Phase Distortion Statistics as a Representation of the Glottal Source: Application to the Classification of Voice Qualities

Gilles Degottex, Nicolas Obin

Phase Distortion Statistics as a Representation of the Glottal Source: Application to the Classification of Voice Qualities

Gilles Degottex[1] *and Nicolas Obin*[2]

[1]University of Crete and Foundation for Research and Technology-Hellas, Heraklion, Greece
[2]IRCAM, UMR STMS IRCAM-CNRS-UPMC, Paris, France

degottex@csd.uoc.gr, nobin@ircam.fr

Abstract

The representation of the glottal source is of paramount importance for describing para-linguistic information carried through the voice quality (e.g., emotions, mood, attitude). However, some existing representations of the glottal source are based on analytical glottal models, which assume strong a priori constraints on the shape of the glottal pulses. Thus, these representations are restricted to limited number of voices. Recent progresses in the estimation of the glottal models revealed that the Phase Distortion (PD) of the signal carries most of the information about the glottal pulses. This paper introduces a flexible representation of the glottal source - based on the short-term modelling of the phase distortion. This representation is not constrained by a specific analytical model, and thus can be used to describe a larger variety of expressive voices. We address the efficiency of this representation for the recognition of various voice qualities, with comparison to MFCC and standard glottal source representations.

Index Terms: glottal source, phase distortion, voice quality.

1. Introduction

The glottal source contains most of the para-linguistic information that convey the expressivity of a voice [1, 2, 3]. Consequently, accurate representation of the glottal source characteristics is of primary importance in many automatic speech processing systems, from the extraction of para-linguistic information (emotions [4, 5], among others [6, 7]), speaker recognition [8], automatic voice casting [9], etc. In particular, the glottal source is closely related to the voice quality [1], which plays a central role in speech communication (linguistic [2] and para-linguistic [3]). Furthermore, the automatic recognition and classification of the voice quality has recently raised as major research topic [10, 11, 12]. For speech classification, the signal representation of the glottal source characteristics, the features, must represent most of the properties of the glottal source.

According to the source/filter model of the speech production, a speech signal results from the filtering of the Glottal Source (GS) by the Vocal Tract Filter (VTF). Since the amplitude spectrum characteristics can be summarized into the widely known MFCC, many research works have focused on the VTF, which carries most of the spectral amplitude properties. However, the GS conveys extra-information to the VTF: the fundamental frequency $f_0(t)$ (mean, jitter); the shape of the glottal pulses (estimated from inverse filtering [13], or phase minimization using an analytical glottal model [14, 15]), the noisiness of the GS (SNR measures [16, 6, 7]).

Recent advances with glottal models have indicated promising results in recognition tasks related to voice quality [4, 17, 18]. However, the definition of analytical models assumes strong a priori constraints which considerably limit the range of GS properties that can be represented. Accurate glottal model estimations [14, 19] are based on the minimization of Phase Distortion (PD) [20] between the speech signal and the glottal model [15, 14] (see also Sec. 2.1). This indicates that the PD of the signal carries all of the crucial information about the glottal pulses' shape, which can be used directly in classification [17], thus, avoiding the glottal model limitations. Additionally, as shown later on, PD can be used to measure the variations of the glottal pulses over time, which might also be relevant to describe the noisiness of the GS.

In this paper, in order to avoid the limitations of the glottal models and obtain a more flexible representation of the GS properties, we suggest to use short-term statistics of the PD. These statistics include: the PD's Mean (PDM) within a short-term slidding window, which is a robust correlate of the pulse's shape [15, 14], and, the PD's standard-Deviation (PDD), which measures the noisiness of the pulses' shape in the same window. The PD's computation require a Harmonic Model (HM) of the speech signal [21]. In our work, we used a full-band HM which avoids voiced/unvoiced decision in the time and frequency domains [22]. This full-band/*full-time* approach allows to avoid voice segmentation, which is usually an error-prone estimation procedure. We assess the efficiency of the suggested GS features for the recognition of various voice qualities (breathy, creaky, hoarse, and pressed), with comparison to MFCC and standard glottal source representations.

