A Speaker Verification System Using SVM over a Spanish Corpus

Juan Gabriel Pedroza Bernal, Alfonso Prieto Guerrero and John Goddard Close
Electrical Engineering Department, Metropolitan Autonomous University, México.
pedrozafm@yahoo.com.mx, apg@xanum.uam.mx, jgc@xanum.uam.mx.

Abstract

This paper presents a description of the principal aspects employed in the development of a speaker verification system based on a Spanish corpus. The main goal is to obtain classification results and behavior using Support Vector Machines (SVM) as the classifier technique. The most relevant aspects involved in developing a Spanish corpus are given. For the front end processing a novel method to suppress silences between words is proposed and successfully applied. The validation to the complete system is made using randomly selected feature vectors and vectors from continuous sequences of the voice signal. Additionally, Gaussian Mixtures Models (GMM) and Artificial Neural Networks (ANN) are also used as classifiers to compare and validate the results.

1. Introduction

A speaker verification system consists of four modules: acquisition, processing, classification and decision [1]. In acquisition, the speech is sampled in order to get a discrete representation of the sound wave. Next, the signal is analyzed and processed in segments to get a numerical representation of each segment (a set of vectors). It is expected that such a numerical representation is unique for every speaker. The classifier module builds one model for each voice and one model for the complement or rest of the speakers (Background). When a sample of voice is given to the verification system, it is processed and compared with the model of the hypothetical claimant speaker. Then the decision module determines, if it is possible with the largest probability, if the sample corresponds to the speaker or not and takes a decision: accept or reject. The paper describes the main aspects involved in designing and creating a Spanish corpus of speech utterances. It then presents the speech processing employed using Mel Frequency Cepstral analysis (MFCC) and the SVM classifiers chosen to build models for every speaker and their background. Finally, we present tests and the results obtained and propose some work can be done in the future.

2. The Spanish corpus

The main challenge in the use of speech utterances as a biometric pattern in classifiers is the variability of the speech in the time and intensity scales [2]. This variability is due to intrinsic and external speaker factors. Some intrinsic factors are the physical condition of the vowel tract and the nervous and mental state of the speaker. External factors are the noise present in the environment, quality of the recording equipment, and other factors which modify the signal-to-noise ratio (SNR), such as the distance between the speaker and the acquisition device. So there are two questions that should be answered before building a corpus: what utterances should be recorded, and, how should they be recorded. In the next section some answers to these questions are given.

2.1. Recording protocol

There is no unique answer to what utterances should be recorded, but we know that a good phonetic representation is necessary [3]. This means that if a fundamental sound is omitted the probability error in the associated verification system will increase, because it will not contain enough information to get a good model of the voice. Based on these ideas we used the following utterances for the recording protocol of the corpus, which was called *MCyTI*:

a) The name of the speaker.
b) Pronunciation of the ten isolated digits.
c) Pronunciation of two chains of five isolated digits randomly generated for all speakers.

d) Pronunciation of two chains of five isolated digits randomly generated for each speaker.
e) Pronunciation of two fifteen-word-phrases randomly chosen for every speaker.
f) Pronunciation of a syllabically and phonetically balanced phrase for each speaker.
g) Repetition of task e) increasing the speed of pronunciation.
h) Repetition of task e) decreasing the speed of pronunciation.

The phrase used in task f) was chosen after a syllabic and phonetic analysis of several phrases. The goal of these analyses was to find a phrase with a similar phonetic and syllabic distribution to that found in the Spanish language [3]. This assures the inclusion of almost all the fundamental sounds of Spanish together with the frequency with which they are spoken. Figures 1 and 2 show the phonetic and syllabic distribution of the phrase chosen in task f).

