IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

CENATAV Voice-Group Systems for Albayzin 2018 Speaker Diarization Evaluation Campaign

Edward L. Campbell Hernández, Gabriel Hernández Sierra , José R. Calvo de Lara

Voice Group, Advanced Technologies Application Center, CENATAV, Havana, Cuba
ecampbell@cenatav.co.cu, gsierra@cenatav.co.cu, jcalvo@cenatav.co.cu

Abstract

Usually, the environment to record a voice signal is not ideal and, in order to improve the representation of the speaker characteristic space, it is necessary to use a robust algorithm, thus making the representation more stable in the presence of noise. A Diarization system that focuses on the use of robust feature extraction techniques is proposed in this paper. The presented features (such as Mean Hilbert Envelope Coefficients, Medium Duration Modulation Coefficients and Power Normalization Cepstral Coefficients) were not used in other Albayzin Challenges. These robust techniques have a common characteristic, which is the use of a Gammatone filter-bank for dividing the voice signal in sub-bands as an alternative option to the classical Triangular filter-bank used in Mel Frequency Cepstral Coefficients. The experiment results show a more stable Diarization Error Rate in robust features than in classic features.

Index Terms: Speaker Diarization, Robust feature extraction, Mean Hilbert Envelope Coefficients, Albayzin 2018 SDC

1. Introduction

This is the first participation of the CENATAV Voice Group in the Albayzin Challenges, participating in the Speaker Diarization Challenge (SDC) task and developing a Diarization System focuses in robust feature extraction. A Speaker Diarization System allows identifying " Who spoke when ? " on an audio stream, which has been of interest for the scientific community since the last century, with the emergence of the first works on speaker segmentation and clustering [1][2]. The diarization can be used as a stage that enriches and improves the results of other systems, for example: a Rich Transcription System uses the diarization for adding the information about who is speaking to the speech transcription, or a Speaker Recognition System uses it when the test signal has several speakers, so diarization allows finding the segments from the test signal with only one speaker [3].

An issue of the diarization is the environment where the speech is recorded, because noise is a natural condition in real applications. The proposed system is focused on robust feature extraction techniques for improving the results in a real application. Robust techniques as Mean Hilbert Envelope Coefficients (MHEC), Medium Duration Modulation Coefficients (MDMC) and Power Normalization Cepstral Coefficients (PNCC) are analysed.

The system was mainly developed on S4D tool [4], with the following structure: robust feature extraction, segmentation (gaussian divergence and Bayesian Information Coefficient), speech activity detection (Support Vector Machine), clustering (Hierarchical Agglomerative Clustering) and the last stage is the Re-segmentation (Viterbi algorithm). A system description is done in the next sections.

2. Robust Feature Extraction

A feature is robust when it has a stable effectiveness both in controlled or uncontrolled environment (noise, reverberation, etc.), being the second condition the most common in the practice [5], so the use of a feature with this characteristic is relevant in real applications. A brief description of several robust feature extraction techniques is provided in this section. These techniques use a gammatone filter-bank (see Fig. 1), the design of which was based on Patterson's ear model [6], defining the impulse response at the $channel_i$ for the equation 1.

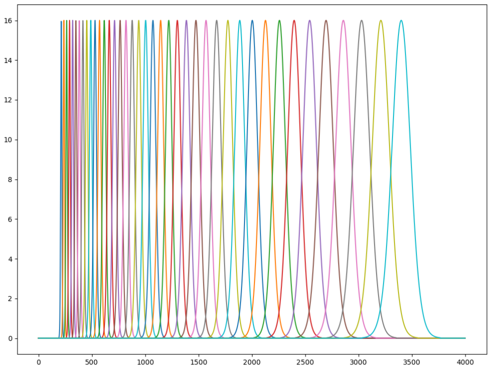

Figure 1: *Gammatone filter-bank of 40 dimension*

$$h(t)_i = \frac{\gamma * t^{\tau-1} * cos(2\pi * fc_i * t + \theta)}{exp(-2\pi * erb_i * t)}, \qquad (1)$$

where:

- γ**:** amplitude.
- τ**:** filter order.
- erb**:** equivalent rectangular bandwidth.
- fc_i**:** center frequency at the $channel_i$.
- θ**:** phase.

