Proceedings of the 22nd Annual EMBS International Conference, July 23-28, 2000, Chicago IL.

Automatic Speaker Identification by means of Mel Cepstrum, Wavelets and Wavelets Packets

Torres Humberto M., Rufiner Hugo L.
Laboratorio de Cibernética - Facultad de Ingeniería, Bioingeniería
Universidad Nacional de Entre Ríos

Abstract - The present work consists on the use of Delta Cepstra Coeficients in Mel scale, Wavelet and Wavelet Packets Transforms to feed a system for automatic speaker identification based on neural networks. Different alternatives are tested for the classifier based on neural nets, being achieved very good performance for closed groups of speakers in a text independent form. When a single neural net is used for all the speakers, the results decay abruptly when increasing the number of speakers to identify. This takes to implement, a system where there is one neural net for each speaker, which provided excellent results, compared with the opposing ones in the bibliography using other methods. This classifier structure possesses other advantages, for example, add a new speaker to the system only requires to train a net for the speaker in question, in contrast with a system where the classifier is formed by a single great net, which should be in general trained completely again.

Keywords - Speaker Identification, Voice Analysis, Mel Cepstrum, Wavelet, Wavelet Packets, Neural Networks

1. INTRODUCTION

The voice signal has different levels of information. Firstly, it carries the words or messages, but in a secondary level, the signal also takes information about the speaker's identity. While the area of the automatic speech recognition is relative to the extraction of the linguistic message in a sentence, the area of the speaker's recognition is concerning with the extraction of the identity of a person [1].

The automatic recognition of the speech is a multidisciplinar field with special linking to computer science and, inside her and in a special way, to pattern recognition and artificial intelligence [2].

According to the sentences used in the speaker's recognition, this it is generally divided in two classes: (1) dependent of the text and (2) independent of the text [2]. Another division possible of the speaker's identification, it is as soon as if it is a problem of closed group or a problem of open group.

The problem of the speaker's identification can be divided in two components: speech analysis and classification.

Artificial Neural Networks (ANN) are excellent classification systems and they specialize in working with noisy, incomplete, overlapped data, etc. The problem of the speaker's identification is a task of clàssification of data that has all these characteristics, making the ANN an attractive alternative to the descripted approach.

2. MATERIAL AND METHODS

2.1 About the data

Voice signal for identification experiments were obtained from TIMIT continuous speech corpus [4]. This database has been made in combined form by Texas Instruments (TI) and the Massachusetts Institute of Technology (MIT). It is one of the multi-speaker databases more employed in the field of the Automatic Continuous Speech Recognition (ACSR) to be the biggest, complete and better documented of its type. It should be observed that, for our work, the original separation of the sentences of TIMIT in training and test it is not valid, since the speakers of one and another are different. So we decided to take three of each four sentences, for each speaker, to train and the remaining one for test. This division diminishes the quantity of sentences to train. Also, we only work with the group of labeled sentences for training in the original TIMIT division, since this presents more sentences per speaker.

Where the opposite is not indicated, records were used of feminine speakers of the same region, that in our case increases the degree of difficulty of the recognition task.

2.2 Signal Processing

In speech recognition, the main objective of the acoustic processing module is to extract characteristic that are invariant to the speaker and the channel of transmission of the signal, and that are representative of the content of the lexicon. On the other hand, the speaker's identification requires the extraction of characteristic related with the speaker, which are independent of the pronounced words. Such characteristics include properties related with the spectral envelope (such as the average position of the formants in some vowels) or average ranges of the fundamental frequency. Unfortunately, since these characteristics frequently are difficult to estimate, the current systems use acoustic parameters that have been developed to be used in speech recognition. In general, characteristic based on some type of short time spectral estimation are used .

