TR-I-292

A Study on Dynamic Speaker Adaptation using HMM-Nets

Edward Willems, Tetsuo Kosaka, Jun-Ichi Takami,
Shigeki Sagayama

January 1993

ABSTRACT

This report describes a new approach to dynamic speaker adaptation, which relies on switching between different methods of adaptation in order to gain maximum performance depending on the amount of speech data obtained through the speech recognition session.

The speech recognition performance of speaker adaptive systems is determined by the specific method used for the adaptation as well as by the amount of available training speech data. Furthermore, the effectiveness of the adaptation often depends on the speakers.

We present a two dynamic features to include in the design of speaker-adaptive recognisers using Hidden Markov Networks : dynamic method selection and dynamic method adaptation. We also present an implementation of the former feature : a system which can switch between three different speaker adaptation techniques, namely, vector field smoothing, speaker-tied weight training and speaker-free weight training. The methods are selected according to the most likely candidate they produce, based on the input speech data. Experimental results show that this dynamic system achieved better results compared to conventional recognisers which use a single adaptation procedure.

ATR Interpreting Telephony Research Labs.
ATR 自動翻訳電話研究所

INDEX

List of Figures

1 Introduction

The activities of the Interpreting Telephony Research laboratories at ATR are mainly concerned with the development of a telephone system to automatically translate speech between English, German and Japanese (ASURA project). The task can be split into the following three main areas of research : speech recognition, linguistic analysis and translation, and speech synthesis. This report looks into the field of speech recognition[6][7], more specifically into that of speaker adaptation[1] of Hidden-Markov Networks[2][3].

1.1 Automatic speech recognition

In order to perform the task of automatic speech recognition using a computer, it is necessary to first build a model which relates the elementary linguistic units of the speech signal, known as the **phonemes**, to their acoustic manifestation, i.e. the way they are spoken, known as the **phones**. There is a wide variation in the acoustic properties of phones, due to factors such as co-articulation, which is the modification in pronunciation due to the influence of neighbouring sounds on the position of the tongue or other articulators, and prosodic features, such as stress, tone and timing.

The acoustic properties of phones also vary from speaker to speaker, due mainly to the differences in the mechanical properties of the vocal tracts.

It is hence difficult to build a general model for speech, which will be equally accurate for any speaker. However, this is necessary if one wishes to use computers to recognise human speech. One approach is to combine the speech characteristics of many different speakers into a single model. This is then called a speaker-independent model. It relies on the fact that it will yield a more *neutral* representation of the speech signal to perform speaker-independent speech recognition. However, due to the fact that it is neutral, it is usually unable to model the user's specific phones very accurately, especially if they differ strongly from those of the speakers originally used to define the model.

An alternative approach is to ask the user to input a known sequence of phones, before he starts using the recogniser. These phones are then used to slightly alter the parameters of the recognition model, to adjust it to this specific user's phones. This process is known as **speaker-adaptation**[1][9]. Different methods exist to extrapolate new model parameters from the input phones, known as the **training data**. The amount of training data required and the increase in recognition gained vis-a-vis the unadapted model depend on the specific method used.

1.2 A new approach

When implementing a speaker-adaptive recognition system, is it desirable to pick the most efficient method of adaptation. The conventional approach is to select one method and impose a sufficient amount of training data for this method to yield its best results. However, our experiments have shown that the most effective method to use on a given set of training data depends on the speaker. This has led us to present a new approach to the design of speaker-adaptive recognition systems, which includes different methods of adaptation, which are dynamically selected depending on the speaker.

The model we used was a Hidden Markov Network, which is a doubly stochastic model

structurally similar to a Hidden Markov Model. It is however a more efficient model in terms of output distributions. Hidden Markov Networks are described in [3] and briefly covered in the following section on speaker-adaptive recognition systems. We considered three adaptation methods : speaker-tied mixture weight training[9], speaker-free mixture weight training[10] and vector field smoothing[11]. These adaptation techniques are described in sections 3 and 4.

Section 5 describes the advantages of a dynamic speaker-adaptive recognition system[5] and outlines two different aspects to consider : dynamic method selection and dynamic method adaptation.

Section 6 presents a system for dynamically switching between the three adaptation techniques, based on the maximum likelihood criterion. The recognition results obtained with this system show that it yielded better results than systems which rely on a single adaptation technique.

2 Model used for speaker-adaptive recognition

This section assumes the reader is familiar with Hidden Markov Models (HMMs), and their application to speech recognition. A good analysis of HMMs can be found in [4].

The model used was a 12-mixture Hidden Markov Network, which will now be described.

2.1 Hidden Markov Network

The general structure of a Hidden Markov Network (HM-Net) is the same as that of ordinary HMMs. Its main features are the following :

- it is a doubly stochastic network, the nodes of which are called *states* and the arcs are called *transitions*
- each state corresponds to a specific probability distribution function in the acoustic space. Acoustic space is defined by the parameters used to represent the speech signal
- the network contains a single **start** state and a single **finish** state
- transitions between the states are defined using probabilities, which are called the **transition probabilities**
- allophones are represented by paths through the network from the start state to the finish state
- a path may be shared by a cluster of similar allophones

The last two points cited above are important. In conventional HMMs, allophones are modelled separately using a specific sequence of states. The general structure of an HMM model for allophones is shown in figure 1.

A reliable model can be created in this way for each allophone. However, this does not take into consideration the fact that some of the states in the model are in fact very close in acoustic space, and could be grouped into a single state, which would then be on the path of both allophones. If this can be done, the total number of output distributions in the model is reduced, although the accuracy is maintained. The model is then more efficient, and

the computational load reduced. This is the underlying principle of HM-Nets, whose general structure is represented in figure 2.

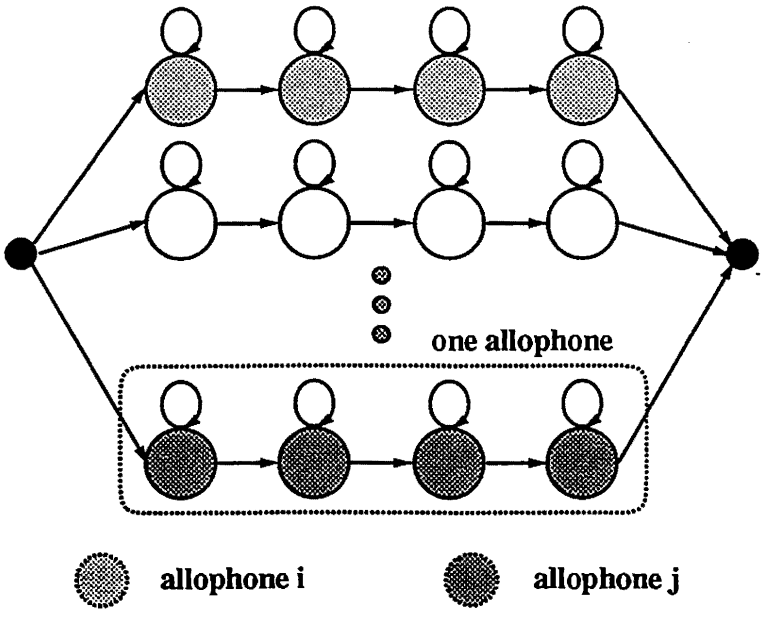


Figure 1. General structure of a Hidden Markov Model for allophone modelling

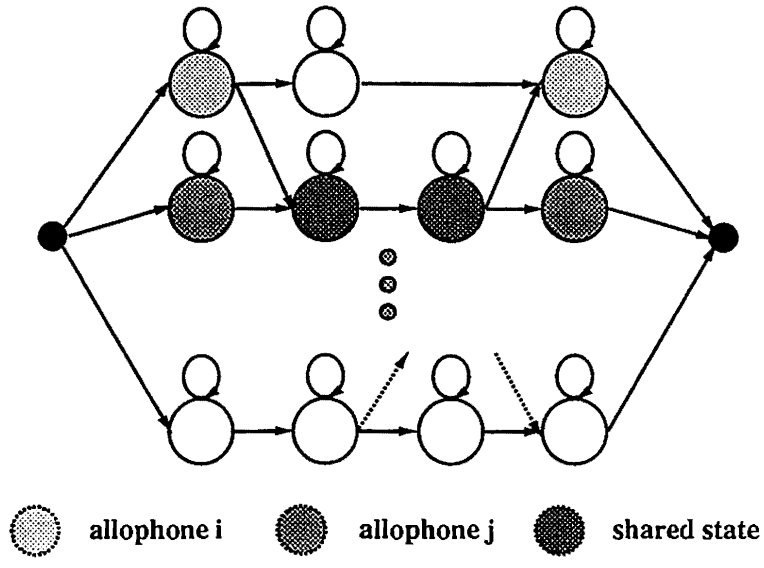


Figure 2. General structure of an HM-Net for allophone modelling

In this case allophones are represented by paths through the network, and states can have more than one predecessor and successor. However, allophones are still described as sequences of states and transitions, governed by probabilities. This means that HM-Nets can be treated in exactly the same manner as conventional HMMs[2], i.e. the same algorithms can be used for computing results. We now explain how HM-Nets are derived.

2.2 Creating a HM-Net

The main problem : how do we establish which states can be grouped into one ?

[3] presents an algorithm (Successive State Splitting) for automatic computation of the most efficient HM-Net, for a given total number of states. It is briefly summarised below :

Step 1 : an initial single state HMM is trained using all of the training data. Step 2 to 4 are then repeated until the model obtains a prescribed number of states.

Step 2 : the spread of the resulting output distribution is calculated using eq (1) in [3].

Step 3 : the state with the largest distribution spread is then split into two single-Gaussian distributions. This can be done in the time-domain or in the contextual domain, as explained in [3].

Step 4 : the newly split state distributions are re-trained to form 2-mixture Gaussian distibutions.

Step 5 : the HM-Net is retrained to change each output probability density distribution to any chosen one, in our case single Gaussian distributions.

These steps are represented in figure 3. [3] showed that the resulting HM-Net yielded better performance than a HMM model consisting of the same number of states, and is thus a more efficient model. This justifies using an HM-Net to model speech in our recognition system.

The SSS algorithm enables a HM-Net to be derived from training data, which is obtained from a single speaker using a large vocabulary. This stage define the internal structure of the model, i.e. the states and the transitions which best represent the different allophones for the given number of states. However, the model parameters are specific to the speaker used for the training, so the resulting model does not cover inter-speaker phonetic variations. We now introduce the concept of a speaker-mixture HM-Net[9], in which the phonetic characteristics of several different speakers are used to determine the output distributions of the network.

2.3 Speaker-mixture HM-Nets

The initial HM-Net is re-trained to the phones of n different speakers using a medium-sized vocabulary. A suitable adaptation technique is used in order to reshape the original output distributions to those corresponding to each speaker. However, the original structure of the HM-Net is not modified. The overall procedure for generating speaker-mixture HM-Nets is shown in figure 4.

We then obtain n HM-Nets, each having identical structures, but different output distributions. These n HM-Nets are called the n mixture components. The output probabilities for each states, $b_{ij}(y)$, are calculated by combining the output probabilities of the mixture components, $b_{ijm}(y)$, according to the following formula :

$$b_{ij}(y) = \sum_{m=1}^{N} \lambda_{ijm} b_{ijm}(y)$$

with $\sum_{m=1}^{N} = 1$ and $\int b_{ijm} dy = 1$

The λ_{ijm} are the weighted contributions from each mixture component. In the unadapted case each speaker-mixture component bears the same weight, so we have for all m:

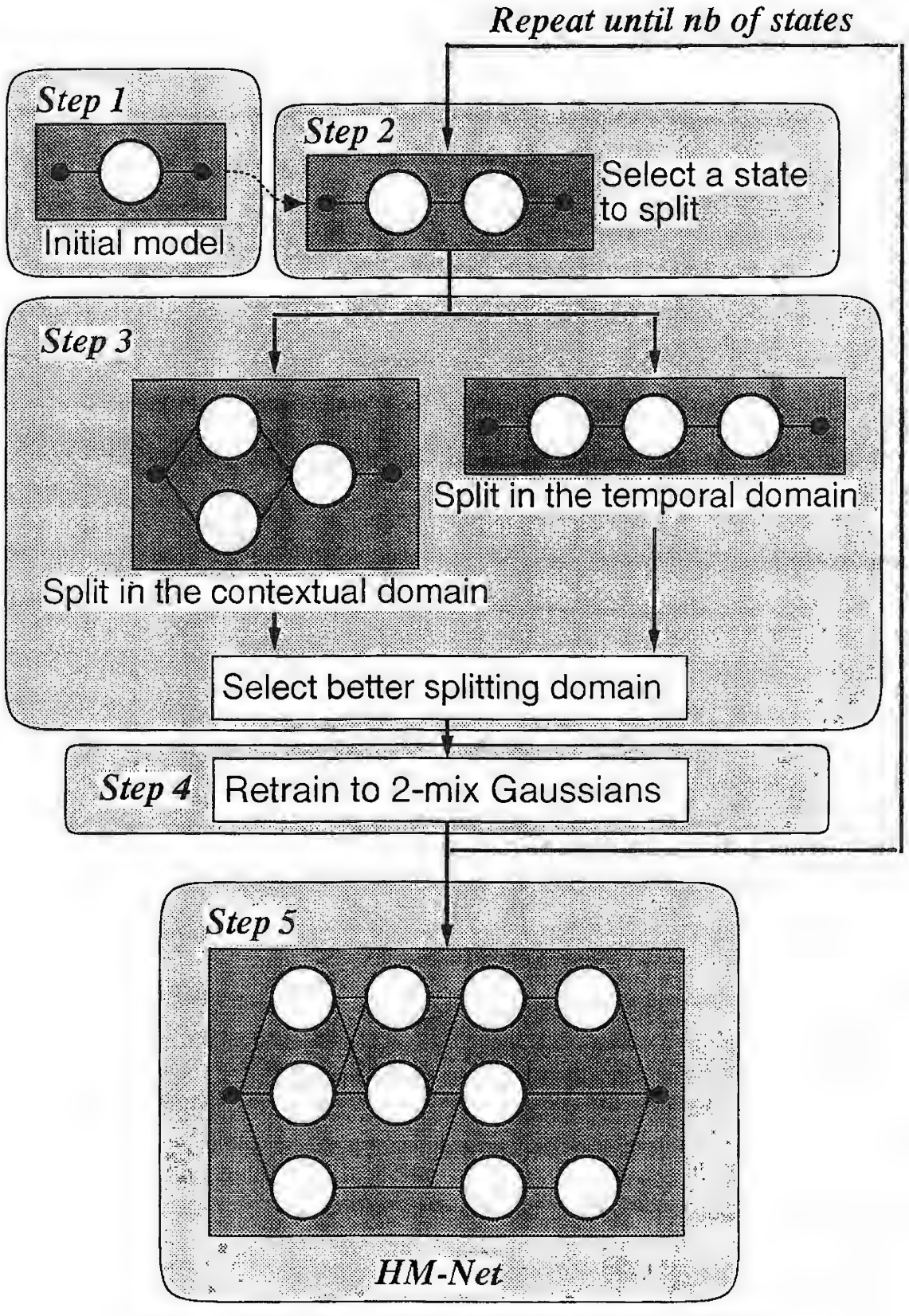


Figure 3. The Successive State Splitting algorithm (SSS)

