Voice Activity Detection: Merging Source and Filter-based Information

Thomas Drugman, *Member, IEEE*, Yannis Stylianou, *Senior Member, IEEE*, Yusuke Kida, and Masami Akamine, *Senior Member, IEEE*

***Abstract*—Voice Activity Detection (VAD) refers to the problem of distinguishing speech segments from background noise. Numerous approaches have been proposed for this purpose. Some are based on features derived from the power spectral density, others exploit the periodicity of the signal. The goal of this letter is to investigate the joint use of source and filter-based features. Interestingly, a mutual information-based assessment shows superior discrimination power for the source-related features, especially the proposed ones. The features are further the input of an artificial neural network-based classifier trained on a multi-condition database. Two strategies are proposed to merge source and filter information: feature and decision fusion. Our experiments indicate an absolute reduction of 3% of the equal error rate when using decision fusion. The final proposed system is compared to four state-of-the-art methods on 150 minutes of data recorded in real environments. Thanks to the robustness of its source-related features, its multi-condition training and its efficient information fusion, the proposed system yields over the best state-of-the-art VAD a substantial increase of accuracy across all conditions (24% absolute on average).**

***Index Terms*—Excitation, information fusion, periodicity, voice activity detection.**

I. Introduction

VOICE ACTIVITY DETECTION (VAD) refers to the problem of distinguishing speech segments from background noise in an audio stream. This is a fundamental task which finds a wide range of applications in voice technology: speech coding [1], automatic speech recognition (ASR, [2]), audio surveillance and monitoring, speech enhancement, or speaker and language identification [3]. In the workflow of these applications, VAD is generally involved as the very first block. As a consequence, the main characteristics expected from a VAD algorithms are generally a high efficiency and robustness to noise, as well as a low computational latency.

Numerous studies have addressed the problem of VAD in the literature. Generally speaking, a VAD method consists of two successive steps: feature extraction and a discrimination model. Early works focused on energy-based features, possibly combined with the zero-crossing rate (ZCR) [4], [5]. These features are however highly affected in the presence of additive noise. Therefore, various other features have been proposed: autocorrelation-based features [6], [7], [8], Mel-Frequency Cepstral Coefficients (MFCCs) [7], line spectral frequencies [9], a cepstral distance [10], the skewness and kurtosis of the linear prediction (LP) residual [11] or periodicity-based features [12], [13], [8]. Some other methods are based on a statistical model of the Discrete Fourier Transform (DFT) coefficients [14], [15]. Other approaches exploit the fact that the speech and noise signals should have different variability properties [16], [17]. Finally, some studies have addressed the use of a combination of multiple features. These works differ by the way the features are combined: using a linear combination where the weights are trained via a minimum classification error in [18], a linear or a kernel discriminant analysis in [19] or a principal component analysis in [8].

The resulting acoustic information is then generally the input of a statistical model whose goal is to draw a decision about the presence or not of speech. Proposed approaches differ in whether they use a supervised framework or not. In the former case, several models have been used: Gaussian Mixture Model (GMM, [20]), Hidden Markov Model (HMM, [21]) or Multi-Layer Perceptron (MLP, [20]). Some other works state that the drawbacks of a supervised method are that large amounts of labeled training data are required and that they are sensitive to a mismatch between training and testing conditions [8], [22]. As a consequence, unsupervised approaches have been recently proposed in [23], [8], [22].

With respect to the state of the art, the main contributions of this letter are the following: *i)* to propose the use of robust source-related features for VAD purpose, *ii)* to assess the relative performance of source and filter-based features, *iii)* to investigate the best strategies to merge information from various feature sets, *iv)* to compare the proposed VAD system with existing algorithms on *real data*, *v)* to examine the generalization capabilities of a supervised approach when trained on a multi-condition dataset. Note that the two last points must be moderated as recent studies conducted VAD experiments on real-life videos [24], [25], possibly with a multi-condition training approach [25].

