TR-I-0291

Variance Spreading for Better Recognition using HM-Nets : a deterministic approach

Edward WILLEMS, Tetsuo KOSAKA, Shigeki SAGAYAMA

October 1992

ABSTRACT

Empirical observation has shown that substantial improvements in speaker-independent recognition using HM-Nets can be achieved, if all the variances in the HM-Net are spread by a same factor. Optimal choice of the value of this factor can increase the recognition rate by as much as 20 percent. However, the value of the factor is currently determined heuristically for each speaker, which undermines the usefulness of this adaptation method. The purpose of the study presented in this report was to investigate the underlying parameters in the speech signal which determine the optimal spread factor value, in order to devise a more robust procedure for speaker-adaptation. The study concentrated on finding a correlation between the optimal spread factor value and the difference between the parameter distributions of the HM-Net, and those determined directly from the speaker's speech data. The data from five out of six speakers showed a correlation between Bhatacharyya distance, which is a measure of both mean and variance difference, and optimal spread factor value.

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

INDEX

1 Introduction

1.1 Speaker independent recognition with HM-Net systems

The stochastic properties of Hidden Markov Networks (HM-Nets) [7] make them a powerful tool for speech recognition, because they can account for the variations in phoneme elocution encountered in everyday speech. The process of determining the probability distributions of the HM-Net is called model *training* [4]. If the model is suitably trained, it is possible to achieve high recognition rates (greater than 90 percent) for speaker-specific systems. However, this performance is strongly degraded and can be as low as 20 percent if this same model is used to recognise a different speaker's speech, as is the case for speaker-independent recognition. This is simply because the stochastic properties of the model no longer describe the speaker's phones accurately, and it illustrates the importance of model training.

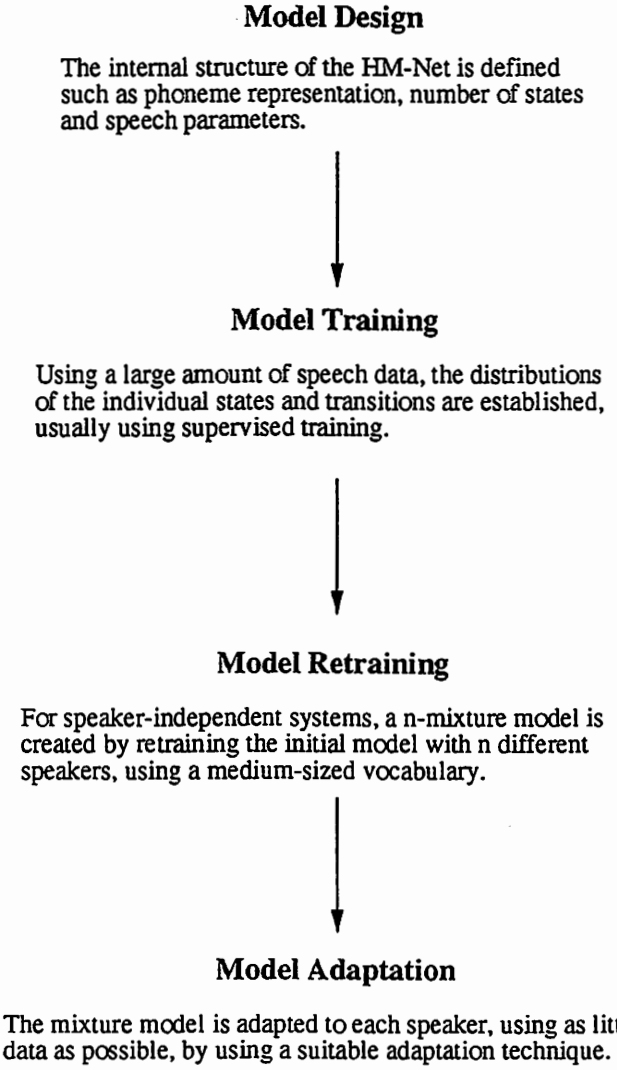

Figure 1. The main steps in the design of a speaker-independent recognition system using HM-Nets

In the case of speaker-independent recognition, it is no longer possible to train the model to the speech characteristics of every user. To achieve high recognition rates, the initial training has to produce a fairly 'neutral' stochastic model of phoneme elocution. To achieve this,

it is common to train n different HM-Nets using n different speakers, and to combine their properties into one 'super-HM-Net', which is then called an *n-mixture* HM-Net [5]. This way higher recognition rates can be achieved. However, the specific characteristics of each individual speaker will still strongly influence the overall performance.

