Robust features fusion for text independent speaker verification enhancement in noisy environments

Conference Paper · May 2017

DOI: 10.1109/IranianCEE.2017.7985357

2 authors:

Mohsen Mohammadi
Iranian Research Institute for Electrical Engineering, ACECR, Tehran, Iran
10 PUBLICATIONS **13** CITATIONS
SEE PROFILE

H. R. Sadegh Mohammadi
Iranian Research Institute for Electrical Engineering, ACECR, Tehran, Iran
70 PUBLICATIONS **145** CITATIONS
SEE PROFILE

Robust Features Fusion for Text Independent Speaker Verification Enhancement in Noisy Environments

Mohsen Mohammadi
PhD student
Iranian Research Institute for Electrical Engineering
ACECR
Tehran, Iran
mohammadi.mohsen@gmail.com

Hamid Reza Sadegh Mohammadi
Associate Professor
Iranian Research Institute for Electrical Engineering
ACECR
Tehran, Iran
mohammadis@acecr.ac.ir

Abstract—**So far, many methods have been proposed for speaker verification which provide good results, but their performances reduce in actual noisy environments. A common approach to partially alleviate this problem is the fusion of several methods. In this paper, four systems based on different speech features, i.e., MFCC, IMFCC, LFCC, and PNCC were combined in score-level to improve verification accuracy under clean and noisy speech conditions. The features pairwise and foursome fusion in a speaker verification system based on speaker modeling through the Gaussian mixture model (GMM) were evaluated. TIMIT and NOISEX92 databases were used to implement as the speech and noise datasets, respectively. The experimental results show that the score-level fusion of different feature vectors enhances the accuracy of speaker verification system and this reduces the equal error rates is in some cases up to 44%.**

Keywords-speech feature vectors; speaker verification; score fusion; noisy speech; Gaussian mixture model.

I. INTRODUCTION

Different methods have been studied for user verification based on their vital features. Some of the most famous of these features are fingerprints and people's voices. The use of each feature has its own advantages and disadvantages based on the desired application and accuracy, and no unique verification system provides the best results under all circumstances. There are elements that distinguish speech from other features. Speech is a natural signal, and it is not easy for one person to intimidate other's speech sounds. In some situations, the only way to get access to a people is his speech, for example in long distance communication with limited bandwidth via telephone. In addition, speaker verification is preferred by users, and does not require expensive special equipment and sensors [1].

Numerous speaker verification systems have been proposed, and many have obtained satisfactory results in laboratory environments; however, the performance of these systems drop significantly in natural environments where intrusive elements, such as different types of noises, channel distortions, and environmental reflections, are present [2]. Hence, a desirable speaker verification system should be environmental noise resilience. Therefore, the uses of methods that make a system noise resistant are very important in speaker verification applications.

Speaker verification systems consist of several main stages, including: Feature extraction, speaker modeling, model comparison, and decision-making. One method for making robust speaker verification system, is to extract and utilize the features that are less sensitive to noise. The use of features such as MFCC [3], utilizing resistant methods for estimating the spectrum, e.g., weighted linear prediction (WLP) [4], employing more robust features like PNCC [5], and applying post-processing technique, such as feature warping [6] are among some of the approaches that have been proposed to achieve superior performances.

Another idea for improving the resilience of speaker verification systems is to combine various methods, which is a common solution in different verification methods based on various biometric systems. Each method only model part of the information, and their combination can better represent the entire information. Combination can occur between different biometrics, e.g., fingerprints and face. In this case, in addition to increasing the accuracy in decision-making, it reduces the possibility of forgery [7]. Combination system can be applied in the form of a single biometric method, and within its various sections, such as feature extraction, modeling, scoring, and decision making. Different feature vectors fusion – with different strategies such as combining final scores [8, 9], joining different vectors to obtain new vectors [10], and combining information on a decision level [11] – are more common and produce better results, since some features have good complementary information for each other [12].

