INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Improving i-Vector and PLDA based Speaker Clustering with Long-term Features

Abraham Woubie[1], *Jordi Luque*[2], *and Javier Hernando*[1]

[1] TALP Research Center, Department of Signal Theory and Communications, Universitat Politecnica de Catalunya, Barcelona, Spain
[2] Telefonica Research, Edificio Telefonica-Diagonal, Barcelona, Spain
abraham.woubie.zewoudie@upc.edu, jls@tid.es, javier.hernando@upc.edu

Abstract

i-vector modeling techniques have been successfully used for speaker clustering task recently. In this work, we propose the extraction of i-vectors from short- and long-term speech features, and the fusion of their PLDA scores within the frame of speaker diarization. Two sets of i-vectors are first extracted from short-term spectral and long-term voice-quality, prosodic and glottal to noise excitation ratio (GNE) features. Then, the PLDA scores of these two i-vectors are fused for speaker clustering task. Experiments have been carried out on single and multiple site scenario test sets of Augmented Multi-party Interaction (AMI) corpus. Experimental results show that i-vector based PLDA speaker clustering technique provides a significant diarization error rate (DER) improvement than GMM based BIC clustering technique.

Index Terms: speaker clustering, i-vector, voice-quality, prosody, GNE, fusion, cosine distance, PLDA

1. Introduction

Speaker diarization is the process of partitioning an input audio stream into homogeneous segments according to the speaker identity. It needs to first classify the speech and non-speech parts of an audio signal. Then, it marks the speaker changes in the detected speech and clusters speech segments which belong to the same speaker [1].

Feature extraction plays a significant role on the performance of speaker diarization systems. It needs to extract features that have large between-speaker variability and small within-speaker variability. Mel Frequency Cepstral Coefficients (MFCCs) are the most widely used short-term acoustic features for speaker diarization [2]. Prosodic features provide useful information for automatic speaker recognition as reported in [3]. Our previous work in [4] has shown that the fusion of jitter and shimmer voice-quality features with the prosodic and spectral ones improves the performance of speaker diarization systems. Fusion techniques also increase the reliability and robustness of a system as reported in [5].

Another factor that affects the performance of speaker diarization system is the speaker clustering technique. Gaussian Mixture Modeling (GMM) based Bayesian Information Criterion (BIC) metric is the most widely used speaker clustering technique for speaker diarization systems [2]. The state-of-the-art i-vector modeling techniques in speaker recognition have recently been successfully used in language identification [6] and speaker diarization systems [7]. It is reported in [8, 9, 10, 11] that i-vector based cosine distance speaker clustering technique provides better DER results than GMM based BIC clustering one. The use of i-vector based PLDA clustering technique also provides better DER result than GMM based BIC and i-vector based cosine distance clustering techniques as reported in the works of [12, 13].

The above mentioned works extract the i-vectors from the short-term spectral features for speaker clustering task. Therefore, we have proposed the use of Glottal to noise excitation ratio (GNE) feature and i-vector based PLDA clustering technique. The main contribution to our recent work in [14] is the use of GNE feature together with the voice-quality and prosodic features. The i-vector based cosine distance clustering technique in [14] is also replaced by i-vector based PLDA clustering one. Two sets of i-vectors are extracted first from the short-term spectral and long-term speech features. The long-term speech features are the concatenation of voice-quality, prosodic and GNE features. The PLDA scores of these two i-vectors are then fused linearly for speaker clustering task.

The experiments have been carried out on selected shows of Augmented Multi-party Interaction (AMI) corpus, a multimodal dataset of meeting recordings [15].

The rest of this paper is organized as follows. The next two sections give an overview of long-term speech features used in our experiment and the speaker diarization system architecture. The fusion techniques are outlined in Section 4. Section 5 and 6 discuss about experimental results and conclusions, respectively.

2. Long-term Speech Features

Although short-term spectral features are the most widely used ones for speaker diarization, our previous work in [4] shows that the fusion of these short-term spectral features with jitter and shimmer voice-quality features improve the performance of speaker diarization systems. The authors in [16, 17] also show that prosodic features have been used with short-term spectral ones to improve the performance of speaker diarization systems.

We have therefore extracted absolute jitter, absolute shimmer, shimmer apq3 and prosodic features (fundamental frequency, acoustic intensity and formant frequencies) based on previous studies of [17] [18]. Each of the voice-quality and prosodic features are extracted by averaging them over a window length of 500ms with 10ms shift. Furthermore, one of the contributions of this paper is the use of GNE feature together with the feature level stacked prosodic and voice-quality ones.

