INTERNATIONAL COMPUTER SCIENCE INSTITUTE

1947 Center St. • Suite 600 • Berkeley, California 94704-1198 • (510) 643-9153 • FAX (510) 643-7684

Simultaneous speech and speaker recognition using hybrid architecture

Dominique Genoud, Dan Ellis, Nelson Morgan

TR-99-012

July 1999

Abstract

This rapport summarize the work that was done this last 6 month at ICSI in speaker recognition and speaker adaptation.

1 Introduction

The automatic recognition process of the human voice is often divided in *speech recognition* and *speaker recognition*. The 2 area use the same input signal (the voice), but not for the same purpose: the speech recognition aims to recognize the message uttered by any speaker, and the speaker recognition want to identify the person who is talking. However, more and more applications need to use simultaneously the 2 kind of information. Some actual examples given below illustrate this tendency.

State-of-the-art speech recognition systems tend to be speaker independent by using models (phonemes, diphones, triphones) estimated on huge databases containing numerous speakers, and also by using parameterization which try to suppress the speaker dependent characteristics (PLP,RASTA-PLP). However, for some types of applications it could be important to readapt the speaker independent speech recognizer to a defined speaker, in order to improve the noise robustness for example, or simply to improve the speech recognition performances by adding some knowledge of the speaker. Some recent results shows that **speaker adaptation** of a speech recognizer improve the performances of the systems [DARPA, 1998].

Nowadays, numerous applications performing speech information retrieval require the automatic extraction of the content of shows and the retrieval of the speech of a particular speaker on a particular subject. In this case a speech recognition and a speaker recognition should be carried on in parallel. Furthermore the detection of speaker change in a conversation (speaker A/ speaker B or speaker/music) may also be very useful for the indexing and the labeling of the huge databases available.

Finaly, a speaker recognition is needed for applications like secured voice access to information (as a bank account or a voice-mail box) . In this case, the speaker recognition can be **text independent** if the content of the utterance is not checked. However, better results are obtained by using **text dependent** speaker recognition, both because a control of what is said can be done and also because more accurate models (phonemes, words) can be built. Anyhow, the text dependent speaker recognition has to be preceded by a speech recognition step to control and split the message properly.

All these applications show the need of a *simultaneous* speaker and speech recognition. This rapport shows that it exists some possibilities exist to carry out this 2 tasks simultaneously.

2 Background

This section will recall some useful notions about speech and speaker recognition domains.

2.1 Some definitions

2.1.1 Type of speakers

The speakers which have to be identified (verified) by our system will be named **registered speakers**, ***RS*** , the speakers which will attempt to impersonate the registered speakers will

be called **impostors**. We also use the voice of many other speakers which will constitute the **world speakers**.

2.1.2 Type of speaker recognition applications

Speaker recognition applications are classified by their text dependency, they can be **text dependent**, **text independent**, or **text prompted**, the two latter cases imply a control of the text. The speaker recognitions applications can also be classified by the way that the identity of the speaker is checked: if the voice of an unknown speaker is used directly to compare to references of enrolled speakers we perform an **identification** of the speaker. If the identity is checked by another mean (password,identification number, etc...) a **verification** that the input voice belongs to the identified speaker.

2.1.3 Type of speech recognition applications

The speech recognition applications can be **speaker dependent** or **speaker independent**, however this notion becomes a little bit fuzzy when a speaker adaptation of the speaker independent speech recognizer is performed. The speech recognizer can be classified by the size of the vocabulary they can handle like **small, medium or large vocabulary applications**.

2.1.4 Into the systems

The speech/speaker recognition systems are built into two parts:

1. A training phase where the parameters of the models of speech recognizer or of the registered speakers are estimated, using *known target sentences*.

2. A test phase where unknown sentences uttered by registered speakers (often called true speaker test) or by impostors are given to the system. The speech recognizer will produce a sequence of words using the models estimated in the training phase. The speaker recognizer will produce a score which will be compared to a threshold, if the score is greater than the threshold, the utterance will be accepted as pronounced by the ***RS*** which is tested.

3. A third phase is often used to set the a *priori* thresholds used to measure the performances for the speaker recognition system ([Bimbot and Genoud, 1997; Pierrot *et al.*, 1998]). However in this rapport a theoretical speaker independent threshold will be used.

2.2 The Log Likelihood Ratio (LLR) as speaker verification score

When using statistical algorithms, the LLR is the main score computed in speaker recognition, because his strong relationship with the statistical models themselves ([Green and Swets, 1988; Scharf, 1991]). The decision of accepting or rejecting the utterance of a registered speaker can be seen as an hypothesis test H_0 (the speech segment belongs to the ***RS***) against H_1 (the speech segment doesn't belong to the ***RS***). Which is identical than

testing the conditional probability of an event X knowing the hypothesis H_0 and H_1 (see equation 1).

$$H_0 : \textit{ Registered speaker, } H_1 = \overline{H_0}$$

$$P(H_0) \underset{reject}{\overset{accept}{\gtrless}} P(H_1), \quad P(H_0|X) \underset{reject}{\overset{accept}{\gtrless}} P(H_1|X) \tag{1}$$

The quantities $P(H_0|X)$ and $P(H_1|X)$ are called a *posteriori* probabilities of the hypothesis H_0, respectively H_1 knowing the event X.

Since it is not possible to model the *"non registered speaker space"* (virtually all the other speaker [past, present and future] of this planet living or having lived at the same time than the ***RS***), the hypothesis H_1 is translated in: *"the segment belongs to a speaker among a large amount of people which are neither registered speakers of the application, nor possible impersonators"*. H_1 will be modeled by a *world model* estimated by the voice of many speakers excluding the registered speakers and the speakers used as tuning impersonators of the application.

