Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 04, April-2018

Speaker Identification by Combining MFCC and Kohonen Neural Networks in noisy Environments

Cătălin Mircea Dumitrescu
Department of System Engineering UPB,
Splaiul Independentei no. 313, Sector 6, Bucharest, 060042,
Romania

Professor, Ph.D. Ioan Dumitrache
Department of System Engineering UPB,
Splaiul Independentei no. 313, Sector 6, Bucharest, 060042,
Romania

***Abstract*— The Mel-Frequency Cepstral Coefficients (MFCC) are widely used as acoustic features for speaker recognition algorithms, but they are not very robust in the presence of additive noise.**

In this paper, we present and test a robust text independent architecture for speaker recognition in noisy environments. The system uses Kohonen neural networks (Self-Organizing Maps - SOM) for speaker modelling and MFCCs as acoustic features. By using the Spearman distance as a metric function for computing the distortion between the known SOM speaker models and the unknown speech input data, we achieve a system recognition rate of 80% in noisy conditions.

We compare the performance of the system for different noise sources and signal to noise ratios.

Keywords*— *Robust speaker recognition; Self-Organizing Maps (SOMs); Mel-Frequency Cepstral Coefficients (MFCCs)

I. Introduction

The speaker identification represents the task of selecting the identity of the speaker from a known population by using their voices, based on the individual acoustic features included in the speech.

There are two categories of speaker identification methods: *text dependent* and *text independent* methods. In *text dependent* systems the speaker is identified by using a specific phrase, like passwords, access code etc. On the other hand, *text independent* methods rely on the characteristics of the speaker's voice for identification.

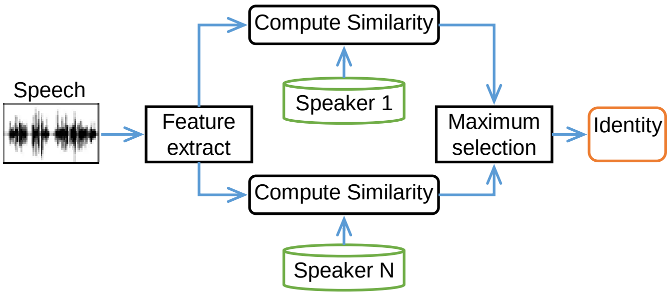


Fig. 1. Speaker identification system

All speaker identification systems contain three modules: acoustic feature extraction module, speaker modelling technique and feature matching module.

The acoustic features extraction module converts the speech signal from a waveform into parametric representation. The speech is a quasi-stationary signal, so when it is examined over a short period of time (usually between 5 – 100 ms), its characteristics are stationary. As a result, a method to characterize the speech signal is the short-time spectral analysis. The most used representation of the short-term power spectrum of a human sound is the Mel-frequency cepstrum (MFC), which represents a linear cosine transform of a log power spectrum on a nonlinear Mel-scale of frequency. Mel-Frequency Cepstral Coefficients (MFCCs) are coefficients that collectively make up an MFC. By using the Mel-scale, which is a perceptual scale of pitches judged by listeners to be equal in distance from one another, the MFC approximates the human auditory system more closely than the normal cepstrum.

This paper describes a speaker identification system that uses as feature the MFCCs combined with DMFCCs and a speaker modelling technique based on SOMs. For determining the similarity, we compute the pair-wise distance between the feature vectors of the speech sample and the known speaker models.

The paper is organized as follows. Section 2 is a presentation for the state of the art speaker identification algorithms and systems. In Section 3 we present the algorithm developed in our research. Section 4 is a case study, in which different parameters are tuned and the proposed algorithm is tested and validated. In this section we determine the optimal frame rate and a noise robust metric. Finally, Section 5 presents or conclusions.

II. State of the Art

The state-of-the-art speaker identification algorithms are based on statistical models of short-term acoustic measurements. The most popular feature extraction methods are: Perceptual Linear Coding (PLC) Linear Predictive Coding (LPC), Mel-Frequency Cepstral Coding (MFCC). Spectral methods, like MFCC, present an interesting property: they mimic the functional properties of the human ear by using a logarithmic scale. And by doing so, methods that use spectral analysis of the speech tends to present better performances.

Modelling techniques used in speaker identification are: Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Support Vector Machines (SVM), Vector Quantization (VQ), Dynamic Time Warping (DTW) and, recently, Self-Organizing Maps (SOMs – Kohonen neural networks).

In order to establish the identity, temple models techniques, like VQ or DTW, directly compare the training against the unknown feature vectors. The distortion between them represents their degree of similarity.

L. Wang et al., in [1], presents a method that combines MFCC and phase information in order to perform speaker identification in noisy environments. The speaker model is computed by combining two GMMs, one for MFCCs and the second one for phase information. By doing so, the error rate is reduced by 20% - 70%. On clean speech, the system has a detection rate of 99.3%, but it drops at 59.7% in a noisy environment with a signal to noise ratio (SNR) of 10dB.

