Automatic detection of creaky voice using epoch parameters

N. P. Narendra, K. Sreenivasa Rao

School of Information Technology, Indian Institute of Technology Kharagpur,
Kharagpur - 721302, West Bengal, India.
narendrasince1987@gmail.com, ksrao@iitkgp.ac.in

10.21437/Interspeech.2015-508

Abstract

This paper proposes a method based on epoch parameters for detection of creaky voice in speech signal. The epoch parameters characterizing the source of excitation considered in this work are number of epochs in a frame, strength of excitation of epochs and epoch intervals. Analysis of epoch parameters estimated from zero-frequency filtering method with different window sizes is carried out. Distinct variations in the epoch parameters are observed for modal and creaky voiced regions. Variances of epoch parameters are used as input features to train a neural network classifier for identifying creaky regions. Performance evaluation results indicate that the proposed method performs significantly better than the existing creaky detection methods on different speech databases.

Index Terms: Creaky voice, vocal fry, creaky detection, zero-frequency filtering, epoch parameters.

1. Introduction

Creaky voice or vocal fry refers to a voice quality characterized by a low and irregular rate of vocal fold vibration. Creaky voice is commonly produced in different speech modes such as read, conversational and expressive speech. Creaky voice has been observed at phrase boundary in American English [1] and Finnish [2]. It has been shown in [3] that the presence of creaky voice is important for communicating attitude and affective states. Creaky voice is frequently produced during hesitations [4] and its detection could be used to identify hesitations. Even though creaky voice occurs in different speech modes, most of the speech technology applications tend to ignore the creaky regions. The main reason for this is due to inability to accurately detect creaky regions [5].

Creaky voice produces dramatically different acoustic characteristics than that of modal voice. The most prominent acoustic features of creaky voice include: (i) long glottal pulse duration and as a result, little or no superposition of formant oscillations between successive glottal cycles, (ii) occurrence of secondary excitations and (iii) extremely long glottal closed phases [6][7]. Fundamental frequency (F_0) in creaky voice is lower than in modal voice and F_0 values will be below the lower limit of F_0 range of speaker. Standard F_0 tracking algorithms tend to either output spurious values in creaky regions or detect creaky regions to be unvoiced. As a result of this, creaky regions will be poorly modeled. In order to efficiently extract proper acoustic features in creaky regions, the first task should be to correctly identify creaky regions in speech.

In literature, there are very few approaches to automatically detect the creaky regions [8, 9, 10], though several methods exist for detecting the broader class, i.e., irregular phonation. In [8], an extension of the Aperiodicity, Periodicity and Pitch (APP) detector is proposed for automatic detection of irregular phonation including creaky voice. In the first step, irregular frames are separated from periodic frames using the periodicity measure of the APP detector. In the second step, using the dip profile of the Average Magnitude Difference Function (AMDF) in various frequency bands, creaky regions are identified. Ishi et al., [9] computed short term power and Intraframe Periodicity (IFP) strength contours from the speech signal for differentiating modal and creaky voiced regions. Interpulse similarity (IPS) measure is used to differentiate unvoiced and creaky regions. In [10], a creaky voice detection method is proposed using two new acoustic features. The first feature exploits the occurrence of secondary peaks in the LP residuals of creaky regions, and the second feature captures strong impulse-like peak and long glottal pulse duration properties of creaky regions.

In this paper, the source of excitation represented in terms of set of epoch parameters is analyzed in modal and creaky regions. Epoch parameters are extracted directly from speech signal using zero-frequency filtering (ZFF) method with different window sizes [11, 12]. Using the variance of epoch parameters, a neural network classifier is trained to detect the creaky regions. The paper is organized as follows. In section II, ZFF method used for extracting epoch parameters is discussed. The proposed creaky voice detection method is described in section III. In section IV, the proposed method is compared with the existing creaky detection methods. The summary of work presented in the paper and future issues that need to be addressed are given in section V.

