Creaky Voice and the Classification of Affect

Ailbhe Cullen[1]*, John Kane*[2]*, Thomas Drugman*[3]*, Naomi Harte*[1]

[1]Department of Electronic and Electrical Engineering, Trinity College Dublin, Ireland
[2]Phonetics and Speech Laboratory, School of Linguistic, Speech and Communication Sciences, Trinity College Dublin, Ireland
[3]TCTS Lab, University of Mons, Belgium

cullena3@tcd.ie, kanejo@tcd.ie, thomas.drugman@umons.ac.be, nharte@tcd.ie

Abstract

Creaky voice, a phonation type involving a low frequency and often highly irregular vocal fold vibration, has the potential both to indicate emotion and to confuse automatic processing algorithms, making it both friend and enemy to emotion recognition systems. This paper explores the effect of creak on affect classification performance using hidden Markov models (HMMs) trained on a variety of voice quality features. The importance of "creak-robust" features is demonstrated, and it is shown that features designed to capture creak may be particularly useful for the classification of power. The SEMAINE database of emotional speech is used, thus allowing us to compare with the AVEC 2011 challenge, and we find that the voice quality features match the performance of the best reported classifier on the challenge.

Index Terms: emotion recognition, voice quality, vocal creak, hidden Markov model

1. Introduction

It is widely accepted that the ability to understand emotion or affect from speech is central to the design of more natural human-computer interfaces [1]. There has been much research into the relationship between emotion and voice quality [2, 3, 4]. This has focused on human expression and perception of emotion. There has been little investigation into the effect of voice quality on automatic emotion classification, other than the incorporation of voice quality features to aid discrimination.

A particularly interesting voice quality for emotion recognition is creaky voice (also referred to as vocal fry, glottal fry, or laryngealisation). Creak is widely accepted as being an indicator of emotion (in particular boredom and low activation anger) and should thus be expected to improve emotion classification. Despite this potential, the irregular periodicity often associated with creak can have a severely damaging effect on acoustic feature extraction [5].

This paper explores the effect of creak on affect (as opposed to emotion) classification by comparing the performance of a binary (High/Low) classifier on creaky and creak-free words for four affective dimensions: activation; expectation; power; and valence. While a number of different features have been applied to the emotion/affect recognition problem, there is little consensus over which are most relevant [6]. Prosodic features are the most popular, but Teager energy operator (TEO) [7], voice quality [8, 9], and a number of spectral measures [10, 11], have also been applied to the task. In this study we focus on voice quality features for two reasons. Firstly, because we expect certain glottal source parameters to be sensitive to creak, and secondly because they have been demonstrated to be useful in classifying emotions and affective dimensions which are not well defined by prosodic features [8, 9].

To enable comparison with previous work, we use the SEMAINE database of natural emotional speech [12], which has recently been used in the 2011 Audio-Visual Emotion Challenge (AVEC 2011) [13]. We compare our results with those of Meng et al. [14] who achieved the best performance on the audio sub challenge. Furthermore, we repeat all analysis with a set of Mel frequency cepstral coefficients (MFCCs), and a set of prosodic and spectral features, as they have been widely used in speech and emotion recognition.

The results reported offer important new insights into the role of creak in affect classification, demonstrating the benefit of creak-robust features for power. The final classifier results, using voice quality features, slightly outperform the classifier of Meng et al. [14], the first reported results to do so.

The rest of the paper is structured as follows: Section 2 introduces the phenomenon of vocal creak. The database used and our labelling for creak is discussed in Section 3. Section 4 outlines the features and classifier used in this study. Results are presented in Section 5, with a final discussion and summary in Section 6.

2. Creak

Creak is a mode of phonation (a strong determinant of perceived voice quality), also known as vocal fry, glottal fry, or laryngealisation, which has a low fundamental frequency and produces a distinctive crackling sound [3].

Creak has been associated with boredom [2, 3], sadness [2], and also with some forms of anger (touchy, reprimanding [16], suppressed rage [3]). We would thus expect creak (or some feature capturing it) to aid discrimination between emotion classes. Batliner [16] argues that creak is at best a weak indicator and is certainly not strong enough to aid automatic recognition. However Gobl et al. [2] suggest that voice quality indicates subtle changes in affect, rather than strong emotions. Thus, while creak may not be useful for N-class emotion recognition, it may be more important for classification along affective dimensions.

