BULETINUL INSTITUTULUI POLITEHNIC DIN IAŞI
Publicat de
Universitatea Tehnică „Gheorghe Asachi” din Iaşi
Volumul 65 (69), Numărul 4, 2019
Secția
ELECTROTEHNICĂ. ENERGETICĂ. ELECTRONICĂ

FEATURE FUSION BASED ON AUDITORY AND SPEECH SYSTEMS FOR AN IMPROVED VOICE BIOMETRICS SYSTEM USING ARTIFICIAL NEURAL NETWORK

BY

YOUSSOOUF ISMAIL CHERIFI[1,*] and ABDELHAKIM DAHIMENE[1,2]

[1]University M'hamed Bougara, Boumerdes, Algeria
[2]Signal and System Laboratory, Boumerdes, Algeria

Received: December 20, 2019
Accepted for publication: April 28, 2020

Abstract. In today's world, identifying a speaker has become an essential task. Especially for systems that rely on voice commands or speech in general to operate. These systems use speaker-specific features to identify the individual, features such as Mel Frequency Cepstral Coefficients, Linear Predictive Coding, or Perceptual Linear Predictive. Although these features provide different representations of speech, they can all be considered as either auditory system based (type 1) or speech system based (type 2).

In this work, a method of improving existing voice biometrics system is presented. Fusing a type 1 feature with a type 2 feature is evaluated and an artificial neural network is trained and tested on in-campus recorded data set. The results confirm the ability for such an approach to be utilized for improving voice biometrics system, regardless of the underlying task being speaker identification or verification.

Keywords: speech processing; neural network; pattern recognition; speaker recognition; feature extraction.

1. Introduction

It is a well-known fact that speech itself contains several levels of information conveyed to the auditor. Mainly, speech is used to communicate a

[*]Corresponding author: *e-mail*: youcef.ismail.cherifi@univ-boumerdes.dz

message to the auditor, a message that is carried in words. However, speech also contains information about the speaker himself, this is due to the way it is produced. This information could be for instance the speaker's gender, emotions, age, origin and obviously identity. Recognizing a person's identity by analysis of a portion of his speech is known as Speaker Recognition (Reynolds, 2002).

Depending on the number of identities involved, speaker recognition can be one of two tasks: speaker verification or speaker identification. In speaker verification, the objective is to verify that the speaker is the exact person he claims to be. On the other hand, speaker identification goal is to assert the identity of the speakers from a closed set of known speakers. In addition to the number of speakers involved, speaker recognition tasks can be either text-dependent or text-independent based on the level of control and cooperation of the user.

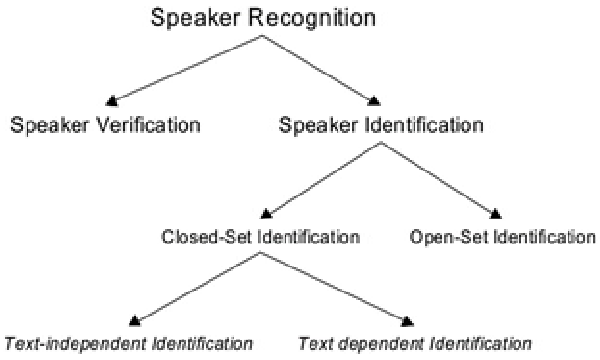

Fig. 1 − Different tasks of speaker recognition.

The development of speaker recognition methods and techniques has been an active field of research for well over five decades and it continues to be. These methods have extended from using spectrogram comparisons to basic template matching, to dynamic time-warping algorithms, to more advanced statistical pattern recognition algorithms such as Artificial Neural Networks (ANNs) and Hidden Markov Models (HMMs). In fact, most of the research that has been done in order to improve the accuracy of a speaker recognition tasks focuses on the development of high-performance pattern recognition algorithms or the adjustment of existing algorithms. An example of this would be the use of Gaussian Mixture Model (GMM) in order to obtain a better recognition rate (Chakroum *et al.*, 2016) or Artificial Neural Network (ANN) (Srinivas *et al.*, 2014). Continuous research and effort are ongoing, involving the combination of two modelling techniques (Al-Shayea & Al-Ani, 2016; Chakroborty & Saha, 2009; Singh *et al.*, 2016; Awais *et al.*, 2014) or the implementation of specific hardware (Gaafar *et al.*, 2014).

Having a better recognition rate does not depend only on the modelling techniques, as was proven by Paliwal *et al.*, 2010, who obtained a better rate by

adjusting the width of the window at which the feature was extracted. This makes sense, given that it is the extracted features that are used as input for the modelling technique. A similar conclusion was deducted by Gulzar *et al.*, 2014 and Dave, 2013, who instead of adjusting the frame width, studied the effect of changing the set of features entirely. In their work, Eringis and Tamulevicius, 2014 combined both changes and proved that by adjusting the frame width and increment for different features, an improvement of 4.15% (from 88.75% to 92.9%) can be achieved.