A method that uses Self-Organizing Maps for speaker modelling is presented in [2] by E. Monte et al. The system uses the VQ function of the SOM and its topological property, i.e. the fact that neighbouring code-words in the feature space are neighbours in the topological map. The feature vectors were computed using two methods: LPC and MFCC (24 coefficients for both methods), and the analysis window has a duration of 30ms, with 10ms overlap between them. In the training stage, for each speaker an occupancy histogram is computed and then filtered, using a low pass filter. In the identification stage, the occupancy histogram of the unknown speaker is computed and compared with the known models from the database. Using the relative entropy, the system determines the degree of similarity between the unknown speaker's histogram and the reference.

The system performances were as follows: 98.2% for clean speech using LPC, 100% for clean speech using MFCC, 8% in noisy environment with a SNR of 10dB using LPC and 19.5% in noisy environment with a SNR of 10dB using MFCC.

Another method was proposed by R. Hasan in [3]. It uses MFCC for feature vector extraction and VQ to create the speaker model. The system was tested on clean speech from 21 speakers with different code book sizes and framing windows. It had a performance of 100% recognition rate for a code book size of 64.

III. PROPOSED ALGORITHM

The system that we propose uses the Mel-Frequency Cepstral Coefficients as feature vectors and Self-Organizing Maps for creating the speaker model. The training of the neural network is done using noise free recordings, while the testing of the system is performed using noise corrupted speech files.

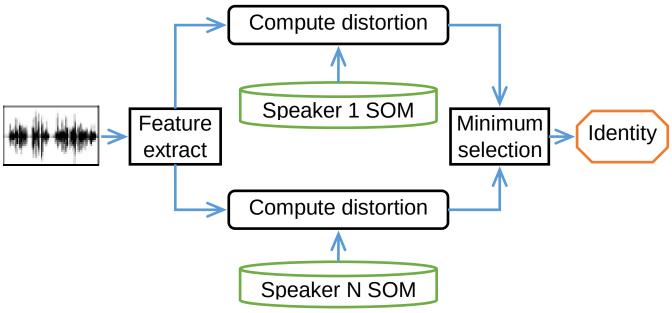


Fig. 2. Proposed algorithm

The algorithm uses the VQ functionality of the self-organizing maps in order to establish the identity of the speaker. In order to establish the identity of the unknown speaker, we compute the mean quantization distortion between the unknown speech input data and all the known speaker models from the database. The identity of the unknown speaker corresponds to the SOM for which the quantization distortion presents the smallest value.

A. Feature extraction

The feature extraction module generates a parametric representation for the speech waveform. In the proposed algorithm, we use Mel-Frequency Cepstral Coefficients (MFCCs) and delta MFCC (DMFCC), as speech features.

There are many implementations of the original MFCC algorithm, introduced in 1980 by B. Davis and P. Mermelstein [4]. They mainly differ in the number of filters, the shape of the filters, the way the filters are spaced, the bandwidth of the filters and the manner in which the spectrum is warped. The algorithm used in this work is the one implemented by Slaney in [5].

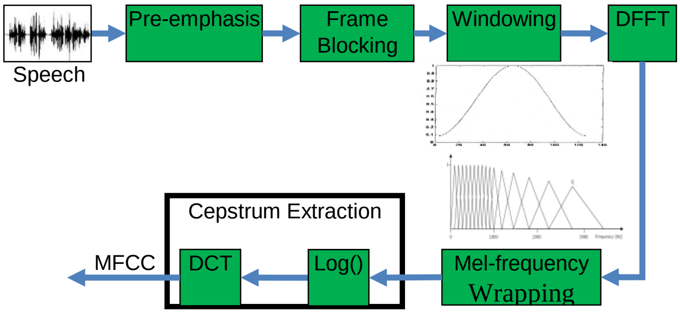


Fig. 3. Block diagram of MFCC extraction

The speech waveform, sampled at 8 kHz 16 bit, is sent to a second order high pass filter, in the pre-emphasis stage. The aim of this stage is to increase the high-frequency components of the human voice that are suppressed during speech production. The used pre-emphasis filter is given by the following transfer function:

$$H(z) = (1 - 0.96 * z^{-1})(1 - 0.97 * z^{-1}) \qquad (1)$$

After that, the signal is blocked into frames of N samples with a frame overlap of M samples. Typical values are 30ms frame width and 10ms overlap. The proposed algorithm uses a 32ms frame, but the optimal overlap between frames is determined through testing.