2. Representations of the Glottal Source

In order to compute the voice source features, the signal is first represented by a Harmonic Model (HM) at instants t_i [21, 22]. The HM parameters consist of frequencies $h \cdot f_0(t_i)$ (integer multiples h of the fundamental frequency $f_0(t_i)$), the amplitude $a_{i,h}$ and the instantaneous phase $\phi_{i,h}$. As suggested in [22], the HM model is employed in this work for the full-band of the signal, for the voiced segments and the unvoiced segments (assuming that f_0 values can be obtained in unvoiced segments within a controled range (70-700Hz in our work)). Note that it has been shown that this approach can be used for resynthesis of the speech signal [22]. Thus, the perceived elements are fully represented by the HM parameters. In our work, we used the SWIPEP method for the estimation of $f_0(t)$ [23]. The parameters $a_{i,h}$ and $\phi_{i,h}$ are estimated using the Least Square (LS) solution [21]. For estimation of the Rd-glottal parameter [24] presented later on, t_i is defined at regular intervals of 5ms. To estimate the short-term statistical model, an intermediate pitch synchronous analysis is first necessary, as detailed later on. The next sections describe the mean to estimate the voice source features using the harmonic parameters.

2.1. The Phase Distortion (PD)

The phase difference between two frequency components is called Phase Distortion (PD) in [20], whose perceived characteristics are already known [20, 25, 26]. Using a harmonic model, the phase difference between consecutive harmonics $\phi_{i,h+1} - \phi_{i,h}$ can be seen as a discrete approximation of a frequency derivative. Thus, the PD of a harmonic model is similar to the group-delay, whose perceived characteristics are also known [27] and whose applications are numerous [28, 29, 30].

In order to reveal the phase characteristics of the voice source and not those of the whole speech signal, the PD can be computed on a minimum-phase residual (e.g. linear prediction residual). For this purpose, similarly, we first estimate a spectral amplitude envelope $A_i(f)$ at each instant i through linear interpolation of the amplitude parameters $a_{i,h}$. The minimum-phase response of $A_i(f)$ is then computed through the real cepstrum [31] and the residual phase is obtained by: $\widetilde{\phi}_{i,h} = \phi_{i,h} - \angle A(hf_0(t_i))$.

In order to focus on the impulses' shape of the voice source, it is also necessary to remove the linear phase of the signal. Because of the frequency derivative effect in the PD, the linear phase becomes a constant. Thus, the PD computation can be simply normalized by its value at the 1st harmonic:

$$\text{PD}_{i,h} = \widetilde{\phi}_{i,h+1} - \widetilde{\phi}_{i,h} - \widetilde{\phi}_{i,1} \tag{1}$$

where we enforce $\widetilde{\phi}_{i,0} = 0$, since the DC is unreliable for acoustic signals. In [14], we have shown that (1) is directly linked to the maximum-phase component of the voice source[1]. This sole property allows to estimate glottal model parameters, as shown in [14, 15], and summarized in next section.

2.2. Analytical model: The Rd-Liljencrants-Fant model

As shown in [15], the PD can be used to estimate the Rd shape parameter of the Liljencrants-Fant (LF) model [24]. Compared to the full set of the LF parameters $\{t_e,t_p,t_a\}$, the Rd parameter represents also a limitation of the shapes of the LF model. However, because of numerical dependency between $\{t_e,t_p,t_a\}$, their estimation is far from robust [32, p.77]. Consequently, in this work, we used Rd, which currently offers the most reliable estimate of a glottal model parameter. Basically, high values of Rd indicate a laxed voice whereas low values indicate a tense voice [24]. The best method suggested in [15] estimates Rd by error minimization with respect to candidate Rd values:

$$\epsilon(Rd_i) = \frac{1}{H}\sum_{h=1}^{H}\left(\text{PD}_{i,h} - \text{LF}_h^{Rd_i}\right)^2 \tag{2}$$

with $H = 7$ according to [14] and $\text{LF}_h^{Rd_i}$ are the PD values of the LF model parametrized by Rd_i. A minimization of (2) leads to an estimate of Rd_i. Additionally, the error $\epsilon(Rd_i)$ can be used to compute a confidence value, in $[0, 1]$, of this estimate:

$$c(Rd_i) = 1 - \sqrt{\epsilon(Rd_i)}/\pi \tag{3}$$

which describes how well the glottal model fits the PD of the signal. As discussed in the introduction, glottal models are limited in flexibility whereas the useful information for estimation of Rd is fully available in $\text{PD}_{i,h}$.

