ÖHÜ Müh. Bilim. Derg. / OHU J. Eng. Sci.

ISSN: 2564-6605
doi: 10.28948/ngumuh.516805
Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi, Cilt 8, Sayı 1, (2019), 48-58
Omer Halisdemir University Journal of Engineering Sciences, Volume 8, Issue 1, (2019), 48-58

Araştırma / Research

TWO-STAGE DECISION MAKING ALGORITHM FOR SPEAKER VERIFICATION WITH TRAINING SET OPTIMIZATION

Efe Tankut YAPAROĞLU[1] (ORCID: 0000-0003-1537-1237)
Yavuz ŞENOL[1] (ORCID: 0000-0002-3686-5597)[*]

[1]*Dokuz Eylül University, Department of Electrical and Electronics Engineering, Izmir, Turkey*

Geliş / Received: *08.02.2018*
Kabul / Accepted: *26.09.2018*

ABSTRACT

In this paper, a two-stage decision making algorithm is proposed for the task of speaker verification. This two-stage algorithm aims to eliminate the first-stage qualifying impostors by the help of impostor-resistant structure in the second stage. First, a baseline system is formed using mel-frequency cepstral coefficients (MFCC) as features and, a radial basis function (RBF) neural network for speaker modelling. Then, the investigations have been realized for optimizing the training set by means of two issues: (1) the ratio of impostor features to genuine speaker features, (2) the ratio of same gender features to opposite gender features (in respect of the genuine speaker) within the impostor speakers' set. Last, the two-stage decision making algorithm is presented, and the performance enhancement provided by the two-stage system is given with the test results.

Keywords: Speaker verification, Training set optimization, RBF neural network, MFCC, Cohort

KONUŞMACI DOĞRULAMA İÇİN EĞİTİM SETİ OPTİMİZASYONLU İKİ AŞAMALI KARAR VERME ALGORİTMASI

ÖZ

Bu çalışmada, konuşmacı doğrulama görevi için iki aşamalı bir karar verme algoritması önerilmiştir. Bu iki aşamalı algoritma, ikinci aşamada sahtekarlara dayanıklı yapı sayesinde ilk aşamayı geçen sahtekârları ortadan kaldırmayı amaçlıyor. Birinci aşamada, öznitelik olarak mel-frekanslı sepstral katsayılar (MFCC) kullanılarak temel bir sistem oluşturulmuş ve bir radyal taban fonksiyonu (RBF) sinir ağı kullanılarak konuşmacı modellemesi gerçekleştirilmiştir. Ardından, eğitim setini iki kısımda optimize etmek için araştırmalar gerçekleştirildi: (1) taklitçi konuşmacı özniteliklerinin gerçek konuşmacı özniteliklerine oranı, (2) taklitçi konuşmacı kümesi içinde aynı cinsiyet özniteliklerinin zıt cinsiyet özniteliklerine oranı (gerçek konuşmacıya bağlı olarak). Son olarak, iki aşamalı karar verme algoritması sunulmuş ve iki aşamalı sistem tarafından sağlanan performans artışı test sonuçlarıyla birlikte verilmiştir.

Anahtar kelimeler: Konuşmacı doğrulama, Eğitim kümesi optimizasyonu, RBF yapay sinir ağları, MFCC, Cohort

1. INTRODUCTION

Speaker recognition is concerned with the problems of identification and verification, each of which may in turn be text-dependent or text-independent [1, 2, 3]. In speaker identification, the aim is to determine which of

[*]Corresponding author / Sorumlu yazar. Tel.: +90 232 37170; e-mail / e-posta: yavuz.senol@deu.edu.tr

the registered speakers a given utterance comes from. The test utterance is scored against all possible speaker models, with the best score determining the speaker identity. In speaker verification, which will be focused on in this work, the aim is to give acceptance or rejection decision for the identity claim of a speaker. The claimant speaks the phrase into a microphone and his/her voice is analyzed by a verification system that makes the binary decision to accept or reject the user's identity claim or possibly to report insufficient confidence and request a new trial before making the accept/reject decision.

The general approach to automatic speaker verification (ASV) consists of: i) acquisition of digital speech, ii) extraction of speaker-specific features, iii) pattern matching, and iv) giving the accept/reject decision. A block diagram of this procedure is shown in Figure 1. The upper part of the figure is the "enrolment" stage, where the speaker-specific model is created.

