CREATIVE MATH. & INF. **Online version at** http://creative-mathematics.ubm.ro/
17 (2008), No. 3, 406 - 411 **Print Edition: ISSN 1584 - 286X Online Edition: ISSN 1843 - 441X**

Dedicated to Professor Iulian Coroian on the occasion of his 70^{th} *anniversary*

The influences of distances measures for speaker identification

MARIETA GÂTA AND GAVRIL TODEREAN

ABSTRACT. This paper is a study which compare statistical features of long term text-independent speaker identification .We compare by statistical methods four distances: the City block, Euclidean, Weighted Euclidean and Mahalanobis distance measures. Experiments confirm our assumption that Weighted Euclidean distance performs better then the other three distances.

1. INTRODUCTION

Speaker identification is the process of finding and attaching a speaker identity to the voice of an unknown speaker. Automated speaker identification do this by comparing the voice with stored samples in a database of voice models. Speaker recognition is a synonym for speaker identification, means a generic term referring to many spoken technologies applied to speakers, including speaker identification and speaker verification. Speaker verification is the process of determining whether a person is who she / he claims to be. It determines a one-to-one comparison between a newly input voiceprint (by the claimant) and the voiceprint for the claimed identity that is stored in the system. While performance of speaker verification is unaffected by the population size, performance of speaker identification decreases as the population size increases. Text dependent is a variant of speaker verification that requires the use of a password, pass phrase, or another pre-established identifier (e.g. the speaker's name). Text independent is a variant of speaker verification that can process freely spoken speech (an unconstrained utterance). Text prompted is a variant of speaker verification that asks users to repeat random numbers and / or words. A typical prompt might be "Say 1 2 3 4." Some developers consider text prompting to be a kind of text-independent technology. It is also called challenge response. Speaker recognition is a classification in which pattern matching is done between reference model and test patterns.

From speech utterances are extracted (by statistical or dynamic methods [2]) test patterns and reference patterns (acoustic feature vectors). It can be used different statistical acoustic features: linear prediction coefficients, cepstral coefficients, reflection coefficients, and log area ratio coefficients [3].

Received: 2.02.2009. In revised form: 05.03.2009. Accepted: 23.05.2009.
2000 *Mathematics Subject Classification.* 51K05, 68T10.
Key words and phrases. *City block distance, Euclidean distance, Weighted Euclidean distance, Mahalanobis distance.*

In training phase reference models are trained or generated from the reference models by different methods like: general statistical methods, Vector Quantization, Hidden Markov Models and Neural Networks. In general statistical methods a reference model can be formed by obtained statistical parameters from reference speech data. We compare a test model with the reference models in the phase of pattern matching. This comparison is made by distance measure [4] [5]. After this comparison, in the decision phase, we labeled test model to a speaker model. In the labeling phase we use the minimum risk criterion [6]. Text-dependent case use the same text for training and testing stage in utterance. Text-independent case implied that training and testing phases involve different texts in utterance. In this process we need more utterance for better results which means higher accuracy. In the text-independent case it can be obtained better results with statistical features [1].

In this paper we implemented speaker identification, this way: we extracted statistically acoustic feature vectors of reflection coefficients and then we calculate average over a long period [7]. We formed a text-independent reference model for each speaker generating a reference template (which means a covariance matrix and a mean vector) from reference feature vectors. We compared then each test feature vector with a reference model by distance measure.

We used four distance measure for comparing and studying, using different usage of covariance matrix [4]. These distances are: City block, Euclidean, Weighted Euclidean (WED) and Mahalanobis distance measures. We obtained a match if a test vector is labeled to the right speaker. In the case of distance measures this means that intra-speaker is smaller then inter-speaker distance. We intend to obtain the minimum risk criterion. We obtained accuracy by calculating the percentage of matches.

2. THEORY

One of the classification methods is distance measure. This classification method is based on the assumption that the underlying probability has a Gaussian distribution. The Gaussian-distributed probability density function for a speaker (a class) is:

$$p(x) = \frac{1}{(2^{\pi})^{N/2} |W|^{1/2}} \exp\left\{ -\frac{1}{2} (x - \bar{x})^T W^{-1} (x - \bar{x}) \right\} \tag{2.1}$$

where x is a Gaussian random vector with dimension N , $\bar{x}$ and W are the mean vector and the covariance matrix for the class model, and T means the transposition of a vector. If we desired to plot the curve surface of $p(x)$ let $p(x) = C$, where C is a constant. Then equation (2.1) become

$$(x - \bar{x})^T W^{-1} (x - \bar{x}) = C' \tag{2.2}$$

where C' is another constant related to C in an obvious way. The quantity on the left-hand side of equation (2.2) explains the property of distance measure. Intra-speaker distance is generally smaller than inter-speaker distance, due to the normal distribution. Classification is then reduced to the process of finding a

speaker model nearest to a given test vector and then labeling this vector to its speaker.

