MULTI-SPEAKER TEXT-INDEPENDENT SPEAKER VERIFICASTION USING WAVELET AND NEURAL NETWORK

Dr. Bayez Al-Sulaifanie[*] , Dr. Ahmed M. Al-Kababji[**]

[*]Computer Engineering Department, University of Duhok, Iraq
[**] Computer Engineering Department, University of Mosul, Iraq
bayez_k@yahoo.com , ahmedalkababji@yahoo.com

ABSTRACT

Automatic speaker recognition systems use machines to recognize a person from a spoken phrase. These systems can operate in two modes: to identify a particular person or to verify a person's claimed identity. Personal identity verification is an essential requirement for controlling access to protected resources or in forensic applications.

One of the still challenging fields in speaker recognition is to verify a person in a multi-speaker environment. The multi-speaker recognition task has been in the NIST (National Institute of Standards and Technology) evaluation plan since 1999. However, the researches done in this field are not sufficient and there are very few as compared the single speaker recognition systems.

In this paper, a new speaker recognition system was proposed and tested for one and multi-speaker task. The system was constructed of a wavelet decomposition front end followed by a linear predictive coding cepstral feature extractor. The matching process was accomplished by a probabilistic neural network (PNN) with a background model as an imposter reference. The proposed system was able to reach an equal error rate (EER) of 2.35% for a one speaker male gender dependent system and an EER of 3.35% for a one speaker female gender dependent system. For the two speakers and at a target to imposter ratio (TIR) of 3db the male gender dependent system had an EER of 8.7%, while for female system the EER value was 17%. The TIMIT corpus was used as the system evaluation database.

Keywords: *Speaker recognition, Speaker verification, Wavelet, Probabilistic neural network, Multi speaker*

1. INTRODUCTION

Although many speech processing tasks, like speech and speaker recognition, reached satisfactory performance levels on specific applications, and even though a variety of commercial products were launched in the last decade. Many problems remain an open research area due to the fact that there is an increasing need for person authentication in the world of information, applications ranging from credit card payments to border control and forensics.

In short, contemporary speech/speaker recognition systems are composed of a feature extraction stage, which aims at extracting speech/speaker's characteristics while evading any sources of adverse variability, and a classification stage, that identifies the feature vector with certain class. The feature extraction phase converts the input speech signal in a series of multi-dimensional vectors, each corresponding to a short segment of the acoustical speech input. The resulting feature vector makes use of information from all spectrum bands, and therefore, any inaccuracy of representation and any distortion induced to any part of the spectrum is spread to all features forming the vector. The classification stage that is based on the probability density function of the acoustic vectors is seriously confused in case of impaired features.

Historically, the following speech features dominated the speech and speaker recognition areas in consequent periods: LPC, LPCC, and MFCC. Other speech features like, PLP, ACW, wavelet-based features, although presenting reasonable solutions for the same tasks, did not gain widespread practical use, often due to their relatively more sophisticated computation. Nowadays many earlier computational limitations are overcome, in view of the significant performance boost up of contemporary microprocessors. That opens possibilities for revaluation of the traditional solutions when speech features are selected for a specific task. MFCC model the human auditory system, since they account for the nonlinear nature of pitch perception, as well as for the nonlinear loudness perception. That makes them more adequate features for speech recognition than other formerly used speech parameters like CC, LPC, and LPCC. That success of MFCC, combined with their robust and cost-effective computation, turned them in almost "a must" in the speech recognition area. Because of that, MFCC became widely used on speaker recognition tasks, too, although they might not represent the speaker's voice individuality with a sufficient accuracy.

In an attempt to find out a more suitable representation of speech signal for the task of speaker verification, we investigate alternative ways to represent speaker's voice individuality. In this study, by a combination of wavelet analysis and the well known

LPCC speech parameters, we seek a more general approach, which allows easy handling of the spectral content of speech signal leading to a new speaker verification system. This new system was tested in the one as well as in the multi-speaker verification area.

Multi-speaker detection (verification) is the task of determining whether a particular known speaker is present in a speech segment containing speech from multiple speakers. It may be viewed as an extension of the basic one-speaker detection task. The 1998 NIST speaker recognition multi speaker evaluation was held during the summer of 1998, which was a special supplement to an ongoing series of yearly evaluations conducted by NIST [9, 11]. The multi-speaker detection task was added to the NIST evaluations in 1999 [12]. It is a new and challenging area of research, in this paper the performance of the new system was found for different TIR (Target to Imposter Ratio).The corpus used was the American English TIMIT corpus.

