# COVAREP – A COLLABORATIVE VOICE ANALYSIS REPOSITORY FOR SPEECH TECHNOLOGIES

*Gilles Degottex*[1]*, John Kane*[2]*, Thomas Drugman*[3]*, Tuomo Raitio*[4]*, Stefan Scherer*[5]

[1]Computer Science Department, University of Crete, Heraklion, Greece
[2]Phonetics and Speech Laboratory, Trinity College Dublin, Ireland
[3]TCTS Lab - University of Mons, Belgium
[4]Aalto University, Department of Signal Processing and Acoustics, Espoo, Finland
[5]Institute for Creative Technologies, University of Southern California, United States

## ABSTRACT

Speech processing algorithms are often developed demonstrating improvements over the state-of-the-art, but sometimes at the cost of high complexity. This makes algorithm reimplementations based on literature difficult, and thus reliable comparisons between published results and current work are hard to achieve. This paper presents a new collaborative and freely available repository for speech processing algorithms called COVAREP, which aims at fast and easy access to new speech processing algorithms and thus facilitating research in the field. We envisage that COVAREP will allow more reproducible research by strengthening complex implementations through shared contributions and openly available code which can be discussed, commented on and corrected by the community. Presently COVAREP contains contributions from five distinct laboratories and we encourage contributions from across the speech processing research field. In this paper, we provide an overview of the current offerings of COVAREP and also include a demonstration of the algorithms through an emotion classification experiment.

***Index Terms***— Speech processing, toolkit, glottal source, voice quality, sinusoidal modeling, spectral envelope

## 1. INTRODUCTION

Over the past few decades, a vast array of advanced speech processing methods have been developed, often offering significant improvements over the existing state-of-the-art. Such methods can have a reasonably high degree of complexity, which has certain implications. Firstly, as methods often consist of large blocks of code, reimplementations by researchers may not be consistent with the original version. This can significantly affect the behaviour of the method. Secondly, dependencies between methods can be complex. By combining methods together and building methods upon others, the results become increasingly difficult to analyze and understand. Indeed, the extensions and settings of the base methods are often superficially described in new publications in favor of descriptions of the newest contribution. Unfortunately, this can seriously affect the comparability of evaluations between publications.

Existing toolboxes partly address this issue [1, 2, 3, 4, 5]. Thanks to them, researchers can access the developments of a given laboratory and safely build extensions based on known and reproducible work. However, each existing toolbox is mainly the result of a single laboratory's work. From this approach, conventional and straightforward methods can be re-implemented, as well as the complex methods that are developed by the laboratory hosting the toolbox. The drawback is that such toolboxes do not actively look to compile complex algorithms developed by other laboratories. As a consequence, the base of complex methods which are shared and openly available is still fragile and many promising developments have been under-exploited or discarded in the past, with researchers tending to prefer conventional methods. To address this issue, a collaborative approach for sharing methods seems necessary.

In this paper, we present the COVAREP project[1], a collaborative and freely available repository initially gathering works from five laboratories. We envisage a range of benefits to the repository:

- *Strengthening of complex methods implementations*: Researchers are welcome to include their original implementations, thus resulting in a single de facto version for the speech community to refer to. The openly available methods will build a solid basis encouraging researchers from a wide range of speech-related disciplines to exploit them for their own research works.
- *More reproducible research*: COVAREP will allow fairer comparison of algorithms in published articles. Former methods can also be re-implemented while allowing users to discuss settings and method constituents which are not easily reproducible from original publications.
- *Participation and Feedback*: As a *GitHub* project, users are able to raise issues about bugs, suggest improvements, and add novel methods. We welcome contributions from a wide range of speech processing areas, including, but not limited to: speech analysis, synthesis, recognition, conversion, transformation, enhancement, voice quality analysis, expressive speech processing, speaker recognition, etc.

In terms of Intellectual Property (IP), getting contributing institutions to agree to a homogeneous IP policy would be close to impossible. As a result, COVAREP is a repository and not a toolbox, and each method has its own license associated with it. Though flexible to different license types, contributions need to have an open-source license which is compatible with the repository (e.g. GPL/LGPL, Apache). Further, to maintain a high standard, only published works in well-known speech conferences and journals can be added to the repository.