The usual way of circumventing this problem is to use the actual speech data produced by the user to adapt the model to his voice. The idea is to use as little data as possible for the adaptation, so that the system remains practical to use. Various adaptation techniques have been successfully implemented and provide good results (Vector Field Smoothing [6], Speaker-tied weight training [5], Fixed weight training [5]). The adaptation technique discussed here is the so-called *variance spread adaptation*, which is a simple and efficient means of initially adapting mixture HM-Nets. It is discussed in more detail in the following section. Figure 1 reviews the different stages in the design of a speaker-independent speech recogniser.

1.2 An empirical observation : spreading the variances for better recognition in Hidden Markov Networks

The principle of variance spread adaptation is very simple : inter-speaker differences are ironed out by broadening the distributions in the HM-Net. This is illustrated in figure 2. The model's distribution and the 'true' distribution differ both in mean and variance. The black regions show the joint probability of the two distributions (P1(x)∩P2(x)). By increasing the variance of the model's distribution, this region is increased, hence the probability of recognising the actual speaker's speech is also increased. However, the distribution becomes less sharp.

The adaptation is carried out by defining a single factor, the *spread factor*, by which we multiply every distribution in the model. This has obvious limitations : if the factor is too small, the adaptation is insufficient, if it is too big, there is no discrimination between the state outputs and the recognition is worse. No distinction is made between speech parameters, so the adaptation would seem to be coarse and ineffective.

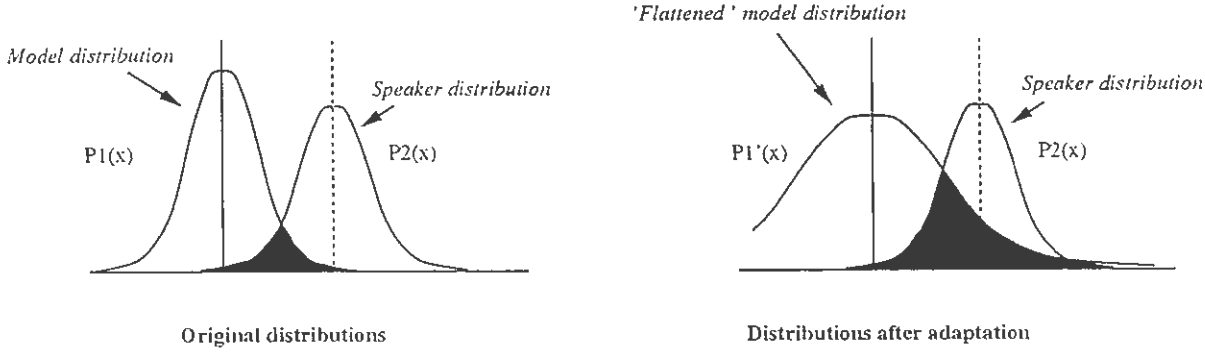

Figure 2. Principle of variance spread adaptation

However, experience has shown that an optimal choice of spread factor can actually lead to improvements of up to 20 percent in the recognition rate for speakers. It is therefore on the contrary very effective, which together with its extreme simplicity make it very attractive as

an initial adaptation technique.

The current method results from empirical observation, and the actual spread factor used is derived heuristically. We do not know how to obtain the optimal spread value from speaker input data, so a heuristic value is chosen and applied for every speaker. This method is far from satisfactory for many reasons of which the following are the most important :

- the optimal factor value varies substantially from speaker to speaker, and there is no *a priori* means of deriving it.
- there is no discrimation between parameter type, i.e. the same factor is applied to every distribution, regardless of the actual differences between model and speaker distributions.
- the heuristic value is always sub-optimal.

In order the remove these limitations, it would be extremely interesting to look into the precise characteristics of the speech data which determine the spread factor value, for the following reasons :

Reason 1 variance spreading is a simple means of adaptation, which enables high recognition rates to be achieved. If it can be derived directly from the input speech data, it could prove to be powerful adaptation technique for independent speech recognition using HM-Nets.

Reason 2 the method is currently very crude, so its analysis could reveal a more sophisticated method of adapting the model, producing better results.

Reason 3 the current method relies on heuristics, so is of limited interest for robust system design.

This report investigates the variance spread adaptation technique, by studying the correlation between the optimal spread factor value and the dissemblance between the model and speaker-based distributions.

The research was carried out using 6 male speakers, uttering phrased speech. A heuristic procedure to determine the optimal spread factor value for each of these speakers was first carried out. This is the topic of the next section. The next four sections present the four different procedures for automatic estimation of the spread factor which were studied : direct variance calculation, Mahalanobis distance, Bhatacharyya distance and log likelihood.