2.3. Short-term statistical model: Mean and Deviation

In this paper, we suggest to characterize $\text{PD}_{i,h}$ in short-term windows across the speech signal. In voiced time-frequency regions, $\text{PD}_{i,h}$ is mainly related to the shape of the glottal pulse. In unvoiced time-frequency regions (e.g. fricatives and mid-high frequencies of vowels), we assume that $\text{PD}_{i,h}$ can also be used to characterize the voice source. More precisely, we assume that the time evolution of $\text{PD}_{i,h}$, which also represents the source impulses throughout adjacent frames, can reveal the randomness of the voice source. Therefore, in this paper, we suggest to characterize PD statistically, in short-term sliding windows. For this purpose, at each t_i, we extract a feature related to the average shape of the source impulses and another feature representing the local variation, the randomness, of this source shape.

The procedure of the feature extraction starts the following. Firstly, a constant number of periods is necessary in each short-time window used for computing PD's moments, as described in the next subsections. Thus, $\text{PD}_{i,h}$ values are computed pitch synchronously, using 4 analysis instants per period: $t_i = t_{i-1} + 0.25/f_0(t_{i-1})$ with $t_0 = 0$. According to experiments, 4 analysis instants are sufficient for an estimate of PD's moments with sufficient accuracy. Secondly, to remove the dependency of $\text{PD}_{i,h}$ from the harmonic structure, we interpolate $\text{PD}_{i,h}$ on a linear frequency scale, like a phase spectral envelope [33, 34], using 512 frequency bins up to the Nyquist, i.e. $\text{PD}_{i,h} \Rightarrow \text{PD}_i[k]$.

Since $\text{PD}_i[k]$ is defined on a wrapped support (e.g. $(-\pi, \pi]$), we estimate the average shape and the randomness of the voice source through estimation of the mean and standard-deviation of the wrapped normal distribution [35, 36] over a few periods. The mean, called Phase Distortion Mean (PDM) in the following, is estimated using:

$$\text{(PDM)} \qquad \mu_i[k] = \angle\left(\frac{1}{L}\sum_{l\in C} e^{j\text{PD}_l[k]}\right) \tag{4}$$

where $C = \{i - \frac{L-1}{2}, \cdots, i + \frac{L-1}{2}\}$ and we used $L = 13$ (3 periods). Three periods have been chosen experimentally, so that C is long enough for obtaining a reliable mean and short enough so that $\mu_i[k]$ can follow the variations of the speech signal across time.

As described above, the variance of $\text{PD}_i[k]$ owns the randomness of the voice source impulses. However, it also owns the smooth evolution, the trend, of the glottal pulse shape. Since this trend is already represented by $\mu_i[k]$, it is necessary to remove it, prior to the estimation of PD's standard-deviation, i.e. $\text{PD}_i[k] - \mu_i[k]$. Thus, the standard-deviation, called Phase Distortion Deviation in the following, is estimated by [35]:

$$\text{(PDD)} \qquad \sigma_i[k] = \sqrt{-2\log\left|\frac{1}{L}\sum_{l\in C} e^{j(\text{PD}_l[k]-\mu_l[k])}\right|} \tag{5}$$

with C and L as above.

Finally, the feature are resampled each 5ms through linear interpolation, in order to keep a time sychronicity with the other features used in the experiments. Fig. 1 shows examples of features extraction for various voice qualities. Through visual inspection, one can see that the PDD of pressed voice is smaller than of breathy voice, below 5 kHz. This can be explained by the quantity of noise, which is more important in the breathy voice, thus, resulting in a higher variance. The PDM of the creaky voice seems close to zero in most regions, which means that the frequency components are phase synchronous, like those of a Dirac delta function. This supports the idea that a glottal cycle in creaky voice mainly contains a simple impulse concentrated at a single time instant. However, even though

[1] The rather complicate definition of PD in [15](Eq. 3-5) is actually equal to Eq. (1).

