants (as in modal speech [27]), but also displays secondary (and sometimes even tertiary) excitation peaks.

In [12], we have proposed an extension of the Deterministic plus Stochastic Model (DSM) [28] which integrates a proper modeling of these secondary excitation peaks. The resulting parametric vocoder was shown to provide a much better perceptual rendering of the creaky voice quality [12]. In the following, this vocoder will be used to enhance the creaky voice synthesis.

There are three crucial points to ensure correct perceptual creaky rendering. First, voicing decision method should be robust enough to deal with the acoustic characteristics of creaky voice. If this is not the case, the use of an unvoiced excitation in creaky segments will dramatically affect the quality of the produced voice. Secondly, the F0 estimation technique should provide tangible F0 trajectories even in creaky voice. The third criterion is a proper modeling of the creaky excitation which importantly differs from the excitation in modal speech.

4.3. Voice building

Creaky voices were built using the standard HTS method [24, 25] with the addition of 1-dimensional stream of creaky probability [13]. First, F0 was estimated with two methods: GlottHMM vocoder [18] and TEMPO [23]. SPTK 3.6 [29] was used to extract the spectrum of speech by 30th order mel-generalised cepstral analysis with $\alpha = 0.42$ and $\gamma = -1/3$ [30]. Generalised mel-cepstrum was then converted to line spectral frequencies (LSF) [31] for better parameter representation for HMM training. In synthesis, parameters were generated considering global variance [32] except for the spectrum. Creaky parts were determined according to the generated creaky voice probability. Excitation was generated using the DSM vocoder [28] with the extension that creaky parts were rendered with the creaky excitation waveform [12]. Finally, the excitation was filtered with the mel-generalised log spectral approximation (MGLSA) filter [33]. The following voices were built both for MV and BDL speakers using the previous methods:

1. Conventional (STRAIGHT F0)
2. Proposed (GlottHMM F0)
3. Proposed (GlottHMM F0 and creaky excitation)

4.4. Evaluation

To evaluate the three synthesis systems we carried out a subjective online two-part listening test. For the stimuli used in the listening test we randomly selected 20 sentences (synthesised using the 3 systems, as well as the natural speech utterance) from the 100 held-out test sentences of the American (BDL) and the Finnish speaker (MV). Note we included natural utterances as a check, but as participants rated these almost unanimously as completely natural we will not consider these in the results.

The first part was a standard mean-opinion score (MOS) style test, where participants rated the naturalness of synthesised stimuli on a scale of 1 to 5. 29 participants (22 of whom are engaged in speech research) carried out the first test. Participants were presented with 48 stimuli (i.e. 2 speakers with 6 sentences by the 4 systems). Note that the 6 sentences were randomly selected from the set of 20 for each participant, and stimuli were presented in a randomised order each time.

The second part was a pairwise preference test, where participants were presented with two synthesised stimuli and were required to indicate their preference of the two in terms of synthesising creaky voice effectively. Participants could also choose "no preference". Note that 3 of the 29 participants did not complete the preference test part. As some participants may not be familiar with the term creaky voice we included a range

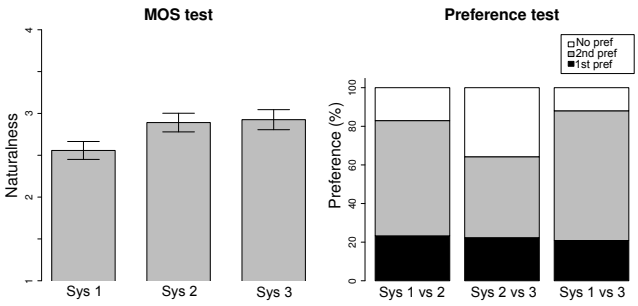

Figure 2: *Results from (a) MOS and (b) preference test. Data for (a) is displayed as means and 95% confidence intervals.*

of natural utterances as references with which to familiarise themselves with the voice quality. Participants were presented with 30 pairs of stimuli, 5 sentences for the 2 speakers and the 3 possible pairwise comparisons (i.e. 1 vs 2, 2 vs 3 and 1 vs 3). Note again that the 5 sentences were randomly selected from the set of 20 for each participant and pairs of stimuli presented were presented in a randomised order.