To minimize discontinuities between frames, a windowing function is applied to each frame. We use the Hamming window in this work:

$$\mathrm{w(n)} = 0.54 - 0.46 \cos\left(\frac{2\pi \mathrm{n}}{\mathrm{N}-1}\right) \qquad (2)$$
$$0 \le \mathrm{n} \le \mathrm{N} - 1$$

After applying the windowing function, each frame is converted from time domain into frequency domain using the DFFT algorithm. The power of the spectrum obtained in the last step is mapped on the Mel-scale using a filter bank. The algorithm uses a filter bank of 40 equal area filters (see **Error! Reference source not found.**), implemented by Slaney in [5]. It covers the frequency range of 133 – 6854 Hz and it consists of:

- 13 linearly spaced filters, range 100 – 1000 Hz, with a step of 133.33 Hz

- 27 logarithmically spaced filters, range 1071 – 6854 Hz, with a logarithmic step of 1.0711703

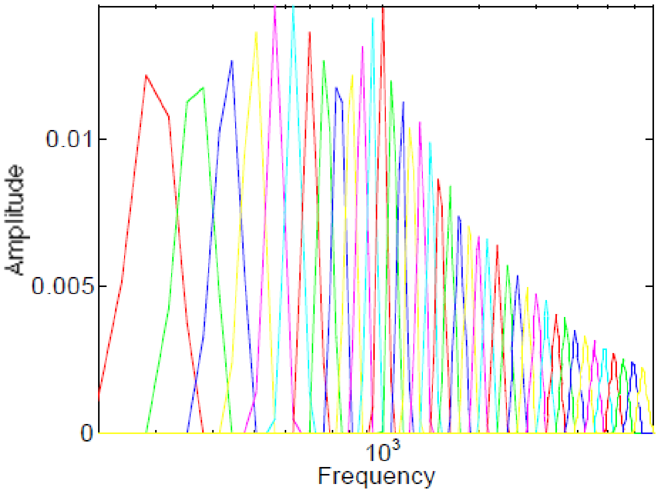


Fig. 4. MFCC filter bank

In the final step, the logarithmic Mel-spectrum is converted back into time domain using the discrete cosine transformation (DCT).

The proposed algorithm uses a feature vector of 64 components, made up of the first 32 MFCCs and their corresponding DMFCCs, in order to add dynamic information for static Cepstral features.

B. *Speaker modeling*

For the speaker model, we use Kohonen neural networks or Self-Organizing Maps (SOMs), so for each of the speakers we train an 8x8 SOM, using the 64 components feature vectors.

Self-Organizing Maps provide a way of representing multidimensional data in much lower dimensional spaces - usually one or two dimensions. They consist of components called nodes or neurons, arranged in 1D, 2D, usually hexagonal or rectangular grid, or higher order (not common) structures. Each node has a weight vector associated with it and a specific topological position in the output map space; the dimension of the weight matches the dimension of the data input vector. Distances between neurons are calculated from their positions with a distance function.

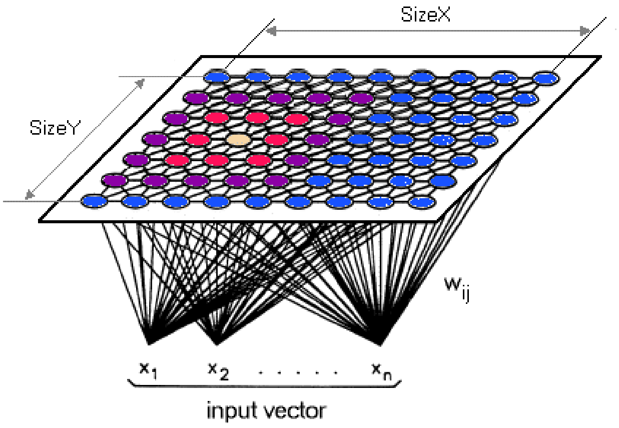


Fig. 5. SOM architecture

The main advantage of using Self-Organizing Maps is that they preserve the topological properties of the input space.

Being a neural network, a SOM needs to be trained. The training is unsupervised and it uses a competitive learning algorithm. At each training step, an input vector is chosen and fed to the network. Using a distance function, a "winning" neuron is identified, and then the weights of the "winning" neuron and its neighbours are updated. Thus, the "winning" neuron and its close neighbours move towards the input vector. The training stage ends after a predefined number of epochs.

The proposed algorithm associates each speaker an 8x8 hexagonal SOM with 64 inputs. As distance function we use "`linkdist`", the number of training epochs is set at 500.

IV. CASE STUDY

In the following section, we are going to determine the optimal frame step for MFCC extraction algorithm, identify a robust metric for computing the similarity between the unknown test speaker and the speaker models from the database and test the proposed algorithm for different type of noise and SNRs.

The database used in this work is The CHAINS corpus [6]. It contains 36 speakers, with 2 recordings for each speaker. The different recording sessions are about two months apart and present different speaking styles. For our algorithm, we train the SOMs using the solo reading recordings, where subjects simply read a prepared text at a comfortable rate. The testing of the system is done using the retell recordings, where, after reading the Cinderella fable in the solo condition, subjects were asked to retell the story in their own words. The sound files are down-sampled from 44.1 kHz to 8 kHz, 16 bit mono PCM.

A. *Determining the optimal frame rate*

In order to compare two Self-Organizing Maps, a performance index must be defined. In this work, we are using the mean quantization error of the SOM as performance index. One of the goals of a SOM is to quantize the input values into a finite number of centroids or output vectors. The quantization error is defined as the distance between an input vector x^j and its nearest centroid. The sum of the quantization error over the input data represents the network distortion:

$$Dis(x, G) = \sum_{j=1}^{N} \min_{G} d(x^j, G^i) \tag{3}$$

As mentioned earlier, we are using the mean value of the distortion as performance index for the speaker model:

$$P(x, G) = \frac{\sum_{j=1}^{N} \min_{G} d(x^j, G^i)}{N} \tag{4}$$

Where: N represents the number of input vectors, d a distance function.

