Using Voice-quality Measurements with Prosodic and Spectral Features for Speaker Diarization

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

10.21437/Interspeech.2015-110

Abstract

Jitter and shimmer voice-quality measurements have been successfully used to detect voice pathologies and classify different speaking styles. In this paper, we investigate the usefulness of jitter and shimmer voice measurements in the framework of the speaker diarization task. The combination of jitter and shimmer voice-quality features with the long-term prosodic and short-term spectral features is explored in a subset of the Augmented Multi-party Interaction (AMI) corpus, a multi-party and spontaneous speech set of recordings. The best results have been obtained by fusing the voice-quality features with the prosodic ones at the feature level, and then fusing them with the spectral features at the score level. Experimental results show more than 20% relative DER improvement compared to the spectral baseline system.

Index Terms: speaker diarization, spectral features, jitter, shimmer, prosody, fusion

1. Introduction

Speaker diarization segments a multi-speaker audio segment into homogeneous parts, and then clusters these segments into groups where each cluster contains the speech of a single speaker [1]. Commonly, a speaker diarization system consists of the following three major modules: feature extraction, speaker segmentation and speaker clustering. *Feature extraction* extracts specific information from the raw audio signal allowing subsequent speaker modeling and classification. *Speaker segmentation* partitions the audio data into acoustically homogeneous segments according to speaker identities. *Speaker clustering* groups the homogeneous segments of the speaker segmentation task and displays a single cluster for each speaker in the audio signal.

One of the factors that critically affect the performance of speaker diarization approaches is the extraction of relevant speaker features. Mel Frequency Cepstral Coefficients (MFCC) are the most widely used short-term speech features in speaker diarization [2]. Despite its broadly employment in speech processing applications, it is shown in the works of [3, 4, 5] that fusing short-term features with long-term ones yield better results since the later features provide discriminative power among different speakers. Long-term conversational features have been used by [6] to improve the acoustic feature based overlap detector in speaker diarization task.

Jitter and shimmer voice-quality measurements discern variations of fundamental frequency and amplitude, respectively. Studies show that these measurements can be used to detect voice pathologies [7], speaking styles and emotions [8], and also identify age and gender [9]. For example, the authors in [10] report that fusing jitter and shimmer voice-quality measurements along with the baseline spectral features improve the performance of speaker recognition systems. It is also shown in [8] the use of jitter and shimmer measurements improves the classification accuracy, with respect to the baseline spectral features, by conveying complementary information which aids to discriminate among different arousal levels. The work of [11] also shows that fusion of voice-quality with prosodic features is able to effectively discriminate different emotions in Chinese speech emotion identification. The importance of voice-quality features in emotion identification is also discussed in [12, 13]. It is also shown in [7] that these voice-quality measurements can be used to characterize voices such as breathy, tense, harsh, whispery, creaky and hoarse.

Based on these studies, we propose the use of jitter and shimmer voice-quality measurements in the task of speaker diarization. The main contribution of this work is the fusion of jitter and shimmer voice-quality features both with the long-term prosodic and short-term spectral features. We have carried out feature selection of both the voice-quality and prosodic features. The fusion of voice-quality with the prosodic and spectral features is implemented both at the feature and score level. To the best of our knowledge, this is the first reported work applying voice-quality features in speaker diarization task. The development and test experiments are conducted on AMI corpus [14], a multi-party and spontaneous speech set of recordings, and assessed in terms of speaker diarization error (DER).

The rest of this paper is organized as follows. The next sections give an overview of voice-quality measurements followed by agglomerative hierarchical clustering. Section 4 discusses about fusion of spectral, voice-quality and prosodic features. Experimental results and conclusions are finally presented in Section 5 and Section 6, respectively.

2. Voice-quality measurements

Although short-term spectral features (MFCC) are the most widely used ones for speaker diarization, it is shown in [4, 5, 6] that the state-of-the-art speaker diarization system can be improved by combining spectral features with prosodic and other long-term features. Prosody is estimated capturing the evolution in time of fundamental frequency, acoustic intensity and formant frequencies. The appropriate characteristics related to the human speech prosody are conveyed through intonation, rhythm and stress. Encouraged by work of [5], we have extracted features related to the evolution in time of pitch, acoustic intensity and the first four formant frequencies to validate

their performance in this work. This leads to a six dimensional feature vector.