Translated into the speaker verification problem, the test of a sequence of observations O_t (a sequence of parameters) knowing a statistical model M_C for each ***RS*** and a model M_W for the world, the equation 1 can be rewritten as (equation 2):

$$P(M_C|O_t) \underset{reject}{\overset{accept}{\gtrless}} P(M_W|O_t) \tag{2}$$

Using the first Bayes rule (equation 3) and if the *a priori* probability $P(M_C)$ and $P(M_W)$ are known (there are often assumed equal) **the *a posteriori* probabilities** $P(O_t|M_C)$ and $P(O_t|M_W)$ can be estimated (equation 4):

$$P(B/A) = \frac{P(A/B)P(B)}{P(A)} \qquad \text{(Bayes)} \tag{3}$$

$$P(M_C|O_t) = \frac{P(O_t|M_C) \cdot P(M_C)}{P(O_t)} \underset{reject}{\overset{accept}{\gtrless}} \frac{P(O_t|M_W) \cdot P(M_W)}{P(O_t)} = P(M_W|O_t) \tag{4}$$

Which can be transformed in:

$$\frac{P(O_t|M_C)}{P(O_t|M_W)} \underset{reject}{\overset{accept}{\gtrless}} \frac{P(M_W)}{P(M_C)} \tag{5}$$

The estimation of the 2 a *posteriori* probabilities $P(O_t/M_C)$ and $P(O_t/M_W)$ is carried out by using a **maximum likelihood** estimation function [Scharf, 1991]. The quantity LR is the Likelihood ratio of the observation sequence O_t knowing the 2 models M_C and M_W. The equation 5 becomes:

$$LR = \frac{\mathrm{L}\ (O_t, M_C)}{\mathrm{L}\ (O_t, M_W)} \underset{reject}{\overset{accept}{\gtrless}} \frac{P(M_W)}{P(M_C)} \tag{6}$$

In order to decrease the computing resources, and due to the normality properties of the logarithm, the *log* of the equation 6 is taken (equation 7) which define the **Log Likelihood Ratio (LLR)**:

$$LLR(M_C, M_W, O_t) = \log \mathrm{L}\ (O_t, M_C) - \log \mathrm{L}\ (O_t, M_W) \underset{reject}{\overset{accept}{\gtrless}} \log\left(\frac{P(M_W)}{P(M_C)}\right) \tag{7}$$

With $\log \mathrm{L}\ (O_t, M_C)$ and $\log \mathrm{L}\ (O_t, M_W)$ the log likelihood of the ***RS*** , respectively of the world computed on the observation sequence O_t.

2.2.1 Error function in speaker recognition

In a practical application, we try to minimize the *total cost of error*. For this purpose a *cost function* can be defined as the sum of the errors done by accepting wrongly speech utterances which do not belongs to the ***RS*** (**False Acceptance, FA**) and rejecting wrongly speech utterances which belongs to the ***RS*** (**False Rejection, FR**):

$$c_{tot} = c_{fr} \cdot P(C) \cdot E(FR|C) + c_{fa} \cdot P(\overline{C}) \cdot E(FA|\overline{C}) \tag{8}$$

With c_{fr} and c_{fa} the cost of a false rejection, respectively a false acceptance. These costs are established according to the application needs. $P(C)$ and $P(\overline{C})$ are the a *priori* probabilities that the test sequence belongs to the ***RS*** or not.

$E(FR|C)$ and $E(FA|\overline{C})$ are the false rejection (i.e. reject falsly an utterance belonging to the registered speaker) and false acceptance (i.e. accept falsly an utterance of an impostor) error rate done by the system. It can be shown that minimizing the cost function c_{tot} comes down to add the error costs c_{fr} et c_{fa} to the equation 4. Moreover, if we accept that $P(O_t|M_{\overline{C}})$ can be approximated by $P(O_t|M_W)$, the equation 4 can be re-written as:

$$\frac{P(O_t|M_C) \cdot P(M_C)}{P(O_t)} \cdot c_{fr} \underset{reject}{\overset{accept}{\gtrless}} \frac{P(O_t|M_W) \cdot P(M_W)}{P(O_t)} \cdot c_{fa} \tag{9}$$

Which, when using the equations 5, 6 and 7, eventually leads to the estimation of the following equation:

$$LLR(M_C, M_W, O_t) = \log \mathrm{L}\ (O_t, M_C) - \log \mathrm{L}\ (O_t, M_W) \underset{reject}{\overset{accept}{\gtrless}} \log\left(\frac{P(M_W)}{P(M_C)} \cdot \frac{c_{fa}}{c_{fr}}\right) \quad (10)$$

$$\log\left(\frac{P(M_W)}{P(M_C)} \cdot \frac{c_{fa}}{c_{fr}}\right) = \log(R) = \Theta \quad (11)$$

The quantity R, which is the decision threshold, is often called the **risk ratio**. This ratio is independent of the registered speakers, it only depend on *a priori* quantities, determined by the application conditions. If the threshold is computed *a posteriori* when knowing the distributions of the LLR, a point of **Equal Error Rate, EER** (FA=FR) can be computed. If the threshold is set *a priori* , then, the **Half Total Error Rate, HTER=(FA+FR)/2** is computed.

2.3 Speech recognition using Hidden Markov Models (HMM), Multi Layer Perceptrons (MLP) and Maximum *a posteriori* Probabilities (MAP)

Classical state-of-the-art speech recognition systems use stochastic models as Hidden Markov Models, with a Maximum Likelihood Estimation (MLE) criterion to estimate the proper sequence of words (phonemes) [Jelinek, 1976; Bahl and Mercer, 1983]. The modeling of the distributions in each HMM state is actually done by estimating a Gaussian mixture [Scharf, 1991; Rabiner and Juang, 1993; Kleinrock, 1975]. However, the MLE criterion has several drawbacks which can be corrected by using a Maximum a posteriori Probabilities criterion (MAP) leading to a discriminant modeling [Bourlard and Wellekens, 1990]. The MAP criterion can be elegantly solved by the use of Multi Layer Perceptron (MLP) as probability estimator. Nevertheless, it should be noted that the posterior estimation performed by the MLP directly estimate the probability that a particular input vector belongs to a particular output class, but doesn't give any information on how likely this particular vector can be observed, which would have been the case when using a full joint density estimation.