Supports are: G. Degottex by Swiss National Science Foundation (grants PBSKP2_134325, PBSKP2_140021), J. Kane by the Science Foundation Ireland (grant 09/IN.1/I2631 FASTNET), T. Drugman by the Fonds de la Recherche Scientifique and T. Raitio by EU FP7 Simple4All (grant 287678).

[1]http://covarep.github.io/covarep

A coding convention, though flexible enough not to discourage participations, also ensures the intelligibility of the code and the normalisation of its documentation. COVAREP is initially written in the Matlab® language, a widely used language in speech analysis and voice manipulation. However we strongly encourage authors to make the code compatible with GNU Octave (`octave.org`) to maximise usability. For more information, the reader is very welcome to visit the official website[1] which details the procedure for new contributions, the requirements for the license and the management of the repository as a whole.

The impact of the COVAREP concept will obviously be demonstrated over time. Nevertheless, in order to illustrate the current offering of the project, experiments in emotion recognition are shown in Section 3 using various analysis methods available in COVAREP. Prior to this application, Section 2 will first present a non-exhaustive overview of methods currently available in the repository.

## 2. ANALYSIS ALGORITHMS OF COVAREP

This section gives a description of the algorithms which have been implemented so far in COVAREP. The interconnection between these methods is shown in the workflow of Figure 1. Information necessary to perform pitch-synchronous analysis is first extracted from the speech signal: pitch tracking, polarity detection and determination of the Glottal Closure Instants (GCIs). These techniques are presented in Section 2.1. The resulting information is generally required to guarantee the high performance of subsequent methods: spectral envelope estimation and formant tracking (Section 2.3), sinusoidal modeling (Section 2.2), glottal analysis (Section 2.4) and phase processing (Section 2.5).

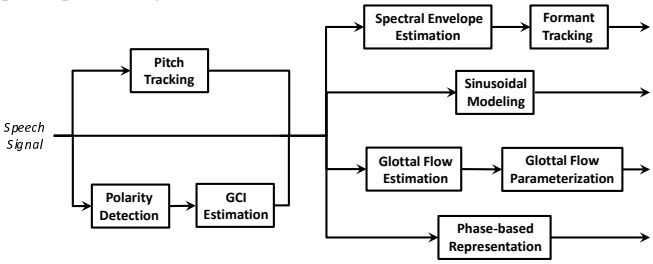


**Fig. 1**. Workflow of the methods implemented so far in COVAREP.

### 2.1. Periodicity and Synchronisation

*2.1.1. Pitch tracking*

Pitch tracking is one of the most fundamental problems in speech analysis. Fundamental frequency ($F_0$), which is the primary acoustic correlate of pitch, is mainly affected by the frequency of vocal fold vibration at the glottis and is used to represent the periodicity of the speech signal.

$F_0$ estimation from the speech signal is a non-trivial task considering the requirement of robustness that is needed in many speech processing applications. Given the fact that $F_0$ tracking has a very long history, there exists a vast number of different estimation algorithms. In COVAREP, a simple and robust pitch tracking algorithm is included: the *Summation of the Residual Harmonics* (SRH) method [6]. This method exploits the harmonic structure of the Linear Prediction (LP) residual signal to estimate both $F_0$ and voicing boundaries. SRH has been shown to be very robust to additive noise [6]. A large selection of different $F_0$ estimation algorithms will enable better possibilities for the development of other analysis tools.

*2.1.2. Speech polarity detection*

Speech polarity stems from the asymmetric excitation signal generated at the glottis, where the closure of the vocal folds creates a sharp discontinuity to the waveform. This discontinuity, represented as a sharp peak in the differentiated glottal flow signal, has a negative amplitude if the speech polarity is positive. Speech polarity has no perceptually relevant effect for humans, but it may have a dramatic impact on the performance of various analysis and synthesis techniques [7]. That is, among others, the case for the majority of approaches for GCI estimation or glottal analysis. For this purpose, COVAREP includes a speech polarity detection method based on the skewness of the LP residual signal [7].