Voice-quality are another crucial long-term features. Jitter and shimmer voice-quality measurements measure variations of fundamental frequency and amplitude of speakers voice, respectively. They are normally used to measure long sustained vowels where measured values above a certain threshold are considered as pathological voices [7]. Since pathological voices are unique to a particular speaker, they can be considered to differentiate speakers. Studies show that voice-quality measurements have been successfully used in speaker recognition [10], stress and emotion classification tasks [8] and emotion recognition [11] . The authors in [7] also report that jitter and shimmer measurements provide significant differences between different speaking styles.

Although there are different types of jitter and shimmer measurements, we extract only two of them called absolute jitter and absolute shimmer encouraged by previous work of [10]. Two different voice-quality features are extracted, *absolute jitter* and *absolute shimmer*, creating a two dimensional feature vector. The voice quality and prosodic features are extracted over 30ms duration with 10ms shift and averaged over 500ms duration.

2.1. Jitter measurement

- *Jitter (absolute):* It measures the periodic deviation or pitch perturbation of voice signal from one pitch period to another.

$$\text{Jitter (absolute)} = \frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i+1}| \quad (1)$$

where T_i are the extracted pitch period lengths and N is the number of extracted pitch periods.

2.2. Shimmer measurement

- *Shimmer (absolute):* It measures amplitude perturbation of voice signal from one pitch period to another.

$$\text{Shimmer (absolute)} = \frac{1}{N-1}\sum_{i=1}^{N-1}\left|20\log\left(\frac{A_{i+1}}{A_i}\right)\right| \quad (2)$$

where A_i are the extracted peak-to-peak amplitude data and N is the number of pitch extracted periods.

3. Agglomerative hierarchical clustering (AHC) of speakers

Our speaker diarization system is based on bottom-up clustering which is one of the the most successful approaches to address the problem of speaker diarization [15, 16]. Input feature vectors are partitioned in a set of segments, i.e., homogeneously splitting the whole feature set or using techniques like k-means. In the first iteration, clusters are initialized through previous segments and a Gaussian model is built on them. Next, a distance among cluster models is computed and then a pairwise comparison is performed aiming to group similar regions. This is a critical step in AHC since it accounts for most of the computation time. After distance matrix analysis and minimum distance are computed, those candidate clusters with minimum distance among them are merged. This process is iterated several times and clusters are merged until some condition is fulfilled, e.g., a threshold on the previous distance matrix. Finally, each remaining cluster is expected to represent an ensemble of the data based on the selected distance measure.

In Figure 1, it is depicted a more detailed scheme of the AHC algorithm adapted to speaker diarization task. The translation mainly consist of keeping main previous steps jointly with the key idea that at the end of the process each cluster C_i should be composed exclusively by speech from the same speaker.

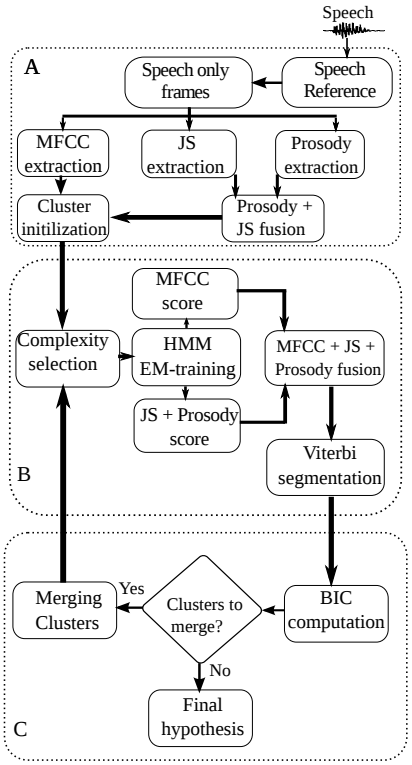

Figure 1: *Speaker diarization scheme based on AHC Hierarchical Clustering with automatic complexity selection.*