To determine the optimal frame rate, we will generate speaker models using several frame overlaps (12.5 ms, 10 ms, 8 ms, 6 ms, 5 ms, 4 ms and 3 ms) and, afterwards, compute for each of the models, the mean quantization error for the training data. The SOM performance index is computed using the following metrics: Euclidean distance, Cityblock distance, Chebychev distance and Spearman distance.

Given a data matrix $X \in R^{m_x \times n}$, treated as m_x row vectors $x_1, x_2, \dots x_{m_x} \in R^{1\times n}$, and a data matrix $Y \in R^{m_y \times n}$, treated as $m_y \in R^{1\times n}$ row vectors, the distance between the vectors x_s and y_t is defined as:

- Euclidean distance

$$d_{st}^2 = (x_s - y_t)(x_s - y_t)' \qquad (5)$$

- Cityblock distance

$$d_{st} = \sum_{j=1}^{n} |x_{sj} - y_{tj}| \qquad (6)$$

- Chebychev distance

$$d_{st} = \max_j\{|x_{sj} - y_{tj}|\} \qquad (7)$$

- Spearman distance

$$d_{st} = 1 - \frac{(r_s - \overline{r_s})(r_t - \overline{r_t})'}{\sqrt{(r_s - \overline{r_s})(r_s - \overline{r_s})'}\sqrt{(r_t - \overline{r_t})(r_t - \overline{r_t})'}} \qquad (8)$$

Where:

r_{sj} is the rank of x_{sj} take over $x_{1j}, x_{2j,\dots}$
r_{tj} is the rank of y_{tj} take over $y_{1j}, y_{2j,\dots}$

$$r_s = (r_{s1}, r_{s2} \dots, r_{sn})\ and\ r_t = (r_{t1}, r_{t2} \dots, r_{tn})$$

$$\overline{r_t} = \frac{1}{n}\sum_j r_{tj} = \frac{(n+1)}{2}\ \overline{r_s} = \frac{1}{n}\sum_j r_{sj} = \frac{(n+1)}{2}$$

The evolution of the average quantization error for each of the speaker models is presented in figures 6 to 9. By analysing them, we can observe that increasing the frame rate (decreasing the frame overlap), the quantization errors for each metric functions decreases. But with the increase of the frame rate, the number of MFCCs also increases, thus the system's processing time is affected. In order to determine an optimal frame rate, we must define an optimization criterion.

We propose the following optimization criterion:

$$Q = \frac{dP}{dN} \qquad (9)$$

Where: dP represents the variation (in %) of performance index defined in (4); dN the variation (in %) of MFCCs due to frame rate increase