*2.1.3. Glottal Closure Instant Detection*

Glottal Closure Instants (GCIs) are defined as the pseudo-periodic instants of significant excitation of the voice source [8]. Having knowledge of the precise GCI locations is crucial to perform many pitch-synchronous analysis procedures. COVAREP includes a state-of-the-art GCI detection algorithm called SEDREAMS (Speech Event Detection using the Residual Excitation And a Mean-based Signal) [9], which was shown to be one of the most accurate and robust GCI detection methods in [8]. Also a GCI estimation method (SE-VQ) dedicated to detecting GCIs for non-modal phonation is included [10]. A selection of GCI detection methods in COVAREP will enable further work on pitch-synchronous analysis.

### 2.2. Sinusoidal Modeling

The periodicity resulting from the glottal excitation translates to a harmonic structure in the speech spectrum. Thus, in the discrete Fourier transform (DFT) of a short time window of voiced speech signal ($\sim$3 periods), peaks appear in the amplitude spectrum corresponding to integer multiples of the fundamental frequency. These peaks carry the perceptually most significant spectral content for voiced speech and many models have been suggested for their representation. The Sinusoidal Model (SM) [11] directly extracts the amplitude peaks of the DFT spectrum; the Harmonic Model (HM) makes use of a time domain least square solution [12]; the quasi-harmonic model [13] assumes imperfect harmonicity; the adaptive quasi-harmonic model [14], the Adaptive Harmonic Model (aHM) and the extended adaptive quasi-harmonic model [15] allow frequency and amplitude demodulation during the estimation of the parameters, which provides accurate sinusoidal parameter estimates and high perceived quality of the reconstructed signal [16]. In COVAREP, a simple and unified interface allows the representation of speech signal using SM, HM or aHM models. Resynthesis is also possible through overlap-add or harmonic synthesis. Finally, by exploiting sinusoidal and harmonic model parameters, more abstract models can be built, such as spectral envelopes [17] or glottal flow parametrization [18, 19].

### 2.3. Spectral Envelope Estimation and Formant Tracking

Estimation of the amplitude spectral envelope is a recurrent subject in speech processing for approximation of the vocal tract filter response. COVAREP includes an estimator of the so called "true-envelope" (TE) [20, 21, 22]. This envelope is computed directly on the spectrum of the DFT of a windowed speech signal. On the same spectrum, amplitude peaks can also be extracted providing sinusoidal or harmonic representation, as mentioned above. Sinusoidal and harmonic representations can then be used to estimate discrete envelope, like the Discrete All-Pole (DAP) [17] (also available in COVAREP) which assumes that the vocal tract response obeys an Auto Regressive (AR) model (see illustration in Fig. 2, top panel). Also temporally weighted LP methods are included in COVAREP, such as WLP [23], SWLP [24], and XLP [25], which aim at temporal emphasis of those parts of a speech frame that are most likely

to correspond to the vocal tract response, thus being more robust to additive noise or the interfering effect of the excitation harmonics.

Although the aforementioned techniques for spectral envelope estimation could be used to determine formant trajectories, COVAREP integrates a dedicated formant tracker whose performance has been shown to outperform the state-of-the-art [26]. This algorithm is based on processing the negative derivative of the argument of the chirp-z transform (termed as the differential phase, or group-delay spectrum). Note that no modeling is included in the procedure, but only peak picking on group delay spectrum. This method is effective at tracking high-order formants due to its enhanced resolution.

### 2.4. Glottal Analysis

#### *2.4.1. Glottal Flow Estimation*

Glottal flow (GF) estimation, also referred to as source-filter separation, is the process of estimating the vocal-tract (VT) and GF components from a speech signal. Separating these contributions is important as it enables their distinct characterisation and modeling, which is motivated by both physiological and perceptual considerations. It is important to note that GF estimation is different from the process of estimating the LP spectrum of speech and then using the inverse of the LP filter to get a residual signal. The difference is that LP residue spectrum is white, whereas the GF excitation exhibits the spectral characteristics of the voice source (e.g. spectral tilt and glottal formant). This distinction is important in many fields, such as the study of speech production and voice quality characterisation.