2. Zero-frequency filtering method for extracting epoch parameters

In this work, epoch parameters which include number of epochs in a frame, strength of excitation of epochs and epoch interval between successive epochs are extracted by using zero-frequency filtering method [11]. The ZFF method is based on the principle that the discontinuity due to impulse excitation is reflected across all frequencies including zero-frequency [11]. By designing a resonator at zero frequency, information regarding impulse excitation can be obtained by removing the vocal tract response. The system function of such a resonator is given by

$$H(z) = \frac{1}{1 + a_1 z^{-1} + a_2 z^{-2}} \quad (1)$$

where a_1 = -2 and a_2 = 1. The above resonator de-emphasizes the characteristics of vocal tract system. A cascade of two such resonators, given by system function $G(z)$=$H(z)H(z)$ is used to damp out all the resonances of vocal tract system. Let $s[n]$ denote the input speech signal. The output of cascade of two resonators is given by $x_s[n] = s[n] * g[n]$. The output of

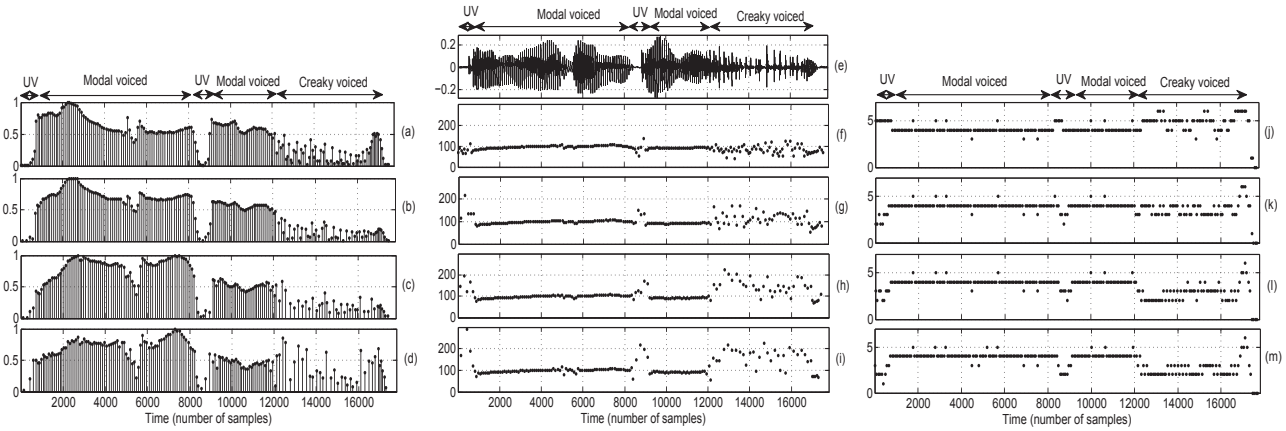

Figure 1: *Illustration of variation of epoch parameters in different voicing regions with different window sizes. Strength of excitation ((a),(b),(c),(d)), epoch interval ((f),(g),(h),(i)) and number of epochs ((j),(k),(l),(m)) computed from the speech signal ((e)) with window sizes of 8, 10, 12 and 14 ms.*

zero-frequency resonator $x_s[n]$ decays or grows as a polynomial function of time. Hence, it is difficult to directly detect the effect of discontinuities due to impulse excitation in the filtered output. The characteristics of discontinuities due to impulse excitation are extracted by computing the deviation between zero-frequency filtered output and the local mean. The window length used for computing the local mean is chosen to be around average pitch period. The resulting signal obtained after subtracting the local mean is called zero-frequency filtered (ZFF) signal [11] and is given by

$$y[n] = x_s[n] - \frac{1}{2N+1} \sum_{m=-N}^{N} x_s[n+m] \qquad (2)$$

where $2N$+1 represents the length of window in terms of number of samples. The time instants of negative to positive zero crossings of the ZFF signal are called as the instants of significant excitation or epochs. The strength of excitation is computed as the slope of ZFF signal at each epoch location [12]. The strength of excitation indicates the rate of closure of the vocal folds in each glottal cycle [13]. Sharper closure of the vocal folds corresponds to stronger excitation to the vocal tract system. Epoch interval is computed as the time duration between successive epochs.

3. Proposed creaky voice detection

Proposed creaky voice detection is performed based on the variation of epoch parameters, namely, number of epochs, strength of excitation of epochs and epoch intervals for different voicing regions. The epoch parameters vary with the size of window used for local mean subtraction. In [14], variation of strength of excitation for different window size is analyzed for detecting voiced and unvoiced speech. In this approach, variation of epoch parameters for different window sizes is systematically examined for creaky voice detection.

