SPEAKER AND LANGUAGE INDEPENDENT VOICE QUALITY CLASSIFICATION APPLIED TO UNLABELLED CORPORA OF EXPRESSIVE SPEECH

[1]*John Kane,* [2]*Stefan Scherer,* [3]*Matthew Aylett,* [2]*Louis-Philippe Morency,* [1]*Christer Gobl*

[1]Phonetics and Speech Laboratory, School of Linguistic, Speech and Communication Sciences, Trinity College Dublin, Ireland
[2]Institute for Creative Technologies, University of Southern California, United States
[3]School of Informatics, University of Edinburgh, UK and CereProc Ltd., UK

ABSTRACT

Voice quality plays a pivotal role in speech style variation. Therefore, control and analysis of voice quality is critical for many areas of speech technology. Until now, most work has focused on small purpose built corpora. In this paper we apply state-of-the-art voice quality analysis to large speech corpora built for expressive speech synthesis. A fuzzy-input fuzzy-output support vector machine classifier is trained and validated using features extracted from these corpora. We then apply this classifier to freely available audiobook data and demonstrate a clustering of the voice qualities that approximates the performance of human perceptual ratings. The ability to detect voice quality variation in these widely available unlabelled audiobook corpora means that the proposed method may be used as a valuable resource in expressive speech synthesis.

Index Terms— Voice quality, glottal source, speech synthesis, expressive speech, audiobooks

1. INTRODUCTION

Voice quality is the perceptual colouring of a person's voice [1]. It is affected by long-term physiological and anatomical settings of the vocal system as well as dynamic shifts in phonation type used for various communicative purposes. Voice quality is an important factor in the perception of emotion in speech [2] and makes a key contribution to the style and uniqueness of a speakers voice [3].

As voice quality plays a critical role in portraying our individuality as well as internal affective states, the ability to detect, and synthesise voice quality changes without compromising naturalness is a key objective for current speech technology. The effective characterisation of shifts in voice quality has potential in speech technology applications like: speech synthesis with flexible voices [4, 5, 6], emotion identification [7, 8], speech recognition [9], speaker identification [10]. However, successful analysis of voice quality faces a number of challenges: As much of the variation in voice quality is brought about by substantial change in phonation type, efforts are usually made to decompose the speech signal into glottal source and vocal tract constituents. This process is non-trivial, both because of the difficulty in applying effective methods for modelling the vocal tract system and because of the limitations of the source-filter model [11]. Separating within- and across-speaker differences requires the analysis of large labelled corpora, which are not generally available. Also, perceptual responses to voice quality changes can vary by subject which makes verification of classification more complex.

A further challenge is that the terminology used to describe voice quality is rather varied in the literature. In this study we adopt the terminology described in [1]. We are concerned here with lax, modal and tense voice, which is perhaps the most researched dimension of voice quality. The term modal voice is used to depict a neutral phonation type with low to moderate laryngeal tension involving periodic vocal fold vibration with a minimum of pulse-to-pulse irregularities and efficient glottal excitation. Tense voice typically involves elevated tension throughout the entire vocal system, compared to modal voice. In terms of laryngeal settings, tense voice typically involves moderate increases in longitudinal and adductive tension [12]. Contrastingly, lax voice is described as having reduced levels of tension in the vocal system and particularly at the laryngeal level.

A considerable amount of research attention has been applied to the design of effective acoustic parameters for characterising voice quality on a lax-tense dimension, with some notable new developments [13, 14, 15]. However, using multiple features for the classification of targeted voice qualities has received considerably less attention. The study described in [16], presented an approach with glottal gradients [17] as voice quality features used with a hidden Markov model (HMM) setup for classifying breathy, modal, rough and creaky voice qualities. However, the data includes speech from just four speakers. A recent study [18] involved the design of a classifier for breathy, modal and tense voice using a range of voice quality features as input to the so-called fuzzy-input fuzzy-output support vector machine classifier. The study demonstrated how classification performance could be improved by leveraging the information contained within the disagreement of multiple voice quality annotators. Although this study involved a reasonably high number of speakers, the data were restricted to steady vowels and all-voiced sentences. Some notable work has also been carried out in [19] where the author classifies different registers of vocal effort in video-game recordings.

