Bulletin of Electrical Engineering and Informatics
Vol. 10, No. 1, February 2021, pp. 374~382
ISSN: 2302-9285, DOI: 10.11591/eei.v10i1.1782

Towards an objective comparison of feature extraction techniques for automatic speaker recognition systems

Ayoub Bouziane, Jamal Kharroubi, Arsalane Zarghili
Laboratory of Intelligent Systems and Applications, University Sidi Mohamed Ben Abdellah, Morocco

Article Info

Article history:

Received Sep 7, 2019
Revised Dec 28, 2019
Accepted Mar 16, 2020

Keywords:

GFCCs
MFCCs
Speaker features
Speaker recognition

ABSTRACT

A common limitation of the previous comparative studies on speaker-features extraction techniques lies in the fact that the comparison is done independently of the used speaker modeling technique and its parameters. The aim of the present paper is twofold. Firstly, it aims to review the most significant advancements in feature extraction techniques used for automatic speaker recognition. Secondly, it seeks to evaluate and compare the currently dominant ones using an objective comparison methodology that overcomes the various limitations and drawbacks of the previous comparative studies. The results of the carried out experiments underlines the importance of the proposed comparison methodology.

Corresponding Author:

Ayoub Bouziane,
Laboratory of Intelligent Systems and Applications,
Sidi Mohamed Ben Abdellah University,
P.B: 2202 Immouzer road, Fez, Morocco.
Email: ayoub.bouziane@usmba.ac.ma

1. INTRODUCTION

Like the majority of biometric systems, contemporary speaker recognition systems are composed of two main building components: The feature extraction component and the speaker modeling & scoring component. The feature extraction component involves the processing of speech signal and the extraction of speaker-specific and discriminative characteristics as shown in Figure 1. The modeling & scoring block aims to train a reference model for each client speaker on the basis of its extracted features, as well as, to score the test utterances [1, 2].

The speaker features are usually categorized, as shown in Figure 2, into low-level and high-level features. The low-level features, also known as physical features, include the features influenced by the physical structure of the speaker's vocal tract. On the other hand, high-level features, also known as behavioral features, comprise the features influenced by the speaker's behavioral characteristics [3, 4]. Typically, high-level features are often used just as complementary features to the low-level features. In this study, we focus only on the low-level features (physical features) [5].

Although the performance of feature extraction techniques depend mainly on how the extracted features are modeled, the overall previous comparative studies on features extraction techniques were carried out independently of the used speaker modeling technique, neither its parameters (see the literature review in section two). With this in mind, the present paper aims firstly to review the most significant advancements in feature extraction techniques used for automatic speaker recognition. Secondly, it seeks to evaluate and compare the currently dominant ones using an objective comparison methodology that overcome the limitation of the previous studies.