3.1. Cluster initialization

First, the speech signal is equally partitioned as shown in (Fig. 1 block A) which generates the initial clusters. The initial number of clusters depends on meeting duration but it is constrained in the range $[35, 65]$ clusters. This methods enables to deal with common issues of AHC such as over-clustering and high computational cost due to combinatorial explosion in pair-wise distance computation. Initial number of clusters is defined as:

$$K_{\text{init}} = \frac{N}{G_{\text{init}}\, R_{\text{CC}}}, \quad (3)$$

where N stands for the number of total frames in a recording and G_{init} is the number of Gaussians initially assigned to each cluster. The complexity ratio, R_{CC}, stands for the minimum amount of speech data in frames needed per each Gaussian mixture in the cluster model. They are fixed to 5 Gaussians and 7 seconds (per Gaussian) respectively. This method of cluster partitioning allows to build a "pure" enough initial cluster segmentation which is a key point in AHC algorithm [17, 18].

3.2. Acoustic modeling

Set of acoustic features in a cluster is independently modeled using HMM/GMM which is iteratively refined, (Fig. 1 block B). Refinement is performed in each clustering iteration through a two step training and decoding process. Each state of the

HMM is composed by a mixture of Gaussians, fitting the probability distribution of the features by the classical expectation-maximization (EM) algorithm. Note that independent HMM models are estimated for each feature stream, as explained in Section 4. The number of mixtures is chosen as a function of available seconds of speech per cluster in the case of MFCC features isolated. They are kept fixed for all the rest of feature streams: shimmer and jitter, prosodic, and feature level fused voice-quality with prosodic features. Finally, a time constraint, as in [19], is also imposed on the HMM topology. It constrains the minimum duration of the speaker turn to be greater than 3 seconds.

3.3. Distance metric

Agglomerative distance among clusters is based on the Bayesian Information Criterion (BIC). This distance measures the difference among each pair of clusters. The stopping criterion is also driven by a threshold on the same matrix of distances, (Fig. 1 block C). A modified BIC-based metric [19] is employed to select the set of cluster-pairs candidates with smallest distances among them. Cluster-pair (i, j) is merged depending on whether its BIC_{ij} fulfills

$$BIC_{ij} > \max\left(\gamma, BIC_{\mu} + \frac{3}{2}BIC_{\sigma}\right), \quad (4)$$

where BIC_{ij} is the BIC estimation between the clusters i and j performed as in [19] and γ is a threshold tuned on development data. The BIC_{μ} is the mean of BIC_{ij} for $i \neq j$ and the BIC_{σ} stands for the standard deviation of the same BIC set. Once clusters are merged, a two-step training and decoding iteration is performed again to refine the model statistics and align them with the speech recording, block B (see Fig. 1).

The automatic selection of the model complexity applied only in MFCC stream has shown a successful performance while avoiding the use of the penalty term in the classical BIC formulation [20, 21]. It is computed at each iteration and cluster by the following equation:

$$M_i^j = \left\lfloor \left(\frac{N_i^j}{R_{\mathrm{CC}}}\right) + \frac{1}{2} \right\rfloor, \quad (5)$$

where N_i^j is the number of frames belonging to the cluster i and j is the number of iterations. A more detailed description of the system can be found in [16, 22]. The model complexity M_i^j stands for the number of mixtures composing the model associated to cluster i at iteration j. It is updated according to the R_{CC} value only for the MFCC stream. In all other streams and feature combination cases, Gaussian complexity is fixed manually and different values are explored.

4. Fusion of spectral, voice-quality and prosodic features

The two dimensional voice-quality and the six dimensional prosodic feature vectors are stacked in the same feature vector generating an eight dimensional vector which can be considered as fusion at the feature level.

