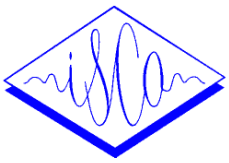

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

SPEAKER IDENTIFICATION THROUGH A MODULAR CONNECTIONIST ARCHITECTURE: EVALUATION ON THE TIMIT DATABASE

Younès BENNANI

L.R.I. U.A. 410 C.N.R.S.
University of Paris-Sud
Bât. 490, 91405 Orsay, France

ABSTRACT

We present a connectionist system for text independent speaker identification. We have developed an architecture based on the cooperation of several connectionist modules to achieve this identification. The system is composed of a typology detector and a set of expert modules. Each expert module of the system is concerned with the discrimination between speakers of the same typology. The score used in the final decision is obtained weighting the scores of the typology detection module with those of the expert modules. The system has been tested on a population of 102 speakers extracted from the DARPA-TIMIT database. Perfect identification has been observed, specifically, an interval of confidence 95% for [99.9%,100.0%] recognition with a precision of 0.1%. The performances of our system are compared with those of a system based on multivariate auto-regressive models.

10.21437/ICSLP.1992-178

1. INTRODUCTION

We present in this paper a system for speaker identification which is operational for a large number of speakers, and which we have tested on a part of the TIMIT database. It concerns a complex system which integrates modules for the extraction of characteristics and modules for classification.

To develop this, we have used an approach which is novel in the connectionist domain. It concerns the decomposition of a principle task into a set of sub-tasks , such that the solution of the principle can be achieved through the solution of the sub-tasks. This leads us to conceive of a modular system where different networks cooperate on the same task. In such systems, the computational power is not a function of a single network with a large number of nonlinear units, instead it is a function of the cooperation between different modules operating on the same task.

This allows us to use networks of small size, and when the complexity of the problem increases, to simply add extra modules of the same simplicity. In the non-modular approach, the only response to an increase in the difficulty of a task is to use models of increasing complexity. Due to the increasing number parameters, the resulting systems become very difficult to train.

Our approach is totally general and can be used to solve problems that occur in completely different domains. In using this methodology we have simply transposed to the level of modularity, the philosophy of connectionism, and have created a kind of network of networks. In what follows, the advantages of this approach will emerge in detail.

The principle characteristics of our system are:

• A connectionist module based on an architecture of the type TDNN is used to extract characteristics.
This system works directly with the parameterized signal, thus performing a dynamic extraction of characteristics.

• The ability to handle a large number of speakers is made possible because of the modular architecture. This reflects the nature of the task being solved, and accounts for the typology of the speakers.

• The system functions in a mode independent of text. The training and testing are based on the internationally used DARPA-TIMIT database.

The advantages of our system are:

• The extraction of characteristics and classification are done by the same module and hence at the same time. The two stages are therefore optimal with respect to each other, which is rarely a feature of classical recognition systems.

• The architecture used for the extraction of characteristics accounts for the dynamic nature of the speech signal. This architecture is capable of representing the temporal relationship between acoustic vectors.

• The modules used are both simple and small, due to the modular architecture.

• The system in the identification phase, functions in "real time", which is not the case of other approaches.

• We have developed the system for a population which is statistically representative (102 speakers: 33 female and 69 male). This constitutes the most significant system (in terms of population size) that has so far been developed and tested. The philosophy underpinning its construction also allows the number of speakers recognised to be increased relatively painlessly.

We present firstly a brief description of the database used for evaluating the system. We then describe the preprocessing and analysis techniques used with the speech signal. A new TDNN architecture termed STDNN will be presented. After a description of an experimental study, we will present the results plus a comparison of the performances of our system with another technique (M-ARM).

2. GENERAL CONSIDERATIONS

2.1 Description of the Database

The database is composed of American English sentences spoken by 420 speakers representing 8 dialects [5].

Each speaker utters 10 sentences: 5 phonetically rich, which are denoted (SX), 3 represent natural sentences (SI), and 2 reflecting the dialect of the speaker (SA). The last two sentences are the same for all the speakers.

We have chosen the 5 phonetically balanced sentences for training our system, and the 5 other sentences for testing our system. The (SI) set are different for each speaker, hence the system is text independent.

2.2 Preprocessing and Analysis of Acoustic Signal

The speech has been recorded on a CDROM, at a sampling rate of 16 kHz and with a precision of 16 bits.

A LPC analysis of order 16 was preformed. Each frame of analysis was weighted by a Hamming window of length 26.5msec. The first 16 correlation coefficients of the signal were calculated every 10 msec. A pre-emphasis of 94% was applied to the speech samples.