2. WAVELET ANALYSIS

Over the last few decades, wavelet analysis has been proven an effective signal processing technique for a variety of problems.

Wavelet is a type of the tree structure nonuniform filter bank in which speech signal is filtered in stages, and the sampling rate is successively reduced at each stage. Wavelets are based on mathematical constructs that deal with the linear expansion of a signal into contiguous frequency bands, instead of analyzing a signal with a single fixed window, as with short-time Fourier transform techniques, wavelets enable analysis with multiple window durations that allow for a coarse to fine multi resolution perspective of the signal (see Figure 1).

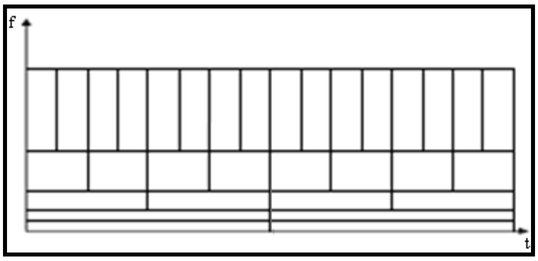

Figure 1: The time frequency plane defined by a wavelet basis.

The multi-resolution analysis implementation is based on the wavelet decomposition algorithm developed by Mallat [7]. Multi-resolution analysis of a signal decomposes it into a hierarchical system of subspaces that are one-dimensional and square integrable. Each resolution of the decomposition consists of a multi-resolution subspace and an orthogonal subspace. These subspaces can also be referred to, respectively, as the *discrete approximation* and the *detail signal* at a particular resolution. Orthogonality implies that no correlation exists between subspaces of different resolutions. Each subspace is spanned by basis functions that have scaling characteristics of either dilation or compression, depending on the resolution. The implementation of these basis functions is incorporated in a recursive pyramidal algorithm, in which the discrete approximation of a current resolution is convolved with quadrature mirror filters in the subsequent resolution.

Quadrature Mirror Filters (QMF) consist of a pair of filters whose frequency responses are complementary [7, 14]. Essentially, they are high- and low-pass filters that define the bandwidth for a particular resolution. A particular resolution in the decomposition process can also be referred to as an octave (see Figure 2). Figure 3 shows an example of a speech signal and its four decomposition levels, (the detail only). The low-pass filter is denoted by *h(n)* and the high-pass filter by *g(n)* which can be derived from *h(n)* as follows:

$$g(n)=(-1)^{1-n}h(1-n) \qquad (1)$$

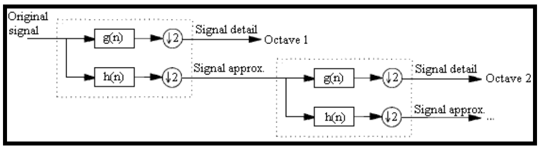

Figure 2: Successive recursive stages of multi-resolution decomposition using QMF.

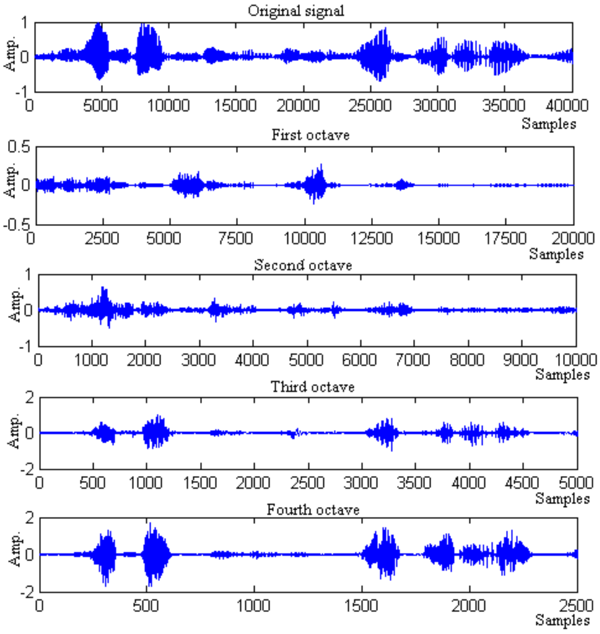

Figure 3: wavelet representation of a speech signal.