The letter is organized as follows. The proposed method is described in Section II. The protocol used throughout our experiments is presented in Section III and the results are discussed in Section IV. Section V finally concludes the letter.

Manuscript received December 14, 2014; revised February 16, 2015; accepted October 15, 2015. Date of publication October 27, 2015; date of current version January 14, 2016. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Björn Schuller.

The authors are with Toshiba Cambridge Research Laboratory, Cambridge CB4 0GZ, UK (e-mail: yannis.stylianou@crl.toshiba.co.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2015.2495219

II. Proposed Technique

According to the mechanism of voice production, speech is considered as the result of a glottal flow (also called *source* or *excitation* signal) filtered by the vocal tract cavities [26]. This physiological process motivates the goal of this letter as we believe it to be essential that a VAD exploits information from both these two complementary components of speech. The proposed VAD approach will be shown in Section IV to carry out a significant improvement over the best state-out-the-art approach. It is worth emphasizing on that this would not be possible without the *combined* effect of 4 main factors: the joint use of filter and source-related information, the design of robust source features, an efficient strategy of information fusion and a multi-condition training.

A. Filter-Based Features

According to the source-filter model of speech, the *spectral envelope*, defined as a smooth function passing through the prominent peaks of the spectrum [27], is the transfer function of the filter. Various ways to parameterize the spectral envelope have been proposed in the literature. In this work, the following representations are considered: the Mel Frequency Cepstral Coefficients (MFCCs, [28]), the Perceptual Linear Prediction coefficients (PLP, [29]) and the Chirp Group Delay (CGD) of the zero-phase signal which is a robust high-resolved representation of the filter resonances [30]. A vector of 13 coefficients is used for each feature type. The advantage to use such parameters is that they have been already shown to be efficient for ASR or speaker recognition purpose [30], [31], and can generally be of high interest in any speech technology application following the VAD. Most of the time, their computation is therefore already required and their integration in a VAD system can therefore be achieved at a very low computational cost.

B. Source-Related Features

The glottal flow has been recently shown to be useful in various voice technologies [32], [26]. Despite promising advances, it has been acknowledged in [26] that the weakest point of current glottal source processing algorithms is clearly related to their lack of robustness. It is therefore a challenging and still open problem to design source-related features for applications in adverse environments. One issue is the strong degradation of glottal flow estimation techniques when the speech signal gets noisier [33]. When working in adverse conditions, it is consequently preferable to use indirect measurements derived either from the speech signal or from the LP residue. In this work, we aim at using robust source-related features which are compatible with the noisy environments targeted by our VAD.

Various existing studies have already used excitation information for VAD. The periodicity of the speech signal has been exploited in [6], [7], [12], [13], [8]. Furthermore, features extracted from the LP residual have been used in [11]. In this work, we consider some of these features already proposed for VAD purpose, as well as some new other source-related measurements. Two popular features used from the early attempts are the log-energy of the speech signal [4] and the zero-crossing rate (ZCR, [5]). In [11], Nemer *et al.* proposed the use of high-order statistics of the LP residual. As suggested in that study, we included the skewness and kurtosis of the LP residual which are known to respectively characterize the polarity of the speech signal [34] and the sparsity of the excitation [35] at the glottal closure instants. Sadjadi recently proposed in [8] a VAD system using 4 voicing measures: the so-called harmonicity and clarity features derived from the average magnitude difference function (AMDF), the normalized LP error [11] and the Harmonic Product Spectrum (HPS, [8]).

In addition to the aforementioned features, we include three other source-related measurements. These latter features were proposed in previous studies and were here selected for their robustness properties. The first is the Cepstral Peak Prominence (CPP) which was originally proposed in [36] for the prediction of breathiness ratings. CPP is a measure of the amplitude of the cepstral peak at the hypothesized fundamental period. The two other features are extracted from the Summation of the Residual Harmonics (SRH) algorithm [37], a robust pitch tracker. The SRH criterion quantifies the level of voicing by taking into account the harmonics and inter-harmonics of the residual spectrum. The two features used in this work, referred to as SRH and SRH^* differ by the energy normalization or not of the residual spectrum. Note that the implementations of CPP and SRH are available from the COVAREP project [38].