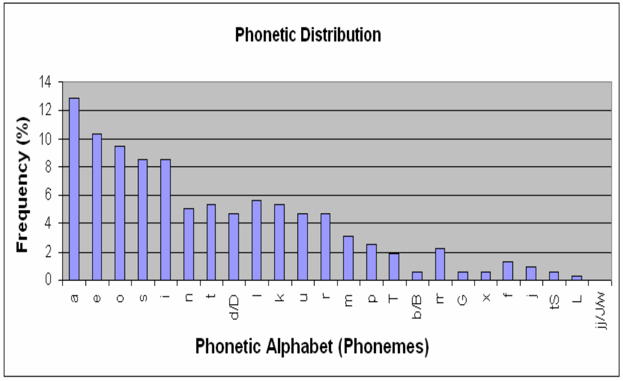


Figure 1. Phonetic distribution of the task f) in the recording protocol, this distribution includes the 28 sounds established by SAMPA for the Spanish. In some cases the phonemes were analyzed together.

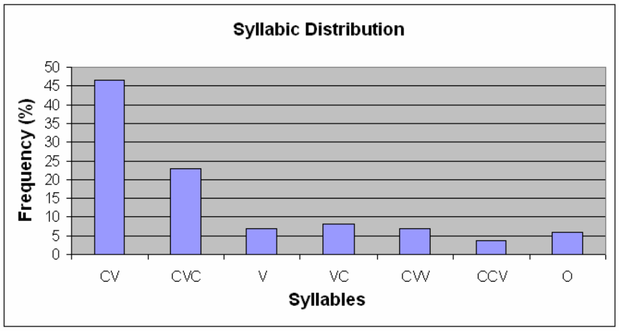


Figure 2. Syllabic distribution of task f) in the recording protocol. The C label represents a consonant sound and the V label represents a vowel sound. In the O label are included all other kinds of syllabic combinations.

2.2. Recording sessions

As to how the speech should be recorded, it is necessary to include in the recorded speech samples as much noise as that found in the environment in which the speaker verification system will be eventually used. Preferably, the kind of noise should be the same. It was considered that the *MCyTI* corpus should manage the noise of an office environment, so the recording sessions were conducted in a low noise office (SNRmax=28dB). We did not care about sound disturbances or other kinds of sound influences encountered during the recording sessions. The acquisition was made using a medium quality PC microphone at a sampling rate of 8000Hz. The distance between the speaker and the microphone was not controlled once it was initially adjusted at the beginning of the session. All speech samples were acquired in a single session for each speaker but at different dates and times. There were no gender restrictions for participation in the corpus. In this first version of the corpus, 17 speakers were included, 11 males and 5 females. The total length of the sound registers was between 110 and 150 seconds.

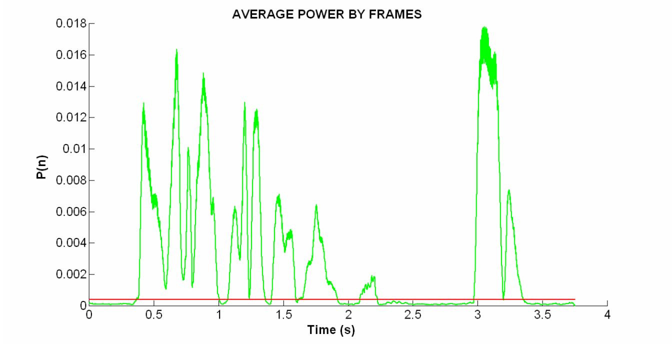
3. Speech processing

The ideal of the speech processing is to obtain a numerical representation of the voice of every speaker which is unique. A successful technique used to achieve this goal is Mel Frequency Cepstral analysis [4,5] which provide the so called Mel Frequency Cepstral Coefficients (MFCC). This analysis is based on the extraction of the frequency components from speech segments or frames. The first stage of the processing consists in applying a FIR filter in a band of 300-3300Hz to eliminate low frequency noise and limit the high frequency spectrum. The next stage is a silence suppressor. The goal of this stage is drop the pauses between words and phrases. The suppressor takes a sample of noise of r_i seconds at the beginning of each sound signal $S_i(n)$. This sample is used to find the mean and variance of the noise power $S_i^2(n)$ in the environment. Then the average power by frame of the signal is calculated using the Equation 1:

$$P_i(n) = \sum_{k=n}^{n+M} S_i^2(k), \qquad \textbf{(1)}$$

where M is the length of each frame. Figure 3 shows the effect of applying Equation 1 to the first 4 seconds of $S_1(n)$.