The multi resolution theory of orthogonal wavelets proves that any conjugate mirror filter characterizes a wavelet that generates an orthonormal basis of any energy signal [7]. From Figure 3 it can be seen that some segments of the original signal appear in some octaves and not in others. Therefore, from the point of view of a verification system dealing with the multi speaker problem, using wavelets, a speech signal for a target speaker contaminated with speech from an imposter can be processed in a multi resolution analysis. In this multi resolution analysis the contamination appears in some octaves and not in others, so the features extracted from this multi resolution system will be less contaminated than if they are directly extracted from the original speech signal.

3. PRE-PROCESSING

The front-end preprocessing stage consists of three parts. The first part in the preprocessing stage is the silence removal, silence and any low energy periods (including noise like unvoiced periods) are removed from the speech signal using an energy based speech detector explained in the following steps:

1. The (16000 samples/s) speech signal is divided into segments of 100 samples (each representing duration of 6.25ms).
2. The energy of each segment (100 sample) is calculated using the following equation:

$$energy = \sum_{n=1}^{100} (s(n))^2 \qquad (2)$$

If the energy of a segment is greater than a specific value, which was determined experimentally, the segment is left otherwise the segment is removed.

The second part in the preprocessing stage is the four levels wavelet decomposition (explained previously) that enables the proposed system to examine and extract features from the speech signal in a new form of four octaves.

The third part in the preprocessing stage is windowing. Each of the four octaves was windowed into frames of 20 ms with an overlap of 50% (10 ms) using Hamming window.

4. FEATURE EXTRACTION

At this stage each octave level consists of N frames where N:

$$N= [(D_{sp}\text{-mod}(D_{sp},10\text{msec}))/10\text{msec}]\text{-}1 \qquad (3)$$

D_{sp} is the duration of the input speech signal. In the feature extraction stage the LPC coefficients are calculated for each frame using a 16th LPC predictor. Kinnunen [5, 6] after a brief study of features for speaker recognition stated that an LPC predictor order should be larger than 15, but at the same time if the number of coefficients is chosen too high, this can lead to long calculation time.

Therefore, a vector of 16 LPC coefficients represents each frame of each octave resulting that a vector of 64 LPC coefficients represents each 20ms of the speech signal; this vector is called *code vector*. In the next step, Equation (4) is used to convert the LPC coefficients of each octave to its corresponding LPC cepstrum coefficients (LPCC). At this point the speech signal of duration D_{sp} is represented by a 64*N matrix (N code vectors) as shown in Figure 4.

$$c[n] = \begin{cases} a[n] + \sum_{k=1}^{n-1} \frac{k}{n} c[k] a[n-k], & 1 \le n \le p \\ \sum_{k=n-p}^{n-1} \frac{k}{n} c[k] a[n-k], & n > p \end{cases} \qquad (4)$$

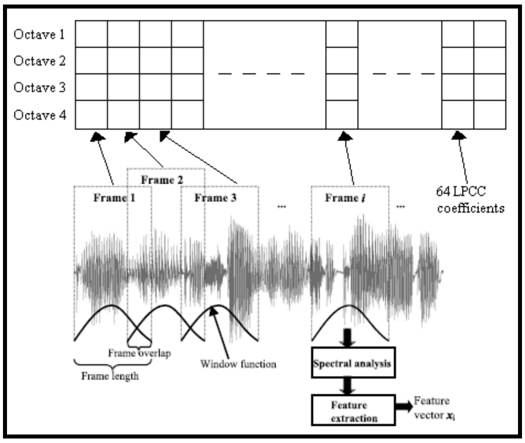

Figure 4: Driving the 64*N LPCC coefficients matrix from the speech signal.

5. SPEECH CORPUS

The speech corpus used to examine the performance of this proposed speaker recognition system was the standard American English TIMIT provided by Linguistic Data Consortium [2]. From the 630 speakers available in the TIMIT corpus a subset of 105 male speakers and a subset of 105 female speakers were used in this work. The selection of the subsets was arbitrary. The male speaker's subset contained 71 speakers from the dialect region DR7 and 34 speakers from the dialect region DR4. The female speaker's subset contained 31 speakers from the dialect region DR4, 35 speakers from the dialect region DR5, 13 speakers from the dialect region DR6 and 26 speakers from the dialect region DR7. There were 10 speech files for each speaker; two of the files had the same linguistic content for all speakers, whereas the remaining eight files were phonetically diverse.