In this work the aim is to form a high performance text-independent speaker verification system that can be used especially for security purposes. For this, a baseline verification system is proposed and, several experiments and investigations are made to improve the verification performance of the proposed system.

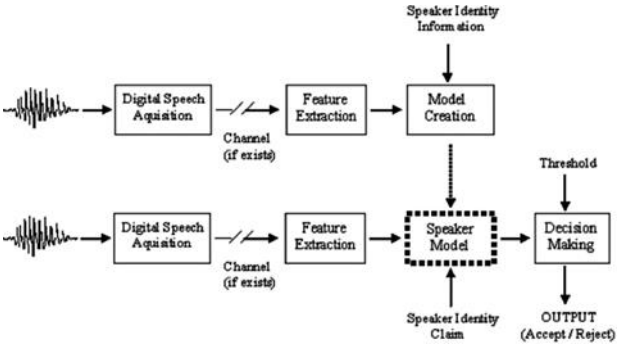

Figure 1. Block diagram of generic speaker verification system

In various literatures, for the speaker verification problem, many different methods have been used and analyzed at the phases of feature extraction and pattern classification. For the pattern classification task, widely used techniques are Gaussian mixture modelling [4], vector quantization [5], hidden Markov models [6, 7], support vector domain descriptions [8], and types of artificial neural networks [9-11]. The most used feature selection techniques in literature are linear predictive analysis methods [12], mel-frequency cepstral coefficients (MFCC) [13, 14], and discrete wavelet coefficients (DWTC) [15]. In this work, MFCC features and RBF neural networks are used because of the reported success of RBF in natural signal representation [16, 17].

2. DESIGN AND IMPLEMENTATION OF THE BASELINE SYSTEM

For the experiments, the IViE corpus [18], formed by 58 male and 58 female speakers, each uttering 20 sentences in various lengths which equals to 2320 utterances in total, is used. The recordings were made in English language and with sampling rate of 16 kHz..

Tests were performed using MATLAB software. First, the experiments were done for the baseline speaker verification system which is described below together with detailed results. In the second part, balancing of the training data is investigated for obtaining optimum performance. Then, by generating an additional cohort model, a new scoring technique is proposed, and results were reported.

The implemented baseline system consists of: pre-processing, MFCC features extraction, RBF neural network implementation, and performance evaluation. Each individual has his/her own model, created in the training phase. The experiments were done using these models and verification scores for the baseline system were obtained accordingly.

2.1. Extraction of Features

Feature extraction, by definition, is the estimation of variables called feature vector from another set of variables, at a considerably lower information rate. Pre-processing is an essential part of feature extraction and as the name implies, pre-processing involves the conditioning of digital speech signal prior to extracting the speaker-specific features from the speech signal. In this work, the speech signal sampled in 16 kHz in digital

format is taken and passed from a pre-emphasis filter in the form: *"y(n) = x(n) – 0.95 x(n-1)"* to suppress low frequency components. Then, the pre-emphasized speech utterance is divided into 256 sample frames, overlapping by 128 samples (16 msec. frame size, 8 msec. overlapping). Next step is to remove the silence frames according to Rabiner and Sambur method, since the silence frames contain no speaker-specific information. Then, the remaining non-silence frames are windowed by a Hamming window to minimize the signal discontinuities at the beginning and at the end of each frame. Next process is to convert the pre-processed speech frames into a spectral-domain representation by extracting the MFCC features. The procedure of MFCC feature extraction, which shows an example MFCC feature vector calculated from a preprocessed speech frame of 16 msec., is shown in Figure 2. In the baseline system of this work, 12 mel-filters are used and 11 mel-frequency cepstral coefficients are extracted as feature vector for each frame. The first coefficient component, C_0, is excluded since it does not carry significant speaker-specific information.