Independent HMM models are estimated for each feature stream. The fusion of the short term spectral features with the voice-quality features is carried out at the score level, that is, combining the log likelihood scores corresponding to each stream to create a joint single score . The fusion of short term spectral features with the prosodic ones is also carried out at the score level. When the three streams are used, the voice quality and prosodic features are fused at the feature level, and then they are fused with the spectral ones at the score level.

Given a set of input features vectors, $\{\mathbf{x}\}$ and $\{\mathbf{y}\}$, the log-likelihood score is computed as a joint likelihood between features distributions as follows:

$$\log P(\mathbf{x}, \mathbf{y}) = \alpha \log P(\mathbf{x}|\theta_{ix}) + (1 - \alpha) \log P(\mathbf{y}|\theta_{iy}), \quad (6)$$

where $\log P(\mathbf{x}, \mathbf{y})$ is the fused GMM score for cluster i, θ_{ix} is the model created for cluster i using the spectral feature vectors $\{\mathbf{x}\}$, and θ_{iy} is the model created for the same cluster i using feature vectors $\{\mathbf{y}\}$. The feature vector $\{\mathbf{y}\}$ can be voice-quality, prosodic or feature level fused voice-quality with prosodic features. The weight of the spectral feature vector is α and $(1 - \alpha)$ is the weight of the other features used together with the spectral ones. The later features can be voice-quality, prosodic or feature level fused voice-quality with prosodic ones.

5. Experiments

5.1. Database and experimental setup

Wiener filtering technique has been applied to remove noises of the input signal. Manually annotated speech references have been employed to extract the speech frames and discard non-speech regions both for the development and test sets. A feature vector of MFCC features is computed with 30ms frame length at 10ms frame shift. The extracted MFCC have 20 dimensions and they are without deltas. The two voice-quality and the six prosodic features are extracted over 30ms frame length and 10ms frame shift using Praat software [23]. Then, we calculate the mean of each of the voice-quality and prosodic features over a window length of 500ms with 10ms step to smooth out these feature estimation and synchronize them with the short term spectral features.

The experiments have been developed and tested on AMI corpus, a multi-party and spontaneous speech set of recordings [14].

- **Development database:** We have selected 11 shows from AMI corpus as a development set to tune the weight values and number of Gaussians for the voice-quality and prosodic features. The total duration of the development set is 297 minutes with average number of four speakers. The development database is based on far-field microphone array channels sampled at 16KHz.
- **Test database:** We have carried out the test experiments on 12 AMI shows. We have used only the best weight values and optimum number of Gaussian of the development database on the test sets. The total duration of the test set is 324 minutes with average number of four speakers. The test database is also based on far-field microphone array channels sampled at 16KHz.

5.2. Experimental results

Diarization Error Rate (DER) metric is computed to assess the performance of different fusion approaches. DER is measured as the fraction of time that is not attributed correctly to a speaker, non-speech or speech. [1]

As shown in Figure 2, we have carried out different experiments on the development database to find out the optimum

[1] The scoring tool is the NIST RT scoring used as: ./md-eval-v21.pl -1 -nafc -o -R reference.rttm -S system hypothesis.rttm

set of weight values that gives us the best result in terms of DER. Figure 2 shows that the best weight value for MFCC set is 0.95 when they are fused both with the voice-quality and with prosodic features. However, the optimum weight value of MFCC when they are combined with the feature level fused voice quality and prosodic features is 0.98.

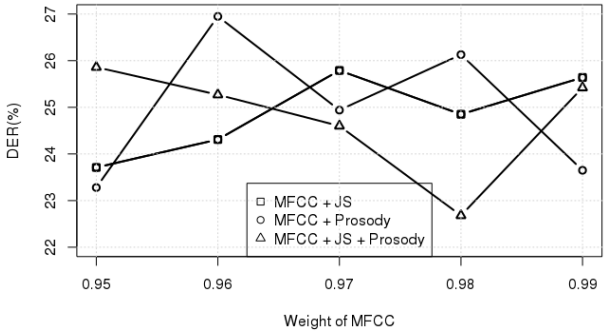

Figure 2: *DER of MFCC with JS, MFCC with Prosody, and MFCC with JS and Prosody on the development database*

Once we got the best weight values, we have carried out experiments to explore the best number of Gaussians per each feature set and combination. We have found out that the optimum number of Gaussians for voice-quality features is 2. The optimum number of Gaussians both for the prosodic, and feature level fused voice-quality with prosodic features is 3.