Figure 3. Graph of the average power by frame of the signal *S*$S_i(n)$. The horizontal line can be used as a threshold to drop silences or pauses between words.



A threshold u_i can be estimated as a maximum level of noise power using the following equation:

$$u_i = \overline{N_i} + c_i\sqrt{\sigma^2_{Ni}}, \qquad \textbf{(2)}$$

where c_i is a positive factor which can be set to improve the results of the suppression stage. The threshold u_i is used to discriminate points of $S_i(n)$ which correspond to noise according to the following criterion:

$$\text{If } P_i(n) \geq u_i \text{ then } S_i(n) \text{ is a voice point.} \qquad \textbf{(3)}$$

$$\text{If } P_i(n) < u_i \text{ then } S_i(n) \text{ is a noise point.}$$

Once the pauses have been suppressed from $S_i(n)$, the resultant signal $S_i(n)_{voice}$ is divided into frames of 30ms with an overlap of 10ms between them. Each frame is transformed using the FFT with 512 points and then the canonical norm is applied to the complex components. It gives a set of symmetric values which represent the frequency components of the frame, so only the first 256 are considered. In order to improve the resolution of each spectrum, a Mel filter bank is applied. This bank consists of 31 triangular filters overlapped and distributed along the frequency scale according to the equation:

$$f_{MEL}(k) = 1000\frac{\log\left(1+\dfrac{f_{lin}(k)}{1000}\right)}{\log(2)}. \qquad \textbf{(4)}$$

In Equation 4, $f_{lin}(k)$ corresponds to the frequencies of the 31 filters equally spaced in the band 0 to 3300Hz. After that a *20log* transformation is applied which converts the results to the dB scale. In this way, a 31-dimensional vector V is obtained for each frame which is called the Spectral vector. Finally, the Cepstral Transformation is applied to the set of Spectral vectors as in Equation 5,

$$x(n) = \sum_{k=1}^{31} V(k)\cos\left[\frac{n\pi}{31}\left(k-\frac{1}{2}\right)\right], n = 1,..,L, \qquad \textbf{(5)}$$

where $V(k)$ is the k component of the vector V and L is the number of Cepstral coefficients to be calculated. In this case we set $L=13$, the final length of the feature vectors.

4. Classification using SVM

In a speaker verification system two random variables can be considered; one represents the voice to be identified, the second one represents the voices of the rest of the speakers. When a speaker claims to be an authorized one, the system has to decide whether this is true or not. Hence, the classifier module requires a representation of the voice of each valid user. Different classification techniques, such as SVM [6], GMM [7] and ANN [8], provide alternate ways to construct a model from a voice.

4.1. Support vector machines

Suppose we have $C_1=\{x_1, \dots , x_k\}$, $C_2=\{x_{k+1}, \dots , x_l\}$ two disjoint subsets of points in $\Re^n$ which are called classes and which were generated from two different random variables. The goal is find a hyper-plane determined by a vector w, which divides the space in such way that:

$$w \cdot x_i + b > 0, \forall x_i \in C_1, \qquad \textbf{(6)}$$
$$w \cdot x_j + b < 0, \forall x_j \in C_2.$$

In this case we call the two sets linearly separable. It can be shown that if such a hyper-plane exists it is not unique, so some criterion is required to select w. Firstly we associate a value $y_i \in \{1,-1\}$ to each vector x_i to denote the class it belongs to. It is supposed the vector w satisfies:

$$w \cdot w + b = -\varepsilon, \qquad \textbf{(7)}$$

$$w \cdot \left(w + \frac{w}{\|w\|}m\right) + b = \varepsilon. \Rightarrow m = \frac{2\varepsilon}{\|w\|}, \varepsilon > 0.$$

and that we wish to maximize the margin m. In this case the norm of the vector w must be minimized.