Although using combination methods in speaker verification is not a new idea, still more efficient algorithms regarding these combinations are presented. Moreover, the implementation of common combination techniques for new speaker verification methods can improve the results of the systems. Researchers have also tried to improve the speaker identification systems performances by using fusion of MFCC with other features as in [13], [14].

This paper investigates the combination of four different speech features and evaluation of their performance in the presence of several types of additive noise in different signal to noise levels that was not reported before. The structure of the article is as follows. First, in Section II, four feature vectors of

speech signal are briefly reviewed. Section III introduces the various strategies for verification systems combination. In Section IV, the experiments are described including, the train and test datasets, structure of the test system, and the performance evaluation measures. It also provides the test results and discussions. Section V concludes the paper.

II. SELECTED FEATURES

Four common spectral speech feature vectors, i.e., mel-frequency cepstral coefficients (MFCC), inverted mel-frequency cepstral coefficients (IMFCC), linear frequency cepstral coefficients (LFCC), and power-normalized cepstral coefficients (PNCC) have been widely applied in speaker verification applications for clean and noisy speech. The use of MFCC has been motivated by the auditory properties of human ear, which is more sensitive to variations at lower frequencies. Accordingly, a filter bank was designed to put more emphasis on low frequencies.

Undergoing the logarithmic compression and discrete cosine transform (DCT), the output of the filter bank results in the MFCC coefficient. The log-compressed filter outputs will be relatively uncorrelated using the DCT [3]. Assuming that the outputs of an M-channel filter bank is $Y(m), m = 1, \dots, M$, the MFCC coefficients, c_n, are given by

$$c_n = \sum_{m=1}^{M} \left[log\left(Y(m)\right) \right] \cos\left[\frac{\pi n}{M} \left(m - \frac{1}{2} \right) \right] \qquad (1)$$

The extraction of coefficients is almost similar for IMFCC and MFCC coefficients; the only exception is that the former one put more emphasis on high signal frequencies. Hence, the only is in the specification of the applied filter bank. IMFCC put more concern on higher frequencies information, which are less important to MFCC. So improves distinguishability of speaker specific characteristics available in the higher frequency zone [15]. In a similar way, LFCC and MFCC are largely identical in terms of coefficient extraction scheme. The only difference is in the filter bank, since the LFCC filter bank coefficients equally cover all speech frequency ranges and consider them of equal importance. LFCC consistently outperforms MFCC in the female trials and there is some advantage of LFCC over MFCC in reverberant speech [16]. Fig. 1 illustrates the similarities and differences of the structures used in the extraction of four feature vectors.

Similar to MFCC, the feature vector extraction in PNCC are based on human auditory processing and making an utmost effort to simulate that. The traditional logarithmic nonlinearity used in MFCC extraction is replaced by a power-law nonlinearity. According to psychoacoustic observations, the power value is set to 1/15. Other main characteristic of PNCC is noise removal asymmetric filter algorithm including temporal masking effects. The PNCC coefficients are extracted by a 30-channel gammatone filter bank within 100 to 4000. The bandwidths of filters were designed to put more emphasis on lower frequencies similar to that of MFCC. In order to enhance the system's performance, the area under the curve are normalized to one for each channel [5]. Fig. 2 shows the overall structure of feature extraction for the features.