The results of the subjective evaluation are illustrated in Figure 2a. For the MOS test, a one-way ANOVA indicates a significant effect of system type on participant ratings [$F_{(2,1041)} = 12.52$, $p < 0.001$]. Pairwise comparison using Tukey's Honestly Significant Difference (HSD) test reveals both system 2 and 3 to have higher ($p < 0.001$) participant ratings than system 1. These findings indicate a clear improvement in naturalness when using the GlottHMM F0 tracker compared with STRAIGHT. However, participants did not notice any clear difference in overall naturalness if the creaky excitation was included or not.

For the preference test (illustrated in Figure 2b) participants clearly signalled a preference for system 2 compared to system 1 (60 % preference) and system 3 compared to system 1 (67 % preference). Around twice as many ratings favoured system 2 (42 % preference) to system 1 (22 % preference), however a large proportion of the ratings (36 %) indicated no preference for either. The findings here clearly show a lower preference for the synthesis system using STRAIGHT F0 in terms synthesising creaky voice. They also indicate a preference for the synthesis system using the creaky excitation (i.e. system 3) compared to the one without (i.e. system 2).

5. Adding creak for non-creaky speaker

The following four systems were assessed in terms of transforming a non-creaky voice (AWB) to a creaky one:

1. Baseline AWB
2. AWB with BDL creaky excitation
3. AWB with BDL F0 and BDL creaky excitation
4. AWB with F0 transformation and BDL creaky excitation

For system 1, a normal baseline voice was trained according to the description in Section 4.3, without creaky voice prediction and rendering. For system 2, creaky voice regions are predicted and a creaky excitation pulse from another speaker (i.e. BDL) is used to render the excitation. This, however, may impose problems since the combination of the original F0 curve and artificially added creaky excitation may not sound natural. System 3 uses F0 stream substitution and creaky excitation from another creaky speaker (BDL). Note that a similar approach of feature stream substitution has previously been shown to be effective for reconstruction of the timbre of individuals with degenera-

tive diseases using HMM-based synthesis with an average voice model [34]. The F0 curve is, hence, in line with creak, but with the cost that the prosody of the original speaker is affected by the substitution of another speaker's F0. Finally, system 4 tries to overcome this problem by transforming the original F0 curves so that they decline appropriately in the creaky regions. This is achieved by applying a data-driven transformation to the original F0 curve in the region preceding, around 500 ms (approx. 2 syllables), and including the creaky segment. This transformation is learnt from the analysis of F0 trajectories from a creaky speaker. More precisely, a set of F0 curves preceding a creaky segments are collected, converted to a logarithmic scale, normalised so that they start with a zero value, and then simply averaged. The original F0 trajectory of the non-creaky speaker is then transformed such that it matches the trends extracted from the creaky speaker, in the 500ms region preceding the predicted creaky region.

5.1. Evaluation

To assess the effectiveness of the creaky transformation as well as the overall naturalness of the synthetic utterances we carried out a further online subjective evaluation of the 4 systems described in previous section. 14 participants carried out the listening test where they were presented with 28 synthesised stimuli (i.e. 7 sentences for each of the 4 methods) and were required to rate the stimuli on two scales. The stimuli we used were again randomly selected 20 test sentences and the corresponding synthetic signals produced using the 4 systems. Participants were presented with 28 stimuli (the 4 system versions of 7 test sentences randomly selected from the 20 for each participant) in a randomised order. The first scale was a standard MOS scale, with naturalness rated on a score of 1 to 5. The second scale was also from 1 to 5, with 1 being *"does not sound like creaky voice"* and 5 being *"sounds exactly like creaky voice"*. Again reference samples of natural utterances containing creaky voice were given at the beginning of the test to allow participants to familiarise themselves with the voice quality. Note that in this test these references utterances were randomly selected (for each participant) from a set of utterances taken from the conversational data. This was done to avoid biasing participants to one particular form of creaky voice.