When building the background model all of these ten files were used. For a target speaker, eight of its ten files available including one of the phonetically identical files made the training set for this speaker. The remaining two files were used for testing.

The background model was constructed as follows:

1. For each speaker included in the construction of the background model a feature matrix was found for each of its ten speech files.
2. A matrix describing the speaker features is obtained by horizontally conceiting and clustering the formal ten feature matrices into a 128*64 feature matrix.
3. The last step is to cluster all of the describing matrices of all the speakers used to build the background model into one matrix which is called the background model matrix.

6. TRAINING AND TESTING OF THE SYSTEM

This text-independent SV system is built on a modular structure with an individual Probabilistic Neural Network (PNN) [11] for each enrolled user.

A useful interpretation of the network outputs under certain circumstances is to estimate the probability of class membership, in which case the network is actually learning to estimate a probability density function (PDF). This is the case of the probabilistic neural network (PNN). The PNN is a special type of neural network using a kernel-based approximation to form an estimate of the PDFs of the categories in a classification problem. This particular type of ANN provides a general solution to pattern classification problems by following the probabilistic approach based on the Bayes decision theory. The network paradigm uses the Parzen-Cacoulos estimator to obtain the corresponding PDF of the classification categories. PNN uses a supervised training set to develop probability density functions within a pattern layer [3]. Figure 5 shows the architecture of the probabilistic neural network.

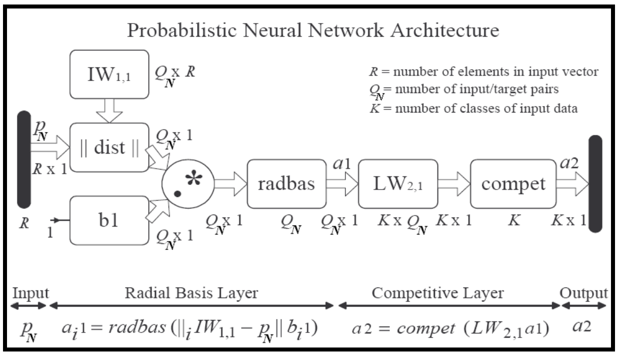
Probabilistic Neural Network Architecture
IW1,1
Q_N x R
R = number of elements in input vector
Q_N = number of input/target pairs
K = number of classes of input data
p_N
R x 1
|| dist ||
Q_N x 1
b1
R
1
Q_N x 1
.*
Q_N x 1
radbas
Q_N
$a1$
Q_N x 1
LW2,1
K x Q_N
K x 1
compet
K
$a2$
K x 1
Input
Radial Basis Layer
Competitive Layer
Output
p_N
$a_i1 = radbas(\|_i IW_{1,1} - p_N\| b_i1)$
$a2 = compet(LW_{2,1}a1)$
$a2$

Figure 5: Architecture of the Probabilistic Neural Net

In the training phase, a training matrix was built by horizontally conceiting the background model matrix with a describing feature matrix for a certain target speaker. The describing feature matrix of the target speaker was obtained from eight of its ten files available including one of the phonetically identical files.

First, the feature matrix was found for each of the eight speech files. Then, the eight feature matrices are horizontally conceited and clustered to get the describing feature matrix for the specific target speaker. This training matrix is used to train the neural network. The output of one is obtained if a feature vector belonging to the target speaker is presented to the neural network input and a zero otherwise.

The testing phase was performed after training the neural network. In the testing phase, the speech sample from the trial speaker is used to find its corresponding feature matrix. Then, a verification rate (VR) score is calculated for this trial speaker. The average verification rate (VR) score is measured as the percentage of the input vectors that score an output of one to the total vectors in the feature matrix of the speaker under trial. The value of VR obtained varies between 0% and 100%.

The last step is to compare the verification score with a certain threshold, the speaker is considered a target if the calculated verification score is larger or equal to this threshold and an imposter otherwise.

The size of the background model was 384 code vectors; while a codebook size of 128 code vectors represented the target speaker. Therefore, the training matrix that was used to train the neural network had a total size of 512 code vectors.

7. EXPERIMENTS AND RESULTS

7.1 ONE SPEAKER SYSTEM

Series of experiments were performed to obtain the DET plots that measure the performance of the proposed system in the one speaker verification domain. For the one speaker verification system, there are two ways of dealing with this system. One way gives importance to the sex of the targeted speaker (gender dependent), while the other (gender independent) does not.