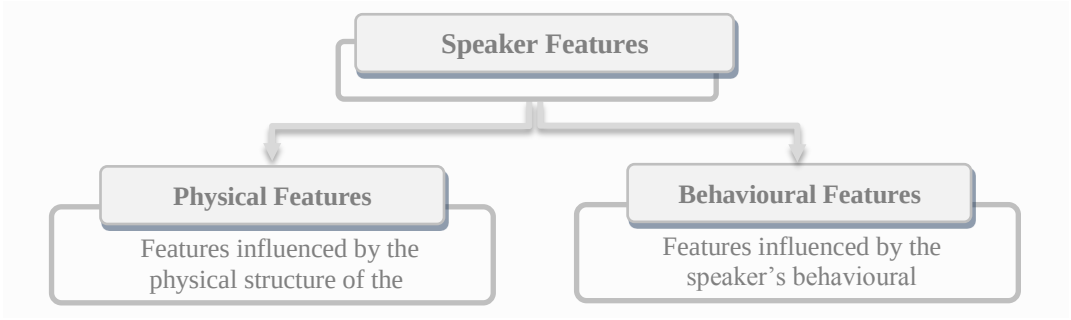

Figure 1. The two main categories of speaker features

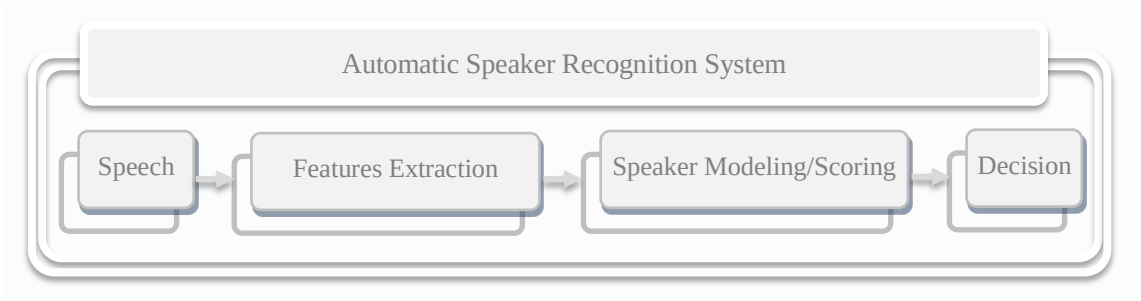

Figure 2. The two main blocks of speaker recognition systems

The remainder of this paper is organized as follows. The second section presents a concise literature review of the most significant advancements in feature extraction techniques. Next, the outlines of the proposed comparison methodology are summarized in the third section. Afterward, the experiments are reported and discussed in the fourth section. Finally, conclusions and future research directions are drawn in the last section.

2. LITERATURE REVIEW

The earlier low-level features used for speaker recognition systems were firstly, as shown in Figure 3, based on spectral energy patterns [6, 7]. Further, an improved features have been proposed based on the variance analysis of the spectral energy patterns [8]. Later on, fast fourier transform (FFT)-based cepstral coefficients were presented and yielded good performance, confirming their usefulness for speaker recognition. In 1974, the concept of linear prediction was introduced for speaker recognition tasks [9]. Instead of being used by themselves, the linear prediction coefficients were transformed into a set of robust, less correlated features such as, the linear prediction cepstral coefficients "LPCCs" [9], the perceptual linear prediction coefficients "PLP" [10], the perceptual linear predictive cepstral coefficients "PLPCC" [11-15] and the line spectral frequencies "LSF" [16] etc. In early 1980s, the so-called Mel-frequency cepstral coefficients "MFCCs" were introduced and yielded the best results compared to contemporary used features for speaker recognition [5, 11]. One year later, the concept of dynamic features has been introduced to incorporate some temporal information to the extracted features [16, 17]. Later on, seen that the original idea of Davis and Mermelstein does not provide an explanation about the choice of several parameters within the calculation process of the MFCCs, a numerous variations and improvements of the original proposed idea have been proposed in the literature [13, 15, 18, 19]. Furthermore, a number of discrete wavelet packet transform (DWPT)-based features have been proposed for speech and speaker verification systems [20, 21].

On another side, several features have been proposed for speaker recognition in specific scenarios [22]. Some of them have been proposed for noisy scenarios, such as minimum variance distortion less response features "MVDR" [23, 24], mean Hilbert envelope coefficients "MHEC" [25, 26], medium duration modulation cepstrum "MDMC" [27], and power normalized cepstral coefficients "PNCC" [28]. Some others have been proposed for reverberant scenarios, such as the frequency domain linear prediction features "FDLP" [29, 30]. Others been proposed to address the feature distribution distortion caused by transmission channel effects, such as the multitaper MFCC features [31, 32]. Recently, the gammatone

frequency cepstral coefficients (GFCC) were proposed and achieved a promising recognition performance in some speaker recognition applications compared to the Mel-frequency cepstral coefficients, especially in noisy acoustical environment [33–39].

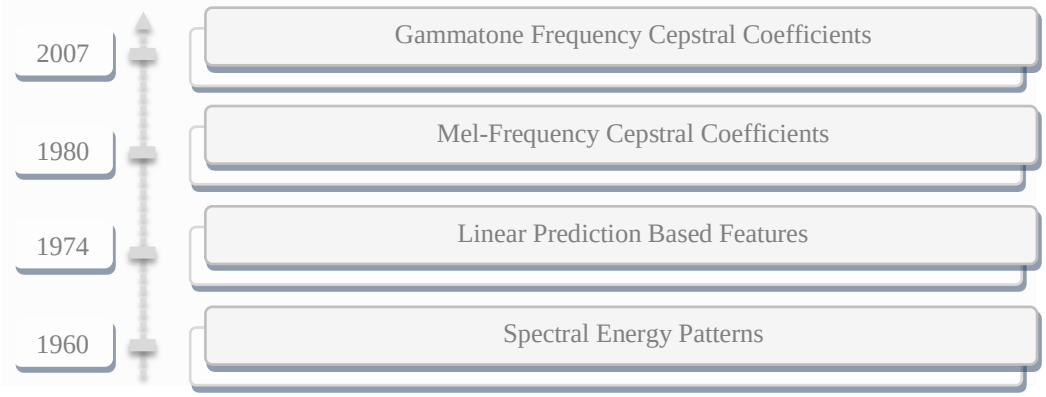

Figure 3. Timeline view of the most significant developments in features extraction techniques