2 First stage : Heuristic determination of the Variance Spread Factor

2.1 General Description

To establish the optimal value for the spread factor for each speaker, the following procedure was adopted : for each speaker in turn, we adapted the model using spread values ranging from 1.0 (no adaptation) to 3.0, the usual heuristic optimum being normally situated around 2.0. We tested each adapted model using the same test data, and plotted the recognition rate for each value. The optimal value is then the value which achieves the highest recognition rate.

The experimental conditions are the following :

model used 12-mixture HM-Net, consisting of 200 states, with 2 to 4 states per phoneme.

speech parameters 34 parameters are used : log power + 16 cepstral coefficients, and the associated delta coefficients.

training data phrased speech, uttered by six professional male speakers (code names : MMS, MMY, MSH, MAU, MHT, MTT).

test data 279 phrases, uttered by the same professional male speakers

test software left-right phrase parser (SSS-LR, ATR laboratories)

2.2 Results

The results are plotted in figure 3. The optimal spread factor values are tabulated below :

Speakers	MMS	MMY	MSH	MAU	MHT	MTT
Spread Factor	2.4	2.2	1.4	2.4	2.2	2.6
Recognition Rate	80.58	74.10	78.42	75.27	85.87	82.01

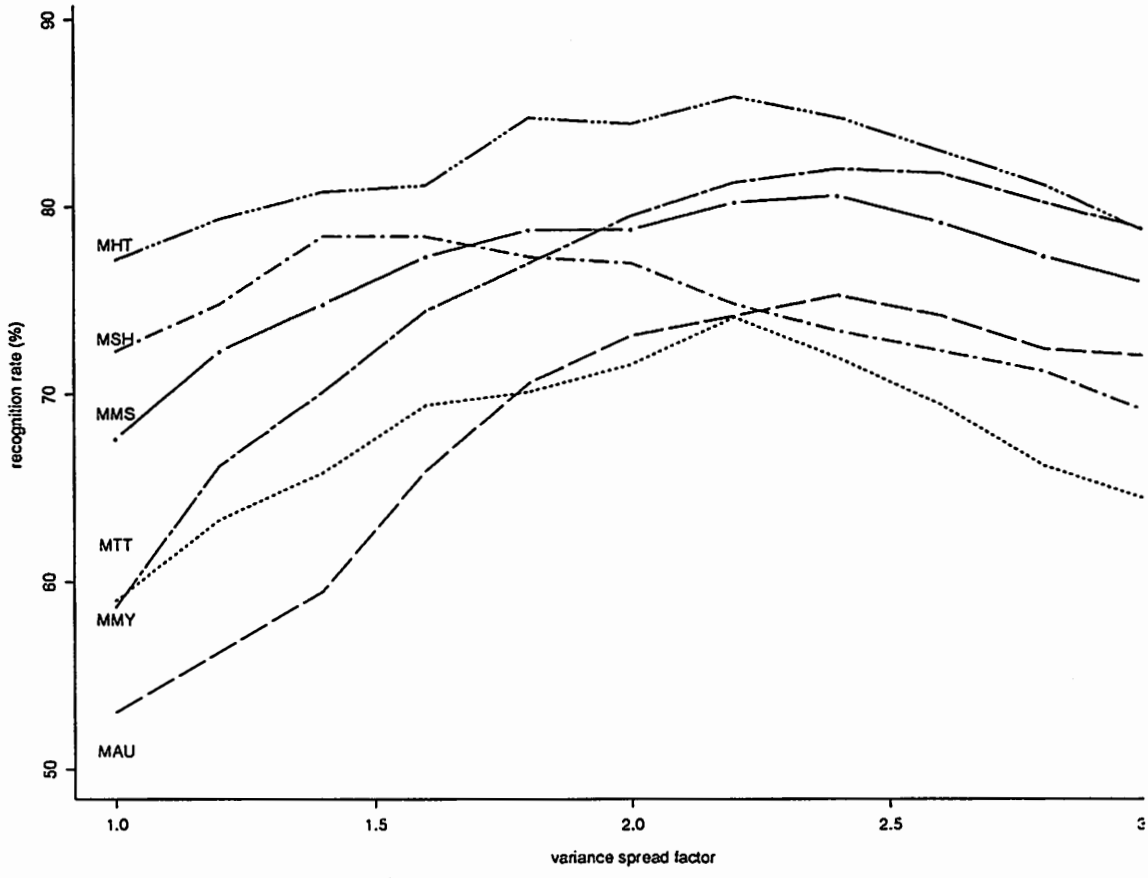

Figure 3. Heuristic determination of the optimal spread factor

As can be seen, the optimal spread factor values vary from speaker to speaker : 1.4 for MSH, 2.6 for MTT. These values are used as the reference data for evaluation of the procedures described in the next four sections.