C. ANN-Based Classification and Information Fusion

For our classification experiments, we opted for an ANN for its discriminant properties, its ability to model non-linear relations and for the convenience of the posterior probabilities it generates. Each ANN is made of a single hidden layer consisting of neurons whose activation function is an hyperbolic tangent sigmoid transfer function. As any parameter used by the proposed technique, the number of neurons was set on the development data. Performance was very similar using between 32 and 128 neurons, and we fixed this parameter to 32 in the remainder of this letter. The output layer is a simple neuron with a sigmoid function suited for a binary VAD decision. Note that we also tried to make use of recurrent neural networks. This however did not lead to a particular gain in performance while it increased the computational load.

Before being fed to the ANN, the feature vector x_t at time t goes through two processing steps. First, the feature trajectories are smoothed using a median filter with a width of 11 frames (5 on each side). Working with a frame shift of 10 ms, this roughly corresponds to the phone scale. This operation allows to remove possible spurious values. Secondly, contextual information is added by including the first and second derivatives, computed using the following finite difference equation: $x'_t = \frac{1}{N}\sum_{i=1}^{N} x_t - x_{t-i}$. To keep working at the phone level, the number N of contextual frames is set to 10. When in test, the ANN outputs the posterior of speech activity. As a last post-process, the posterior trajectories are smoothed out by a median filter whose width is again set to 11 frames so as to remove possible erroneous isolated decisions.

Our goal being to combine various sets of features, we consider two strategies to merge their information: feature fusion and decision fusion (also called early and late fusion). In the

TABLE I
CHARACTERISTICS OF THE TESTING DATABASE

Environ.	Kitchen	Mall	Station	Living	Street	Overall
Dur. (min)	49	19.8	20	42.3	20.5	151.6
Av.SNR (dB)	7.3	6.9	13.5	18.2	15.1	12.2
% of speech	12.3	20.2	20.5	22.6	18.9	18.9

feature fusion case, synchronous feature vectors are simply concatenated and a single ANN is trained. In the decision fusion case, one specific ANN is trained for each feature set. Each ANN outputs a trajectory of posteriors, and the trajectories from the various ANNs are further merged to derive one final posterior value. Several strategies to combine the posteriors have been proposed in [39]. In this work, we have tried the arithmetic and the geometrical mean (corresponding to the sum and product rule in [39]). The differences in performance that we noticed were however negligible, and the geometrical mean is used throughout the rest of this letter.

III. EXPERIMENTAL PROTOCOL

A. Speech Databases

For the training of the proposed technique, our goal was to use a corpus containing a large diversity of speakers and noisy conditions. We chose a subset of 1500 files from the TIMIT database [40] from 300 speakers. As the original utterances were recorded in clean studio conditions, the advantage of this approach is that the labels can be easily obtained by using a simple energy threshold to extract the speech endpoints. For each file, noise was then artificially added at two SNR levels: 0 and 10 dB, leading to a total of 3000 files. For each file, the noise was randomly selected among 4 types from the Noisex-92 database: babble, car, factory and jet noises. Note that we added 2 seconds of noise before and after each utterance so that the database is roughly balanced between speech activity and background noise. We expect that this multi-condition training set is sufficiently diversified for the classifier to be effective in various (possibly unseen) environments and with new speakers. The development set consists of a 5% held-out portion of the training set.

The testing corpus is a manually annotated proprietary database containing real data recorded in 5 places: mall, kitchen, street, station and living room. Various sources of noise are therefore covered and encompass TV in the background, people talking nearby, cooking, cars passing by, etc. The data consists of Japanese read speech from 5 speakers using either a tablet or a smartphone. The main characteristics of the testing database are summarized in Table I. Note that the averaged SNR only reflects one aspect of the noise, and that other characteristics such as its dynamics and its spectral shape might be a preponderant source of performance degradation.