Althrough according to [Renals and Morgan, 1992], the MLP doesn't provide a full probability model, but draw the limit of the output classes, which is what we are interested in the speech recognition problem. The estimation mechanism of these probabilities is shown in figure 1. The context dependency is added by using a window of 9 vectors as input. In the experiments described here, these vectors are constituted of 12 mel cepstral parameters (see section 3.6). 2000 hidden nodes set up the hidden layer, and of the 54 nodes of the output layer, 53 are used for the phoneme probability estimation and one output for the *"non speech"* information. The output layer activation function of the MLP are adapted in order to guarantee that the sum of all the output equal one. This function is called *softmax* [Bridle, 1989] and is given in equation 12.

$$g(i,l) = \frac{e^{x(i,l)}}{\sum_{m=1}^{K} e^{x(i,m)}} \quad (12)$$

Where $x(i,l)$ is the output value of the unit l before the non linearity for an input value i.

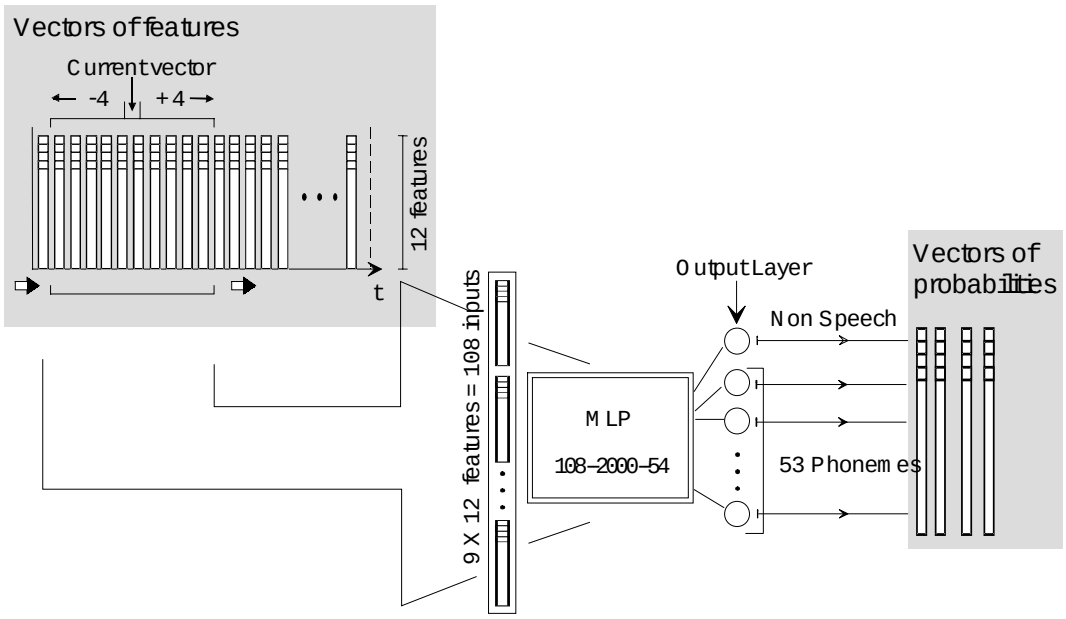

Figure 1: The MLP a *posteriori* probabilities estimator

2.3.1 Speech decoding

The frame by frame emission probabilities are then used in the same way than in classical HMM speech recognition approach using a Viteribi decoder, and a language model. The decoder used here is exactly the same the one used in the SPRACH system [Cook *et al.*, 1999], a principle diagram is shown in figure 2.

3 Database and parameterization

3.1 The Broadcast news database

The database used for all the experiments is the 1997 Broadcast news evaluation database from the Defense Advanced Research Project Agency (DARPA). This database was designed for the Hub-4E speech recognition evaluations organized by DARPA . The first corpus of 100 hours of speech (denoted bntrain97) is annotated to support "evaluation focus condition". The focus condition are based on the speaking mode (planned or spontaneous), the dialect (native or non native), the fidelity of the acoustic channel (high medium or low), and the background noise (music, speech or other). Further details about this database can be found in [Cook *et al.*, 1999] or [Fisher *et al.*, 1998]. This database was designed to train the speech recognition systems which were in competition. The signal is sampled at 16 [kHz], giving a bandwidth of 8 [kHz].

The database is constituted of 10 different types of shows on 4 different broadcast stations (ABC,CNN,CSP,NPR).

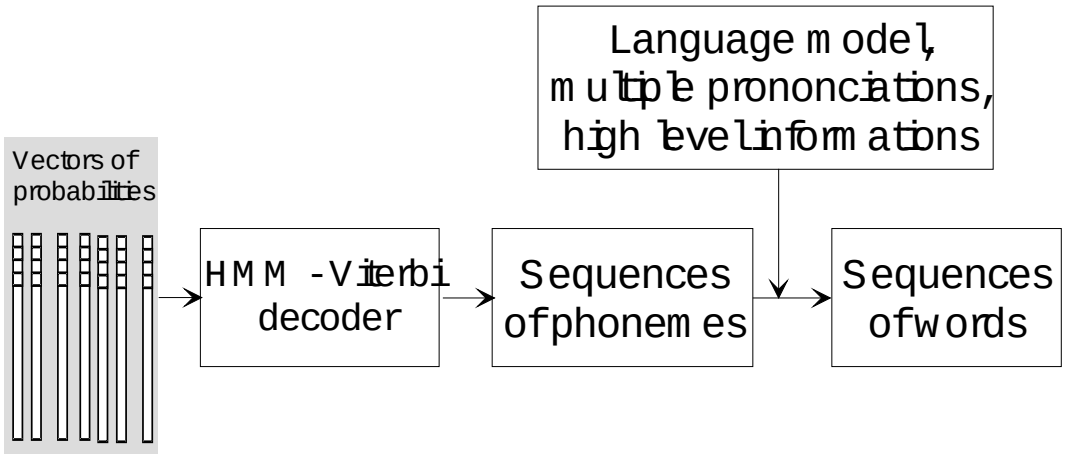

Figure 2: The speech decoding part of a speech recognizer.

3.2 Use of the database

To fit our purposes the database is divided in 3 different speaker sets:

1. **The registered speaker set** which will contain the speakers who will be used for the various speaker evaluations (speaker adaptation, recognition, detection), and also for the speech recognition evaluations. The voice of the speakers of this set are also used during the test to impersonate the other registered speakers.