The Cepstral Coefficients (LPCC) calculated from the linear prediction coefficients were also conserved.
After the LPCC parameterisation, each sentence was represented by a matrix of dimension 16xN, where N represents the number of frames in the sentence.

This parameterisation is the norm in Automatic Speaker Recognition (ASR). One of the reasons is that the parameters based on the analysis of linear prediction contain information about the formants, the glottal excitation and the lip transfer function [1]. We have used a significant part of the TIMIT database containing around one hundred speakers.

3. A MODULAR CONNECTIONIST ARCHITECTURE FOR ASR

3.1 Decomposition of the ASR task by module

ASR is in general much more complicated if text independent rather than text dependent. Moreover, contrary to verification, the applications of ASR require for the most part that the recognition be independent of text. Therefore, it is not possible to use simple techniques such as the matching of references.

The training demands a large amount of data, usually of the order of tens of seconds, and the testing requires several seconds of speech signal. Naturally, the time taken to train such a system grows with the number of speakers to identify. In non-discrminant systems , it grows in general, linearly with this number. In discriminant systems this growth is usually exponential. It should be noted that this is not necessarily the case in the recognition phase.

Concerning the extraction of characteristics from the parameterised signal, it is possible to utilize a representation based on the long term spectra as we have done in our first system [3].

It needs to be noted that because of the length of the training database, a requirement of the need to ensure text independence, the training stage is very slow with this type of representation.

Further, during the recognition phase, it is necessary to analyse the whole utterance before submitting it to the system. We therefore decided to treat the signal dynamically, i.e. to use analysis frames based on fragments of an utterance.

We wished to conserve a very important property of connectionist networks, which is the possibility of accounting for the inter-class structure of the data, and thus to achieve a discriminant training. Moreover, as we have already mentioned, the inclusion of this ability causes the training time to grow faster than linear as a function of the number of speakers.

It thus rapidly becomes impossible to implement a network as the number of speakers increases.

One response to this difficulty is to decompose the problem into sub-problems that are simpler to solve and to implement each sub-problem with a network. These networks are combined to form a modular architecture, cooperating on the global problem. The number of modules grows with the complexity of the problem to solve, but each one remains of a limited size, and is thus easy to train. Compared with a single system, the training time is greatly reduced and grows in a quasi linear fashion. Naturally the decomposition of the problem must represent its inherent structure. If this is the case, then the modular structure corresponds to the a prori knowledge we have of the problem. This task decomposition can be done automatically in an optimal fashion, as part of the training of the system.

In the first version of the system [4], we have tested the implementation of a simple modular architecture. The idea of even using the cooperation of different connectionist modules was completely new at that time, and there was no insurance of feasiblity.

During our work with ASR, we have noticed that certain characteristics permit the formation of homogenous and separable classes. As an example, the fundamental frequency F0 can be divided into two classes: female and male. This a prori knowledge about the global task can be incorporated into the sytem to implement two separate modules [4].

3.2 Incorporation of a prori knowledge in the architecture of the system

Within the classes of females and males we have further been able to distinguish many sub-classes. These group together the speakers who have similar vocal characteristics. In order to automate the determination of these speaker typologies by a connectionist network, we have labelled the training database by the speaker typology instead of their identity. The labelling was performed by a using a clustering technique of the type k-means followed by a majority vote.

In effect, the training acoustic vectors (SX) were grouped together into homogenous classes by a non-supervised k-means algorithm. These classes represent, thus, the different typologies of the population studied. After this regrouping of the acoustic vectors, we proceed to a majority vote by speaker and class. A speaker is then classified as belonging to the group to which the majority of the acoustic vectors belong.

After the regrouping of the speakers into homogenous classes, we have not encountered a single case of a female being classified as a male or vica-versa. On the other hand in most of the groups, a mix between the dialects has been observed, with one being dominant.

We have then implemented a modular system for ASR using this a prori knowledge of the task.

3.3 Architecture of the Connectionist System

The sytem is composed of a typology detector and a set of expert modules. Each expert module of the system is dedicated to the discrimination between speakers of the same typology (the confusions between members of the same typology are more important then those between different typologies).

The specialized module for the detection of typology plays a role of an information gating network. The architecture at this level of the system can be viewed in two forms [2].

The first case is where the typology detection module contributes to the final score in the form of a weight factor for the scores of the expert modules.

The second case is where the typology detection module serves to orientate the input towards the appropriate expert module.

It is, however, necessary to note that the identification time is noticeably less in the second case than the first. In effect the pre-selection of the expert module, avoids the need to compute the outputs of the other expert modules and hence considerably reduces the calculation time. However, with the first architecture, an error commited during the detection of typology will be compensated for by the expert modules, which is not the case with the second architecture.