3 Direct Variance Calculation

3.1 Principle

The idea in this method is to use the speaker's training data to calculate directly a new variance value $\sigma^2_{speaker}$, which is computed for each distribution as follows :

$$\sigma^2_{speaker} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_{model})^2$$

where μ_{model} is the mean of the model's distribution and the x_i are the new data values. Since the distributions are, in general, slightly offset, we could expect this new variance value to be representative of the dissimilarity between the distributions. The overall factor is computed as :

$$F = \frac{1}{M}\sum_{i=1}^{M}(\frac{\sigma^2_{ispeaker}}{\sigma^2_{imodel}})$$

3.2 Experimental Results

The original program for matrix adaptation, *Adapt.sh*, was modified to implement this method of variance factor calculation. The new program was called *Var-adapt.sh*. It was

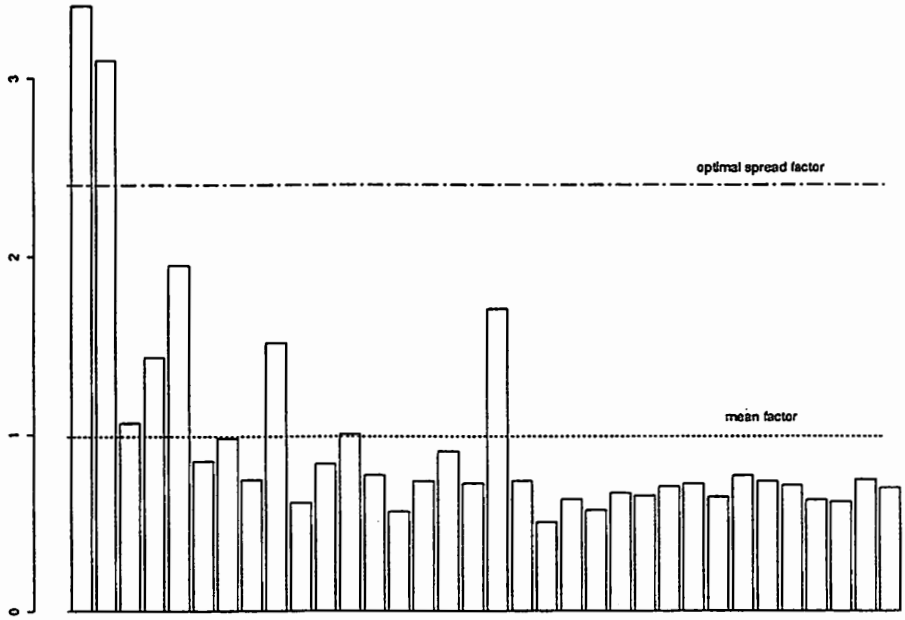

Figure 4. Parameter factors for speaker MMS

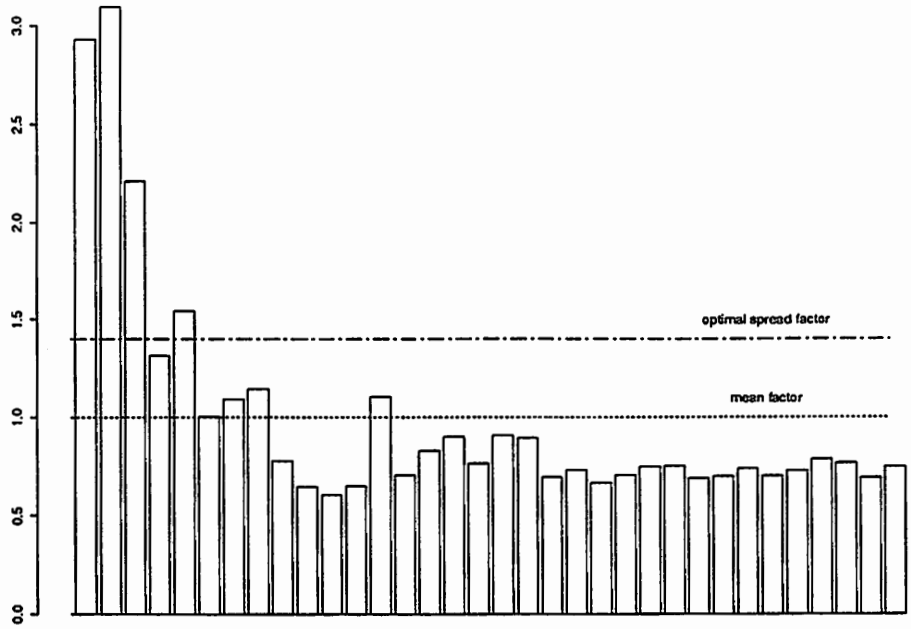

Figure 5. Parameter factors for speaker MSH

tested and ran on all six speakers, using 50 phrases to train the model, to ensure sufficient data. The results were very poor :

- the computed factors were all close to 1.0, which is way below the heuristic optimal values.
- there was no correlation between computed value and observed value.