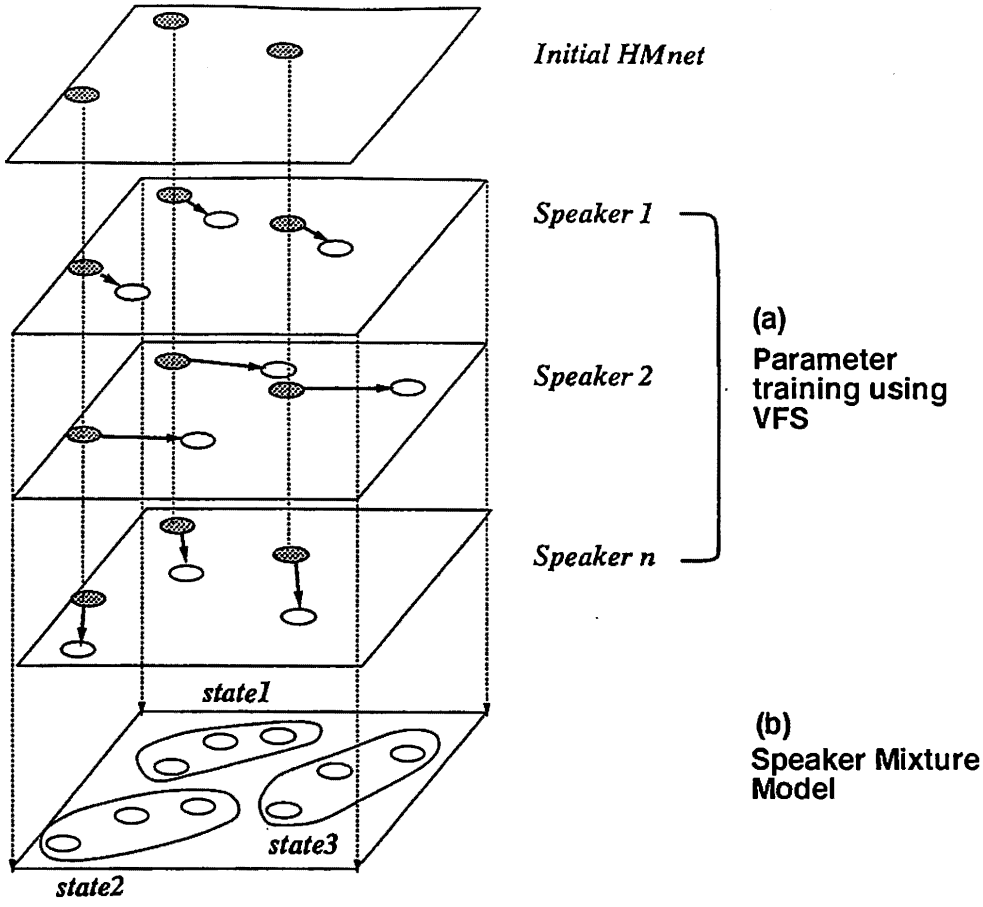


Figure 4. Procedure for generating speaker-mixture HM-Nets

$$\lambda_{ijm} = \frac{1}{n}$$

where n is the number of mixture components.

The entire model consisting of the n mixture components and the combination of their output probabilities is called a n-mixture HM-Net. If the phones of the mixture components provide a good coverage of allophonic variation, speaker-mixture HM-Nets form a neutral model for speech. Such a model is well suited for use in speaker-adaptive recognisers.

The following two sections look into two fundamentally different families of adaptation techniques which can be used to adapt the speaker-mixture HM-Net to the input speaker's phones.

3 Speaker-mixture weight training

3.1 General principle

Speaker-mixture weight training relies on altering the weighted contribution of the mixture components to adapt the model to the user. The mixture components which are close to the input speaker are given a stronger weight than those which differ from him. The output probability is still computed in the same way, so the model is effectively shifted in acoustic

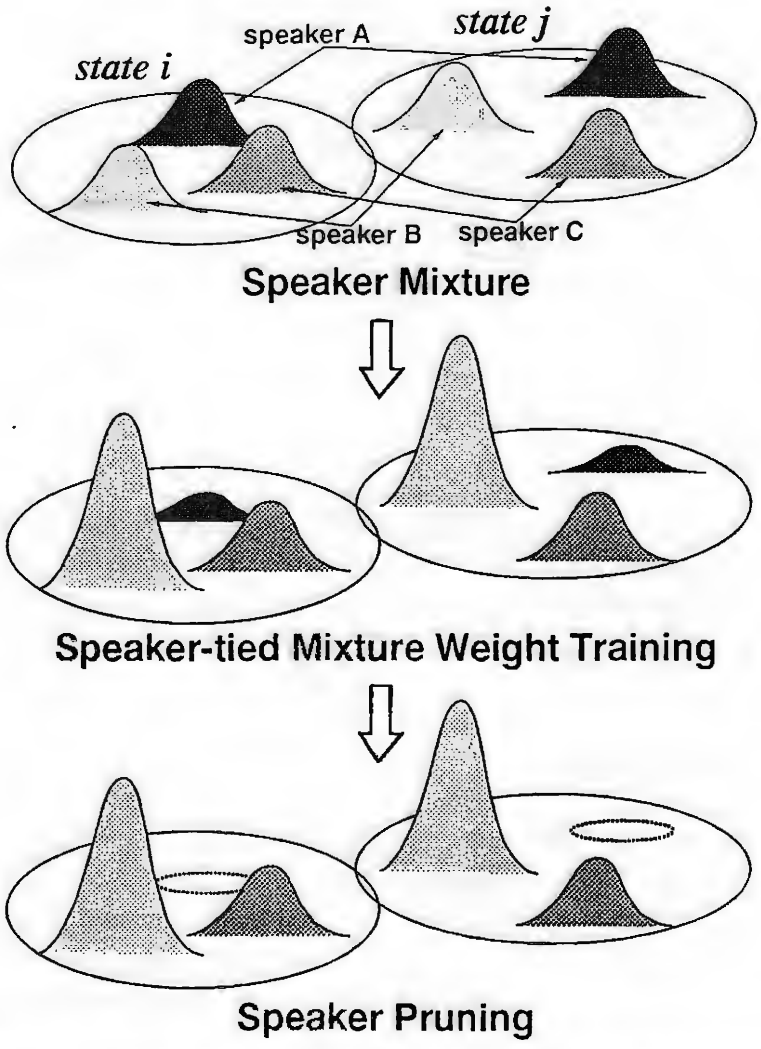


Figure 5. General principle of speaker-tied weight training

space towards the mixture components which are the closest to the input speaker. The new weights are calculated using the Baum-Welch algorithm.

We present two such adaptation techniques : speaker-tied mixture weight training and speaker-free mixture weight training. A detailed description of both of these can be found in [9] and [10]. The new weights are calculated using the Baum-Welch algorithm.

3.2 Speaker-tied mixture weight training

In this technique, the adaptation training data is used to determine the mixture components which are closest to the current input speaker. The evaluation is carried out globally for the whole model, i.e. a weight is given to each mixture components and apllied to each states. So we have the following equation for the weight λ_{ijm} for all i,j (states) :

$$\lambda_{ijm} = \lambda_m$$

where m refers to the mixture components, with $\sum_{m=1}^{n} = 1$.

The geometric structure of the model in acoustic space is left unchanged, it is merely translated towards those mixture components which have the strongest weights. Figure 5

shows how speaker-tied weight training affects the mixture components of two states i and j in the model.

(3.2.1) Speaker pruning

An interesting characteristic of this adaptation technique is that it is possible to discard the mixture components which differ too strongly from the input speaker. Since new weights are calculated for all of the mixture components, those which are two far from the input speaker's phones will be given a very small weight. Their influence on the output probability is then insignificant and it is desirable to eliminate them from the calculation, since this reduces the computional load. This process is known as **speaker pruning** and it is illustrated in figure 5.

(3.2.2) Data requirements

In this adaptation procedure, only the weights of the mixture components are altered. In a n-mixture HM-Net, there are therefore n free parameters to be derived from the training data. So the degrees of freedom for the adaptation are very limited, which means this adaptation techniques requires a small amount of training data to be effective. This is the main advantage of this technique.

(3.2.3) Increase in recognition

Precisely because there are few degrees of freedom, the performance of the adapted model is limited to how well the linear combination of the mixture components can represent the input speaker's phones. The adapted model can be seen as the nearest approximation to the speaker phones in the sub-space formed by the set of mixture components.

As a result of this, the increase in recognition will be large if the input speaker is close to the mixture components, and his phones are sufficiently closely covered by them. However, if he differs too strongly from each of the components, the adaptation will yield poor results. In any case, the limited degrees of freedom means this adaptation procedure cannot reach very high recognition rates.

3.3 Speaker-free mixture weight training

In this case, weights are calculated for the components at each state accessed by the training data. Unbiased weighting is maintained for those states which are not accessed. The adaptation is therefore only carried out for the accessed states.

This adaptation procedure is more flexible than speaker-tied mixture weight training, since the degrees of freedom are number given by :

$$\text{number of accessed states} \times \text{number of mixture components}$$

The adapted model is no longer a simple translation in acoustic space. It can be seen as a *re-sharing* the output distributions between the mixture components at each state of the mixture HM-Net. The actual 'geometric' shape is therefore altered, and the adaptation is more sophisticated and therefore more accurate.

(3.3.1) Data requirements

Since there are more free parameters to set, this method requires more training data before it can become effective. This is also due to the fact that the adaptation is only carried out for accessed states. This is a drawback.

(3.3.2) Increase in recognition

The accuracy of the adapted model is limited by the ability of the mixture components for each state to model the input speaker's phone. This adaptation technique is threrefore a superset of speaker-tied mixture weight training. It can cover more acoustic space and is therefore able to approximate the input speaker's phones more accurately. Given sufficient training data, it therefore achieves higher recognition rates than the previous method.

3.4 Note on method combination

We have seen that the main drawback of speaker-free mixture weight training is that the adaptation is only carried out for the states which are accessed by the training data. Sufficient data is needed to access the majority of the states in the HM-Net before the adaptation becomes effective. This was not the case for speaker-tied mixture weight training, in which a global weight was estimated ands applied to every state.

An interesting possibility to consider is the combination of both technique, as follows :

- derive a global weight for mixture components using the speaker-tied method, apply it to all states not accessed by the training data
- for those states which were accessed, use speaker-free mixture weight training for estimating the speaker's phones

This adaptation would then combine the finer adaptation of the second technique with the rapid adaptation of the first.

4 Adaptation using Vector Field Smoothing (VFS)

Vector field smoothing is an adaptation technique which acts in a totally different manner to speaker-mixture weight training. In VFS, new means are calculated for the distribution of every speech parameter in the network. This means the adapted HM-Net can assume nearly any structure within recognition space. However, it does not alter the shapes of the distributions.

This enables an accurate representation of any speaker's speech to be achieved, quite independently of how close the speaker's phones are to the mixture components. High recognition rates can therefore be attained. Three steps are carried out :

mean calculation : new means are calculated for the distributions in the states which are accessed by the training data.

mean interpolation : news means for the distributions of the states which are not accessed are interpolated from the calculated ones.

smoothing : interpolated values are constrained within limits to avoid chaotic model definition when little training data is used.

The two last steps characterise the VFS adaptation technique. The interpolation is based on the assumption that the adaptation process is an isomorphic one, i.e. that the relative positions of the states in recognition space are preserved. From the new positions of the calculated means, the other means can then be easily interpolated. However, if few states are accessed, this interpolation is at a risk of producing worse values than if the states were left unadapted. It is indeed possible to *decrease* the recognition rate if too little data is used. The smoothing step prevents this, by introducing a constraint on the permissible values for interpolated means. This is controlled by introducing the notion of fuzziness, which is used to determine a critical parameter : the smoothing factor.

4.1 Fuzziness - the smoothing factor

The interpolation error is the difference between interpolated mean values and the true values for the input speaker. This error is a function of the number of available calculated means. The interpolation process is less likely to yield good results if few calculated means are used. We regroup the variation in interpolation accuracy under the generic term *fuzziness*. Fuzziness is said to be large when little data is available, and many means values have to be interpolated. It is on the contrary low when very few means are interpolated. So fuzziness is an expression of the uncertainty of obtaining accurate values in the interpolation process.

The smoothing factor is an important parameter in VFS. It was introduced to correct the interpolation, by assuming that the transfer field from the unadapted model to the adapted one is uniform. We use the term *uniform* to describe a transfer field in which the following constraint exists : if two points are neighbours in acoustic space before the transformation, then their transformed values cannot differ by more than a given threshold value. This limiting value means the overall characteristics of the network are preserved, hence the transfer field is not chaotic. This means that interpolated values which would exceed this limit are modified ('smoothed') to fit the constraint. It is the degree of uniformity of the transfer field which is controlled by the smoothing factor. A smoothing factor value of infinity produces a linear transfer field, a value of zero places no constraint, and allows chaotic interpolation.

Fuzziness and smoothing factor are closely related. If the fuzziness is large, we require a large value for the smoothing factor. This corresponds to the case when little training data is available. As the amount of data increases, the interpolation becomes more reliable, so the fuzziness decreases. The value of the smoothing factor should gradually decrease as we increase the amount of training data. The current approach is to fix its value heuristically, which limits the performance of the adaptation, as will be discussed in the next section.

5 Dynamic considerations for speaker adaptation

5.1 Choosing the adaptation technique

The aim of any adaptation technique is to achieve the highest increase in recognition using as little training data as possible. Various techniques have been developed, each relying on its own specific procedure to adapt the model to the input speaker's phones. As was described

in section 3 and 4, the final increase can be estimated by analysing the underlying principles. This way we can predict that in general the final recognition rate will be lowest for speaker-tied mixture weight training and highest for VFS. We can also predict that the former method requires less data than the latter. Generally speaking, we know the *relative* behaviour of adaptation techniques. This is illustrated in figure 6.