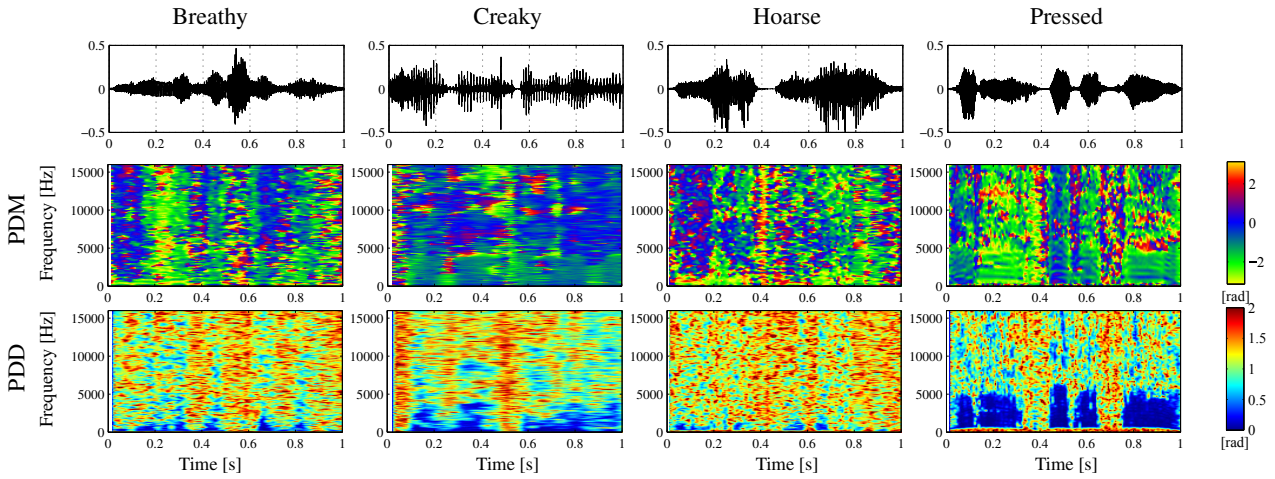

Figure 1: Examples of Phase Distortion's Mean (PDM) and Phase Distortion's Deviation (PDD) for 4 voice qualities. Whereas Breathy and Hoarse voices show substantial phase variance through PDD, the pressed voice has low PDD values below 5kHz in voiced segments. Also, Creaky's PDM has many time-frequency regions with values around zero. This suggests a phase synchronous frequency content, similarly to Dirac delta functions.

a hoarse voice is made of erratic glottal pulses and a breathy voice contains a substantial amount of additive noise, they look similar in Fig. 1. This can be explained by the fact that noisiness is described by PDD as a random variation of the source pulses' shape, like in a hoarse voice. Therefore, future works should complete the suggested features with a measurement of additive noise.

2.3.1. Compression

For this work, PDM and PDD features are used in a classification System. Thus, the dimensionality of the features (512 bins each 5ms) has to be reduced. PDD values are defined in $[0, \infty)$, like an amplitude spectral envelope. For this reason, we suggest to compress PDD, as if it was an amplitude envelope, using a Mel-CEPstral compression (MCEP) [37] of order 24. Because of the wrapped support of PDM, its compression is less straightforward. We suggest to compress it in a similar way to Mel-Frequency Cepstral Coefficients (MFCC). The first step of the traditional MFCC computation consists in energy measurement in mel-scale triangular frequency windows. Instead, because we are interested in compressing $\mu_i[k]$ and not in its absolute value (i.e. its "energy"), we compute the average $\mu_i[k]$, using Eq. (4), weighted by the mel-scale triangular windows. Finally, the resulting averaged $\mu_i[k]$ values are converted to cepstral coefficients (of order 24), similarly to the MFCC or MCEP computations.

3. Experiment

3.1. Speech Database

The speech database used in this study is the French version of the MASS EFFECT 3 video game containing 20,000 speech recordings of professional actors, around 500 roles, around 50 speakers, and a total of 20 hours of expressive speech. Each speech recording was recorded in professional conditions, and encoded into a 48 kHz-16 bits uncompressed format. A subset of 4,000 speech recordings was used for the manual labelling of speech into a large variety of classes in the context of automatic voice casting: from age/gender, voice quality, to emotions. The final representation includes: 6 dimensions, 14 classes, and 68 labels (see [9] for details). Among the number of classes, four voice quality classes were selected for this preliminary study: breathy, creaky, hoarse, and pressed.