Then the classification problem can be stated as the following optimization problem with constraints:

$$\min\left\{ f(w) = \frac{1}{2\varepsilon}\|w\|, w \in R^{n} \right\}, \quad \textbf{(8)}$$
$$y_i\left(w \cdot x_i + b\right) \geq \varepsilon, i = 1,\ldots,l,$$
$$\varepsilon > 0, b \in R.$$

A vector *w* that satisfies Equation 8 is called a support vector for the classes. The most interesting case is when a solution does not exist for the problem, for then the sets are not linearly separable. A procedure to manage this case is the use of a function ϕ which maps the classes to $\Re^{m}$, $m \geq n$ or $m = \infty$. The intention of this mapping is to transform the problem into a linearly separable or, in the worst case, decrease the number of vectors classified incorrectly. Instead of working directly with the function ϕ, a kernel function of two variables is introduced which is related to ϕ through the inner product in $\Re^{m}$. The goal is to obtain a non-linear separation of the classes. A vector *x* is then essentially classified according to the region it belongs to, although viewed in $\Re^{m}$ this corresponds to the side of the hyper-plane its image falls in. Several kernel functions have been proposed in classification tasks and one widely used is the Radial Basis Function (RBF) [9] defined by:

$$K(x,y) = C\exp\left(-\gamma\|x-y\|^{2}\right), C \in \Re. \quad \textbf{(9)}$$

In this paper we use it to investigate its behavior in speaker verification tasks over the *MCyTI* Spanish corpus.

5. Tests and results

After the suppression of silences, discussed in section 3, the 17 sound registers were reduced to between 33 and 39 seconds each. To have uniform sets the first 32.15s of each file was processed which gave us 1606 feature vectors for each speaker. They were included in a set T_i and from it a training set E_i was formed with 1252 vectors and two validation sets: P_i^{1} with 252 vectors and P_i^{2} with 102 vectors. Each vector was uniformly and randomly selected from the total set T_i. The background B_i for each user U_i was made by picking the first 200 vectors of each T_j, $j \neq i$. Here, the feature vectors from U_{16} and U_{17} were not considered because they were used to validate the system with unknown claimants. After some tests *C* and γ were set to $C=32$ and $\gamma=1$. The sets $E_i \cup B_i$ were normalized to the interval *[-1,1]* to generate a model M_i considering the max and min values of each entrance of the vectors separately.

Three sets of tests were developed: In the first one the randomly generated sets P_i^{1} and P_i^{2} were classified with every model M_j, $i, j = 1,..,15$. In the second one a set of vectors R_i generated from continuous sequences of voice were used. Finally the sets P_i^{1} were used to classify with GMM and ANN.

5.1. Classifying randomly selected vectors

The randomly generated sets P_i^{1} and P_i^{2} were classified with every model M_j, $i, j = 1,..,15$, then the percentage of vectors assigned to the model of each user was obtained. When P_i^{1} was classified with M_j and $i=j$ the percentage obtained is a value which represents the ability of the classifier to recognize the user. These values are known as True Scores. When P_i^{1} is classified with M_j and $i \neq j$ the percentage obtained signifies the error of the classifier to reject invalid claimants. These scores are known as False or Impostor Scores. The distribution obtained for True and False Scores shows that the best percentage of classification is achieved when P_i^{1} or P_i^{2} are classified with M_i in all cases. The Table 1 shows the scores were obtained for both sets.

Table 1. Best percentages of classification for the randomly generated sets P_i^{1} and P_i^{2} against the models M_j. The table shows the variation when the number of classification vectors is reduced. The Impostor Scores are lower than 40% in all cases.

Model M_i	**True Scores**	
	P_i^{1}(252 vectors)	P_i^{2} (102 vectors)
M_1	71.428	71.568
M_2	82.142	67.647
M_3	70.634	63.725
M_4	68.254	64.705
M_5	58.330	62.745
M_6	55.555	53.921
M_7	51.873	50.000
M_8	62.301	71.568
M_9	67.857	69.607
M_{10}	58.333	57.843
M_{11}	63.095	61.764
M_{12}	70.634	56.862
M_{13}	55.158	60.784
M_{14}	62.698	67.647
M_{15}	74.603	82.352

5.2. Classifying continuous sequences of voice

To get the performance of the SVM system with real time utterances, a test with vectors extracted from continuous sequences was made. The signal containing the name of the speaker (task a) was extracted and processed for each user U_i, $i=1,\ldots,17$ and the sets of feature vectors obtained were denoted by R_i. Again, we classify every set R_i with each model M_j, $j=1,\ldots,15$. The main aspects in this test are that the system is dealing with utterances with different lengths (and so with a variable number of vectors from 29 to 97), and with claimants never included in the system. The performance was compared using the scores obtained in this test against those from 5.1 by means of DET curves. The results are shown in Figure 4.