0-7803-6465-1/00/$10.00 ©2000 IEEE

2.2.1 Mel Cepstrum

The objective of signal processing is to extract important information of a signal by means of some transformation type. An analysis commonly employee in the speaker's recognition is the Mel Frequency Cepstrum Coefficients (MFCC) [5]. Previous techniques of extraction of characteristics work on the power spectrum and the Cepstrum coefficients of the signal. However, the power spectrum and the Cepstrum are not always advisable for pattern recognition since the amplitude and the form change with a simple microphone change. A simple alternative that provides a bigger robustness in the patterns constitutes the Delta Cepstrum.

2.2.2 Wavelet Transform

In the case of the digital signals the Discrete Wavelet Transform (DWT or WT) has several attractive characteristics that have contributed to their recent increase of popularity in mathematics and signal processing fields and it also have a fast implementation denominated Fast Wavelets Transform (FWT).

The FWT is implemented with a tree of subsampled by two diadic filters. The number of scales in the FWT is limited by the number of elements of the input vector. We use a vector with 512 elements (corresponding to 32 mseg. of voice signal), that which restricts us to use a maximum of 9 scales. As in this situation the last scale and the residual is only represented by 1 element, we decided to use a transformation with 8 scales (Figure 1).

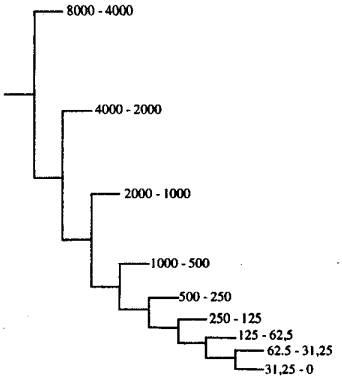


Figure 1: Tree of filters for the FWT. The numbers indicate the pass band of each filter, in Hz.

The WT has an inconvenience: if we entered a vector of 512 coefficients, the transformation returns the same number of elements, while when Mel Cepstrum was used the number of elements of the input vector was reduced to 32 coefficients. Evidently in this case the number of weigths of the neural network grows excessively. On the other hand it is supposed that in these coefficients, redundant information exists in order to make the classification, for what several coefficients could be compressed or integrated in only one, therefore based on previous works [6][7], we intended to integrate the coefficients of the WT per bands . But, since we don't know a priori how to carrying out the division of the vector, we proceeded to take as vector of training the energy of each one of the scales of the WT.

2.2.3 Wavelet Packet Transform

The DWT is really a subset of the Wavelet Packet Transform (WPT) [8]. It generalizes the time-frequency analysis carried out by the DWT, giving as a result a family of orthonormal bases, one of which is the DWT. In the same way that the FWT, a fast algorithm exists (derived from this) for the calculation of the WPT. This way each of the bases of the WPT can be seen as a tree of filters. A problem consists on choosing the tree or appropriate base for a particular application.

All these characteristics and a series of recent and successful applications of this technique to our field, have taken us to consider it like an interesting alternative to the classic methods for speech analysis.

The first trouble to solve it is what base (or tree of filters) to use. Presently work intends to use a perceptually guided base, which took of [6] (Figure 2).

2.3 Artificial neural networks

Since the patterns to classify are dynamic, makes necessary the use of neural nets with temporal delays (TDNN) [7][11]. On the other hand our initial experiments confirmed this hypothesis, for that that finally a TDNN was used, with a hidden layer and a delay in the input layer.

Also, in previous works [7] we had demonstrated that a system that uses one neural net per speaker (Figures 4), it is superior to one that uses a single great net for all the speakers (Figure 5).

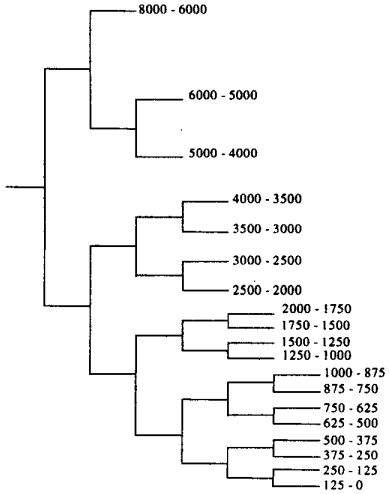


Figure 2: Chosen base to carry out the experiments. The numbers indicate the pass band of each filter, in Hz.