2. **The world speaker set** consists of speakers other than the first set and will be used to build the *world model*. The voice these speakers will also be used to create the speaker independent models for the *speech recognition*.

3. **The tuning impersonator set**, these speaker are different than the 2 first sets and are used to setup a decision threshold for some applications. This set is not used in the experiments described in this rapport, and is reserved for future experiments.

3.3 Creating the registered speakers set

In order to perform speaker verification tests, the first step was to select speakers with enough data for our purpose.

The following criterions are used to elect a speaker as a registered speaker:

- At least 2 recording sessions.
- More than 1200 seconds of recording.
- the recordings can be on different shows.

Twelve speakers match this criterion:

Noah-Adams, Peter-Jennings, Thalia-Assures, Mark-Mullen, CSP-WAJ-Susan, Linda-Wertheimer, Brian-Lamb, Lou-Waters, Kathleen-Kennedy, Andrea-Arsenault, Chris-Wallace, Katherine-Calloway

3.3.1 training set

For each speaker the training set is constituted of around 1000 seconds of speech (see table 1).

3.3.2 test set

For each speaker the test set is constituted of at least 200 seconds (see table 1).

Speaker name	Training duration (in seconds)	Test duration (in seconds)
Noah-Adams	1000.46	3354.45
Peter-Jennings	1000.14	2304.69
Thalia-Assures	1006.25	2203.23
Mark-Mullen	1002.41	2104.18
CSP-WAJ-Susan	1002.96	1646.42
Linda-Wertheimer	1000.05	1972.75
Brian-Lamb	1002.97	1646.78
Lou-Waters	1005.27	1465.14
Kathleen-Kennedy	1013.06	1296.28
Andrea-Arsenault	1000.01	1092.85
Chris-Wallace	1004.91	549.20
Katherine-Calloway	1002.11	203.88
TOTAL	12040.61	19839.85

Table 1: Training and Testing amount of data available for each speaker.

3.4 Creating the *"World"* set

This subset is created using the voice of 100 speakers, constituting 14h20' of speech divided on 495 segments.

3.5 Creating the tuning impersonators set

The tuning impersonators are 19 speakers chosen with enough data and sessions there are:

Tom-Wicker, Bill-Hemmer, John-Kyle, Marshall-Freidy, Daniel-Ellsberg, Austin-Bay, Phil-Borgess, Bill-Maynes, Michael-Barone, Sheila-Jackson-Lee, Corva-Coleman, Robert-Francis, Vincent-Thompson, Anne-Garrols, David-Fromm, Gary-Hart, Ron-Elving, Chris-Beary, Cory-Flintoff

There are different than the speakers used as registered speaker or world speakers.

3.6 Parameterization

The input signal is sampled each 10 [ms] using a window of 25 [ms]. each window is then passed through a filter bank, and using a mel conversion function given in equation (13), 12 cepstral coefficients are extracted. No derivatives or accelerations are used.

$$M = 2595 \cdot \log_{10}(1 + \frac{frqHz}{700}) \tag{13}$$

With $frqHz$ the input frequency.

4 Speaker adaptation experiments

This section will talk about adapting an speaker independent speech recognition system to a particular speaker. Some previous experiments carried out by [J.Neto *et al.*, 1996] on MLP-hybrid architectures had already shown the possibility to adapt a MLP to a particular speaker.

However, the approach chosen here is much simpler. Indeed, the MLP probability estimator is first trained with the *world training set* (this model will be named M_W). Then for each registered speaker i ($i = 1 \ldots N$, with N the number of speaker registered into the application), a new model M^i_{RS} is created, using M_W as bootstrap and re-trained with the training data of each ***RS*** i.

4.1 Test protocol

10 sentences of each registered speaker are selected as a test set, constituting 110 test sentences[1]. Three experiments are then carried out:

1. The 110 sentences are given to the speech recognition system using the speaker independent MLP M_W trained on the *world speaker data*. This experiment gives the reference performances for a recognizer non particularly adapted to a particular speaker.

2. The sentences of a ***RS*** i are given to the recognition system using his own model M^i_{RS}. In this case, the model is adapted to the test sentences.

3. The sentences of all the registered speakers j with $j \neq i$ are given to the recognition system using the model M^i_{RS}. This experiment use models which are adapted to other speaker than the one who said the test sentences.

4.2 Results

The results obtained for these three different experiments are given in the **table 2**

These results show that some adaptation to a speaker by re-training the speaker independent MLP improves the speech recognition performances (i.e. decrease the word error rate). This imply that some specific information belonging to a particular speaker can be added to the MLP, helping to a better decoding (principally less substitutions). The line 3

[1] Unfortunately the training step for one of the ***RS*** failed, so for this experiment only 11 registered speakers are available.

Experiment	#Snt	#Wrd	Corr	Sub	Del	Ins	Err	S.Err
1.World Model M_W	110	4698	**77.29**	17.43	5.28	3.34	26.05	88.18
2.Adapted Model M^i_{RS}	110	4698	**80.99**	**14.41**	4.60	2.94	21.95	90.00
3.Impersonator Models $M^{j\neq i}_{RS}$	1100	46980	**65.57**	28.43	6.01	5.65	40.09	95.64

Table 2: performance of the speech recognition with adaptation to a speaker.

of the table 2 show also that adapting the recognizer to a particular speaker decrease the recognition performances of the system when ***it is not the proper speaker*** who speaks indicating that the generalization of the MLP is partly lost in the adaptation process.

5 Text dependent speaker recognition using hybrid architecture

The state-of-the-art statistical algorithms used in speaker recognition, and especially in **text dependent speaker verification**, use 2 kind of models, one which tries to capture the intra-speaker variabilities, and one which tries to model the *world* (see section 2.2). Then a LLR of an utterance is computed using the output probabilities of the 2 models (see section 2.2).

A first attempt was made by using the same procedure with MLPs as probability estimators: First a *world model* was build (exactly the same than under section 4). Then, for each ***RS*** , a *speaker dependent model* was created by re-estimation of the *world model* using the training data of the speaker. During the test, an unknown sentence is passed trough the speaker dependent model and the world model. To split the problem of speech and speaker errors, we assume that the phoneme alignment of the sentence is known (the aI5 alignment is used, see [Cook *et al.*, 1999]). Then, we subtract the log probability of the selected phoneme of the world model from the log probability of the same phoneme out of the speaker dependent model, giving a total score which should be higher for the true speaker tests.