For our simulations, we have not noticed a difference in performance between the two types of architecture, given that the typology detection module perfectly distinguishes between the speaker typologies.

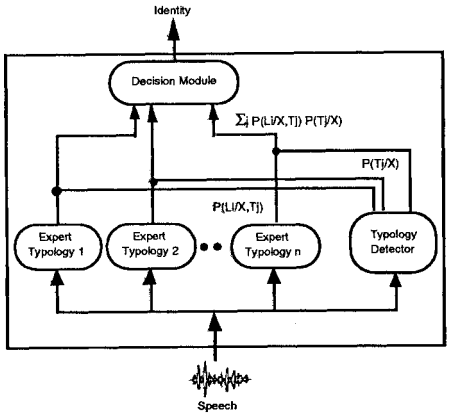

Fig. 1: Architecture of the Connectionist System.

System Composition. In automatic speech recognition, all the input data is indexed by time. To not take this into account imposes an extra heavy burden on the system.

The initial idea of taking into account the notion of time in a multi-layer perceptron, was introduced [8], under the name of TDNN (Time Delay Neural Network).

The components of the architecture of our ASR system are composed of TDNN type connectionist modules. The size of input of these networks is fixed, as in the majority of connectionist models. This poses a problem with speech data, where the sentences are not all of the same size.

In our previous system [3], the modelling was performed by a statistical technique. This technique allowed us to represent each sentence by the average spectrum and the principle eigenvector of the covariance matrix. This technique of modelling is independent of the size of the sentence and produces fixed sized input for the connectionist classifier.

Here, we proceed differently by sliding a fixed sized window over each sentence. Each position of the window produces an input for the system. The set of data concerning one sentence will be characteristic to each speaker. More specifically we proceeded in the following fashion.

We divided each sentence (the LPCC parameters calculated from the signal) into a set of successive windows. Each window is composed of 25 spectral vectors or frames, with an overlap of 20 frames. These windows constitute the input of the three modules described above.

The Training Phase of the System. During the training phase, the successive windows obtained from the decomposition of the sentence will be used to determine the parameters of the system. The training set of a module will be composed of the set of windows obtained in this fashion from the training sentences of the database.

The Identification Phase. For the identification, all the frames comprising a sentence are successively presented to the system in the form of the 25 acoustic vectors. At each presentation of a window, the system produces a response (the identity of the most probable speaker). The successive activations of the system are accumulated over the duration of the sentence. The final decision is accorded to the speaker with the largest score.

We will call this type of TDNN, STDNN (for shift TDNN). We have noticed in our experiments that after training, a set of three successive windows is sufficient for a perfect identification. If this is compared with existing systems, these three windows correspond to an extremely short utterence (less than a second) for an identification independent of text. The majority of systems use utterenecs of the order of several seconds, for the TIMIT database, [9] for example have used 8 second utterences.

4. MULTIVARIATE AUTO-REGRESSIVE MODELS (M-ARM) FOR ASR

This technique was used for the first time for speaker identification by Y. Grenier [6].

He considered that "the auto regressive model calculated from the voice of a speaker , models the articulatory capacities of a speaker, at least to a first approximation".

Based on the concatenation of 5 training sentences, a model is constructed. Each speaker thus possess one model. In the identification phase, a model is calculated from the test sentence.

The final decision consists of comparing the test sentence model with all the reference models of the speakers and choosing the closest model to the test, to determine the identity of the speaker. The order of the model is difficult to determine; we have fixed it at 2, taking into consideration the amount data available.

The description of this technique is given in [2].

5. RESULTS AND COMPARISONS

The results in tab. 1 show the superiority of the connectionist multi-module approach in comparison to the M-ARM technique.

Approach	*Score*	*Precision*	*95% Conf. Int.*
STDNN	100 %	± 0.1 %	[99.9 %, 100 %]
M-ARM	97.5 %	± 6.1%	[91.4 % , 95.6 %]

Tab. 1: Comparison of the two approaches: Connectionist and M-ARM

It can be seen that there is a very large difference between the size of the confidence intervals for the two techniques. This is due to the fact that M-ARM models require the entire sentence for model estimation and for identification, which results in the size of the test set being equal to the number of test sentences per speaker (5 per speaker). In contrast, the connectionist approach requires only two or three windows of 25 frames for perfect identification, which results in a large number of test examples (approx. 100 per speaker). The confidence interval is computed as a function of the test size and the rate of recognition (the exact method is given in the Appendix). As a consequence of the large test size, the confidence interval is small and conversely a small test size results in a large interval of confidence. The size of this interval can be viewed as a reflection of the validity of the system's performance.

