A Modular Connectionist Architecture
For Text-Independent Talker Identification

Younès BENNANI and Patrick GALLINARI

Laboratoire de Recherche en Informatique
Université de Paris-Sud, Centre d'Orsay Bât. 490
U.A. 410 C.N.R.S. 91405 ORSAY FRANCE
e-mail : younes@lri.fr

ABSTRACT

Only a few experiments have been reported in the literature on the use of connectionist models for Talker Recognition. We propose here a new model for text-independent talker identification which uses TDNN extracted features information. This model has been tested on 20 speakers half males and females from the TIMIT database using an LPC parameterization. An average identification of 98% has been observed.

1. INTRODUCTION

Talker Recognition (TR) has received a great deal of attention among speech researchers. Talker Recognition is a term generally used to refer to three related functions: Talker Verification (TV), Talker Identification (TI) and Talker Change Detection (TCD). In TV, an identity is claimed by the user, and a binary decision is made whether to accept or reject this claim. In TI, the machine is required to make an absolute identification from a population of talkers. Finally, in TCD, the computer locates the transitions between talkers in a stream of speech (e.g. conversation). A review of TR can be found in [9].

We are interested here in TI. Unlike verification, this problem has not yet received a tremendous amount of attention in the literature. This is probably due to the fact that TV is considered to be a more tractable problem and of more practical value than TI or TCD. Also in TI, the decision logic for identification is much more complicated, and whereas TV is characterized by cooperative subjects, this is not necessarily the case for TI or TCD. Only a few experiments have been reported on the use of connectionist models for TR. The earliest references seem to be [8] which describes a system using MLPs with one Neural Network (NN) per speaker, and [3,4] where each speaker is represented by a codebook in the LVQ algorithm. These two approaches are text dependent.

We describe in this paper work on TI, modifying and extending previous work [3,4] which used a connectionist classification system.

The extension concerns the following points:
- a connectionist system based on Time-Delay Neural Networks (TDNN) [11] is used for feature extraction instead of classical statistical techniques.
- the basic text-dependent system has been extended to provide a text-independent mode using part of the TIMIT database [5].

The goal of this system is thus more ambitious and realistic than that of our previous system. The advantages of our approach are the following:
- because feature extraction is performed together with decision making, it is optimal for the TI task. This was not the case in our previous system which used long term averaged statistical features. This is also an advantage of connectionist systems over classical TI approaches.
- the TDNN architecture used for feature extraction takes into account part of the "dynamic nature of speech". It is able to represent temporal relationships between successive acoustic frames, while providing some invariance under time translation.
- the system is fast when operating in recognition mode. Furthermore, TDNN is a very efficient data reduction pre-processor. The connectionist classifier operates onto a low dimensional space, which greatly reduces the size of its codebook.

We first present in 2 the database, in 3 we describe our system, in 4 and 5 we present our results and discuss some perspectives.

2. GENERAL CONSIDERATIONS

2.1. The Database

We have used for our experiments a part of the TIMIT database. A full description of this database can be found in [5]. It consists of 630 speakers, 438 males and 138 females. The speakers are categorized into one of eight "dialect regions" that approximately map to speech dialects in American English.

Each speaker spoke ten utterances, each being one sentence. Two of the ten sentences are "dialect sentences" that were spoken by every speaker, other sentences are different for each speaker. Five of them are "MIT" sentences, iteratively designed by hand with the goal of providing a rich variety of phonetic segments and phonetic contexts. The remaining three utterances are "TI" sentences. They were designed by taking a large corpus of written text.
2.2. Preprocessing of the analogic signal

The analogic recordings have been digitized at 16 KHz on 16 bit PCM. The samples are then pre-emphasized by a first order digital filter with transfer function:

\[ H(z) = 1 - 0.94.\downarrow \]

2.3. Digital signal analysis

A 25.6 ms Hamming window is used to perform short time analysis, resulting in a parameter vector sequence. The time interval between two successive analysis windows was 10 ms.

We have used a 16th order autocorrelation Linear Prediction Coding (LPC) [1] as parameter space. Each of the 200 sentences is thus converted into a 16xN array, where N is the number of frames in the sentence.

This parameterization technique has been widely used for automatic speaker recognition. One of the most important reason for this is that it contains a lot of information about the signal: the predictor coefficients represent the combined information about the formant frequencies, their bandwidth, and the glottal wave [2].

3. SYSTEM ARCHITECTURE

3.1. A modular architecture

Because the TI task we are dealing with involves a very important amount of data, training a system to perform identification on such data is very time consuming. One possible solution is to break the TI task into smaller subtasks and to use a NN for each of these subtasks. These NN will be smaller than what would have been necessary for the global task and therefore easier to train.

We have thus built a three modules architecture (figure 1). Data is first processed through a first net M1 which discriminates between males and females. This is usually an easy task. Depending upon this decision, the token is then sent to one of two other nets, M2 or M3 which is specialized respectively in the identification of male and female speakers.

Motivations for this modular architecture are multifold. First it has long been observed that speech spectra tend to form two clusters according to the sex of the speaker. This phenomenon has been observed in our first TI system [3,4].

Second, training large nets is computationally intensive. Finally this approach is very general and allows to decompose a global task into subtasks which are easier to solve. This allows us to incorporate into the architecture some knowledge about the task and therefore helps the network find a solution.

