SPEAKER RECOGNITION EXPERIMENTS ON A BILINGUAL DATABASE

Marcos Faundez-Zanuy, Antonio Satué-Villar

Escola Universitària Politècnica de Mataró, Universitat Politècnica de Catalunya (UPC)
Avda. Puig i Cadafalch 101-111, E-08303 Mataró (BARCELONA), SPAIN
e-mail: {faundez, satue}@eupmt.es http://www.eupmt.es/veu

ABSTRACT

This paper presents some speaker recognition experiments using a bilingual speakers set (49), in two different languages: Spanish and Catalan. Phonetically there are significant differences between both languages. These differences have let us to establish several conclusions on the relevance of language in speaker recognition, using two methods: vector quantization and covariance matrices.

1. INTRODUCTION

This paper deals with speaker recognition [1] (identification and verification) with fully bilingual speakers. Thus, we extend our previous results published in 1999 [2].

We have done a set of experiments with a bilingual database in order to establish if the language of the speaker has relevance in a speaker identification and verification application (mainly if it is more suitable one language than other, and if it is possible to recognize with different training and testing languages).

Phonetically there are significant differences between both languages. Mainly, the Catalan language has eight vowels (see figure 1) and Spanish only five. Although there are only nine million people of Catalan speakers in front of four hundred million people of Spanish, both languages can be used for our purpose. The differences between both languages have let us to establish several conclusions on the relevance of language in speaker recognition.

Another important question is that for bilingual speakers in conversational speech is quite common the change from one language to the other, so it is interesting to evaluate if this fact can affect a speaker recognizer.

For these experiments we have used our previous database [3]. An interesting fact is that the Spanish sentences have been balanced, but the Catalan ones have been merely translated from Spanish. Thus, the database consists of the same texts recorded in both languages in the same day, one language after the other. Speaker could freely choose which the first recording language was.

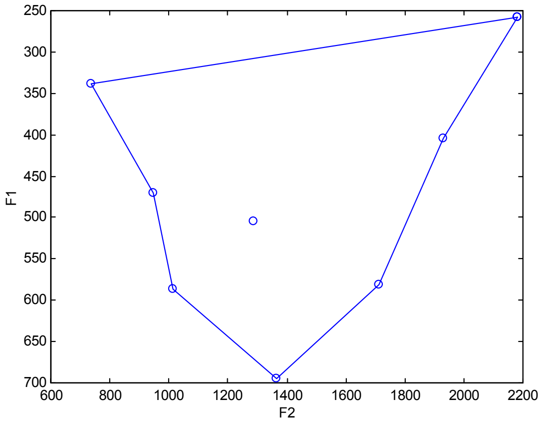

Figure 1: Formants of the Catalan vowels

2. DATABASE

The main characteristics of the database are:

- 4 sessions with different tasks in each session (isolated numbers, connected numbers, sentences, text, conversational speech, etc.)
- In each session, tasks were sequentially collected in both languages (Catalan and Spanish), uttered from the same speaker. Each task was simultaneously acquired with two microphones (SONY ECM-66B and AKG C-420).

This paper presents results of the fourth session using the common text (aprox. 1 min) and the first five sentences (approx, 4 seconds lasting each one).

3. SPEAKER RECOGNITION EXPERIMENTS

With this database we have made several tests:

- Speaker recognition with each language: train and test in Catalan (CC), train and test in Spanish (SS)
- Speaker recognition with different train and test conditions: train in Catalan and test in Spanish (CS), train in Spanish and test in Catalan (SC).

Two speaker recognition methods have been used:
1. Vector quantization [4] with LBG [5] algorithm for codebook generation (1 codebook for each speaker). The number of parameters used in each model is:

$$parameters = 2^{No} \times P \qquad (1)$$

where P is the analysis order of the parameterization (dimension of LPCC vectors) and No is the number of bits of the codebook ranging from 0 to 8.
2. Arithmetic-harmonic sphericity measure [6], which implies the computation of a covariance matrix for each speaker, and the following measure distance:

$$\mu\left(C_j C_{test}\right) = \log\left[tr\left(C_{test}C_j^{-1}\right)tr\left(C_j C_{test}^{-1}\right)\right] - 2\log\left(P\right) \quad (2)$$

where C_j and C_{test} are covariance matrices and P is its size.

The trace of the matrices can be computed as:

$$tr\left(YX^{-1}\right) = 2\sum_{i=1}^{P}\sum_{j=1}^{i-1} y_{ij}\tilde{x}_{ij} + \sum_{k=1}^{P} y_{kk}\tilde{x}_{kk} \qquad (3)$$

where, x_{ij} , $\tilde{x}_{ij}$, y_{ij} , $\tilde{y}_{ij}$ are respectively the elements of the matrices X, X^{-1},Y and Y^{-1}.