Due to the irregular periodicity often associated with creak, pitch trackers often output either spurious values (as is sometimes the case for YIN) or an unvoiced state (as often happens with Talkin's RAPT algorithm) [5]. Furthermore, creak is a speaker-dependent phenomenon. Not all speakers use creak, and while it may be used to signal emotion. Creak is also used to mark word or phrase boundaries, it often occurs as a speaker runs out of breath, or it may be part of the speakers accent or another idiosyncrasy. Thus automatic creak processing may cause confusion rather than aid discrimination.

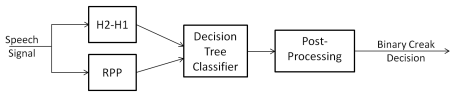

Figure 1: Block Diagram of Creak Detector

3. Emotional Speech Database

3.1. The SEMAINE database

In this study we perform word level binary classification of four affective dimensions: activation; valence; power; and expectation. The database used is the Solid SAL portion of the SEMAINE database of emotionally coloured character interactions [12]. This is an English language database, consisting of conversations held between an operator and a user. The operator adopts the role of one of four characters, and by acting emotionally attempts to induce natural emotional responses in the user. This has resulted in a rich database of over 12 hours of natural emotional speech.

The SEMAINE database was recently used as the challenge data for AVEC 2011 challenge [13]. The challenge was part of the 2011 International Conference on Affective Computing and Intelligent Interaction. The task for this challenge was binary(High/Low) classification of words along four affective dimensions, using either audio or video information, or both. This provides a useful reference with which to compare our results. We consider two partionings of the database in this study, one in which the data is separated into folds containing set levels of creak, and the second in which we use the official AVEC partitions. We compare results on the AVEC partitions with the highest performing classifier in the audio sub-challenge, that of Meng et al. [14].

3.2. Annotation for Creak

The manual annotation of a database this large would be time-consuming. Instead, creak labels are generated automatically using the algorithm of Kane et al. [5]. A block diagram is shown in Figure 1 This method relies on the observation that the Linear Prediction (LP) residual contains strong secondary peaks proceeding the glottal closure instants in creaky regions. These peaks cause strong harmonics when the LP residual signal is applied to a resonator centred at the speakers mean f0. The first creaky feature measures the difference between the first and second harmonics obtained from this process (H2-H1). The second feature is the residual peak prominence (RPP), which measures the amplitude difference between the two strongest peaks in the LP signal. These two features are then used to classify creaky speech using a binary decision tree classifier. This method has been proven to outperform the state of the art on a wide range of speakers, genders, languages, and noise conditions [5].

Creak is labelled at 10ms intervals. The AVEC 2011 classification task is on the word level, therefore we create word level creak rating by labelling words containing any creak as creaky. Overall, approx 18% of words in the database were flagged as containing some creak, with the 17% of the training set, 18% of the development set, and 19% of the test set containing creak. In normal (non-pathological) speech we expect creak to account for 3-5% of speech. On average less than 30% of frames in each creaky word contain creak, so an average of 18% word level creak is reasonable.

4. Experimental Setup

4.1. Feature Extraction

Typically, voice quality is measured via a number of time and frequency parameters from the glottal waveform [17, 18, 19]. Our first set of glottal features are extracted in four steps. First the pitch and voicing decision are obtained using the RAPT pitch detection algorithm [20] in the VOICEBOX toolbox [21]. The voiced segments within each word are then divided into frames sampled every 10 ms, with a variable frame length corresponding to four times the pitch period. The glottal waveform is estimated via the iterative adaptive inverse filtering (IAIF) method [17] from the Aparat toolbox [18], and 11 time and frequency parameters from Aparat are recorded (see Table 1). Henceforth this feature set will be referred to as the Glot feature set.

Table 1: Full list of voice quality features used. *extracted using Aparat [18].

		Glot	VQ	Glot+VQ
OQ1, OQ2, OQa, AQ, CLQ, SQ1, SQ2 *	[18]	✓		✓
NAQ, QOQ, H1-H2 *	[18]	✓	✓	✓
HRF *	[18]	✓		✓
Peak Slope (Wavelet, Glottal)	[22]		✓	✓
RPP, H2-H1	[5]		✓	✓
MDQ	[23]		✓	✓
F0, VUV	[24]		✓	

A second set of voice quality features, referred to as the VQ feature set, contains a mixture of measurements from the glottal waveform, wavelet decomposition, and LP residual, including those used by Kane et al. for creak detection [5]. The inclusion of features which are not derived from the glottal waveform should make this feature set more robust in areas where the pitch or glottal waveform are difficult to estimate [22]. A full list of the features included in this set and relevant references can be found in Table 1.

