





Limited-data automatic speaker verification algorithm using band-limited phase-only correlation function

Ángel David PEDROZA RAMÍREZ* , José Ismael DE LA ROSA VARGAS, 
José de Jesús VILLA HERNÁNDEZ , Aldonso BECERRA SÁNCHEZ 
Faculty of Electrical Engineering, Autonomous University of Zacatecas, Zacatecas, Mexico

Received: 17.05.2018

Accepted/Published Online: 08.05.2019

Final Version: 26.07.2019

Abstract: In this paper, a new method to deal with automatic speaker verification based on band-limited phase-only correlation (BLPOC) is proposed. The aim of this study is to validate the use of the BLPOC function as a new limited-data automatic speaker verification technique. Although some speaker verification techniques have high accuracy, efficiency usually depends on the extraction of complex theoretical information from speech signals and the amount of the data for training the algorithms. The BLPOC function is a high-accuracy biometric technique traditionally implemented in human identification by fingerprints (through image-matching). When applying the BLPOC function in automatic speaker verification through the proposed algorithms (under limited-data conditions), a 98.24% true acceptance rate (TAR) and 87.17% true rejection rate (TRR) in a custom database (and 93.75% TAR and 67.05% TRR in the ELSDSR database) were obtained. The proposed algorithm is a theoretically simple method for automatic speaker verification whose main advantage is that it can provide identification under limited-data conditions. In this sense, the BLPOC function could be applicable in other limited-data biometric identifications by sound signals.

Key words: Biometrics, limited data, speaker verification, text-dependent

1. Introduction

Speech communication is an elementary process that involves some organs' coordination, word articulation, and brain processes, by which human beings evolved as a society. Automatic speech recognition is the machine recognition of words spoken by speakers. Automatic speaker identification focuses on the identification of a speaker by classifying a speech signal as being from a specific speaker among speakers in the dataset. Since the system knows the impostors, this is closed-set identification [1–4]. The speaker verification task is to confirm if a speaker is who he/she claims to be (a yes or no decision) by extracting and analyzing some speech parameters from speakers. Since the system does not know the impostors because they can not be defined, this is open-set identification [1].

A speaker verification system requires, among other steps, to extract the relevant information from speech signals and an appropriate training and test scheme. Among others [5, 6], one of the usual features extracted from voice signals are mel frequency cepstral coefficients (MFCCs). MFCC features are based on the human auditory system and are a common front-end for a speaker identification system [7]. On the other hand, to compare the extracted information from speech, statistical models are commonly used [8, 9]. Modeling requires extracting some voice distribution parameters from the speech signals uttered by the speakers in the dataset and

*Correspondence: p.a.d_16@hotmail.com

the use of a training scheme to use this information as a statistical pattern for recognition. Although statistical models have been widely used due to their high performance in speech recognition tasks, the availability of large-scale datasets is frequently assumed. Nevertheless, with limited data or small datasets, some of the high-performance traditional speaker verification methods can provide inferior performance [10].

Recently, new methods have been developed to extract the relevant information from speech signals [11, 12]. The band-limited phase-only correlation (BLPOC) function is a high-accuracy image-matching biometric technique traditionally implemented in human identification by fingerprints [13]. In this sense, with the aim of validating its use as a new limited-data automatic speaker verification technique, this paper proposes to use the BLPOC function as a simple matching algorithm. The paper is organized as follows: Section 2 explains the basic theory of the BLPOC function. Section 3 details the proposed automatic speaker verification algorithm. Section 4 summarizes the algorithm evaluation results and discussions. Finally, Section 5 gives some conclusions.

2. Band-limited phase-only correlation (BLPOC)

The BLPOC function is an image-matching technique traditionally implemented in human identification by fingerprints whose main advantage is that it is a brightness invariance method [14–18]. Since the origin of the BLPOC function is the phase-only correlation (POC) function, it is briefly described next [13, 19].

Let $f(n)$ and $g(n)$ be two N -point 1D sequences (with index range $n = 0, \dots, N - 1$) and let the vectors $F(k)$ and $G(k)$ be the 1D discrete Fourier transform (1D DFT) of each sequence given by:

$$F(k) = \sum_{n=0}^{N-1} f(n)W_N^{kn} = A_F(k) \exp(j\theta_F(k)), \tag{1}$$

$$G(k) = \sum_{n=0}^{N-1} g(n)W_N^{kn} = A_G(k) \exp(j\theta_G(k)), \tag{2}$$

where $W_N = \exp(-j2\pi/N)$ and k is the frequency index defined as $k = 0, \dots, N - 1$. $A_F(k)$ and $A_G(k)$ denote amplitude components and $\theta_F(k)$ and $\theta_G(k)$ are phase components.

The cross-spectrum ($R_{FG}(k)$) between $F(k)$ and $G(k)$ can be calculated as:

$$R_{FG}(k) = F(k)\overline{G(k)} = A_F(k)A_G(k) \exp(j\theta(k)), \tag{3}$$

where $\overline{G(k)}$ denotes the complex conjugate of $G(k)$ and $\theta(k)$ denotes phase difference $\theta_F(k) - \theta_G(k)$.

A normalized version of the cross-spectrum can be calculated by the cross-phase spectrum ($Rn_{FG}(k)$) as:

$$Rn_{FG}(k) = R_{FG}(k)/\|R_{FG}(k)\|, \tag{4}$$

where $\|R_{FG}(k)\|$ is the norm of the cross-spectrum.

Finally, the POC function (r_{fg}) is the 1D inverse discrete Fourier transform (IDFT) of the cross-phase spectrum and is given by the following equation:

$$r_{fg}(n) = \frac{1}{N} \sum_{k=0}^{N-1} Rn_{FG}(k)W_N^{-kn}. \tag{5}$$

The calculated POC function is a measure of similarity between the two sequences whereby, in the case that they are the same sequence, the resulting POC function is Kronecker's delta function ($\delta(n)$). On the other hand, in the case that they are different sequences, the resulting POC is another function different from $\delta(n)$. Since the relevant information in the frequency spectrum of a sequence is usually clustered in a frequency band, some of the irrelevant phase components of the 1D DFT can reduce the effectiveness of the matching algorithm. In this sense, the BLPOC function (\hat{r}_{fg}) proposes to eliminate the irrelevant information in the sequence by analyzing its inherent frequency band to calculate the cross-phase spectrum only in an effective frequency band [19, 20]. In other words, it is intended to eliminate the irrelevant information in the calculation of $Rn_{FG}(k)$ depending on the frequency spectrum of a given input sequence.