In this work, the proposed system was examined for all of the probable situations as follows:

1. Male targeted speaker with a background model that was built of male speakers only.
2. Female targeted speaker with a background model that was which is built of female speakers only.
3. Male-female targeted speakers with a background model that was which is built of male and female speakers.

7.1.1 MALE ONE SPEAKER SYSTEM

From the male subset 30 speakers each having 10 speech files were chosen to build a background model. This model was used in the target and imposter trials. The choice of the 30 speakers from the 105 speakers in the subset was arbitrary (the 105 speakers files were arranged in alphabetic order and the first 30 files were chosen). The remaining 75 speakers of the subset were used in the target trials. The verification process was carried out by building a neural network for each targeted speaker. Therefore 75 neural networks were built. Each speaker had ten files, 8 of these files were used for training the network of that speaker and the other two for testing. Therefore, 150 target trials were performed. The total imposter trials were 900. For each of the 75 targeted speakers 12 imposter trials were performed. The 12 imposter trials belonged to three imposter speakers chosen arbitrary each contributing with four speech files. Figure 6 shows the DET plot for the performance of the male one speaker verification system.

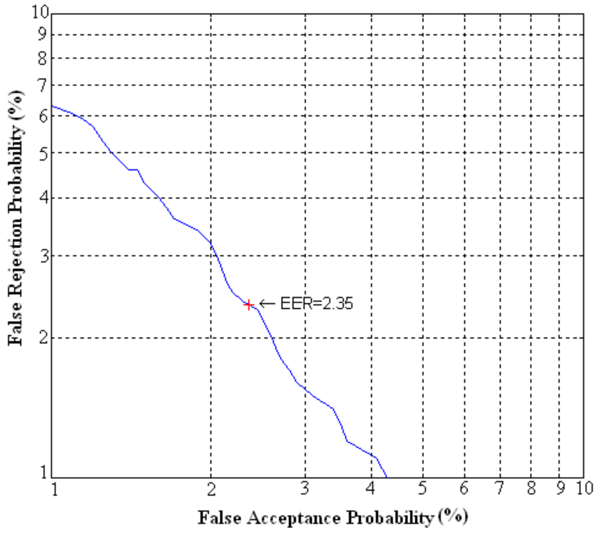

Figure 6: DET plot for the performance of the male one speaker verification system.

From Figure 6, it is clear that the system has almost linear performance with a line slop of -1 as the threshold is varied from 0% to 100% to calculate the false alarm and miss probabilities. Another important point to be noticed is the EER, which is the measure of the system performance. The EER of (2.35%) was obtained for male one speaker system

7.1.2 FEMALE ONE SPEAKER SYSTEM

The target and imposter trials in the female one speaker verification system were performed on the female subset in a way similar to that for the male one speaker verification system (i.e. 150 target trials and 900 imposter trials). Figure 7 shows the DET plot for the performance of the female one speaker verification system.

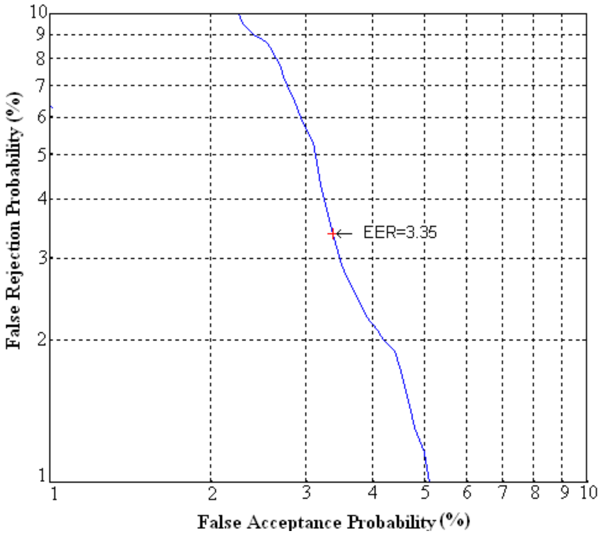

Figure 7: DET plot for the performance of the female one speaker verification system

Comparing the results obtained for the female system (Figure 7) to that of the previous male system (Figure 6), it can been seen that the female system has a higher EER (3.35%) than that of the male system. This result (male system better than female) is found in most gender dependent recognition systems (see [8, 13]), while there are systems where the opposite is true [11] or there is no difference in the performance [1]. A gender dependent system is a system where the targeted (hypothesized) speaker is always of the same sex as the test speaker.