The multiplicity and diversity of the proposed feature extraction techniques raised the necessity of directly comparing them in a common experimental setup. Accordingly, several comparative studies were therefore conducted to assess the performance of the different proposed features extraction techniques. In [12], a comparative study of various features (MFCC, LFCC, LPCC and PLPC features) have been conducted under a GMM-based speaker identification system of 32 Gaussian components. The findings of the study reported somewhat better performance of the MFCC and LPCC features than the other features. On similar lines, several feature extraction techniques (LPC, LPCC, the standard MFCC features and its recent variant, commonly known as the human factor cepstral coefficients "HFCC") were evaluated and compared under a ANN-based speaker identification system [40]. The reported results confirmed the superior performance of the MFCC features compared to the LPC and LPCC features. Later on, Ganchev and his colleagues were conducted a comparative study of the most popular variants of the MFCC features for speaker verification using a probabilistic neural network based system [14]. The obtained results revealed that the most successful variant was the Slaney's variant [19]. On the other hand, several features have been assessed and evaluated in specific scenarios. Senthil and Dandapat have compared several speech features (MFCC, LPC, LPCC …) for both speaker verification and identification under stressed condition [41]. The comparison was done under tow systems: a sixteen-order GMM based speaker recognition system and a VQ based system using sixteen-codeword codebook. The finding of the study revealed that the MFCC features still outperforms the others features in stressed condition. The reported results revealed that the MFCC features still outperform the other features, even in stressed conditions. In a further study [42], the MFCC and the LPCC futures were also assessed in speech resynthesized conditions. The assessment was done using a GMM-UBM based speaker verification system of 128 Gaussians. The authors concluded that the MFCC features remain the best choice for speaker recognition in both clean and resynthesized conditions. Conversely, it was shown in noisy conditions that the recognition performance using the MFCC features degrades significantly compared to the system performance using the GFCC features [43].

3. TOWARDS AN OBJECTIVE COMPARISON METHODOLOGY OF SPEAKER FEATURES

The adopted comparison methodology in the previously mentioned comparative studies consists in comparing the obtained performance in each feature extraction technique using the same speaker modeling technique and the same modeling parameters in most of them. In other words, the followed methodology consists in comparing the corresponding system-performances of the feature extraction techniques independently on the used modeling technique and its parameters. However, the system performance depends on both the robustness of the extracted features and the manner in which they are modeled. By way of illustration, a feature extraction technique based on a filter-bank of many filters tends to produce many acoustic classes that require many Gaussians components to model them perfectly, and vice versa. Additionally, a feature extraction technique may give the better performances in conjunction with some speaker modeling techniques and relatively decreased performance than the other feature extraction

techniques when using other speaker modeling techniques. Moreover, a feature extraction technique with an appropriate modeling could possibly give the best performance within a segment of speakers, whilst in another segment of speakers, its performance mightn't be the best using the previously selected modeling parameters**.**

Taking into account these considerations and assumptions, we suggest an objective methodology for comparing feature extraction techniques in speaker recognition systems. The outlines of the proposed methodology can be summarized as follows: (1) the performance of feature extraction techniques should be assessed in conjunction with several speaker modeling techniques, (2) the produced features through each feature extraction technique must be modeled using the appropriate modeling parameters and, finally, (3) The comparison process has to be done by means of a cross-validation strategy.

Acting on this methodology, we have conducted a comparative study on the most popular variants of traditional MFCC features, the recently introduced GFCC features, as well as, on their corresponding dynamic features. The comparison of those techniques is done in conjunction with the most three widely used speaker modeling techniques in the speaker recognition community: (1) the GMM-UBM, (2) the hybrid GSV-SVM and (3) the state of art i-vectors/CSS based speaker-modeling techniques. Furthermore, the extracted features of each feature extraction technique are modeled using the most appropriate modeling parameters pertaining to the used modeling technique. Moreover, the evaluation procedure was done through two experiment series: the first series attempts to compare the different features extraction techniques on the overall evaluation dataset using the most appropriate modeling parameters, whereas the second series aims to assess their performances through cross-validation strategy (see the experimental protocol). During this study, our focus is put on the traditional and the widely used MFCC features, the recently introduced GFCC features, as well as, on their corresponding dynamic features [44-46].