The current research builds on existing studies and develops a classifier of voice quality on a lax-modal-tense scale using a large volume of speech data produced by a range of speakers of several languages. The study then aims at testing the generalisation of the classification approach to unlabelled corpora of expressive speech. The research questions can be written explicitly as:

RQ1: Do voice quality features contribute significantly to the classification of these voice qualities compared to using f_0 and standard spectral features?

RQ2: Can we develop a classifier for lax, modal and tense voice and validate on a range of speakers of different languages?

This research was supported by the Science Foundation Ireland Grant 09/IN.1/I2631 (FASTNET) and the Irish Department of Arts, Heritage and the Gaeltacht (ABAIR project). The authors would like to thank Irena Yanushevskaya for providing some of the reference voice quality recordings.

RQ3: Can such a classifier be used to cluster voice qualities in unlabelled corpora of expressive speech to a level which approximates human perceptual ratings?

2. PROPOSED METHOD

The proposed method combines commonly used spectral features and f_0 with a set of voice quality features which have been specifically designed to characterise lax, modal and tense voice. This feature set is used with a classifier recently shown to be effective for classifying these voice quality classes [18].

2.1. Acoustic features

The first three features are derived from the glottal source signal estimated by glottal inverse filtering. For many glottal-based features the correct signal polarity is required, and this is determined automatically [20] and corrected when necessary. Glottal closure instants (GCIs) are then detected using SE-VQ algorithm [21]. Iterative adaptive inverse filtering (IAIF, [22]) is carried out on GCI centred Hamming windowed frames of a duration twice the length of the local glottal period, T_0. The IAIF method works by successive vocal tract all-pole model estimation following the removal of the estimated glottal source contribution modelled with a prediction order which increases at each iteration. The output is then the differentiated glottal flow (provided the effect of lip radiation has not been cancelled). The normalised amplitude quotient (**NAQ**, [23]) is calculated using:

$$\text{NAQ} = \frac{f_{ac}}{d_{peak} \cdot T_0} \tag{1}$$

where d_{peak} is the negative amplitude of the main excitation in the differentiated glottal flow pulse, while f_{ac} is the peak amplitude of the glottal flow pulse. The quasi-open quotient (**QOQ**, [24]) is also derived from amplitude measurements of the glottal flow pulse. The quasi-open period is measured by detecting the peak in the glottal flow and finding the time points previous to and following this point that descend below 50 % of the peak amplitude. The duration between these two time-points is divided by the local glottal period to get the QOQ parameter. The final parameter measured from the glottal source estimate is the difference between the first two harmonics (**H1-H2**), from the narrowband spectrum of the glottal source signal.

The following three features do not rely on explicit glottal inverse filtering. A recently developed method for deriving the global shape parameter of the Liljencrants-Fant (LF, [25]) model was described in [13]. The algorithm considers the mean squared phase difference (MSPD). One can utilise the outcome of minimising:

$$\text{MSPD}^2(\theta, N) = \frac{1}{N} \sum_{k=1}^{N} \left(\Delta^{-1} \Delta^2 \angle R_k^{\theta} \right) \tag{2}$$

where $N = |f_{lim}/f_0|$, k is the harmonic index, f_{lim} is the maximum harmonic frequency and θ is the shape parameter of the glottal model (i.e. Rd). The computation of this objective function involves applying the second order phase difference (Δ^2) and the anti-difference operator (Δ^{-1}) to the *convolutive residual*, R_k^{θ}. R_k^{θ} is the deconvolution of the given speech spectrum, S_k, by the speech model. $\text{MSPD}^2(\theta, N)$ is minimised with respect to θ (i.e. Rd) using the algorithm described in [26].