The number of parameters for each speaker is (the covariance matrix is symmetric):

$$parameters = \frac{P^2 + P}{2} \qquad (4)$$

Although these methods are not the state-of-the-art in speaker recognition, they require lower computational time than GMM. On the other hand, we are interested on relative comparisons, the recognition algorithm not being a critical issue.

4. RESULTS

The results have been obtained with the following parameters:

- 49 bilingual speakers.
- 1 read text (about 1 minute and the same text for all speakers) for computing the models.
- 5 different test sentences (the same for all speakers).
- *No*=number of bits of the codebook, from 0 to 8.
- Silence removal
- Frames of 240 samples with an overlap of 2/3.
- Hamming window and pre-emphasis of 0.95.

We have evaluated the identification and verification results. Verification systems can be evaluated using the False Acceptance Rate (FAR, those situations where an impostor is accepted) and the False Rejection Rate (FRR, those situations where a user is incorrectly rejected), also known in detection theory as False Alarm and Miss, respectively. This framework gives us the possibility of distinguishing between the discriminability of the system and the decision bias. The discriminability is inherent to the classification system used and the discrimination bias is related to the preferences/necessities of the user in relation to the relative importance of each of the two possible mistakes (misses vs. false alarms) that can be done in verification. This trade-off between both errors has to be usually established by adjusting a decision threshold. The performance can be plotted in a ROC (Receiver Operator Characteristic) or in a DET (Detection error trade-off) plot [7]. DET curve gives uniform treatment to both types of error, and uses a scale for both axes, which spreads out the plot and better distinguishes different well performing systems and usually produces plots that are close to linear. DET plot uses a logarithmic scale that expands the extreme parts of the curve, which are the parts that give the most information about the system performance. For this reason the speech community prefers DET instead of ROC plots.

We have used the minimum value of the Detection Cost Function (DCF) for comparison purposes. This parameter is defined as [7]:

$$DCF = C_{miss} \times P_{miss} \times P_{true} + C_{fa} \times P_{fa} \times P_{false} \qquad (5)$$

where C_{miss} is the cost of a miss (rejection), C_{fa} is the cost of a false alarm (acceptance), P_{true} is the a priori probability of the target, and P_{false} = 1 − P_{true}. C_{miss}= C_{fa} =1.

4.1 Vector quantization results

Table 1 summarizes the results for a vector quantization speaker identification method, with parameterizations LPCC-12, 16 and 20, and for codebooks ranging from 0 to 8 bits.

Although vector quantization (VQ) performs well for identification task, reaching identification rates up to 100%, the verification task is not so successful, and it is outperformed by the covariance matrices (CM) method.

4.2 Covariance matrices

The parameter that can be adjusted for modeling the speakers is the prediction order (*P*). That is, the dimension of the LPCC vectors.

We have studied several P values (table 2). It is important to see that a frame length of 240 samples is used, so for a correct LPC parameter estimation, the prediction order must not be higher than 24, because then the autocorrelation used in the Levinson-Durbin recursion can not be properly estimated. For this reason, the recognition rates drop for high *P* values.

Another important fact is that a covariance matrix assumes that the modeled distribution is symmetrical. This assumption is not made in the VQ approach. Thus, for nonsymmetrical distributions the VQ approach could be more accurate.

No	Num. Parameters (aprox.)	P
0	12	4
1	24	6
2	48	9
3	96	13
4	192	19
5	384	27
6	768	39
7	1536	55

Table 3: # of parameters used in VQ (P=12) and CM

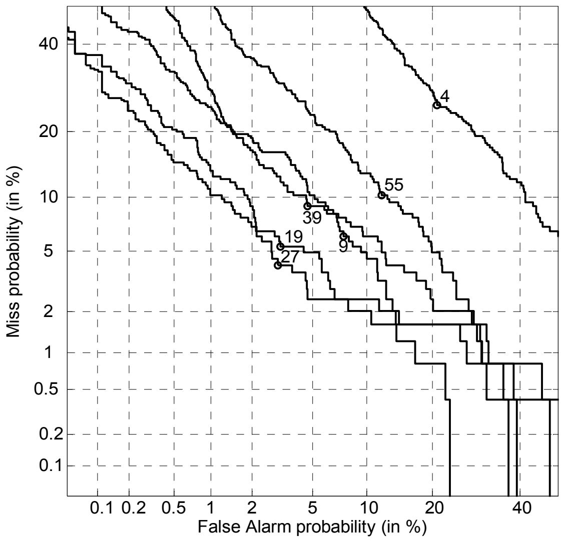

Figure 2. DET plots for Covariance matrices of sizes 4, 9, 19, 27, 39 and 55.