So the method had to be altered. We noted that the method averages the $\frac{\sigma^2_{speaker}}{\sigma^2_{model}}$ over all states and over all parameters. It is unlikely that all parameters are equally strongly influenced by new speaker data. We also computed $\sigma^2_{speaker}$ for each state that was accessed in the model, regardless of the number of new data samples available. It is obvious that a $\sigma^2_{speaker}$ obtained using, say, 2 samples is statistically unsound. To investigate the influence of these two points, we imposed a minimum of 50 samples for variance calculations and averaged over states, but not over parameters. The results were equally poor, showing no correlation between the calculated values and the observed ones. This was the case for all speakers. Figures 4 and 5 show the results for speakers MMS and MMY.

The results for the other speakers were similar to those in figures 4 and 5. The table below summarises the results for direct variance calculation.

Speakers	MMS	MMY	MSH	MAU	MHT	MTT
Optimal Factor	2.4	2.2	1.4	2.4	2.2	2.6
Computed Factor	0.988	1.269	0.998	1.578	1.138	1.417

3.3 Conclusion

We can conclude that there is no correlation between any one parameter factor and the optimal spread factor, nor is the mean of all such factors a good estimate for the variance spread factor. Another method must be used.

4 Mahalanobis Distance Calculation

4.1 Principle

The previous method was derived rather haphazardly, which could explain the poor results. The method presented here and in the next section rely both on statistical analysis of the two distributions (model and speaker) for each parameter in each state. The hypothesis we are trying to justify is still that the spread factor is linked with the extent in which the distributions are dissimilar. The first assumption we make is that the distribution have the same shape, but that their means are different. If this is the case, then a statistical measure of mean difference should correlate with spread factor value. The Mahalanobis distance [2, pages 22–24] [1, pages 204–208]is such a measure.

(4.1.1) Definition of the Mahalanobis distance

The general definition of the Mahanalobis distance r between two multivariate Gaussian distributions is the following :

$$r^2 = (\mu_1 - \mu_2)^{\mathrm{T}} \Sigma^{-1} (\mu_1 - \mu_2)$$

where μ_1 and μ_2 are the mean vectors and Σ is the covariance matrix. The Mahalanobis metric assumes both distributions share covariance matrix.

In our case, we are dealing with univariate, independent distributions, so the distance simplifies to :

$$r^2 = \frac{(\mu_1 - \mu_2)^2}{\sigma^2_{model}}$$

so the distance becomes simply the difference in means, normalised in terms of the variance of the model's distribution.

4.2 Experimental Results

A procedure to compute the Mahalanobis distance was included in the *Var-adapt.sh* program. The results are tabulated below :

Speakers	MMS	MMY	MSH	MAU	MHT	MTT
Optimal Factor	2.4	2.2	1.4	2.4	2.2	2.6
M-distance	0.011	0.012	0.005	0.017	0.006	0.009