GNE is an acoustic measure that can be used to assess the amount of voice excitation by vocal-fold oscillations versus excitation by turbulent noise. It indicates whether a given voice signal originates from vibrations of the vocal folds or from tur-

http://dx.doi.org/10.21437/Interspeech.2016-339

bulent noise generated in the vocal tract [19]. The main advantage of GNE is its computation is independent of variations of fundamental frequency and amplitude [20, 21]. GNE is closely related to breathiness, and is considered as a reliable measure for the relative noise level even in the presence of strong amplitude and frequency perturbations. It is shown in the work of [21] that GNE parameter has a significant potential to screen voices since it quantifies the amount of voice excitation and turbulent noise. It is also reported in [22] that GNE provides reliable measurements in terms of discrimination among normal and pathological voices more than other classical long-term noise measurements, such as Normalized Noise Energy and Harmonics to Noise Ratio. It has also been used successfully to screen voice disorders in [22].

The GNE is then stacked together with the three voice-quality features and features related to the evolution in time of pitch, acoustic intensity and the first four formant frequencies. This generates a ten dimensional feature vector. From now, for the sake of clarity, we shall refer the stacked GNE, voice-quality and prosodic features as long-term features.

3. Speaker Diarization Architecture

Feature extraction is first carried out for the short-and long-term speech features only for the speech frames as shown in Figure 1. The voice-quality, prosodic and GNE features are then stacked together in the same feature vector. The initial number of clusters depends on the duration of audio signals, but it is limited to the range $(10, 65)$. This is done to reduce the common issues of agglomerative hierarchical clustering (AHC) such as over-clustering and the high computational cost in pair-wise distance computation.

GMM modeling technique have been used to model the acoustic features for speaker segmentation. Each state of the Hiddem Markov Model (HMM) is composed of a mixture of Gaussians, fitting the probability distribution of the features by the classical expectation-maximization (EM) algorithm. Independent HMM models are used for the short- and long-term speech features. The number of Gaussian mixtures is automatically computed based on of the size of speech segments for the short-term spectral features. But, it is manually set for the long-term features. Finally, a time constraint is imposed on the HMM topology as in [23]. The minimum duration of the speaker turn is set to be greater than 3 seconds.

Note that the main contribution of this paper is the use of i-vector based PLDA clustering technique based on short- and long-term speech features. (See Figure 1).

Factor analysis techniques have been used to extract the i-vectors from the outputs of Viterbi segmentation as proposed by [24]. Two sets of i-vectors are extracted from the short- and long-term speech features as shown in Figure 1. The i-vectors are first extracted from the outputs of Viterbi segmentation at each iteration.

Once the i-vectors are extracted from the short- and long-term speech features, PLDA models the i-vectors as follows:

$$w_{ij} = \mu + Fh_i + \Sigma_{ij} \quad (1)$$

where w_{ij} represent the *j*'th segment of i-vector *i*, μ is the overall speaker and segment independent mean of the i-vectors in the training dataset, and the columns of the matrix F define the between-speaker variability. Any unexplained data variation is represented by Σ_{ij}. The components of the vector h_i are the eigen-voice factor loadings. The term Fh_i depends only on the identity of the speaker, not on the particular segment.