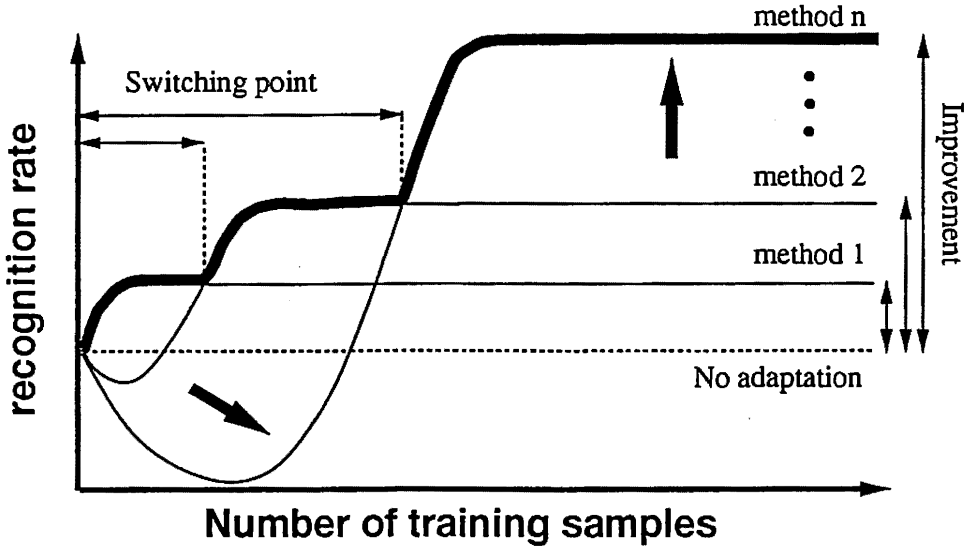


Figure 6. General recognition curves for different adaptation techniques

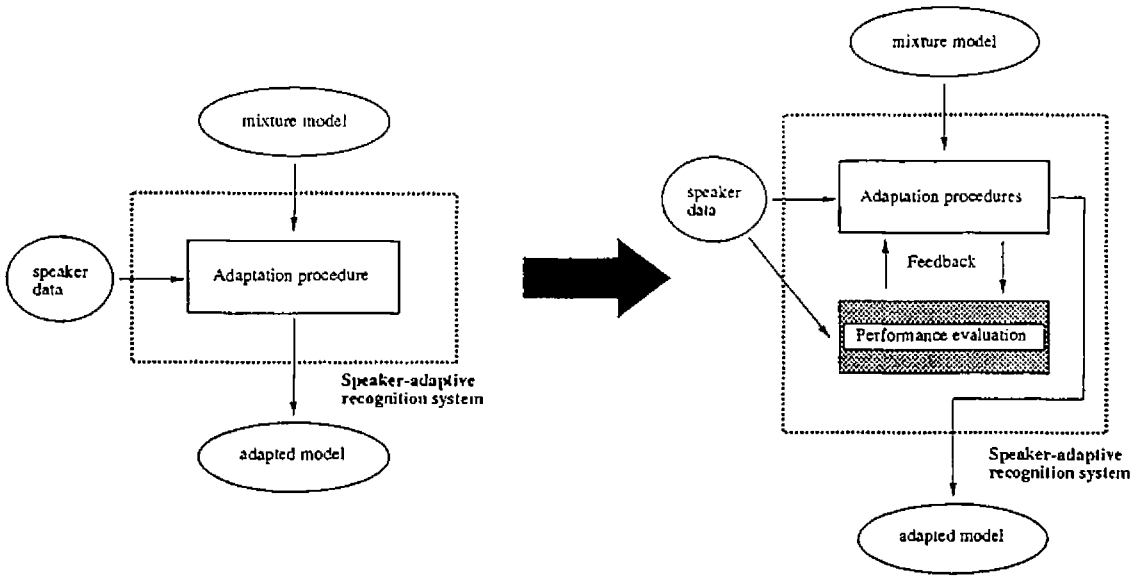
Figure 6 shows the expected recognition rates achieved by n methods for different numbers of training samples. The methods have been classified according to their ability to adjust the model to the input speaker's phones : method 1 puts a strong constraint on the final adapted model, method n allows any configuration in acoustic space to be derived from the training samples. They are therefore classified in increasing order of free parameters.

A direct consequence of this is that the expected improvement will be highest for method n and lowest for method 1. However, method n is at risk of producing a worse model if to little data is used, since more parameters need to be set. From figure 6, we can see that there is a trade-off between improvement and minimal data requirement. The best recognition performance would be achieved if the method used for the adaptation varied with the amount of training data, so that as soon as a more accurate adaptation procedure becomes more effective that the current one, this procedure is used. If we simply select the most efficient method, the recognition curve would be the one represented by the thick line in figure 6, which is clearly an improvement over choosing any one of the methods.

A far more intricate problem is that of predicting absolute values for final rates and minimum data requirements. The question we would like to answer is : given the current training data, which method would give the highest recognition rate ?

The obvious solution would be to find out heuristically where the switching points are and use it to select the desired method. This is not possible in practice for the following reasons :

- the efficiency of adaptation techniques varies from speaker to speaker, due to their specific phonetic characteristics. Therefore the switching points vary from speaker to speaker.



Current Systems **Proposed Dynamic System**

Figure 7. Block diagrams of current and dynamic speaker-adaptive systems

- the minimum data requirement varies from speaker to speaker, so fixing its value is not an optimal solution.

Both of these statements are justified by experimental results described later in this report.

A more interesting approach is to implement more than one technique, in our case three, within the same system, and choose the most efficient one based the actual input speaker utterance. In other words, we extract the information from the speech entered by the current user to determine the best method to use. We refer to this as *dynamic method selection.*

5.2 Setting the parameter values

Of the three adaptation techniques described previously, only VFS requires a parameter to be set : the smoothing factor. This is therefore the subject of the discussion. However, the concept presented here could be applied to any technique which relies on a given parameter to perform its adaptation.

Since the smoothing factor is a function of the fuzziness, which is itself determined by the training data, by heuristically fixing the value of this factor, we hinder the efficiency of the adaptation in the following manner :

- when too little training data is used, the smoothing factor is too large, thus the interpolation is not sufficiently constrained. The adaptation procedure risks producing a worse model than the unadapted one.
- when a lot of data is used, the constraint is too strong and thus limits the accuracy of adapted model

The solution to this would be to determine the smoothing factor value directly from the input speaker utterance, thus ensuring that the VFS adaptation is carried out under the best possible conditions. We refer to this generally as *dynamic method adaptation.*

5.3 Dynamic adaptation system

The previous two points have led us to suggest a new type of speaker-adaptive recognition system, enabling the implementation of both *dynamic method adaptation* and *dynamic method selection*. The block diagram of current adaptation systems is represented in figure 7. The system incorporates one adaptation technique and does not make use of the training data to determine parameter values.

The block diagram of the proposed dynamic adaptation system is also represented in figure 7. In this case, the adaptation process is more complex. The system consists of two modules : the model adaptation module and the performance evaluation module. The former uses any of the techniques to produce an adapted model to the current input speaker. The quality of this model is evaluated using the available training data. The results of this evaluation are used to control the adaptation module. There is therefore a feedback loop between the modules, and once the most suitable method/parameter value has been determined, the adapted model is output.

The key points to bear in mind while attempting to implement such a system are:

- the system should use the speaker data to evaluate the model
- the system also uses the same data for the adaptation, so we are in a closed data configuration
- our analysis assumes supervised training

In the next section we analyse more closely the implementation of dynamic method selection. The dynamic method adaptation is not discussed any further.

6 The proposed system for dynamic method selection

6.1 Hypothesis : one method for a given region

Dynamic method selection is only justified if each the adaptation techniques used is likely to be the best suited for different amounts of training data or different speakers. That is to say, there is clearly no point in including techniques in the system which always yield worse results than another one. One approach to ensure this is not the case is to use techniques which can be classified as subsets and supersets of each other, i.e. that they differ only by the constraints placed on the adaptation procedure. In our case, speaker-tied mixture weight training is clearly a subset of speaker-free mixture weight training. Although the underlying principle behind VFS is fundamentally different from that of mixture weight training, we can consider mixture weighting as a special kind of smoothing function, in which case VFS can be treated as a superset of speaker-free mixture weight training.

Based on these considerations, we can expect the speaker-tied method to be initially the most effective, then speaker-free one and finally VFS.

6.2 Region definition : switching between methods

The efficiency of any of the adaptation techniques is strongly dependant on the speaker's phonetic characteristics. This is clearly the case for mixture weight adaptation techniques,

depending on how close the speaker's phone are to the mixture components' ones. This is illustrated in figures 8 and 9, which show the recognition performance of the three adaptation techniques for two different male speakers. As can be seen, the efficiency of the methods is fundamentally different.

Because of this, it is impossible to determine in advance the regions within which a given method is most effective. It depends on the input speaker, and hence must be determined dynamically for each speaker.

Since the efficiency of the adaptation is related to the phonetic characteristics of the speaker, it seems normal to derive the best method by examining the phonetic properties of the training data. To assess the quality of an adapted model, we need measure how close it is to the original input speaker's phones. If we can do this, then we simply select the adaptation procedure which yields the closest model to the speaker's phones.

However, the following two problems have to be solved :

Problem 1 : in order to measure how close two models are, we need to define a **distance** in acoustic space

Problem 2 : we do not have a model of the input speaker's phones, so we must estimate one from the available speech data

Both of these problems and the proposed solutions are now discussed.

6.3 Suggested procedure for evaluating the performance

The quality of an adaptation procedure is determined by the accuracy of the new values which are computed for the speech parameters of the states in the network. These values are extrapolated from the values for the states which were accessed by the training data. In order to assess their accuracy, we have no other means than to test the performance of the model when recognising speech data from the speaker. To do this, we need to have *a priori* knowledge of that data, i.e. carry out supervised testing.

We are however in a closed data configuration, i.e. the only data available is the training data, which does not cover all states. The testing also has to be carried out using a different set of data, if it is to be an effective representation of performance. Although this could be simulated by splitting the training data recursively into two sets, one which is used for the training and the other for the testing, we suggest a different solution to the problem of estimating the speaker's phones, which uses the log-likelihood of first candidates produced by all three model as a measure of accuracy, and therefore as the selection criterion. Before we describe the system, we justify the use of log-likelihood to measure accuracy.

(6.3.1) Log-likelihood

During the recognition process, the particular phoneme/word/phrase which is chosen as valid is the one which yield the highest probability. This is the fundamental principle of Hidden Markov Models/Nets. Now the total probability P_{tot} is simply computed as the product of the transition probabilities $\alpha_{i,j}$ and state output probabilities β_i on the path in the HM-Net:

$$P_{tot} = \prod_{i}^{path} \alpha_{i,j}\beta_i$$

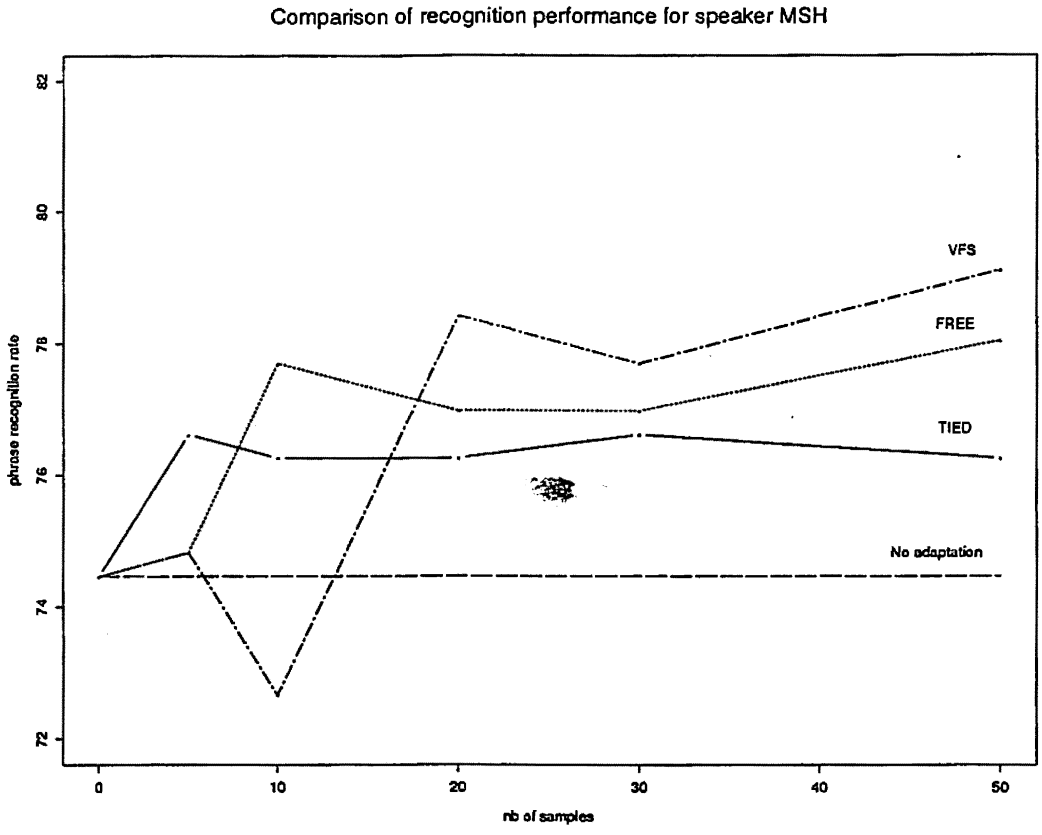


Figure 8. Recognition performance of the three methods for speaker MSH

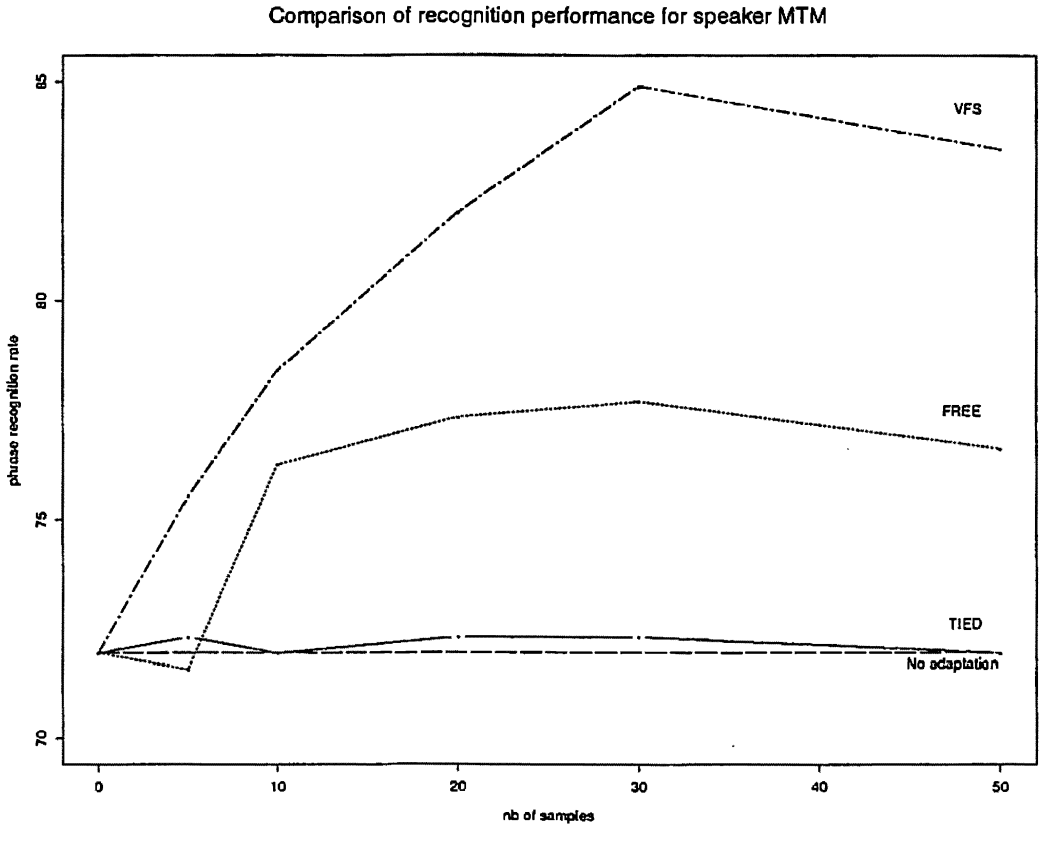


Figure 9. Recognition performance of the three methods for speaker MTM

The adaptation procedure only affects state output distribution, and ouput probability is given by the standard equation for multi-dimensional Gaussian probability density functions. However, it is common to take the logarithm of this equation for computation of probability. This value is called the log-likelihood and is given by the following equation :

$$L(y, \mu_{ij}, \Sigma_{ij}) = -\frac{1}{2}n\log(2\pi) - \frac{1}{2}\log\left|\Sigma_{ij}\right| - \frac{1}{2}(y - \mu_{ij})^t\Sigma_{ij}^{-1}(y - \mu_{ij})$$

where the symbols have their usual meaning.