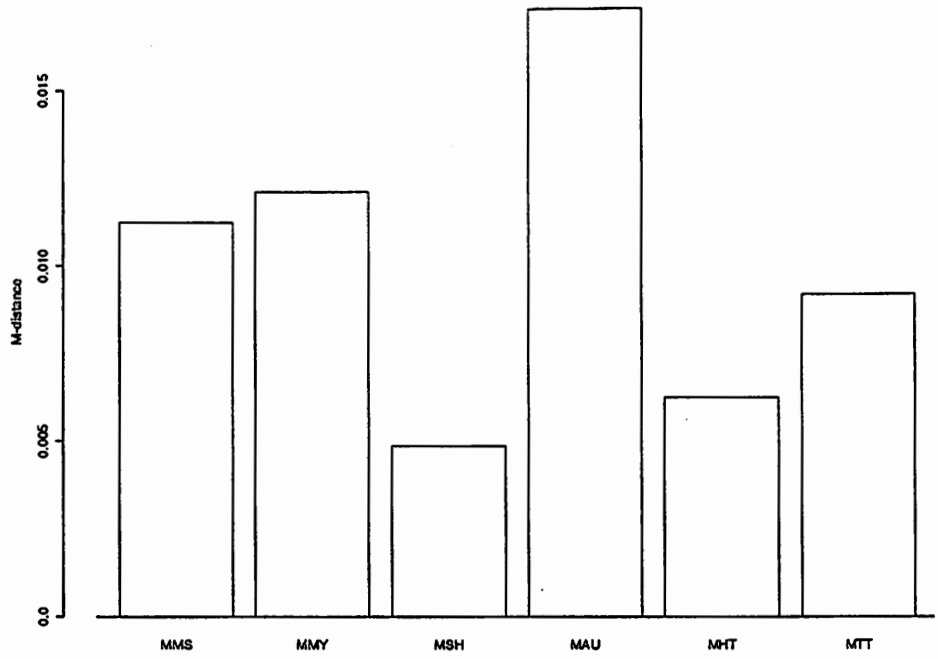

Figure 6. Average Mahalanobis distance for each speaker

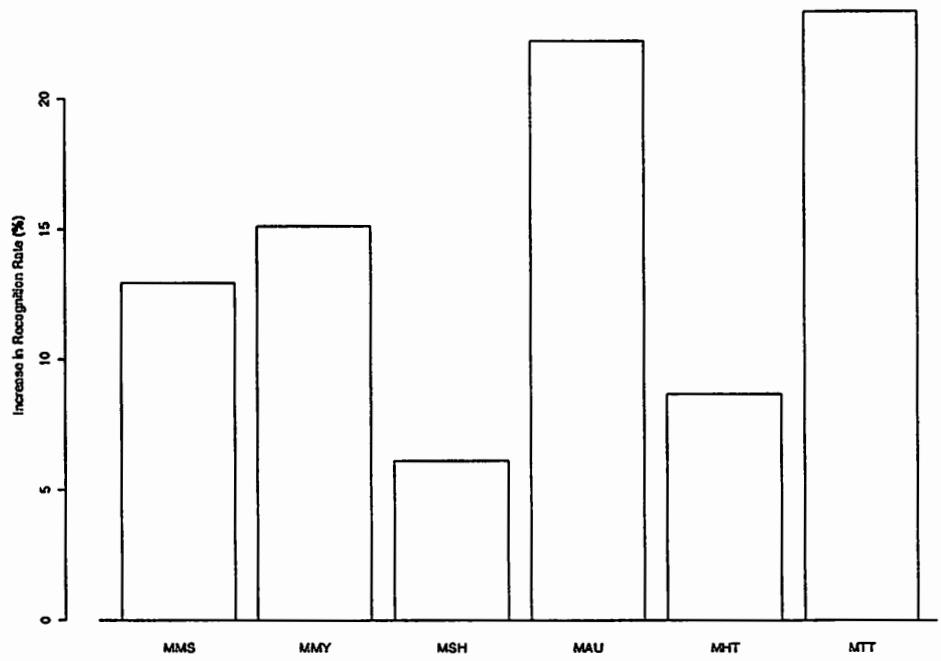

Figure 7. Increase in recognition rate achieved when using the optimal spread factor value

So again no direct correlation was observed. The average Mahalanobis distance for all speakers in shown in figure 6. It is interesting to compare these distances with the increase in recognition rate achieved using the optimal spread factor for each speaker. This is represented

in figure 7. As can be seen, there exists a strong correlation between the two for speakers MMS, MMY, MSH, MAU and MHT. In other words, this means that the greater the normalised mean difference, the more efficient the variance adaptation. However, the value obtained for speaker MTT does not correspond to the expected value, and the initial recognition rate for an unadapted model is not related to the distance.

4.3 Conclusion

The Mahalanobis distance is not correlated with the variance spread factor. It does seem to show some correlation with the increase in recognition rate, but this was contradicted totally by one of the speakers. The correlation is therefore weak, if it exists at all. However, as stated above, we assumed both distribution to have the same variance. We should now question this assumption, by using a measure which incorporates both mean and variance differences. This is the topic of the next section.

5 Bhattacharyya Distance Calculation

5.1 Principle

The Bhattacharyya distance is a popular measure of the separability of two distributions [3, pages 188–201]. Its value depends both on mean and variance differences, so it seems well suited to describe the general difference between model and speaker characteristics.