There are various ways to estimate the glottal flow signal (for a review, see e.g. [27, 28, 29]). COVAREP includes two of the most representative techniques for GF estimation: Iterative Adaptive Inverse Filtering (IAIF, [30]) and Complex Cepstrum-based Decomposition (CCD, [31]). IAIF is based on repetitively applying low and high-order LP and using inverse of the filters to estimate the GF signal and the VT filter. IAIF has been used and evaluated in various experiments (see e.g. [32, 33, 29]) and it has been shown to yield rather robust estimates of the GF. IAIF can be performed either synchronously to given GCIs or asynchronously. COVAREP includes implementations of both variants. Fig. 2 (bottom panel) shows an extracted GF derivative waveform along with GCIs detected using SEDREAMS. The second technique, CCD [31], is a non-parametric approach which exploits the phase properties of the speech signal to separate its GF and VT components: while the VT is a minimum-phase system, the GF open phase is know to be a maximum-phase (i.e. anticausal) signal [34].

GF estimation is a difficult blind separation problem since neither the VT response nor the GF contribution are actually observable. The research field of GF estimation in particular may benefit from the COVAREP philosophy. Indeed, there is a clear need for easy and fair comparisons to allow a better understanding of the complexity of the voice source and to encourage the development of more robust voice source analysis algorithms.

#### *2.4.2. Glottal Flow Parameterisation*

Since GF has an important contribution to the supra-segmental characteristics of speech and is known to significantly vary with changes in phonation type, glottal flow parameterisation finds useful applications in many areas of speech research. COVAREP includes algorithms for extracting several commonly used GF parameters: NAQ [35], QOQ [36], H1–H2 [37], HRF [38], and PSP [39]. Also a collection of new algorithms are included, such as the estimation of maxima dispersion quotient (MDQ) [40] and peak slope parameter [41], and the estimation of the $R_d$ shape parameter [42] of the Liljencrants-Fant (LF) glottal model [43] using the Mean Squared Phase (MSP) method based on MSPD2 [18, 19]. Also, a method for detecting creaky voice from speech signal is included and arose from a sequence of developments [44, 45, 46]. The specific algorithm included here is that described in [44] and involves the extraction of features related to the residual excitation with classification carried out using an artificial neural network (ANN).

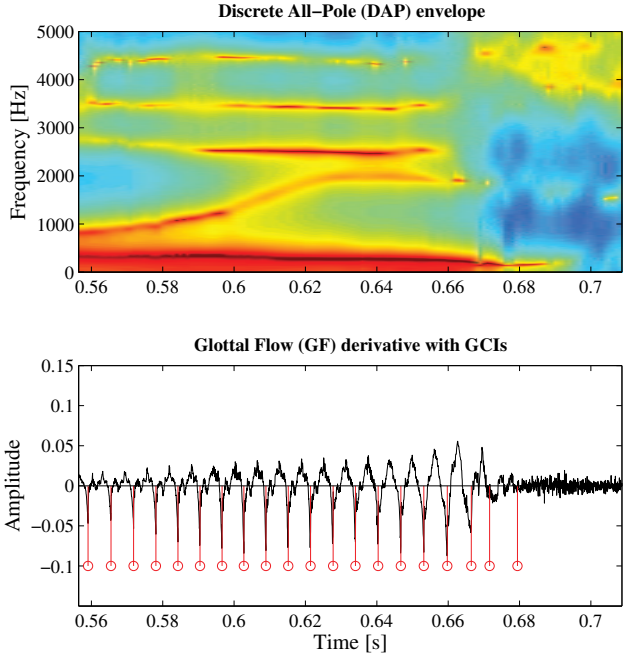


**Fig. 2**. Example showing analysis of a short speech segment using basic functionalities of COVAREP: DAP spectral envelope representation (top panel) and glottal flow derivative estimation along with glottal closure instants (bottom panel).