Assuming that the zero frequency component of Rn_{FG} is shifted at the middle of the spectrum, the range of the effective region is calculated by $k = F_1, \dots, F_2$, where $0 \leq F_1 \leq F_2$ and $F_1 \leq F_2 \leq N - 1$ (thus, the effective size of the region is $L = (F_2 - F_1) + 1$). Finally, the BLPOC function is defined as:

$$\hat{r}_{fg}(n) = \frac{1}{L} \sum_{k=F_1}^{F_2} Rn_{FG}(k)W_L^{-kn}. \tag{6}$$

3. Limited-data automatic speaker verification algorithm using BLPOC function

The goal of a speaker verification system is, as mentioned earlier, to identify if a speaker is who he/she claims to be. The system can be a text-dependent system in the case of cooperative speakers repeating preestablished phrases or a text-independent system in the case of noncooperative speakers uttering independent phrases [21]. The proposed system is a text-dependent speaker verification algorithm that compares a prestored speech signal from an authorized speaker with another input speech signal of the same word being uttered and identifies whether the input speech signal is from the same speaker (in the case of genuine matching) or not (in the case of impostor matching). Traditionally, the magnitude of the max peak in the BLPOC function is used as a measure of similarity in human identification by image matching; for speaker matching, some other features were proposed.

The main steps of the proposed algorithm are: 1) speech region identification, 2) common region adequacy, 3) speaker matching, and 4) speaker verification decision.

3.1. Speech region identification

From a traditional voice activity detection (VAD) algorithm [22], a new speech region identification method is proposed based on an energy threshold. After analog to digital conversion of a speech signal, the first step is to analyze and identify which segments of a speech signal correspond to speech information (speech region) and which correspond to noise or silence (see Figure 1). Let $f(n)$ be a speech signal. The speech region (SF) is extracted as follows: i) calculate the average energy of the whole speech signal as $E = \frac{1}{N} \sum_{n=0}^{N-1} |f(n)|^2$, ii) divide the signal $f(n)$ into M frames ($fr_j(i)$) of m samples according to sampling frequency to ensure stationary signal features, iii) calculate the average energy of each j th frame as $e(j) = \frac{1}{m} \sum_{i=1}^m |fr_j(i)|^2$, iv) determine a suitable whole energy threshold value (thr_E), and v) identify the speech frames by the following equations:

$$SF_s = first(\{fr_j(i) | e(j) \geq thr_E, 1 \leq j \leq M, 1 \leq i \leq m\}), \tag{7}$$

$$SF_f = \text{last}(\{fr_j(i) | e(j) \geq thr_E, 1 \leq j \leq M, 1 \leq i \leq m\}), \quad (8)$$

where SF_s and SF_f are the first and last speech frames that contain the speech information. Finally, the speech region is:

$$SF = [SF_s, \dots, SF_f]. \quad (9)$$

In other words, once the speech signal is framed, by using Eqs. (7), (8), and (9) to extract the speech region (SF) and remove the silence/noise segments from the signal, it is possible to automatically identify which frames ($fr_j(i)$) in the signal have the quantity of energy ($e(j)$) above an energy threshold value (thr_E). The optimal whole energy threshold value in the experiments was $thr_E = 0.001 * E$. However, in some cases, to distinguish between the end of the speech signal and a middle pause due to phonetic pronunciation, an analysis of consecutive speech energy frames and calculation of a frame acceptance criterion based on sampling rate is needed.

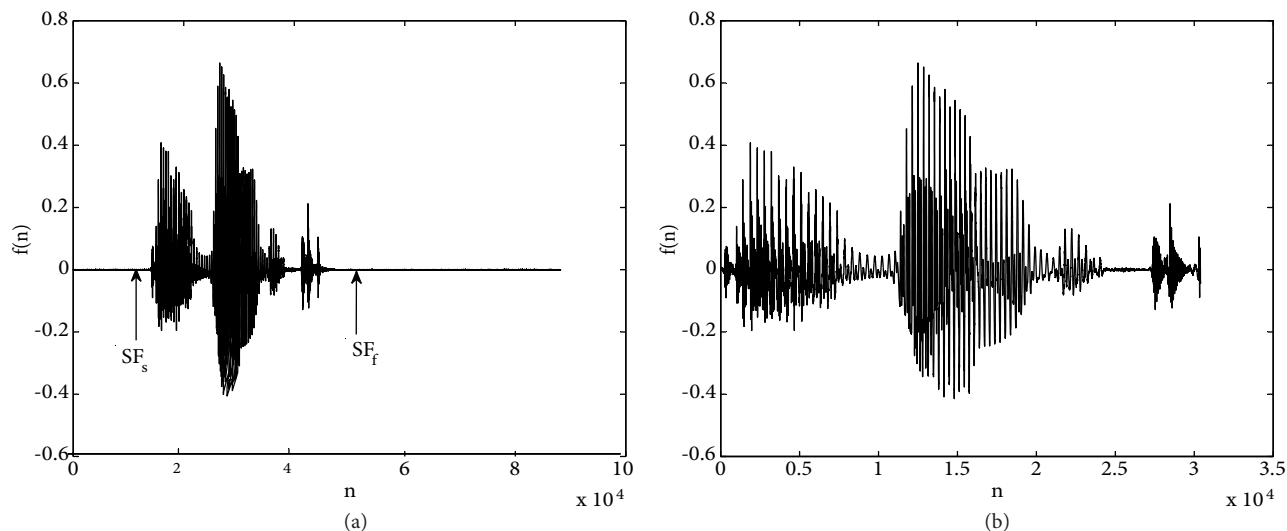


Figure 1. Speech region automatic identification: a) speech signal before and b) after speech region extraction.

3.2. Common region adequacy

Due to time domain speech variability introduced by the speaker (variability of speech speed), the speech signal requires time domain adequacy. This process is as follows: i) compare the number of samples between the prestored and the input speech signals and ii) adjust time domain variability by zero padding at the end of the shortest speech signal according to the number of samples of the largest (see Figure 2).