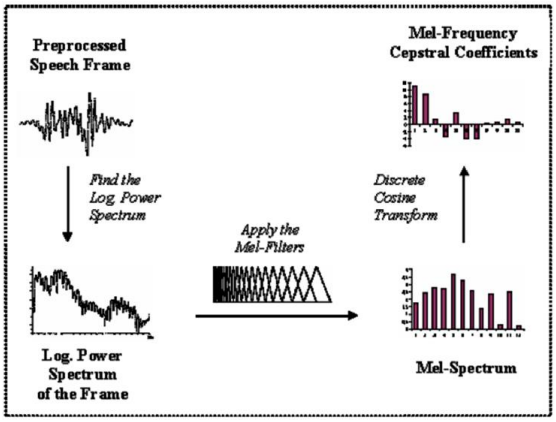

Figure 2. Schematic representation of MFCC feature extraction

2.2. Creation of the Speaker Model

In this work, for speaker modelling, a radial basis function (RBF) network is used. RBF neural network is a multidimensional function that depends on the distance between the input vector and a center vector [19]. The input layer has neurons with a linear function that simply feeds the input signals to the hidden layer. Moreover, the connection between the input layer and the hidden layer are not weighted, that is, each hidden neuron receives each corresponding input value unaltered. The hidden neurons are processing units that perform the radial basis function. In this work, Gaussian function is used as the transfer function of the hidden neurons, since it is common knowledge that Gaussian approximation provides good results in modelling the natural signals such as speech signals.

The output of the jth hidden neuron with Gaussian transfer function can be calculated

$$h_j = \exp(-\|x - c_j\|_2^2/\sigma^2) \quad (1)$$

where, h_j is the output of the j^{th} neuron , $x \in \mathcal{R}^{nx1}$ is an input vector, $c_j \in \mathcal{R}^{nx1}$ is the jth RBF center in the input vector space, σ is the center spread parameter which controls the width of the RBF, and $\| \ \|_2^2$ denotes the Euclidean norm. The output of any neuron at the output layer of RBF network is calculated as

$$y_i = \sum_{j=1}^{M} w_{ij}\, h_j \quad (2)$$

where, w_{ij} is the weight connecting hidden neuron j to output neuron i and M is the number of neurons in the hidden layer. The training of the RBF network can be realized through the weights in the output layer, the centers of the RBF NN, and the spread parameter of the gaussian function. The simplest form of RBF network training can be obtained with fixed number of centers. If the number of centers is made equal to the number of input vectors, namely exact RBF network, then the error between the desired and actual network outputs for the training data set will be equal to zero. In this work, Radial basis networks can be used to approximate functions.

As training algorithm Matlab newrb function was used. This training algorithm adds neurons to the hidden layer of a radial basis network until it meets the specified mean squared error goal.

In the baseline system described above, each speaker model is trained using RBF network, with a set of 27 different non-speakers. The output layer of the RBF network had 2 neurons, where the first output neuron represents the probability of being a true speaker feature, and the second output neuron represents the probability of being a non-speaker feature. The proposed RBF network with 11 inputs and 2 outputs is given in Figure 3.

During the training, the RBF network was trained to give output values of 1 and 0, respectively, for true speaker feature. In the same manner, output values were set to 0 and 1, respectively, for non-speaker feature value.

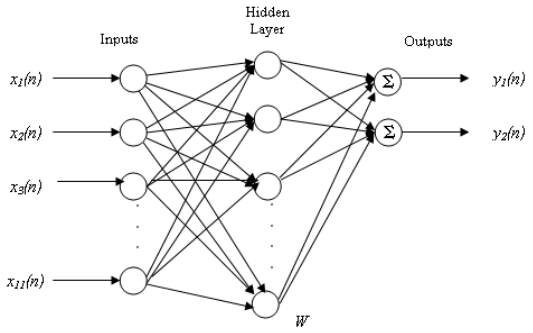

Figure 3. The proposed RBF neural network

The score of each test utterance (M is the number of feature vectors in the test utterance) was evaluated as,

$$S = \frac{1}{M}\sum_{n=1}^{M}\left((y_1(n) - y_2(n)\right) \quad (3)$$

2.3. Evaluation Of The Baseline System

The tests have been done with 288 non-speaker utterances and 26 true speaker utterances. In this stage of the work, the performance was evaluated in terms of equal error rate (EER). Eight different models are created for four male speakers and four female speakers. In the training set the amount of true speaker features were chosen to be equal to the amount of non-speaker features. Also, the amount of male non-speaker features was equal to the amount of female non-speaker features in the world set. The test scores for the un-optimized baseline system are given in Table 1. Note that the speaker names given in Table 1 are in the abbreviation form.

Table 1. Test scores in terms of EER for the baseline system

	Female Speakers					**Male Speakers**				
Speaker	**wsc**	**wkt**	**mmm**	**cmf**	**Average**	**wlh**	**wer**	**mpm**	**cmc**	**Average**
eer (%)	9.7	9.3	6.4	6.9	**8.1**	3.2	5.5	4.9	6.7	**5.1**

Figure 4 and Figure 5 are the FAR-FRR (False Acceptance Rate - False Rejection Rate) graphs for two different speakers that are trained and tested according to the baseline model. The equal error rate (EER) is the point where FAR equals to FRR.