The final two features involve a dyadic wavelet transform using $g(t)$, a cosine-modulated Gaussian pulse similar to that used in [27] as the mother wavelet:

$$g(t) = -cos(2\pi f_n t) \cdot exp\left(-\frac{t^2}{2\tau^2}\right) \tag{3}$$

where the sampling frequency f_s = 16 kHz, $f_n = \frac{f_s}{2}$, $\tau = \frac{1}{2f_n}$ and t is time. The wavelet transform, $y_i(t)$, of the input signal, $x(t)$, at the i^{th} scale, s_i, is calculated by:

$$y_i(t) = x(t) * g\left(\frac{t}{s_i}\right) \tag{4}$$

where $*$ denotes the convolution operator and $s_i = 2^i$. This functions essentially as an octave band zero-phase filter bank. For the **peakSlope** feature [14], the speech signal is used as $x(t)$ in Eq. (4). Maxima are measured across the scales, on a fixed-frame basis, and a regression line is fit to log_{10} of these maxima. The slope of the regression line for each frame provides the peakSlope value. The feature is essentially an effective correlate of the spectral slope of the signal. Finally, the measurement of the maxima dispersion quotient (**MDQ**, [15]) uses the Linear Prediction (LP) residual as $x(t)$ in Eq. (4). Then using the GCIs, located using SE-VQ, the dispersion of peaks in relation to the GCI position is averaged across the different frequency bands and then normalised to the local glottal period. For tense voice, where the sharp closing of the glottis is analogous to an impulse excitation the maxima are tightly aligned to the GCI, whereas for laxer phonation the maxima become highly dispersed.

We also include 12 Mel-cepstral coefficients (MFCCs) extracted from the speech signal using Hanning windowed 32 ms frames with a 10 ms shift. Fundamental frequency (f_0) is measured using the Summation of Residual Harmonics (SRH, [28]) method. All feature contours are resampled to be updated every 10 ms.

2.2. Classification

For the classification we utilise a further development of support vector machines (SVM, [29]), the so called fuzzy-input fuzzy-output SVMs (F^2SVM) which are capable of receiving soft labelled data and producing soft outputs with predicted memberships m_i assigned over multiple classes for each presented sample [18]. This method allows us to identify samples for which the F^2SVM is (a) certain about a predicted voice quality, (b) if there is some confusion between classes, or (c) an intermediate sample is observed. This proposed method has already been successfully utilised for the classification of voice qualities in previous work [18]. Here, we train the F^2SVM on the training corpora in the three voice qualities (i.e. lax, modal, tense) and categorise each sample from the audiobook data without knowing its ground truth value. The F^2SVM then generates predictions for the class memberships of each sample from the audiobook data. These predicted membership assignments m_i for each sample are ideal for a post-classification clustering. We use the standard k-means as the clustering algorithm of choice and utilise the clustered samples in the perception tests to subjectively evaluate the capability of the approach to extract targeted voice qualities from the audiobook data.

3. EXPERIMENTAL PROTOCOL

3.1. Speech data

The CereVoice speech synthesis system uses sub-corpora of neutral, tense and lax voice quality data in order to produce subtle changes in emotion [30, 6]. These sub-corpora have been recorded over a five

year period across several languages (English, French, German, Italian, Japanese), and covering different accents of English (RP, General American, Scottish accent, Irish accent, Northern England, Midlands). This totalled 8 hours and 11 minutes of speech data. All of the first set of recordings (**Acted**) were carried out in a recording studio with a high performance head mounted microphone. Voice talents were shown by an expert phonetician how to articulate lax voice and tense voice and were supervised during the recordings. Despite this, there still exists variability in terms of the articulation of the voice qualities. Along with the lax and tense recordings, we also use the recordings of the voice talents speaking with a neutral voice quality, which we consider as our modal voice set. Admittedly, a person's neutral speaking voice can be inherently lax or tense, and may not strictly correspond to the definition of modal voice given above (Section 1). As a result this voice quality class inevitably contains the largest variability.

The second set of speech data (**Audiobook**) consisted of a selection of two open-source audiobooks (librivox.org). The audiobooks were selected on the basis that the books were read in a lively fashion involving frequency variation in voice quality. The first audiobook was *A tramp abroad*, by Mark Twain, read by John Greenman (2166 utterances, approx. 2 hrs). The second was *Pride and Prejudice* by Jane Austin, read by Karen Savage (2548 utterances, approx. 2.5 hours). Audio was aligned with the text, and segmented into individual utterances using the method described in [31]. All speech data was down-sampled to 16 kHz for analysis. For the audiobook data the utterances selected all had a duration between 2 and 5 seconds.