(5.1.1) Definition of the Bhattacharyya distance

The Bhattacharyya distance is often considered as a upper bound for the Bayes error, less accurate but simpler than the *Chernoff* distance [3, pages 99–110]. For this reason it is sometimes referred to as the Bhattacharyya bound. Its expression is :

$$D^2 = \frac{1}{8}(\mathbf{M}_2 - \mathbf{M}_1)^{\mathrm{T}} \left(\frac{\Sigma_1 + \Sigma_2}{2}\right)^{-1} (\mathbf{M}_2 - \mathbf{M}_1) + \frac{1}{2} ln \frac{\left|\frac{\Sigma_1+\Sigma_2}{2}\right|}{\sqrt{|\Sigma_1|}\sqrt{|\Sigma_2|}}$$

where the symbols have their usual meaning (cf. Mahanalobis distance).

The distance consists therefore of two terms. The first one depends on the means, and cancels out if they are equal. The second term is a function of variances only, and also cancels out if these are equal. It is interesting to note that if the variances are equal, the equation becomes :

$$D^2 = \frac{1}{8}(\mathbf{M}_2 - \mathbf{M}_1)^{\mathrm{T}} \Sigma^{-1} (\mathbf{M}_2 - \mathbf{M}_1) = \frac{1}{8} r^2$$

,

where r is the Mahalanobis distance.

In our case the distributions are univariate independent Gaussian distributions, so the Bhattacharyya distance simplifies to :

$$D^2 = \frac{1}{4} \frac{(\mu_2 - \mu_1)^2}{\sigma_1^2 + \sigma_2^2} + \frac{1}{2} ln \frac{\sigma_1^2 + \sigma_2^2}{2\sigma_1\sigma_2}$$

5.2 Experimental results

A procedure to calculate the average Bhattacharyya distance for each parameter was implemented in program *Var-adapt.sh*. The overall distance was computed by first calculating the average distance over all states for each parameter, then taking the average of these 34 distances. The results are summarised below :

Speakers	MMS	MMY	MSH	MAU	MHT	MTT
Optimal Factor	2.4	2.2	1.4	2.4	2.2	2.6
B-distance	0.054	0.031	0.040	0.057	0.052	0.059

These results appear in figure 8. The correlation between spread factor and B-distance is promising for speakers MMS, MSH, MAU, MHT and MTT. However, speaker MMY contradicts the correlation.

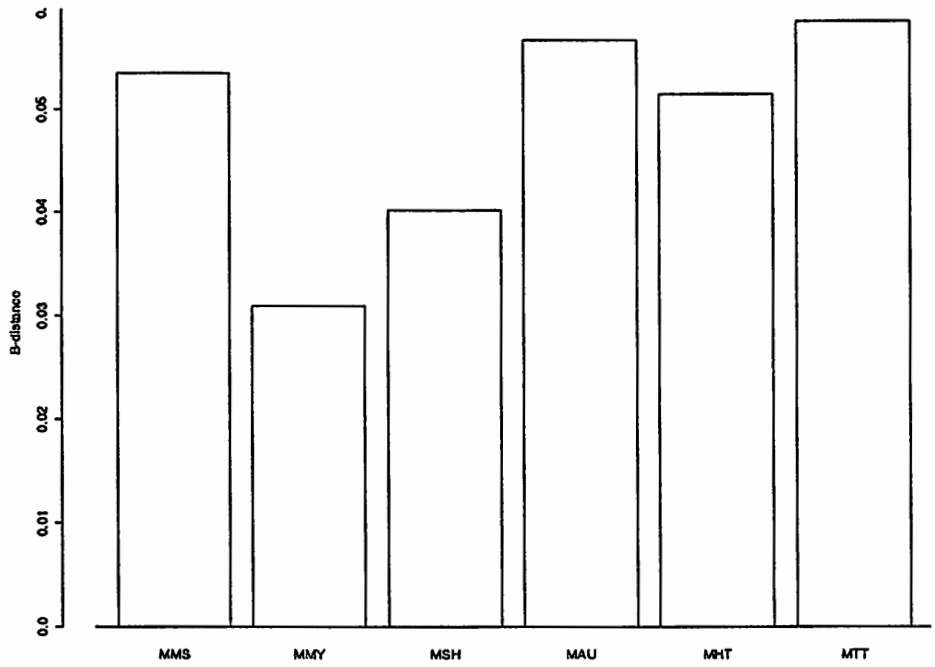

Figure 8. Average Bhattacharyya distance for each speaker

Figures 9 and 10 show the average first term and second term respectively. These have been included to show the difference from speaker to speaker. They confirm that the Mahalanobis distance, which assumed variances to be equal, was not well suited, since figure 9 is very different from figure 6.