7.1.3 MALE-FEMALE ONE SPEAKER SYSTEM

Before examining the performance of the male-female one speaker system, a background model of 30 speakers was built. This background model consisted of 15 male speakers and 15 female speakers taken from the male and female subsets (the speaker selection was arbitrary).

After building the background model, 37 male target trials and 38 female target trials were performed making 75 target trials. For the imposter trials 280 imposter trials were performed covering all the four possible probabilities equally (male network male imposter trial, male network female imposter trial, female network male imposter trial and female network female imposter trial).

Figure 8 shows the DET plot for the performance of the male-female one speaker verification system.

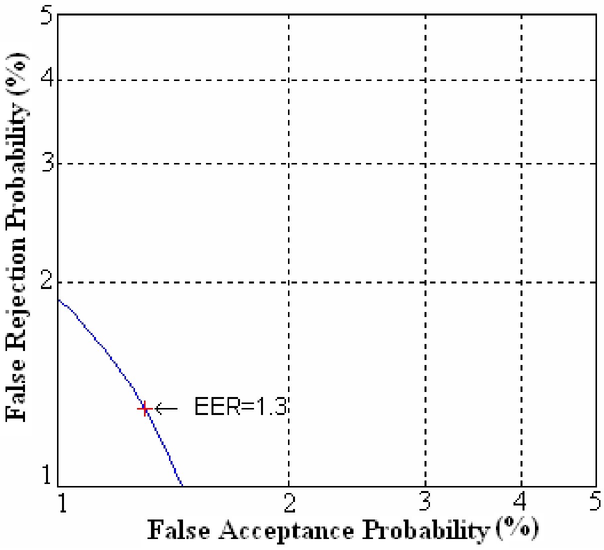

Figure 8: DET plot for the performance of the male-female one speaker verification system.

7.2 MULTI-SPEAKER VERIFICATION SYSTEM

In the multi-speaker system, the training phase does not differ from that of the one speaker system. The main difference between the one speaker system and the multi-speaker system is that the speech signal in the testing phase belongs to more than one speaker. In other words, the speech signal is a sum of two or more speech signals each signal belonging to a specific speaker.

The procedure of adding the speech signals was carried out as follows (supposing two signals):

1. Assume that the first speaker has a digitized speech signal $s_1(u)$ where $1 \le u \le U$ and the second speaker has a digitized speech signal $s_2(v)$ where $1 \le v \le V$.
2. The values of U&V are compared and the minimum one is taken to be the value of the upper range for the two signals. Therefore, s_1 and s_2 will have the same length ($\min_{U,V}$). The out-of-range samples of the longer signal will be discarded.
3. The final step is to find the new signal which is a result of adding sample by sample of the two speech signals s_1&s_2.

$$s_{new}(n) = s_1(n) + s_2(n) \qquad n=1,2,\ldots,\min_{U,V} \qquad (5)$$

In the test phase there are two types of trials, the target and imposter trials. For the target trial in the multi-speaker system a new expression must be introduced that is the target to imposter ratio (TIR). The TIR is calculated as the ratio of target speech power to the imposter speech power [4, 15]:

$$TIR = 10 \log \sum_{1}^{\min_{U,V}} [s_1(n)]^2 \Big/ \sum_{1}^{\min_{U,V}} [s_2(n)]^2 \qquad (6)$$

7.2.1 MALE MULTI-SPEAKER SYSTEM

The steps performed in examining the performance of the male multi-speaker system are the same as those in the male one speaker system. First, a background model was built using 30 reference speakers. Then 75 neural networks were built and trained. Each network belongs to one of the remaining 75 speaker from the male subset.

The only different step is the last step (testing). Here, the input signal is the sum of two signals from two speakers of the same sex (the targeted speaker signal and the imposter speaker signal), with different TIR values. The TIR values were 25dB, 10dB, 3dB, 0dB and -3dB. Figure 9 shows the multi-speaker system performance for numerous TIR values.