B. Assessment Metrics

As a first metric to quantify the discrimination power of each feature individually, we use the normalized mutual information [41], defined as the mutual information (MI) of the feature with the class labels divided by the class entropy. The normalization ensures an intuitive interpretation with values ranging between 0 and 1. This measure has also the advantage to be independent from the subsequent classifier. The computation of mutual information is here carried out via a histogram approach [41]. The number of bins is set to 50 for each feature dimension, which results in a trade-off between an adequately high number for an accurate estimation, while keeping sufficient samples per bin. Class labels correspond to the presence (or not) of speech.

To assess the perfomance after classification, two metrics are used. These two measures respectively characterize the frame and the utterance levels. By varying a decision threshold θ, a Receiving Operating Characteristics (ROC) curve can be obtained. The first metric is the so-called *equal error rate* (EER), which corresponds to the location on a ROC curve where the false accept rate and false reject rate are equal. The second metric quantifies the ability to detect the endpoints of speech utterances. For this purpose, we use the F1 score (maximized over θ in the dev set) as a single measure combining both precision and recall. The F1 score ranges from 0 to 1, where 1 implies a perfect classification. The correctness of a speech segment with regard to a reference is conform to the CENSREC-1-C criteria defined in [42]. Note that before being assessed at the utterance level, the vector of binary decisions goes through an hangover scheme [43] consisting of a morphological closing (i.e. a dilatation followed by an erosion) with a time constant of 600 ms and a length extension of 200 ms on each side. Note that the same hangover scheme was applied to all techniques for the computation of the utterance-level results.

C. Comparison with State-of-the-Art Techniques

Four state-of-the-art VAD systems are used for comparison purpose: the G.729B algorithm [1], Shon's statistical model-based VAD [14], Ying's unsupervised technique based on sequential Gaussian mixture models [23] (whose code was kindly shared by Dongwen Ying), and Ghosh's VAD using long-term signal variability [16]. As for the proposed technique, each of these methods makes use of a decision parameter which was tuned to optimize the EER and F1 scores, as discussed in Section III-B.

IV. RESULTS

A. Mutual Information-Based Assessment

The results of the MI-based assessment are presented in Table II. For the filter-based features (MFCC, CGD and PLP), MI values have been averaged across the 13 coefficients. Note also that, for each feature, these results are averaged across static, first and second derivatives values. It can be seen that CGD gives the best results among the spectral envelope representations. Among the source-related features, the three proposed features interestingly provide the best results. They are followed by 3 features used in [8]: HPS, harmonicity and clarity. This latter feature achieves a MI value comparable to that of the LP kurtosis and of the log-energy. As mentioned in Section II-B, designing robust source-related features is a challenging problem. The fact that the 3 proposed features yield better performance can be explained as follows: *i)* time-domain features are expected to be more sensitive to noise and working either in the spectral or cepstral domain turns out to be more appropriate, *ii)* SRH features outperform HPS because they exploit interharmonics as well as the LP residue which allows

TABLE II
MUTUAL INFORMATION-BASED FEATURE ASSESSMENT

Feature	MFCC	CGD	PLP	Energy	ZCR	Kurt.	Skew.
MI (%)	14.3	17.2	13.6	27.3	22.6	27.3	19.0
Feature	Harm.	Clar.	LP err.	HPS	CPP	SRH	SRH*
MI (%)	29.7	27.2	17.1	32.9	36.3	38.7	51.8

TABLE III
CLASSIFICATION RESULTS USING THE 5 FEATURE SETS

Feature Set	MFCC	CGD	PLP	Sadjadi	New
1-EER (in %)	87.9	90.2	87.6	93.7	94.0
F1 score (in %)	77.1	79.2	75.5	86.8	86.7

TABLE IV
CLASSIFICATION RESULTS (1-EER, IN %) USING A COMBINATION OF FEATURE SETS

Combination	MFCC+S	MFCC+N	S+N	MFCC+S+N
Feature fusion	93.6	89.8	94.9	90.9
Decision fusion	94.8	95.4	95.3	95.8

to minimize the effects of both the vocal tract resonances and of the noise [37].