To compare performance, we also include a standard set of MFCCs, and a set features based on the baseline features provided for the AVEC challenge. We use a set of 12 MFCCs which are extracted over 25 ms frames, spaced 10 ms apart. The full AVEC feature set consists of word-level functionals of 31 frame-level features. Our reduced AVEC (AVEC-r) feaure contains set only the frame-level features, as listed in [13], less the 10 MFCCs. First derivatives of all features are computed over a three frame window and are included in HMM training and testing.

4.2. Classifier

The majority or participants in the AVEC 2011 challenge used static classifiers. A notable trend in the challenge is the sharp decrease in performance on the test set for all participants when compared with the development set. This suggests that classifiers may have been over trained, or were unable to generalise well to the unseen data.

The Hidden Markov Model (HMM) is used in an attempt to address this issue. HMMs have proven to perform better than static classifiers for discrete emotion recognition [25, 26], and have also been shown to perform well on speaker independent classification tasks [11, 27]. Despite this there has been relatively little interest in using HMMs for affect recognition. We use a left/right HMM classifier, in which HMMs were trained using frame level features. Experiments suggested that a single state HMM is sufficient to model activation and valence, while power and expectation benefit from the temporal modelling ability of the HMM, thus for expectation and valence a five state model is used. All models contain twenty Gaussian mixtures within each state. The HMMs were implemented

Table 2: Creak ratio for each speaker in the AVEC database, using the SEMAINE speaker ids, and presence in AVEC partitions.

Speaker id	2	3	4	5	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Creak ratio	0.15	0.28	0.13	0.09	0.22	0.08	0.09	0.19	0.79	0.36	0.27	0.09	0.18	0.11	0.12	0.20	0.28	0.07	0.17	0.18
Train		✓			✓		✓	✓			✓			✓	✓					
Devel	✓	✓			✓					✓		✓		✓		✓			✓	
Test			✓	✓		✓			✓				✓				✓	✓		✓

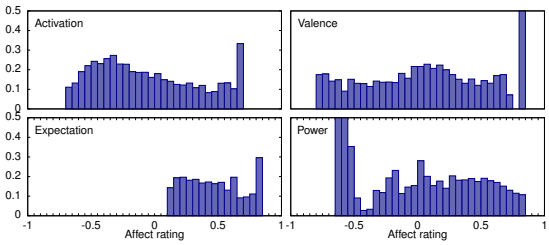

Figure 2: Distribution of creak with respect to affect ratings

using the Hidden Markov Model Toolkit (HTK) [28], and were trained using word level labels. For each dimension, two HMMs were trained, corresponding to the High and Low binary labels.

A noted drawback of the HMM is its inability to model supra-segmental information [6]. That is to say, the HMM classifier models the evolution of emotion within words, but does not account for long term dependencies between words. Emotion varies slowly, so rapid switching between high and low affective states is unlikely. Thus a post processing stage is included in which a median filter is applied over a seven word window to de-noise the output labels.

5. Results

5.1. Creak and Affect Labels

Within the English language, creak has been associated with boredom, and to a lesser extent anger and sadness. However, there is little mention in the literature of the relationship between creak and any affective dimensions. Therefore, we first compare the creak and affective labels to see if there is any relationship between them.

The SEMAINE database is labelled on a continuous scale from -1 to +1 for all dimensions, except for expectation which is labelled from 0 to 1. The distribution of creak across each dimension is shown in Figure 2. We expect creak to occur more frequently with low activation, as low subglottal pressure is likely to be a necessary condition for creaky voice. This is confirmed by Figure 2. Creak occurs with negative power. The distribution is more flat for expectation and valence. Disregarding the final peak (due to a lack of extremely high valence ratings), creak occurs most frequently with weak positive valence (within the range [0,0.25]). Similarly, disregarding the final peak creak occurs more frequently with weak positive expectation. This agrees with Gobl et al.'s observation that voice qualities express subtle changes rather than strong emotions [2].