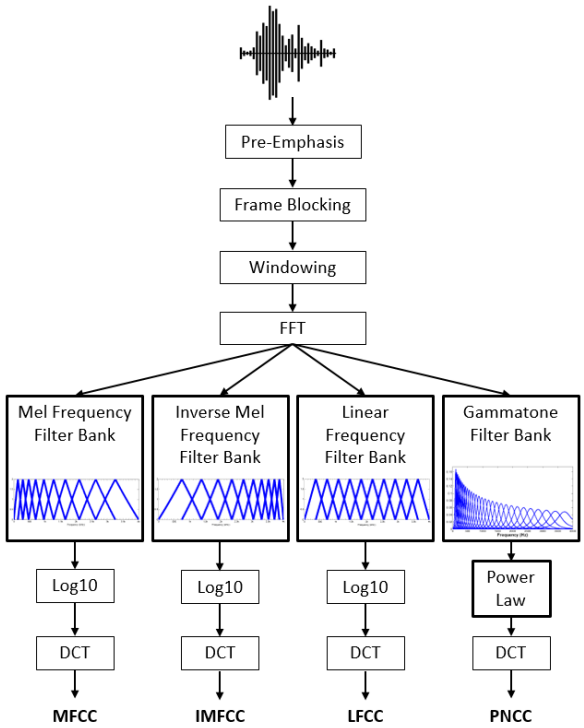

Figure 1. Comparison of four feature extraction methods MFCC, IMFCC, LFCC, and PNCC in terms of structure.

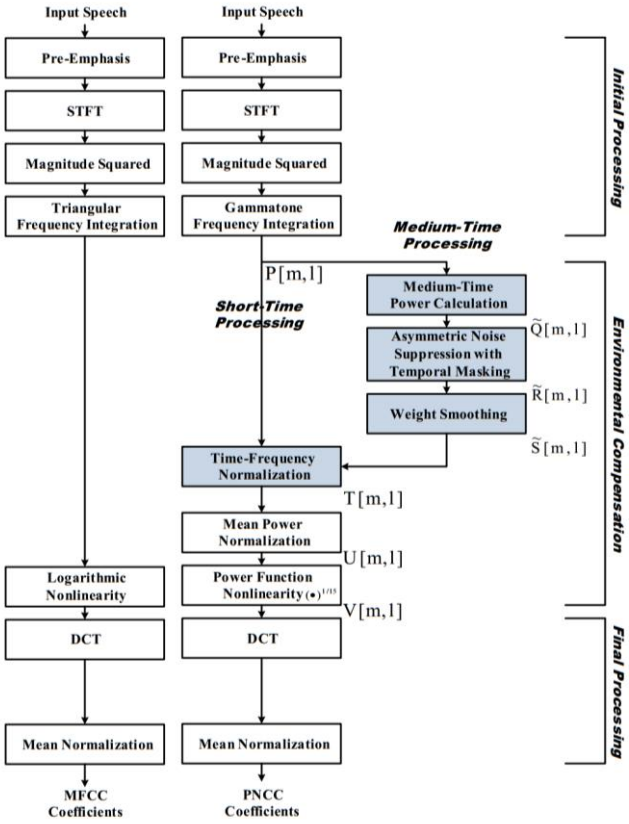

Figure 2. Comparison of two feature extraction methods MFCC and PNCC in terms of structure [5].

III. FUSION TECHNIQES

A. Features Level Fusion

Feature vectors Combination in a speaker verification system may be implemented in different stages, i.e., formation of feature vector, assigning scores, and decision-making; and various strategies are proposed for each approach. Fig. 3 shows three different combination patterns based on feature vectors. Combining feature vectors usually improve the system performance. However, conditions such as the independency of these features increases this improvement, since in such cases the entire model is better represented.

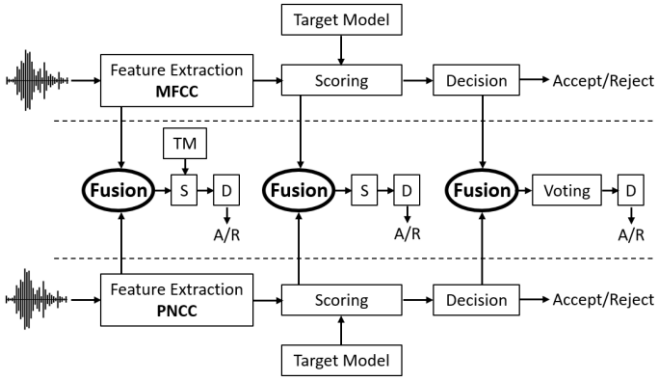

Figure 3. Features fusion in various level of speaker verification system.