The baseline system of the development database which is based only on MFCC has 26.88% DER. Figure 2 shows that weighting the MFCC by 0.95 and the voice-quality features by 0.05 gives us a DER of 23.71%, which represents a 11.79% relative improvement compared to the baseline. Similarly, weighting the MFCC by 0.95 and the prosodic features by 0.05 gives us a DER of 23.28%, which represents a 13.39% relative improvement compared to the baseline. Figure 2 also shows that the fusion of the three feature sets shows the best DER when the weight of MFCC is set to 0.98. It shows a DER of 22.68%, which represents a 15.63% relative DER improvement compared to the baseline.

We have then used the tuned weight values and optimum number of Gaussians of the development database on our test data. It is seen in Table 1 that the baseline system of the test set which is based only on MFCC has a DER of 17.62%. Table 1 shows that fusion of spectral features with the prosodic ones gives us a DER of 16.5%, which is a 6.37% relative improvement compared to the baseline spectral features. Likewise, Table 1 shows that fusing the spectral features with the voice-quality ones shows a DER of 16.22%. This represents a 7.94% DER improvement compared to the baseline spectral features. Finally, the table shows that the fusion of spectral features with the voice-quality and prosodic features shows a DER of 13.44%, which represents a **23.72%** relative DER improvement compared to the baseline.

Figure 3 reports results of the test set. It shows the minimum, lower quartile, median, upper quartile, and maximum DER of different shows for different feature sets. Note that the figure reports the DER ranges of the test set. Such range decreases and it becomes lower when the spectral features are used together with the voice-quality and prosodic features. The figure also shows a similar trend in the value of the median. When the spectral features are joined with the voice-quality and prosodic information, the median DER values decrease. The combination of the different fusion systems reduce the DER error for most of the shows in the test set. However, error rate increases for some recordings compared to baseline system and reasons for this effect should be explored in the future.

Feature set	**MFCC Weight**	**DER(%)**
MFCC (Baseline)	1.0	17.62
MFCC + Prosody	0.95	16.5
MFCC + JS	0.95	16.22
MFCC + JS + Prosody	0.98	**13.44**

Table 1: *DER of MFCC, MFCC with Prosody, MFCC with JS and MFCC with JS and Prosody on the test database using optimum weight values and number of Gaussians of the development database*

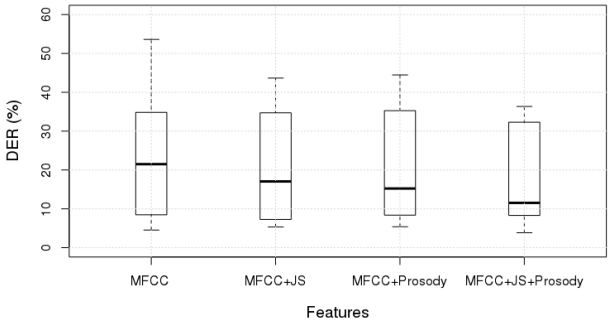

Figure 3: *DER box plots of test set shows for the different feature sets: MFCC, MFCC with JS, MFCC with prosody and MFCC with JS and prosody*

In overall, the use of voice-quality and prosodic features together with spectral ones increase the robustness and reliability of speaker diarization systems.

6. Conclusions

In this work, we have proposed the use of jitter and shimmer voice-quality measurements as complementary source of information to both long-term prosodic and short-term spectral features within speaker diarization task.

Experimental results on AMI corpus show that the fusion at the score level of both voice-quality and prosodic features with the spectral-based baseline system increases diarization performance. Results reported show a 7.9% and 6.4% relative DER improvement compared to the spectral baseline system, respectively. Moreover, the fusion of voice-quality with long-term prosodic information at the feature level and, their fusion with the spectral MFCC at the score level show a 23.72% relative DER improvement compared to the spectral baseline system.