Based on the aforementioned work, an improved speaker recognition scheme is suggested, in this paper. This scheme has a high accuracy that makes it well suited for applications in various fields such as security and crime investigations

2. Proposed Approach

The adopted approach for improving the recognition rate is to fuse two different types of features extracted with proper parameters using an advanced pattern recognition algorithm. The two sets of features that will be fused are the Mel Frequency Cepstral Coefficients (MFCC) and the Linear Predictive Coefficients (LPC), since this combination provides the best results in comparison with different combinations that we tried.

In order to model the voice print, an Artificial Neural Networks (ANN) algorithm is selected, for its superiority over conventional algorithms in terms of pattern recognition and flexibility in handling inputs. This implies that the fused features are kept isolated and are not grouped under the same vector space, due to the fact that ANNs can have more than one input layer.

As shown in Fig. 2, the proposed scheme consists of a feed-forward neural network that is trained to model voice prints based on two inputs: a type 1 feature and a type 2 feature. Each input vector is connected to its own input layer. The two input layers are then merged/fused at the hidden layer. This latter layer is connected to the output layer.

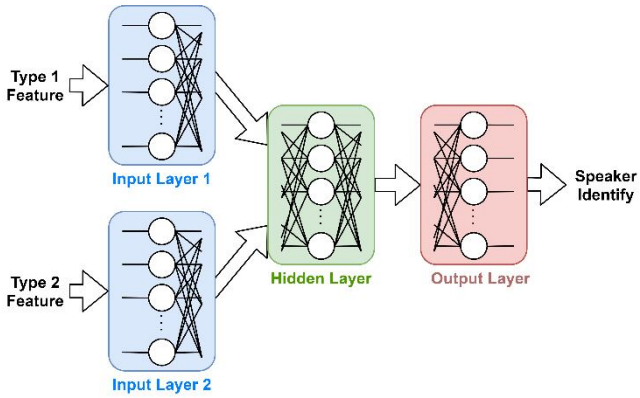

Fig. 2 − The proposed system for feature fusion.

In order to recognize the identity of the person speaking, two main steps are always involved. First is enrolment and second identification. In the first step a speaker database is built based on patterns or models that are deducted from different speech segments. This database is then used in the second step to identify an unknown speaker based on the comparison of results against the existing models within the database.

3. Data Collection

To carry out this work, a data set of speech segments was collected at the Institute of Electrical and Electronic Engineering (IGEE). These speech segments contain 51 to 54 English sentences that are read by 16 IGEE PhD students (8 males, 8 females; mean age 25 years). All speakers were non-native English speakers from Algeria.

The recordings were carried out in a room of the IGEE building, (6.0m(L) × 3.5m (W) × 4m(H) as shown in Fig. 3). The speaker (S) was sitting on a chair, facing a wall at a distance of 0.75m with a monitor in front of him displaying the sentence to be read and a recording device (Honeywell CN51 Personal Digital Assistant (PDA)) was placed between the monitor and the speaker.

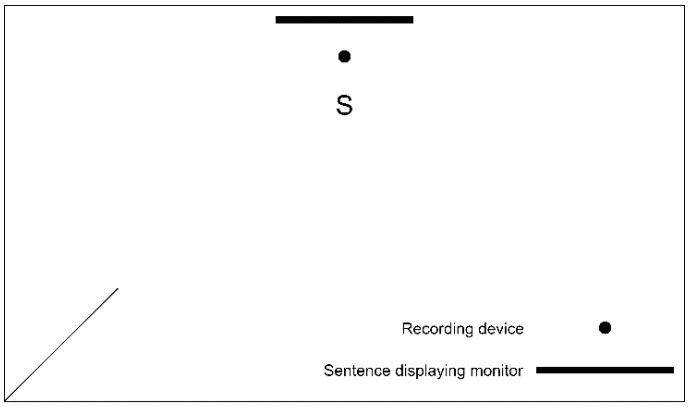

Fig. 3 − Recording room layout.

The recorded sentences differ from one student to the other to ensure that the task remains text-independent and the collaboration of the speaker to a minimum.

4. Speaker Specific Feature Extraction

Although more than two speaker-specific features were used in this work, only the Mel Frequency Cepstral Coefficients (MFCC) and the Linear Predictive Coding (LPC) are discussed in this section. This is because this particular combination of fused features gave the best results.

Mel frequency cepstral coefficient. The MFCC are a set among the most dominant and common type 1 features extracted, either for pattern recognition or media compression. To extract MFCC from an audio signal, the signal is divided into frames of a short duration since the vocal tract changes slowly and during a frame of 20 to 30 ms it is assumed to be "stationary". In addition to being short in time, the frames are also overlapped to ensure a smooth transition between the frames. The resulting segmented signals are then multiplied by a Hamming window to eliminate uncertainty in the amplitude of the frequency spectrum that may occur due to the side lobes of the window (Messikh & Bedda, 2011).