Although in most cases combination improves the results, the rate of improvement is not always the same. For example, in cases where both classification methods decline a test, we cannot expect their combination confirms the test [8].

IV. EXPERIMENTS

To compare the performance of the aforementioned feature vectors, a series of experiments were conducted as follows.

A. Speech Database

TIMIT speech dataset was used in this study. It consists of speech samples from 630 speakers, i.e., 438 males and 192 females, using eight dialects of English. There are 10 short sentences for each speaker that has been uttered under clean conditions. Each utterance is about 3 seconds and is phonetically diverse [17]. Only, the speech samples of male speakers were used in this study. The entire speech samples of 368 male speakers were used to develop the UBM model, while for the other male speakers 9 sentences used for speakers' model adaptation and the last utterances were used in the test stage. In overall, the total number of speaker verification tests was 4900 which means testing the speech utterance of a male speaker against all male speakers not used in the UBM training.

The system's performance under noisy speech input data was evaluated by adding the NOISEX92 noise data to the TIMIT clean data at signal-to-noise ratios of 0, 5 and 10 dB 18]. The types of noises applied in this study for the system's performance evaluation were white, babble, car, and factory noises.

B. Experimental Setup

The speech signals were segmented through a 30-ms sliding hamming window with15-ms overlap. A first-order high pass pre-emphasis filter with $\alpha = 0.97$ is applied. . Twenty six mel channels were used in the filterbank. The sizes of feature vectors were set to 36. It comprises of 12 main coefficients along with their first and second derivatives forming a 36-D feature vector. The zero cepstral coefficient was excluded.

Silence removal was performed by low energy frames elimination from the speech data through statistical method [19]. The effects of any potential mismatching was reduced by the feature warping through a 3-second window applied on the feature vectors, whose distribution was transformed into a standardized normal distribution, i.e., $N(0,1)$ [6].

Speaker modeling done with GMM-UBM approach. First a total of 256 Gaussian mixture were trained with EM algorithm to create the universal background model (UBM). Then the speaker GMM model for each speaker was adapted by the MAP based on UBM and its training speech samples. As shown in Fig. 4, UBM is a speaker independent distribution of feature vectors that was adapted with a new speaker speech sample distribution in enrolment stage. This approach leads to reducing the system's computational load than the model parameters are prepared dependently [3].

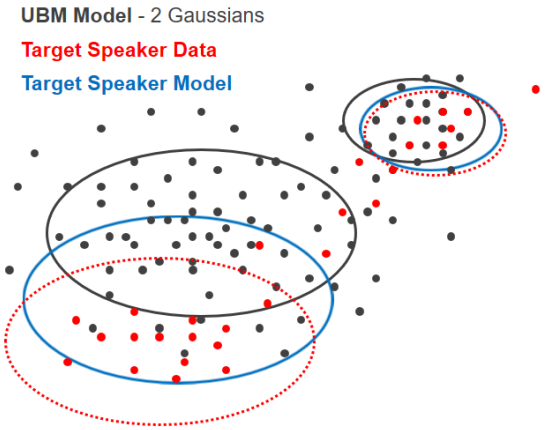

Figure 4. Example of GMM-UBM modeling, create GMM speaker model from UBM [3]

Provisional experiments carried out with different numbers of Gaussian mixtures indicated that 256 was the optimal order option for the aforementioned speech dataset and also better results can be obtained by training two separate models for men and women [20].

In the fusion stage, after the experiments was done for four feature vectors alone and before decision phase, the scores of each feature vector were combined with each other and was made a new score. Final score is a mean of scores that is obtained from two system based on two different feature vector.

.