### 2.5. Phase Processing

Whereas the spectral amplitude information varies smoothly across time, the phase information is more complex to handle. Indeed, the integral of the frequency constantly wraps the phase values across time which makes the extraction of meaningful phase information difficult. Nevertheless, the perceived components of the phase information have been studied and measured [47, 48] and, by necessity, phase modeling is a recurrent subject in all speech synthesis applications. Recently, the Relative Phase Shift (RPS) has been suggested for the evaluation of perceptual importance of phase information in speech and also for speaker verification [49, 50, 51]. Its frequency derivative, the Phase Distortion (PD), which also assumes removal of the amplitude envelope minimum-phase, has been also suggested, first for glottal parameter estimation [52], then for emotional valence detection [53]. COVAREP includes both RPS and PD measurements in a simple function. The Chirp Group Delay (CGD) representation, proposed in [54], relies on a chirp (i.e the Fourier transform is evaluated on a contour in the $z$-plane different from the unit circle) analysis of the zero-phase version of the speech signal. This approach was shown to provide a high-resolved representation of formant peaks, and is the basis of the formant tracker presented in Section 2.3. Note that variations of the CGD were shown to be particularly suited to highlighting irregularities in phonation, and therefore for detecting voice disorders [55]. Pooling advances in phase processing in an open repository will definitely help to better understand and handle phase information in all speech processing techniques.

## 3. EXPERIMENTAL WORK – EMOTION RECOGNITION

In order to illustrate the usefulness of the features of COVAREP presented in Section 2, we carry out feature assessment and classification experiments on an emotion speech database. Note that this serves purely as an illustration and not as a comprehensive evaluation of the offerings of the repository.

### 3.1. Speech data

The speech data used here is the Berlin database of emotional speech. The database contains utterances spoken in 7 different acted emotions (neutral, boredom, sadness, disgust, fear, anger, happiness) by 10 professional actors (both male and female) and can be downloaded from: `http://pascal.kgw.tu-berlin.de/emodb/`. Two separate labelling schemes are used: 1) emotion vs neutral and 2) three levels of activation {passive, neutral, active}.

### 3.2. Feature extraction and assessment

A set of features based on COVAREP algorithms are extracted from each utterance (using COVAREP v1.1.0). As a baseline comparison, we include 12 MFCCs (excluding the energy-related $0^{th}$ coefficient) extracted on 25-ms frames with a 5-ms shift. We also extract an alternative set of MFCCs (TE-MFCCs) which are extracted from the *True Envelope* spectral representation rather than from FFT. A set of features are derived from the glottal source signal estimated by glottal inverse filtering (based on GCI-synchronous IAIF). These include: NAQ, QOQ, PSP and H1–H2. Two wavelet based features (peakSlope and MDQ) as well as $R_d$ derived by phase minimisation are also extracted. Note that the $R_d$ confidence measure is also included. Finally, the posterior probability of the creaky voice detection algorithm is included as an additional feature. Note that parameter contours are sampled simply as the median value in voiced regions (as detected by the SRH algorithm) for each sentence.

In order to investigate the discriminative power of the features included, we carry out an initial mutual information based feature assessment using the method described in [56]. Figure 3 shows the top 10 discriminative features (of emotion vs neutral) in terms of their relative intrinsic information. The peakSlope is found to be the most discriminative feature, followed by the first TE-MFCC coefficient. Interestingly, the confidence measure associated with the $R_d$ parameter is also found to have discriminative qualities for emotion vs neutral speech.

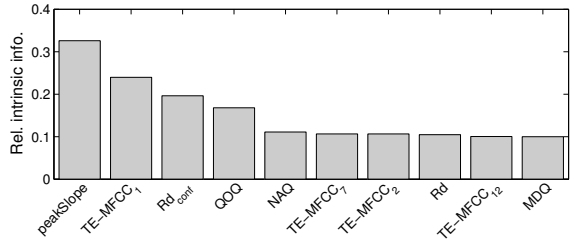


**Fig. 3**. Top 10 features in terms of relative intrinsic information.

### 3.3. Classification experiments

We carry out two speaker-independent classification experiments, one for emotion vs neutral (binary) and one using the activation labels (3-class). Five feature set variants are used: MFCCs, TE-MFCCs (MFCCs derived from *True Envelope* spectral representation), glottal/VQ (glottal source and voice quality related features), ALL (TE-MFCC and glottal/VQ combined), SEL (top 10 discriminative features selected from ALL set). We use support vector machines (SVMs) as our classifier, utilising a radial basis function (RBF) kernel. For the 3-class classification, a one-against-one setup is employed. All classification experiments involve a speaker independent, leave-one-speaker-out procedure where the classifier is trained on all but one speaker's data, and are then tested on the held-out data. The held-out speaker is then rotated until all speakers have been covered.