3.3. Speaker matching

The next step is to calculate the effective frequency band in the cross-phase spectrum as follows (see Figure 3): i) using Eqs. (1) to (4), compute the cross-phase spectrum ($Rn_{FG}(k)$) between the prestored and the input speech signals, and ii) calculate the normal power spectral density ($NPSD$) of vector $Rn_{FG}(k)$ by the following

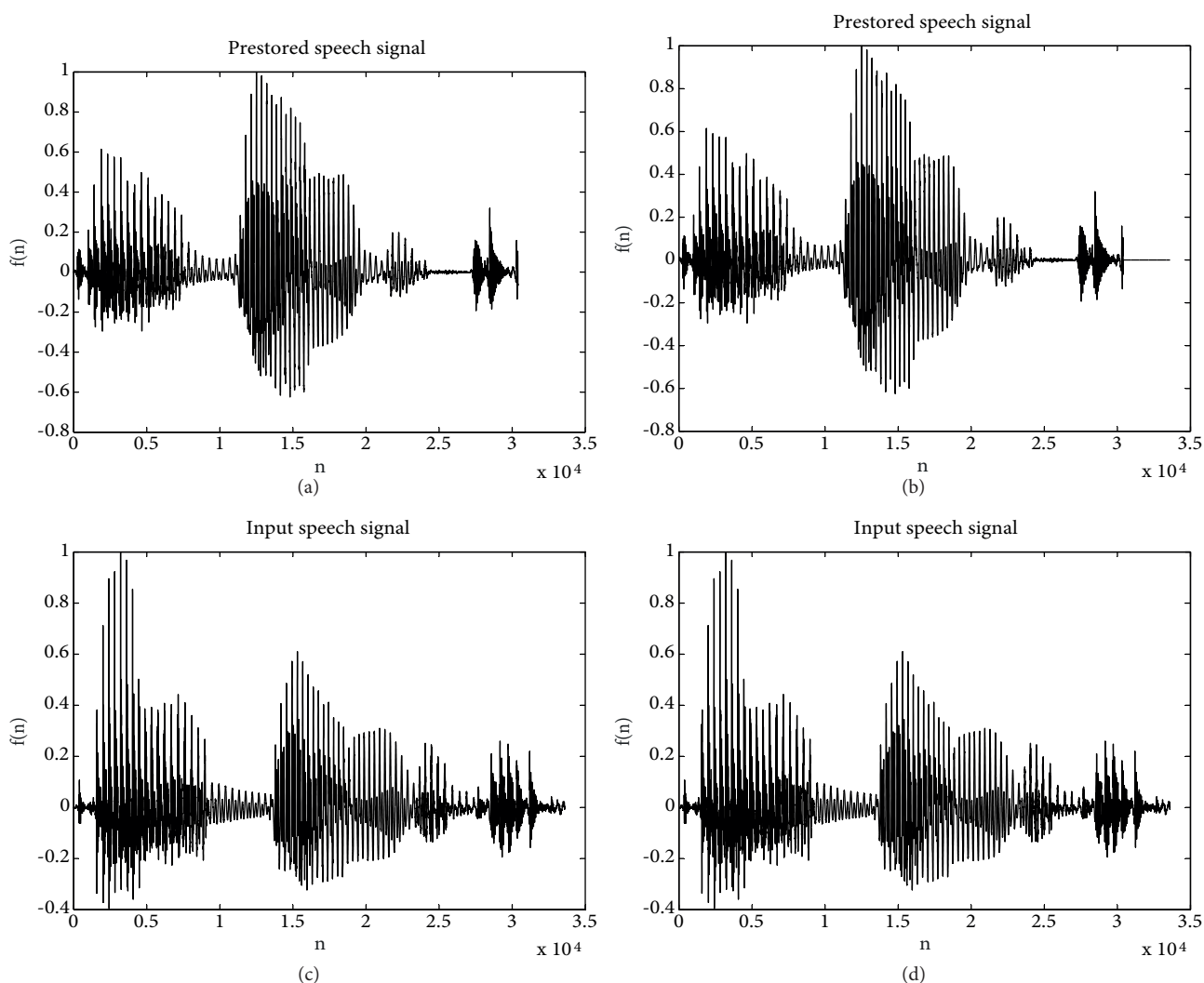


Figure 2. Common region adequacy: prestored speech signal a) before and b) after time domain adequacy and input speech signal c) before and d) after time domain adequacy.

equation:

$$NPSD(k) = \frac{Rn_{FG}(k)\overline{Rn_{FG}(k)}}{\max(Rn_{FG}(k)\overline{Rn_{FG}(k)})}, \quad (10)$$

where $\overline{Rn_{FG}(k)}$ denotes the complex conjugate of the $Rn_{FG}(k)$ sequence. Then, iii) compute and store the frequency band index by the following equations:

$$F_1 = \min(\{i | NPSD(i) \geq thr_p, 0 \leq i \leq N - 1\}), \quad (11)$$

$$F_2 = \max(\{i | NPSD(i) \geq thr_p, 0 \leq i \leq N - 1\}), \quad (12)$$

where thr_p is a calculated power threshold (see Table 1). The thr_p value is the threshold that allows the algorithm to eliminate frequency components with less energy than the minimum acceptable (set by the

threshold). To calculate thr_p , a comparison of TAR and TRR performance is made by changing the thr_p value. From Table 1 it can be observed that $thr_p=0.001$ is the value with the highest TAR and TRR performance.

Table 1. Example of performance of algorithm for different thr_p values.

thr_p value	TAR (%)	TRR (%)	thr_p value	TAR (%)	TRR (%)
0.0005	100	90	0.005	100	60
0.001	100	100	0.055	100	60
0.0015	100	90	0.006	100	10
0.0020	100	90	0.0065	100	10
0.0025	66	90	0.007	100	10
0.003	66	100	0.0075	100	20
0.0035	66	90	0.008	100	10
0.004	100	60	0.0085	100	20
0.0045	100	60	0.009	100	10

Finally, iv) calculate the BLPOC function using Eq. (6) and assume the zero frequency component of the \hat{r}_{fg} sequence shifted at the center of the spectrum (see Figure 4).

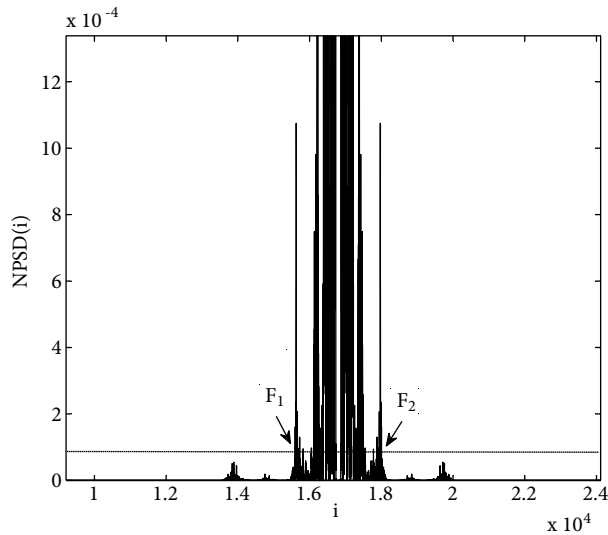


Figure 3. Effective frequency band identification from amplitude spectrum of $NPSD(k)$. Dashed line denotes the power threshold (thr_p).