4. EXPERIMENTS, RESULTS AND DISCUSSION

4.1. The experimental protocol

The performed experiments in this study were conducted on the THUYG-20 SRE database [47]. In the training phase, the whole THUYG-20-SRE's development-set is used for training the system hyper-parameters (UBM, TVM …), whereas in the enrolment phase of the system, one minute of active speech per speaker is used for the building the speaker's models. In the testing phase, the client speakers are tested against each other, resulting in total of 5862 target and 891024 impostor trials of 4 seconds and 2886 target and 438672 impostor trials of 8 seconds. The MFCCs features are pre-processed as follows. The emphasizing step is firstly performed using a simple first order digital filter with transfer function H(z)=1-0.95z. Next, the emphasized speech signal is blocked into Hamming-windowed frames of 25 ms (400 samples) in length with 10 ms (160 samples) overlap between any two adjacent frames [5, 48, 49]. As regards the GFCCs, the features are extracted using a filter bank of 64 Gammatone filters and a down sampling frequency of 100 Hz (yielding frame rate of 10 ms), as recommended by [36]. The performance of the system is measured in the first experiment series using the equal error rate (EER) values, whereas in the second experiment series where a pre-adjusted threshold is fixed, the verification performance of the system is assessed using the half total error rate (HTER) values. Besides, the verification threshold is adjusted to the EER point where the false rejection and the false acceptance rates are equal.

The performed experiments in this paper were divided into two series. The experiments of the first series were carried out on the overall evaluation dataset using the most appropriate modeling parameters (model size, relevance factor …), whereas the experiments belonging to the second series were performed through a 2-fold cross-validation strategy. The specifics of the cross-validation strategy are as follows. Firstly, the client speakers of the evaluation dataset are divided into two roughly equal-sized sets, say D_1 and D_2. Next, the first dataset D_1 is used to estimate the appropriate modeling parameters and the verification threshold, whereas the other dataset D_2 is used to evaluate the system performance based on the previously estimated modeling parameters and verification threshold (*Scenario I*). Afterwards, the two datasets alternate their roles and the second dataset D_2 is therefore used for estimating the modeling parameters and the verification threshold, whilst the dataset D_1 is used to evaluate the system performance (Scenario II). Finally, the obtained results in the two scenarios are averaged out. The most appropriate modeling parameters of each speaker modeling technique are selected through an exhaustive grid search over the followings typical values:

- The UBM size (GMM-UBM, GSV-SVM, i-vectors): 32, 64, 128, 256, 512 and 1024.
- The MAP relevance factor (GMM-UBM, GSV-SVM): 0, 4, 8, 12, 16 and 20.
- The dimension of the total variability matrix (TVM): 200, 400 and 600.
- The optimal N° of iterations used for TVM training (i-vectors): 1, 2, 3, …, 20.
- The dimension of the PLDA speaker subspace (i-vectors): 400, 800, 1200, 1600 and 2000.

4.2. The first experiment series

The obtained results in the first experiment series for speaker verification and identification tasks are shown in Figure 4. Primarily, the findings revealed that best performances across the various used speaker-modeling techniques were achieved in the both tasks using the HTK's variant of the MFCC features. Furthermore, it can be seen that the overall MFCCs variants demonstrates better performances than the GFCC features, regardless of the used speaker modeling technique. Surprisingly, the obtained results indicate also that the combination of the static and dynamic features doesn't bring any gain in performance compared to the obtained results using the static features only. Quite the contrary, the results show that the system performance is degraded when the static features are combined with their respective dynamic ones.