The results of the experiments in terms of error (%) are shown in Figure 4. For both classification experiments, the TE-MFCCs provide lower mean classification error compared to standard MFCCs. The iterative cepstral smoothing of the TE spectral envelope representation may be beneficial for making the spectral coefficients less biased towards harmonics and, hence, more independent of variation in $F_0$. For the emotion vs neutral (binary) experiment, the glottal/VQ features provide the highest mean classification error. However, closer inspection reveals that the MFCC classification is extremely biased towards the emotion class (mean error for neutral: 48 %, emotion: 82 %), whereas the Glottal/VQ result is much more balanced (mean error for neutral: 82 %, emotion: 73 %). Similarly, the ALL set results in highly biased results, whereas the SEL set is more biased but with a comparatively low error (18 %). Note that attempts to force a balance between the two classes in the data did not alleviate this bias problem.

For the activation (3-class) experiment, the glottal/VQ features provide the lowest mean error (26 %). Combination with the spectral features, even with feature selection, does not improve the classification. This demonstrates that the glottal/VQ features are effective at discriminating laxness (typically found in low activation emotion) and tenseness (in high activation) in the speech data.

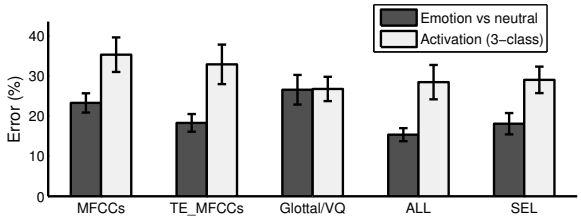


**Fig. 4**. Classification error (%) for emotion vs neutral (binary) and activation (3-class) classification experiments, plotted as a function of feature set variants. Data is expressed as mean ± standard error of the mean.

## 4. CONCLUSION

This paper presented a new repository for open-source speech processing algorithms called COVAREP. The main motivations for developing this repository are: to facilitate fair and reproducible research by having single de facto code versions of published algorithms, to improve the visibility and availability of newly developed state-of-the-art algorithms, and to encourage feedback and bug reports to improve the overall quality of the algorithms. An overview of the main algorithms currently available in COVAREP was presented, that arose from the contributions of 5 separate laboratories. Although the success of COVAREP will be judged over time, a small experimental procedure based on emotion recognition illustrated the potential of both glottal and voice quality related features as well as novel spectral envelope representations contained within COVAREP. Finally, we encourage researchers to consider contributing their published open-source algorithms to this project. We envisage that this platform can bring about significant improvements in the effectiveness and impact of speech processing research.

## 5. REFERENCES


[1] M. Brookes et al., "VOICEBOX: Speech processing toolbox for MATLAB," [Online], 2005, http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.

[2] K. Tokuda and K. Oura et al., "Speech signal processing toolkit (SPTK)," [Online], recent version 2012, http://sp-tk.sourceforge.net.

[3] F. Eyben, M. Wöllmer, and B. Schüller, "Opensmile: the Munich versatile and fast open-source audio feature extractor," in *Proc. ACM Multimedia (MM)*. 2010, pp. 1459–1462, ACM, http://www.openaudio.eu.

[4] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011.

[5] F. Eyben, M. Wöllmer, and B. Schüller, "OpenEAR - Introducing the Munich open-source emotion and affect recognition toolkit," in *Intl Conf. on Affective Comp. and Intell. Interaction and Workshops*, 2009, http://www.openaudio.eu.

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

[7] T. Drugman, "Residual excitation skewness for automatic speech polarity detection," *IEEE Sig. Proc. Lett.*, vol. 20, no. 4, pp. 387–390, 2013.

[8] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit, "Detection of glottal closure instants from speech signals: A quantitative review," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 20, no. 3, pp. 994–1006, 2012.

[9] T. Drugman and T. Dutoit, "Glottal closure and opening instant detection from speech signals," in *Proc. Interspeech*, 2009.