TABLE I. COMPARISON OF FOUR FEATURE VECTORS PERFORMANCE AND THEIR FUSIONS BASED ON EERS

Feature(s)	Clean	White Noise SNR [dB]			Babble Noise SNR [dB]			Car Noise SNR [dB])			Factory Noise SNR [dB]		
		10	5	0	10	5	0	10	5	0	10	5	0
MFCC	5.71	10.00	14.29	16.96	5.71	9.42	11.43	4.29	4.29	3.56	8.36	10.62	14.16
IMFCC	2.86	8.57	11.43	14.29	5.61	5.71	6.29	3.31	2.22	3.71	3.46	5.71	8.57
LFCC	5.13	8.57	14.29	16.92	5.57	7.16	8.57	4.29	3.71	4.29	5.88	8.55	11.43
PNCC	5.71	7.14	10.00	15.07	4.45	5.16	13.81	5.16	3.83	4.29	7.14	7.43	12.86
MFCC + IMFCC	2.90	7.14	9.69	12.51	4.29	4.62	6.46	2.86	2.24	**2.71**	**3.08**	5.71	8.57
MFCC + LFCC	4.29	7.14	11.49	14.29	4.64	5.88	8.53	2.86	1.80	2.86	5.45	8.57	11.10
MFCC + PNCC	4.29	7.62	10.35	12.86	4.16	4.93	9.38	4.29	3.13	2.86	5.61	7.14	11.76
IMFCC + LFCC	3.06	7.14	10.33	17.14	4.29	4.29	5.71	2.86	**1.72**	3.11	3.91	5.71	7.87
IMFCC + PNCC	**2.65**	7.14	**5.76**	13.27	2.86	**4.29**	**5.71**	2.22	2.57	2.84	3.23	**3.98**	**5.71**
LFCC + PNCC	2.86	7.14	8.34	14.10	**2.80**	4.70	7.23	2.86	2.13	2.96	4.29	5.71	11.01
MF+IMF+LF+PN CCs	2.94	**5.71**	7.14	**11.49**	2.86	4.29	5.84	**2.15**	2.11	2.86	3.27	4.29	8.41

C. Evaluation Criteria

The evaluation of speaker verification system was done based on the error rate in the tests and the detection error Tradeoff (DET) curves that was illustrating the system's performance at different threshold values. The DET is a tradeoff of false alarm (FA) and false rejection (FR) errors. Furthermore, the detection cost function (DCF) was also used as defined by NIST:

$$DCF = C_{miss}\, E_{miss} P_{target} + C_{fa}\, E_{fa}\left(1 - P_{target}\right) \quad (3)$$

where E_{miss} and E_{fa} represent miss alarm (rejection of target speaker) and false alarm errors (acceptance of imposters), respectively. P_{target} is the prior probability of real speakers, C_{miss} is the cost of missing false, and C_{fa} is the cost of false accepting. The proposed NIST values for the three parameters.

are 0.01, 10, and 1, respectively [21]. The optimal point is where the value of the DCF is minimized. Given the values of fixed parameters, it can be argued that the optimal point inclines toward lower false accept error rate. In addition to the above, the equal error rate (EER) was calculated to provide a more comprehensive evaluation of the system's performance.

D. Results and Discussion

The speaker verification system was implemented based on Voicebox® [22] and MSR Identity Toolbox [23]. The performances of MFCC, IMFCC, LFCC, and PNCC and fusion of their scores were compared through GMM-UBM modeling. The tests were carried out on speech signals under clean conditions and contaminated with white, babble, car, and factory noises at signal-to-noise levels of 0, 5 and 10 dB.

Table I shows the performances of four feature vectors, their pairwise fusions, and the foursome fusion in the terms of equal error rate (EER). As can be seen in the results, accuracy of the system increases in both clean and noisy conditions. For each column (test condition), the best feature vector and the best fusion mode was specified in terms of minimum error rate. In case of same EER, the best method was specified based on minDCF. It is noteworthy that the length of all feature vectors are equal to 36, and all the fusions are implemented in the score level and not in the feature level.

As expected in most cases the fusion of the features provides better results than the individual one. The best performance in clean conditions achieved using the fusion of IMFCC and PNCC. The fusion of these two features also provides the best performance in more than 50% of the noisy conditions. It is noteworthy that the best performance not necessarily results from the fusion of those features which provide the best outcome individually.