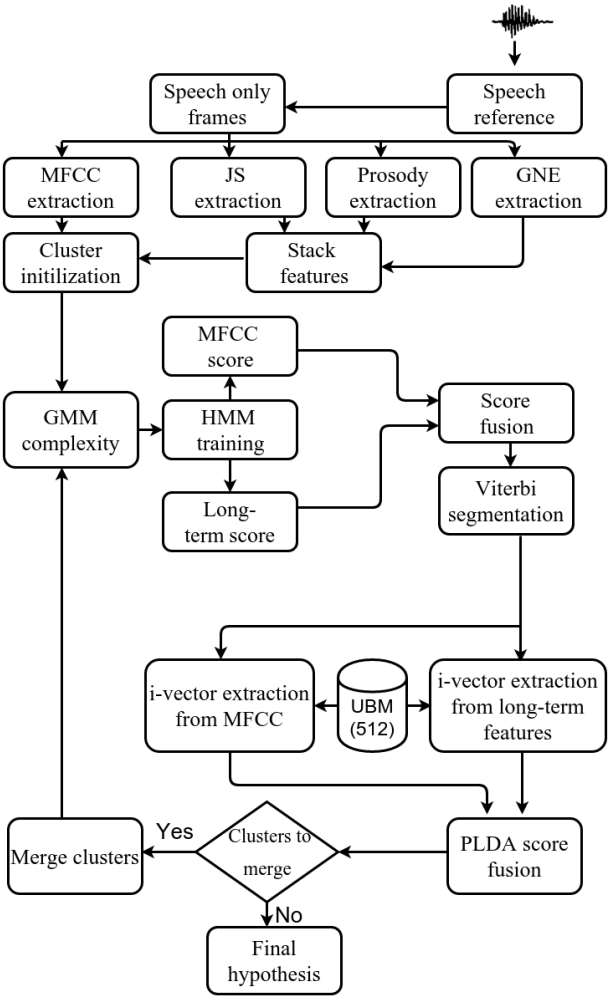

Figure 1: *Speaker diarization architecture based on HMM modeling for speaker segmentation and i-vector based PLDA speaker clustering based on short- and long-term speech features.*

The parameters $\{\mu,\ F$ and $\Sigma\}$ should then be estimated from a set of training data assuming that the speech samples of an individual consist of different number of sessions. The recognition phase checks whether two i-vectors belong to the same speaker or different speakers.The parameter estimation is done using expectation maximization(EM) algorithm.

Once the similarity measure between i-vectors is computed, the two sets of cluster with the highest PLDA score are merged at each iteration. A new i-vector is extracted at each iteration from the outputs of the new segmentation. Note that i-vectors are used only for speaker clustering task. The speaker segmentation is carried out using HMM based Viterbi segmentation as in [4, 14].

A manual threshold value λ is used as a stopping criterion on the matrix of distances of clusters. When the PLDA score among all pair of clusters is less than λ, the merging process stops. Finally, the speaker diarization system outputs the final speaker segmentation hypothesis.

4. Fusion Techniques

Short-term spectral and long-term speech features have been used in our work. The long-term features are the concatenation of voice-quality, prosodic and GNE features. The concatenation of the long-term features can be considered as fusion at the feature level.

The fusion of the short-term and long-term speech features is carried out differently in speaker segmentation and speaker clustering. While the fusion of speaker segmentation is based on log-likelihood ratios, the fusion of speaker clustering is based on PLDA scores extracted from i-vectors.

Given a set of input features vectors, $\{\mathbf{x}\}$ and $\{\mathbf{y}\}$, the joint log-likelihood score for speaker segmentation is carried out as follows:

$$\alpha \log P(\mathbf{x}|\theta_{ix}) + (1-\alpha) \log P(\mathbf{y}|\theta_{iy}), \tag{2}$$

where θ_{ix} is the model of cluster i from spectral features, and θ_{iy} is the model for the same cluster i from long-term features. The GMM scores of short- and long-term speech features are weighted by α and $(1-\alpha)$, respectively.

Once the i-vectors are extracted from the outputs of Viterbi segmentation, the fused PLDA score is computed as follows:

$$\begin{aligned} \gamma . \log \frac{p(w_i, w_j|H_1)}{p(w_i|H_0)p(w_j|H_0)} + \\ (1-\gamma) . \log \frac{p(w'_i, w'_j|H_1)}{p(w'_i|H_0)p(w'_j|H_0)}, \end{aligned} \tag{3}$$

where w_i and w_j represent the i-vectors extracted from the short-term spectral feature for cluster *i* and cluster *j*, respectively. The i-vectors extracted from long-term speech features for cluster *i* and cluster *j* are represented by w'_i and w'_j, respectively. Hypothesis H_1 and H_0 assume that the two i-vectors belong to the same and different speakers, respectively. The PLDA scores of i-vectors extracted from the short-term spectral features are weighted by γ.

5. Experiments and Results

5.1. Database

The experiments have been developed and tested on AMI corpus, a multi-party and spontaneous speech set of recordings [15]. The AMI shows were recorded in English using three different setup rooms with different acoustic properties. The recordings were performed at Idiap, Edinburgh, and TNO sites.

We have selected 10 shows from AMI corpus as a development set to tune the optimum parameters. In order to generalize the results of the development dataset, we have defined two experimental scenarios. The first one is a single site scenario which comprises 10 shows from only Idiap site. The second one is a multiple-site scenario that consists of 10 shows from Idiap, Edinburgh and TNO sites. The total duration of the single-site and multiple-site scenario dataset are 307.36 and 294.01 minutes, respectively. The number of speakers are 4 both in the development and test sets.