[10] J. Kane and C. Gobl, "Evaluation of glottal closure instant detection in a range of voice qualities," *Speech Commun.*, vol. 55, no. 2, pp. 295–314, 2013.

[11] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 34, no. 4, pp. 744–754, 1986.

[12] Y. Stylianou, *Harmonic plus Noise Models for Speech combined with Statistical Methods, for Speech and Speaker Modification*, Ph.D. thesis, TelecomParis, France, 1996.

[13] Y. Pantazis, O. Rosec, and Y. Stylianou, "On the properties of a time-varying quasi-harmonic model of speech," in *Proc. Interspeech*, 2008, pp. 1044–1047.

[14] Y. Pantazis, O. Rosec, and Y. Stylianou, "Adaptive AM-FM signal decomposition with application to speech analysis," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 19, no. 2, pp. 290–300, 2010.

[15] G.P. Kafentzis, Y. Pantazis, O. Rosec, and Y. Stylianou, "An extension of the adaptive quasi-harmonic model," in *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Proc. (ICASSP)*, 2012.

[16] G. Degottex and Y. Stylianou, "Analysis and synthesis of speech using an adaptive full-band harmonic model," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 21, no. 10, pp. 2085–2095, 2013.

[17] A. El-Jaroudi and J. Makhoul, "Discrete all-pole modeling," *IEEE Transactions on Sign. Proc.*, vol. 39, no. 2, pp. 411–423, 1991.

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

[19] A. Roebel G. Degottex and X. Rodet, "Function of phase-distortion for glottal model estimation," in *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Proc. (ICASSP)*, 2011, pp. 4608–4611.

[20] S. Imai and Y. Abe, "Spectral envelope extraction by improved cepstral method," *Electronics and Communication*, vol. 62-A, no. 4, pp. 10–17, 1979, in Japanese.

[21] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," in *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Proc. (ICASSP)*, apr 1983, vol. 8, pp. 93–96.

[22] A. Roebel, F. Villavicencio, and X. Rodet, "On cepstral and all-pole based spectral envelope modeling with unknown model order," *Pattern Recognition Letters*, vol. 28, no. 11, pp. 1343–1350, 2007.

[23] C. Ma, Y. Kamp, and L.F. Willems, "Robust signal selection for linear prediction analysis of voiced speech," *Speech Commun.*, vol. 12, no. 2, pp. 69–81, 1993.

[24] C. Magi, J. Pohjalainen, T. Bäckström, and P. Alku, "Stabilised weighted linear prediction," *Speech Commun.*, vol. 51, no. 5, pp. 401–411, 2009.

[25] J. Pohjalainen, R. Saeidi, T. Kinnunen, and P. Alku, "Extended weighted linear prediction (XLP) analysis of speech and its application to speaker verification in adverse conditions," in *Proc. Interspeech*, 2010, pp. 1477–1480.

[26] B. Bozkurt, B. Doval, C. d'Alessandro, and T. Dutoit, "Improved differential phase spectrum processing for formant tracking," *Proc. ICSLP*, 2004.

[27] J. Walker and P. Murphy, "Advanced methods for glottal wave extraction," in *Nonlinear Analyses and Algorithms for Speech Processing*, M. Faundez-Zanuy et al., Eds., pp. 139–149. Springer Berlin/Heidelberg, 2005.

[28] P. Alku, "Glottal inverse filtering analysis of human voice production – A review of estimation and parameterization methods of the glottal excitation and their applications," *Sadhana*, vol. 36, no. 5, pp. 623–650, 2011.

[29] T. Drugman, B. Bozkurt, and T. Dutoit, "A comparative study of glottal source estimation techniques," *Comp. Speech & Lang.*, vol. 26, no. 1, pp. 20–34, 2012.

[30] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Commun.*, vol. 11, no. 2–3, pp. 109–118, 1992.

[31] T. Drugman, B. Bozkurt, and T. Dutoit, "Causal-anticausal decomposition of speech using complex cepstrum for glottal source estimation," *Speech Commun.*, vol. 53, no. 6, pp. 855–866, 2011.