The figure 3 shows the LLR scores for the true speaker test and the impostors attempts on the 10 most frequent words in the training set of the speaker Noah-Adams. This results show that no real differences between the distributions of the true access scores and the impostors scores can be made. Which make sense because the MLP, according to section 2.3 use a MAP criterion and not a Maximum Likelihood criterion. Thus, these Maximum *A posteriori* Probabilities depend on the speech content, but shouldn't depend on the speakers, if we admit that the *a priori* probabilities of the phonemes are the same for all the speakers. Moreover, as the outputs of the MLP give no information on the likelihood of a vector to belong to particular class (in this case to belongs to a particular speaker **and** a particular phoneme), using the subtraction of the scores coming out from the registered model minus the world model should be **zero**. However due to the imperfections of the modeling and the segmentation, what we obtain is distribution of scores centered in zero, with no differences between true speaker and impostor tests.

However, some experiments done for the NIST-97 text independent speaker verification

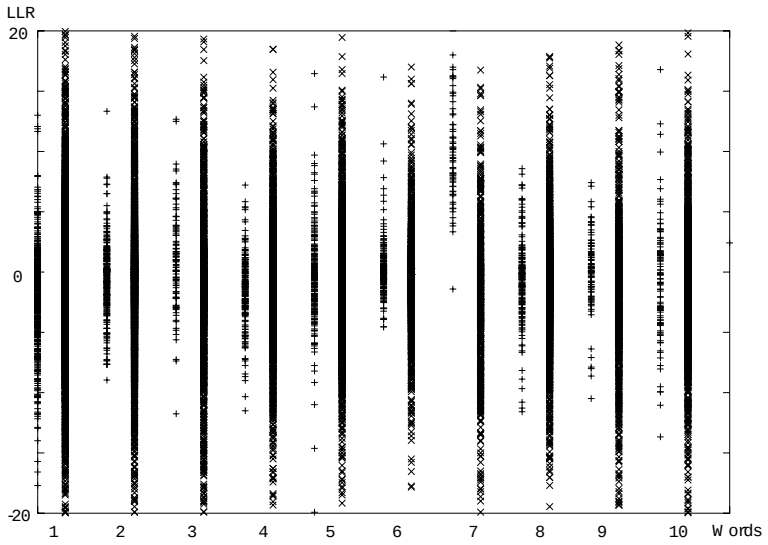

Figure 3: Distributions of true access (+) and impostor access (×) scores using a model for each registered speaker and one model for the *world*.

evaluation [Martin, 1997] have shown that MLP can be used successfully for speaker verification [Genoud and Caloz, 1997]. The way the MLP was used in that case is shown on the figure 4. One MLP is built for each registered speaker, and each MLP has only 2 nodes on the output layer, one is targeted on the ***RS*** data and the other on the world data. The frame by frames *log* scores of each output is summed on one sentence, and then subtracted $(O_1 - O_2)$ to perform in a same way than a LLR. The results were close to the state-of-the art GMM systems, even if no special channel normalization was performed [Martin, 1997].

Here the 2 output classes give the posterior probability to belongs to the speaker or to the word class, and the log ratio of these probabilities gives a good estimator of the log likelihood ratio used with a statistical classifier, because we have only 2 output classes (which correspond to the 2 hypothesis H_0 and H_1, see section 2.2).

6 Next step

From another point of view than the statistical one, a MLP can be seen as a black box which cluster the input vectors space into the output classes. The fact that the output become probabilities is due to the re-normalization carried out by the softmax function (see section 2.3). So, to perform simultaneously a phoneme clustering **and** a *registered speaker/world* partition, the output task must be changed. This can be performed by merging the 2 approaches. Thus, for each ***RS*** , a MLP with 2 times the number of phonetic classes is generated. One set of output will now be targeted on the phonetic clustering of the ***RS*** , and the other set will be targeted on the phonetic clustering of the world. As it makes no sense to have 2 times the *non-speech* output, this one is not duplicated. The MLP has now the structure shown in figure 5, and it will be called Twin-Outputs MLP,

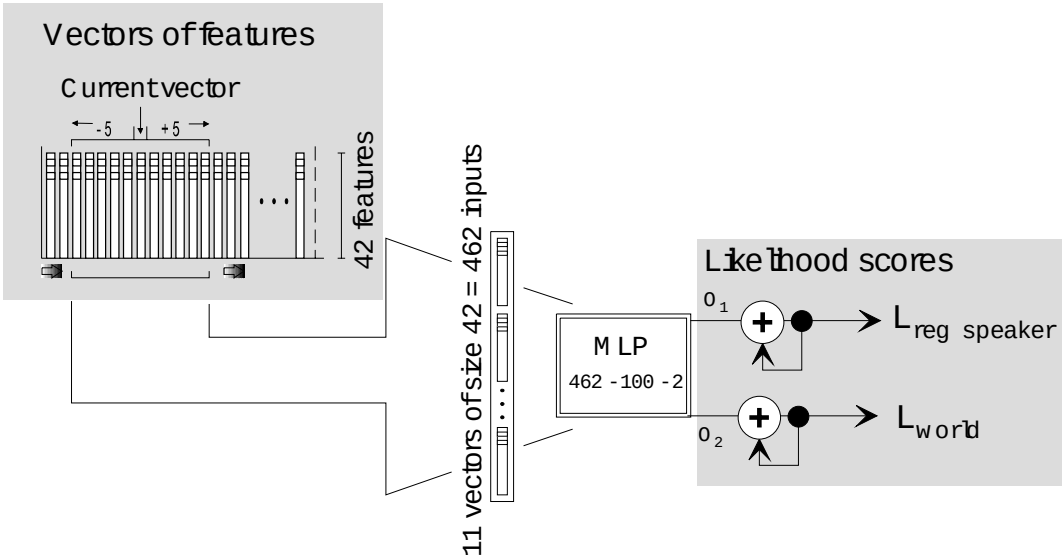

Figure 4: Distributions of true access scores and impostor access scores using a "statistical-like" approach with a MLP.