The next gammatone filter parameters were set in the proposed system, following Glasberg and Moore's recommendation [6], where:

- $fc_i = -(EarQ * minB) + \frac{(f_{max} + EarQ * minB)}{exp^{(i*0.5)/EarQ}}$
- $erb = (\frac{fc_i}{EarQ}^{\tau} + minB^{\tau})^{1/\tau}$
- $EarQ = 9.26449$
- $minB = 24.7$

10.21437/IberSPEECH.2018-47

$Earq$ is the asymptotic filter quality at high frequencies and $minB$ is the minimum bandwidth for low frequencies channels. The parameters θ, γ, τ were set to 0, 1 and 4 respectively.

2.1. Mean Hilbert Envelope Coefficients

A gammatone filter modulates a signal in amplitude and frequency [7], and to demodulate the output signal is a way for recovering the information transmitted. Mean Hilbert Envelope Coefficients (MHEC) extract this information by applying Hilbert Transform for estimating the analytic signal, separating the AM component from the modulated signal and assuming that the FM component does not exist [7]. The extraction process is shown in the figure 2.

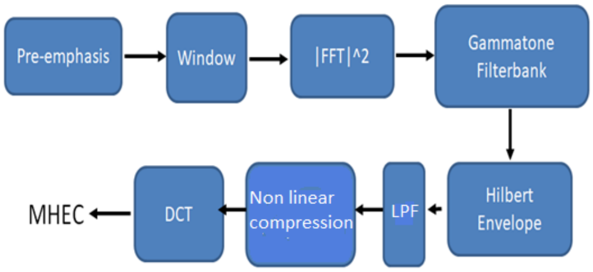

Figure 2: *Extraction process of MHEC*

A low-pass filtering is done on the estimated envelope in order to eliminate rapid changes of the signal, which are usually attributed to noise [8].

2.2. Medium Duration Modulation Coefficients

Medium Duration Modulation Coefficients (MDMC) is called " medium duration " because it employs a window of 52 ms, a bigger length than the traditional windowing of 20-30 ms. However, a length of 25 ms is used in this proposal for efficiency purposes. This feature takes the same approach that MHEC. It demodulates the gammatone output signal [9]. However, in MDMC is assumed that the FM component exists, applying the Teager's Operator (TEO) for estimating the AM component [9]. TEO is a non-linear operator that tracks the energy of a signal, which is a function of the amplitude and frequency [10]. The figure 3 shows the process for extracting the MDMC.

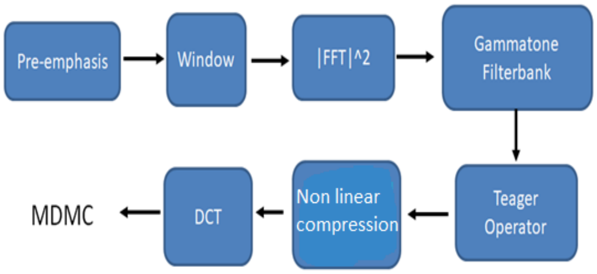

Figure 3: *Extraction process of MDMC*

2.3. Power Normalization Cepstral Coefficients

Power Normalization Cepstral Coefficients (PNCC) does not apply any technique for demodulating or separating the AM-FM component at the gammatone filter-bank output, rather the power at each sub-band is computed and transformed into the cepstral domain. There are two main approaches of PNCC, the short and medium time approaches, being the first the approach used in this paper. For more details about this technique, see [11]. The figure 4 shows the process for computing PNCC.

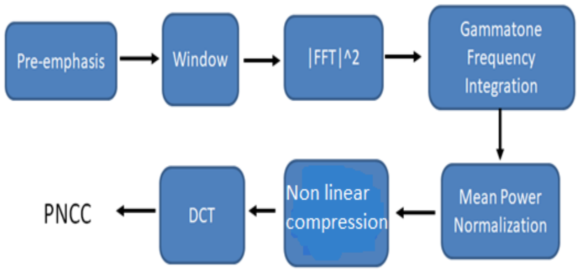

Figure 4: *Extraction process of PNCC*

3. Proposed Diarization System

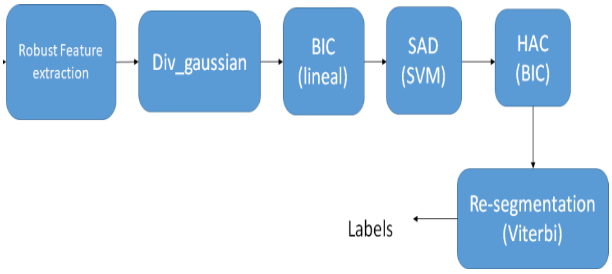

Figure 5: *Proposed Diarization System*

The proposed Diarization System (see figure 5) has six stages:

1. Robust feature extraction: based on MHEC, MDMC and PNCC as test features.
2. Gaussian Divergence: a sliding window of 2.5 seconds is applied along the signal, two Gaussians with diagonal covariance are computed from left and right half of the window in each shift. A change point of acoustic classes (different speakers, music and noise) is detected on a middle point of the window when the gaussian divergence score reaches a local maximum.
3. Bayesian Information Coefficient (BIC): Segmentation stage for decreasing the over-segmentation originated in the previous stage. A BIC approach is used for comparing consecutive segments and merging the couple segments that belong to the same acoustic class.
4. Support Vector Machine (SVM): A SVM [12] is applied as speech/non-speech classifier of the resulting segments from the previous stage. The non-speech segments are deleted.
5. Hierarchy Agglomerative Clustering (HAC): The segments that belong to the same speaker are clustered for a traditional HAC, using BIC as comparative measure.

6. Re-segmentation: a HMM is trained on the whole signal and a Viterbi re-segmentation is done for redefining the change points.

4. Experiment

The tested robust feature extraction, LFCC and LPCC algorithms are self-implementations, while MFCC implementation is from SIDEKIT tool [13] and the SVM from pyAudioAnalysis tool [12]. The Gaussian Divergence, BIC, HAC and Re-segmentation algorithms are is from S4D tool [4]. The proposed systems were submitted at closed condition and they do not use training data, with the exception of SVM, for which a portion of Albayzin SDC 2016 Database was used. The experiment was developed on RTVE-2018 SDC Development Database, tuning the default thresholds of BIC and HAC in S4D, and comparing robust (MHEC, MDMC, PNCC) and classic (MFCC, LFCC, LPCC) feature extraction techniques. The tables 1 and 2 show the best configurations at each feature. The pre-emphasis (0.97), length window (0.025 sec), shift window (0.01 sec), compression (logarithmic) and normalization (Cepstral Mean Normalization) are equal in each feature.

Table 1: *Robust features configuration*

Parameters	MHEC	MDMC	PNCC
Filter-bank dimension	40	40	40
Bandwidth (Hz)	0 - 7000	0 - 7000	0 - 7000
Cepstral coefficients	$12 + \Delta + \Delta\Delta$	$12 + \Delta + \Delta\Delta$	$15 + \Delta + \Delta\Delta$

Table 2: *Classic features configuration*

Parameters	MFCC	LFCC	LPCC
Filter-bank dimension	40	40	-
Bandwidth(Hz)	0 - 7500	0 - 7500	-
Prediction order	-	-	60
Cepstral coefficients	$19 + \Delta$	$19 + \Delta$	$19 + \Delta + \Delta\Delta$

5. Results

The figure 6 presents the results of the proposed diarization system, using several feature extraction techniques and separating these result for each TV-show of the SDC Development Database, ”La noche en 24H ” and ” Millennium ”.

The objective is to find a stable feature, where the Diarization Error Rate (DER) changes little between different signals. The figure 6 shows that the robust features are more stable that the classics features and MHEC is the most stable between these robust features. The general system proposed is showed in figure 5, and the primary and contrastive systems are based in the general system, using the following features:

- **Primary system:** MHEC. The most stable feature and the best performance on the ”La noche en 24H” group.
- **Contrastive system-1:** MDMC. The best general performance.
- **Contrastive system-2:** LFCC. The best performance on the ” Millennium ” group.

Despite MFCC is the most used feature extraction in speech processing, this algorithm does not present a better performance than the robust features proposed in this paper.