The mathematical expression of a Hamming window of length N is as follows:

$$W(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1 / =0 \text{ otherwise} \quad (1)$$

After framing the signal with short time Hamming windows, the frequency domain components of the signal are then extracted using an FFT. A logarithmic Mel-Scaled filter bank is applied to the extracted frequency components. This scale is almost linear for frequencies up to 1 kHz, and logarithmic for higher frequencies. The conversion from Herts (Hz) to the Mel scale is the following:

$$f(mel) = 2595\frac{\log(1+f(Hz))}{700} \quad (2)$$

The Mel-scale filter banks that are used have a greater bandwidth for high frequencies and a smaller bandwidth for low frequencies. However, their temporal resolutions are equal.

The final step in extracting the MFCC is to compute the Discrete Cosine Transformation (DCT). The DCT will sort the resulting coefficients according to their significance, where the 0th coefficient is neglected since it carries no reliable information (Lahouti et al, 2006). The overall procedure of MFCC extraction is shown in Fig 4.

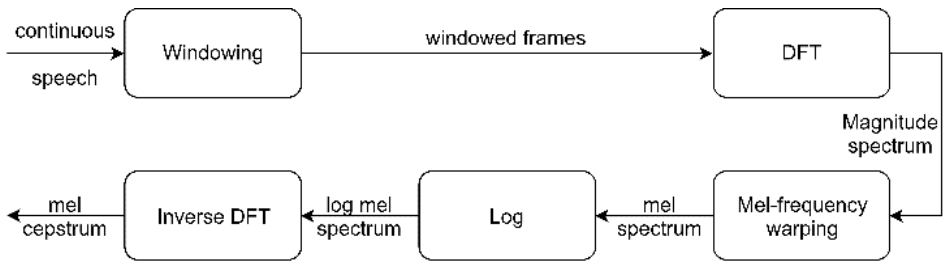

Fig. 4 − MFCC derivation.

Linear Predictive Coding. LPC is one of the most powerful and useful type 2 feature set for speech analysis (Buza *et al*., 2006). LPC extracts the power spectrum of the speech signal in formant analysis (Yuhas et al, 1990). LPC is a good candidate for speech analysis due to the nature of the speech generation process. The entire process can be represented by a digital filter as modelled in Fig. 5.

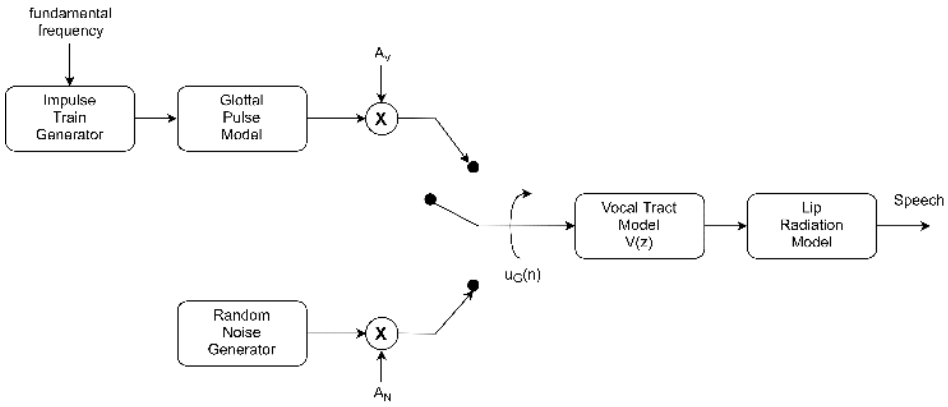

Fig. 5 − Speech Production Model. A_v is the voiced sound gain, A_n is the unvoiced sound gain and $u_G(n)$ is voiced/unvoiced switching function.

To extract LPC from an audio signal, the signal is first framed using short time Hamming windows. For each frame 14 formants or coefficients are extracted. This ensures that all possible speech segments (voiced and unvoiced) are covered and for both genders as well. Formants are extracted by computing coefficients that link the current speech sample with the previous samples with the same window. This can be expressed by the following equation:

$$s(n) = \sum_{k=1}^{p} \alpha_k s(n-k), \tag{3}$$

where: s is the speech signal, n – sample point, α – formant and p – number of the required formants (14 in our case).

5. Modelling Voice Print

To model the voice print of each speaker in the data set we used Artificial Neural Networks. Two different network structures were implemented. The first has one input layer and the second has two. The first structure is used for tuning the parameters of the extracted features while the second is used to fuse features. Both structures are represented in Figs. 6 and 7 respectively.