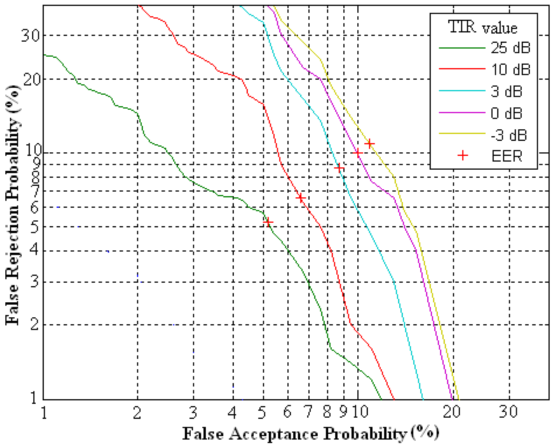

Figure 9: Multi-speaker system performance for numerous TIR values (male system).

7.2.2 FEMALE MULTI-SPEAKER SYSTEM

As in the male multi-speaker system a female only background model was built from the speech files of 30 female speakers taken from the female subset (the selection was arbitrary). This background model was used to train 75 neural networks belonging to the remaining 75 speaker.

The test input signal in this case is the sum of two signals from two female speakers (the targeted speaker signal and the imposter speaker signal), with different TIR values. The TIR values were 25dB, 10dB, 3dB, 0dB and -3dB. Figure 10 shows the multi-speaker system performance for numerous TIR values.

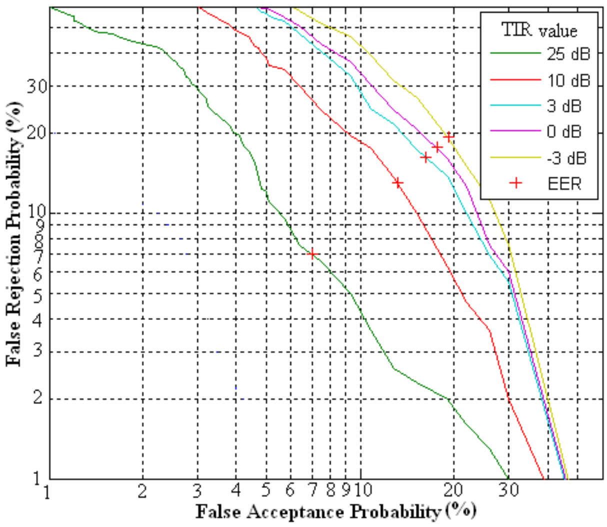

Figure 10: Multi-speaker system performance for numerous TIR values (female system).

For all the systems mentioned above (male, female, male-female, one or multi-speaker) the construction of the background model and the training of the neural networks are done off time. It was found that the execution time needed by the system to accept or reject a given speech signal was in the range of 1 to 1.5 sec. on a P4, 2.8MHz, 256Mbytes personal computer. From this we conclude that this verification system is capable of working in real time conditions.

8. CONCLUSIONS

There is a very important point that must be mentioned, that is the fact that our work is on the TIMIT corpus, while most of the recent works have used the Switchboard telephone conversational speech corpus. This point makes our work different from others as discussed below:

1. In our work, we used the TIR representation that was calculated from the target and imposter speech signals. These speech signals were present for all the durations in the input signal (a duty cycle of 100%). While in the Switchboard corpus the test segment consisted of summed two-sided intervals of a conversational speech segment having the two

speakers (the target and imposter speakers) recorded on it with a duty cycle of target speakers varying from close to zero to 100% [10]. This point leads to two important differences between our work and any other work that uses the Switchboard corpus. Firstly, in this work the value of the TIR is known and its effect was studied and taken into consideration, while in other works where the Switchboard corpus is used there is no care of the TIR value because the two speech signals (target and imposter signals) are prerecorded as a sum. Secondly, the 100% duty cycle of target and imposter speech in our work make it a harder job and worst-case situation than that in the Switchboard corpus.

2. The duration of the speech segments in our work was (8 to 20 sec.), while in the NIST Switchboard corpus it was about 59 to 60 sec. [10]. The duration of the speech signals used to examine a particular verification system affects its performance in a way that the system has a better performance as the duration of speech signal increases [10, 13]. This is true for one and multi-speaker systems.

In this work, it is the first time that a multi-speaker system is tested and its performance was evaluated against the level of corruption. The ratio of the energy of the targeted speech signal to the energy of the imposter speech signal (TIR) was used as a measure of this corruption. The duty cycle of the corruption was taken as 100% of the speech signal.