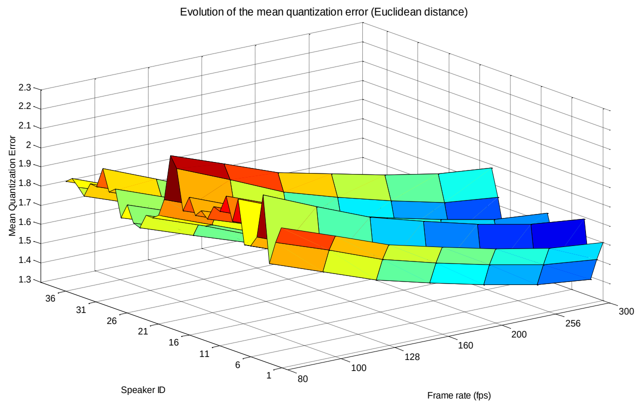


Fig. 6. Quantization error using Euclidian distance

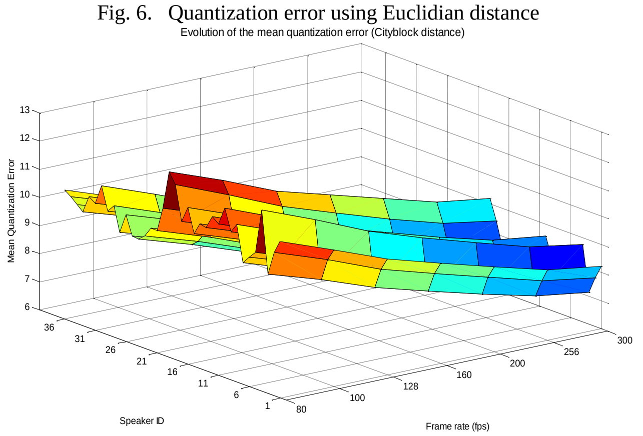


Fig. 7. Quantization error using Cityblock distance

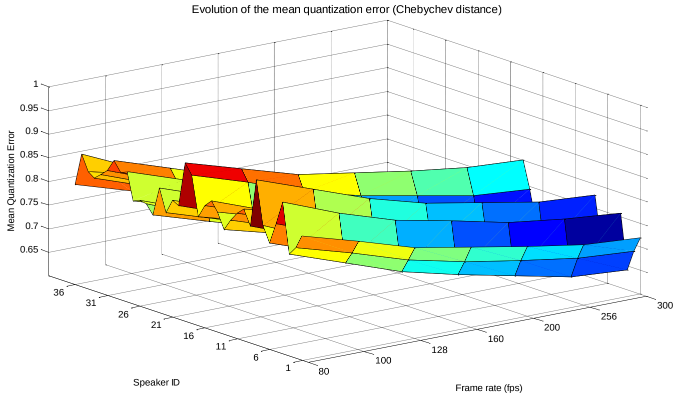


Fig. 8. Quantization error using Chebychev distance

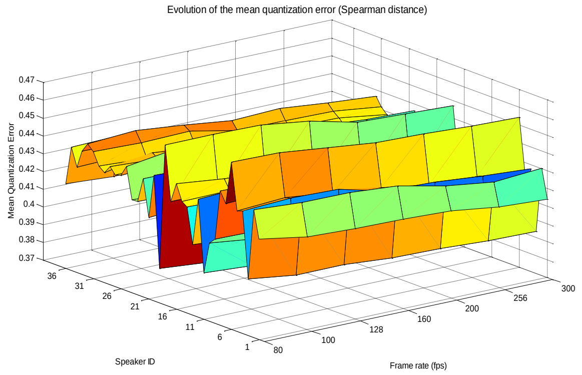


Fig. 9. Quantization error using Spearman distance

The mean quantization error for each speaker model, as can be seen from the previously figures, decreases with the frame rate (for higher frame rates we have a lower quantization error). This fact can also be observed from figure 10, where the evolution of the mean performance index, for all the speaker models, is shown.

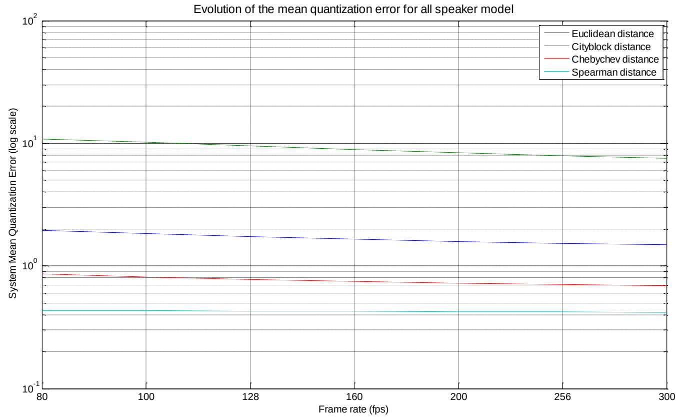


Fig. 10. Optimisation criteria vs. frame rate

From figure 11, which illustrates the evolution of the optimization criterion in relation to the frame rate, we can conclude that the optimal frame rate for extracting the MFCCs is 200 fps (frame overlap of 5ms). The Spearman metric presents a maximum when the frame rate is 200 fps; all the other metrics wave a downward trend, the performance increases slower than the number of Mel-Frequency Cepstral Coefficients.

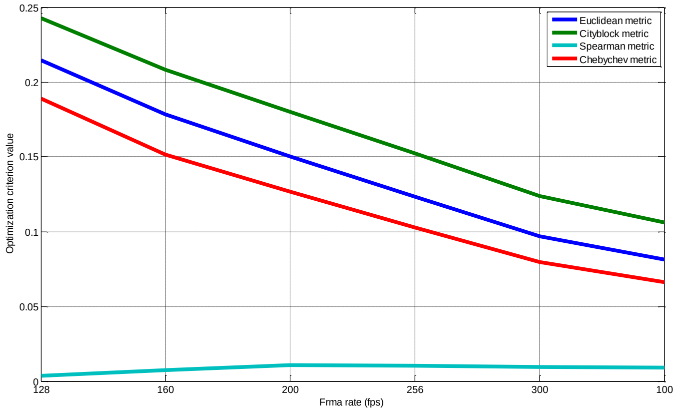
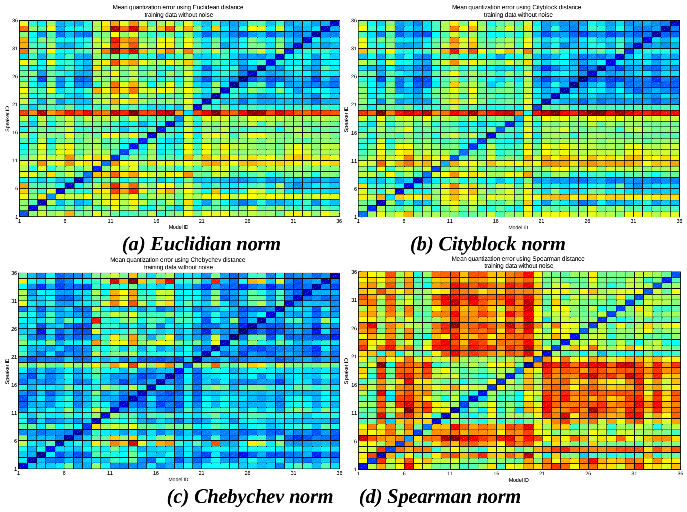