In order to measure the relative "creakiness" of each speaker we will define the Creak Ratio (CR) as follows:

$$CR = \frac{N_{ck}}{N_{wd}} \quad (1)$$

where N_{ck} is the number of words from a speaker detected as creaky, and N_{wd} is the overall number of words spoken by the speaker. Table 2 reports the CR of each speaker in the database, and shows their allocation to the official AVEC training, development, and test partitions. Creak is unevenly distributed throughout the different partitions. The test partition has a mean CR of 0.23 and standard deviation 0.24, while the train set has mean 0.18 and standard deviation 0.08. Both the most ($CR = 0.79$) and least ($CR = 0.07$) creaky speakers are contained in the test partition. The 0.79 CR obtained for speaker 11 may seem surprisingly high, suggesting a potentially high frequency of false positives by the creaky detector, however close listening to the recordings for this speaker revealed a habitual creaky voice quality, and confirms that 0.79 is an accurate CR.

5.2. Performance on Creaky and Creak-free partitions

We explore how well a classifier can cope with creak if it has not previously encountered it in training. Two classifiers are trained, one in which the training data contains no creaky words, and one which contains equal amounts of creaky and creak-free words in the training data. Both classifiers are tested using the same data, which contains an equal number of creaky and creak free words. The total size of the two training partitions is equal, with the test partition being slightly smaller. All folds contain a balanced distribution of High and Low labels for each dimension.

Figure 3 shows the accuracy when the classifiers are trained on a creak free partition and on a partition in which 50% of the words contain some creak. The naïve assumption is that by incorporating creak in the training data, we will improve classification of creak and so improve overall classification performance. However, Figure 3 shows that only three out of the 16 classifiers benefit from having seen creak in training. This is due to a number of factors. First, the creaky speech introduces more variance in the features, which the classifier is unable to model without more data. Second, creak is a very personal trait. Different speakers use different levels of creak in different ways, so the creak patterns for the training speakers may not match the creak patterns of the test speakers. Finally, no precautions were taken to ensure that the creak samples used are caused by emotion rather than word/utterance ends which, as discussed in Section 2, may confuse the classifier.

Of the five classifiers tested, the voice quality based classifiers (Glot, VQ, and GlotVQ) are less affected by the presence of creak in the training data. In particular, the GlotVQ features give practically the same performance on activation and power, regardless of whether they have seen creak or not in training. Thus these features appear more robust to creak.

5.3. Performance on the AVEC partitions

Figure 4 shows the performances on the test (red points) and development (blue points) partitions of the AVEC 2011 challenge. Results are shown separately for creaky and creak-free words therein. An overall trend is that performance drops on the test set, where new speakers are introduced. The low performance, close to 50%, is typical of this challenging data.

On activation, expectation, and power, the Glot features perform better on creak free words than on creaky words, while the VQ set tends to perform better on creaky words. This implies that the LP-residual features designed by Kane et al. [5] better capture and exploit creak. On valence, all voice quality features perform better on creaky samples, indicating that creak

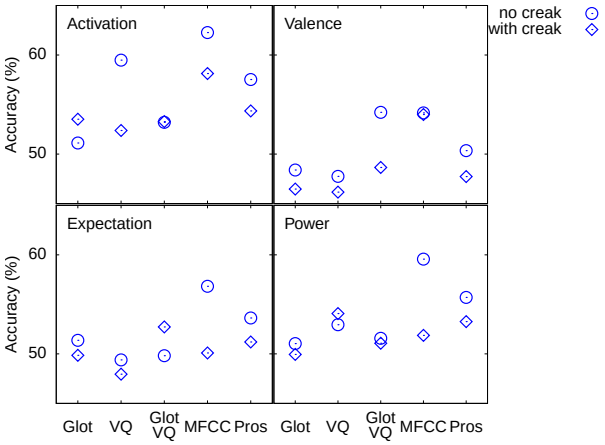

Figure 3: Accuracy of classifiers trained in the absence of creak (○) and trained on equal parts creaky and creak free speech (◇)

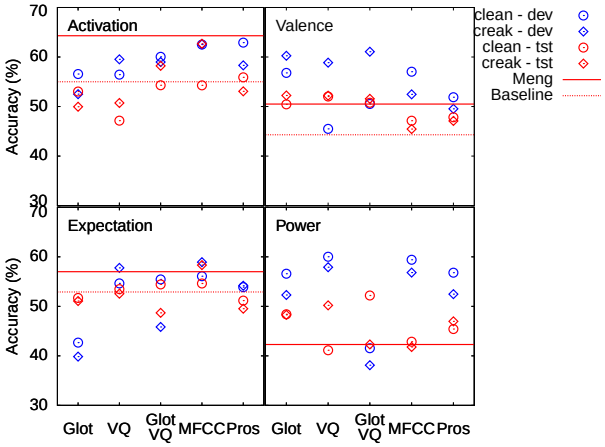

Figure 4: Performance on creaky and clean (i.e. non creaky) words in AVEC development and test partitions.