5.2. Experimental Setup

The Mel Frequency Cepstral Coefficients (MFCCs) are extracted over 30ms frame length at 10ms frame shift without the deltas. The total number of coefficients extracted for the spectral features are 20. The voice-quality, prosodic and GNE features are extracted over 30ms frame length at 10ms frame rate using Praat software [25]. Then, the voice-quality, prosodic and GNE features are averaged over a window length of 500ms at 10ms. This is done to smooth out their estimation and synchronize them with the short-term spectral features.

The size of i-vectors extracted from the short- and long-term speech features are 100 and 50, respectively. Length normalization is applied on i-vectors before PLDA scoring. The i-vectors are extracted using ALIZE open source software [26].

We have selected 100 shows from AMI corpus with duration of 60 hours to train the universal background model (UBM) and T matrix. Two different UBMs of 512 Gaussians components have been trained for the short- and long-term speech features. The UBM of the short-term spectral features is trained on 20 cepstral coefficients. The feature level stacked voice-quality, prosodic and GNE features are used to train the UBM of long-term features. The size of the total variability matrix is 100 and 50 for the short- and long-term speech features, respectively.

The PLDA system of the short-term and long-term speech features use a 40 and 20 dimensional speaker space. The PLDA is trained on the same data used to train the UBM and T-matrix but the audio signals are chopped into pieces of 3 second segments.

Our speaker diarization system results are evaluated using Diarization Error Rate (DER) metric. In its purest form, DER represents the sum of false alarm speech, missed speech and speaker errors. Since we have used the speech references (Oracle SAD), DER results reported in our work have only speaker errors.

5.3. Experimental Results

As it is shown in Table 1, the baseline system of the single site scenario that uses GMM based BIC clustering technique and MFCC feature set has a DER of 15.87%. The use of same modeling technique and fusion of jitter and shimmer voice-quality (JS), prosodic and GNE features reduces the DER to 14.48%. This corresponds to a 8.76% relative DER improvement more than the baseline system. Replacing the GMM based BIC clustering technique that uses MFCC feature sets with i-vector based PLDA clustering techniques on the same feature set reduces the DER to 13.64%. This represents a 14.05% relative DER improvement more than the baseline system.

The table also shows that that applying i-vector based PLDA clustering techniques on MFCC, voice-quality, prosodic and GNE feature sets provides a DER of 12.93%. This represents a 10.7% relative DER improvement over the GMM based system. It also represents a 5.21% relative DER improvement more than the system that applies i-vector based PLDA clustering technique and uses only MFCC feature set.

The table also shows that baseline system of the multiple site test scenario which is based on GMM modeling and MFCC feature set has a DER of 24.66%. The use of same modeling technique and MFCC, voice-quality and prosodic feature sets reduces the DER to 22.96%. This amounts to 6.89% relative DER improvement more than the baseline system. Similarly, the use of same modeling technique and fusion of voice-quality, prosodic and GNE features yields a DER of 20.83%. This corresponds to a 15.53% relative DER improvement than the baseline system. Applying i-vector based PLDA clustering technique on MFCC feature set shows a DER of 20.11%. This represents a 18.45% relative DER improvement more than the baseline system.

Finally, the table shows that applying i-vector based PLDA clustering technique based on MFCC, voice-quality, prosodic

		Single site	Multiple site
Features	**Clustering**	**DER(%)**	**DER(%)**
MFCC	GMM/BIC	15.87	24.66
MFCC + JS + Prosody	GMM/BIC	15.02	22.96
MFCC + JS + Prosody + GNE	GMM/BIC	14.48	20.83
MFCC	i-vector/PLDA	13.64	20.11
MFCC + JS + Prosody	i-vector/PLDA	13.07	19.23
MFCC + JS + Prosody + GNE	i-vector/PLDA	**12.93**	**18.78**

Table 1: *DER of single- and multiple-site scenarios for GMM based BIC and i-vector based PLDA clustering techniques using different feature sets. JS denote absolute jitter, absolute shimmer and shimmer apq3. JS + Prosody represent the feature level stacked JS and Prosodic features. JS + Prosody + GNE represent the the feature level stacked JS, prosodic and GNE features.*

and GNE features provides a DER of 18.78%. This corresponds to a 6.6% relative DER improvement more than the system that applies same clustering technique and uses only MFCC feature set. It also represents a 9.84% relative DER improvement more than the system that uses same feature sets and applies GMM modeling technique.