The advantage of our connectionist approach compared to those of other techniques is essentially:

• The time required for an identification, at less than one second, it effectively real time.

• Only a short duration of the speech signal is required.

• The validity of the system performance.

6. INTERPRETATION OF THE FUNCTIONS CALCULATED BY THE MODULAR NETWORK

We briefly give below a probabilistic interpretation of the function of our modular networks. We will denote by L_i (i=1 .. m) the identity of the speaker i, and by T_j (j=1 .. n) the jth typology of the population under study.

If, for example, the architecture in fig. 1 is considered, the typology detection network is going to calculate, for an acoustic pattern X, an approximation of $P(T_j/X)$ for j=1..n. The expert networks which are each trained independently, are going to compute an approximation of $P(L_i/X,T_j)$ for network j (for the function calculated it is considered that $P(L_i/X,T_j)=0$ if L_i is not in T_j.

Thus for output i, overall the system computes an approximation of:

$$\sum_{j=1}^{n} P(L_i/X,T_j)\, P(T_j/X) \qquad (1)$$

which is recognisable as $P(L_i/X)$.

The system in fig. 1 thus computes through its modular architecture an estimation of $P(L_i/X)$ [2] by the means of successive approximations of the probabilites in equation (1). This also presents the following problem: for the values calculated at the output, the approximation errors of successive networks are accumulated.

It is possible in theory to train the same network to predict directly at the output $P(L_i/X)$. In this case an estimation of $P(L_i/X)$ would be obtained which is the best approximation in a least squared sense for the architecture used and consequently will be different to that produced by (1).

7. GLOBAL TRAINING ALGORITHM

To implement a global training of the system in fig. 1, an algorithm of adaptive gradients similar to back propagation, can be used. The only differnce between the classical MLP architecture and that of the system depicted in fig. 1 is the presence of the final multiplicative connections between the typology module and the expert modules.

An algorithm which permits an implementation of such connections has been, for example, proposed by [7]. Its derivation is very simple and is similar to that of back-propagation.

A good solution, without doubt, consists of initializing the system with the local algorithms, and then attempting to improve the system by using a global algorithm. This approach has already proved fruitful for other problems and other systems.

8. DISCUSSION

We have performed a set of experiments which have demonstrated the validity of the connectionist approach for ASR independent of text. These models extract the necessary traits for the discrimination between speakers whilst at the same time performing the classification. Our modular approach is very general, and has allowed us to apply a priori knowledge to the problem, and to solve the global task by decomposing it into simpler subtasks. This approach also considerably reduces the calculation time in the system and allows identification to be performed in real time.

Finally, the comparison of our system with other different techniques has revealed the power and interest of the connectionist approach to ASR. This comparison has given us the idea to use the cooperation between these different techniques so that their respective merits can be combined in the conception of hybrid systems.

9. REFERENCES

[1] Atal B.S. *"Effectiveness of LPC characteristics of the speech wave for A.S.I and A.S.V"*, JASA-Vol.55.1974.

[2] Bennani Y. *"Approches Connexionnistes Pour la Reconnaissance Automatique du Locuteur : Modélisation & Identification"*, Ph.D. thesis, University of Paris-Sud, 01-92. 1992.

[3] Bennani Y., Fogelman F., Gallinari P. *"Text-Dependent Speaker Identification Using Learning Vector Quantization"*, proc. of INNC, July Paris, France. 1990.

[4] Bennani Y. & Gallinari P. *"On The Use Of TDNN-Extracted Features Information In Talker Identification"*, proc. of ICASSP, S6.5, Toronto, Canada. 1991.

[5] Fisher W., Zue V.,Bernstein J., Pallett D. *"An Acoustic-Phonetic Data Base"*, J. Acoust. Soc. Amer. Suppl. (A), 81, S92. 1987.

[6] Grenier Y. *"Utilisation de la Prédiction Linéaire en Reconnaissance et Adaptation au Locuteur"*, proc. of XIème JEP, Strasbourg, pp. 163-171. 1980.

[7] Hampshire J.B. & Waibel A.H. *"The Meta-Pi Network: Connectionist Rapid Adaptation For High-Performance Multi-Speaker Phoneme Recognition"*, proc. of ICASSP, S3.9, NM, USA. 1990.

[8] Lang K. & Hinton G. *"The development of the Time Delay Neural Network Architecture for Speech Recognition"*, Carnegie Mellon University TR CMU-CS, N° 88-152. 1988.

[9] Rudasi L. & Zahorian S.A. *"Text-Independent Talker Identification With Neural Networks"*, proc. of ICASSP, S6.6, Toronto, Canada. 1991.