B. Classification Results

For these experiments, we consider various sets of features: 13 MFCCs, 13 CGDs, 13 PLPs, the 4 voicing features (Harmonicity, Clarity, LP error and HPS) used in Sadjadi's paper [8], and the 3 new source-based features (CPP, SRH and SRH*) which have not been used for VAD purpose yet. The two last sets of features will be referred to as *Sadjadi* and *New* in the following. The performance of these 5 feature sets is shown in Table III. Two main conclusions, which corroborate our observations from Section IV-A, can be drawn from these results: *i)* for VAD purpose, source-related features are more relevant than those characterizing the filter. Among them, the *Sadjadi* and *New* feature sets achieve similar performance; *ii)* across the filter representations, the CGD features, whose robustness was already highlighted in [30] for ASR purpose, turn out to be the most efficient. Nonetheless, since MFCCs are widely used in various speech technology applications and that their extraction is likely to be required anyways, we chose to use them as filter-based features in the rest of this letter.

In the second part of our experiments, we investigated the combination of different feature sets either at the feature or the decision level (see Section II-C). The results are displayed in Table IV, where N and S respectively stand for the *New* and *Sadjadi* feature sets. Note that Table IV only shows the EER-based results; similar conclusions could be however drawn from the F1 scores. Interestingly, it can be observed that in all cases the decision fusion scheme outperforms feature fusion, by 3% in absolute on average. Feature fusion even led to a degradation in 3 out of the 4 cases. This is important because feature concatenation is conventionally used in most existing approaches. One possible reason to explain this is the curse of dimensionality [44]: as the dimensionality of the feature vector increases, it becomes more and more difficult to accurately model the data, as an ever increasing number of samples is required. Although the association of the two excitation-based feature sets (S+N) yields already a high performance, the best results are obtained when they are combined with MFCCs. This is however only true when using the decision fusion. In the rest of our experiments, the system based on these 3 feature sets and using decision fusion will be referred to as the proposed VAD system.

TABLE V
COMPARISON WITH STATE-OF-THE-ART METHODS (F1 SCORES, IN %)

	G.729B [1]	Sohn [14]	Ying [23]	Ghosh [16]	Prop.
Kitchen	37.8	43.9	51.2	44.6	**89.2**
Mall	33.2	69.6	70.9	67.0	**89.4**
Station	67.6	85.5	84.2	93.6	**95.6**
Living	28.0	46.9	47.8	45.1	**93.3**
Street	72.1	84.5	81.8	94.9	**97.7**
Average	47.7	66.1	67.2	69.1	**93.0**

The comparative evaluation with state-of-the-art techniques is summarized in Table V for the 5 different environments and using the F1 score. Note that all the observations that will be made hereafter were also corroborated using the EER metric. Three main conclusions can be drawn from Table V. First, it can be noticed that across all conditions the proposed system clearly outperforms existing methods, sometimes by a large increase of the F1 score. This is especially true in the kitchen, living room and mall environments, where existing algorithms tend to fail dramatically. This is mostly due to the fact that the corresponding recordings contain sporadic impulsive noises such as cough, laughter or cooking, whose dynamics can sometimes be similar to that of speech. These environments are therefore much more challenging than the street and station conditions which are rather stationary. Secondly, it is worth reminding that the four state-of-the-art techniques used in this comparison are based on the power spectral density, and therefore discard any source-related information. This further supports our results from Tables II and III that excitation-based features are necessary in an efficient VAD system. Finally, despite the mismatch between training and testing data, the proposed algorithm works well in all environments. This makes us think that the generalization capabilities of the proposed system are high, and that it can potentially adapt to any new environment, speaker, language or sensor. This is likely due to the robustness of the source-related features as well as the ability of the ANN to capture the speech patterns through the multi-condition training.