The score-level fusion of foursome feature vectors with equal weight did not reduce the error rate in most cases compared to pairwise features combination. Even for the cases that the former provides better result the improvement is marginal.

Figs. 5 and 6 show comparatively the DET curves for the speaker verification system using PNCC, IMFCC, PNCC+IMFCC, and foursome fusion of features for clean and 5 dB white noise contaminated speech, respectively. The figures show the performance of the system with combined features outperforms the one with a single feature vector both in clean and noisy environments.

V. CONCLUSIONS

In this paper, the performance of a speaker verification system based on GMM-UBM modeling with fusion of four common speech feature vectors, i.e., MFCC, IMFCC, LFCC, and PNCC are evaluated and compared for clean and noisy speech signals. Four types of noise, namely, white, babble, car, and factory noises on three signal-to-noise level were used in the experiments. The results indicate that score-level fusion in most cases improve the system performance. The enhancement of IMFCC and PNCC combination outperforms the other.

The future direction of this research is toward the used of more advanced speaker modeling techniques as well as speech and/or feature enhancement methods.

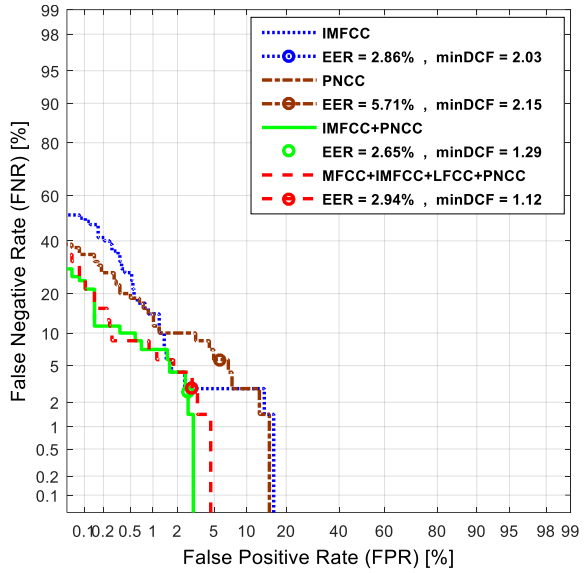

Figure 5. The performances comparisonof the systems based on IMFCC, PNCC, PNCC+IMFCC, and foursome fusion of features for clean speech.

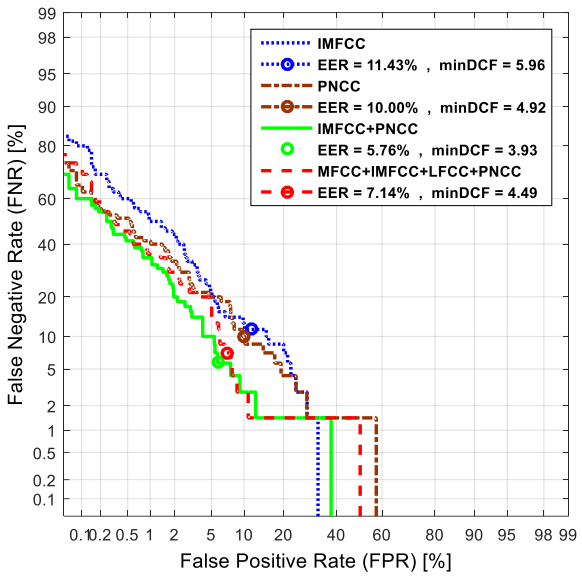

Figure 6. The performances comparison of the systems based on IMFCC, PNCC, PNCC+IMFCC, and foursome fusion of features in the presence of white noise at signal-to-noise ratio of 5 dB.

REFERENCES

[1] R. de Luis-García, C. Alberola-López, O. Aghzout, and J. Ruiz-Alzola, "Biometric identification systems," Signal Processing, vol. 83, pp. 2539-2557, 2003.