3.4. Speaker verification decision

Since the BLPOC function in speaker verification has multiple peaks, some discrimination features to determine genuine and impostor matching need to be calculated. As an evaluation criterion, this method proposes to use a set of decision thresholds: 1) highest peak relationship, 2) variance, and 3) average energy.

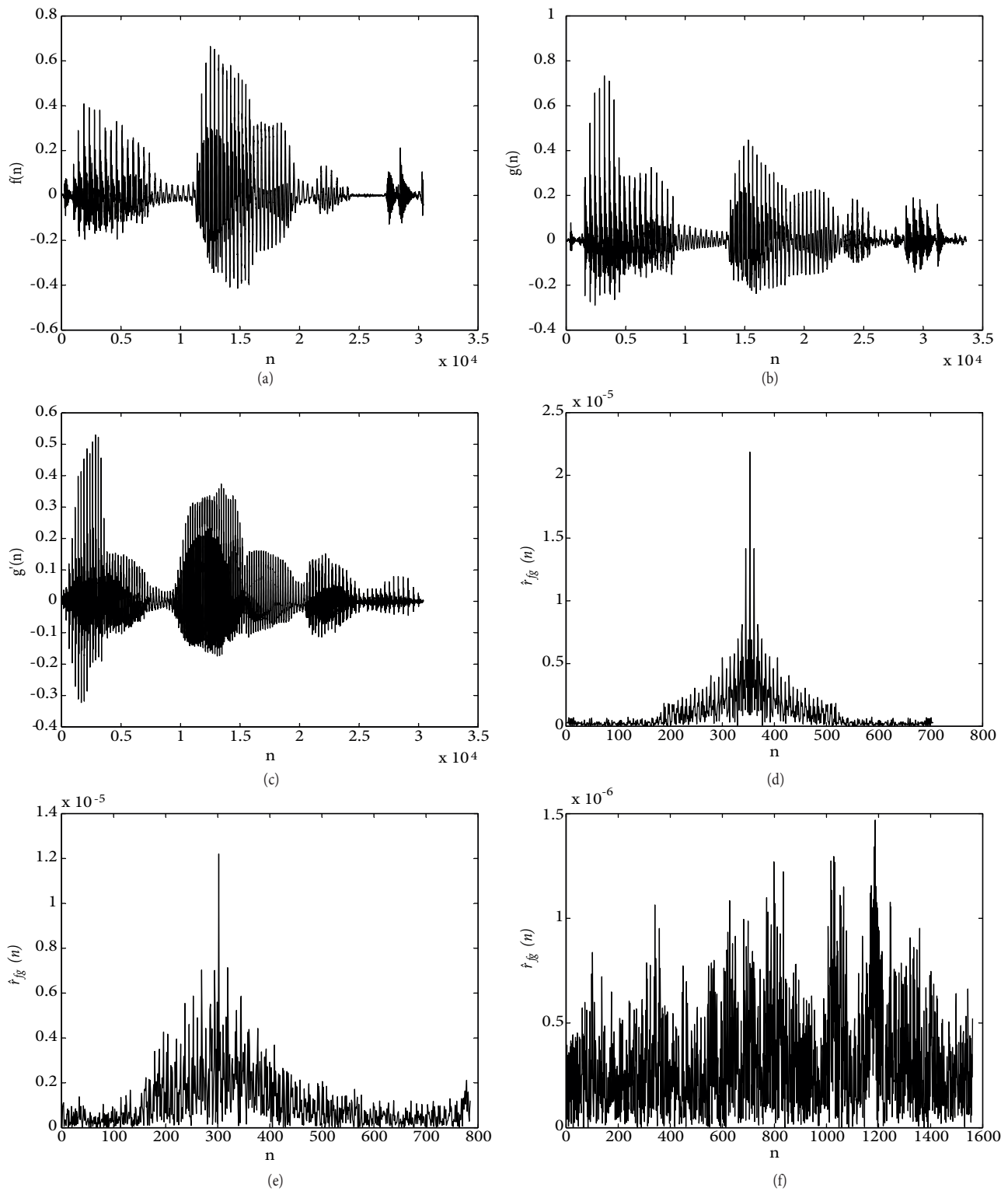


Figure 4. Examples of genuine and impostor matching from the same word uttered: a) speech signal from an authorized speaker ($f(n)$), b) another speech signal from an authorized speaker ($g(n)$), c) speech signal from an impostor speaker ($g'(n)$), d) BLPOC function between two identical speech signals (speech signal $f(n)$) from authorized speaker, e) BLPOC function between two speech signals from authorized speaker ($f(n)$ and $g(n)$), and f) BLPOC function between speech signals from the authorized ($f(n)$) and the impostor ($g'(n)$) speakers.

3.4.1. Highest peak relationship

The highest peak relationship (Hp) is calculated to evaluate the ratio between the max peak in the BLPOC function (P_M) and the other less significant peaks by calculating its mean energy (\bar{S}) using the following equations:

$$P_M = \max|\hat{r}_{fg}|, \tag{13}$$

$$\bar{S} = \frac{1}{N} \sum_{i=1}^N |\hat{r}_{fg}(i)|, \tag{14}$$

$$Hp = \frac{\bar{S} * 100}{P_M}. \tag{15}$$

An Hp acceptance section (s_{Hp}) is calculated. An acceptance Hp value is a relation percentage between the magnitude of P_M and the \bar{S} of the BLPOC function. s_{Hp} in the experiments was calculated automatically by selecting the maximum and minimum Hp values. These values change depending on the word uttered, the amount of data available for training, the speaker, and the quality of the recordings. Figure 5 shows an example of the delimitation of s_{Hp} (dotted horizontal lines) and an utterance test where only the magnitudes of Hp on the acceptance section defined by s_{Hp} are accepted.