3.2. Classification System

This section summarizes the main paradigms of the speech classification system used for the experiment. First, short-term speech characteristics are extracted from a speech recording. Then, statistical processing (UNIVERSAL BACKGROUND MODEL, TOTAL VARIABILITY SPACE) are used to summarize the statistical information of a speech recording (GMM SUPERVECTOR, I-VECTOR). Finally, a classifier (SUPPORT VECTOR MACHINE) is used for the classification of a speech recording. The remaining of this section details the paradigms of the classification system. The classification system is based on the IRCAMCLASSIFIER [38], which includes the ALIZÉE 3.0 [39] and the LIBSVM [40] libraries.

3.2.1. Acoustic Space Modeling: Universal Background Model and GMM supervector

The Universal Background Model (UBM) is used to model the distribution of the entire acoustic space, which is usually achieved with a standard Gaussian Mixture Model (GMM-UBM) [41]. Then, the means parameters of the UBM are adapted to each speech recording by using maximum a posteriori (MAP) adaptation [41]. Finally, each speech recording is represented by the mean vectors of the adapted mixture components:

$$\boldsymbol{\mu} = [\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_M]^T \tag{6}$$

where $\boldsymbol{\mu}$ - referred to as a GMM SUPERVECTOR - is the concatenation of all the mean vectors of the M mixture components.

3.2.2. Factor Analysis: Total Variability Space and i-vector

An i-vector is the compact representation of a high-dimensional representation of speech recording (GMM SUPERVECTOR) into a low-dimensional space called Total Variability space [42] (TV) - assuming an affine linear model (i.e., factor analysis):

$$\boldsymbol{\mu} = \boldsymbol{m}_{\boldsymbol{\mu}} + \mathbf{T}\mathbf{x} \tag{7}$$

where $\boldsymbol{\mu}$ is the adapted GMM-supervector of a speech recording, $\boldsymbol{m}_{\boldsymbol{\mu}}$ is the GMM-supervector corresponding to the UBM mean parameters, $\mathbf{T}$ is the $(M \times p)$ total variability matrix, and $\mathbf{x}$ is a p normally-distributed vector - referred to as an I-VECTOR. The total variability matrix $\mathbf{T}$ is modelled by Maximum-Likelihood (ML) and Expectation-Maximization (EM). The i-vector of a speech recording is determined by MAP adaptation [42].

	BREATHY	CREAKY	HOARSE	PRESSED	TOTAL
MFCC	76.2	73.6	82.9	77.4	**77.5**
RD	66.3	67.2	64.3	66.5	**66.0**
ΔGCI	70.0	71.0	75.1	67.4	**71.0**
PDM	**77.9**	71.6	81.2	**79.2**	**77.5**
PDD	**75.0**	72.0	82.3	**76.0**	**76.5**
MFCC+RD	77.7	75.2	83.5	77.8	78.6
MFCC+ΔGCI	77.4	74.8	85.0	76.7	78.4
MFCC+RD+ΔGCI	77.8	75.2	84.6	77.9	**78.9**
MFCC+PDM	78.7	75.7	84.7	81.1	80.0
MFCC+PDD	79.2	76.3	85.1	79.1	79.9
MFCC+PDM+PDD	80.2	76.7	85.4	81.8	**81.0**

Table 1: *Average balanced accuracy obtained for the classification of voice qualities.*

3.2.3. Classifier: Support Vector Machine

Among the number of existing classifiers, the Support Vector Machine (SVM) is a standard for speaker recognition and speech classification [43]. For each class (creaky, breathy, hoarse, pressed), the classification of a vector $\mathbf{x}$ (e.g., supervector, i-vector) corresponding to a speech recording is obtained with regard to the decision function:

$$f(\mathbf{x}) = \sum_{i=1}^{N} \omega_i K(\mathbf{x}, \mathbf{x}_i) + b \ , \qquad f \in [-1, 1] \qquad (8)$$

where: $< w_i, \mathbf{x_i}, b >_{i=1}^{N}$ are the parameters of the maximum-margin hyperplane determined during training (respectively: weights, support vectors, and offset), and $K(.,.)$ the SVM kernel [44].