The result of the the test sets show that the use of GNE feature together with the voice-quality and prosodic features provides better DER improvement more than the system that uses only voice-quality and prosodic features. The improvements are both for GMM and i-vector modeling techniques as shown in the table.

The optimum set of parameters have been tuned on 10 shows of AMI development dataset. For GMM based BIC modeling technique, the optimum α weight value of the short-term spectral features is 0.975 . The optimum set of γ weight value for short-term spectral features is 0.98 for i-vector based PLDA clustering technique.

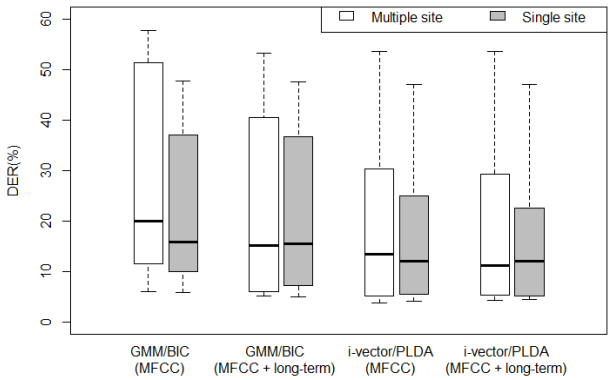

Figure 2: *Box plots of single- and multiple-site scenario test sets for GMM based BIC and i-vector based PLDA clustering techniques using short- and long-term (JS + Prosody + GNE) feature sets.*

The box plots in Figure 2 show the DER distribution of different audio recordings of the single- and multiple-site scenario test sets for different feature and clustering techniques. It shows the minimum, lower quartile, median, upper quartile, and maximum DER of the shows. As shown in the figure, the range of DER values decreases and becomes lower for both scenarios when short- and long-term features are used together. The median DER and the range of DER variations among the different shows becomes lowest for both scenarios when i-vector based clustering technique is used with the short- and long-term features.

Although the use of i-vector based PLDA clustering technique reduces the DER of most of the shows in the test sets, the DER values increase for some of the shows compared to the GMM based BIC clustering technique. Reasons for these should be studied in the future.

6. Conclusions

We have proposed the use of GNE feature and i-vector based PLDA clustering technique within the framework of speaker diarization. The clustering technique is based on the fusion of PLDA scores of i-vectors extracted from short- and long-term speech features.

Our experimental results show that i-vector based PLDA clustering technique provides a substantial relative DER improvement more than GMM based BIC clustering one. Experimental results also show that the extraction of i-vectors from the short- and long-term speech features provides better DER result than extracting i-vectors only from the short-term spectral features. Finally, the results show that the use of GNE features together with the voice-quality and prosodic ones provides better DER result more than the system that uses only the latter features both for GMM and i-vector based speaker clustering techniques.

The results of our work show the usefulness of replacing GMM based BIC clustering technique with the i-vector based PLDA clustering one. The experimental results also show the usefulness of voice-quality, prosodic and GNE features for speaker diarization.

7. Acknowledgements

This work has been supported by the Spanish Ministry of Economy and Competitiveness under contract PCIN-2013-067, the Generalitat de Catalunya under contract 2014-SGR-1660, and by the project TEC2015-69266-P (MINECO/FEDER, UE). This project has also received funding from the EU's Horizon 2020 research and innovation programme under grant agreement No 645323. This text reflects only the authors' view and the Commission is not responsible for any use that may be made of the information it contains.

8. References

[1] S. Meignier, D. Moraru, C. Fredouille, J.-F. Bonastre, and L. Besacier, "Step-by-step and integrated approaches in broadcast news speaker diarization," *Computer Speech & Language*, vol. 20, no. 2, pp. 303–330, 2006.

[2] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker Diarization: A Review of Recent Research," *IEEE Transactions on Audio, Speech and Language Processing*, 2011.

[3] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, "Modeling Prosodic Feature sequences for Speaker Recognition," *Speech Communication*, vol. 46, no. 3, pp. 455–472, 2005.

[4] A. Woubie, J. Luque, and J. Hernando, "Using Voice-quality Measurements with Prosodic and Spectral Features for Speaker Diarization," in *INTERSPEECH*, 2015.

[5] X.-G. Wang and H. C. Shen, "Multiple hypothesis testing fusion method for multisensor systems," in *IEEE International Conference on Intelligent Robots and Systems*, vol. 2, 1999.