The results for the subjective evaluation of creaky transformation are presented in Figure 3. For the MOS, (panel a), a one-way ANOVA (with participant rating as the dependent variable) indicates significant effect of system type on the MOS naturalness score [$F_{(3,388)} = 2.93$, $p < 0.05$]. Pairwise comparison using Tukey's Honestly Significant Difference (HSD) test reveals that system 3 was rated as lower ($p < 0.05$) than system 1. However, there were no other significant pairwise differences. For the creaky scale, a one-way ANOVA again indicates (but in this case a more pronounced) significant effect of system type on participant ratings [$F_{(3,388)} = 33.43$, $p < 0.01$]. Tukey's HSD post-hoc test this time reveals a significant difference ($p < 0.001$) between system 2, 3 and 4 compared with system 1, with no other pairwise significant differences.

These findings demonstrate that systems 2–4 clearly achieve incorporation of creaky voice into the utterance of a non-creaky speaker. For system 3, which utilises F0 stream substitution, the altered prosody of the speaker brings about a degradation in naturalness, and is, hence, somewhat less effective. A further possibility for the degradation may be due to the higher spectral coefficients from the non-creaky speaker being unsuitable and, hence, the need for further feature substitution of these coefficients (as is done in [34]). Interestingly, the method with the highest mean creaky rating, which does not degrade the naturalness, is a relatively straightforward post-processing of the F0 contour (i.e. system 4).

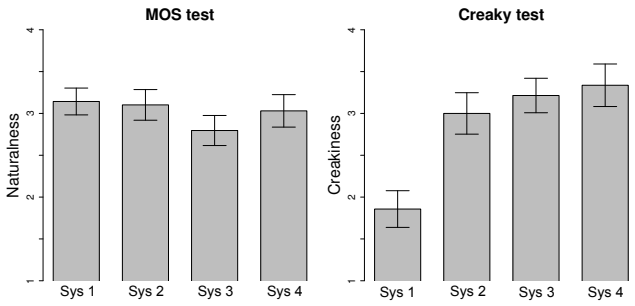

Figure 3: *Results from the MOS (a) and creaky preference (b)* ***transformation*** *subjective evaluations. Data displayed as means and 95 % confidence intervals.*

6. Conclusion

The goal of this paper was two-fold. First, we investigated methods for the HMM-based synthesis of creaky voice. Compared to the synthesis of modal voice, this purpose requires the development of specific and necessary modules: i) at analysis time, a robust pitch tracker which copes with the inherent less regular periodicity of creaky voice should be used, ii) at generation time, the segments where creaky voice should be used have to be predicted from contextual factors, iii) at synthesis time, a dedicated vocoder integrating the presence of secondary peaks in the creaky excitation should be used to allow a proper rendering of creaky voice. The inclusion of these modules into a HMM-based speech synthesiser was shown to provide a substantial improvement over the standard approach. Our subjective evaluation revealed a significant improvement in terms of naturalness, as well as a clear preference towards the proposed system. These experiments also highlighted the need to use appropriate creaky voice analysis tools.

The second goal of the paper was to investigate the possibility of applying voice transformation techniques so as to produce creaky voice by a speaker who initially only used modal speech. Three techniques were proposed for this purpose. Compared to the standard HMM-based speech synthesiser for such a speaker, these methods were shown to maintain the level naturalness (i.e. they did not introduce any artifacts) while they clearly induced a proper creaky rendering perceived by listeners. Interestingly, the best method for this purpose did not involve any statistical manipulation and could be used as a post-process in any (i.e. not necessarily statistical) speech synthesis method. There is of course the risk that different languages (in particular), but also possibly different dialects may have different systematic usage of creaky voice. Nevertheless, using a model of predicting creaky voice from an American English speaker applied to a Scottish English speaker was deemed to be effective.

7. Acknowledgements

This research was supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 287678, by FNRS, by the Science Foundation Ireland Grant 09/IN.1/I2631 (FASTNET), and by the Irish Department of Arts, Heritage and the Gaeltacht (ABAIR project).

8. References

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

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

[3] Gobl, C. and Ní Chasaide, A., "Acoustic characteristics of voice quality", Speech Commun., 11(4–5):481–490, 1992.

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

[5] Kane, J., Drugman, T. and Gobl, C., "Improved automatic detection of creak", Computer Speech & Language, 27(4):1028–1047, 2013.