In this study, a multi-label classification system is used to process each class separately. The multi-label system is constructed by converting the classification of multiple classes into multiple binary classifications [45]. First, each class is turned into a binary representation (i.e., yes/no). Then, a classifier is trained for each class separately, which results into C independent classifiers. This assumes that a speech recording can be simultaneously assigned to multiple classes (e.g., breathy and creaky) instead of a single exclusive class (e.g., breathy vs. all).

3.3. Experimental Setups

The front-end processing consisted in the extraction of short-term acoustic features: Mel-frequency cepstral coefficients (MFCC using 20ms Hanning window with 50% overlapping and 13 cepstral coefficients determined on 25 Mel-frequency bands); the Rd parameter and its confidence value (thus, 2 coefficients for the Rd feature, as summarized in Sec. 2.2); the jitter of Glottal Closure Instants (GCI) [6](a single coefficient); and our suggested glottal source representation derived from phase distortion (PDM and PDD, 25 coefficients determined on 25 Mel-frequency bands, each).

The classification system setups were defined as follows: N_{GMM} = 8 to 2048 (GMM-UBM) with diagonal covariance matrices, and p = 10 to 400 (Total Variability Space). For the classification, a standard SVM system with a Gaussian kernel [46] was used. During the training, each feature set is considered as a separate stream for the determination of the GMM-supervector, the i-vector, and the SVM parameters. During the classification, the decision is made by fusing the score obtained for each stream using average decision fusion.

The experiment was conducted in the form of a 10-fold cross validation. The balanced accuracy [47] was chosen for the classification performance, which is an alternative to the F-measure in the context of binary classification.

4. Results and Discussion

Table 1 reports the average recognition score for the four voice qualities. From the comparison of the single contribution of each acoustic feature: the standard GS representation (Rd and ΔGCI) presents recognition scores that are substantially lower (66.0% and 71.0%, respectively) than the baseline MFCC (77.5%). Conversely, the suggested PDM and PDD present recognition scores that are comparable (PDM, 77.5%, PDD, 76.5%) with MFCC (77.5%). This shows the efficiency of the phase distortion to capture the GS characteristics that convey voice qualities. Also, this supports the idea that PDM and PDD overcome the limitations of the glottal model and offer a more flexible representation of the GS properties.

From the comparison of the combination of the acoustic features: the standard GS representation combined with the MFCC (MFCC+Rd+ΔGCI) presents a recognition score that is slightly higher (78.9%) than MFCC (77.5%). However, the suggested GS representation derived from phase distortion (MFCC+PDM+PDD) presents a recognition score that is substantially higher (81.0%) than MFCC (77.5%). This shows the following two points. Firstly, this confirms that PDM and PDD better represent the voice properties than Rd (MFCC+Rd vs. MFCC+PDM+PDD) as mentionned above. Secondly, this shows that the phase distortion provides complementary information to that contained in the MFCC.

One may expect that PDD better describes the breathiness than PDM. On the contrary, Tab. 1 shows the opposite for the breathy (PDM:77.9%, PDD 75.0%) and the pressed qualities (PDM:79.2%, PDD 76.0%). This can be explained by the fact that the noisiness is implicitely represented in PDM through its variation between time-frequency regions, as shown in Fig. 1. Thus, PDM can partly encompass information measured by PDD and, in addition, it carries the average pulse's shape which is removed in PDD.

5. Conclusion

This paper presented a flexible representation of the glottal source based on the short-term statistics of the phase distortion (PD). This representation presents the advantage of not being constrained by a specific analytical model of the glottal source, and thus can be used to describe a large variety of voices. The efficiency of this representation has been proved for the recognition of various voice qualities. Further studies will address the use of PD information from para-linguistic information extraction (e.g., age/gender, emotions), voice casting, voice conversion and speech synthesis.