[2] J. H. L. Hansen and T. Hasan, "Speaker Recognition by Machines and Humans: A tutorial review," Signal Processing Magazine, IEEE, vol. 32, pp. 74-99, 2015.

[3] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," Speech Communication, vol. 52, pp. 12-40, 1// 2010.

[4] C. Ma, Y. Kamp, and L. F. Willems, "Robust signal selection for linear prediction analysis of voiced speech," Speech Communication, vol. 12, pp. 69-81, 1993.

[5] C. Kim and R. M. Stern, "Power-Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, pp. 1315-1329, 2016.

[6] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," 2001.

[7] A. Ross and A. Jain, "Information fusion in biometrics," Pattern recognition letters, vol. 24, pp. 2115-2125, 2003.

[8] T. Kinnunen, V. Hautamäki, and P. Fränti, "Fusion of spectral feature sets for accurate speaker identification," in 9th Conference Speech and Computer, 2004.

[9] F. Răstoceanu and M. Lazăr, "Score fusion methods for text-independent speaker verification applications," in 2011 6th Conference on Speech Technology and Human-Computer Dialogue (SpeD), 2011, pp. 1-6.

[10] M. Sarria-Paja, M. Senoussaoui, D. O. Shaughnessy, and T. H. Falk, "Feature mapping, score-, and feature-level fusion for improved normal and whispered speech speaker verification," in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 5480-5484.

[11] C. C. Lip and D. A. Ramli, "Comparative Study on Feature, Score and Decision Level Fusion Schemes for Robust Multibiometric Systems," in Frontiers in Computer Education, S. Sambath and E. Zhu, Eds., ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 941-948.

[12] M. Faundez-Zanuy, "Data fusion in biometrics," IEEE Aerospace and Electronic Systems Magazine, vol. 20, pp. 34-38, 2005.

[13] R. S. S. Kumari, S. S. Nidhyananthan, and A. G, "Fused Mel Feature sets based Text-Independent Speaker Identification using Gaussian Mixture Model," Procedia Engineering, vol. 30, pp. 319-326, 2012/01/01 2012.

[14] M. T. S. Al-Kaltakchi, W. L. Woo, S. S. Dlay and J. A. Chambers, "Study of fusion strategies and exploiting the combination of MFCC and PNCC features for robust biometric speaker identification," 2016 4th International Conference on Biometrics and Forensics (IWBF), Limassol, 2016, pp. 1-6.

[15] S. Chakroborty, A. Roy, S. Majumdar, and G. Saha, "Capturing Complementary Information via Reversed Filter Bank and Parallel Implementation with MFCC for Improved Text-Independent Speaker Identification," in Computing: Theory and Applications, 2007. ICCTA '07. International Conference on, 2007, pp. 463-467.

[16] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in Sixteenth Annual Conference of the International Speech Communication Association, 2015.

[17] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," Speech Communication, vol. 9, pp. 351-356, 1990.

[18] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," Speech Communication, vol. 12, pp. 247-251, 7// 1993.

[19] S. Jongseo, K. Nam Soo, and S. Wonyong, "A statistical model-based voice activity detection," Signal Processing Letters, IEEE, vol. 6, pp. 1-3, 1999.

[20] H. Sadegh Mohammadi and R. Saeidi, "Efficient implementation of GMM based speaker verification using sorted Gaussian mixture model," in Signal Processing Conference, 2006 14th European, 2006, pp. 1-4.

[21] "The NIST Year 2008 Speaker Recognition Evaluation Plan," [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/sre/2008/, 2008.
[22] M. Brookes, "Voicebox: Speech processing toolbox for MATLAB," Software, available [Mar. 2011] from www. ee. ic. ac. uk/hp/staff/dmb/voicebox/voicebox. html, vol. 47, 1997.
[23] S. O. Sadjadi, M. Slaney, and L. Heck, "MSR Identity Toolbox v1. 0: A MATLAB Toolbox for Speaker Recognition Research," Speech and Language Processing Technical Committee Newsletter, 2013.