Fig. 11. Optimization criterion evolution

B. Noise robust metric identification

To determine a noise robust metric, we test the speaker models with a noise corrupt version of the training data. Because the CHAINS corpus database is recorded in a sound-proof booth (no noise is present), we have manually altered the training recordings by adding Gaussian white noise with different SNRs (40 dB, 35 dB, 30 dB, 25 dB, 20 dB, 15 dB, 10 dB, 5 dB and 0 dB). In these tests we use a frame width of 32ms and a frame overlap of 5ms, as determined in section A.

By analysing the system outputs, figure 12 (blue represents lower values - red higher values) for the case where the input data is without noise, we can observe that the minimum values for the mean quantization error are obtained when the speech data and the speaker model are from the same person (the minimum values are on the minor diagonal), thus we have a recognition rate of 100 %.

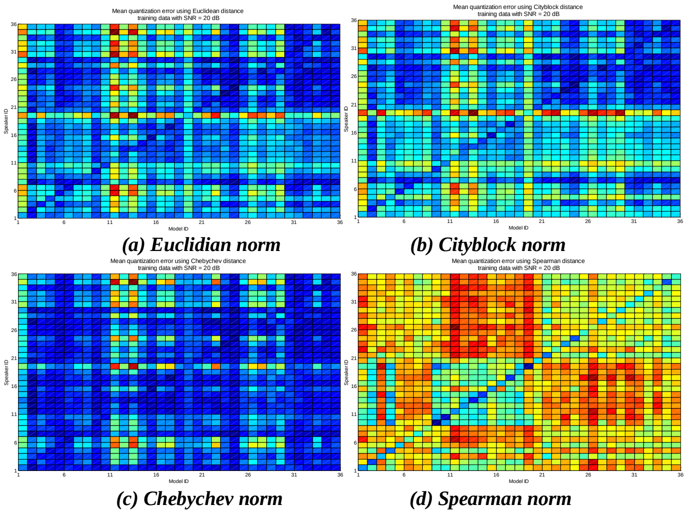


(a) Euclidian norm *(b) Cityblock norm*

(c) Chebychev norm *(d) Spearman norm*

Fig. 12. Systems mean quantization error no noise

By adding noise, the minimum values for the performance index no longer appear on the minor diagonal (where the speaker model and speech sample correspond to the same person), as can be seen from figure 13 where we have the system outputs for a SNR of 20 dB, and so the identification rate decreases.



(a) Euclidian norm *(b) Cityblock norm*

(c) Chebychev norm *(d) Spearman norm*

Fig. 13. Systems mean quantization error SNR 20 dB

By analyzing figure 12(a – d) and figure 13(a – d) we can establish that one of the four norms that are used to compute the performance index, is less influenced by the noise. When the mean quantization error is computed using the Spearman metric as a distance function, the systems present roughly the same distribution of the output values for the two presented cases (no noise figure 12 (d), 20db SNR figure 13 (d)).

The identification rate for a SNR of 20 dB is 97% when Spearman metric is used; for the other metrics the rate is below 50%.

Figure 14 shows the evolution of the identification rate, for all the four metric functions, in relation to the SNR value. It can be easily seen that:

- the performance of the system, when the Chebychev metric is used, it is heavily influenced by the Gaussian white noise
- using the Cityblock metric, the system outperforms the case when the performance index is computed using the Euclidian metric
- the Spearman metric is the least affected by noise.

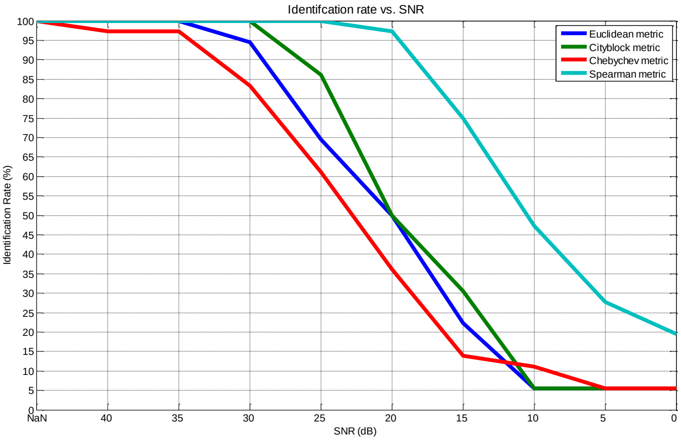


Fig. 14. Identification rate vs. SNR using training data as input

C. *System validation*

We evaluate the system performance using the retell recordings of the CHAINS corpus database and different noise sources (Gaussian white noise, airport noise, restaurant noise and street noise) with multiple SNRs. The Mel-Frequency Cepstral Coefficients are extracted using a frame width of 32ms and a frame overlap of 5ms (section A). The quantization error is computed using the Spearman metric as distance function.