TO-MLP.

6.1 Training of the TO-MLP

As the TO-MLP has know to carry out two tasks simultaneously, it will be trained in 2 passes:

1. (a) The first phase estimates the proper phonetic classes for each input vector. So a normal MLP with one output for each of the 53 phonetic classes and one non-speech output is trained on the *world set* of the broadcast news database (See [Ellis and Morgan, 1999] for further details on the MLP training scheme).

   (b) The outputs of the MLP previously trained are then duplicated, leading to a MLP with the structure shown in figure 5. The connection between the hidden layer and the output layer are simply copied.

2. The second phase will train the clustering *registered speaker/world*. Thus, one T O-MLP is created for each ***RS*** . The speaker phonemes output are targeted on sentences belonging to the training set of the ***RS*** , the world output are targeted on the speech of the speaker used for building the speaker independent (*world*) MLP of the previous phase. In order to obtain a good learning effect during the training phase, the input sentences belonging to the ***RS*** and to the world speakers are presented alternatively to the TO-MLP. An heuristic is used to select sentences from the world set which have the same duration than the sentences of each ***RS*** set. Anyhow, the sentences of the world set used to train each of the ***RS*** T O-MLP are roughly the same, as the training

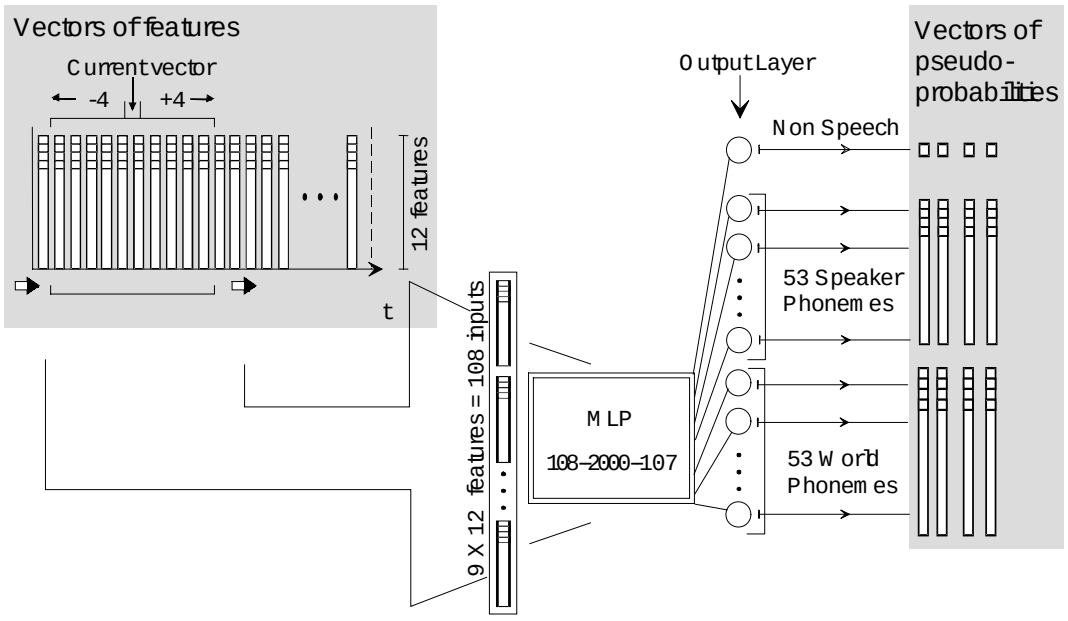

Figure 5:

duration of for each ***RS*** is nearly the same. The softmax function is still used on the output layer.

6.2 Using a TO-MLP

As two tasks are performed simultaneously, the outputs of the TO-MLP can give different estimates depending on the task. The outputs cannot anymore be used directly, and a post-processing has to be added. The following section will describe some examples of post-processing to estimate the frame by frame MAP used for the speech recognition, and how to compute the speaker recognition scores.

6.2.1 Speech decoding with TO-MLP

With the TO-MLP each phonetic class is represented 2 times in the output layer, so it is possible to use one of the 2 output sets or a combination of the two to compute the *a posteriori* probabilities. The table 3 gives the results obtained when using the speaker outputs, the world outputs and the mean of the 2 outputs. In all of these cases, a re-normalization of the output scores of each set is performed to have the output summed to one (see the Abbot system [Cook and Robinson, 1998] description for the details on re-normalization).

Model type	output type	#Snt	#Wrd	Corr	Sub	Del	Ins	Err	S.Err
World MLP		120	4950	76.85	18.08	5.07	3.35	26.51	90.83
Adapted TO-MLP	Reg speaker	120	4950	**80.97**	14.44	4.59	3.03	22.06	91.67
Impersonator TO-MLP	Reg speaker	1320	54450	**51.02**	26.67	22.31	3.18	52.16	97.42
Adapted TO-MLP	World	120	4950	**74.40**	19.27	6.32	3.39	28.99	95.00
Impersonator TO-MLP	World	1320	54450	**72.49**	21.77	5.74	4.30	31.82	95.61
Adapted TO-MLP	Mean	120	4950	**79.94**	15.23	4.83	2.97	23.03	92.50
Impersonator TO-MLP	Mean	1320	54450	**72.42**	21.68	5.90	4.01	31.59	95.23

Table 3: Speech recognition performances using different TO-MLP output combinations.

6.2.2 Speaker recognition with the TO-MLP

As the TO-MLP give the a *posteriori* probabilities of the ***RS*** against the world, a LLR-like score can be computed. First, the best phoneme is choosen by the speech decoder, and then LLR-like score is computed by subtracting the log probabilities output of the world from the output of the speaker for this elected phoneme. Of course, all the speech decoding possibilities explained in section 6.2.1 are available to choose the proper phoneme used to compute the LLR-like scores.

The speaker recognition can be performed at different level. The simpliest way would be to sum the LLR-like score over a sentence,take the mean and compare it to a threshold in order to decide if the utterance is belonging to the ***RS*** or not. The table 4 gives the results of this approach for all the ***RS*** , when using the aI5 (see section 5) phonemes decoding, and table 5 when using the mean decoding. These results are only exploratory, and should be balanced with a reference system. Howerver, they indicate that the principle of MLP retraining and simlutaneous speech-speaker decoding works raisonably.