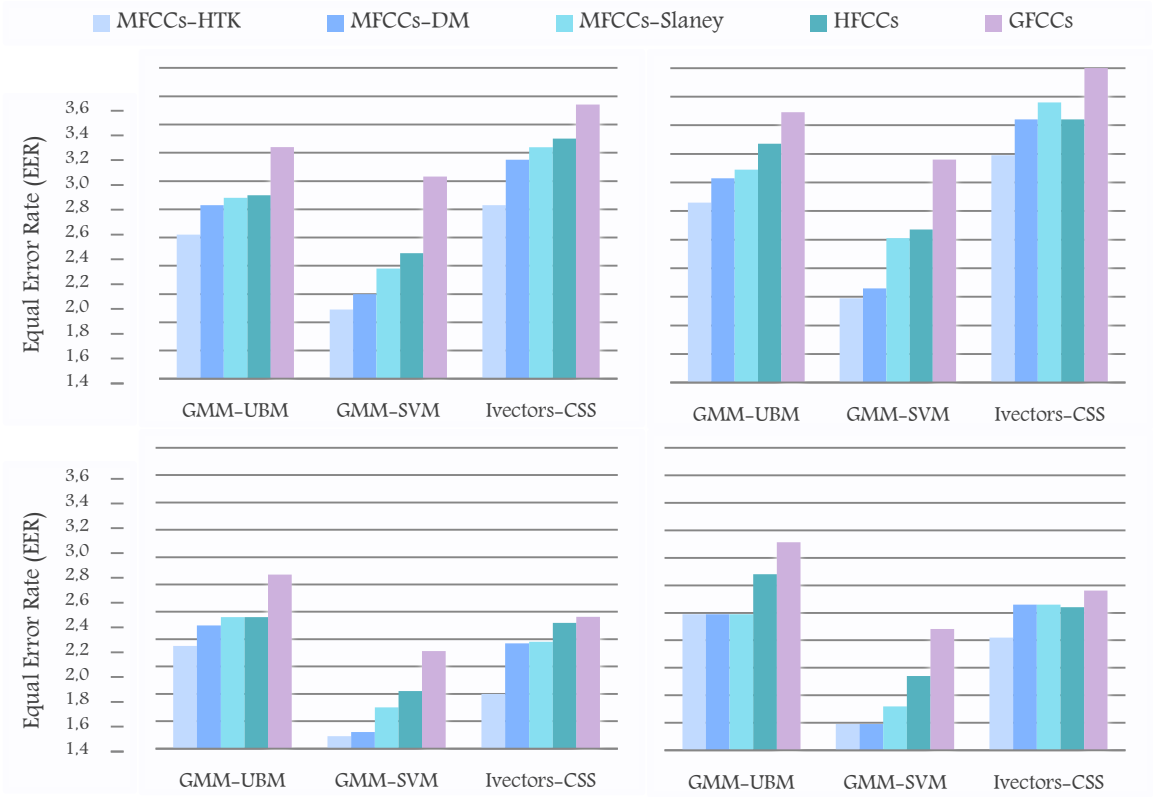

Figure 4. The obtained EERs across the various features extraction and speaker modeling techniques (verification task), the upper and bottom bar charts reflect the obtained EERs using test utterances of 4s and 8s, respectively, the left bar charts represent the obtained EERs using the static features only, whereas the right bar charts represent the obtained EERs using a combination of the static features with their corresponding dynamic features

On another side, although the best performances of the various features extraction techniques were obtained mostly in association with the GSV-SVM based speaker modeling technique, it can be noticed that the choice of the speaker modeling technique influences on the relative performance of each feature extraction technique compared to the others. The exhibited results demonstrate also that the increase in the test-utterances' duration is translated into an increase in the recognition performance of the system. Accurately, it seems that the amount of the achieved performance gain depends mainly on the used speaker modeling technique. In an attempt to underline this observation, we have computed the relative error reduction rates obtained through the increase of the test utterances from 4s to 8s, for instance, from the obtained results using static features and test utterances of 4s. The obtained results are reported in Table 1.

The results revealed that the achieved gain in performance using the i-vectors-CSS based speaker modeling technique is greater than that achieved using the GSV-SVM based speaker modeling technique by roughly 24 and 13 percent, as well as, greater too than the gain achieved using the GMM-UBM based speaker modeling technique by roughly 133 percent. As regards the feature extraction techniques, it can be seen that the amount of the test duration doesn't have a great influence on the relative performance of the various features extraction techniques.

Table 1. The relative EER reduction rates obtained through the increase of the test utterances from 4 to 8 seconds

	MFCCs-HTK	MFCCs-DM	MFCCs-Slaney	HFCCs	GFCCs
GMM-UBM	11.16 %	12.55 %	11.94 %	12.59 %	12.17 %
GMM-SVM	21.16 %	24.00 %	11.94 %	20.52 %	25.44 %
I-vectors-CSS	31.56 %	26.44 %	28.29 %	25.16 %	29.34 %
Means	11.16 %	12.55 %	11.94 %	**12.59** %	12.17 %

The selected modeling parameters for each feature extraction technique in each speaker modeling technique are shown in Table 2. The most important result to emerge from the table is that the various features extraction techniques don't share the same modeling parameters in each speaker modeling technique. Additionally, it can be noticed that there is such relation between the design of the used filter-bank (the shape, the number of filters …) in each feature extraction technique, the number of Gaussian and the relevance factor used for speaker modeling. Except the GFCC features and the HFCC features where the filters that compose the filter-bank are more interfered, it can be remarked more precisely that the greater the number of used filters, the greater the number of Gaussian components required or the lower the relevance factor (i.e. the higher the influence of the training speech data on the adapted models) is.

Table 2. The selected modeling parameters for each feature extraction technique/modeling technique

	GMM-UBM		GSV-SVM		ivectors-CSS	
	GMM sizes	Relevance factors	GMM sizes	Relevance factors	GMM sizes	Number of iterations
GFCCs	512	12	1024	0	1024	2
HFCCs	512	4	512	4	1024	3
MFCCs-HTK	1024	8	1024	8	512	3
MFCCs-Slaney	1024	4	1024	4	1024	4
MFCCs-DM	128	0	256	4	256	1

4.3. The second experiment series

The obtained results in the second experiment series are shown in Table 3. The results are expressed as means and variances of the obtained performances in the two cross-validation scenarios. The average values represent the system performance, whereas the variance values reflect the performance variability across speakers segments. The findings revealed that the best performances, i.e. lower average values of the HTER and IER metrics, were obtained using the HTK's MFCC features together with the GSV-SVM based speaker modeling technique.