As benchmark, to compare the degradation of the systems performance with noise, we determine the systems mean quantization error using the retell speech files without noise corruption. Afterwards, the input files are mixed with different noise sources using multiple signal to noise ratios. All the files are down-sampled to 8000 Hz, near telephone line quality.

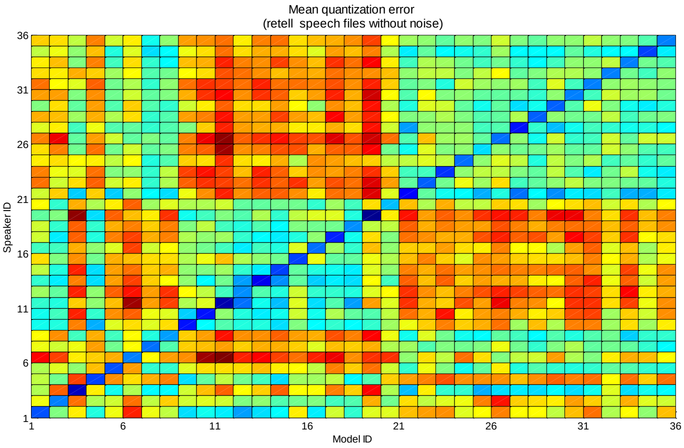


Fig. 15. Mean quantization error for retell recordings (without noise)

Looking at figures 15 and 12 (d), we see that the systems quantization error is nearly the same in the two cases (input data retell sound files and solo reading sound files).

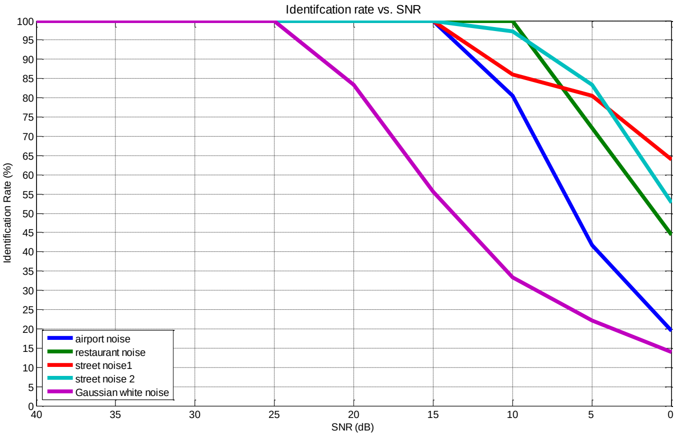
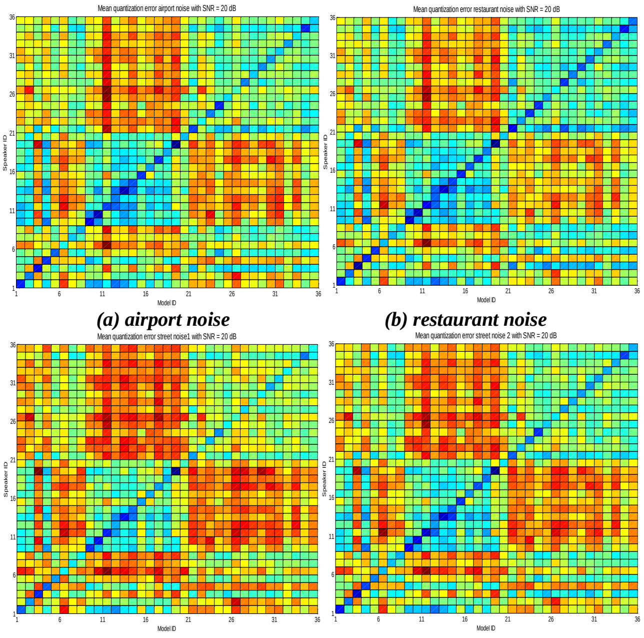


Fig. 16. Identification rate vs. SNR using retells recordings

Figure 16 shows the evolution of the systems identification rate for different types of noise sources and signal to noise ratios. We can see that the system presents a 100% identification rate, even for a SNR value of 25 dB and for all types of noise sources. The system is influenced more by the Gaussian white noise, the identification rate drops at 85% at SNR of 20 dB and the performance decrease is more rapid for this type of noise. In the other cases, the system presents an identification rate of 100% until a SNR of 15dB … 10 dB. The highest rate at 0 dB SNR is nearly 65%, using street noise as corruption source; the lowest value is 15%, when using GWN.