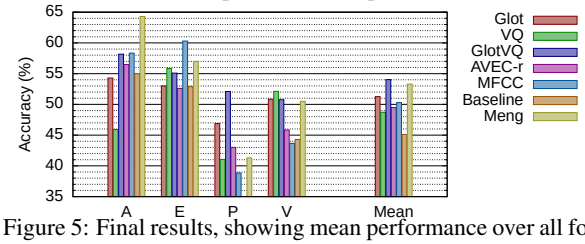

Figure 5: Final results, showing mean performance over all four dimensions

is particularly useful in classifying this dimension.

Combining the Glot and VQ features we improve performance over both that of the Glot and the VQ features alone for activation and power. A slight degradation occurs on expectation and valence, but averaging over the four dimensions the combined GlotVQ feature set provides the best performance.

Figure 5 presents the overall results on the AVEC partitioning, with the post-processing stage included, and compares them with the challenge baseline and the results obtained by Meng et al. [14]. The voice quality features compare well with the prosodic (AVEC-r) features, which are known to capture activation well. For expectation and valence, the voice quality features achieve comparable results to Meng and the challenge baseline, and clearly outperform the AVEC-r features, reinforcing the observations in [8, 9] that voice quality features can discriminate between emotions which are poorly modelled by prosodics. On power we achieve a noticeable improvement on previous results, using the GlotVQ feature set. This is consistent with the GlotVQ performance on power in Figure 3. The best overall classifier is the GlotVQ classifier which at 54% accuracy slightly exceeds the previous AVEC 2011 winners, on 53.3%.

6. Discussion & Conclusion

This paper has explored the effect of a particular type of voice quality, known as vocal creak, on the performance of binary affect classification, on four affective dimensions. From an initial analysis of the creak affect labels we found that creak tends to occur with low activation and expectation and high power and valence. For different dimensions and feature sets, significant correlations ($p < 0.05$) were found between speaker creak ratio and classification performance on that speaker. "Creaky" speakers appear to be more difficult to classify for expectation and power but are more accurately classified for activation and valence.

Given this information it may seem surprising that classification of creaky samples is not improved by increasing the amount of creak in the training data. This is due to the adverse effects of creak on the system, increasing the variance and essentially noise in the feature vectors. However, we found that certain features are more robust to this effect than others. For valence, power, and expectation the Glot and VQ features suffer very little degradation in performance when creak is encountered in training. There is no effect on the classification of activation and power using the GlotVQ features.

The same features which are robust to creak in the training set also provide the best overall performance on the official AVEC partitions. In Figure 5, the GlotVQ features clearly surpass all other features when classifying power. Looking back to Figure 3 the classification of power using the GlotVQ features is not affected by creak. Similarly, the best performing features for expectation and valence are the least affected by creak for these dimensions in Figure 3. Clearly there is a benefit to choosing creak-resistant features.

With an accuracy of 54%, the combined GlotVQ features, which contains far fewer features than the AVEC set (16 vs. 1941), gave the best overall performance on the AVEC 2011 task, just surpassing the 53.3% accuracy reported by Meng et al. [14], and in proving particularly beneficial for the power dimension. Given that this is a two class problem, a 54% accuracy is still quite low. As previously mentioned, performances reported on the development tended to be much higher than what was reported on the test set. For example, the official challenge baseline (obtained by an SVM classifier) is 63% on the development set, but only 45% on the test set. However, it is encouraging to note, that while the overall results are low, the difference between performances on development and test samples shown in Figure 4 is smaller than many previously reported solutions, in particular for the Glottal and VQ features.

An interesting next step would be the combination of the GlotVQ features and the stacked k-NN/HMM classifier of Meng et al. [14]. This should retain the Meng's high performance on activation while improving performance on power and valence.

7. Acknowledgements

Ailbhe Cullen is funded by an Embark postgraduate scholarship from the Irish Research Council (IRC). John Kane is supported by the Science Foundation Ireland Grant 09/IN.1/I2631 (FASTNET). Thomas Drugman is supported by FNRS.

8. References

[1] R. Picard, *Affective Computing*. Cambridge, Massachussets: MIT Press, 1997.