Also, to the same as in the case of the WT, we proceeded to integrate for bands, like sample the Figure 3, to obtain a vector of 32 components.

The nets were trained using Backpropagation algorithm. Training stopped in the generalization pick. To fix the learning coefficients and moment for each experiment type, they were carried out a series of experiments with the purpose of determining them.

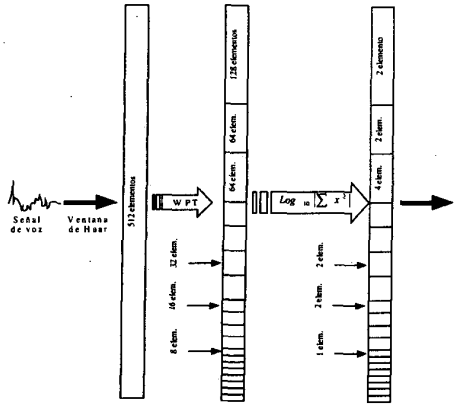


Figure 3: Integration for scales of the WPT's coefficients.

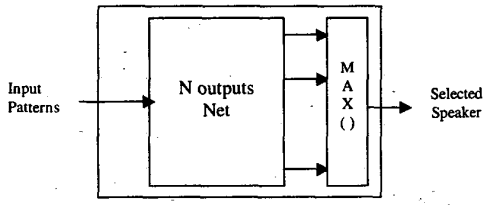
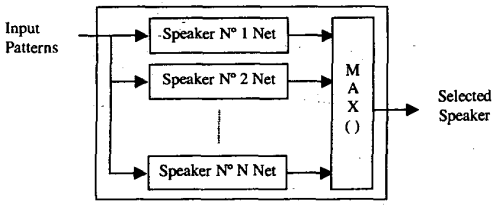


Figure 4: Structure of the traditional classifier.



Figures 5: Structure of proposed classifier

3. RESULTS

There is a great number of variables in a possible speaker's identification system, so much in signal processing of speech, like in the possible configurations of the classifier. This leads to the realization of a great number of experiments to be able to fix the most appropriate values in these variables. On the other hand, since the results obtained by the neural nets depend on their initial conditions, it is convenient to repeat the experiments to obtain an average that is more representative of the results. Also, if for example it is sought to identify a speaker of among a group of 5 speakers, the obtained results will depend on the selected group of speakers. Therefore, it is convenient to take different groups of speakers and to find an average of the results for each group. Presently work, all the presented results correspond to average values, so for a particular case the results can be even better (or worse).

We were carried out experiments with three different types of processing: Mel Cepstrum (MFCC), Wavelet (WT) and Wavelet Packet Transforms (WPT) .

For the experiments, silence fragments of speech signal were removed, since a series of tests demonstrated that this increases the recognition percentages. In all the cases temporal windows of 32 mseg of duration were used, overlapped in 50%. The window selected for the experiments with MFCC was Hamming, and for the other experiments a Haar window was used. For the experiments with Mel Cepstrum we used a preenfasis, with m = 0.97, because this was indicated in the bibliography, while for WT and WPT we don't use this type of pre-processing.

To form the vector of patterns with MFCC, 16 coefficients Mel Cepstrum were calculated and then added the Delta Cepstrum, to form a vector of 32 coefficients for each pattern.

The first carried out experiments had as objective to determine the efficiency of each one of the signal processing proposed for speech. Once certain the most appropriate processing, we carried out experiments with 50 speakers to determine how the system responded when the number of speakers increased.

The results of the carried out experiments are presented in form of charts, where we detailed results obtained with training and test patterns. In these charts appears the percentages of bad classified frames, labeled as"% of bad classified frames", the percentage of good classified sentences,"% of good classified sentences", where a sentence is considered good classified if most of its frames were correctly classified by the system.