Table 3. The means and variances of the obtained HTERs in the two scenarios of the cross-validation based evaluation strategy (Using static features only and both 4s and 8s-test-utterances)

		GMM-UBM	GSV-SVM	ivectors-CSS
GFCCs	μ	2.89	11.24	**3.24**
	σ^2	0.003	156.5	**0.080**
HFCCs	μ	2.56	2.36	3.54
	σ^2	0.003	0.005	0.708
MFCCs-HTK	μ	**2.40**	**2.14**	3.50
	σ^2	**0.000**	**0.001**	2.531
MFCCs-Slaney	μ	2.63	2.54	3.27
	σ^2	0.186	0.036	0.744
MFCCs-DM	μ	2.87	2.17	3.88
	σ^2	0.238	0.048	2.856

Additionally, it can be seen more clearly that the used speaker modeling technique has a significant influence on the relative performances of the various features extraction technique. As an illustration, it can be seen that the GFCC features yielded concurrently the best and the worst verification performances, depending upon the used speaker modeling technique. By using the i-vectors based speaker modeling technique, it can be seen that the GFCC features demonstrate the best verification performance. Conversely, the GFCC features give the worst verification performance when using the GMM-UBM based speaker modeling technique.

5. CONCLUSION AND FUTURE RESEARCH DIRECTIONS

In this paper, an objective comparison methodology was proposed and applied for the comparison of the currently dominant feature extraction techniques in automatic speaker recognition systems, namely the most popular variants of traditional MFCC features, the recently introduced GFCC features, as well as, their corresponding dynamic features. The findings of the study can be summarized in two main points. On the one hand, the obtained results throughout the study underlined the importance of the proposed methodology. On the other hand, it has been shown using the proposed methodology that the best performances were obtained using the HTK's MFFC variant.

REFERENCES

[1] A. Bouziane, J. Kharroubi, and A. Zarghili, "Probabilistic Self-Organizing Maps for Text-Independent Speaker Identification," *TELKOMNIKA Telecommunication Computing Electronics and Control*, vol. 16, no. 1, pp. 250–258, February 2018.

[2] A. Bouziane, J. Kharroubi, and A. Zarghili, "Towards an Optimal Speaker Modeling in Speaker Verification Systems using Personalized Background Models," *International Journal of Electrical & Computer Engineering (IJECE)*, vol. 7, no. 6, pp. 3655–3663, December 2017.

[3] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, Jan. 2010.

[4] S. Singh and P. Singh, "High Level Speaker Specific Features as an Efficiency Enhancing Parameters in Speaker Recognition System," *International Journal of Electrical & Computer Engineering (IJECE)*, vol. 10, no. 4, pp. 2443-2450, Aug. 2020.

[5] B. Ayoub, K. Jamal and Z. Arsalane, "An analysis and comparative evaluation of MFCC variants for speaker identification over VoIP networks," *2015 World Congress on Information Technology and Computer Applications (WCITCA)*, Hammamet, pp. 1-6, 2015.

[6] P. Denes and M. V. Mathews, "Spoken Digit Recognition Using Time- Frequency Pattern Matching," *The Journal of the Acoustical Society of America*, vol. 32, no. 11, pp. 1450–1455, Nov. 1960.

[7] S. Pruzansky, "Pattern- Matching Procedure for Automatic Talker Recognition," *The Journal of the Acoustical Society of America*, vol. 35, no. 3, pp. 354–358, Mar. 1963.

[8] S. Pruzansky and M. V. Mathews, "Talker- Recognition Procedure Based on Analysis of Variance," *The Journal of the Acoustical Society of America*, vol. 36, no. 11, pp. 2041–2047, Nov. 1964.

[9] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *The Journal of the Acoustical Society of America*, vol. 55, no. 6, pp. 1304–1312, Jun. 1974.

[10] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, Apr. 1990.

[11] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357-366, August 1980.

[12] D. A. Reynolds, "Experimental evaluation of features for robust speaker identification," in *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 639-643, Oct. 1994.

[13] M. D. Skowronski and J. G. Harris, "Exploiting independent filter bandwidth of human factor cepstral coefficients in automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 116, no. 3, pp. 1774–1780, Sep. 2004.

[14] T. Ganchev, N. Fakotakis, and G. Kokkinakis, "Comparative evaluation of various MFCC implementations on the speaker verification task," in *Proceedings of the SPECOM 2005*, vol. 1, no. 2005, pp. 191–194, 2005.

[15] S. Young *et al.*, "The HTK book (for HTK version 3.4)," *Cambridge university engineering department*, vol. 2, no. 2, pp. 2–3, 2006.