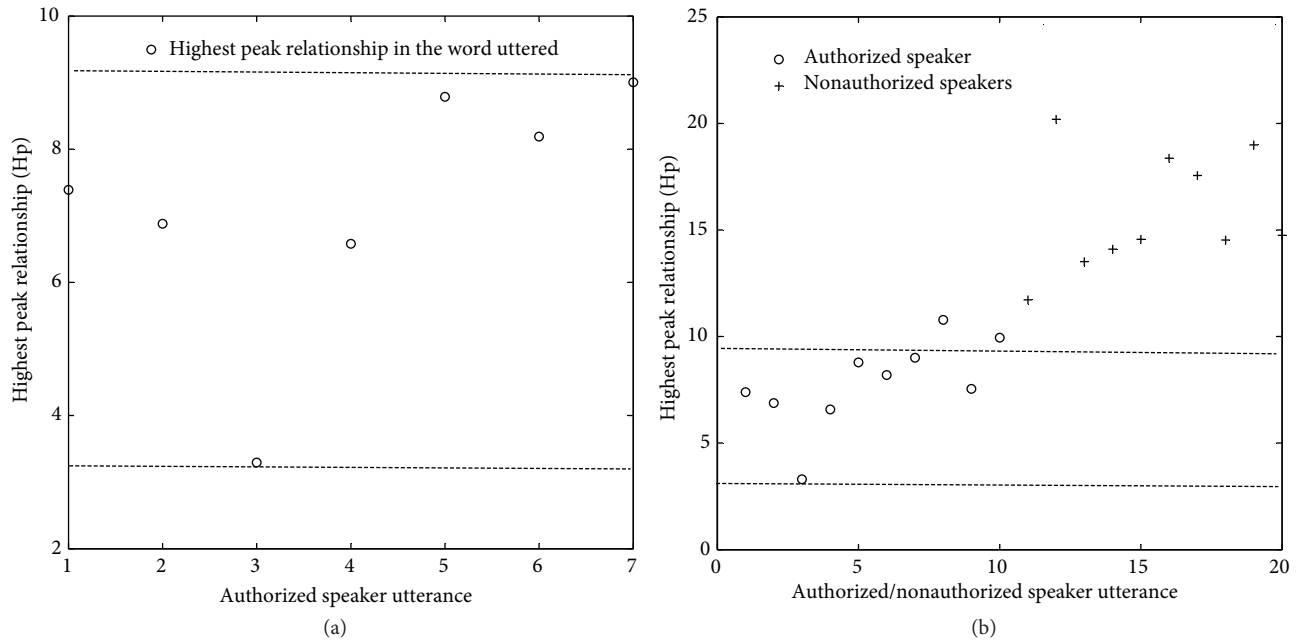


Figure 5. BLPOC Hp : a) authorized speaker utterance delimitation and b) utterance speaker test.

3.4.2. Variance

A variance feature (σ^2) is proposed to determine how scattered the samples are in the BLPOC function. This feature can be obtained by:

$$\sigma^2 = \frac{\sum_{i=1}^N (\hat{r}_{fg}(i) - \bar{S})^2}{N - 1}. \tag{16}$$

The σ^2 acceptance section (s_{σ^2}) sets the maximum and minimum (σ^2) in the BLPOC function considered as an acceptance value. The s_{σ^2} used in the experiments was calculated automatically by determining the maximum and minimum σ^2 values (these values also change depending on the word-uttered speaker-test). Figure 6 shows an example of the delimitation of s_{σ^2} (dotted horizontal lines) and an utterance test where only the magnitudes of σ^2 on the acceptance section defined are accepted.

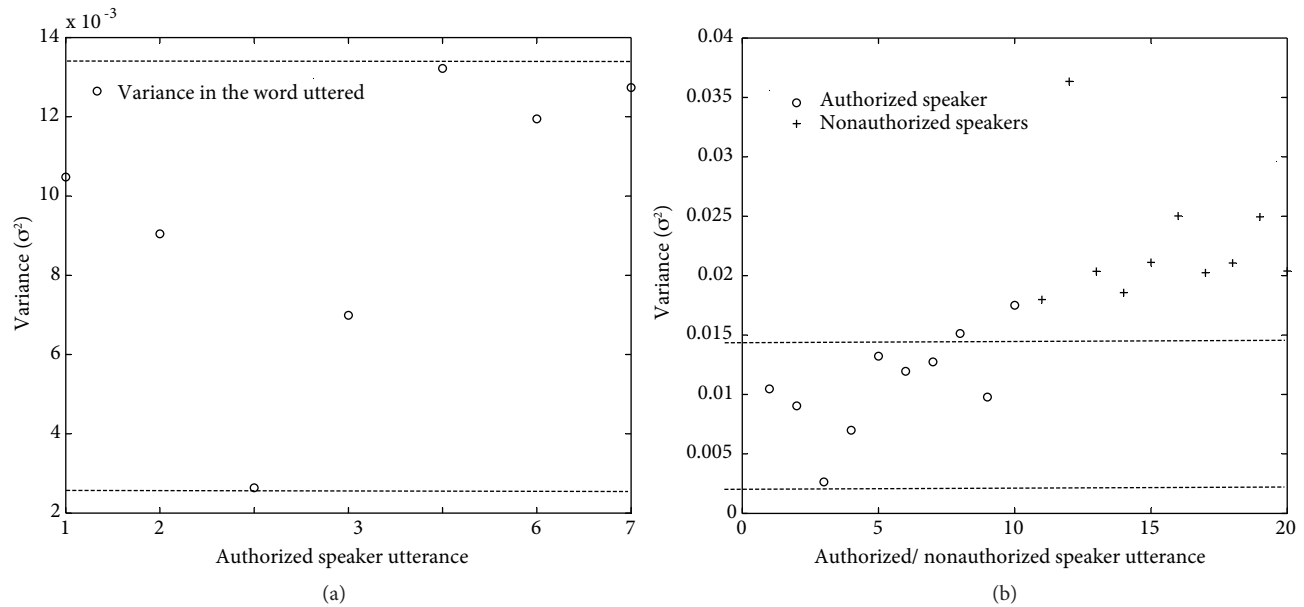


Figure 6. BLPOC σ^2 : a) authorized speaker utterance delimitation and b) utterance speaker test.

3.4.3. Average energy

The BLPOC function at genuine matching has (according to experiments) more average energy (E) than in impostor matching. This feature can be calculated by:

$$E = \frac{1}{N} \sum_{i=1}^N |\hat{r}_{fg}(i)|^2. \tag{17}$$

The E acceptance section (s_E) sets the maximum and minimum of E in the BLPOC function considered as an acceptable value. The s_E used in the experiment was calculated automatically by determining the maximum and minimum E values (these values also change depending on the word-uttered speaker-test). Figure 7 shows an example of the delimitation of s_E (dotted horizontal lines) and an utterance test where only the magnitudes of E on the acceptance section are accepted.