3.2. Objective evaluation

The classification approach is evaluated in two steps. First we need to confirm the classifier's capabilities to accurately predict voice qualities over a large variety of speakers. Second, we evaluate if the audiobook data contains varied enough samples that spread over the full range of targeted voice qualities.

For the first step of the evaluation, we train and test the F^{2}SVM following a leave one speaker out protocol on the acted speech data. For this we train the classifier on 18 out of the 19 available speakers and test on the remaining one speaker. This process is repeated for all of the speakers and leads to a speaker independent classification assessment. For the classifier-training and to answer RQ2, we utilise two different sets of features; one set contains f_0 and MFCC features, and the second set additionally includes the voice quality features introduced in Section 2.1. The performance of the F^{2}SVM trained on both feature sets is compared and the overall results for the acted data classification are reported in Section 4.

For the second step of evaluation, we train the F^{2}SVM on all the available 19 speakers and run the classification for the samples available in the two analysed audiobooks, for which no ground truth labels are available. The predicted fuzzy outputs are clustered using k-means with k = 5. The three dimensional centroids of the k-means algorithm are initialised with the three extreme cases (i.e. one of the three dimensions is 1 and two are set to 0) - the voice quality has clearly been detected - and two intermediate samples where there is considerable overlap of membership assignments between the neighbouring classes lax and modal as well as modal and tense (i.e. two of the entries are set to 0.5 whereas the third is 0). The centroids and number of assigned samples for each cluster for both audiobooks is illustrated in Section 4. Samples that are assigned to the three clusters that were initialised to be either clearly lax, modal, or tense are then selected to be evaluated subjectively.

3.3. Subjective evaluation

In order to be able to compare human perception with the automatic clustering, and thereby address RQ3, we opted to conduct a web-based listening test in which we present two groups of stimuli, the first a set of acted utterances from the three voicing types, the second a set of utterances from audiobook data categorised into the three voicing types by the classifier. We then test inter-rater agreement with the baseline acted voice types - did the voice talent succeed in creating the voice type, and the classifier - have we successfully classified the voice types to reflect human classification. Further, we are interested in the effects and sources of disagreement.

For the web-based listening test, participants were initially presented with reference lax, modal and tense voice utterances produced by John Laver [1] and also by a female phonetician trained in the Laver labelling scheme. The participants were further asked to use headphones in order to reduce the influence of environmental noise. Participants then were presented with 36 utterances and had to select the most appropriate voice quality on a five-point scale: lax (1), lax-modal (2), modal (3), modal-tense (4), and tense (5), for each utterance. The mixed classes lax-modal and modal-tense allowed participants to signal uncertainty between the two class labels. The presented stimuli were selected from 12 sets of utterances: 2 datatypes $\{$acted and audiobook$\}$ $\times$ 2 genders $\{$male and female speakers$\}$ $\times$ 3 voice qualities $\{$lax, modal, and tense$\}$. Each set contained 10 unique utterances, resulting in 120 utterances in total. For each participant in the evaluation 3 utterances were randomly selected from each set giving 36 stimuli for each participant. The order of the stimuli was randomised.

As mentioned above, audiobook samples automatically clustered into the three confident voice quality clusters (cf. Figure 1) were qualified to be selected for the perceptual evaluation. For the acted speech data, one male and one female were selected. The ten utterances for each set were selected randomly from the available utterances.

4. RESULTS

4.1. Objective evaluation

As mentioned in Section 3.2, we evaluate the performance of the classifier on the acted speech. Here, we can observe a performance of 56.78% (σ = 12.57) accuracy for the F^{2}SVM based on MFCCs and f_0 only. When including the voice quality features introduced in Section 2.1, the performance significantly (pairwise t-test: T(18) = 3.46, p $<$ 0.003) rises 7.62% to 64.40% (σ = 13.96) accuracy for the speaker-independent leave one speaker out validation. This demonstrates the value of the voice quality features, with an even increase in accuracy for the three classes, and clearly answers RQ1.