It is possible, using the same decoding approach to perform a real text dependent speaker verification, using the world level decoding, the figure 6 shows the LLR-like scores of the registered speakers versus the impostors ordered by word length (the longer first). Of course there is a dependency of the length (more phonemes) but there is still a nice discrimination even for shorter words.

6.2.3 Speaker detection with TO-MLP

In some speaker detection problems, it is important to know when the speaker changes in a conversation, or how long a speaker has to speaker to be detected by the recognition system. The figure 7 gives an example for the speaker Noah-Adams: the mean LLR are sorted by utterance length and it shows that after around 8 seconds of speech, the speaker is detected without any errors.

The TO-MLP approach allow to compute simultaneously and extemely fast the speaker/world decoding. Some limited attempt to perform the same thing with conventional multigaussian HMM (see for example [Mariéthoz *et al.*, 1999]) are, for the moment, very limited, the synchronous decoding cannot be performed in one pass because a slience dectection as to be made.

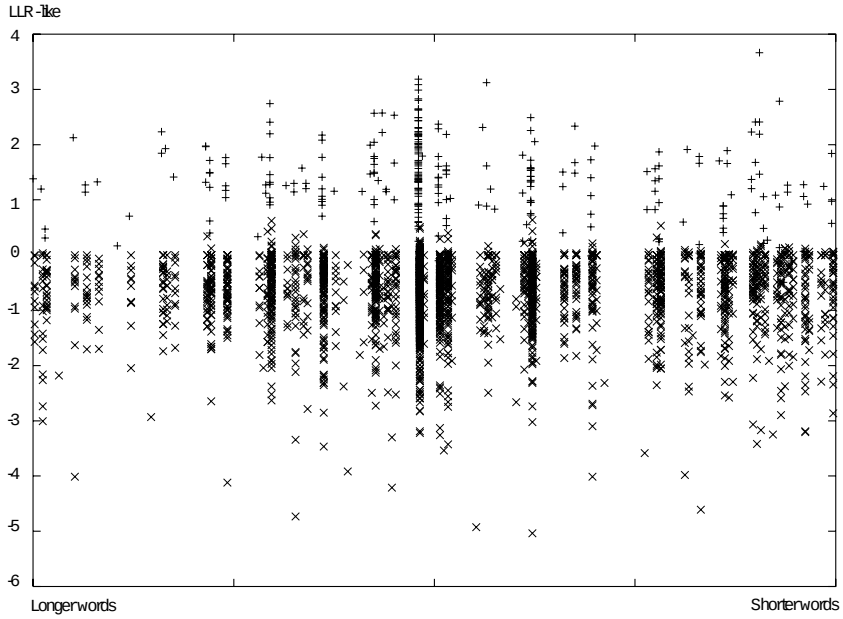

Figure 6: Values of true speaker (+) and impostors (×) mean-LLR-like scores sorted by length of sentences of the test set.

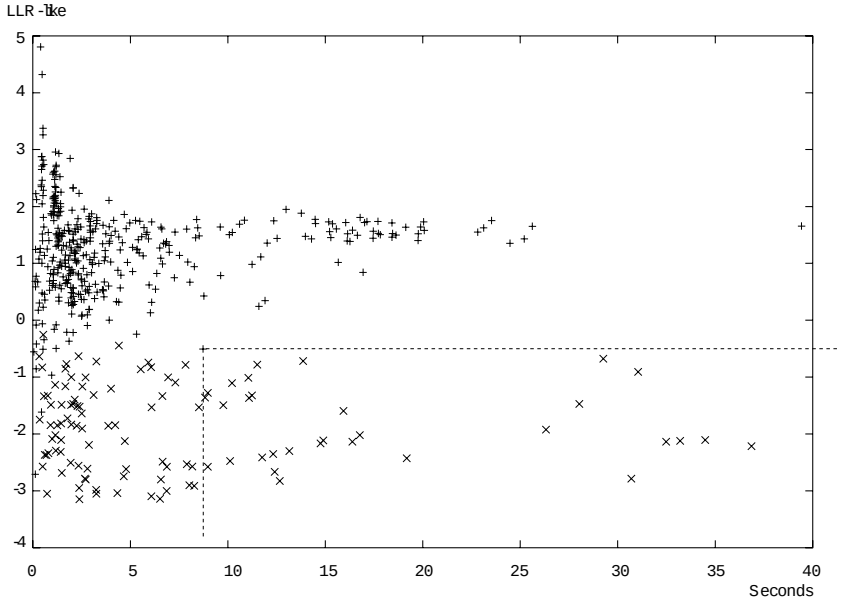

Figure 7: Values of true speaker (+) and impostors (×) mean-LLR-like scores sorted by length of sentences of the test set.

Speaker Name	EER%	Θ-EER	FR% (hardΘ=0)	FA% (hardΘ=0)	HTER% (FA+FR/2)
Noah-Adams	2.07	-0.50	4.91	0.00	2.45
Peter-Jennings	4.47	0.02	4.47	4.55	4.51
Thalia-Assures	4.55	-0.36	7.58	1.82	4.70
Mark-Mullen	5.84	-0.15	9.42	2.73	6.10
CSP-WAJ-Susan	13.85	-0.91	48.21	1.82	25.01
Linda-Wertheimer	4.61	-0.59	6.91	0.00	3.45
Brian-Lamb	5.00	-0.61	8.33	1.82	5.07
Lou-Waters	9.56	-0.52	13.97	1.82	7.89
Kathleen-Kennedy	8.92	-0.05	10.83	8.18	9.50
Andrea-Arsenault	5.74	0.30	2.46	13.64	8.05
Chris-Wallace	1.92	0.37	0.00	3.64	1.82
Katherine-Calloway	5.56	0.68	0.00	11.82	5.91
TOTAL(12)	6.01	-	9.76	4.32	7.04

Table 4: Speaker recognition performances with a threshold a *posteriori* or an a *priori* (hard) threshold set to the theoretical value $\Theta = 0$, when the aI5 decoding is used.