6. Acknowledgements

N. Obin was funded by the European FEDER project VOICE4GAMES. G. Degottex was funded by the Swiss National Science Foundation (grants PBSKP2_140021) and FORTH.

7. References

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

[2] N. Campbell and P. Mokhtari, "Voice Quality: the 4th Prosodic Dimension," in *International Congress of Phonetic Sciences*, 2003, pp. 2417–2420.

[3] C. Gobl and A. N. Chasaide, "The Role of Voice Quality in Communicating Emotion, Mood and Attitude," *Speech Communication*, vol. 40, no. 12, pp. 189–212, 2003.

[4] R. Sun, E. Moore, and J. F. Torres, "Investigating Glottal Parameters for Differentiating Emotional Categories with Similar Prosodics," in *Proc. ICASSP*, 2009.

[5] S. Koolagudi and K. Rao, "Emotion Recognition from Speech using Source, System and Prosodic Features," *International Journal of Speech Technology*, vol. 15, no. 2, pp. 265–289, 2012.

[6] N. Obin, "Cries and Whispers - Classification of Vocal Effort in Expressive Speech," in *Interspeech*, 2012.

[7] N. Obin and M. Liuni, "On the Generalization of Shannon Entropy for Speech Recognition," in *IEEE workshop on Spoken Language Technology*, 2012.

[8] P. De Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga, "Evaluation of Speaker Verification Security and Detection of HMM-Based Synthetic Speech," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, no. 8, pp. 2280–2290, 2012.

[9] N. Obin, A. Roebel, and G. Bachman, "On Automatic Voice Casting for Expressive Speech: Speaker Recognition vs. Speech Classification," in *Proc. ICASSP*, 2014.

[10] S. Scherer, J. Kane, C. Gobl, and F. Schwenker, "Investigating Fuzzy-Input Fuzzy-Output Support Vector Machines for Robust Voice Quality Classification," *Speech Communication*, vol. 27, no. 1, pp. 263–287, 2013.

[11] J. Kane, T. Drugman, and C. Gobl, "Improved Automatic Detection of Creak," *Computer Speech & Language*, vol. 27, no. 4, pp. 1028–1047, 2013.

[12] J. Kane and C. Gobl, "Wavelet Maxima Dispersion for Breathy to Tense Voice Discrimination," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 21, no. 6, pp. 1170 – 1179, 2013.

[13] P. Alku, "Glottal Wave Analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering," *Speech Communication*, vol. 11, no. 2-3, pp. 109–118, 1992.

[14] 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.

[15] ——, "Function of Phase-Distortion for Glottal Model Estimation," in *Proc. ICASSP*, 2011, pp. 4608–4611.

[16] D. G. Childers and C. K. Lee, "Vocal Quality Factors: Analysis, Synthesis, and Perception," *The Journal of the Acoustical Society of America*, vol. 90, no. 5, pp. 2394–2410, 1991.

[17] M. Tahon, G. Degottex, and L. Devillers, "Usual Voice Quality Features and Glottal Features for Emotional Valence Detection," in *Proc. Speech Prosody*, 2012, pp. 693–696.

[18] J. Kane, S. Scherer, M. Aylett, L.-P. Morency, and C. Gobl, "Speaker and language independent voice quality classification applied to unlabelled corpora of expressive speech," in *Proc. ICASSP*, May 2013, pp. 7982–7986.

[19] S. Huber, A. Roebel, and G. Degottex, "Glottal Source Shape Parameter Estimation using Phase Minimization Variants," in *Proc. Interspeech*, 2012.

[20] S. P. Lipshitz, M. Pocock, and J. Vanderkooy, "On the Audibility of Midrange Phase Distortion in Audio Systems," *J. Audio Eng. Soc*, vol. 30, no. 9, pp. 580–595, 1982.

[21] Y. Stylianou, "Harmonic plus Noise Models for Speech combined with Statistical Methods, for Speech and Speaker Modification," Ph.D. dissertation, Telecom Paris, France, 1996.

[22] G. Degottex and Y. Stylianou, "Analysis and Synthesis of Speech Using an Adaptive Full-Band Harmonic Model," *IEEE Audio, Speech and Lang. Proc.*, vol. 21, no. 10, pp. 2085–2095, 2013.