Figure 1 shows strength of excitation ((a),(b),(c),(d)), epoch interval ((f),(g),(h),(i)) and number of epochs ((j),(k),(l),(m)) computed from the speech signal with window sizes of 8, 10, 12 and 14 ms. The speech signal shown in Figure 1(e) contains three types of regions, namely, unvoiced, modal and creaky voiced regions. In unvoiced regions, vocal folds do not vibrate and there is no impulse-like excitation. As a result, in unvoiced regions the epochs are located at random instants and the strength of excitation is very low for different window sizes. For modal regions, vocal folds vibrate regularly and the most significant impulse-like excitation occurs during glottal closure instant (GCI). Hence, in modal regions the epochs are located at regular instants and the strength of excitation is high for different window sizes. In creaky regions, vocal folds vibration is low and irregular, and impulse-like excitation to the vocal tract system occurs at two instants of glottal cycle. One impulse-like excitation occurs at glottal closing instant (primary excitation) and other at glottal opening instant (secondary excitation), following long glottal closed phase. Hence, the epochs are located at uneven locations and the strength of excitation is higher than unvoiced regions but lower than modal regions. On careful analysis of epoch parameters for modal and creaky regions with different window sizes, three main observations can be drawn.

1. **Strength of excitation (Figure 1(a),(b),(c),(d)):** In creaky regions, in addition to GCI, secondary excitation is also present within a single glottal cycle. The strength of excitation at GCI is relatively high compared to secondary excitation. As a result for different window sizes, the strength of excitation in successive epochs is varying abruptly. In modal regions, slow variation in the strength of excitation can be observed across different window sizes.
2. **Epoch interval (Figure 1(f),(g),(h),(i)):** In modal regions, successive epoch intervals are almost equal or vary slowly. For different window sizes, variation in epoch intervals are not significant. In creaky regions, due to presence of secondary excitation, successive epoch intervals are unequal. Hence the epoch intervals vary significantly for different window sizes.
3. **Number of epochs (Figure 1(j),(k),(l),(m)):** In creaky regions, the secondary excitations having very low strength are detected for lower window sizes and missed for higher window sizes. As a result, the number of epochs in a frame varies for different window sizes. In modal regions, the secondary excitations are not present. Hence, for different window sizes, the number of epochs in modal voiced regions does not vary significantly.

From these observations, we can conclude that for different

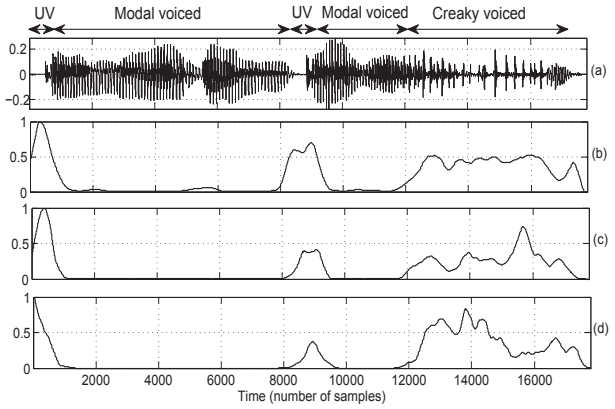

Figure 2: *(a) Speech signal. Variances of (b) strength of excitation, (c) epoch interval and (d) number of epochs computed with the window sizes varying from 8 to 14 ms in steps of 1 ms.*

window sizes, the epoch parameters vary significantly in creaky regions compared to modal regions. To differentiate modal and creaky regions, variance of epoch parameters is computed for every frame of speech. The procedure for finding the variance of epoch parameters is as follows. **(1) Variance of strength of excitation:** For every frame, the strength of excitation obtained from every window size is normalized between 0 to 1 and its variance is calculated. Average variance of strength of excitation is computed from the variances obtained for different window sizes. **(2) Variance of epoch interval:** For every frame, epoch intervals are collected from different window sizes. From the epoch intervals obtained from different window sizes, the variance of epoch interval is computed. **(3) Variance of number of epochs:** For every frame, by considering the number of epochs obtained from different window sizes, variance of number of epochs is computed. All three variances determined from the speech utterance are normalized between 0 to 1. Figure 2 shows the speech signal and the variances of strength of excitation, epoch interval and number of epochs computed with the window sizes varying from 8 to 14 ms in steps of 1 ms. The procedure for choosing the optimum range of window size is detailed in Section 4. From the figure, it can be observed that the variances have high values in creaky and unvoiced regions, and zero or very low values in modal regions. Using voicing detection method at the first stage, unvoiced regions are removed and epoch parameters are extracted only in voiced regions consisting of modal and creaky regions. For voicing detection, recently proposed method based on the strength of instants of significant excitation [14] is used. This method is shown to efficiently identify both modal and creaky regions as voiced.