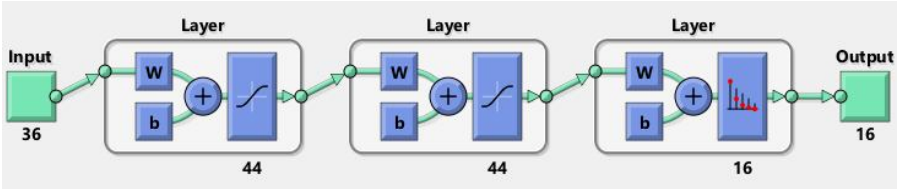

Fig. 6 − The single feature ANN structure.

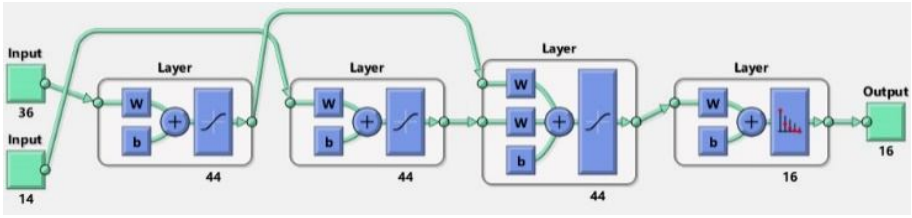

Fig. 7 − The feature fusion ANN structure.

We included 16 perceptrons in the output layer of each structure, equivalent to the number of speakers in our data set. For the remaining layers we used 44 perceptrons, equivalent to the number of phonemes in the English language.

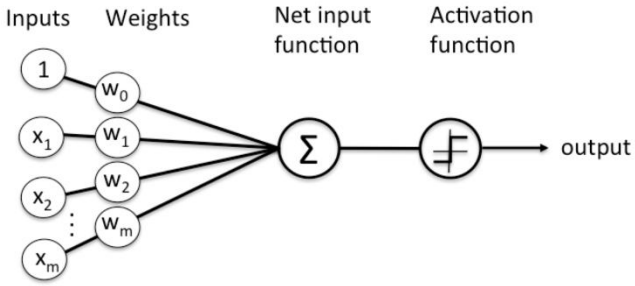

Fig. 8 − Perceptron structure where x_i is an input value and w_i is its associated weight.

In both structures, the perceptrons in the output layer use the SoftMax activation function to reduce the influence of extreme values or outliers in the dataset without having to remove them. For input layers and hidden layers perceptrons, the tangential sigmoid activation function was used. This function has a steeper derivative which makes it a good candidate for extracting intermediary features (Meena *et al.*, 2011).

To train the structures shown above, the conjugate gradient backpropagation algorithm is used to reduce the Sum of Square Errors (SSE) between the outputs of the network and a vector of desired targets. This algorithm has a better accuracy when compared with other algorithms as shown in Table 1.

Table 1
Performance of Different Training Functions in Matlab's NN Toolbox (Vacic, 2015)

Function name	Training		Validation		Testing		Time	
	Mean	Stdev	Mean	Stdev	Mean	Stdev	Mean	Stdev
trainb	0.6456	0.7246	0.6302	0.6946	0.6386	0.7081	2.511	3.3835
trainbfg	0.0096	0.0032	0.0199	0.0084	0.0209	0.0046	7.3219	4.5702
trainbr	7.6088	3.5328	18.9761	10.219	149.8294	32.2893	18.5063	8.927
trainc	0.0072	0.0015	*	*	0.0374	0.0066	466.072	163.5241
traincgb	0.0102	0.0026	0.0193	0.0069	0.0203	0.0059	4.3389	1.886
traincgf	0.0112	0.0033	0.0199	0.0091	0.0202	0.0051	4.9752	2.4127
traincgp	0.0114	0.003	0.0213	0.0093	0.0216	0.0045	4.0544	1.9337
traingd	0.0265	0.0055	0.0332	0.0099	0.0323	0.0029	13.003	4.4432
traingdm	0.5528	0.34	0.5556	0.3221	0.5592	0.3499	1.2875	0.3697
traingda	0.0244	0.0063	0.0293	0.0084	0.0310	0.0037	5.2	2.222
traingdx	0.0394	0.0312	0.0448	0.0317	0.0445	0.0274	5.4219	3.526
trainlm	0.0065	0.0027	0.0199	0.0066	0.0231	0.0037	8.5762	3.494
trainoss	0.013	0.0038	0.0204	0.0081	0.0205	0.0035	5.1703	2.8221
trainr	0.0077	0.0014	*	*	0.3319	0.0042	422.3888	148.2313
trainrp	0.0137	0.0045	0.0207	0.0059	0.0229	0.0035	7.4954	3.8277
trains	2.0723	1.5461	*	*	2.1834	1.6277	0.1893	0.0188
trainscg	0.0114	0.0035	0.0213	0.0109	0.0218	0.0073	4.3171	1.7394

6. Results & Discussion

We have described the proposed approach for improving the accuracy for speaker recognition tasks along with the methods used for extracting speaker-specific features and modeling the speaker's voiceprint. In this section we detail the assessment of the performance of our approach.