[23] A. Camacho, "SWIPE: A Sawtooth Waveform inspired Pitch Estimator for Speech and Music," Ph.D. dissertation, University of Florida, USA, 2007.

[24] G. Fant, A. Kruckenberg, J. Liljencrants, and M. Bavegdrd, "Voice Source Parameters in Continuous Speech, Transformation of LF-Parameters," in *Proc. ICSLP*, 1994, pp. 1451–1454.

[25] V. Hansen and E. R. Madsen, "On Aural Phase Detection: Part 1," *J. Audio Eng. Soc*, vol. 22, no. 1, pp. 10–14, 1974.

[26] ——, "On Aural Phase Detection: Part 2," *J. Audio Eng. Soc*, vol. 22, no. 10, pp. 783–788, 1974.

[27] H. Banno, K. Takeda, and F. Itakura, "The Effect of Group Delay Spectrum on Timbre," *Acoustical Science and Technology*, vol. 23, no. 1, pp. 1–9, 2002.

[28] B. Yegnanarayana, D. Saikia, and T. Krishnan, "Significance of Group Delay Functions in Signal Reconstruction from Spectral Magnitude or Phase," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 32, no. 3, pp. 610–623, Jun 1984.

[29] D. Zhu and K. Paliwal, "Product of Power Spectrum and Group Delay Function for Speech Recognition," in *Proc. ICASSP*, vol. 1, May 2004, pp. I–125–8 vol.1.

[30] T. Drugman, T. Dubuisson, and T. Dutoit, "Phase-based Information for Voice Pathology Detection," in *Proc. ICASSP*, 2011, pp. 4612–4615.

[31] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*, 2nd ed. Prentice-Hall, 1978.

[32] G. Degottex, "Glottal Source and Vocal-Tract Separation," Ph.D. dissertation, UPMC-Ircam, France, 2010.

[33] Y. Shiga and S. King, "Estimation of Voice Source and Vocal Tract Characteristics Based on Multi-frame Analysis," *Proc. Eurospeech*, 2003.

[34] J. Bonada, "High Quality Voice Transformations Based On Modeling Radiated Voice Pulses In Frequency Domain," in *Proc. Digital Audio Effects (DAFx)*, 2004.

[35] N. I. Fisher, *Statistical Analysis of Circular Data*. U.K.: Cambridge University Press, Oct. 1995.

[36] Y. Agiomyrgiannakis and Y. Stylianou, "Wrapped Gaussian Mixture Models for Modeling and High-Rate Quantization of Phase Data of Speech," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 17, no. 4, pp. 775–786, 2009.

[37] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An Adaptive Algorithm for Mel-Cepstral Analysis of Speech," in *Proc. ICASSP*, vol. 1, 1992, pp. 137–140 vol.1.

[38] G. Peeters, "A Generic System for Audio Indexing: Application to Speech/Music Segmentation and Music Genre," in *Proc. Int. Conf. on Digital Audio Effects*, 2007.

[39] A. Larcher, J.-F. Bonastre, B. Fauve, K. A. Lee, C. Levy, H. Li, J. S. Mason, and J.-Y. Parfait, "ALIZE 3.0 - Open Source Toolkit for State-of-the-Art Speaker Recognition," in *Interspeech*, 2013.

[40] C.-C. Chang and C.-J. Lin, "LIBSVM: a Library for Support Vector Machines," 2001.

[41] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[42] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, , and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[43] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support Vector Machines using GMM Supervectors for Speaker Verification," *IEEE Signal Proc. Letters*, vol. 13, no. 5, pp. 308–311, 2006.

[44] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[45] J.-J. Burred and G. Peeters, "An Adaptive System for Music Classification and Tagging," in *International Workshop on Learning the Semantics of Audio Signals*, 2009.

[46] N. Dehak and G. Chollet, "Support Vector GMMs for Speaker Verification," in *Odyssey: The Speaker and Language Recognition Workshop*, San Juan, Puerto Rico, 2006, pp. 1–4.

[47] D. G. Altman and J. M. Bland, "Statistics Notes: Diagnostic Tests 1: Sensitivity and Specificity," *British Medical Journal*, vol. 308, p. 1152, 1994.