V. CONCLUSION

The goal of this letter was to investigate the joint use of source and filter-based features for VAD purpose. The main conclusions of this study are the following: *i)* source-related features, and especially the 3 proposed features, have a better discrimination power and their use in an efficient VAD system is necessary, *ii)* as a strategy to merge different sources of information, decision fusion outperforms feature fusion, *iii)* the resulting proposed system, combining source and filter-based information, gives a significantly better performance compared to state-of-the-art methods, *iv)* the robustness of source-related features combined with the generalization capabilities of neural networks makes the proposed approach perform very well in unseen conditions. Features used in this letter can be extracted with the following toolkit: tcts.fpms.ac.be/~drugman/files/VAD.zip.

REFERENCES

[1] A. Benyassine, E. Shlomot, H. Su, D. Massaloux, C. Lamblin, and J. Petit, "ITU-T recommendations G.729 Annex B: A silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications," *IEEE Commun. Mag.*, vol. 35, pp. 64–73, 1997.

[2] D. Valj, B. Kotnik, B. Horvat, and Z. Kacic, "A computationally efficient mel-filter bank VAD algorithm for distributed speech recognition systems," *EURASIP J. Appl. Signal Process.*, no. 4, pp. 487–497, 2005.

[3] I. McCowan, D. Dean, M. McLaren, R. Vogt, and S. Sridharan, "The delta-phase spectrum with application to voice activity detection and speaker recognition," *IEEE. Trans. Audio Speech Lang. Process.*, vol. 19, pp. 2026–2038, 2011.

[4] F. Lamel, R. Rabiner, E. Rosenberg, and G. Wilpon, "An improved endpoint detector for isolated word recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 29, pp. 777–785, 1981.

[5] B. Kotnik, Z. Kacic, and B. Horvat, "A multiconditional robust front-end feature extraction with a noise reduction procedure based on improved spectral subtraction algorithm," in *Proc. 7th Eurospeech*, 2001, pp. 197–200.

[6] B. Kingsbury, G. Saon, L. Mangu, M. Padmanabhan, and R. Sarikaya, "Robust speech recognition in noisy environments: The 2001 IBM SPINE evaluation system," in *Proc. ICASSP*, 2002, pp. 53–56.

[7] T. Kristjansson, S. Deligne, and P. Olsen, "Voicing features for robust speech detection," in *Proc. Interspeech*, 2005, pp. 369–372.

[8] S. O. Sadjadi and J. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *IEEE Signal Process. Lett.*, vol. 20, pp. 197–200, 2013.

[9] M. Marzinzik and B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," *IEEE Trans. Speech Audio Process.*, vol. 10, pp. 109–118, 2002.

[10] J. Haigh and J. Mason, "A voice activity detector based on cepstral analysis," in *Proc. Eurospeech*, 2003, pp. 1103–1106.

[11] E. Nemer, R. Goubran, and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the LPC residual domain," *IEEE Trans. Speech Audio Process.*, vol. 9, pp. 217–231, 2001.

[12] R. Tucker, "Voice activity detection using a periodicity measure," *Proc. Inst. Elect. Eng.*, vol. 139, pp. 377–380, 1992.

[13] K. Ishizuka and T. Nakatani, "Study of noise robust voice activity detection based on periodic component to aperiodic component ratio," in *Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition*, 2006, pp. 65–70.

[14] J. Sohn, N. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, pp. 1–3, 1999.

[15] J. Ramirez, J. Segura, M. Benitez, L. Garcia, and A. Rubio, "Statistical voice activity detection using a multiple observation likelihood ratio test," *IEEE Signal Process. Lett.*, vol. 12, pp. 689–692, 2005.

[16] P. Ghosh, A. Tsiartas, and S. Narayanan, "Robust voice activity detection using long-term signal variability," *IEEE Trans. Audio Speech Lang. Process*, vol. 19, pp. 600–613, 2011.

[17] J. Ramirez, J. Segura, M. Benitez, A. de la Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Commun.*, vol. 42, pp. 271–287, 2004.

[18] Y. Kida and T. Kawahara, "Voice activity detection based on optimally weighted combination of multiple features," in *Proc. Interspeech*, 2005, pp. 2621–2624.