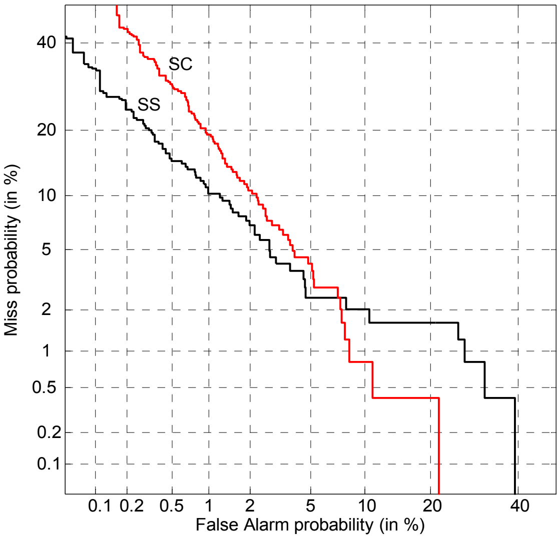

Figure 4. DET plots for Covariance matrices of size 27 and two training/testing scenarios (SC and SS).

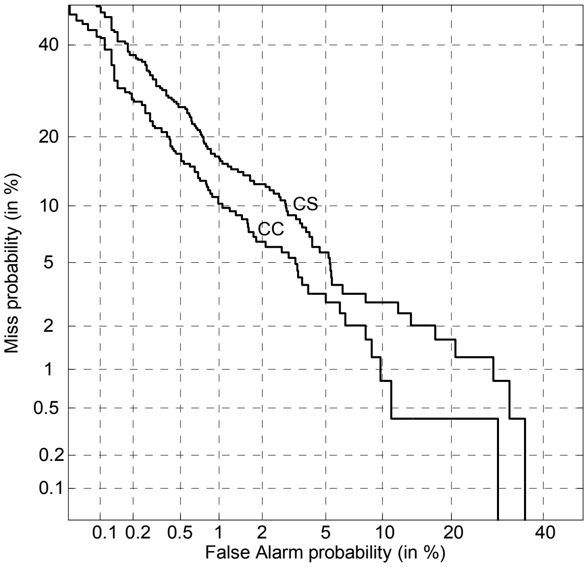

Figure 3. DET plots for Covariance matrices of size 27 and two training/testing scenarios (CC, CS).

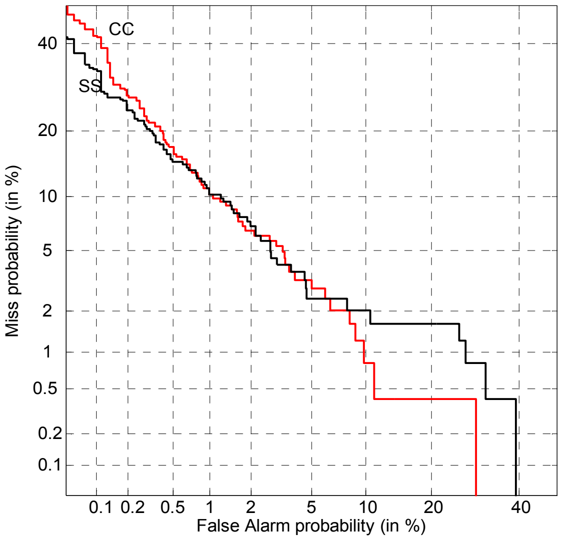

Figure 5. DET plots for Covariance matrices of size 27 and two training/testing scenarios (SS and CC).

Figure 2 compares DET plots for CM-4, 9, 19, 27, 39 and 55 using training and testing in Spanish. Figures 3, 4 and 5 compare several training and testing scenarios (CC, CS, SC and SS) when using CM-27.

Figures 3 and 4 show a slight degradation when training and testing languages are different. In fact, the minimum Detection Cost Function (DCF) increases from 3.6% to 4.5% and 4% respectively in figures 3 and 4.

Figure 5 shows that both plots intersect. Thus, we cannot affirm that one language produces always better results than the other one.

For comparing both methods (vector quantization and covariance matrices), we have used parameters No and P that require the same storage memory, as we see in table 3

5. CONCLUSIONS

In this paper we have studied speaker identification and verification tasks using a bilingual speaker data set. The main conclusions are:

- The Catalan database yields higher identification rates than the Spanish one for a high number of parameters. Otherwise the Spanish language achieves better rates. We think that this is due to the higher number of vocalic phonemes (8 in Catalan against 5 in Spanish).
- With different test and train conditions there is a little decrease in identification rate (about 1% for

high resolution codebooks, and greater values for other models and methods)

- For VQ better results are obtained when increasing the codebook size. Thus, best results are obtained for the larger size: 8 bits. On the other hand, for CM, the model size is related with the parameterization order (*P* value) and the optimal is obtained around P=27 for both tasks, identification and verification.
- Although VQ achieves the highest identification rates, the CM method is faster and in most cases requires less parameters for modeling each speaker. Additionally, CM provides better verification results, evaluated with the minimum DCF value.