The following points have been reached at from the system proposed and results obtained:

- It has been found that the system is capable of working in real time due to the fact that each targeted speaker model including the construction of the background model is done off time. The processing time required to verify a person is approximately 1 to 1.5 seconds for all systems (male, female, male-female, one or multi-speaker).
- A male gender dependent system has a better performance than for female system, while a gender independent system outperformed the gender dependent system as expected. This result supports the policy of the NIST not to include cross-sex tests.
- In spite of the fact that a laboratory-recorded corpus is used to examine the proposed system, the worst EER value obtained for our proposed system is 3.35% (for the female system) which is considered a low value as compared to other existing systems.
- In this work the multi-speaker system is tested and its performance is evaluated against the level of corruption. The ratio of the energy of the targeted speech signal to the energy of the imposter speech signal (TIR) was used as a measure of this corruption. The duty cycle of the corruption was taken as 100% of the speech signal.
- For a value of TIR of 25dB the EER is approximately twice its value for a pure input speech signal (2.35% rises to 5.1% for male system and 3.35% rises to 7% for female system).
- In the multi-speaker system, the male system was less affected than the female system and the raising of the EER as the value of TIR increased was slower especially in the range of (25-10 dB) of TIR.

Investigation of the system performance for more than two speakers is recommended along with the following future work:

- More development on the system is recommended to improve the system performance for the female speaker recognition case wavelet packet is recommended to focus more on high frequencies in the spectrum of the speech signal.
- The performance of the system for database other than the TIMIT can be measured and the results obtained can be compared. Adding to that the system performance can be measured for corpus of other languages (Arabic for example) if the necessary corpus is available.

REFERENCES

[1]Douglas A ., Thomas F., Robert B., "Speaker Verification Using Adapted Gaussian Mixture Models", *Digital Signal Processing*, Vol. 10, No. 1-3, pp. 19-51, (2000).

[2]Garofolo J., Lamel L., Fisher W., *Darpa TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM Manua*l, National Institute of Standards and Technology (NIST), (1993).

[3]Gorunescu F., "Benchmarking Probabilistic Neural Network Algorithms", *International Conference on Artificial Intelligence and Digital Communication,* Research Center for Artificial Intelligence, (2006).

[4]Iyer A., Smolenski B., Yantorno R., Shah J., "Speaker Identification Improvement Using The Usable Speech Concept", *European Signal Processing Conference*, pp. 341-352, (EUSIPO 2004).

[5]Kinnunen T., "Optimizing Spectral Feature Based Text-Independent Speaker Recognition", *Ph.D. thesis*, University of Joensuu, Finland, (2005).

[6]Kinnunen T., "Spectral Features for Automatic Text-Independent Speaker Recognition", *Licentiate's Thesis*, University of Joensuu, (2003).

[7]Mallat S., *A Wavelet Tour of Signal Processing*, Academic Press, New York, (1999).

[8]Martin A., Przybocki M., "NIST's Assessment of Text Independent Speaker Recognition Performance", *The Advent of Biometrics on the Internet*, A COST 275 Workshop in Rome, Italy, Nov. 7-8 (2002).

[9]Martin A., Przybocki M., "The NIST 1998 Development Evaluation of Speaker Recognition on Multi Speaker Telephone Channels", *The Audio-and Video-based Biometric Person Authentication Conference*, Washington D.C, 1999.

[10]Martin A., Przybocki M., "The NIST 1999 Speaker Recognition Evaluation — An Overview", *Digital Signal Processing*, Vol. 10, No. 1-3, pp. 1-18, January/April/July (2000).
(http://dx.doi.org/10.1006/dspr.1999.0355).

[11]Martin A., Przybocki M., "The NIST Speaker Recognition Evaluations, Using Summed Two-Channel Telephone Data for Speaker Detection and Tracking", *Euro Speech Proceedings*, Vol.5, Pages 2215-2218, (1999).

[12]Martin A., Przybocki M., "The NIST Speaker Recognition Evaluations: 1996-2000", *Proceeding of Odyssey Workshop*, Crete, June (2001).

[13]Martin A., Przybocki M., Doddington G., Reynolds D., "The NIST Speaker Recognition Evaluation - Overview, Methodology, Systems, Results, Perspectives", *Speech Communication*, Vol.31, pp. 225-254, (2000).

[14] Rabiner L., Juang B., *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, New Jersey, (1993).

[15]Shao Y. and Wang D-L., "Co-channel Speaker Identification Using Usable Speech Extraction Based on Multi Pitch Tracking", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, pp. 205–208, (2003).