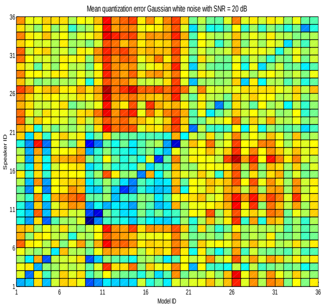


(a) airport noise

(b) restaurant noise

(c) street noise

(d) street noise



(e) Gaussian white noise

Fig. 17. System mean quantization error using retell sound files and SNR 20 dB

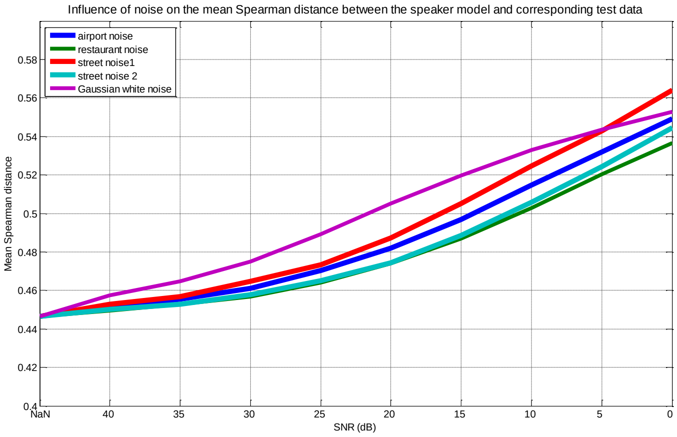


Fig. 18. Influence of noise on the mean Spearman distance between the speaker model and corresponding test data

According to (8), the Spearman distance is one minus the sample Spearman's rank correlation between observations, treated as sequences of values. The rank correlation coefficient is between -1 and 1:

- −1 if the correlation between the two rankings is perfect; one ranking is the reverse of the other.
- 0 if the rankings are completely independent.
- 1 if the correlation between the two rankings is perfect; the two rankings are the same

An increased rank correlation coefficient implies a high correlation between the rankings of the two observations. The two observations, the speaker model and the speech feature vectors, have similar shapes.

V. CONCLUSIONS

Mel-Frequency Cepstral Coefficients (MFCCs) are the most popular speech features, used in speaker identification algorithms, but they are not very robust with noise. The degradation of the MFCC values, due to noise, implies the degradation of the system performance.

Using the Spearman distance, in order to compute the mean distortion between the known speaker models and the unknown speech input data, the systems identification rate is less influenced by noise. This is due to the fact that the Spearman distance is computed on the ranks of the vectors, not their values.

For a low SNR (15 dB … 0 dB), the mean distance is greater than 0.5, which indicates a moderate correlation between the input data (speaker retell recording) and the speaker model.

Analysing the above-mentioned results, we can conclude that the system presented in this work is noise robust and has a high identification rate.

When the input data corruption is performed using real noise sources (like airport, restaurant and street noise), the system outperforms those presented in [1] and [3].

Depending on the noise source and SNR value, the system presents an identification rate greater than 80% (SNR greater than 10dB). The identification drops to 20%, when the speech is corrupted using airport noise with SNR value of 0 dB.

REFERENCES

[1] Wang, L., Minami, K., Yamamoto, K., Nakagawa, S., 2010. Speaker Identification by Combining MFCC and Phase Information in Noisy Environments. Acoustics Speech and Signal Processing (ICASSP), Dallas.

[2] Monte, E., Hernando, J., Miró, X., Adolf, A., 1996. Text Independent Speaker Identification on Noisy Environments by Means of Self Organizing Maps. Fourth International Conference on Spoken Language (ICSLP), vol. 3, pp. 1804 – 1807.

[3] Hasan, R., Jamil ,M., Rabbani, G., Rahman, S., 2004. Speaker Identification Using Mel Frequency Cepstral Coefficients. 3rd International Conference on Electrical & Computer Engineering ICECE.

[4] Davis, S. B., Mermelstein, P., 1980. Comparison Of Parametric Repre-Sentations For Monosyllabic Word Recognition In Continuously Spoken Sentences. IEEE Transactions on Acoustic, Speech and Signal Processing, vol. 28, no. 4, pp. 357 – 366.

[5] Slaney, M., 1998. Auditory Toolbox. Version 2 - Technical Report #1998-010. Interval Research Corporation.

[6] Cummins, F., Grimaldi, M., Leonard, T., Simko, J., 2006. The CHAINS corpus: CHAracterizing INdividual Speakers. Proceedings of SPECOM'06, St. Petersburg.

[7] Faundez-Zanuy, M., Monte-Moreno, E., 2005. State Of The Art On Speaker Recognition. IEEE Aerospace and Electronic Systems Magazine, vol. 20, no. 5, pp. 7-12.

[8] Kohonen, T., Honkela, T., 2007. Kohonen network [Online]. Available: http://www.scholarpedia.org/article/Kohonen_network.

[9] Bodt, E., Cottrell, M., Verleysen, M., 2002. Statistical tools to assess the reliability of self-organizing maps. Neural networks: the official journal of the International Neural Network Society, vol. 15, pp. 967-978.

[10] Ganchev, T., Fakotakis, N., Kokkinakis, G., 2005. Comparative Evaluation of Various MFCC Implementations on the Speaker Verification Task," in SPECOM.

[11] Kumar, S. C., Rao, P. M., 2011. Design Of An Automatic Speaker Recognition System Using MFCC, Vector Quantization And LBG Algorithm. International Journal on Computer Science and Engineering (IJCSE), vol. 3, no. 8, pp. 2942 - 2954.