[2] C. Gobl and A. Ni Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Communication*, vol. 40, no. 12, pp. 189–212, 2003.

[3] J. Laver, *The Phonetic Description of Voice Quality*. Cambridge: Cambridge Uinversity Press, 1980.

[4] K. R. Scherer, "Vocal affect expression: A review and a model for future research," *Psychological Bulletin*, vol. 99, pp. 143–165, 1986.

[5] J. Kane, T. Drugman, and C. Gobl, "Improved automatic detection of creak," *Computer Speech & Language*, vol. 27, no. 4, 2013, http://dx.doi.org/10.1016/j.csl.2012.11.002.

[6] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 910, pp. 1062–1087, 2011.

[7] G. Zhou, J. H. L. Hansen, and J. F. Kaiser, "Nonlinear feature based classification of speech under stress," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, no. 3, pp. 201–216, 2001.

[8] M. Lugger and Y. Bin, "Cascaded emotion classification via psychological emotion dimensions using a large set of voice quality parameters," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 2008, pp. 4945–4948.

[9] S. Rui, E. Moore, and J. F. Torres, "Investigating glottal parameters for differentiating emotional categories with similar prosodics," in *Acoustics, Speech and Signal Processing, IEEE International Conference on*, 2009, pp. 4509–4512.

[10] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech Communication*, vol. 53, no. 5, pp. 768–785, 2011.

[11] S. Ntalampiras and N. Fakotakis, "Modeling the temporal evolution of acoustic parameters for speech emotion recognition," *Affective Computing, IEEE Transactions on*, vol. 3, no. 1, pp. 116–125, 2012.

[12] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *Affective Computing, IEEE Transactions on*, vol. 3, no. 1, pp. 5–17, 2012.

[13] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, *AVEC 2011The First International Audio/Visual Emotion Challenge*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2011, vol. 6975, pp. 415–424.

[14] H. Meng and N. Bianchi-Berthouze, *Naturalistic Affective Expression Classification by a Multi-stage Approach Based on Hidden Markov Models*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2011, vol. 6975, pp. 378–387.

[15] M. Glodek, S. Tschechne, G. Layher, M. Schels, T. Brosch, S. Scherer, M. Kchele, M. Schmidt, H. Neumann, G. Palm, and F. Schwenker, *Multiple Classifier Systems for the Classification of Audio-Visual Emotional States*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2011, vol. 6975, pp. 359–368.

[16] A. Batliner, S. Steidl, and E. Nöth, "Laryngealizations and emotions: How many babushkas," in *Proceedings of the International Workshop on Paralinguistic Speechbetween Models and Data (ParaLing07)*, 2007, pp. 17–22.

[17] P. Alku, B. Story, and M. Airas, "Evaluation of an inverse filtering technique using physical modeling of voice production," in *Proceedings of International Conference on Spoken Language Processing, Jeju Island*, pp. 497–500.

[18] M. Airas, "Tkk aparat: An environment for voice inverse filtering and parameterization," *Logopedics Phoniatrics Vocology*, vol. 33, no. 1, pp. 49–64, 2008.

[19] Y. Tet Fei, J. Epps, E. H. C. Choi, and E. Ambikairajah, "Glottal features for speech-based cognitive load classification," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 5234–5237.

[20] D. Talkin, "A robust algorithm for pitch tracking (rapt)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[21] M. Brookes, "Voicebox: Speech processing toolbox for matlab," 2007, http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.

[22] J. Kane and C. Gobl, "Identifying regions of non-modal phonation using features of the wavelet transform," in *Proceedings of Interspeech*, 2011, pp. 177–180.

[23] ——, "Wavelet maxima dispersion for breathy to tense voice discrimination," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 6, pp. 1170–1179, 2013.

[24] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," *Proc. Interspeech, Florence, Italy*, pp. 1973–1976, 2011.

[25] C. M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "Emotion recognition based on phoneme classes," in *Interspeech*, 2004, pp. 889–892.

[26] B.-S. Kang, C.-H. Han, S.-T. Lee, D.-H. Youn, and C. Lee, "Speaker dependent emotion recognition using speech signals," in *Sixth International Conference on Spoken Language Processing*.

[27] J. Wagner, T. Vogt, and E. Andr, *A Systematic Comparison of Different HMM Designs for Emotion Recognition from Acted and Spontaneous Speech*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2007, vol. 4738, pp. 114–125.

[28] "Hidden markov model toolkit (htk)," http://htk.eng.cam.ac.uk/.