Creaky/non-creaky classification can be performed by applying threshold on the variance of epoch parameters. Instead of using threshold method, neural network classifier is used. The classifier is configured as a feed forward network consisting of a single hidden layer. All neurons (fixed to 16 in this work) present in the hidden layer utilize a *tanh* transfer function. The output layer consists of a single neuron with a logarithmic sigmoid function suited for a binary decision. The training is performed using a standard error back-propagation algorithm [15]. In this study, the output of neural network classifier is approximated as a posterior probability. Post processing is carried out on the binary decision (creaky or non-creaky) of the classifier. The detection of creaky regions in very short regions is removed and nearby adjacent creaky regions are merged by performing a 5-point median filtering to the binary decision. Assuming a minimum possible F_0 value of creak to be 62.5 Hz, the epoch parameters are extracted for a frame length of 32 ms and a frame shift of 10 ms.

4. Performance evaluation

In order to evaluate the detection performance of the proposed method, different speech databases were considered. First three databases include an American English male speaker (BDL [16]), a Finnish male speaker (MV [17]) and a Finnish female speaker (HS [18]). All three databases were developed to build text-to-speech synthesis systems. A creaky database was developed in two Indian languages, namely, Hindi and Bengali. In each language, single female and male speakers (native voice talents) were used for recording the speech corpus. For both Hindi and Bengali, the text corpus consists of 100 sentences obtained from children stories. Speakers were asked to utter the sentences in news reading style and were asked to intentionally produce creaky regions at the end of the utterances. 100 speech utterances obtained from BDL, MV, HS, Hindi (H-F1 and H-M1) and Bengali (B-F1 and B-M1) speech databases were used in evaluation. All speech utterances were downsampled to a sampling frequency of 16 kHz. To evaluate the detection performance, a reference indicating creaky and non-creaky regions in the speech database was required. Annotation of creaky regions was performed manually following a similar approach as described in [9]. Manual annotation of creaky regions was performed based on the auditory criterion "a rough quality with the additional sensation of repeating impulses". In addition, an inspection of waveforms, spectrograms and F_0 contours was also performed to ensure correct annotation.

To assess the performance of the proposed method, three standard frame level metrics are used, namely, True Positive Rate (TPR, also called recall), False Positive Rate (FPR), and F1 score. TPR is the proportion of actual creaky frames that are correctly identified. FPR is the proportion of actual non-creaky frames that are wrongly detected as creaky. F1 score is a single metric (bound between 0 to 1) computed using true positives, false positives and false negatives. If the technique is better, then TPR and F1 score are higher and FPR is lower.

In the neural network classifier, for a given input example the output is the posterior probability, P1. The standard binary decision is class 1 (i.e., creaky) if P1 $> \alpha$ (otherwise class 0) and α is typically set to 0.5. For skewed datasets (e.g., creak or laughter) which consists of sparse occurrence of a given class to be detected, this setting may not be optimal. Hence, α is varied in the range [0, 1] and set to the value which maximizes the F1 score on the training set. The threshold setting had very low inter-database sensitivity, as all speakers had their best F1 score for α in the vicinity of 0.3. This kind of optimal threshold setting was followed in [10][19].

Optimum range of window size: In Figure 2, epoch parameters are computed by varying the window size from 8 to 14 ms in steps of 1 ms. With this range of window size, high distinction of epoch parameters is observed for modal and creaky regions. For different speakers, we need to find the optimum range of window size which results in higher F1 score and hence better creaky detection. First, by varying the window size from 1 ms to 20 ms in steps of 1 ms, epoch parameters are computed. With one window size as centre, the epoch parameters of centre window size and the epoch parameters of two window sizes before and after the centre window size (similar to 5-gram model)