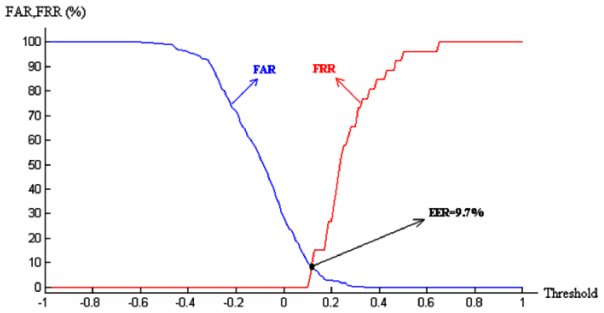

Figure 4. Performance of speaker "wsc" (without training set optimization)

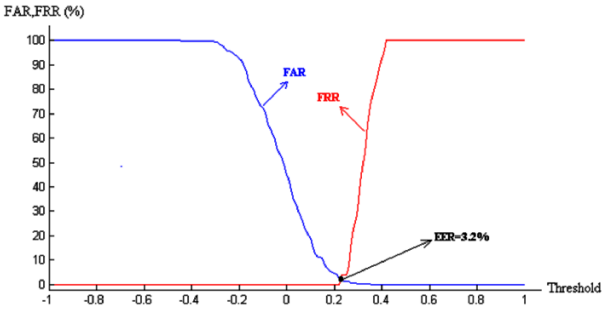

Figure 5. Performance of speaker "wlh" (without training set optimization)

2.4. Optimization of the Training Set

In the baseline system, the feature vectors in the training set are composed of two types of features: *i)* True speaker features (composed of 12 different utterances, total 540 feature vectors); *ii)* Non-speaker features (composed of 27 non-speakers, 20 feature vectors from each, total 540 feature vectors).

2.4.1. Ratio of Same-Gender Non-Speaker Features to Opposite-Gender Non-Speaker Features in the Training Set

The concept in the title of this part can be explained as: if the verification model belongs to a male speaker, the male non-speakers are the same-gender non-speakers, the female non-speakers are the opposite-gender non-speakers. A series of experiments have been performed to determine the optimum ratio of speaker genders in the non-speaker set. The obtained results are shown in Table 2.

Table 2. EER performances for different percentages of same-gender non-speakers in training set

EER(%)		**Speaker Name**								
		wsc	**wkt**	**mmm**	**cmf**	**wlh**	**wer**	**mpm**	**cmc**	**Average EER**
Percentage Of Same Gender Non-Speakers	**33%**	10.1	11.6	7.6	8.9	4.5	7.2	6.3	10.3	**8.3**
	50%	9.7	9.3	6.4	6.9	3.2	5.5	4.9	6.7	**6.6**
	67%	7.1	6.9	6.0	5.9	1.8	4.1	4.3	6.4	**5.4**
	80%	6.6	6.2	4.9	5.7	1.3	4.5	4.3	5.6	**4.9**
	90%	**6.9**	**6.1**	**4.6**	**4.5**	**1.2**	**3.3**	**3.2**	**4.6**	**4.3**
	100%	7.4	7.2	4.8	4.9	2.4	3,4	3,4	4,8	**4.8**

Figure 6 and Figure 7 are the FAR-FRR graphs for the situation when there are 90% same-gender non-speakers and 10% opposite-gender non-speakers. When Figure 6 is compared with Figure 4, and when Figure 7 is compared with Figure 5, the improvement in performance can easily be observed.

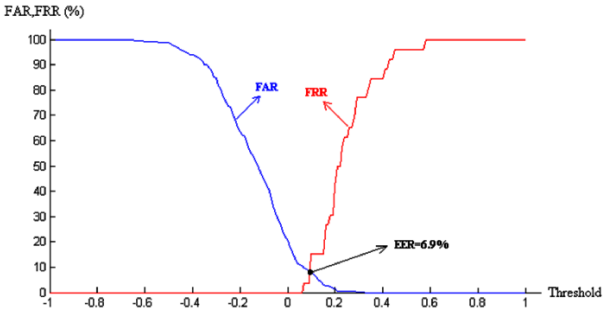

Figure 6. Performance of speaker "wsc" when trained with 90% same-gender and 10% opposite-gender non-speakers.