6. ACKNOWLEDGEMENTS

This work has been supported by FEDER and the Spanish grant MEC TIC2003-08382-C05-02

7. REFERENCES

[1] M. Faundez-Zanuy "State-of-the-art in speaker recognition". IEEE Aerospace and Electronic Systems Magazine. Vol.20 nº 5, pp 7-12, May 2005

[2] A. Satue-Villar, M. Faundez-Zanuy "On the relevance of language in speaker recognition" Proc. EUROSPEECH'99 Budapest, Vol. 3 pp.1231-1234, September 1999.

[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 (2000), pp. 255-264, June 2000

[4] F. K. Soong et alt. "A vector quantization approach to speaker recognition". pp.387-390. ICASSP 85

[5] A. Gersho and R.M. Gray. Vector Quantization and Signal Compression, Kluwer Academic Publishers, 1991.

[6] F. Bimbot, L. Mathan "Text-free speaker recognition using an arithmetic-harmonic sphericity measure. pp.169-172. EUROSPEECH 91

[7] A. Martin et alt. "The DET curve in assessment of detection performance", V. 4, pp.1895-1898, European speech Processing Conference Eurospeech 1997

		No=0		No=1		No=2		No=3		No=4		No=5		No=6		No=7		No=8	
P	Train/test	Iden	ver	Iden	ver	Iden	ver	Iden	ver	Iden	ver	Iden	ver	Iden	ver	Iden	ver	Iden	ver
12	CC	58.8	36.6	69.8	27.9	79.6	21.1	93.1	13.8	98.4	11.7	98	10.3	98.8	9.1	99.6	8.0	100	7.1
12	CS	49.8	33.7	65.7	25.4	68.6	18.9	87.8	13.4	93.9	12.0	96.3	10.5	97.6	10.0	98.0	9.5	98.8	9.0
12	SS	64.1	33.0	74.7	24.8	83.7	18.1	93.5	12.4	97.1	11.1	98.8	9.7	99.2	8.2	98.8	7.6	99.2	7.3
12	SC	47.3	36.9	59.2	28.4	73.1	21.0	90.2	12.2	95.1	12.8	95.9	12.0	98.4	10.7	98.8	10.1	98.4	10
16	CC	62.4	35.6	72.7	27.0	83.3	20.1	93.5	13.4	99.6	11.9	99.2	11.3	99.6	9.7	100	8.4	100	7.5
16	CS	55.9	32.6	68.6	24.3	73.9	18.9	88.6	13.6	95.5	12.4	96.7	11.7	97.6	10.1	99.2	9.9	99.2	9.7
16	SS	68.2	31.7	77.6	23.0	84.5	17.4	94.3	12.5	96.7	11.1	98.4	10.3	98.8	8.4	98.8	8.0	99.2	7.4
16	SC	53.5	35.3	64.9	26.9	77.6	20.6	92.7	14.7	96.7	13.7	97.1	12.3	98.4	11.7	98.4	10.5	98.8	10.2
20	CC	64.9	34.6	74.3	26.4	86.1	20.9	93.1	15.1	99.6	12.7	100	11.0	100	10.6	100	9.0	100	8.3
20	CS	57.1	31.9	70.6	24.3	77.1	19.7	88.6	14	96.3	12.9	97.1	11.3	98.8	10.5	99.6	10.0	99.6	9.8
20	SS	69.0	30.4	80.0	23.1	85.3	17.8	94.3	13.2	97.6	11.1	98.4	9.9	98.8	9.1	99.2	8.1	99.6	7.3
20	SC	54.3	34.8	64.5	26.6	80.4	20.9	93.9	16.4	97.6	14.6	98.0	12.6	98.0	11.3	98.8	11.0	99.2	10.2

Table 1: Identification rates and DCF (verification) using VQ (S=Spanish C=Catalan),

	P=4		P=6		P=9		P=13		P=19		P=27		P=39		P=55	
Train/test	Iden	ver	Iden	ver	Iden	ver	Iden	ver	Iden	ver	Iden	ver	Iden	ver	Iden	ver
CC	22.0	25.8	55.9	13.9	82	7.4	91.0	6.4	96.7	5.2	99.2	3.6	99.2	6.6	92.7	9.7
CS	22.4	28.1	52.7	15.4	77.6	7.6	84.9	6.5	91.8	5.4	95.9	4.5	92.7	8.0	86.1	12.5
SS	27.3	23.1	65.3	12.3	87.3	6.8	92.7	5.1	97.1	4.2	98.8	3.6	95.9	6.8	90.6	11.0
SC	22.0	28.4	50.6	16.9	78.4	7.6	88.2	7.4	95.5	6.0	98.4	4.0	95.9	7.5	88.6	11.4

Table 2: Identification rates and DCF (verification) using CM (S=Spanish C=Catalan)