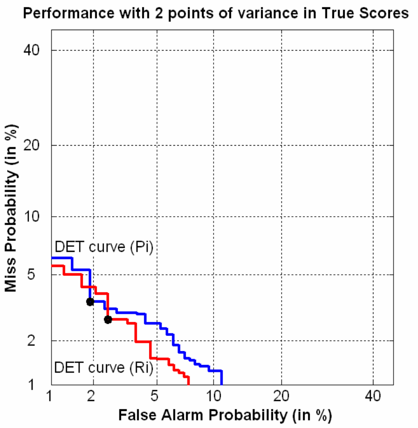


Figure 4. Performance of the system using randomly generated sets P_i^1 and continuously generated sets R_i with a variable length. A variance of 2 is applied to True Scores. The curves show a similar behavior and very close min DCF points.

The results of the minimum detection cost function points (min DCF points) and the associated performance for the DET curves in Figure 4 are shown in the Table 2.

Table 2. Results of the minDCF and Performance of the SVM system using randomly generated sets of vectors against real time utterances and two speakers not included in the system. The average length of the utterances was 2.695s and the average time taken to process them was 5.406s.

True Scores	**Classification using P_i^1**		**Classification using R_i**	
Variance	minDCF	Performance	minDCF	Performance
10	0.0262	97.38%	0.0390	96.10%
4	0.0247	97.53%	0.0378	96.22%
2	0.0193	98.07%	0.0250	97.50%

5.3. Classifying using GMM and ANN

To validate the results obtained in sections 5.1 and 5.2 two commonly used classification techniques were also employed: Gaussian Mixture Models (GMM) and Artificial Neural Networks (ANN). More details on these methods can be found in [7] and [8]. In this paper only the most relevant parameters are mentioned.

For the GMM classifier 10 Gaussians were used and 25 iterations of the EM algorithm conducted to find the mean vectors and covariance matrices. The vector sets were not normalized because it was found that this decreased the percentage of vectors correctly classified. For ANN, 15 feed forward backpropagation nets were constructed with 13 input units, 4 hidden units and 1 output unit. All units had a logistic activation function and the identity output function. Several learning factors η_i, thresholds θ_i and iterations were experimented with to improve the results obtained. As with the SVM, all the sets of vectors were normalized into the interval $[-1, 1]$. All the True and False scores were obtained using the P_i^1 sets. To generate the DET curves the values of $p(False)=0.985$, $p(True)=0.015$ were assigned, and a penalization of 10 to 1 for a False Alarm with respect to a Misclassification. The results we obtained for the three classifiers are shown in Figure 5 and in Table 3.

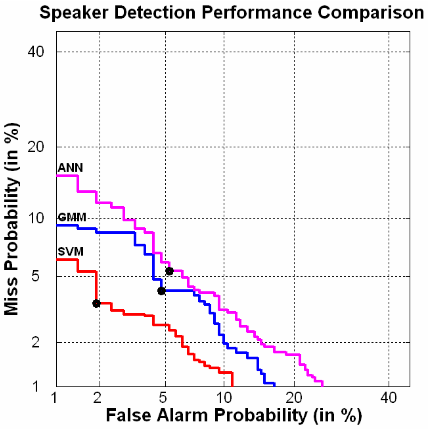


Figure 5. Comparison between DET curves for the detection performance of SVM, GMM and ANN as classifiers over the *MCyTI* Spanish corpus. These graphs were generated applying a variance of 2 points over the True Scores sets and using the randomly generated sets P_i^1. The min DCF point for SVM is closer to the origin which shows a better performance.

Table 3. Comparison between the min DCF values and performances of the SVM, GMM and ANN classifiers with validation sets P_i^1 randomly selected.