[32] P. Alku, J. Horacek, M. Airas, F. Griffond-Boitier, and A.-M. Laukkanen, "Performance of glottal inverse filtering as tested by aeroelastic modelling of phonation and FE modelling of vocal tract," *Acta Acustica united with Acustica*, vol. 92, pp. 717–724, 2006.

[33] P. Alku, B. Story, and M. Airas, "Estimation of the voice source from speech pressure signals: Evaluation of an inverse filtering technique using physical modelling of voice production," *Folia Phoniatrica et Logopaedica*, vol. 58, no. 1, pp. 102–113, 2006.

[34] T. Drugman, B. Bozkurt, and T. Dutoit, "Complex cepstrum-based decomposition of speech for glottal source estimation," in *Proc. Interspeech*, 2009, pp. 116–119.

[35] P. Alku, T. Bäckström, and E. Vilkman, "Normalized amplitude quotient for parameterization of the glottal flow," *J. Acoust. Soc. Am.*, vol. 112, no. 2, pp. 701–710, 2002.

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

[37] I. Titze and J. Sundberg, "Vocal intensity in speakers and singers," *J. Acoust. Soc. Am.*, vol. 91, no. 5, pp. 2936–2946, 1992.

[38] D. Childers and C. Lee, "Voice quality factors: Analysis, synthesis and perception," *J. Acoust. Soc. Am.*, vol. 90, no. 5, pp. 2394–2410, 1991.

[39] P. Alku, H. Strik, and E. Vilkman, "Parabolic spectral parameter – A new method for quantification of the glottal flow," *Speech Commun.*, vol. 22, no. 1, pp. 67–79, 1997.

[40] J. Kane and C. Gobl, "Wavelet maxima dispersion for breathy to tense voice discrimination," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 21, no. 6, pp. 1170–1179, 2013.

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

[42] G. Fant, "The LF-model revisited. Transformations and frequency domain analysis," *STL-QPSR*, vol. 36, no. 2–3, pp. 119–156, 1995.

[43] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 4, pp. 1–13, 1985.

[44] T. Drugman, J. Kane, and C. Gobl, "Automatic analysis of creaky excitation patterns," *Comp. Speech & Lang.*, 2013, submitted.

[45] J. Kane, T. Drugman, and C. Gobl, "Improved automatic detection of creak," *Comp. Speech & Lang.*, vol. 27, no. 4, pp. 1028–1047, 2013.

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

[47] 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.

[48] H. Banno, K. Takeda, and F. Itakura, "The effect of group delay spectrum on timbre," *Acoust. Sc. and Techn.*, vol. 23, no. 1, pp. 1–9, 2002.

[49] I. Saratxaga, I. Hernaez, D. Erro, E. Navas, and J. Sanchez, "Simple representation of signal phase for harmonic speech models," *Electronics Letters*, vol. 45, no. 7, pp. 381–383, 2009.

[50] I. Saratxaga, I. Hernaez, M. Pucher, and I. Sainz, "Perceptual importance of the phase related information in speech," in *Proc. Interspeech*. ISCA, 2012.

[51] P.L. 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 Lang. Proc.*, vol. 20, no. 8, pp. 2280–2290, 2012.

[52] G. Degottex, A. Roebel, and X. Rodet, "Function of phase-distortion for glottal model estimation," in *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Proc. (ICASSP)*, 2011, pp. 4608–4611.

[53] M. Tahon, G. Degottex, and L. Devillers, "Usual voice quality features and glottal features for emotional valence detection," in *Proc. Intl. Conf. on Speech Prosody*, 2012, pp. 693–696.

[54] B. Bozkurt, L. Couvreur, and T. Dutoit, "Chirp group delay analysis of speech signals," *Speech Commun.*, vol. 49, pp. 159–176, 2007.

[55] T. Drugman, T. Dubuisson, and T. Dutoit, "Phase-based information for voice pathology detection," *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Proc. (ICASSP)*, pp. 4612–4615, 2011.

[56] T. Drugman, M. Gurban, and J.-P. Thiran, "Relevant feature selection for audio-visual speech recognition," *IEEE Intl Workshop on Multimedia Signal Processing*, pp. 179–182, 2007.