2.1. **Distance measures.** Let $y^{(i)}$ denote the i^{th} reference speech feature vector of a certain speaker, where $1 \leqslant i \leqslant R$, R is the total number of the reference speech feature vectors for the speaker. The reference template of a certain speaker, y, is:

$$\bar{y} = \frac{1}{R}\sum_{i=1}^{R} y^{(i)} \tag{2.3}$$

and the covariance matrix for the speaker is

$$W = \frac{1}{R}\sum_{i=1}^{R} y^{(i)} y^{(i)^T} - \bar{y} - \bar{y}^T \tag{2.4}$$

where $y^{(i)^T}$ is the transposed vector (in row) of $y^{(i)}$, and y^t is that of y.

Let D denote a diagonal matrix. Its diagonal elements are exactly the same as that of W. A test column vector is denote x. The distance between x and y (where x and y has dimension N) is used in defining the four distance measure methods which are defined respectively:

City block distance: $d_c(x, y)$*or absolute value distance*

$$d_c(x, \bar{y}) = \sum_{i=1}^{N} |x_i - \bar{y}_i| \tag{2.5}$$

Euclidean distance: $d_E(x, y)$

$$d_E(x, \bar{y}) = (x - \bar{y})^T (x - \bar{y}) = \sum_{i-1}^{N} (x_i - \bar{y}_i)^2 \tag{2.6}$$

Weighted Euclidean distance: $d_W(x, y)$

$$d_W(x, \bar{y}) = (x - \bar{y})^T D^{-1} (x - \bar{y}) \tag{2.7}$$

Mahalanobis distance: $d_M(x, y)$

$$d_M(x, \bar{y}) = (x - \bar{y})^T W^{-1} (x - \bar{y}) \tag{2.8}$$

3. Experiment

3.1. **Speech database.** The speech data was obtained in this way: we recorded 10 male speakers. We digitized then these speeches with the sample rate 10 kHz and resolution 12 bit. The test set and the reference utterance set are about 20s after removing silent. We obtained then a set of recordings more accurate than initial set. Each utterance set contained 20 segments, where each segment has duration of 1 second. Each segment contained 20 segments, each frame 256 sampled points. We extracted a vector of 10 reflection coefficients from each segment. These extractions were made this way: computing the mean of these 20 vectors of reflections coefficients in that segment. We have now 20 feature vectors of 10

reflection coefficients for the test set or the reference set. We have other processing specifications of 98% first order pre-emphasis, non-overlapping 250 points Hamming window and analysis filter of order 20.

3.2. **Experimental results.** We generate for each speaker two parameters: covariance matrix and a mean value. We obtained these values from the 20 records of the reference test. These two parameters are the basis for the distance measures or for the Gaussian probability estimation. For each speaker we have a reference model link in equation (2.1). This model is used to calculate Gaussian probability. We applied a test vector to every reference model and we labeled the model with largest probability. We estimate the accuracy on percentage of matches. For calculate distance measure we applied a test vector to each reference models. We calculate thus distance through each one of the equations (2.5), (2.6), (2.7) and (2.8). We labeled, then we count the matches and finally we calculate the accuracy. In Table1 we present the number of speaker matches for the 10 speakers by four methods. Each number from the cells of table represents the number of matches out of 20. Accuracy is the number of matches divide by 20. This ratio (accuracy) is presented in Table 1.