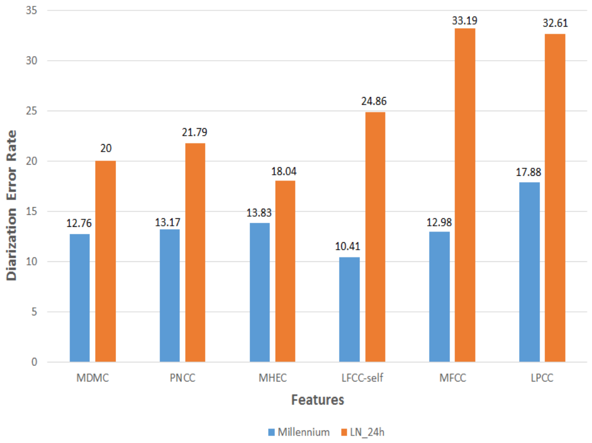

Figure 6: *Diarization Error Rate of the proposed system by feature.*

The computational cost was computed in terms of real-time factor. This measure represents the necessary time for processing a second of signal. The experiment was done on Intel Core i3-3110M CPU 2.40 GHz X 4 with 3.7 GB of memory. The computational cost of each system submitted is shown in the table 3, being the system with LFCC the most efficient.

Table 3: *Real-time factor of each system submitted*

Primary	Contrastive-1	Contrastive-2
0,30	0.23	0,04

6. Conclusion

This paper was reported by the CENATAV Voice-Group, and submitted for Albayzin 2018 Speaker Diarization Challenge. The main proposal was based on robust feature extraction, using MHEC as primary algorithm of feature extraction, with the objective of providing a stable system. The Diarization Error Rate of the primary system on development data was 15.23 %, with a real-time factor of 0.3, being suitable for a real application.

7. Acknowledgments

Finally, we would like to thank the University of Zaragoza, in the person of the professor Eduardo Lleida Solano, for his support during the experiment development. And to thank Wilma Recio for her help during the redaction of this paper.

8. References

[1] M. Diez, L. Burget, and P. Matejka, “Speaker diarization based on bayesian hmm with eigenvoice priors,” in *Odyssey 2018 The Speaker and Language Recognition Workshop, 26-29 June 2018, Les Sables dOlonne, France*, 2018.

[2] X. A. Miró, S. Bozonnet, N. W. D. Evans, C. Fredouille, G. Friedland, and O. Vinyals, “Speaker diarization: A review of recent research,” *IEEE Trans. Audio, Speech & Language Processing*, vol. 20, no. 2, pp. 356–370, 2012. [Online]. Available: https://doi.org/10.1109/TASL.2011.2125954

[3] E. L. Campbell, G. Hernandez, and J. R. Calvo, “Diarization system proposal,” in *IV Conferencia Internacional en Ciencias Com-*

putacionales e Informáticas (CICCI 2018), La Habana, Cuba, 2018.

[4] S. Meignier. (2015) Sd4. [Online]. Available: http://www-lium.univ-lemans.fr/s4d/

[5] N. T. Hieu, "Speaker diarization in meetings domain," Ph.D. dissertation, School of Computer Engineering of the Nanyang Technological University, 2014.

[6] M. Slaney, "An efficient implementation of the patterson-holdsworth auditory filter bank," Apple Computer, Tech. Rep., 1993.

[7] S. O. Sadjadi and J. H. L. Hansen, "Mean hilbert envelope coefficients (MHEC) for robust speaker and language identification," *Speech Communication*, vol. 72, pp. 138–148, 2015.

[8] E. L. Campbell, G. Hernandez, and J. R. Calvo, "Feature extraction of automatic speaker recognition, analysis and evaluation in real environment," in *International Workshop of Artificial Intelligent and Pattern Recognition 2018,Lecture Note of Computer Science*, 2018.

[9] V. Mitra, H. Franco, M. Graciarena, and D. Vergyri, "Medium-duration modulation cepstral feature for robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014*, 2014, pp. 1749–1753.

[10] A. Potamianos and P. Maragos, "A comparison of the energy operator and the hilbert transform approach to signal and speech demodulation," 1995.

[11] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 24, no. 7, pp. 1315–1329, 2016.

[12] T. Giannakopoulos. (2018) pyaudioanalysis. [Online]. Available: https://github.com/tyiannak/pyAudioAnalysis

[13] A. Larcher, S. Meignier, and K. A. LEE. (2017) Sidekit. [Online]. Available: http://www-lium.univ-lemans.fr/sidekit/