3.5. MFCC-HMM speaker verification technique

The mel frequency cepstral coefficient-hidden Markov model (MFCC-HMM) technique is summarized below to compare the proposed method with a traditional speaker verification technique. First, to calculate the MFCCs [7]: i) frame and window (using a Hamming window) a word-uttered speech signal and ii) for each frame ($y(n)$)

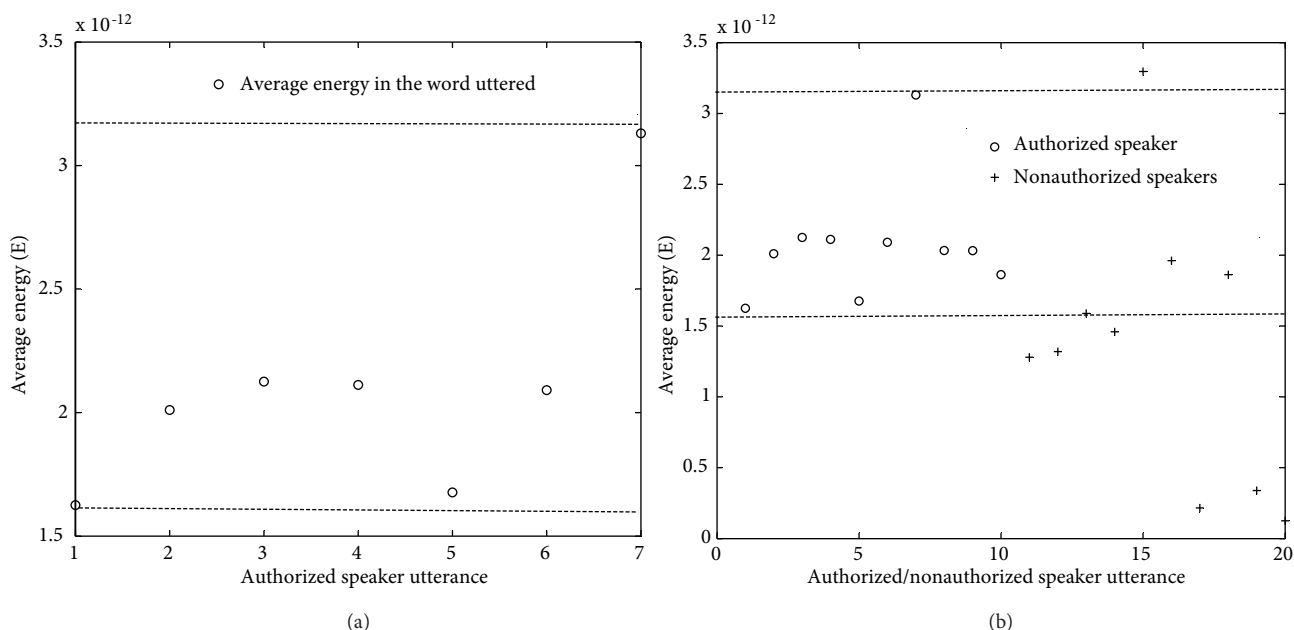


Figure 7. BLPOC E : a) authorized speaker utterance delimitation and b) utterance speaker test.

calculate the power spectrum ($y(k)$) according to:

$$y(k) = \frac{1}{N_s} \left| \sum_{n=1}^{N_s} y(n) W_N^{kn} \right|^2, \tag{18}$$

where N_s is the number of points of the DFT and $1 \leq k \leq N_s$. Then, iii) calculate a filter bank of p mel spaced triangular filters. Next, iv) compute the energy of each filter ($E(j)$) by:

$$E(j) = \sum_{k=1}^{N_s/2-1} y(k) M_j(k), \tag{19}$$

where $M_j(k)$ denotes the amplitude of the m_j triangular filter at frequency k , j is the index of each filter in the filter bank, and $0 \leq j < p$. Next, v) take the discrete cosine transform (DCT-II) according to:

$$C_l(t) = \sum_{j=0}^{p-1} \log [E(j)] \cos \left[t \left(j - \frac{1}{2} \right) \frac{\pi}{p} \right], \tag{20}$$

where $C_l(t)$ is the t th order MFCC of the l th frame and $t = 1, \dots, 13$. Finally, vi) extract the mean value of each frame coefficient to form a feature vector. The process repeats for each word-utterance from the speaker.

Next, the extracted information is trained and tested by using the HMM. The training step obtains the parameters of the model according to [23, 24]:

$$\lambda = (A, B, \pi), \tag{21}$$

where λ is a specific model (i.e. speaker-word model name), A is a probability matrix of the transitions between states, B is a probability matrix for the emissions given the states, and π is a vector of probabilities of each

state in the sequence. This parameter set is trained by an iterative process using the Baum–Welch algorithm. The results of the Baum–Welch algorithm are the optimized A , B , and π parameters. In other words, for each speaker-word, an HMM is trained that models the information of the given sequences. Finally, at the test step, given a sequence (word uttered by a speaker), the probability that the sequence was generated by a specific speaker (λ) is calculated.

The proposed speaker verification was evaluated using likelihood ratio sets [21]. Given N background speaker models ($\lambda_1, \lambda_2, \dots, \lambda_N$), the hypothesis model is calculated as:

$$p(X|\lambda_{hyp}) = f(p(X|\lambda_1), p(X|\lambda_2), \dots, p(X|\lambda_N)), \quad (22)$$

where λ_{hyp} is the hypothesized speaker and $f(\cdot)$ is the maximum of the likelihood values from the background speaker set. In other words, it is to set “nonauthorized speakers” in speaker tests (to cover the set of alternative model speakers) and classify if a speech signal is from who he/she claims to be (a yes or no decision). The HMM algorithm was implemented using free online software.¹

4. Results and discussion

Two different databases were used in the experiments. The custom database is a limited-data speaker-dependent voice corpus (in Spanish from Mexico) from an isolated-words speech recognition application [25]. The database contains 400 utterances (10-word utterance repetition by each of the 4 Mexican speakers in the database). Speech samples were collected using a digital Portastudio TASCAM DP-008 with a 44.1 kHz sampling rate in .wav format (there were no special considerations in the microphone features). Each sample was recorded in a video conference room to guarantee high SNR. The second database, the ELSDSR corpus of reading speech, was a joint effort of the faculty and PhD and master students from the Department of Informatics and Mathematical Modeling (IMM) and Technical University of Denmark (DTU) [26, 27]. The speech corpus is in English and was read by 20 Danes, 1 Icelander, and 1 Canadian (10 females, 12 males). The database contains 198 utterances (suggested 154 and 44 word-utterances for the training and test set, respectively). Although traditionally this database was implemented for automatic speaker recognition tasks, recently it was applied to speaker verification systems [28]. In this sense, the experiments collected only the most repeated uttered words by the speakers.

