A REDUCED SEARCH ALGORITHM FOR SPEAKER RECOGNITION.

G. Miramontes, J. I. de la Rosa, E. García, J. J. Villa, M. A. Araiza, and C. Sifuentes
Universidad Autónoma de Zacatecas.
Av. López Velarde # 801 Centro
C.P. 98000 Zacatecas, Zac..
gmiram@cantera.reduaz.mx , vargasj@cantera.reduaz.mx

ABSTRACT

In this work, a reduced search algorithm for vector quantization codebooks is applied as a way to reduce the risk of wrong decisions in an automatic speaker recognition system.
Instead of a full search method, the algorithm is based on the geometrical properties of the vector space, reducing the search to those codebooks which are closer to the vector under test.
The speaker recognition system is intended to identify a suspect, between a small group of persons, using low quality recordings, working as a text independent automatic speaker recognition system. It was found that the alternative search algorithm can be used to reduce the risk of wrong decisions, which is specially important in forensic applications.

1. INTRODUCTION

Automatic speaker recognition (ASR) is the process of recognize, automatically, a person based on the information included in a speech sample.
ASR can be classified in Automatic Speaker Verification (ASV), and Automatic Speaker Identification (ASI). ASV is the process of determining, among a registered number of speakers, the one who claims his identity. On the other hand, ASI is the process of finding between a group of persons the speaker under test. In addition, we can have a closed group ASI, if it is known a priori that the person belongs to the group, or open group ASI, if there is the possibility that the person under test doesn't belong to that group. This last classification is very important in forensic cases, since there is the possibility of an incorrect identification of a suspect. A false identification in this case will be extremely costly and has to be avoided always.

Many decision methods are based on a minimal distance, for example, the Euclidian distance, and when the suspect (speech sample) doesn't belong to the group, the decision method depends on a threshold to avoid a false identification. This threshold is very difficult to determine and it is still an open problem. In this work, the main objective of the search algorithm is to restrict the search space so to reduce the risk of false identification.

2. EUCLIDEAN DISTANCE CLASSIFIER

The recognition based on the Euclidean distance is used in many practical cases. The Euclidean distance between two points i, j, which is the minimal distance between them, is given by

$$d(i,j)=\mathrm{sqrt}\{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2\} \quad (1)$$

So, the distance between a vector to be classified X and a centroid Z_i, is given by:

$$\begin{aligned} d(X,Z_i) &= \|X-Z_i\| \\ &= \mathrm{sqrt}\{(X-Z_i)^T(X-Z_i)\} \\ &= \mathrm{sqrt}\{\Sigma_{j=1}^{N}(X_j-Z_{ij})^2\}, \end{aligned} \quad (2)$$

where T indicates transpose operation.

2.1. The reduced search algorithm

The geometric properties of the algorithm were discussed in [1],[2]. Having a codebook $C=\{C_i, i=1,...,N\}$ of size N, C_i is a centroid of dimension k, i.e., $C_i=\{c_{i,1},c_{i,2},...,c_{i,k}\}$. Given an input vector $X=(x_{i,1},x_{i,2},...,x_{i,k})$, instead of finding the minimum distance between the cepstral coefficients of the input vector through the whole codebook, we find the minimum total distorsion between the input vector and a subset S of the codebook.

To define the subset S, the first step of the algorithm is to find the centroid of the input vector, Y(X), which would be close to X. The distance between X and the centroid C_i will be denoted as d(X,C). The algorithm consists of two parts. In the first part, a region given by a circle with center C_i and radius $2h_i$ is defined. Any other centroid that falls outside this region is excluded in the search.
In the second part, another region limited by two circles, one of radius r_x-h_i and r_x+h_i, both centered

around the origin is defined. Any centroid that falls outside this regions is excluded in the search. Figure 1 shows the final region where the search will be done. In this way, only the closest centroids of the codebook will be considered for the recognition of the speaker.

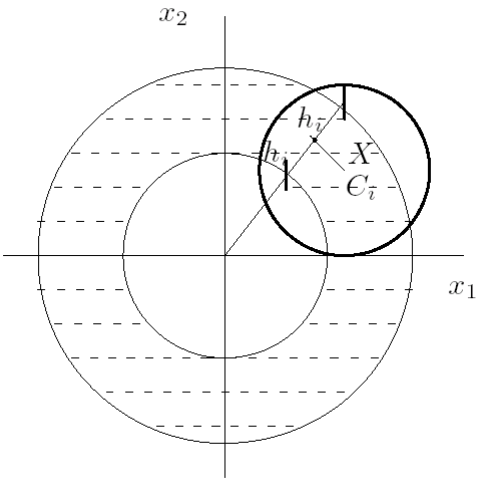

Figure 1. Final Search Region of the algorithm.