Despite this improvement, the overall classification score is only moderately good. Unsurprisingly this seems to be mainly due to the difficulty in accurately classifying modal voice (as shown in Table 1). This may be largely explained by the high degree of variability in the speakers neutral voice quality, i.e. someone's 'normal' speaking voice may be inherently lax or tense. Nevertheless, there is rather effective classification of lax and tense voice with relatively little confusion between the classes lax and tense. These findings correspond to RQ2 and although reasonably effective speaker independent classification of lax and tense is demonstrated, further efforts need to be applied in order to improve classification performance on the modal voice quality class.

The second step of the objective evaluation aims towards identifying if the audiobooks contain samples with varying voice quality. For this we predict the voice quality of all the speech samples in

	MFCC+f_0			MFCC+f_0+VQ		
	L	**M**	**T**	**L**	**M**	**T**
Lax	60.93	26.61	12.44	**66.82**	24.62	8.54
Modal	29.60	41.02	29.36	23.62	**50.86**	25.50
Tense	13.62	17.98	68.39	7.89	16.58	**75.51**

Table 1. Confusion matrices for the leave one speaker out classification for both feature sets with and without voice quality features. Values are in %.

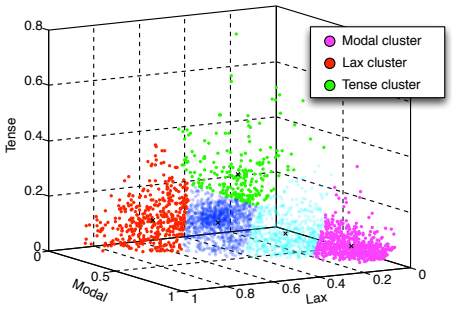

Fig. 1. Sample k-means clustering of audiobook *Pride and Prejudice*. The outer clusters in the corners are corresponding to the clear voice qualities (i.e. lax, modal and tense). The inner two clusters are mixed and not used in the perception test. The cluster centroids are marked with a black cross.

the audiobook data and cluster them in five distinct clusters, as visualised in Figure 1, using the F^{2}SVM trained on the acted speech with all 19 speakers. The corners of the triangle shaped space seen in Figure 1 represent the areas where the classifier clearly could identify a dominant voice quality to be present, the inner two clusters (dark and light blue) are intermediate observations with mixed voice qualities. In Table 2, the cluster centroids for the three main clusters (i.e. lax, modal, and tense) are listed. The values correspond to the mean assigned memberships by the F^{2}SVM for samples belonging to the corresponding cluster. The cluster initialisations correspond to the five possible answers in the perception tests introduced in Section 3.3. The cluster centres strongly correspond to the initialised centroid weights, i.e. the target voice quality for the cluster, except for the tense cluster of the audiobook data of *Pride and Prejudice*. In total 436 samples of the audiobook *Pride and Prejudice* were associated to the lax cluster, 855 of the *A Tramp Abroad* audiobook. 606 samples of *Pride and Prejudice* were assigned to the modal cluster and 717 of the *A Tramp Abroad* audiobook. Lastly, 235 samples of *Pride and Prejudice* were assigned to the tense cluster and only 44 of the *A Tramp Abroad* audiobook.

4.2. Subjective evaluation

30 participants completed the subjective evaluation described in Section 3.3 and display high inter-rater reliability, as assessed using Krippendorff's α (All samples: $\alpha = 0.87$, Acted samples: $\alpha = 0.95$, Audiobook samples: $\alpha = 0.75$). We also compute α for the hard classification output (i.e. the class with maximal membership assigned $\arg\max_i(m_i)$) and the majority vote over all the participants for all the audiobook samples. This gave an $\alpha = 0.74$ and demonstrates strong agreement between human ratings and the classifier output. This is further emphasised by the observation that 88 % of audiobook samples were rated within 1 numerical rating point of the hard classification output (i.e. this includes, for example, a sample classified as lax and rated as lax-modal).

We further conduct several ANOVA in order to identify the sources of disagreement within the above observations. For the first set of analysis we use the absolute difference between the majority vote of the human rating and the hard classification output as the dependent variable. The independent variables are **(1.i)** gender of the speaker **(1.ii)** data source (i.e. audiobook data or acted data). We observe the following: **(1.i)** the mean difference between human ratings and classifier output is significantly [F(1,118) = 26.42, p $<$ 0.0001] higher for female speakers ($\mu = 1.06$; $\sigma = 0.89$) than for male speakers ($\mu = 0.36$; $\sigma = 0.55$). **(1.ii)** the mean difference between human ratings and classifier output is not significant [F(1,118) = 0.44, p = 0.507] when comparing acted speech ($\mu = 0.76$; $\sigma = 0.94$) and audiobook samples ($\mu = 0.66$; $\sigma = 0.68$).