The following parameters were used as performance metrics to evaluate the proposed algorithms:

- True acceptance rate (TAR): The probability to verify as accepted (authorized) an authorized speaker (person).
- True rejection rate (TRR): The probability to verify as rejected (unauthorized) an impostor speaker (person).

For each database, 70% of audio files for the training set (7 utterances) were considered for the TAR test (per word-uttered authorized speaker-test), and the rest of the audios were used as the test set. On the other hand, for the TRR test, all the audio files from other speakers were used (word-uttered from nonauthorized speakers in each TAR test). The total number of genuine and impostor trials per each database (and per each proposed method) was 38 and 73 trials for the custom database and ELSDSR, respectively.

¹Murphy K. Hidden Markov Model (HMM) toolbox for MATLAB, 1998. Available online at <http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html>.

The proposed BLPOC algorithm was compared using four different feature threshold configurations: highest peak relationship, highest peak relationship-variance, highest peak relationship-average energy, and a total discrimination feature (highest peak relationship-variance-average energy).

Figure 8 shows the comparison of the algorithms using the custom database. The general *TAR* performance was nearly uniform for the first three proposed BLPOC feature threshold configurations. From highest peak relationship performance tests, an improvement (especially in *TRR* performance) is shown by adding the variance to the highest peak relationship. Thus, an unauthorized speaker had a probability of 85% to be rejected. Also, in comparison to highest peak relationship performance tests, an improvement of 9% is shown in *TRR* performance by using the highest peak relationship-average energy feature. On the other hand, taking into account all the decision thresholds (highest peak relationship, average energy, and variance), 98.24% *TAR* performance and 87.17% *TRR* performance can be observed. Finally, the MFCC-HMM algorithm has 1.75% and 100% *TAR* and *TRR* performances, respectively (being, from comparative Table 2, the technique with the lowest probability to correctly verify a speaker as authorized).

On the other hand, Figure 9 shows the comparison of the algorithms using the ELSDSR database. It shows a performance improvement by adding threshold features to the highest peak relationship feature. In this sense, taking into account all feature thresholds, 93.75% *TAR* and 67.05% *TRR* can be observed. Finally, the MFCC-HMM algorithm has 8.97% *TAR* and 99.83% *TRR* performance (see comparative Table 3).

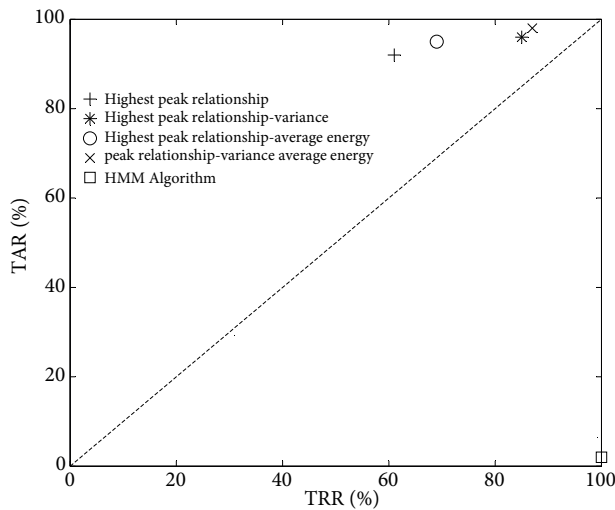


Figure 8. Algorithms' performances with custom database.

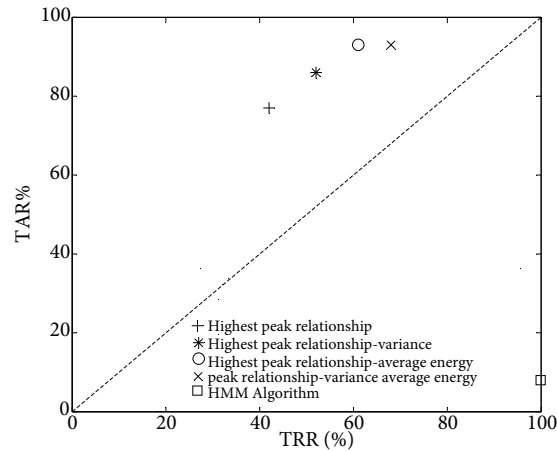


Figure 9. Algorithms' performances with ELSDSR.

Table 2. Performance of algorithms with custom database.

Highest peak relationship-variance-average energy performance	
TAR	TRR
98.24%	87.17%
MFCC-HMM algorithm performance	
TAR	TRR
1.75%	100%

Table 3. Performance of algorithms with ELSDSR database.

Highest peak relationship-variance-average energy performance	
TAR	TRR
93.75%	67.05%
MFCC-HMM algorithm performance	
TAR	TRR
8.97%	99.83%

As mentioned above, from Figure 8 and Figure 9, a performance improvement can be observed by applying different feature threshold configurations. Thus, as expected, the appliance of appropriate decision thresholds to the BLPOC function determines when to authorize a speaker or not.

The inherent frequency band is traditionally extracted by analyzing the DFT of the prestored speech signal [19]. However, to avoid inconsistencies in the BLPOC function depending on which is the prestored and the input speech signal, the proposed algorithm detects the inherent frequency band in the cross-phase spectrum (see Section 3.3).

From comparative Table 2 and Table 3, the superior configuration of the proposed algorithm has a better performance with the custom database than the ELSDSR database. The MFCC-HMM algorithm has the lowest probability of correctly verifying a speaker in both databases. Thus, the amount of training data plays an important role in the efficiency of traditional statistical methods. In this sense, since there is not a “formal training step” in the proposed algorithm, the selected acceptance section at every decision threshold represents a limited data training step. From the results, the BLPOC function, by setting some threshold configurations, can be offered as a limited data automatic speaker verification technique.

5. Conclusion

In this paper a new method was proposed for limited-data automatic speaker verification using the BLPOC function. Taking into account the performance tests, about 98% *TAR* and 87% *TRR* in the custom database and 93% and 67% respectively in the ELSDSR database can be achieved. The MFCC-HMM algorithm rejects almost every speaker. This specific result supports that with limited data, especially for training, the HMM traditional speaker verification method can provide an inferior performance. Therefore, a fair comparison of the proposed algorithm could be against other limited-data speaker verification methods.