3.1 Comparison of the Results with Different Processings

Next we place in a single chart (Chart 1) the best results in the three prosecutions, for a closed group of 10 speakers. Observing the Table 1, we concludes that the best results were obtained with WPK.

Table 1: Comparison among different processings.

PROC.	% WELL CLASSIFIED FRAMES		% WELL CLASSIFIED SENTENCES	
	TRN	TST	TRN	TST
MFFC + Δ	86.22	79.65	99.50	98.00
WPK	86.84	80.14	97.50	100.00
WT	81.09	73.46	97.50	94.00

Table 2: Results for 50 speakers.

SPEAKERS	% WELL CLASSIFIED FRAMES		% WELL CLASSIFIED SENTENCES	
	TRN	TST	TRN	TST
FEMENINE	85.62	76.70	96.60	92.00
MASCULINE	84.66	85.65	91.80	96.00
AVERAGE	85.14	79.55	94.20	94.00

Table 3: Results for 50 speakers, after tunnning the system.

SPEAKERS	% WELL CLASSIFIED FRAMES		% WELL CLASSIFIED SENTENCES	
	TRN	TST	TRN	TST
FEMENINE	86.85	76.61	99.60	98.67
MASCULINE	88.52	81.50	99.80	100.00
AVERAGES	87.69	79.05	99.70	99.33

3.2 Experiments with 50 Speakers

In this experiment 50 speakers were used (25 feminine and 25 masculine), taken at random of the eight regions of TIMIT. The obtained results can be seen the Table 2. The dispersion of the results was calculated using the standard deviation.

Since exist diverse parameters and conditions of the system that can alter (for well or for bad) the results of the Table 2, in the following experiments some of these conditions are varied with the purpose of finding the system that produces the best possible results.

As the results shown in the Table 2 depends on initial conditions we repeated the previous experiment 3 times, varying the initial conditions, taking the best result in all the runs.

Observing the curves of error of the previous experiment, we can observe in some cases that the net falls very quickly in a local minimum (before 500 epochs), well-known fact as "premature convergence". This takes to that finded solution is not necessarily the most appropriate. Therefore, we repeated the experiment, with the difference that save the best net only after epoch 500.

It has been observed that a parameter that influences on the recognition is the threshold of discretization of the continuous output of the net [12]. In previous results, the threshold was 0,5. Keeping this in mind we proceeded to vary the threshold (between 0,1 and 0,9 in steps of 0,1) in independent way for each one of the nets. The obtained result is shown in the following graphics.

Analyzing the obtained results we could conclude that, on the average, the appropriate threshold would have a value of 0,5. But, if the individual results are observed for each speaker, the threshold that maximizes the recognition, varies from speaker to speaker (or from net to net). Keeping this in mind, we take the best threshold for each speaker. In the Table 3 the results are shown after tunning the system.

4. Discussion

Presently work was carried out a comparison between three types of speech processing and fourteen types of classifiers (seven types of neural nets, using a net for all the speakers or a net per speaker) applied to the problem of the speaker's automatic identification.

When a net is used for all the speakers, the results decay abruptly when increasing the number of speakers to identify. This takes to implement, a system where there is a net for speaker, which provided excellent results, compared with the opposing ones in the bibliography using other methods [13][14][15][16].

Of the used processings the best results were obtained with Wavelets Packets Transform, with an integration for bands of their scales with the purpose of reducing the dimension from the input of the neural net, since this reduction in great way the numbers of parameters (weigths of the nets) to train on the part of the classifier, and also, it reduces the search space without loss of important information.

There are other factors that influence on the final result given by the classifier. One of these problems is the threshold of discretization of the ouput signal of the neural nets. In this work, we was proven that each net possesses an optimal value threshold, which should be fixed in an independent way for each one of the nets that compose the system.

In this work was used an integration per bands of the WPK and WT with the purpose of reducing the dimension of vector patterns, with which very good results were obtained. In the same way, when we uses Mel Cepstrum we only experiment with 16 triangular filters. But this doesn't mean that the used partition (or one of the experienced ones) it is the one that better adapts to the problem. Therefore, it should be proven with other possible partitions.