[6] D. Martínez, L. Burget, L. Ferrer, and N. Scheffer, "ivector-based Prosodic System for Language Identification," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4861–4864.

[7] J. Villalba, A. Ortega, A. Miguel, and E. Lleida, "Variational Bayesian PLDA for Speaker Diarization in the MGB Challenge," in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, 2015, pp. 667–674.

[8] J. Silovsky and J. Prazak, "Speaker diarization of broadcast streams using two-stage clustering based on i-vectors and cosine distance scoring," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012.

[9] P. Kenny, D. Reynolds, and F. Castaldo, "Diarization of telephone conversations using factor analysis," *in IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 6, pp. 1059–1070, 2010.

[10] J. Franco-Pedroso, I. Lopez-Moreno, D. Toledano, and J. González-Rodríguez, "ATVS-UAM system description for the audio segmentation and speaker diarization Albayzin 2010 evaluation," in *FALA*, 2010.

[11] S. Shum, N. Dehak, E. Chuangsuwanich, D. A. Reynolds, and J. R. Glass, "Exploiting intra-conversation variability for speaker diarization," in *Proceedings of Interspeech, Florence, Italy*, 2011.

[12] J. Prazak and J. Silovsky, "Speaker diarization using PLDA-based speaker clustering," in *6th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS)*, vol. 1, 2011.

[13] G. Sell and D. Garcia-Romero, "Speaker diarization with PLDA i-vector scoring and unsupervised calibration," in *Proceedings of SLT*, 2014.

[14] A. Woubie, J. Luque, and J. Hernando, "Short- and Long-term Speech Features for Hybrid HMM-i-vector based Speaker Diarization System," in *Odyssey 2016-The Speaker and Language Recognition Workshop*, 2016, Paper Accepted.

[15] "The Augumented Multi-party Interaction Project, AMI Meeting Corpus." Website, http://corpus.amiproject.org, 2011.

[16] G. Friedland, O. Vinyals, Y. Huang, and C. Müller, "Prosodic and other Long-Term Features for Speaker Diarization," *IEEE Transactions on Audio, Speech, and Language Processing*, 2009.

[17] M. Zelenák and J. Hernando, "The Detection of Overlapping Speech with Prosodic Features for Speaker Diarization," in *INTERSPEECH*, 2011, pp. 1041–1044.

[18] A. Woubie, J. Luque, and J. Hernando, "Jitter and Shimmer Measurements for Speaker Diarization," in *VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop*, 2014, pp. 21–30.

[19] D. Michaelis, T. Gramss, and H. W. Strube, "Glottal-to-noise excitation ratio–a new measure for describing pathological voices," *Acta Acustica united with Acustica*, vol. 83, no. 4, pp. 700–706, 1997.

[20] D. Michaelis, M. Fröhlich, and H. W. Strube, "Selection and Combination of Acoustic Features for the description of Pathologic Voices," *The Journal of the Acoustical Society of America*, vol. 103, no. 3, pp. 1628–1639, 1998.

[21] N. Sáenz Lechón, V. Osma Ruiz, R. Fraile Muñoz, J. I. Godino Llorente, and P. Gómez Vilda, "Screening voice disorders with the glottal to noise excitation ratio," 2009.

[22] J. I. Godino-Llorente, V. Osma-Ruiz, N. Sáenz-Lechón, P. Gómez-Vilda, M. Blanco-Velasco, and F. Cruz-Roldán, "The effectiveness of the glottal to noise excitation ratio for the screening of voice disorders," *Journal of Voice*, vol. 24, no. 1, pp. 47–56, 2010.

[23] J. Ajmera and C. Wooters, "A Robust Speaker Clustering Algorithm ," in *Proceedings of IEEE Speech Recognition and Understanding Workshop*, St. Thomas, U.S. Virgin Islands, 2003.

[24] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front end factor analysis for speaker verification," *in IEEE Transactions on Audio, Speech and Language Processing*, 2011.

[25] P. Boersma and D. Weenink, "Praat: doing phonetics by computer, version 5.3.69," http://www.praat.org/.

[26] A. Larcher, J. F. Bonastre, B. G. B. Fauve, K. Lee, C. Lévy, H. Li, J. S. D. Mason, and J. Parfait, "ALIZE 3.0 - open source toolkit for state-of-the-art speaker recognition," in *INTERSPEECH*, 2013, pp. 2768–2772.