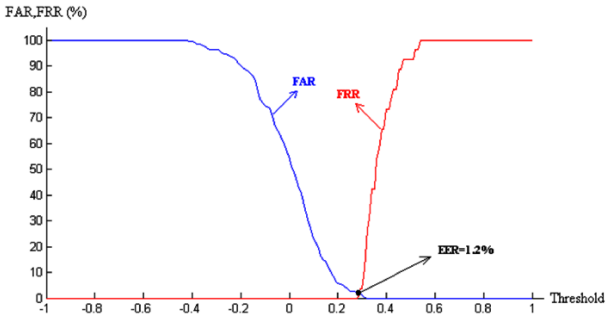

Figure 7. Performance of speaker "wlh" when trained with 90% same-gender and 10% opposite-gender non-speakers.

2.4.2. Ratio of True Speaker Features to Total Non-Speaker Features in the Training Set

As it is known, the RBF neural network should be trained with both true-speaker and non-speaker features in order to discriminate the speaker's specific vocal properties from the other speakers. The ratio of these true speaker features to non-speaker features is also an important parameter that affects the verification performance. The results from experiments for different true-speaker percentages in the training set is as shown in Table 3 (non-speaker features set is composed of 90% same- gender non-speakers, and 10% opposite-gender non-speakers). As seen from Table 3 and Figure 8, the minimum equal error rate value (EER) is achieved when the number of true speaker features in the training set becomes equal to the number of non-speaker features (when %S = 50%). Also, it is obvious from Table 3 that when the percentage of true speaker features in the training set is increased above %50, the error rate is considerably increased.

Table 3. EER performances for different percentages of true speaker features in the training set

EER(%)		**Speaker Name**								
		wsc	**wkt**	**mmm**	**cmf**	**wlh**	**wer**	**mpm**	**cmc**	**Average EER**
Percentage Of True Speaker Features	**33%**	8.1	7.2	5.7	6.2	2.6	4.3	4.0	5.4	**5.4**
	40%	7.3	7.1	5.5	5.4	1.5	3.8	3.8	5.1	**4.9**
	50%	**6.9**	**6.1**	**4.6**	**4.5**	**1.2**	**3.3**	**3.2**	**4.6**	**4.3**
	60%	7.9	6.9	5.4	5.6	2.0	4.1	4.2	5.3	**5.2**
	67%	13.5	11.8	8.6	9.4	6.1	6.9	7.2	9.5	**9.1**

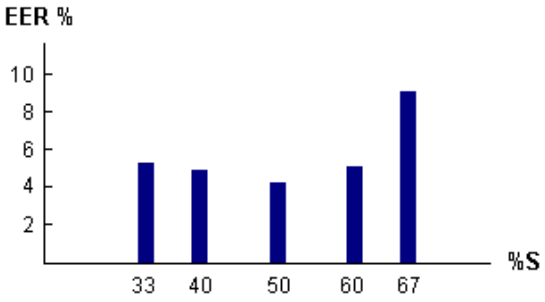

Figure 8. Effect of true speaker features percentage in the training set
(%S = true speaker features / total features in the training set)

3. THE TWO-STAGE DECISION MAKING ALGORITHM AND RESULTS

In a speaker verification system, a decision should be made such that; either the claimant is accepted or the claimant is rejected. Also, in some systems there may be a doubtful region and within this, the claimant may be asked to repeat his/her utterance one more time. The acceptance, rejection or unsure decisions are made according to the predefined threshold(s) of the speaker model. So, determination of threshold(s) is an important process in forming a speaker verification model. With different threshold values, different FAR and FRR values are obtained, and there is a trade-off between them. One of these error values is improved at the expense of the other one. For an application, when a security system is the case, FAR value is kept as low as possible, while keeping the FRR value at an acceptable level. In this part, we propose a two-stage decision making algorithm to eliminate the impostors who qualify from the first stage. The algorithm block diagram representation of the proposed verification system is given in Figure 9.