The third term is proportional to the Mahalanobis distance[8], which is a measure of distance between multi-variate distribution functions. The first two terms are comparable for our three adaptation procedure, so the log-likelihood is a measure of the distance in acoustic space between the speaker's input phones and the closest candidate using a specific adapted model. Since this value is independent of the adaptation procedure, it can be used to compare the *proximity* of a model to the input phones. We assume that a 'good' adapted model will result in states closer to the phones than a 'bad' one, so that the first candidate from the model which has the highest likelihood has the highest probability of being correct, and is selected.

It should be noted that the log-likelihood is a common factor used to determine the best candidate for recognition. We are simply extending this notion to compare candidate from different adapted models, because the computed likelihood values are comparable in absolute value.

(6.3.2) Selection procedure

The selection procedure itself is dynamic : every time the speaker enters a word/phrase, all three models are used to compute first candidates. The most likely candidate is then chosen. So the system does not select one method to use for the given speaker and training samples, but continuously assesses the different models and picks one for each input phrase.

6.4 Dynamic speaker-adaptive recogniser

The recognition is carried out in two separate stages, which are represented in figures 10 and 11 :

First stage : training the training data for the current speaker is used to computed three adapted models, one for each adaptation technique.

Second stage : recognition every time the speaker enters some speech, it is analysed by each of the three models. Each model computes a 'best candidate'. The 'best candidate' with the greatest likelihood is chosen as correct.

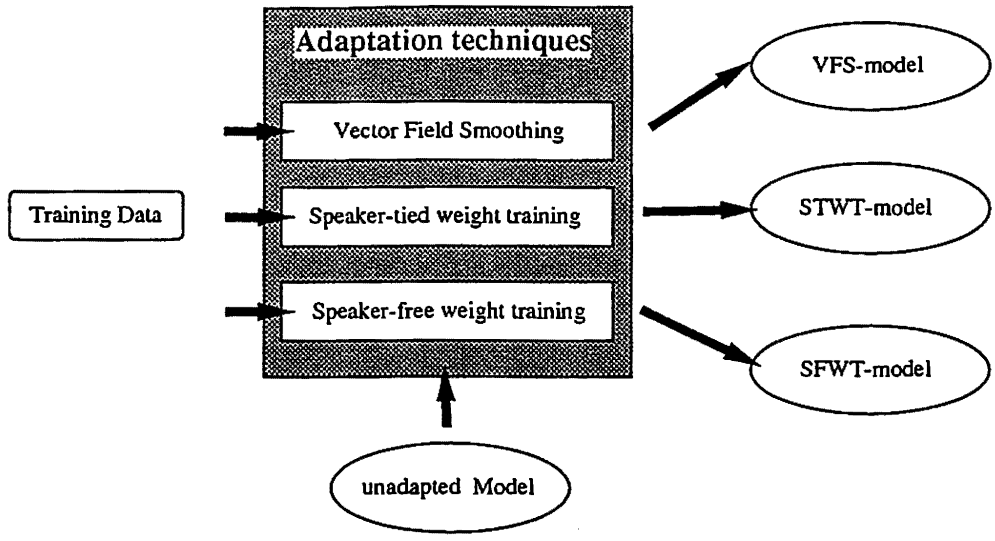


Figure 10. The first stage : obtaining the adapted models

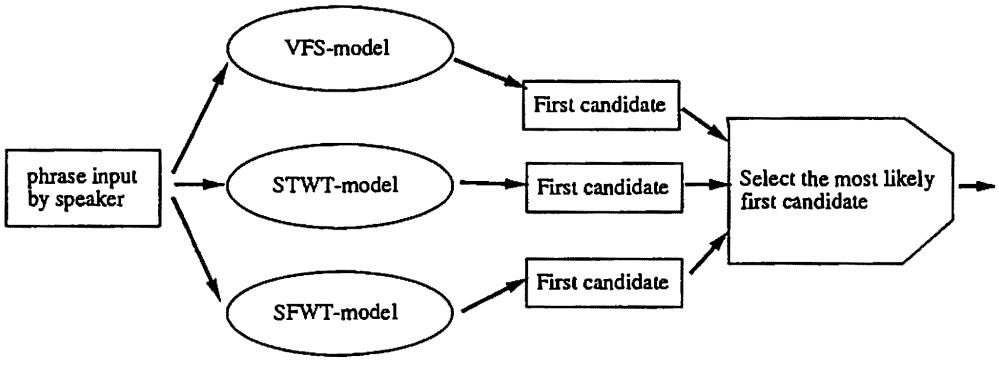


Figure 11. The second stage : speech recognition

6.5 Experimental results

We now describe the experimental results we obtained using for this dynamic recognition system.

(6.5.1) Experimental condition

The following table summarises the speech analysis conditions.

Analysis Conditions	Sampling frequency: 12 kHz Hamming window: 20 ms Pre-emphasis: 0.98 Analysis period: 5 ms 16th-LPC cepstrum + log power + 16th-Δcepstrum + Δlog power

The following experimental setup was used :

model used : mixture HM-Net, 200 states, with a maximum of four states per allophone

mixture components : 12 male professionnal speakers

training data (original SSS): 216 words × 12 males

training data (adaptation): Japanese bunsetsu, (SB1 phrase set of ATR database)

adaptation : three types: VFS, speaker-tied and speaker free

testing data : Japanese bunsetsu (SB3 phrase set of ATR database)

Parser used : SSS-LR parser, ATR Interpreting Telephony

(6.5.2) Results for speaker MSH

The recognition results for speaker MSH are shown in figure 12.

As appears clearly on figure 12, the overall performance of the dynamic system is better than any other technique on its own. In fact, for all sample numbers except 20, it achieved the optimal recognition performance for these three techniques. The final rate achieved is higher than that of VFS. This is simply explained by the fact that the some of the errors of VFS are correctly deciphered by one of the other two techniques. If the system then chooses the output from one of these two, the recognition is higher than that of VFS.

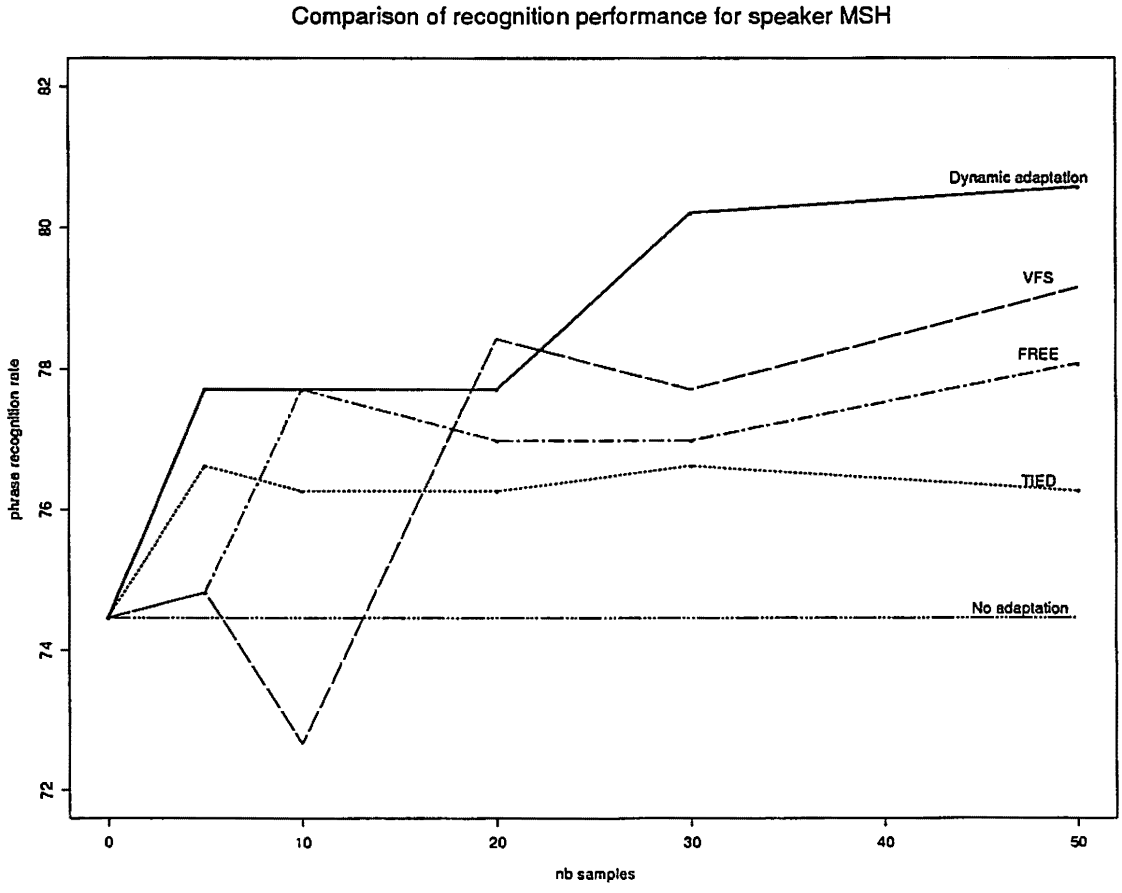


Figure 12. Phrase recognition results for speaker MSH, for different numbers of training samples

The precise selection procedure was further investigated, and is represented in the following table.

Nb of samples	5	10	20	30	50
Speaker-tied	94.6%	86.3%	75.2%	68.7%	47.1%
Speaker-free	4.7%	11.2%	22.3%	24.5%	33.8%
VFS	0.7%	2.5%	2.5%	6.8%	19.1%

As can be seen, speaker-tied is less and less selected, whereas VFS is increasingly selected. The reader should not be surprised that VFS is not the method the most selected for 50 samples, although the recognition rate is higher that VFS. This simply means that the other two methods are selected when they produce correct results.

(6.5.3) Results for speaker MTM

The recognition results for speaker MTM are shown in figure 13.

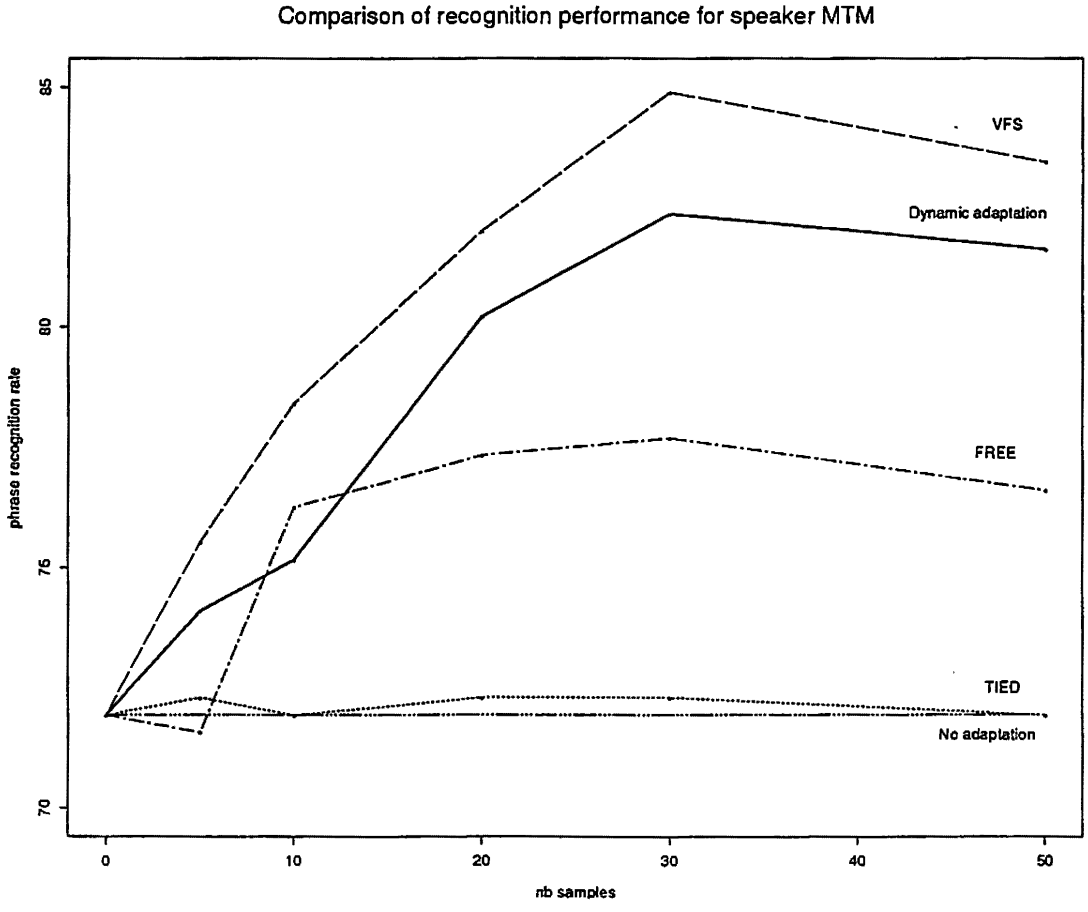


Figure 13. Phrase recognition results for speaker MTM, for different numbers of training samples

Figure 13 shows that the performance was consistenly lower than that of VFS. However, it was much better than both other methods. Although there is some loss compared to using VFS for this speaker, the overall performance of the system for MSH and MTM is better than

if only VFS had considered. The system has detected the difference in efficiency for the three adaptation techniques between speaker MTM and MSH, which is the desired goal.