Tuning the parameters for feature extraction. The first step for improving recognition rate is to adjust the frame width and increment during feature extraction. This improvement has already been shown (Paliwal *et al.*, 2010. Eringis & Tamulevicius, 2014) given that these parameters determine whether the modeling algorithms are getting a sufficient level of information from the speech segment as input.

There is some discrepancy in the literature regarding the frame width values which optimize accuracy (see Fig. 9). To confirm which frame width to choose, the structure shown in Fig. 6 is trained with MFCC features extracted using different frame widths. Thirty random segments for each speaker in the data set are used for training while the remaining 21 segments are used for testing. The width was incremented by an interval of 5 ms with each trial.

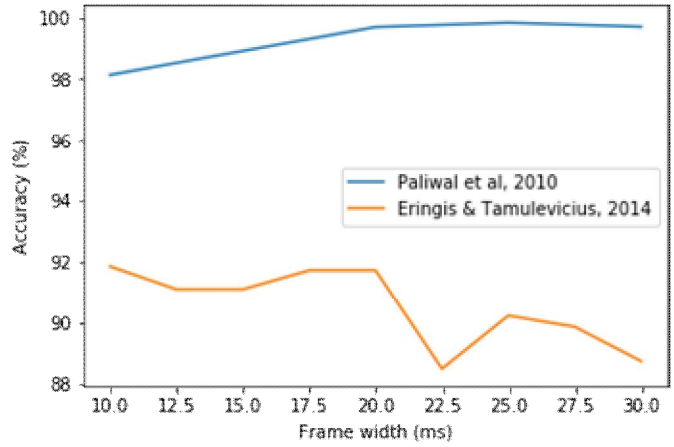

Fig. 9 − The effect of adjusting the frame width during MFCC extraction on the overall accuracy, blue for the work done by Paliwal *et al.*, 2010 and orange for the work done by Eringis and Tamulevicius, 2014.

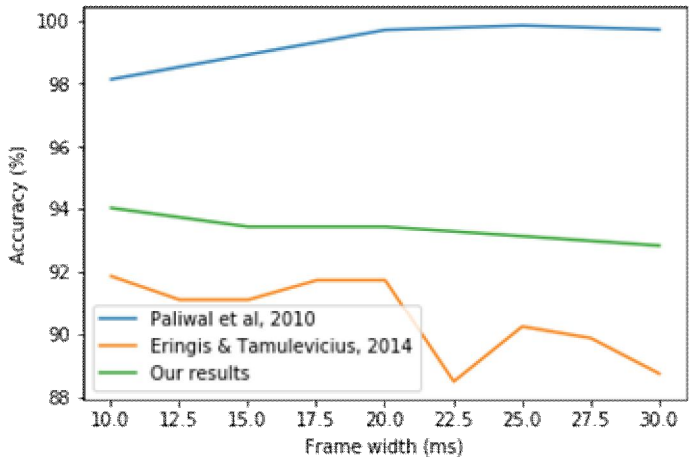

Fig. 10 − The results of frame width tuning tasks.

When extracting features, in addition to frame width, frame increment will also affect the level of information being extracted. When examining the influence of frame width, a fixed frame increment was used at 75% of the frame width. In order to determine whether or not this value provides the best recognition results, we kept the frame width fixed at 10 ms as this value provided the best results while adjusting the frame increment at each trial.

Adjusting the frame width during feature extraction indeed improved the speaker recognition rate. The best recognition rate is obtained for a frame width of 10 ms, which coincides with the results obtained by Eringis and Tamulavicius, 2014 see Fig. 10.

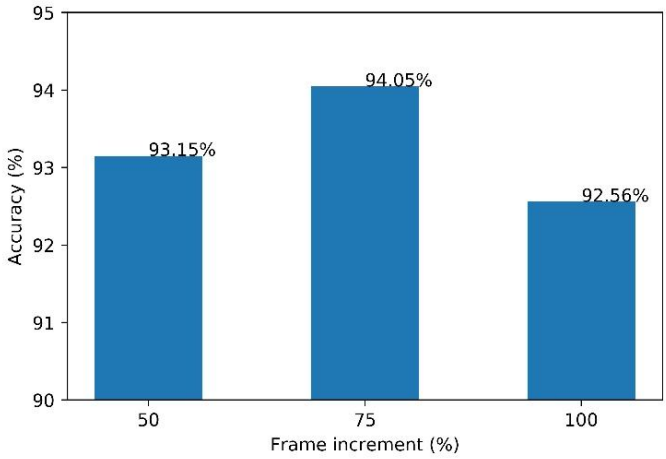

Fig. 11 − The results of frame increment tuning task.