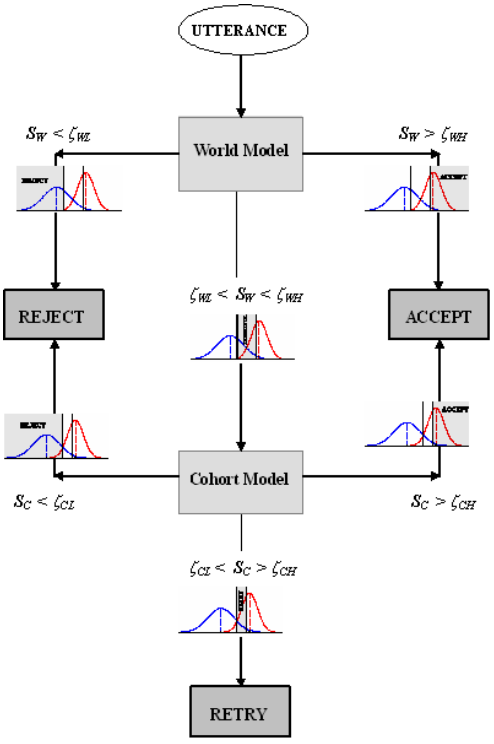

Figure 9. The proposed speaker verification algorithm

First stage aims to test the claimant utterance with the "world model", which is an RBF network formed by a training set according to the results obtained in Section 5. In the first stage, the claimant is directly rejected if his/her score is below the rejection threshold "ζ_{WL}", or directly accepted if his/her score is greater than the acceptance threshold "ζ_{WH}". Therefore, the claimant with a score below the threshold ζ_{WL} or above the threshold ζ_{WH} does not pass into the second stage. If the claimant's score is between these thresholds, the utterance is fed into the "cohort model" which is the second stage of the verification system. Cohort means a group of speakers who are acoustically close to the genuine speaker. The reason behind cohort selection is that a speaker verification system trained with cohort speakers can deal with impostor attacks more precisely. However, the disadvantage of a cohort model is that it may not discriminate well the genuine speaker features from the world set of non-speakers.

The block diagram for the procedure of world model creation, selection of cohort features and training the cohort model is shown in Figure 10.

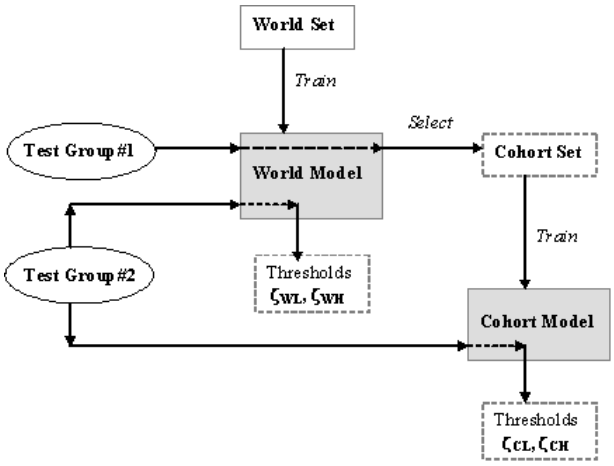

Figure 10. Generation of the world model, cohort model and acceptance/rejection thresholds

3.1. Determination of Thresholds

Speaker rejection threshold of the world model was determined as

$$\zeta_{WL} = \mu_1 + k_1\sigma_1 \tag{4}$$

Speaker acceptance threshold of the world model was determined as

$$\zeta_{WH} = \mu_2 - k_2\sigma_2 \tag{5}$$

where, μ_1 is the mean of impostor scores from the network, μ_2 is the mean of genuine speaker scores from the network, σ_1 is the standard deviation of impostor scores, and σ_2 is the standard deviation of genuine speaker scores. As seen from Figure 11 the doubtful region is between the thresholds ζ_{WL}, and ζ_{WH}. The reject region is on the left-hand side of the doubtful region, whereas the accept region is on the right-hand side.

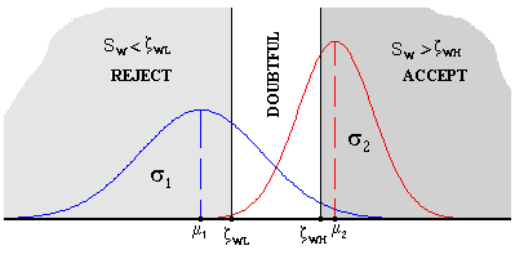

Figure 11. Acceptance and rejection thresholds for the world model

3.2 Generation of the Cohort Model

The reason for using a cohort model is to eliminate the impostors as much as possible when they manage to pass from the first stage without being eliminated.