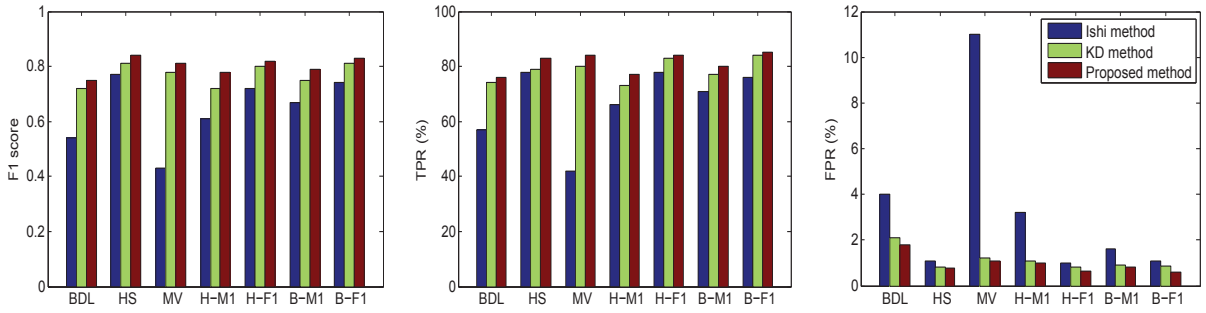

Figure 3: *F1 scores (left), TPR (middle) and FPR (left) values obtained for three creaky detection algorithms on BDL, HS, MV, H-M1, H-F1, B-M1 and B-F1 databases.*

are considered. For example, if centre window size is 8 ms, then the epoch parameters of 8, 6, 7, 9 and 10 ms are considered. From the epoch parameters, variances are computed. Using the variance of epoch parameters, a neural network classifier is trained and F1 score is computed. Similarly, with every window size as center, variance of epoch parameters are computed and subsequently F1 scores are determined. F1 scores obtained with different window sizes as centre for speakers BDL and HS are shown in Figure 4. Average pitch periods of speakers BDL and HS are 5.91 ms and 5.11 ms, respectively. From figure, we can observe that for a window size range of approximately 1.5 to 2 times pitch period of speaker, F1 score is having high values for both BDL and HS speakers. Similar kind of F1 score curves are obtained for other speakers also. Hence in this work, for computing the variance of epoch parameters, the window size is varied from 1.5 to 2 times pitch period of the speaker in steps of 1 ms. Here, instead of considering two window sizes before and after the centre window size, we tried increasing or decreasing the number of window sizes. Increasing the number of window sizes makes the F1 score curve smoother and decreasing the number of window size results in sudden fluctuations in the F1 score curve.

Evaluation of the detection performance was performed using a leave one speaker out strategy, where the speech data of a given speaker was held out for testing and the remaining speech data of all speakers was used for training. This procedure was repeated for each speaker. The proposed method is compared with two existing techniques: **(i) Ishi's method [9]:** In this method, short term power, intraframe periodicty and interpulse similarity measures are extracted to differentiate creaky regions from modal and unvoiced regions. **(ii) Kane-Drugman (KD) method [10]:** Creaky detection is performed by utilizing two acoustic features which are designed to characterize the presence of secondary peaks and prominent impulse-like excitation peaks from the LP residual signal.

F1 score, TPR and FPR obtained for different speech databases are shown in Figure 3. From the figure, it can be observed that the proposed method using epoch parameters performs better than the two existing methods, across all databases. Among all results, Ishi's method displays lowest TPR and F1 score for MV speaker. The main reason for this is that MV speaker has modal regions at low frequency. Intraframe periodicity values extracted from Ishi's method dropped below threshold value in low voiced regions. Hence the modal regions having low frequency regions were wrongly identified as creaky regions (evident by high values of FPR). Proposed method produced high F1 score and TPR of MV speaker, as the extracted

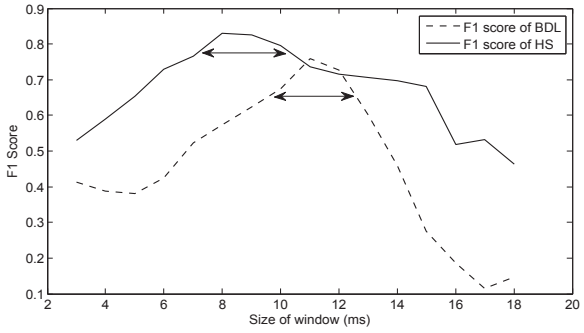

Figure 4: *F1 scores obtained with different window sizes as centre for speakers BDL and HS. Double arrow indicates the approximate range of 1.5 to 2 times pitch period of speaker, where superior performance is observed.*

variance of epoch parameters were independent of pitch of the speech signal. KD-method performed better than Ishi's method for all speech databases, but its performance is inferior, compared to the proposed method. One-way ANOVA is carried out to investigate whether the performance of proposed creaky detection method is significantly better than the two existing methods. Here, F1 score is treated as the dependent variable and detection method as the independent variable. One-way ANOVA indicated that the creaky detection method had a significant effect on the F1 score [F = 16.0, $p < 0.001$] and pair-wise comparisons carried out using Tukeys Honestly Significant Difference (HSD) test showed that the proposed method gave significantly higher F1 scores than both Ishi'method (p <0.001) and KD method (p <0.01).