[16] X. Huang, A. Acero, and H.-W. Hon, *Spoken language processing: a guide to theory, algorithm, and system development*. Upper Saddle River, NJ: Prentice Hall PTR, 2001.

[17] S. Furui, "Cepstral analysis technique for automatic speaker verification," in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 2, pp. 254-272, April 1981.

[18] S. Young, J. Jansen, J. Odell, D. Ollason, and P. Woodland, "The HTK Book (for HTK V2. 0)," *Cambridge university engineering department*, 1996.

[19] M. Slaney, "Auditory toolbox," Interval Research Corporation, pp. 1-52, 1998.

[20] R. Sarikaya, B. L. Pellom, and J. H. Hansen, "Wavelet packet transform features with application to speaker identification," in *IEEE Nordic Signal Processing Symposium*, pp. 81–84, 8-11 June 1998.

[21] M. Siafarikas, T. Ganchev, and N. Fakotakis, "Objective wavelet packet features for speaker verification," in *In Eighth International Conference on Spoken Language Processing*, Jeju Island, Korea, 4-8 October, 2004.

[22] Q. Jin and T. F. Zheng, "Overview of front-end features for robust speaker recognition," *Proc APSIPA*, 2011.

[23] M. N. Murthi and B. D. Rao, "Minimum variance distortionless response (MVDR) modeling of voiced speech," *IEEE International Conference on Acoustics, Speech, and Signal Processing, Munich*, vol. 3, pp. 1687-1690, 1997.

[24] M. Wölfel, J. W. McDonough, and A. Waibel, "Minimum variance distortionless response on a warped frequency scale," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[25] S. O. Sadjadi and J. H. Hansen, "Assessment of single-channel speech enhancement techniques for speaker identification under mismatched conditions," in *Eleventh Annual Conference of the International Speech Communication Association*, pp. 1021-1024, 2010.
[26] S. O. Sadjadi and J. H. L. Hansen, "Hilbert envelope based features for robust speaker identification under reverberant mismatched conditions," *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, pp. 5448-5451, 2011.
[27] V. Mitra, H. Franco, M. Graciarena and A. Mandal, "Normalized amplitude modulation features for large vocabulary noise-robust speech recognition," *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, pp. 4117-4120, 2012.
[28] C. Kim and R. M. Stern, "Power-Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1315-1329, July 2016.
[29] S. Thomas, S. Ganapathy and H. Hermansky, "Recognition of Reverberant Speech Using Frequency Domain Linear Prediction," in *IEEE Signal Processing Letters*, vol. 15, pp. 681-684, 2008.
[30] S. Ganapathy, J. Pelecanos and M. K. Omar, "Feature normalization for speaker verification in room reverberation," *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, pp. 4836-4839, 2011.
[31] M. Hansson and G. Salomonsson, "A multiple window method for estimation of peaked spectra," in *IEEE Transactions on Signal Processing*, vol. 45, no. 3, pp. 778-781, March 1997.
[32] T. Kinnunen, R. Saeidi, J. Sandberg, and M. Hansson-Sandsten, "What else is new than the Hamming window? Robust MFCCs for speaker recognition via multitapering," in *11 Annual Conf of the International Speech Communication Association*, pp. 2734-2737, 2010.
[33] Y. Shao, S. Srinivasan and D. Wang, "Incorporating Auditory Feature Uncertainties in Robust Speaker Identification," *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP '07*, Honolulu, HI, pp. IV-277-IV-280, 2007.
[34] Yang Shao and DeLiang Wang, "Robust speaker identification using auditory features and computational auditory scene analysis," *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, NV, pp. 1589-1592, 2008.
[35] Y. Shao, Z. Jin, D. Wang and S. Srinivasan, "An auditory-based feature for robust speech recognition," *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, Taipei, pp. 4625-4628, 2009.
[36] X. Zhao, Y. Shao and D. Wang, "CASA-Based Robust Speaker Identification," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1608-1616, July 2012.
[37] X. Zhao and D. Wang, "Analyzing noise robustness of MFCC and GFCC features in speaker identification," *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, pp. 7204-7208, 2013.
[38] A. Biswas, P. K. Sahu, A. Bhowmick, and M. Chandra, "Hindi vowel classification using GFCC and formant analysis in sensor mismatch condition," *WSEAS Trans Syst*, vol. 13, pp. 130–143, 2014.
[39] B. Ayoub, K. Jamal and Z. Arsalane, "Gammatone frequency cepstral coefficients for speaker identification over VoIP networks," *2016 International Conference on Information Technology for Organizations Development (IT4OD)*, Fez, pp. 1-5, 2016.
[40] G. Saha, P. Kumar, and S. Chakroborty, "A Comparative Study of Feature Extraction Algorithms on ANN Based Speaker Model for Speaker Recognition Applications," *International Conference on Neural Information Processing*. Springer, Berlin, Heidelberg, pp. 1192–1197, 2004.
[41] G. Senthil Raja and S. Dandapat, "Speaker recognition under stressed condition," *International Journal of Speech Technology*, vol. 13, no. 3, pp. 141–161, Sep. 2010.
[42] D. Yessad and A. Amrouche, "Robust regression fusion of GMM-UBM and GMM-SVM normalized scores using G729 bit-stream for speaker recognition over IP," *International Journal of Speech Technology*, vol. 17, no. 1, pp. 43–51, Mar. 2014.
[43] Fengsong Hu and Xiaoyu Cao, "An auditory feature extraction method for robust speaker recognition," *2012 IEEE 14th International Conference on Communication Technology*, Chengdu, pp. 1067-1071, 2012.
[44] G. Chaudhary, S. Srivastava, and S. Bhardwaj, "Feature Extraction Methods for Speaker Recognition: A Review," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 31, no. 12, Apr. 2017.
[45] G. Dişken, Z. Tüfekçi, L. Saribulut, and U. Çevik, "A Review on Feature Extraction for Speaker Recognition under Degraded Conditions," *IETE Tech. Rev.*, vol. 34, no. 3, pp. 321–332, May 2017.
[46] S. S. Tirumala, S. R. Shahamiri, A. S. Garhwal, and R. Wang, "Speaker identification features extraction methods: A systematic review," *Expert Syst. Appl.*, vol. 90, , pp. 250–271, Dec. 2017.
[47] A. Rozi, Dong Wang, Zhiyong Zhang and T. F. Zheng, "An open/free database and Benchmark for Uyghur speaker recognition," *2015 International Conference Oriental COCOSDA held jointly with 2015 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, Shanghai, pp. 81-85, 2015.
[48] P. Kenny, P. Ouellet, N. Dehak, V. Gupta and P. Dumouchel, "A Study of Interspeaker Variability in Speaker Verification," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 980-988, July 2008.
[49] X. Zhou, D. Garcia-Romero, R. Duraiswami, C. Espy-Wilson and S. Shamma, "Linear versus mel frequency cepstral coefficients for speaker recognition," *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, Waikoloa, HI, pp. 559-564, 2011.