This was confirmed by the selection evolution described in the table below :

Nb of samples	5	10	20	30	50
Speaker-tied	21.22%	32.37%	1.44%	0.36%	0.00%
Speaker-free	38.84%	20.50%	29.50%	24.55%	20.58%
VFS	39.94%	47.13%	69.06%	75.09%	79.42%

It is important to note that for 50 samples, the system no longer selects speaker-tied weight training, which, as can be seen on figure 13, is not effective for this speaker.

(6.5.4) Results for speaker MMY

Recognition results were obtained for a third speaker : speaker MMY and are shown in figure 14.

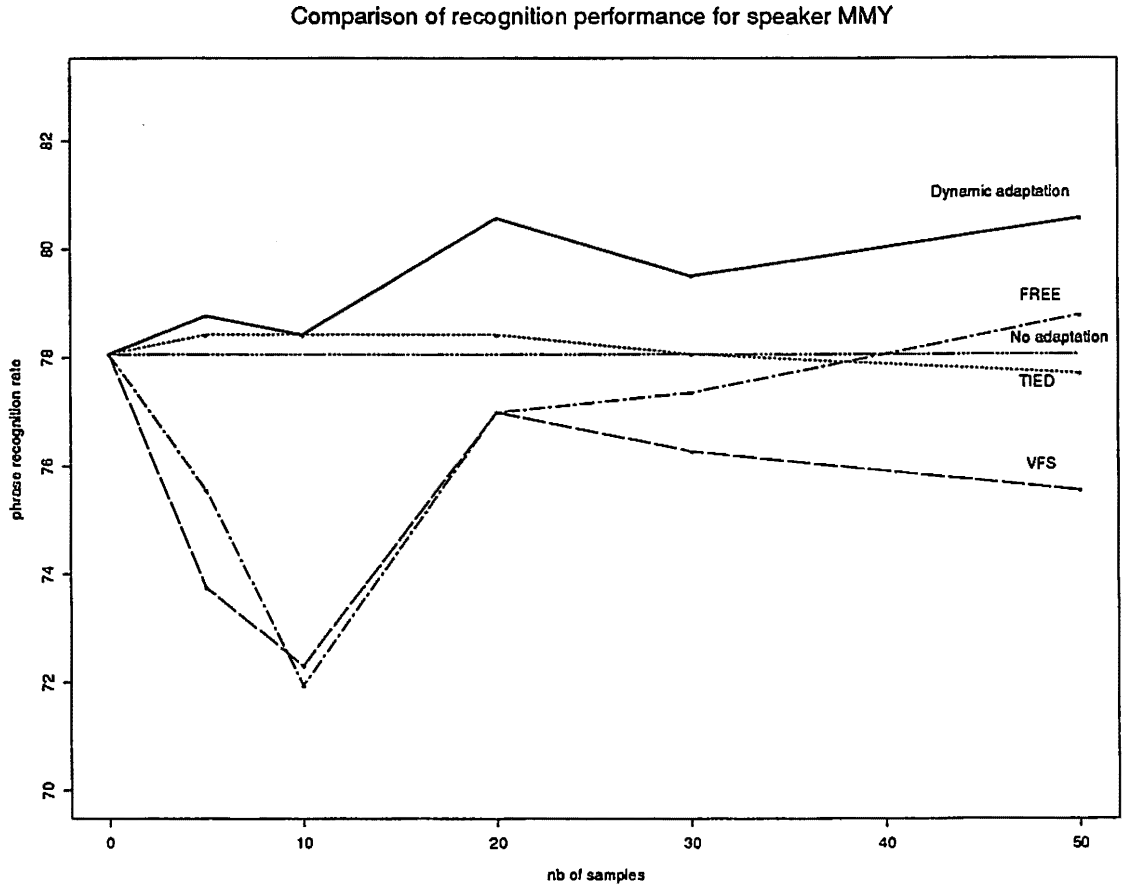


Figure 14. Phrase recognition results for speaker MMY, for different numbers of training samples

Figure 14 shows that none of the three adaptation techniques were particularly effective, due probably to the high recognition rate achieved for the unadapted model. However, once again the dynamic system was optimal, and achieved better results than implementing a single technique.

7 Conclusion and further study

This report has presented a new approach to the design of speaker-adaptive recognition systems. It introduces the concept of dynamic speaker-adaptation, where the system uses the input data from the speaker to determine which adaption to use. There are two ways of doing this : dynamic method selection and dynamic method adaptation. In the former, we include different adaption techniques in the design of the recogniser, and use the input speech to saelect the most effective method. In the latter, we modify the way a particular adaptation technique is carries out, depending on the input speech data.

We presented a system which automatically switches between three adaptation techniques. The structure of this system is parallel, so in this respect, any number of different adaptation techniques could be implemented and available for selection. This could be an initial solution to dynamic method adaptation, where we could implement many different variant of the same techniques, simply altering the value of their parameters (e.g. the smoothing factor value). The system could then select the most effictive one.

The parallel structure also enables the system to be upgraded at a later stage, and enables the system to perform at the same speed as classical ones, since we can run the different recognition processes in parallel.

The results obtained for three speakers showed the system was able to distinguish the differences in the effectiveness of the adaptation techniques for ddifferent speakers. The speaker-adaptive system we suggested was therefore better suited for speaker-independent applications. In all three cases we avoided the undesirable decrease in recognition which can occur if too little data is used for the training.

The research presented in this report is by no means finished. There are many ways we can think of putting dynamic features to use to increase the recognition performance of speaker-adaptive recognisers. The criterion for selection which was used was the log-likelihood. Future research should look into finding an optimal measure for the effectiveness of adaptation procedures.

The dynamic method adaptation has only been presented in this report. It should prove to be a very powerful adaptation technique if it can be successfully implemented. The performance evaluation part in dynamic speaker-adaptive recognisers could perhaps be implemented using neural networks.

Including dynamic features in the design of speaker-adaptive recognition system could provide a solution to inter-speaker variations in the recognition performance of such systems.

8 Acknowledgments

I wish to thank Dr. Habara, Chairman, and Dr. Kurematsu, President, ATR Interpreting Telephony Laboratories, for their continuous support of this research work.

I would also like to acknowledge Mr. Sagayama, head of department, for his help and ideas regarding the project, as well as for his extreme kindness and interest for my work.

The research would have been impossible without the invaluable help from my supervisor, Mr. Kosaka, whose constant suggestions and encouragements enabled me to enjoy the project.

I wish to thank the E.N.S.T., Paris, France for allowing me to come Japan, and Mr. Ventre and Mr. Poirier for their useful advice.

Finally, my stay was made very enjoyable by the smiling faces encountered in the corridors of the laboratory. To all of you, thank you !

REFERENCES

[1] Chin-Hui Lee et al.:"A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models," IEEE Trans. on Signal Processing, Vol.39, No.4, pp.806-814(1991.4).

[2] 鷹見, 嵯峨山：“音素コンテキストと時間に関する逐次状態分割による隠れマルコフ網の自動生成”, 音声研資, SP91-88(1991.12).

[3] Jun-Ichi Takami, Shigeki Sagayama : "A Successive State Splitting algorithm for efficient allophone modelling",Proc. of ICASSP92,pages I-573-ff.

[4] Lawrence R. Rabiner : "A tutorial on Hidden Markov Models and selected applications in speech recognition",Proc. of IEEE, Vol. 77, No. 2, Feb. 1989.

[5] Edward Willems, Tetsuo Kosaka, Jun-Ichi Takami, Shigeki Sagayama : "A dynamic approach to speaker adaptation of Hidden Markov Networks for speech recognition", Proc. of IECE, SP92-102(1992.12).

[6] X. D. Huang, Y. Ariki, M.A. Jack : "Hidden Markov Models for speech recognition",Edinburgh University Press,1990.

[7] J. N. Holmes : "Speech synthesis and recognition",Van Nostrand Reinhold(UK), 1988.

[8] R. Duda, P. Hart : "Pattern classification and scene analysis",Johm Wiley and Sons, 1973.

[9] 小坂, 鷹見, 嵯峨山:“話者混合 SSS による不特定話者音声認識と話者適応”, 音声研資, SP92-52(1992.9).

[10] 松岡, 鹿野：“混合ガウス分布不特定話者 HMM をベースとした重み係数による話者適応化法”, 音学講論, 1-1-6(1992.3).

[11] K.Ohkura et al.:"Speaker Adaptation Based on Transfer Vector Field Smoothing with Continuous Mixture Density HMMs",Proc. of ICSLP92, (1992.10).

A Appendix 1 : Transparencies

The following pages are the transparencies which were used for the final presentation of the project.

音声認識のための隠れマルコフ網の動的話者適応法

ウィレムス・エドワード　小坂 哲夫　鷹見 淳一　嵯峨山 茂樹

ATR 自動翻訳電話研究所

A Dynamic Approach to Speaker Adaptation of Hidden Markov Networks for Speech Recognition

Edward Willems　Tetsuo Kosaka　Jun-Ichi Takami
Shigeki Sagayama

ATR Interpreting Telephony Research Laboratories

Aim of the study

To present a new approach for the design of speaker-adaptive recognition systems

- *Why ?* ⇒ to gain maximum performance from the training data obtained during the speech recognition session

- *How ?* ⇒ by combining the use of different adaptation techniques

Edward Willems, ATR Interpreting Telephony Research Laboratories

The Problem

Common goal : improve the recognition by adapting the model to the speaker's voice

Different adaptation techniques exist, which can be distinguished by :

- minimal data requirements
- expected increase in recognition (efficiency)

In general : high increase in recognition ⇒ large amount of data required

Problems

- minimal data requirement varies from speaker to speaker
- efficiency varies from speaker to speaker
- minimal data requirement not respected ⇝ worse recognition

⟹*the 'best' adaptation technique to use depends on the speaker and the training data*

Edward Willems, ATR Interpreting Telephony Research Laboratories

Proposed Solution

→ Design a recognition system which includes :

- different adaptation techniques
- a control procedure to evaluate the effectiveness of the adaptation
- a feedback loop to pick a technique

→ Obtain an adapted model using each of the adaptation techniques

→ Compare the adapted models produced

→ Choose the most effective model

Principle

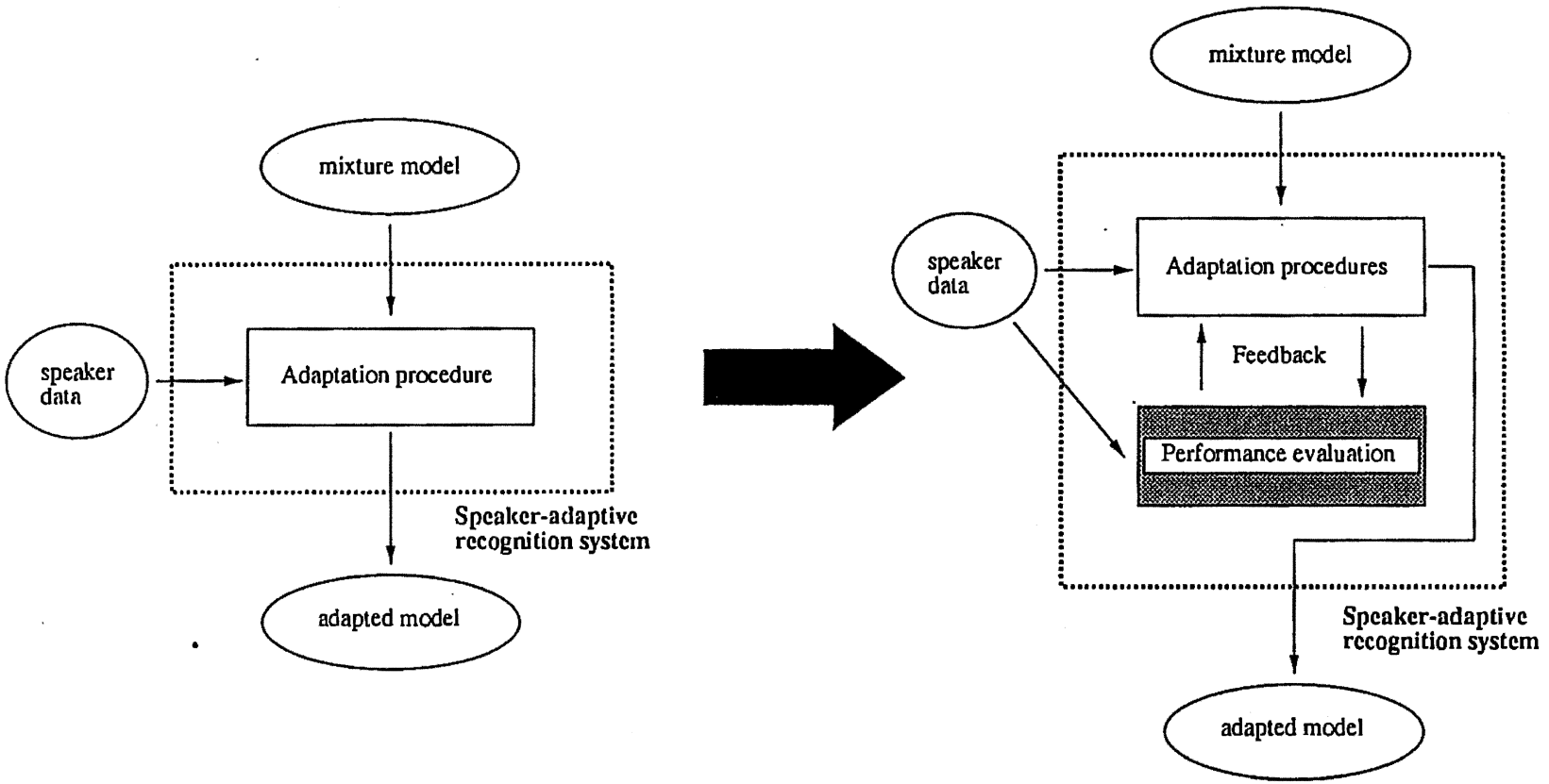
The recognition system always uses the most effective adaptation procedure

⇒ *maximum performance* is achieved

Edward Willems, ATR Interpreting Telephony Research Laboratories

ATR

Dynamic Speaker Adaptation



Current Systems

Proposed Dynamic System

Edward Willems, ATR Interpreting Telephony Research Laboratories

Hidden Markov Network

Equivalent performance to mixture Gaussian density HMMs but for a lower number of output probabilities

⇒ **efficient representation of phoneme context-independent HMMs**

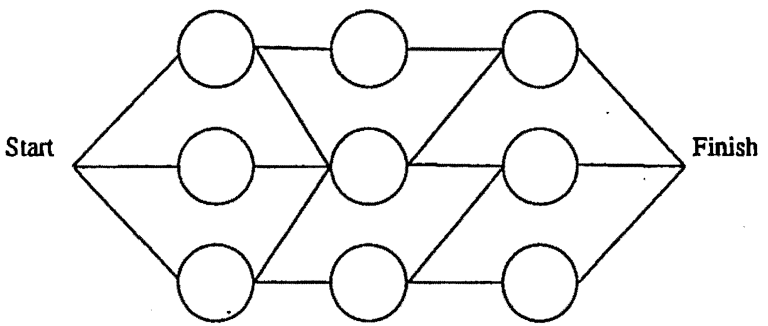


Figure 1: General structure of a HM-net

- network of states
- equivalent structure to common HMMs
- allophones are represented by paths through the states
- a path can be shared by a cluster of allophones

Edward Willems, ATR Interpreting Telephony Research Laboratories

Model used for speech

Problem : how do we obtain a neutral HM-Net model for speech ?