[19] S. Soleimani and S. Ahadi, "Voice activity detection based on combination of multiple features using linear/kernel discriminant analyses," in *Proc. Information and Communication Technologies: From Theory to Applications*, 2008, pp. 1–5.

[20] T. Ng, B. Zhang, L. Nguyen, S. Matsoukas, X. Zhou, N. Mesgarani, K. Vesely, and P. Matejka, "Developing a speech activity detection system for the DARPA RATS program," in *Proc. Interspeech*, 2012.

[21] R. Sarikaya and J. Hansen, "Robust detection of speech activity in the presence of noise," in *Proc. ICSLP*, 1998, pp. 1455–1458.

[22] F. Germain, D. Sun, and G. Mysore, "Speaker and noise independent voice activity detection," in *Proc. Interspeech*, 2013.

[23] D. Ying, Y. Yan, J. Dang, and F. Soong, "Voice activity detection based on an unsupervised learning framework," *IEEE Trans. Audio Speech and Lang. Process*, vol. 19, pp. 2624–2633, 2011.

[24] A. Misra, "Speech/nonspeech segmentation in web videos," in *Proc. Interspeech*, 2012.

[25] F. Eyben, F. Weninger, S. Squartini, and B. Schuller, "Real-life voice activity detection with LSTM Recurrent Neural Networks and an application to Hollywood movies," in *Proc. ICASSP*, 2013, pp. 483–487.

[26] T. Drugman, P. Alku, B. Yegnanarayana, and A. Alwan, "Glottal source processing: From analysis to applications," *Comput. Speech Lang.*, vol. 28, no. 5, pp. 1117–1138, 2014.

[27] T. Drugman and Y. Stylianou, "Fast inter-harmonic reconstruction for spectral envelope estimation in high-pitched voices," *IEEE Signal Process. Lett.*, 2014.

[28] F. Zheng, G. Zhang, and Z. Song, "Comparison of different implementations of MFCC," *J. Comput. Sci. Technol.*, vol. 16, no. 6, pp. 582–589, 2001.

[29] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, pp. 1738–1752, 1990.

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

[31] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Commun.*, vol. 52, pp. 12–40, 2010.

[32] T. Drugman, "Advances in Glottal Analysis and its Applications," Ph.D thesis, Univ. Mons, Mons, Belgium, 2011.

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

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

[35] T. Drugman, "Maximum phase modeling for sparse linear prediction of speech," *IEEE Signal Process. Lett.*, vol. 21, no. 2, pp. 185–189, 2014.

[36] J. Hillenbrand and R. Houde, "Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech," *J. Speech Hearing Res.*, vol. 39, pp. 311–321, 1996.

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

[38] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP - A collaborative voice analysis repository for speech technologies," in *Proc. ICASSP*, 2014, pp. 960–964.

[39] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 20, pp. 226–239, 1998.

[40] DARPA-TIMIT, "Acoustic-phonetic continuous speech corpus," 1990, NIST Speech Disc 1-1.1.

[41] T. Drugman, M. Gurban, and J. Thiran, "Relevant feature selection for audio-visual speech recognition," *IEEE Multimedia Signal Process.*, pp. 179–182, 2007.

[42] N. Kitaoka, K. Yamamoto, T. Kusamizu, and S. Nakagawa *et al.*, "Development of VAD evaluation framework CENSREC-1-C and investigation of relationship between VAD and speech recognition performance," *IEEE Automat. Speech Recognit. Understand.*, pp. 607–612, 2007.

[43] D. Vlaj, M. Kos, M. Grasic, and Z. Kacic, "Influence of hangover and hangbefore criteria on automatic speech recognition," in *Proc. Int. Conf. Systems, Signals and Image Processing (IWSSIP)*, 2009, pp. 1–4.

[44] R. Bellman, *Adaptive Control Processes: A Guided Tour*. Princeton, NJ, USA: Princeton Univ. Press, 1961.