To generate the cohort model, a separate test group of 50 non-speakers (Test Group #1 in Figure 10) were used, and 18 non-speakers out of Test Group #1 were extracted who obtained the highest scores from the world model. Then, these 18 non-speaker features were used to train a new RBF network to form the cohort model. The acceptance and rejection thresholds of the world model and the cohort model were determined by using Test Group #2, and with respect to the genuine speaker features' mean and standard deviation such that:

Speaker rejection threshold of the cohort model was determined as

$$\zeta_{CL} = \mu_4 - k_3\sigma_4 \tag{6}$$

Speaker acceptance threshold of the cohort model was determined as

$$\zeta_{CH} = \mu_4 - k_4\sigma_4 \tag{7}$$

where, μ_4 is the mean of genuine speaker features, σ_4 is the standard deviation of genuine speaker features when tested with the cohort model, k_3 and k_4 are constants such that $k_3 > k_4$.

It can be seen from Equations (4) and (5) that, both the acceptance and rejection thresholds were arranged according to the genuine speaker's statistics. This is for controlling the width of the unsure area (retry region), where claimants are asked to re-utter to the verification system. In this way, the number of false acceptance and false rejections can be reduced at the expense of increased processing time. Here, the retry region as seen in Figure 12 is between the thresholds ζ_{CL} and ζ_{CH}. The k_3 and k_4 coefficient values have the main role in determination of FAR and FRR.

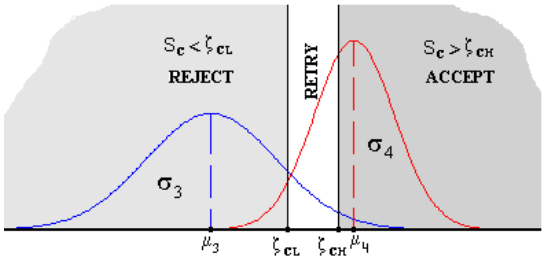

Figure 12. Acceptance and rejection thresholds for the cohort model.

Tests were done with 288 non-speaker utterances and 26 true speaker utterances. The results for the proposed two-stage automatic speaker verification system are given in Table 4.

Various studies for speaker verification have been proposed in the last years. Some of these use the same IViE corpus database. Jhanvar and Raina [20] presented a clustering approach based on the concept of a pitch correlogram for fast speaker identification. The data base used in the research was IViE corpus with 110 speakers (55 male and 55 female). In the experiment only 5% of male and 1.5% of female speakers were misclassified. Djemili et all [21] proposed a speech signal based gender identification system using four classifiers. The algorithm was applied on IViE corpus by selecting 110 speakers (50% male and 50% female). Experimental results gave the best identification rate of 96.4%. Another study [22] tried to improve speaker identification performance of a speaker identification system based on a frame level scoring. This study used IViE corpus constituted by 112 speakers (56 male and the rest are female). Experimental results showed that incorrectly identification error rate was achieved as 3.4%. These results show that two-stage decision making algorithm does improve false acceptance error.

Table 4. Test results for the two-stage speaker verification system

Speaker Model	% False Acceptance Error	% Impostor Retried	% False Rejection Error	% Speaker Retried
Average	**0.65**	**6.44**	**5.75**	**9.6**

4. CONCLUSIONS

In this work, a two-stage decision making algorithm which involves determination of acceptance and rejection thresholds and generation of a cohort set is introduced. The aim of the two-stage decision making algorithm is to eliminate the impostors in the second stage by means of the cohort model. The final output from the system is either accept, or reject, or retry. In addition, the training set is optimized in terms of: the ratio of impostor features to genuine speaker features, and the ratio of same gender features to opposite gender features (in respect of the genuine speaker) within the impostor speakers' set.

The proposed system has indeed decreased the average false acceptance error to 0.65 percent while keeping the false rejection error average at 5.75 percent. The results can be accepted as very promising for a high security speaker verification application.

REFERENCES

[1] WIQAS G., NAVDEEP S., "Literature Review on Automatic Speech Recognition". International Journal of Computer Applications, 41, 42-50, 2012.

[2] SHIKHA G., AMIT P., ACHAL S., "A Study on Speech Recognition System: A Literature Review", International Journal of Science, Engineering and Technology Research (IJSETR), 3, 2192-2196, 2014.

[3] LIU Y, QIAN Y., CHAN N., FU T., ZHANG Y., YU K., "Deep Feature for Text-dependent Speaker verification", Speech Communication, 73, 1–13, 2015.