Speaker Method	1	2	3	4	5	6	7	8	9	10
City Block Distance	18	19	15	16	14	18	15	19	16	14
Euclidean Distance	16	19	13	13	15	18	12	19	16	13
Weighted Euclidean Distance	19	20	17	19	16	18	16	20	17	13
Mahalanobis Distance	17	20	19	18	18	19	18	16	19	16

Table 1 Number of matches (out of 20) for the 10 speakers by four methods

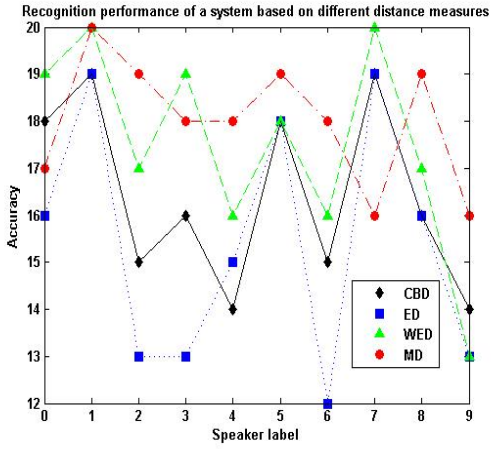

Figure 1. Accuracy for the 10 speakers with the four methods

Where:
CBD - City block distance measure
ED - Euclidean distance measure
WED - Weighted Euclidean distance measure
MD - Mahalanobis distance measure

4. METHODS COMPARISON

For comparing the four measure methods we use a statistical test, multiple comparison approach, applying this test to the accuracy data from the Table 1.

4.1. **Multiple comparison approach.** The multiple comparison approach was used to determine which method had significantly different median total matches. If we have two methods i and j, these methods are considered different if is satisfied the following inequality:

$$|R_j - R_i| > t(\alpha/2)\sqrt{2n(A_f - B_f)/(n-1)(k-1)}$$

Where $R_i,\ R_j,\ A_f$ and $B_f$are given and $t(\alpha/2)$ is a critical value on the $t-$table using $(n-1)(k-1)$ degrees of freedom $(\alpha/2) = P(t_{(n-1)(k-1)} > t(\alpha/2)))$. The total matches of the four methods were ordered in an array, and the rank was assigned to each corresponding value as its order. The rank sums of WED, MD, CBD, and ED were respectively 27.7, 26.3, 18.4, and 12.5. If the rank sums of any two methods were greater than 12.4 units apart (with α=0.05), they might be regarded as having unequal medians total matched. In this situation we concluded that WED and MD might be regarded as superior to CBD and ED. We didn't find any other significant differences.

5. CONCLUSION

In this paper we compared four methods for long term text-independent speaker identification using statistical features. These methods are distance measures: City block, Euclidean, Weighted Euclidean, and Mahalanobis distance measure. All this distance measures are simplified models derived from Gaussian distribution model.

In this experiment we observed by numerical -statistical- calculation that two distances done better performances then the others two. These two methods with better results are Weighted Euclidean distance and Mahalanobis distance. The methods with the weak performances are City block distance and Euclidean distance. We observe that Weighted Euclidean distance and Mahalanobis distance might be regarded as superior to City block distance and Euclidean distance.We don't observe other differences.

REFERENCES

[1] O'Shaughnessy, O., *Speaker recognition*, IEEE ASSP Magazine, 4-17, 1986
[2] Furui, S., *Comparison of speaker recognition methods using statistical features and dynamic features*, IEEE Trans. ASSP, Vol. 29, No. 3, pp. 342-350, 1986
[3] Shridhar, M. and Mohankrishnan, N., *Text-independent speaker recognition:a review and some new results*, Speech Communication., Vol. 1, Nos. 3-4, pp. 257-267, 1982

[4] Wohlford, R. E., Wrench, E. H. and Landell,B. P., *A comparison of four techniques for automatic speaker recognition,* ICASSP-80, pp. 908-911, 1980
[5] Ong S. and Moody, M. P., *Confidence analysis for text-independent speaker identification using statistical feature averaging,* Applied Signal Processing, Vol. 1, No. 3, pp. 166-175, 1995
[6] Basztura, C., *Experiments of automatic speaker recognition in open sets,* Speech Communication, Vol. 10, No. 2, pp. 117-127, 1991
[7] Markel, J. D., Oshika, B. T. and Gray, A.H., *Long-term feature averaging for speaker recognition,* IEEE Trans. ASSP, Vol. 25, pp. 330-337, 1977

NORTH UNIVERSITY OF BAIA MARE
DEPARTMENT OF MATHEMATICS AND
COMPUTER SCIENCE
VICTORIEI 76
430122 BAIA MARE, ROMANIA
E-mail address: marietagata@ubm.ro

TECHNICAL UNIVERSITY OF CLUJ-NAPOCA
FACULTY OF ELECTRONICS
AND TELECOMMUNICATIONS
DEPARTMENT OF COMMUNICATIONS
GEORGE BARIŢIU 26-28
400027 CLUJ NAPOCA, ROMANIA
E-mail address: Gavril.Toderean@com.utcluj.ro