Examining different frame increments for a set frame width (10 ms) we found that a frame increment of 75% provides a higher recognition rate (94.05%) compared to increments of 50% and 100%. This implies that with adjusting only the frame increment 3 more speech segments are correctly recognized.

The two tuning tasks (on frame width and increment) revealed that the best parameters for feature extraction are a frame width of 10 ms with a frame increment of 75%, that is 7.5 ms. Although we discussed the results that are obtained for MFCC only this conclusion holds true also for the other features (LPC and PLP) that were tested, that is why these parameters are used for extracting the different features that are used for feature fusion.

Feature fusion. Using ANNs and adjusting feature extraction parameters can improve the recognition rate. The novelty of this work is the combined use (fusion) of a type 1 feature and a type 2 feature in order to provide more information about the speaker. This will ensure that the suggested modeling algorithm (ANN) is getting information about both how the speech is perceived and how is it produced.

In order to study the effect of fusing features on the overall recognition rate, the structure shown in Fig. 7 is trained using two features extracted at a frame width of 10 ms and a frame increment of 7.5 ms. For each speaker 30 random segments are used for training while the remaining 21 segments are used for testing. The resulting performance is compared with that of training the structure shown in Fig. 6 with a single feature extracted with same parameters from the same speech segments.

The fusion of type 1 and type 2 features improved the accuracy of the recognition tasks. As seen in Fig. 12, any combination of type 1 and type 2 feature would result in a better recognition rate than using just a single feature.

The best result was obtained when combining MFCC and LPC (99.4% accuracy), this is due to the fact that these two sets of features are uncorrelated.

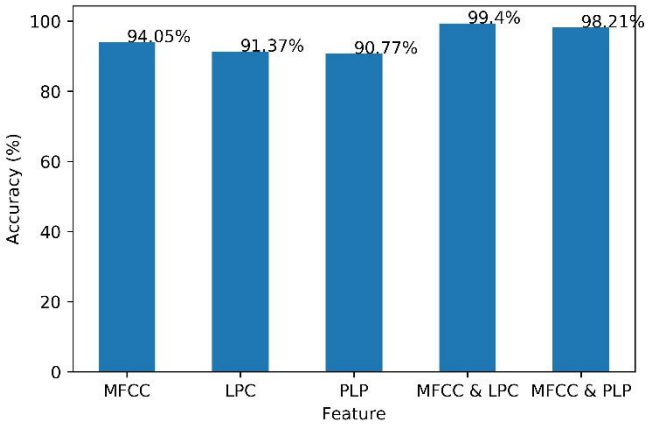

Fig. 12 − Feature fusion effect on the overall accuracy for speaker recognition. MFCC: Mel-frequency cepstral coefficients; LPC: Linear predictive coding; PLP: Perceptual linear predictive.

Although this fusion approach improved the recognition rate, using two sets of features instead of just one, resulted in a slight increase in the training time. Fig. 13 shows the required training time for each of the trials described in section 2.5. These results were obtained using the 4710MG i7 CPU with a RAM of 8GB.

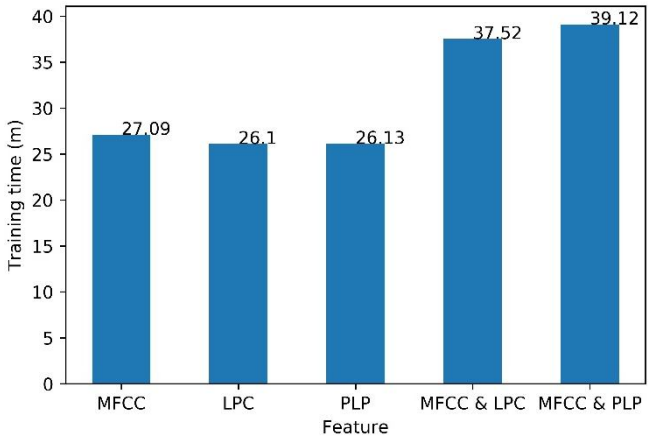

Fig. 13 − Feature fusion effect on the required training time for speaker recognition. MFCC: Mel-frequency cepstral coefficients; LPC: Linear predictive coding; PLP: Perceptual linear predictive.

The difference in training time between the best performing single feature structure and the best performing fused features structure is 10 minutes

and 25 seconds. This means an increase of 38.50% in training time for an improvement of 5.35% in speaker-recognition accuracy if we consider only the effect of feature fusion without the parameter tuning and 6.55% when considering the parameters tuning. This is in fact very significant if we consider and application such as criminal investigation where 6.55% means that 22 cases or suspects are correctly being recognized using for example a phone call.