One way to estimate the distance r_i is to use the norm of each codebook, i.e., $\|C_i\|$, where $\|\bullet\|$ is defined as

$$\|C_i\|=(\Sigma_{j=1}^{k}|C_{ij}|^m)^{1/m}, \quad \text{where } m=1,2,... \qquad (3)$$

Then, for a new input vector we can calculate its r_X and look for the C_i with a r_i closer to r_X. Figure 2 shows a simplified view of the r_i's compared to r_X. It can be seen that r_X, the big dot, is close to r_5, the plus sign. Note also that r_X is also close to r_2 denoted as a circle. So, we can select as good candidates for the search only the closer centroids, and exclude the rest of them.

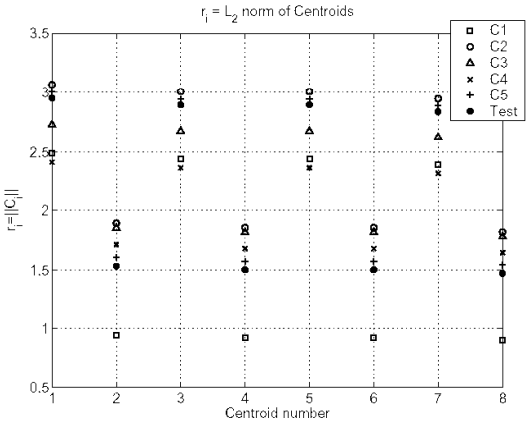

Figure 2. Comparison of r_i and r_x for 8 centroids.

3. RESULTS

Figure 3shows the result for the speech sample L5. It can be seen how r_X is close to r_5 and r_1, which correspond to the L_2 norms of speech sample L5 and the speech sample L1, respectively.

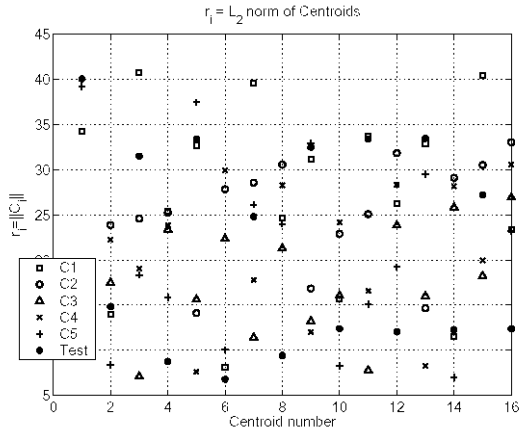

Figura 3. r_X and r_i to find search region for speaker L5.

The next step is to exclude all centroids that do not meet the requirement $d(C_i,C_j)\leq 2h_i$, leaving a subset S of the codebook.
The result, obtained under Matlab, is shown as

```
Closer Centroid to R_x is L1
InD =
       1     5
The identified speaker is L5
The file under test was L5.wav
```

where `InD` is the index of those centroids belonging to the subset S of the codebook where the identification will be applied.
For sake of completeness, we have compared the results using a full search method, but it is clear that once the subset S has been defined, the identification algorithm should run using only $C_i \in S$. Note that in this case r_1 is close to r_X, but r_5 also belongs to the subset, so if the search is done using the subset $S = [C_1, C_5]$ the identification is correct.

It was found that for very low quality speech samples, L4.wav in this case, the identification based on the comparison of r_i and r_X does not provide good results. L4 was recorded with a speaker too far from the recorder, so his voice sounds more like a background sound.
In this case, the subset S does not include the correct centroid, and it can be recognized using only the full search method.
Table 1 summarizes the results of the identification for a particular set of speakers.

Table 1. Recognition results using WAV files.

Spkr	Closest C_i	Subset	Full search
Ismael	L1	L1 L5	L1
Gogo2	L5	L2 L5	L5
Gogo3	L5	L2 L5	L5
Gogo4	L2	L2 L5	L5
CldL	L1	L2 L5	L3
Lalo1	L3	L2 L5	L2
Success	50%	83%	100%

4. CONCLUSIONS

A reduce search algorithm was tested on a text independent ASR system. It was found that high quality recordings are very important to avoid false identification, specially if the reduced search method is used. Future work will be done to test the algorithm under different signal to noise ratios, a greater number of speakers, and a large number of speech samples.

5. REFERENCES

[1] Ramírez Acosta A. A., ***Reconocimiento Automático de Locutor Mediante Técnicas Dependientes e Independiantes del Vocabulario para un Sistema Acotado por el Ancho de Banda Telefónico y Realización de un Sistema Experimental***, Master Thesis, CICESE, Telecommunications and Electronics Department, February 1996.

[2] Huang C. M., Bi Q., Stiles G. S., and Harris R. W., ***Fast Full Search Equivalent Encoding Algorithms for Image Compression using Vector Quantization***, IEEE Trans. on Image Processing Vol. 1, No. 3,pp.- 413-416, (1992).