[4] BHATTACHARYYA S, SRIKANTHAN T, KRISHNAMURTHY P, "Ideal GMM parameters and posterior log likelihood in speaker verification", Proc. IEEE Signal Processing Soc. Neural Networks for Signal Processing XI, 471-480, 2001.

[5] XU Y., SHEN F., ZHAO J., "An incremental learning vector quantization algorithm for pattern classification". Neural Computing and Applications, 21, 1205–1215, 2012.

[6] GALES M., YOUNG S., "The Application of Hidden Markov Models in Speech Recognition". Foundations and Trends in Signal Processing, 1, 195–304, 2007.

[7] PATEL I., SRINIVAS Y. R., "A Frequency Spectral Feature Modeling for Hidden Markov Model Based Automated Speech Recognition" Recent Trends in Networks and Communications, Communications in Computer and Information Science, 90, 134-143. Springer, Berlin, Heidelberg, 2010.

[8] KAMRUZZAMAN S. M., A. N. M. REZAUL KARIM A. N. M., ISLAM S., HAQUE E., "Speaker Identification using MFCC-Domain Support Vector Machine", International Journal of Electrical and Power Engineering, 1, 274-278, 2007.

[9] NAIR P. G., NAIR R., "Efficient Speaker Identification Using Artificial Neural Network", International Journal of Electronics & Communication Technology (IJECT), 6, 27-30, 2015.

[10] SWAMY S., SHALINI T., NAGABHUSHAN S.P., NAWAZ S., RAMAKRISHNAN K.V., "Text Dependent Speaker Identification and Speech Recognition Using Artificial Neural Network" Global Trends in Computing and Communication Systems. Communications in Computer and Information Science, 269, 160-168. Springer, Berlin, Heidelberg, 2012.

[11] YUE X, YE D, ZHENG C, WU X, "Neural networks for improved text-independent speaker identification", IEEE Engineering in Medicine and Biology Magazine, 53-58, 2002.

[12] MUSTA E., KOMINI V., "A Comparative Study Of Linear Predictive Analysis Methods With Application To Speaker Identification Over a scripting programing", Journal of Multidisciplinary Engineering Science and Technology (JMEST), 2, 2881-2885, 2015.

[13] SINGH A. K., SINGH R, DWIVEDI A., "Mel Frequency Cepstral Coefficients Based Text Independent Automatic Speaker Recognition Using Matlab", International Conference on Reliability, Optimization and Information Technology (ICROIT), 524-527, Haryana, India, 2014.

[14] DAS A., JENA M.R., BARIK K. K., "Mel-Frequency Cepstral Coefficient (MFCC) - a Novel Method for Speaker Recognition", Digital Technologies, 1,1-3, 2014.

[15] PANDIARAJ S., SHANKAR KUMAR K. R., "Speaker Identification Using Discrete Wavelet Transform", Journal of Computer Science, 11, 53-56, 2015.

[16] HALDAR R., MISHRA P. K., "Multilingual Speech Recognition Using Radial Basis Function (RBF) Neural Network", International Research Journal of Engineering and Technology (IRJET), 3, 2856-2862, 2016 .

[17] SHARMA S., SHUKLA A., MISHRA P., "Speech and Language Recognition using MFCC and DELTA-MFCC", International Journal of Engineering Trends and Technology (IJETT), 12, 449-452, 2014.

[18] http://www.phon.ox.ac.uk/files/apps/old_IViE/download1.php (erişim tarihi 08.01.2018)

[19] HAYKIN S., Neural Networks a Comprehensive Foundation (2nd ed), Prentice Hall Inc. USA, 1999.
[20] JHANWAR N., RAINA., "Pitch Correlogram Clustering for Fast Speaker Identification", EURASIP Journal on Applied Signal Processing, 17, 2640-2649, 2004.
[21] DJEMILI R., BOUROUBA H., KORBA M.C.A., "A Speech Signal Based Gender Identification System Using Four Classifiers", 2012 International Conference on Multimedia Computing and Systems, 1-4, Tangier, Morocco, 10-12 May 2012.
[22] DJEMİLİ R., BOUROUBA H., KORBA M.C.A., O'SAUGHNESSY D., "Boosting Speaker Identification Performance Using a Frame Level Based Algorithm", International Conference on Communications, Signal Processing, and their Applications (ICCSPA'15), 1-6, Sharjah, United Arab Emirates, 17-19 Feb. 2015.