In addition, the only drawback of the system which the training time can be significantly reduced, if a better hardware is used such as a more performing GPU or/and by reducing the duration of the recordings that are used for training the model as fusing feature allow it to recognize the speaker much earlier as shown in Fig. 14.

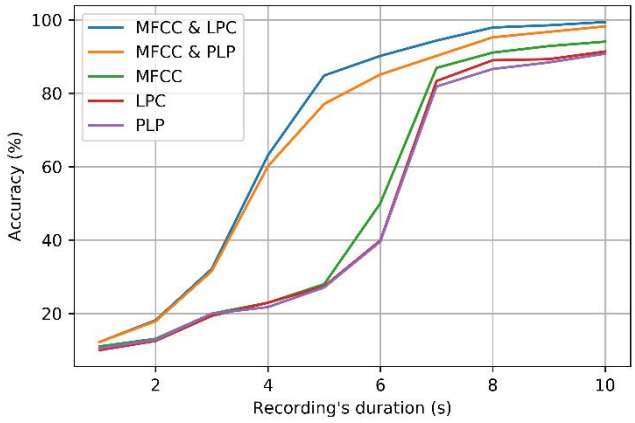

Fig. 14 − Speaker recognition accuracy over time. MFCC: Mel-frequency cepstral coefficients; LPC: Linear predictive coding; PLP: Perceptual linear predictive.

Table 2

Speaker Recognition Accuracy Over Time. MFCC: Mel-frequency Cepstral Coefficients; LPC: Linear Predictive Coding; PLP: Perceptual Linear Predictive

Input	Time (s)									
	1	2	3	4	5	6	7	8	9	10
MFCC	11.01	13.09	19.94	22.91	27.97	50	86.9	91.07	92.85	94.05
LPC	10	12.5	19.34	22.91	27.38	39.88	83.33	88.89	89.28	91.37
PLP	10.5	12.79	19.94	21.72	27.08	39.58	81.84	86.60	88.39	90.77
MFCC & LPC	12.20	18.15	32.14	63.09	84.82	90.17	94.34	97.91	98.57	99.4
MFCC & PLP	12.20	17.85	31.54	60.11	77.08	85.11	90.17	95.23	96.72	98.21

By using 6 seconds of recording the model was able to reach 90.17% when fusing MFCC with LPC. What is even more important about these results is the fact that the fusion approach was able to outperform the single feature approach by utilizing 7 seconds out of the provided 10 seconds of recording, this reduced the training time of the fusion approach from 37 minutes and 32

seconds to 24 min and 41 seconds. This time is less than time required to train the model for any of the features independently. These results are significant when taking into consideration the applications of speaker recognition. For fields such as security, this means less data storage, and for crime investigation this means the ability to identify suspect even if the provided audio is short.

7. Conclusion

The speech signal does not only convey a message, it conveys information about the speaker themselves, their gender, origins, health and age. The aim of this work was to improve the task of recognizing a person based on speech segments.

The approach we used proved to be very effective in improving the speaker recognition rate. This improvement is due to the use of two sets of features instead of just one. These two feature sets are completely uncorrelated and each one represents different characteristics of the speech signal. The drawback of such approach is that it takes a longer time for training the model. Nevertheless, this can be mitigated by using an approach such as deep features, where the input layer and the hidden layers of the trained model are kept. The output layer however is replaced by a less time demanding classification or clustering technique such as Support Vector Machines (SVM) or Gaussian Mixture Models (GMM).

But even as it is the approach improved the recognition rate by 6.55% which means out of the 336 speech segments that were used for testing it recognized 22 more segments. This is was achieved at the cost of a 10 minutes and 26 seconds increment in the training time and no change in the testing time.

REFERENCES

Al-Shayea Q.K., and Al-Ani M.S., *Speaker Identification: A Novel Fusion Samples Approach*, International Journal of Computer Science and Information Security, 14-7, 423-427 (2016).

Awais M., Mansour A., Ghulam M., *Automatic Speaker Recognition Using Multi-Directional Local Features (MDLF)*, Arabian Journal for Science and Engineering, 39-5, 3379-3811 (2014).

Buză O., Toderan G., Nica A., Căruntu A., *Voice Signal Processing For Speech Synthesis*, International Conference on Automation, Quality and Testing, Robotics, 2006, Cluj-Napora, Romania, **1**, 360-364.

Chakroborty S., Saha G., *Improved Text-Independent Speaker Identification using Fused MFCC & IMFCC Feature Sets based on Gaussian Filter*, International Journal of Electrical, Computer, Energetic, Electronic and Communication Engineering, 3-11, 1974-1982 (2009).