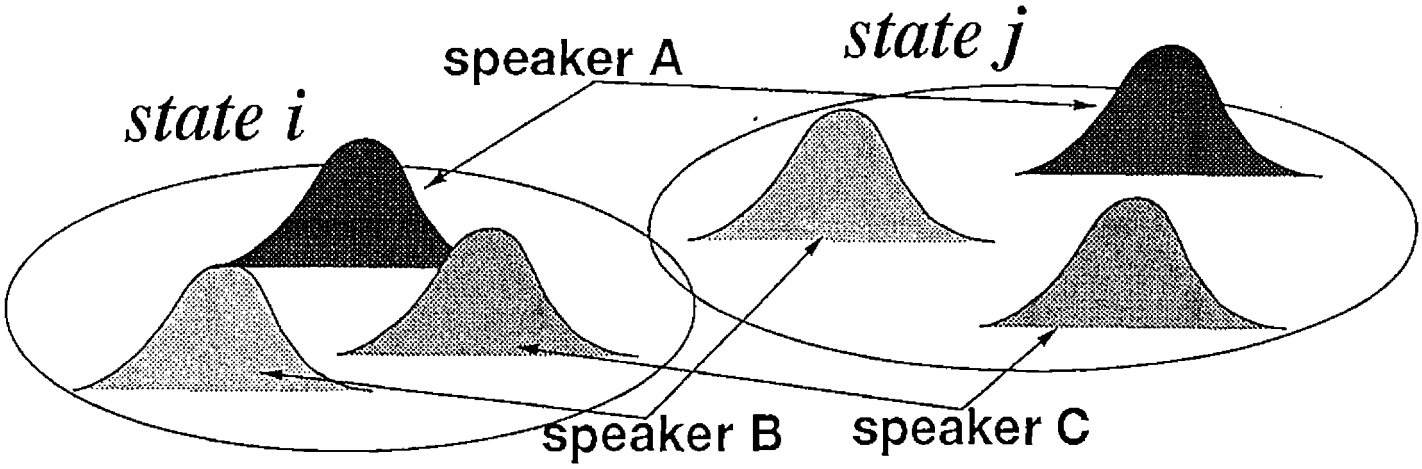
Solution : we use the combined characteristics of different speakers to produce a general model for allophones → *speaker mixture HM-Net*

Obtaining a speaker mixture HM-Net

- train an HM-net to one speaker using a large vocabulary
- adapt this HM-net to n different speakers ⇒ n *mixture components*
- use the combined properties of these mixture components to define the state distributions

Edward Willems, ATR Interpreting Telephony Research Laboratories

Speaker Mixture HM-Net



Speaker Mixture

Speaker Mixture HM-Net

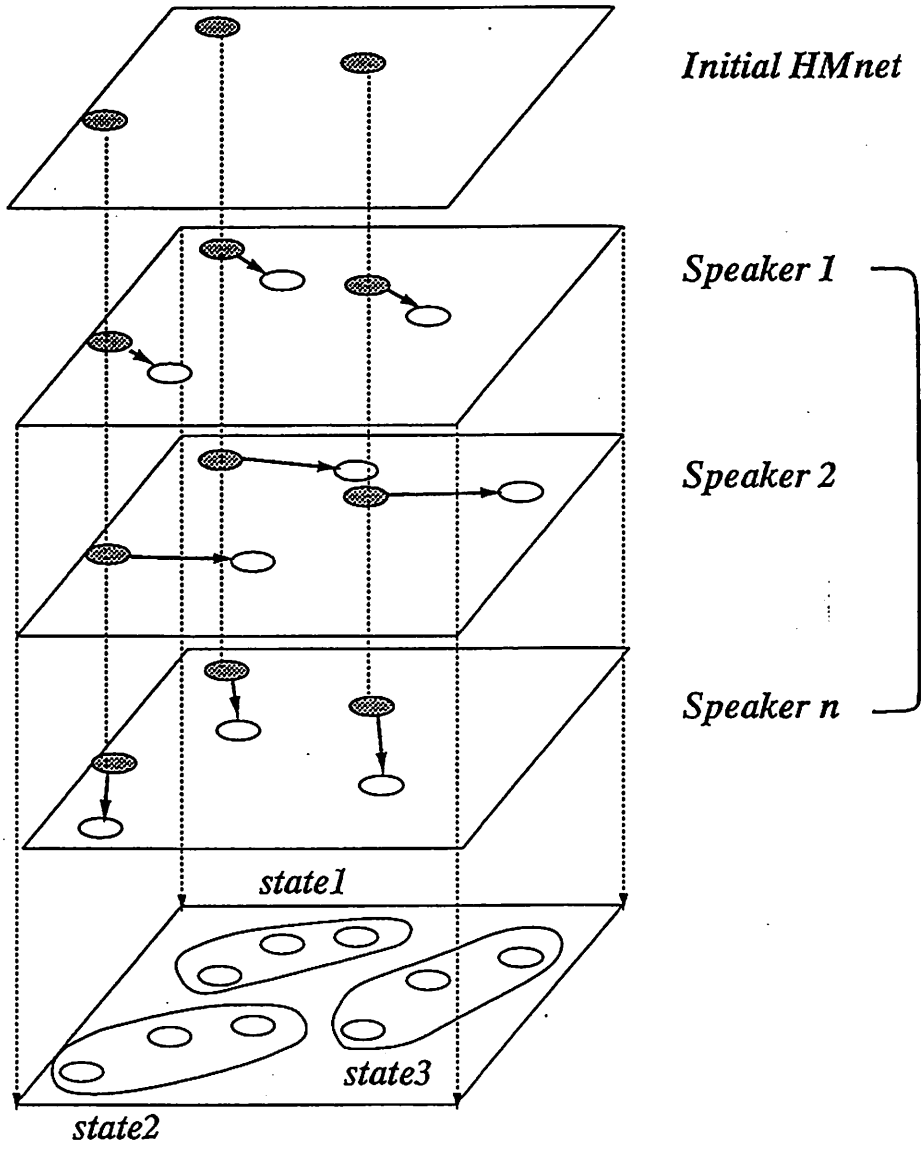
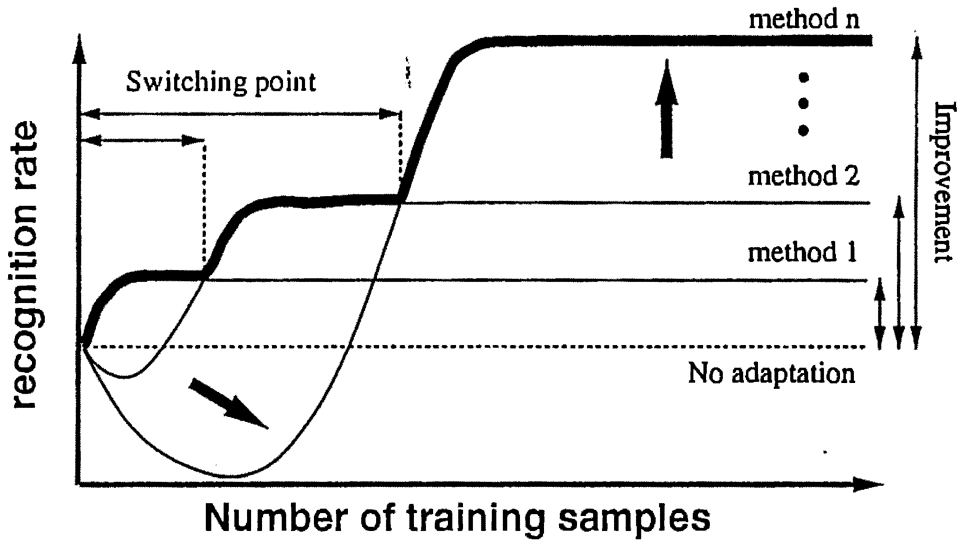


Figure 1: Speaker-mixture HM-Net

Edward Willems, ATR Interpreting Telephony Research Laboratories

Choosing the adaptation techniques

Optimal solution : techniques who differ only in the constraint placed on the adaptation procedure ⇒ the techniques are supersets or subsets of each other.



- arrows indicate the effect of decreasing the constraint on the adaptation
  - → final recognition rate increases
  - → minimal data requirement increases

⟹ *each method is optimal over a range of data*

Edward Willems, ATR Interpreting Telephony Research Laboratories

Speaker mixture weight training

mixture weight training $\Longrightarrow$ the contribution from each mixture component is weighted according to the training data values

Speaker-tied mixture weight training

- a global weight is given to each mixture component

  $\rightarrow$ nb. of free parameters = nb. of mixture components

Speaker-free mixture weight training

- weights are calculated for the mixture components at each state of the HM-Net

  $\rightarrow$ nb. of free parameters = nb. of mixture components
  $\times$ nb. of states in HM-Net

Edward Willems, ATR Interpreting Telephony Research Laboratories

Vector Field Smoothing

Three Steps

- mean calculation for the states accessed by the training data
- interpolation of the means for the other states
- smoothing : constraint on interpolated values

→ nb. of free parameters ≈ nb. of mixture components
× nb. of states in HM-Net
× nb. of speech parameters

Smoothing Factor

- a parameter to control the extent of the smoothing

→ large value ⇒ big constraint on interpolation

Problem : how do we set the value of the smoothing factor ?

Edward Willems, ATR Interpreting Telephony Research Laboratories

Comparing Adaptation Techniques

From what preceeds :

Number of free parameters

$$STWT < SFWT < VFS$$

Constraint on adapted model

$$STWT > SFWT > VFS$$

Classification

$$STWT \subset SFWT \subset VFS$$

$\Longrightarrow$ *these three methods form a suitable set for dynamic switching*

Edward Willems, ATR Interpreting Telephony Research Laboratories

Designing the evaluation procedure

Goal : to automatically evaluate the effectiveness of the different adaptation procedures and select the most effective one.

Points to consider

- system must use closed data (same data for training and evaluation)
- available inputs are the training data, the unadapted model and each of the adapted ones
- output is the most effective adapted model
- the most effective method depends both on the speaker and the amount of training data used

Edward Willems, ATR Interpreting Telephony Research Laboratories

Suggested approach

Hypothesis : the most effective method is the one which produces the model which is the 'closest' in recognition space to the speaker's phones

$\Longrightarrow$ *use this as the selection criterion*

Problems

- 'closest' $\Rightarrow$ we must define a distance in recognition space
  $\Rightarrow$ use log-likelihood
- we do not have a model of the speaker's phones $\rightarrow$ closed data condition
  $\Rightarrow$ use direct likelihood measurements on input data

$\Rightarrow$ *we select the most likely first candidate for each input sentence*

Edward Willems, ATR Interpreting Telephony Research Laboratories

Log-likelihood

Definition

$$L(y, \mu_{ij}, \Sigma_{ij}) = -\frac{1}{2} n \log(2\pi) - \frac{1}{2} \log |\Sigma_{ij}| - \frac{1}{2} (y - \mu_{ij})^t \Sigma_{ij}^{-1} (y - \mu_{ij})$$

Justification

- third term is proportional to the Mahanalobis distance

⇒ log-likelihood is a measure of similarity between distributions

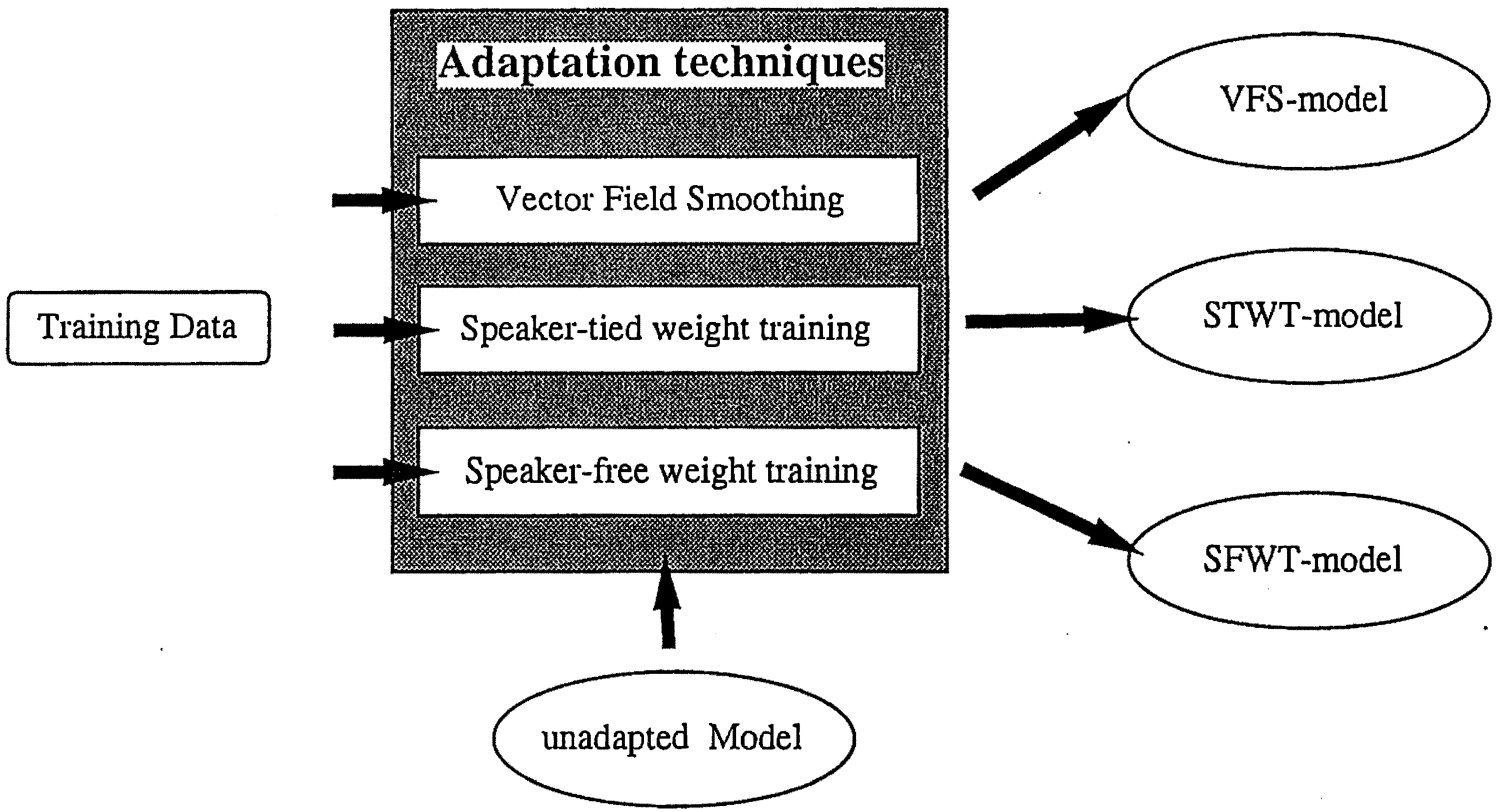
- first two terms are similar for all three adaptation techniques