BIOGRAPHIES OF AUTHORS

Ayoub Bouziane received the M.Sc. degree in intelligent systems and networks from the Faculty of Sciences and Technologies, Fez, Morocco, in 2012. He is currently pursuing the Ph.D. degree in the intelligent systems and applications laboratory, Sidi Mohammed ben Abdullah University, Morocco. His research interests cover signal processing and machine learning, mostly with applications in automatic speaker recognition. In parallel with his research activities, he teaches undergraduate level courses in computer science, at the Faculty of Sciences and Technologies, Fez. Additionally, He is a member of IEEE Signal Processing Society, IEEE Computational Intelligence Society, IEEE Computer Society and the International Speech and Communication Association (ISCA).

Jamal Kharroubi has his B.Sc. in Computer Science from Sidi Mohamed Ben Abdellah University (Fez-Morocco) in 1996. Two years after, he got his postgraduate degree in the domain of Artificial Intelligence from Galilee's Institute - Paris XIII University. In 2002, He received his Ph.D. degree in automatic speaker recognition systems from Telecom ParisTech (Ecole Nationale Supérieure des Télécommunications de Paris-France)". Since January 2003, he is an associate professor in the Department of Computer Science at the Faculty of Science and Technology. In 2008, he received his HDR diploma (Habilitation to conduct research). Additionally, He is currently the coordinator of the Master of Intelligent Systems and Networks. Moreover, he is the author of more than thirty publications in peer-reviewed scientific journals & conference proceedings. His research interests are focused on signal and image processing, pattern recognition, etc.

Arsalane Zarghili is a Doctor of Science from Sidi Mohamed Ben Abdellah University (Fez-Morocco). He received his Ph.D. in 2001 and joined the same University in 2002 as Professor at the computer science department of the Faculty of Science and Technology (FST). In 2007 he was head of the computer sciences department and chair of the Software Quality Master in the FST-Fez. He lectures Programming, Distributed, compilation and Information processing, for both undergraduate and master levels. In 2008 he obtained his HDR in information processing. In 2011, he is the co-founder and the head of the Laboratory of Intelligent Systems and Applications in the FST of Fez. He is a member of the steering committee of the department of computer sciences and was a member of the faculty board. He is also IEEE member since 2011. His main research is about pattern recognition, image indexing and retrieval systems in cultural heritage, biometric, etc.