5. Conclusion

In this paper, we have proposed a creaky voice detection method based on the epoch parameters which displays distinct characteristics for modal and creaky regions. ZFF method is used to extract the epoch parameters from the speech signal. Using the variance of epoch parameters, a neural network classifier is developed to identify creaky regions. Compared to existing methods, the performance evaluation results showed that F1 score and TPR of the proposed method were significantly high for different speech databases. The performance of the proposed method can be evaluated using different speech modes such as conversational and expressive speech. The performance of proposed method may be examined for different noisy environments.

6. References

[1] S. Kushan and J. Slifka, "Is irregular phonation a reliable cue towards the segmentation of continuous speech in American English," in *Proc. of Speech Prosody*, Dresden, Germany, 2006, pp. 795–798.

[2] H. Siln, E. Helander, J. Nurminen, and M. Gabbouj, "Parameterization of vocal fry in HMM-based speech synthesis," in *Proc. Interspeech*, 2009, pp. 1775–1778.

[3] I. Yanushevskaya, C. Gobl, and A. N. Chasaide, "Voice parameter dynamics in portrayed emotions," in *Proc. of the 6th International Workshop on Models and Analysis of Vocal Emissions for Biomet-rical Applications (MAVEBA)*, 2009, pp. 21–24.

[4] R. Carlson, K. Gustafson, and E. Strangert, "Prosodic cues for hesitation," in *Proc. of Fonetik*, 2006, pp. 21–24.

[5] T. Drugman, J. Kane, and C. Gobl, "Resonator-based creaky voice detection," in *Proc. Interspeech*, 2012.

[6] M. Blomgren, Y. Chen, M. Ng, and H. Gilbert, "Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers," *Journal of the Acoustical Society of America*, vol. 103, no. 5, pp. 2649–2658, 1998.

[7] C. Gobl and A. N. Chasaide, "Acoustic characteristics of voice quality," *Speech Communication*, vol. 11, pp. 481–490, 1992.

[8] S. Vishnubhotla and C. Espy-Wilson, "Automatic detection of irregular phonation in continuous speech," in *Proc. Interspeech*, 2006, pp. 949–952.

[9] C. Ishi, K. Sakakibara, H. Ishiguro, and N. Hagita, "A method for automatic detection of vocal fry," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 47–56, 2008.

[10] J. Kane, T. Drugman, and C. Gobl, "Improved automatic detection of creak," *Computer Speech and Language, Elsvier*, vol. 27, pp. 1028–1047, 2013.

[11] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 8, pp. 1602–1613, 2008.

[12] K. S. R. Murty, B. Yegnanarayana, and M. A. Joseph, "Characterization of glottal activity from speech signals," in *IEEE Signal Processing Letters*, vol. 16, no. 6, 2009, pp. 469–472.

[13] P. Alku, T. Bakstrom, and E. Vikman, "Normalized amplitude quotient for parameterization of the glottal flow," *Journal of the Acoustical Society of America*, vol. 112, no. 2, pp. 701–710, 2002.

[14] N. P. Narendra and K. S. Rao, "Robust voicing detection and F0 estimation for HMM-based speech synthesis," *Circuits, Systems, and Signal Processing*, DOI: 10.1007/s00034-015-9977-8, 2015.

[15] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[16] "CMU ARCTIC speech synthesis databases," *[Online]. Available: http://festvox.org/cmu_arctic/*.

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

[18] H. Silen, E. Helander, K. Koppinen, and M. Gabbouj, "Building a Finnish unit selection TTS system," in *Workshop on Speech Synthesis*, 2007, pp. 310–315.

[19] T. Drugman, J. Kane, and C. Gobl, "Data-driven detection and analysis of the patterns of creaky voice," *Computer Speech and Language*, vol. 28, no. 5, pp. 1233–1253, 2013.