⇒ log-likelihood can be used to compare performance of techniques

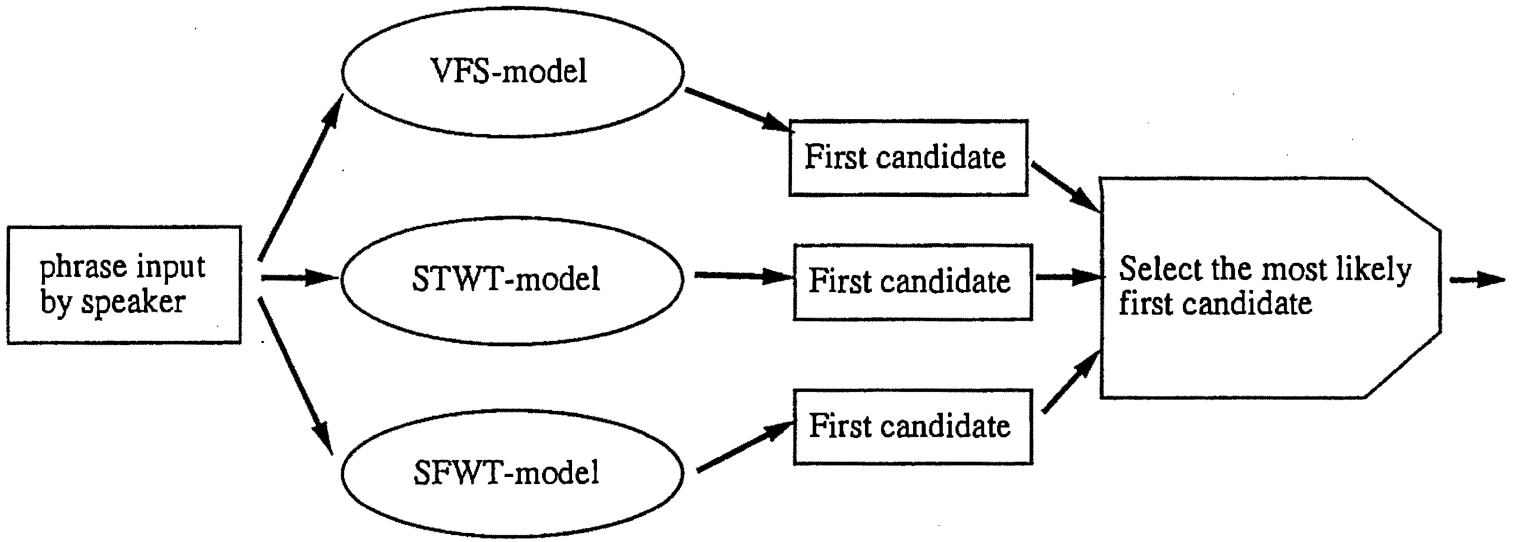
Edward Willems, ATR Interpreting Telephony Research Laboratories

ATR

Proposed System - First stage : adaptation



Proposed System - Second stage : recognition



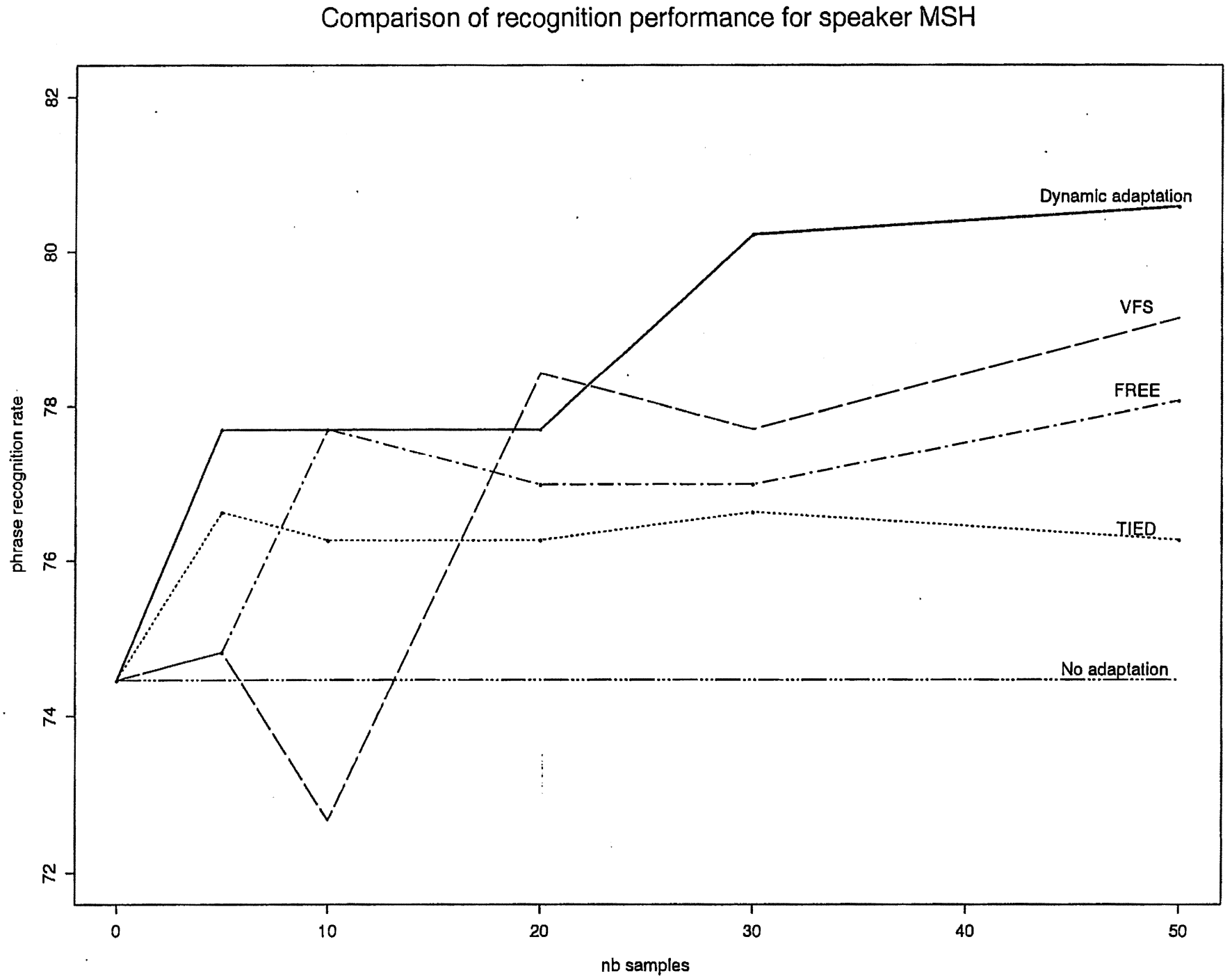
ATR

Experimental Conditions

Analysis Conditions	Sampling frequency: 12 kHz Hamming window: 20 ms Pre-emphasis: 0.98 Analysis period: 5 ms 16th-LPC cepstrum + log power + 16th-Δcepstrum + Δlog power

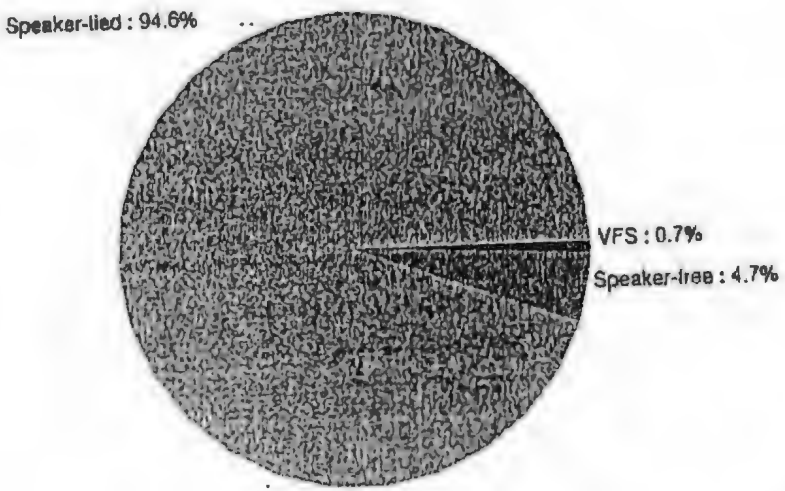
- *model used* : mixture HM-Net, consisting of 200 states, with a maximum of four states per allophone
- *mixture components* : 12 male professionnal speakers
- *training data* : original SSS $\rightarrow$ 216 words $\times$ 12 males
  adaptation $\rightarrow$ Japanese bunsetsu (SB1 of ATR database)
- *adaptation* : three types: VFS, speaker-tied and speaker free
- *testing data* : Japanese bunsetsu (SB3 of ATR database)

Edward Willems, ATR Interpreting Telephony Research Laboratories

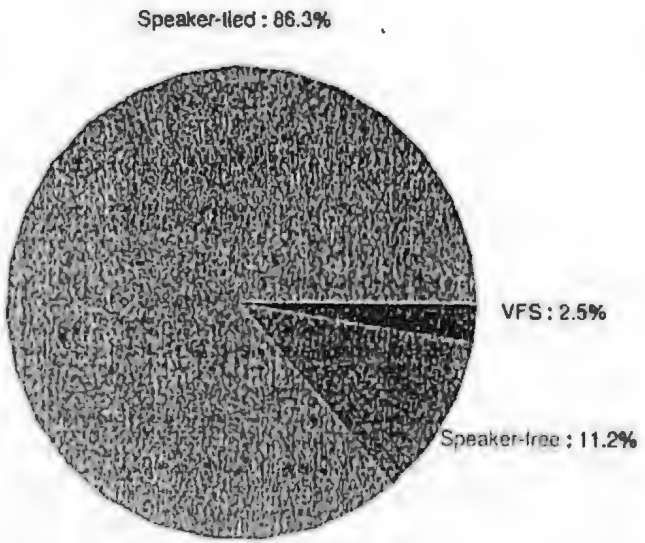

Comparison of recognition performance for speaker MSH
phrase recognition rate
82
80
78
76
74
72
0
10
20
30
40
50
nb samples
Dynamic adaptation
VFS
FREE
TIED
No adaptation