Chakroum R., Zouari L.B., Frikha M., Ben Hamida A., *Improving Text-independent Speaker Recognition with GMM*, International Conference on Advanced Technologies for Signal and Image Processing, 2016, Monastir, Tunisia, **2**, 693-696.

Dave N., *Feature Extraction Methods LPC, PLP and MFCC in Speech Recognition*, International Journal for Advance Research in Engineering and Technology, *1-6*, 1-5 (2013).

Eringis D., and Tamulevicius G., *Improving Speech Recognition Rate through Analysis Parameters*, Electrical Control and Communication Engineering, **5**, 61-66 (2014).

Gaafar T.S., Abo Baker H.M., Abdalla M.I., *An Improved Method for Speech/Speaker Recognition*, International Conference on Informatics, Electronics & Vision, 2014, Dhaka, Bangladesh, **3**, 1-5.

Gulzar T., Singh A., Sharma S., *Comparative Analysis of LPCC, MFCC and BFCC for the Recognition of Hindi Words using Artificial Neural Networks*, International Journal of Computer Applications, *101-112*, 22-27 (2014).

Lahouti F., Fazel A.R., Safavi-Naeini A.H., and Khandani A.K., *Single and Double Frame Coding of Speech LPC Parameters Using a Lattice-Based Quantization Scheme*, IEEE Transactions on Audio, Speech, and Language Processing, *14-15*, 1624-1632 (2006).

Meena K., Subramaniam K., Gomathy M., *Gender Classification in Speech Recognition Using Fuzzy Logic and Neural Network*, The International Arab Journal of Information Technology, *10-15*, 477-485 (2011).

Messikh L., Bedda M., *Binary Phoneme Classification Using Fixed and Adaptive Segment-Based Neural Network Approach*, The International Arab Journal of Information Technology, *8-1*, 48-51 (2011).

Paliwal K.K., Lyons J.G., Wojcicke K.K., *Preference for 20-40 ms Window Duration in Speech Analysis*, International Conference on Signal Processing and Communication Systems, 2011, Gold Coast, Australia, **4**, 1-4.

Reynolds D.A., *An Overview of Automatic Speaker Recognition Technology*, International Conference on Acoustics, Speech, and Signal Processing, 2002, Orlando, Florida, USA, **26**, 4072-4075.

Richardson F., Reynolds D.A., Dehak N., *Deep Neural Network Approaches to Speaker and Language Recognition*, IEEE Signal Processing Letters, *22-10*, 1671-1675 (2015).

Singh S., Assaf M.H., Das S.R., Biswas S.N., Petriu E.M., Groza V., *Short Duration Voice Data Speaker Recognition System Using Novel Fuzzy Vector Quantization Algorithm*, International Instrumentation and Measurement Technology Conference Proceedings, Taipei, Taiwan, **33**, 1-6, 2016.

Srinivas V., Santhi C.R., Madhu T., *Neural Network based Classification for Speaker Identification*, International Journal of Signal Processing, Image Processing and Pattern Recognition, **7**, *1*, 109-120 (2014).

Vacic V., *Summary of the Training Functions in Matlab's NN Toolbox*, MSc, University of California, 2015, Riverside, California, USA.

Yuhas B.P., Goldstein M.H., Sejnowski T.J., Jenkins R.E., *Neural Network Models of Sensory Integration for Improved Vowel Recognition*, Proceedings of the IEEE **78-10**, 1658-1668, 1990.

FUZIUNEA TRĂSĂTURILOR BAZATĂ PE SISTEME AUDITIVE ŞI VOCALE PENTRU UN SISTEM BIOMETRIC VOCAL ÎMBUNĂTĂȚIT UTILIZÂND REȚELE NEURONALE ARTIFICIALE

(Rezumat)

În lumea de astăzi, identificarea unui vorbitor a devenit o necesitate fundamentală, în special pentru sistemele care îşi bazează funcționarea pe comenzi

vocale sau pe vorbire în general. Asemenea sisteme utilizează trăsături specifice vorbitorului pentru identificarea unei persoane, precum Coeficienți Spectral Mel Frequency, Codare Predictivă Liniară sau Predictivă Liniară Perceptuală. Deşi aceste trăsături dau reprezentări diferite ale vorbirii, toate pot fi considerate ca fiind ori bazate pe sisteme auditive (tip 1), ori pe sisteme vocale (tip 2).

In lucrarea de față se prezintă o metodă de îmbunătățire a sistemelor biometrice vocale. Se evaluează fuziunea dintre o trăsătură de tip 1 şi una de tip 2, şi este antrenată şi testată o rețea neuronală artificială pe un set de date înregistrate în campus. Rezultatele confirmă posibilitatea ca o asemenea abordare să fie utilizată pentru îmbunătățirea performanțelor sistemului biometric vocal, independent de identificarea şi verificarea vorbitorului.