In the obtained results it was noticed differences depending on the Wavelet mother used. Another factor that could influence is the base used, the one which presently work it was not varied to see its influence in the results.

In future works it should be considered cases where there are channel differences between training and test data, like it could be the case of a microphone change.

5. References

[1] D. A. Reynolds. R. C. Rose. "Robust Text-Independent Speaker Identification Using Gaussian Speaker Models Mixes".. IEE Trans. on Speech and Audio Processing. Vol. 3 Number 1 January of 1995.

[2] Chi Wei Che. Q. Lin. D.S. Yuk. "An HMM Approach to Text-Prompted Speaker Verificación". CAIP Center. Rutgers United.. Piscataway. NJ. It USES.

[3] H. Silverman. D. Morgan. "The aplication of dynamic programin to conected speech recognition". IEEE Acustic. Speech and Signal Processing Magazine. vol. 7. pp 6-25. Julio 1990.

[4] Garofolo. Lamel. Fisher. Fiscus. Pallett. Dahlgren. DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus Documentation. National Institute of Standards and Technology. February 1993.

[5] J. Deller. J. Proakis. J. Hansen: "Discrete-Time Processing of Speech Signals". Prentice-Hall. 1987.

[6] H. L. Rufiner, H. M. Torres, Classification of Phonemes By means of Packages of Guided Waves Perceptualmente", Annals of the 1° Latin American Congress of Ingeniería Biomédia", Mazatlán, Sinaloa, Mexico, 1998.

[7] H. Torres, H. L. Rufiner, Automatic Identification of the Speaker by means of Redes Neuronales", Annals of the XII Argentinean Congress of Bioingeniería, Buenos Aires, Argentina, 2-4 June of 1999.

[8] M. A. Cody, The Wavelet Packet Transform: Extending the Wavelet Transform", Dr. Dobb's Journal, April 1994.

[9] Mohamad H. Hassoun. Fundamentals Artificial of Neural Networks. The MIT Press. 1995.

[10] J.L. Elman. "Finding structure in it cheats". Cognitive Science 14 (1990) 179-211.

[11] A. Waibel. T. Hanazawa. G. Hinton. K. Shikano. K. Lang; "Phoneme Recognition Using Time-Delay Neural Networks". IEEE Trans. ASSP Vol. 37. Not 3 (1989).

[12] D. H. Milone, J. C. Sáez, G. S., H. L. Rufiner, Trees of nets neuronales autoorganizativas", Annals of the 1° Latin American Congress of Ingeniería Biomédia", Mazatlán, Sinaloa, Mexico, 1998.

[13] J. Koolwaaij, Speaker Identification and Assessment on the YOHO database", Department of Language and Speech, University of Nijmegen, E-mail: koolwaaij@let.kun.nl, November 14, 1996

[14] C. Che, D. Yuk, J. Flanagan, Q. Lin, Development of 1996 RU Speaker Recognition System", Meeting: 1996 NSA Speaker RecognitionWorkshop, Maritime Institute, Maryland, 27 and 28 of March of 1996.

[15] D. Genoud, G. Gravier, F. Bimbot and G. Chollet, Combining methods to improve speaker verification decision", International Conference on Spoken Language Processing, Philadelphia, Octobre 1996.

[16] J.W. Koolwaaij L. Boves, TO new procedure for classifying speakers in speaker verification systems", Department of Language and Speech, Nijmegen University P.O. Box 9103, 6500 HD Nijmegen, the Netherlands, E-mail: koolwaaij,boves@let.kun.nl.

Address for Correspondence: Laboratory of Cybernetics – Univ. Nac. de Entre Ríos (UNER). Ruta 11 Km.10 – Oro Verde (Paraná). E. Ríos.
Electronic mail: Hugo L. Rufiner: lrufiner@arcride.edu.ar
Humberto M. Torres: hmtorres@hotmail.com