Of course the global performances depend on those of the first module M1. As will be seen in section 4, in our case, this net allows for a perfect discrimination.

The modules we have used are NNs with input layers of fixed size, which is the case of most connectionist models. This means that they can only classify static patterns of a given size. Of course, this is a problem when dealing with speech data. There are many solutions to this difficulty.

In our previous system [3,4], identification was performed by first computing long term statistics over a whole sentence. After that, these statistics which correspond to a vector whose dimension is fixed and does not depend upon the length of the sentence, were input to a connectionist system for classification. This results in heavy computations especially for long sentences.

Here, we proceed very differently by considering successive windows of fixed length over each sentence. More precisely, we do as follows:

We divide each sentence (LPC parameters) into successive windows. Each window has 25 frames wide and overlaps with the next one by 24 frames. These windows will be the inputs to the three nets.

During training, we select randomly, for each sentence, a given number of local windows (70 in the experiments we describe here) and train the nets to learn the association between each of these patterns and the corresponding speaker's sex (for M1) or identity (for M2 or M3), depending on the module.

II-858
For the recognition, the whole sentence is scanned, and the successive windows (whose number depends on the length of the sentence) are input to the net one after the other. For each presentation of a window, the net will compute an output vector. The successive output activations are summed and the decision rule is to choose the speaker with the higher score.

Of course, a 25 frames wide window will usually not contain enough information about the speaker identity. However, the successive windows which correspond to a whole sentence contain all the necessary information and can be used during the recognition step. 25 frames correspond approximately to 0.6 s of speech signal, justifications for this particular choice can be found in earlier studies by Atal [2].

3.2. A TDNN Architecture for TI

TDNN is a now well known connectionist technique which has been used extensively for speech recognition [11]. It is a particular case of a MLP with local connections and equality constraints between some of the weights. This architecture provides the network with interesting capabilities for speech analysis. It allows us to deal with time invariance and also with speech variations. Hidden cells learn during training to extract significant features from the speech signal. This network has been trained using a back-propagation algorithm [11].

The two modules M2 and M3 are similar. They are three layer nets with the following architecture: the input layer has 16 x 25 cells which corresponds to 25 successive time frames (~0.6 s) over the signal. The first hidden layer has 12 feature extractors or independent cells which are replicated 21 times. Each cell is connected to 5 consecutive time frames and this local window is shifted one frame to the right onto the input signal to provide the input to the next hidden cell in this layer. Thus hidden cell n°1, for example, is connected to time-frames 1 to 5 and hidden cell n°2 to time-frames 2 to 6. The second hidden layer has 10 feature extractors, connected to 7 consecutive columns of 12 cells in the previous layer with an overlap of 6 columns. The output layer is fully connected to the last hidden layer. We will denote this net by (16*25, 12*21, 10*15, 10) (5*1, 7*1), where numbers in the first parenthesis give the dimension of the successive MLP layers and those in the second term correspond to the size and the shifts of the local windows.

Windows onto the input signal are larger than those classically used for phoneme recognition [11]. The reason is that TI needs global information about the input signal and not only local information as this is the case for phoneme recognition.

The first module M1, for sex discrimination is a TDNN whose architecture is (16*25, 8*19, 4*15, 2) (7*1, 5*1) with the above notations. We have also performed different tests with other architectures but they did not yield better results.

4. EXPERIMENTAL STUDIES AND RESULTS

4.1 Experiments

We report here results of our text-independent system obtained with 20 speakers from the first dialect “dr1”, 10 males and 10 females. We have used the MIT sentences for training because they contain a great variety of phonemes uttered into many different contexts, and the TI and dialect sentences for the test because they are representative of natural language. For each speaker, we thus have 5 sentences for training and 5 for testing.

4.2 Results

Perfect discrimination for the sex of the speaker has been observed for all our experiments. The two other modules identify the speakers with a performance rate of 98% both for males and females. The global performance of the system is then 98%. Figure 2 gives detailed performances for the different speakers and sentences. TI on this corpus is a very difficult task. These results show that our architecture behaves very well on this problem.

5. DISCUSSION

We have reported experiments that demonstrate the interest of connectionist models for text-independent TI. These models allow us to perform some feature extraction together with classification by using an integrated formulation. The error rate is very low for this difficult TI task. We are currently working on larger systems which should allow us to identify an increased number of speakers from different dialects in the TIMIT database. These systems are modular in the sense that they are built from several NNs, each of them performing a part of the TI task. They interact together to solve the global TI problem. We think this is a very general approach for solving large problems. It can be used for more complicated tasks. For example we have begun to work on TI over the 8 dialects of TIMIT. Each module M2 and M3 is built from several modules: one specialized for each dialect. Furthermore, large identification systems like the

---

1 Lippmann demonstrated that three layers can encode arbitrary pattern recognition decision surfaces.
one described above can help improve performances of continuous speech recognition systems as it has been observed in [7].

6. REFERENCES


Figure2: Curves show the outputs of the modules M2 and M3 for the test sentences. There is one graphic per speaker whose name is written along the y axis. The ten graphics on the left part of the figure correspond to male speakers and the ten others to female speakers. The y axis gives the outputs of the corresponding net (male or female) for each test sentence from a given speaker. These outputs have been normalised between 0 and 1 and they sum to 1 for each sentence. The x axis corresponds to the ten different speakers for each net.