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

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

[8] Carlson, R., Gustafson, K. and Strangert, E., "Prosodic Cues for Hesitation", Proc. of Fonetik, pp. 21–24, 2006.

[9] Yuasa, I. K., "Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile American women?", American Speech, 85(3):315–337, 2010.

[10] Yanushevskaya, I., Gobl, C. and Ní Chasaide, A., "Voice parameter dynamics in portrayed emotions", Proc. of the 6th International Workshop on Models and Analysis of Vocal Emissions for Biometrical Applications (MAVEBA), Florence, Italy, pp. 21–24, 2009.

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

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

[13] Drugman, T., Kane, J., Raitio, T. and Gobl, C., "Prediction of creaky voice from contextual factors", accepted for publication in Proc. ICASSP, Vancouver, Canada, 2013.

[14] Silén, H., Helander, E., Nurminen, J. and Gabbouj, M., "Parameterization of vocal fry in HMM-based speech synthesis", Proc. 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] Edmondson, J.A. and Esling, J.H., "The valves of the throat and their functioning in tone, vocal register and stress: laryngoscopic case studies", Phonology, 23(2):157–191, 2006.

[18] Raitio, T., Suni, A., Yamagishi, J., Pulakka, H., Nurminen, J., Vainio, M. and Alku, P., "HMM-based speech synthesis utilizing glottal inverse filtering", IEEE Trans. on Audio, Speech, and Lang. Proc., 19(1):153–165, 2011.

[19] Camacho, A., Harris, J. G., "A sawtooth waveform inspired pitch estimator for speech and music", J. Acoust. Soc. Am., 124(3):1638–1652, 2008.

[20] Talkin, D., "A robust algorithm for pitch tracking (RAPT)", W.B. Klein and K.K. Paliwal (Eds.), Speech Coding and Synthesis, New York, Elsevier, 1995.

[21] [Online] Speech Signal Processing Toolkit (SPTK) v. 3.1, *http://sourceforge.net/projects/sp-tk/*

[22] Kawahara, H, Katayose, H., de Cheveigné A. and Patterson R., "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity", Proc. Eurospeech, pp. 2781–2784, 1999.

[23] Kawahara, H., Masuda-Katsuse, I. and de Cheveigné, A., "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds", Speech Commun., 27(3–4):187–207, 1999.

[24] 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", Sixth ISCA Workshop on Speech Synthesis, pp. 294–299, 2007.

[25] [Online] HMM-based speech synthesis system (HTS), *http://hts.sp.nitech.ac.jp*

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

[27] Drugman, T., Thomas, M., Gudnason, J., Naylor, P. and Dutoit, T., "Detection of Glottal Closure Instants from Speech Signals: a Quantitative Review", IEEE Transactions on Audio, Speech and Language Processing, 20(3):994–1006, 2012.

[28] Drugman, T. and Dutoit, T., "The Deterministic plus Stochastic Model of the Residual Signal and its Applications", IEEE Transactions on Audio, Speech and Language Processing, 20(3):968–981, 2012.

[29] [Online] Speech Signal Processing Toolkit (SPTK) v. 3.6, *http://sourceforge.net/projects/sp-tk/*

[30] Tokuda, K., Kobayashi, T., Masuko, T., and Imai, S., "Mel-generalized cepstral analysis – a unified approach to speech spectral estimation", The 3rd International Conference on Spoken Language Processing (ICSLP), pp. 18–22, 1994.

[31] Soong, F.K. and Juang, B.-H., "Line spectrum pair (LSP) and speech data compression", Proc. ICASSP, vol. 9, pp. 37–40, 1984.

[32] Toda, T. and Tokuda, K., "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis", IEICE Trans. Inf. Syst. E90-D(5):816–824, 2007.

[33] Kobayashi, T., Imai, S. and Fukuda, T., "Mel generalized-log spectrum approximation (MGLSA) filter", Journal of IEICE, J68-A(6):610–611, 1985, (in Japanese).

[34] Veaux, C., Yamagishi, J. and King, S., "Using HMM-based speech synthesis to reconstruct the voice of individuals with degenerative speech disorders", Proc. Interspeech, Portland, Oregon, USA, 2012.