	Pride and Prejudice			A Tramp Abroad		
	L	**M**	**T**	**L**	**M**	**T**
C-Lax	**0.67**	0.22	0.11	**0.80**	0.18	0.02
C-Modal	0.14	**0.80**	0.06	0.17	**0.80**	0.03
C-Tense	0.37	0.35	**0.28**	0.08	0.24	**0.68**

Table 2. Cluster centroids for the samples of the two audiobooks. C-Lax, C-Modal, C-Tense represent the cluster centroids for the respective voice qualities as seen in Figure 1.

The second set of ANOVA investigates the absolute differences between the actual label for the acted samples and the **(2.i)** human majority rating and **(2.ii)** the hard classifier output. We observed the following: **(2.i)** the differences are significantly higher for the female voice ($\mu = 0.66$; $\sigma = 0.95$) than for the male voice ($\mu = 0.13$; $\sigma = 0.35$) with p = 0.006 and F(1,58) = 8.21 for the ANOVA test.

(2.ii) the differences between the automatic ratings for the acted speech of the female speaker are larger ($\mu = 0.60$; $\sigma = 0.93$) than for the male speaker ($\mu = 0.20$; $\sigma = 0.61$), however, the ANOVA does not show strong statistical significance with p = 0.054 and F(1,58) = 3.87 for the ANOVA test. There are no significant differences between the mean absolute differences of **(2.i)** and **(2.ii)** in a paired t-test, which indicates that the automatic classification yields comparable results to human performance.

The combination of the agreement between the human ratings and the classifier output (as shown by α = 0.74) together with the findings from the ANOVA analysis provide ample evidence to address RQ3 and answer that the proposed method indeed does produce a separation of voice quality in unlabelled corpora of expressive speech at a level which closely approximates human ratings.

5. DISCUSSION AND CONCLUSION

This study presented a classification approach involving a feature set consisting of state-of-the-art voice quality features used with fuzzy-input fuzzy-output support vector machines (F^{2}SVM). Speaker independent experiments on acted speech with discrete voice qualities showed effective classification of lax and tense voice. Lower classification of modal voice was observed and this can be largely attributed to high degree of variability in modal class of samples which were read in a neutral voice quality. A formal labelling of the voice quality of the neutral utterances would help to reduce this variability and would inevitably lead to better classification of the modal class. Findings from the subjective evaluation reveal that the clustering approach of the soft output of the F^{2}SVM classifier was highly effective in separating lax, modal and tense voice in unlabelled corpora of expressive speech, with potential for improvement for female voices. This makes the proposed approach an extremely useful method for handling this kind of data and has important implications for expressive speech synthesis, as well as other speech technology applications. It's potential for expressive speech synthesis, in particular, is one which we intend to exploit in future work.

6. REFERENCES

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

[2] C. Gobl and A. Ní Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Communication*, vol. 40, pp. 189–212, 2003.

[3] M. Gordon and P. Ladefoged, "Phonation types: a cross-linguistic review," *Journal of Phonetics*, , no. 29, pp. 383–406, 2001.

[4] J.P. Cabral, S. Renals, J. Yamagishi, and K. Richmond, "HMM-based speech synthesiser using the LF-model of the glottal source," *Proceedings of ICASSP, Prague, Czech Republic*, pp. 4704–4707, 2011.

[5] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, "Utilizing glottal source pulse library for generating improved excitation signal for HMM-based speech synthesis," *Proceedings of ICASSP, Prague*, pp. 4564–4567, 2011.

[6] M. P. Aylett, B. Potard, and C. J. Pidcock, "Expressive speech synthesis: Synthesising ambiguity," in *ICASSP13*, submitted.

[7] M. Lugger and B. Yang, "The relevance of voice quality features in speaker independent emotion recognition," *Proceedings of ICASSP, Honolulu, Hawaii*, vol. 4, pp. 17–20, 2007.