True Scores	**GMM**		**ANN**	
Variance	minDCF	Performance	minDCF	Performance
10	0.0499	95.01%	0.0687	93.13%
4	0.0484	95.16%	0.0672	93.28%
2	0.0475	95.25%	0.0524	94.76%

6. Conclusions

A methodology to develop an automatic speaker verification system based on Support Vector Machines was given. It was validated using randomly selected feature vectors and sequences of voice in real time. The results were compared to those obtained using Gaussian Mixture Models and Artificial Neural Networks as classifiers. The guidelines to make a Spanish corpus and the speech processing for each sample based on MFCC analysis were given. Here the use of a novel suppressor of pauses or silences between words was proposed and successfully proved. The min DCF points and the Performances reported in Table 3 show that the speaker verification system based on SVM is slightly more reliable than GMM and ANN in at least 2%. So it can be concluded that the best of the three classifiers is SVM with more than 97% optimal efficiency. This conclusion is valid for the parameters assigned to GMM and ANN and does not imply their scores cannot be improved. The validation made with real time utterances, which is reported in Table 2, shows that the SVM system can operate with real time speech and so the models for each user are text-independent. It means the models are representative and include information of almost every fundamental sound and their combinations for the Spanish language.

7. Future work

It is proposed investigate the behavior of the system when the speech processing parameters are changed. Mainly, the time of each frame and the overlap time, the resolution of the MEL filter bank, and the number of Cepstral coefficients which determines the length of the feature vectors. These variations are intended to improve the numeric separation between the True and False Scores, and decrease the variance of the distributions which can decrease the error in the classification tasks. The *MCyTI* corpus will be enlarged to include more users. Some of them will be used to develop invalid claimant tests in a wide sense to confirm the results. Some samples of voice will be used to complement the background of each user. It would decrease the false alarm errors. An important future work is to test the system with SVM classifiers using other kernel functions, in order to compare the results and obtain conclusions about their performance. Reducing the average processing time and investigating, in a wider sense, the behavior of the system with real time utterances are tasks that will increase its reliability.

8. Acknowledgements

We thank the UAM for sponsoring us during the development of this investigation for the past year.

9. References

[1] F. Bimbot, J.F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovzka-Delacrétaz, D.A. Reynolds, *A Tutorial on Text-independent Speaker Verification*, EURASIP Journal on Applied Signal Processing 4, pp. 430-451, 2004.

[2] J.P. Campbell Jr., *Speaker Recognition: A Tutorial*, IEEE Proceedings, Vol. 85, No. 9, September 1997.

[3] J. Ortega-García, J. González-Rodríguez and V. Marrero-Aguiar, *AHUMADA: A large Speech Corpus in Spanish for Speaker Characterization and Identification*, Speech Communication, Vol. 31, pp. 255-264, 2000.

[4] S. Furui, *Cepstral Analysis Technique for Automatic Speaker Verification*, IEEE Transactions on Acoustic, Speech and Signal Processing, Vol. ASSP-29, pp. 254-272, 1981.

[5] M. Seltzer, *SPHINX III Signal Front End Specification*, CMU Speech Group, 1999.

[6] S. Raghavan, G Lazarou and J. Picone, *Speaker Verification Using Support Vector Machines*, Center for Advanced Vehicular Systems, Mississippi State University, 2006.

[7] D. Reynolds and R. Rose, *Robust Text-independent Speaker Identification Using Gaussian Mixture Speaker Models*, IEEE Transactions on Speech Audio Processing, Vol. 3, No. 1, pp. 72-83, 1995.

[8] A.K. Jain, J. Mao and K.M. Mohiuddin, *Artificial Neural Networks: A Tutorial*, IEEE Transactions, Vol. 29, pp. 31-44, 1996.

[9] H. Chi-Wei, Ch. Chih-Chung, L. Chih-Jen, *A Practical Guide to Support Vector Classification*, http://www.csie.ntu.edu.tw/~cjlin/papers/guide.

[10] D. A. Reynolds, T.F. Quatieri, R.B. Dunn, *Speaker Verification Using Adapted Gaussian Mixture Models*, Digital Signal Processing, Vol. 10, pp. 19-41, 2000.