Selected methods for Speaker MSH

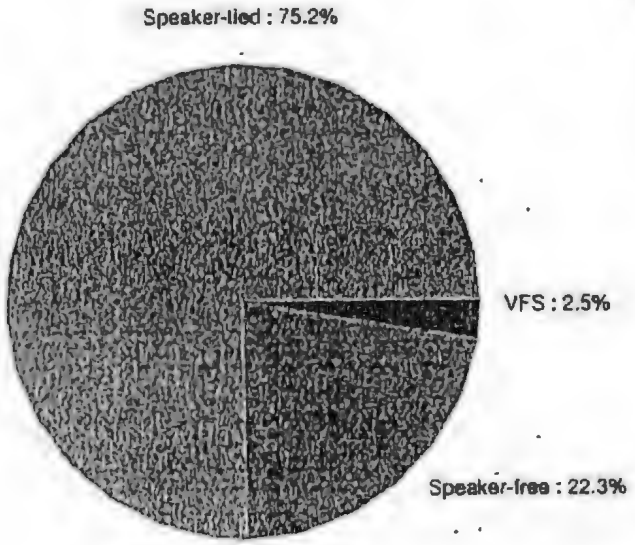
5 phrases



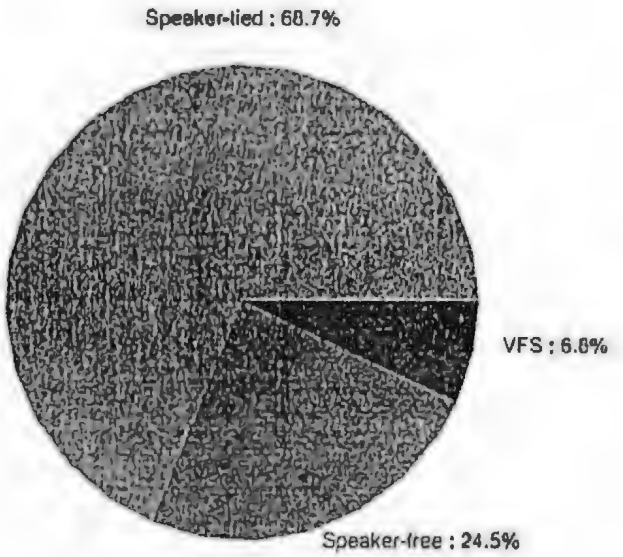
10 phrases



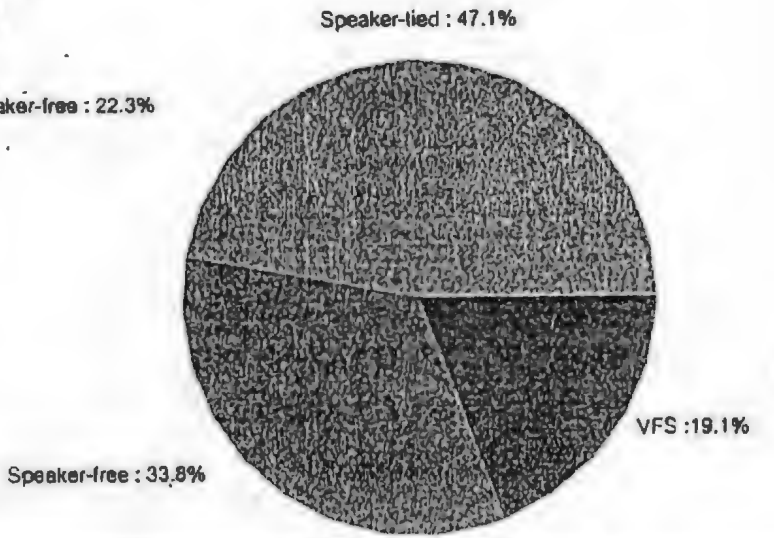
20 phrases

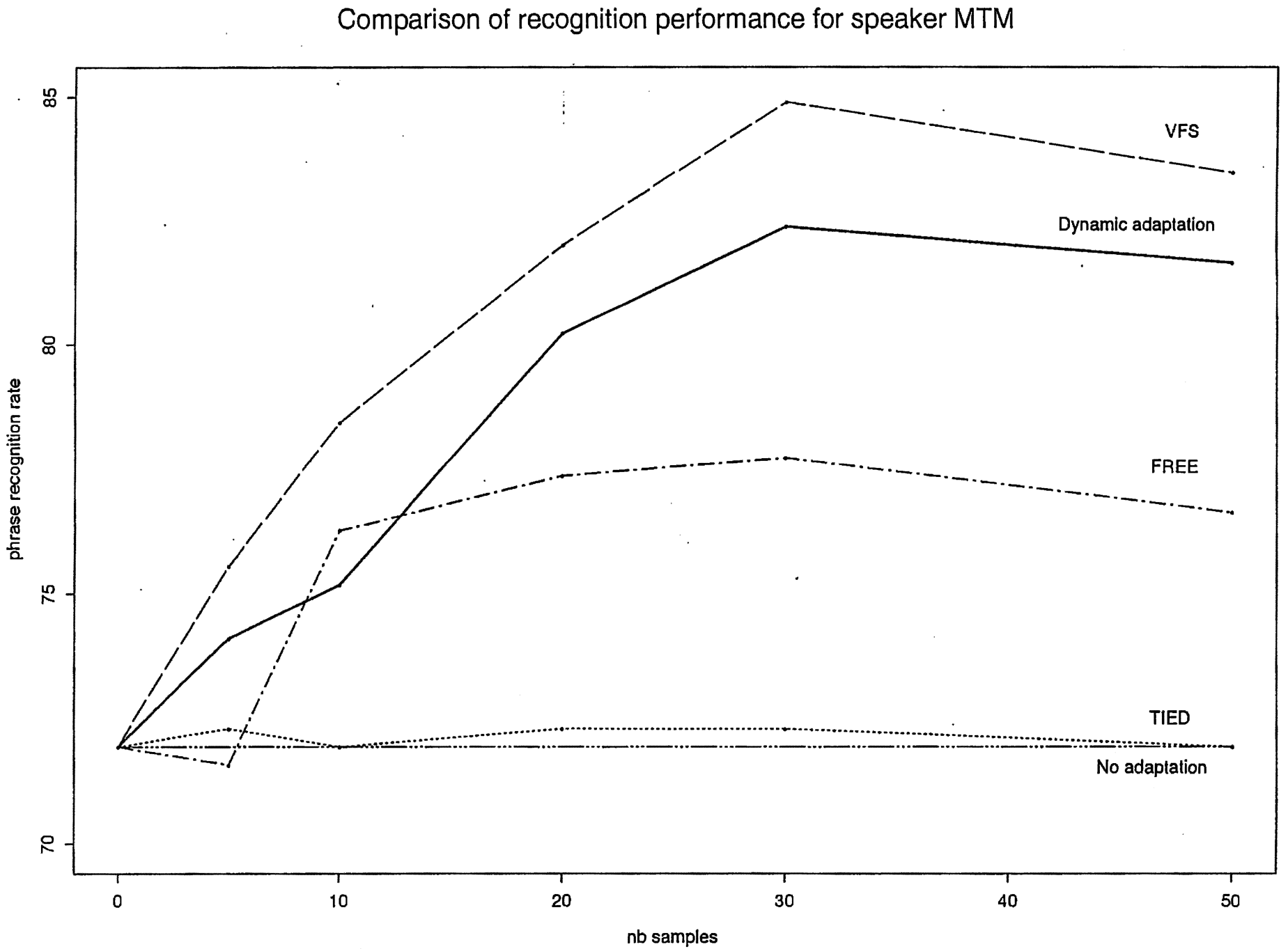


30 phrases



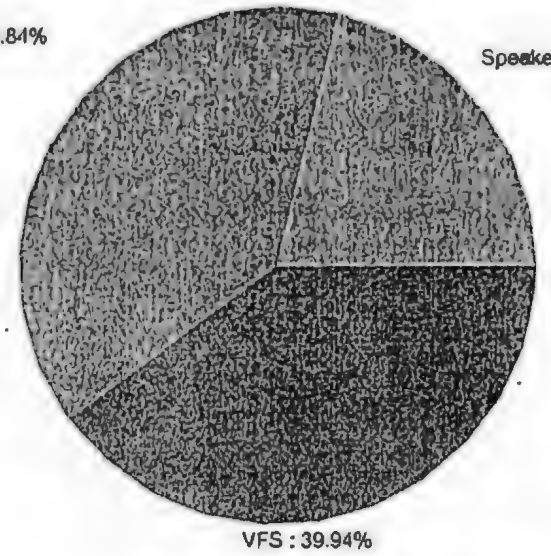
50 phrases



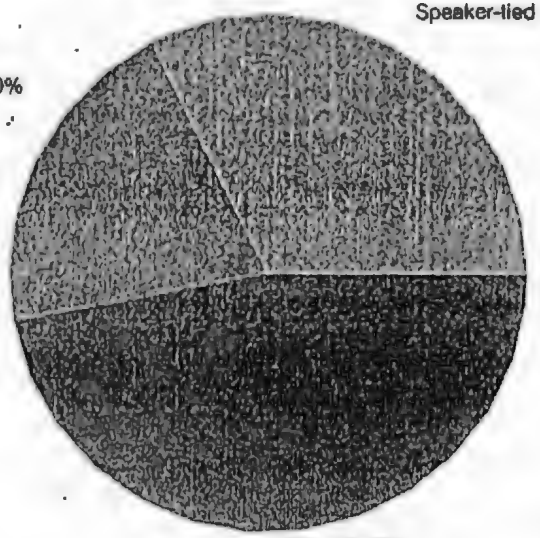

Comparison of recognition performance for speaker MTM
phrase recognition rate
85
80
75
70
0
10
20
30
40
50
nb samples
VFS
Dynamic adaptation
FREE
TIED
No adaptation

Selected methods for Speaker MTM

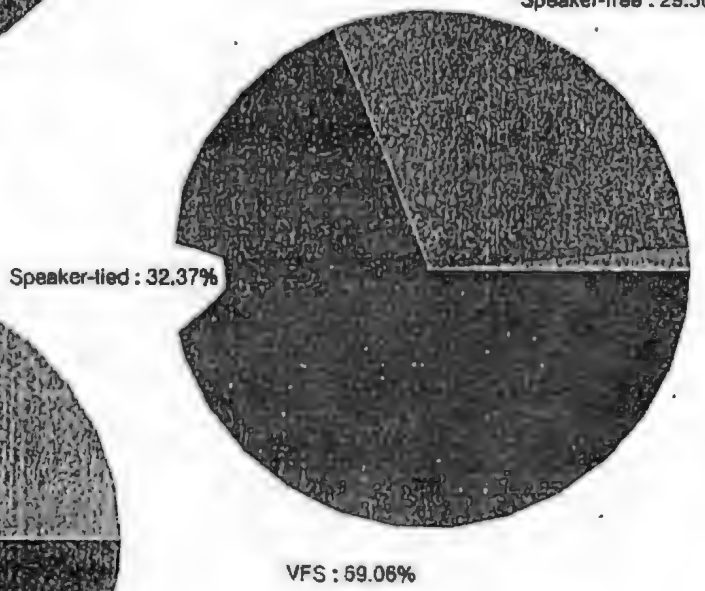
5 phrases



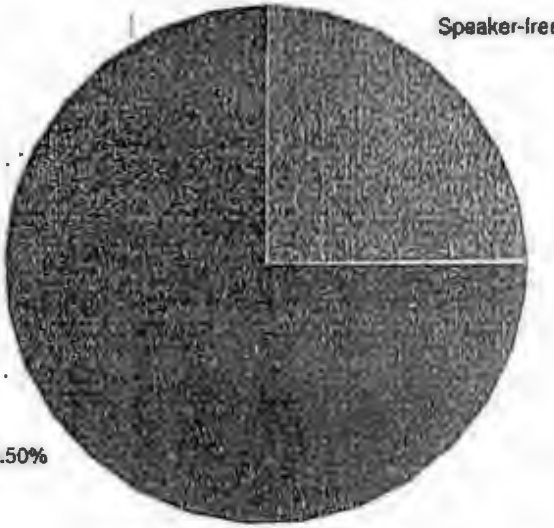
10 phrases



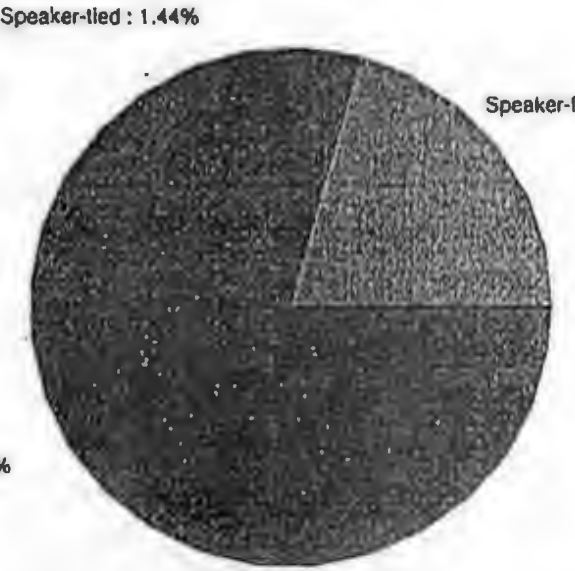
20 phrases

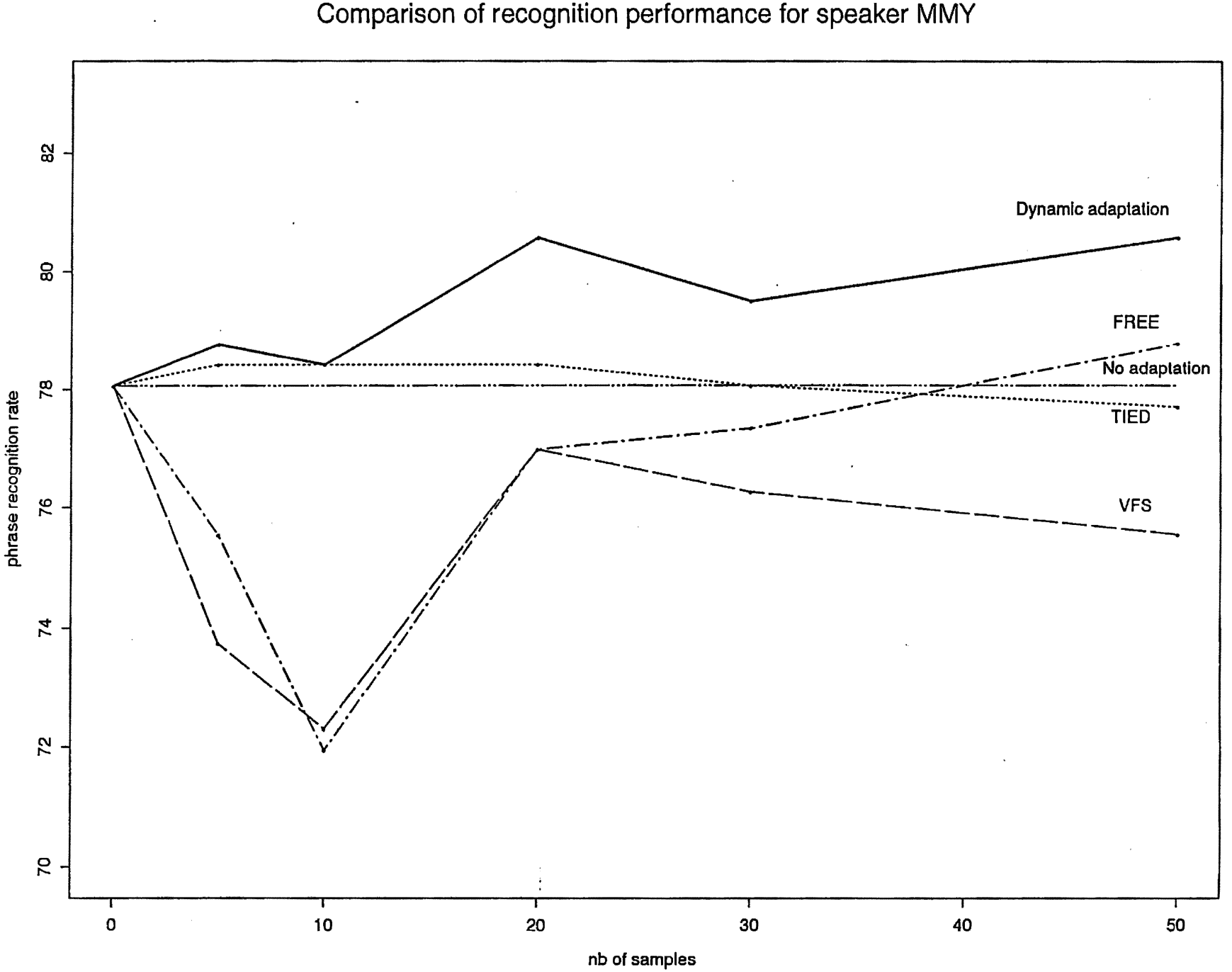


30 phrases



50 phrases




Comparison of recognition performance for speaker MMY
Dynamic adaptation
FREE
No adaptation
TIED
VFS
phrase recognition rate
82
80
78
76
74
72
70
0
10
20
30
40
50
nb of samples

Conclusion and Further Study

- we presented a dynamic approach to the design of speaker-adaptive recognition systems

  → experiments showed it achieved higher recognition rates

Further Study

- dynamic method adaptation
- extend to unsupervised systems