[8] M. Tahon, G. Degottex, and L. Devillers, "Usual voice quality features and glottal features for emotional valence detection," *Proceedings of Speech Prosody, Shanghai, China*, 2012.

[9] P. Zelinka, M. Sigmund, and J. Schimmel, "Impact of vocal effort variability on automatic speech recognition," *Speech Communication*, vol. 54, no. 6, pp. 732–742, 2012.

[10] E. Shriberg, M. Graciarena, H. Bratt, A. Kathol, S. Kajarekar, H. Jameel, C. Richey, and F. Goodman, "Effects of vocal effort and speaking style on text-independent speaker verification," *Proceedings of Interspeech, Brisbane, Australia*, pp. 609–612, 2008.

[11] G. Fant, *The acoustic theory of speech production*, Mouton, Hague (2nd edition 1970), 1960.

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

[13] G. Degottex, A. Roebel, and X. Rodet, "Phase minimization for glottal model estimation," *IEEE Transactions on Audio Speech and Language processing*, vol. 19, no. 5, pp. 1080–1090, 2011.

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

[15] J. Kane and C. Gobl, "Wavelet maxima dispersion for breathy to tense voice discrimination," *IEEE Transactions on Audio Speech and Language processing*, Under Review.

[16] M. Lugger, F. Stimm, and B. Yang, "Extracting voice quality contours using discrete hidden markov models," in *Proceedings of Speech Prosody 2008*, Campinas, Brazil, 2008, pp. 29–32, ISCA.

[17] K. Stevens and H. Hanson, "Classification of glottal vibration from acoustic measurements," *Vocal fold physiology*, pp. 147–170, 1994.

[18] S. Scherer, J. Kane, C. Gobl, and F. Schwenker, "Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification," *Computer Speech and Language*, vol. 27, pp. 263–287, 2013.

[19] N. Obin, "Cries and whispers: Classification of vocal effort in expressive speech," *Proceedings of Interspeech, Portland, Oregon, USA*, 2012.

[20] T. Drugman and T. Dutoit, "Oscillating Statistical Moments for Speech Polarity Detection," *Proceedings of Non-Linear Speech Processing Workshop (NOLISP11), Las Palmas, Gran Canaria, Spain*, pp. 48–54, 2011.

[21] J. Kane and C. Gobl, "Evaluation of glottal closure instant detection in a range of voice qualities," *Speech Communication*, vol. 55, pp. 295–314, 2013.

[22] P. Alku, T. Bäckström, and E. Vilkman, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Communication*, vol. 11, no. 2-3, pp. 109–118, 1992.

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

[24] T. Hacki, "Klassifizierung von glottisdysfunktionen mit hilfe der elektroglottographie," *Folia Phoniatrica*, pp. 43–48, 1989.

[25] G. Fant, J. Liljencrants, and Q. Lin, "A four parameter model of glottal flow," *KTH, Speech Transmission Laboratory, Quarterly Report*, vol. 4, pp. 1–13, 1985.

[26] R. P. Brent, *Algorithms for Minimization Without Derivatives.*, Englewood Cliffs, NJ: Prentice-Hall,, 1973.

[27] C. d'Alessandro and N. Sturmel, "Glottal closure instant and voice source analysis using time-scale lines of maximum amplitude," *Sadhana*, vol. 36, no. 5, pp. 601–622, 2011.

[28] T. Drugman and A. Alwan, "Joint Robust Voicing Detection and Pitch Estimation Based on Residual Harmonics," *Proceedings of Interspeech, Florence, Italy*, pp. 1973–1976, 2011.

[29] K. P. Bennett and C. Campbell, "Support vector machines: hype or hallelujah?," *ACM Special Interest Group on Knowledge Discovery and Data Mining Explorations Newsletter*, vol. 2, no. 2, pp. 1–13, 2000.

[30] M. P. Aylett and C. J. Pidcock, "The cerevoice characterful speech synthesiser sdk," in *AISB*, 2007, pp. 174–8.

[31] N. Braunschweiler, M. Gales, and S. Buchholz, "Lightly supervised recognition for automatic alignment of large coherent speech recordings," *Proceedings of Interspeech, Makuhari, Japan*, pp. 2222–2225, 2010.