Results reported in this work support the usefulness of voice-quality measurements as complementary source of information for speaker diarization task based both on spectral and long-term information.

7. Acknowledgment

This work has been partially funded by the Spanish Government projects TEC2012-38939-C03-02 and PCIN-2013-067 as well as from the European Regional Development Fund (ERDF/FEDER). This project has also received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 645323. This text reflects only the author's view and the Commission is not responsible for any use that may be made of the information it contains.

8. References

[1] S. Tranter and D. Reynolds, "An Overview of Automatic Apeaker Diarization systems," *IEEE Transactions on Audio, Speech, and Language Processing,*, 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] M. Farrús, A. Garde, P. Ejarque, J. Luque, and J. Hernando, "On the Fusion of Prosody, Voice Spectrum and Face Features for Multimodal Person Verification," in *9th International Conference on Spoken Language Processing, ICSLP*, 2006, pp. 2106–2109.

[4] 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.

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

[6] S. H. Yella and H. Bourlard, "Overlapping Speech Detection Using Long-term Conversational Features for Speaker Diarization in Meeting Room Conversations," *IEEE Transaction on Audio, Speech and Language Processing*, 2014.

[7] J. Kreiman and B. R. Gerratt, "Perception of aperiodicity in pathological voice," *The Journal of the Acoustical Society of America*, vol. 117, no. 4, 2005.

[8] X. Li, J. Tao, M. Johnson, J. Soltis, A. Savage, K. Leong, and J. Newman, "Stress and Emotion Classification using Jitter and Shimmer Features," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007.

[9] A. Sadeghi Naini and M. Homayounpour, "Speaker age interval and sex identification based on Jitters, Shimmers and Mean MFCC using supervised and unsupervised discriminative classification methods," in *8th International Conference on Signal Processing*, 2006.

[10] M. Farrús, J. Hernando, and P. Ejarque, "Jitter and Shimmer Measurements for Speaker Recognition," in *INTERSPEECH*, 2007.

[11] S. Zhang, "Emotion Recognition in Chinese Natural Speech by Combining Prosody and Voice Quality Features," in *5th International Symposium on Neural Networks*, 2008.

[12] C. Gobl and A. N. Chasaide, "The Role of Voice quality in Communicating Emotion, Mood and Attitude." *Speech Communication*, vol. 40, pp. 189–212, 2003.

[13] T. Johnstone and K. Scherer, "The Effects of Emotions on Voice quality," in *Proceedings of the XIV Int. Congress of Phonetic Sciences*, 1999.

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

[15] J. Fiscus and et al. (2002-2007) The Rich Transcription Evaluation Project. http://www.nist.gov/speech/tests/rt/.

[16] J. Luque and J. Hernando, "Robust Speaker Identification for Meetings: UPC CLEAR07 Meeting Room Evaluation System," in *Lecture Notes on Computer Science, Springer-Verlag, 2008*.

[17] D. Imseng and G. Friedland, "Tuning-Robust Initialization Methods for Speaker Diarization," *IEEE Transactions on Audio, Speech, and Language Processing*, 2010.

[18] J. Luque, C. Segura, and J. Hernando, "Clustering Initialization based on Spatial Information for Speaker Diarization of Meetings," in *International Conference on Spoken Language Processing, ICSLP*, Brisbane, Australia, 2008, pp. 383–386.

[19] 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.

[20] X. Anguera, C. Wooters, and J. Hernando, "Robust Speaker Diarization for Meetings: ICSI RT06s Evaluation System," in *International Conference on Spoken Language Processing, ICSLP*, 2006.

[21] J. Fiscus and et al., "The Rich Transcription Evaluation Project," Website, http://www.nist.gov/speech/tests/rt/, 2002-2009. [Online]. Available: http://www.nist.gov/speech/tests/rt/

[22] J. Luque and J. Hernando, "On the use of Agglomerative and Spectral Clustering in Speaker Diarization of Meetings," in *Odyssey 2012-The Speaker and Language Recognition Workshop*, 2012.

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