Although the BLPOC function is an efficient method traditionally used in human identification by image matching, this is an effective method for a speaker verification system with limited-data conditions. From the results, the BLPOC function could be applicable in other sound signal biometric tasks where limited data exist.

References

- [1] Reynolds DA. An overview of automatic speaker recognition technology. In: IEEE 2002 International Conference on Acoustics, Speech, and Signal Processing; Orlando, FL, USA; 2002. pp. IV-4072-IV-4075.
- [2] Sharma D, Ali I. The effect of DC coefficient on mMFCC and mIMFCC for robust speaker recognition. In: 2015 International Conference on Advances in Computing, Communications and Informatics; Kochi, India; 2015. pp. 313-317.

- [3] Drgas S, Virtanen T. Speaker verification using adaptive dictionaries in non-negative spectrogram deconvolution. In: Vincent E, Yeredor A, Koldovský Z, Tichavský P (editors). *Latent Variable Analysis and Signal Separation*. Liberec, Czech Republic; 2015. pp. 462-469.
- [4] Li M, Kim J, Lammert A, Ghosh P, Ramanarayanan V et al. Speaker verification based on the fusion of speech acoustics and inverted articulatory signals. *Computer Speech & Language* 2016; 36: 196-211.
- [5] Kinnunen T, Li H. An overview of text-independent speaker recognition: from features to supervectors. *Speech Communication* 2010; 52 (1): 12-40.
- [6] Wu Z, Evans N, Kinnunen T, Yamagishi J, Alegre F et al. Spoofing and countermeasures for speaker verification: a survey. *Speech Communication* 2015; 66: 130-153.
- [7] Chakroborty S, Saha G. Improved text-independent speaker identification using fused MFCC & IMFCC feature sets based on Gaussian filter. *World Academy of Science, Engineering and Technology* 2009; 3 (11): 1968-1976.
- [8] Sharma D, Ali I. A modified MFCC feature extraction technique for robust speaker recognition. In: 2015 International Conference on Advances in Computing, Communications and Informatics; Kochi, India; 2015. pp. 1052-1057.
- [9] Sen N, Basu T, Chakroborty S. Comparison of features extracted using time-frequency and frequency-time analysis approach for text-independent speaker identification. In: 2011 National Conference on Communications; Bangalore, India; 2011. pp. 1-5.
- [10] Jayanna H, Prasanna S. Limited data speaker identification. *Sadhana* 2010; 35 (5): 525-546.
- [11] Mowlae P, Saeidi R, Stylianou Y. Advances in phase-aware signal processing in speech communication. *Speech Communication* 2016; 81: 1-29.
- [12] Schluter R, Ney H. Using phase spectrum information for improved speech recognition performance. In: *IEEE 2001 International Conference on Acoustics, Speech, and Signal Processing*; Salt Lake City, UT, USA; 2001. pp. 133-136.
- [13] Ito I, Kiya H. DCT sign-only correlation with application to image matching and the relationship with phase-only correlation. In: *IEEE 2007 International Conference on Acoustics, Speech and Signal Processing*; Honolulu, HI, USA; 2007. pp. I-1237-I-1240.
- [14] Miyazawa K, Ito K, Aoki T, Kobayashi K, Nakajima H. An efficient iris recognition algorithm using phase-based image matching. In: *IEEE 2005 International Conference on Image Processing*; Genova, Italy; 2005. p. II-49.
- [15] Miyazawa K, Ito K, Aoki T, Kobayashi K, Nakajima H. An effective approach for iris recognition using phase-based image matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2008; 30 (10): 1741-1756. doi: 10.1109/TPAMI.2007.70833
- [16] Shabrina N, Isshiki T, Kunieda H. Fingerprint authentication on touch sensor using Phase-Only Correlation method. In: 2016 7th International Conference of Information and Communication Technology for Embedded Systems; Bangkok, Thailand; 2016. pp. 85-89.
- [17] Rida I, Almaadeed S, Bouridane A. Gait recognition based on modified phase-only correlation. *Signal, Image and Video Processing* 2016; 10 (3): 463-470.
- [18] Teusdea AC, Gabor G. Iris recognition with phase-only correlation. In: *Annals of DAAAM for 2009 Focus on Theory, Practice and Education*; Vienna, Austria; 2009. pp. 189-190.
- [19] Ito K, Nakajima H, Kobayashi K, Aoki T, Higuchi T. A fingerprint matching algorithm using phase-only correlation. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences* 2004; 87 (3): 682-691.
- [20] Wu S, Wang Y, Zhan Y, Chang X. Automatic microseismic event detection by band-limited phase-only correlation. *Physics of the Earth and Planetary Interiors* 2016; 261: 3-16.
- [21] Bimbot F, Bonastre J, Fredouille C, Gravier G, Magrin-Chagnolleau I et al. A tutorial on text-independent speaker verification. *EURASIP Journal on Advances in Signal Processing* 2004; 2004 (4): 101962.
- [22] Kasap C, Arslan M. A unified approach to speech enhancement and voice activity detection. *Turkish Journal of Electrical Engineering & Computer Sciences* 2013; 21 (2): 527-547. doi: 10.3906/elk-1107-30

- [23] Rabiner LR, Juang BH. *Fundamentals of Speech Recognition*. Upper Saddle River, NJ, USA: Prentice-Hall, 1993.
- [24] Dumitru CO, Gavat I, Vieru R. Speaker verification using HMM for Romanian language. In: *Proceedings of ELMAR 2006*; Zadar, Croatia; 2006. pp. 131-134.
- [25] Pedroza A, de la Rosa J, Garcia E, Gamboa H, Becerra A. Diseño de prototipo para mejorar la dicción mediante el uso de Modelos Ocultos de Markov. *Pistas Educativas* 2016; 38 (120): 1020-1038 (in Spanish).
- [26] Jayanth M, Reddy B. Speaker identification based on GFCC using GMM-UBM. *International Journal of Engineering Science Invention* 2016; 5: 62-65.
- [27] Francis F, Vishnu R. A novel noise robust speaker identification system. *ARPJ Journal of Engineering and Applied Sciences* 2015; 10 (17): 7641-7646.
- [28] Lan Y, Hu Z, Soh Y, Huang G. An extreme learning machine approach for speaker recognition. *Neural Computing and Applications* 2013; 22 (3-4): 417-425.