7 Further work

Numerous applications can be derivated from the TO-MLP some example are given here:

- Speaker detection and speech recognition using in parallel the TO-MLP of all the speakers and selecting the more likely. This could make a continuous speech/speaker decoding for broadcast news for example.
- Text prompted speaker verification with control of the speaker answer.
- Speaker monitoring, where the speaker is controlled during the interaction of the system.
- etc...

Speaker Name	EER%	Θ-EER	FR% (hardΘ=0)	FA% (hardΘ=0)	HTER% (FA+FR/2)
Noah-Adams	3.10	-0.36	8.27	0.91	4.59
Peter-Jennings	3.88	-0.01	3.88	3.64	3.76
Thalia-Assures	4.92	-0.31	12.88	1.82	7.35
Mark-Mullen	7.79	-0.16	12.66	3.64	8.15
CSP-WAJ-Susan	12.82	-0.87	55.38	1.82	28.6
Linda-Wertheimer	3.23	-0.45	7.83	0.91	4.37
Brian-Lamb	6.25	-0.48	16.25	0.00	8.12
Lou-Waters	11.03	-0.51	24.26	2.73	13.49
Kathleen-Kennedy	8.92	-0.07	10.19	8.18	9.18
Andrea-Arsenault	9.84	0.12	7.38	10.91	9.14
Chris-Wallace	1.92	0.24	0.00	3.64	1.82
Katherine-Calloway	5.56	0.34	0.00	11.82	5.91
TOTAL(12)	6.60	-	13.25	4.17	8.71

Table 5: Speaker recognition performances with a threshold *a posteriori* or an *a priori* (hard) threshold set to the theoretical value $\Theta = 0$, when the **mean** of the 2 TO-MLP phoneme outputs are used as decoding.

References

[Bahl and Mercer, 1983] L.R. Bahl et R.L. Mercer. A maximum likelihood approach to coninuous speech recognition. *IEEE trans. on Pattern Analysis and Machine Intelligence*, 5(2):179–190, 1983.

[Bimbot and Genoud, 1997] Frédéric Bimbot et Dominique Genoud. Likelihood ratio adjustment for the compensation of model mismatch in speaker verification. In *EUROSPEECH97*, Rhodos Greece, 1997. IEEE.

[Bourlard and Wellekens, 1990] H. Bourlard et C.J. Wellekens. Links between markov models and multilayer perceptrons. *IEEE trans. on Pattern Analysis and Machine Intelligence*, 12:1167–1178, 1990.

[Bridle, 1989] J. S. Bridle. Probabilistic scoring for back-propagation networks with relationships to statistical pattern recognition. In *Conf. Neural Networks for Computing*, SnowBird UT 1989.

[Cook and Robinson, 1998] G.D. Cook et A.J. Robinson. The 1997 abbott system for the transcription of broadcast news, cambridge university. In DARPA [1998].

[Cook *et al.*, 1999] Gary Cook, James Christie, Dan Ellis, Eric Fosler-Lussier, Yoshi Gotoh, Brian Kingsbury, Nelson Morgan, Steve Renals, Tony Robinson, et Gethin Williams. An overview of the sprach system for the transcription of broadcast news. In *DARPA Broadcast News Transcription and Understanding Workshop, Herndon VA*, 1999.

[DARPA, 1998] DARPA, editor. *DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne Conference Resort, Lansdowne, VA, February 1998. NIST, http://www.itl.nist.gov/div894/894.01/proc/darpa98/index.htm.

[Ellis and Morgan, 1999] Dan Ellis et Nelson Morgan. Size matters: An empirical study of neural network training for large vocabulary continuous speech recognition. In *ICASSP99*, Phoenix AZ, 1999. IEEE.

[Fisher *et al.*, 1998] W. M. Fisher, W. S. Liggett, A. Le, J. G. Fiscus, et D. S. Pallett. Data selection for broadcast news csr evaluations. In DARPA [1998].

[Genoud and Caloz, 1997] D. Genoud et G. Caloz. 1997 nist evaluation: Text independent speaker detection (verification). Technical Report IDIAP-Com97-03, IDIAP, 1997.

[Green and Swets, 1988] David M. Green et John A. Swets. *Signal detection theory and pschophysics*. John Wiley & sons, 1988. reprint of the 1966 edition.

[Jelinek, 1976] F. Jelinek. Continuous recognition by statistical methods. *Proc IEEE*, 64(4):532–555, 1976.

[J.Neto *et al.*, 1996] J.Neto, C.Martins, et L.Almeida. An incremental speaker-adaptation technique for hybrid hmm-mlp recognizer. In *ICSLP96*, Philadelphia USA, October 1996. IEEE.

[Kleinrock, 1975] Leonard Kleinrock. *Queueing systems*, volume Volume I: Theory. John Wiley and Sons, 1975.

[Mariéthoz *et al.*, 1999] J. Mariéthoz, D. Genoud, F. Bimbot, et C. Mokbel. Client/world model synchronous alignement for speaker verification. In *to appear in EUROSPEECH99*, Budapest Hungary, 1999. IEEE.

[Martin, 1997] A. Martin. Nist 1997 speaker recognition evaluation plan. Technical report, NIST, 1997. web http://www.itl.nist.gov/div894/894.01/sp_v1p1.htm.

[Pierrot *et al.*, 1998] J.B. Pierrot, J. Lindberg, J. Koolwaaij, H.P. Hutter, M. Blomberg, D. Genoud, et F. Bimbot. A comparison of a priori threshold setting procedures for speaker verification in the cave project. In *ICASSP98*, Seatle USA, 1998. IEEE.

[Rabiner and Juang, 1993] Lauwrence Rabiner et Bing-Hwang Juang. *Fundamentals of speech recognition*. signal processing. Prentice Hall, Englewood Cliffs, NJ, 1993.

[Renals and Morgan, 1992] Steve Renals et Nelson Morgan. Connectionnist probability estimation in hmm speech recognition. Technical Report TR-92-081, International Computer Science Institute, 1992.

[Scharf, 1991] L.L. Scharf. *Statistical Signal Processing. Detection, Estimation and Time Series Analysis*. Addison-Wesley Publishing Company, 1991.