5.3 Conclusion

The Bhattacharyya measure has so far been the closest to correlating with the optimal spread factor value. It has also confirmed that the speaker distributions differ from the model's

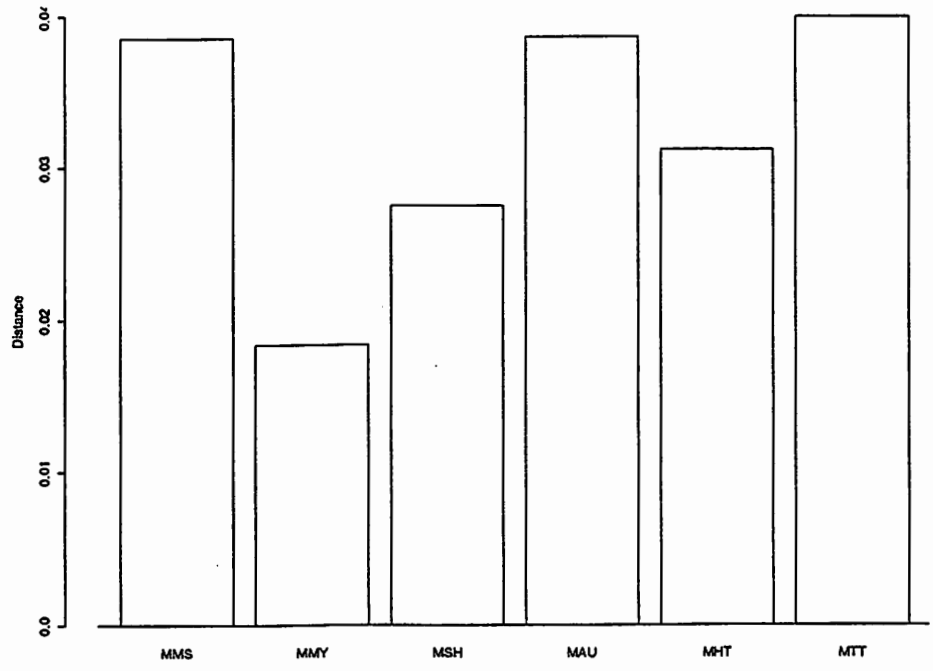

Figure 9. Average first term of the Bhattacharyya distance for each speaker

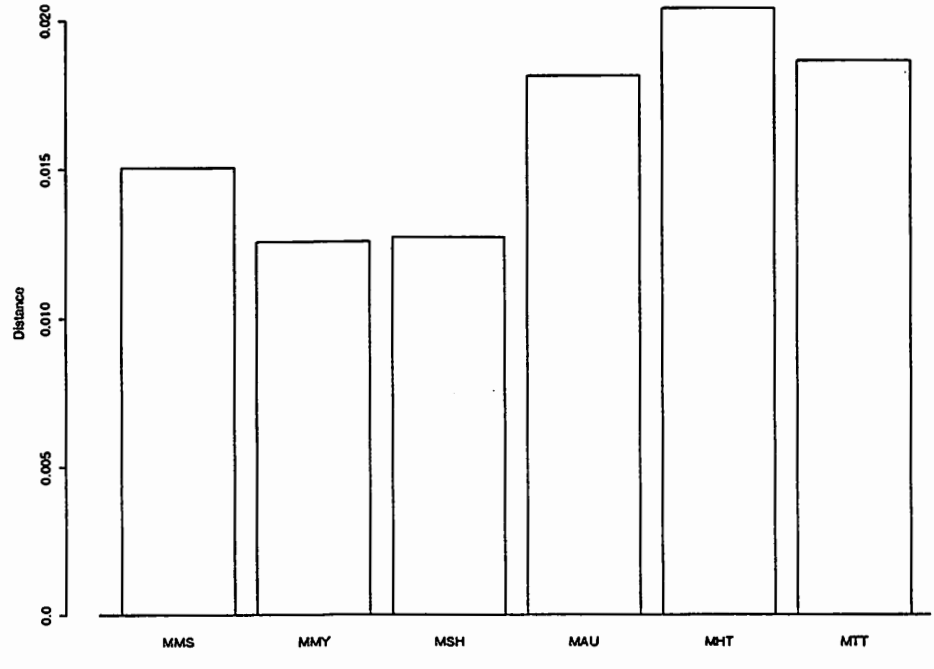

Figure 10. Average second term of the Bhattacharyya distance for each speaker

ones *both* in mean and variance. This would mean that the variance spreading adaptation method could perhaps be improved, if some measure was taken concerning the means of the distributions. However, there is still one speaker for which there was no correlation, so the problem of robust spread factor determination still remains.

6 Studying the Log-Likelihood

6.1 Principle

The final criterion for spread factor estimation that we studied was the correlation between the log-likelihood for each recognised phrase. We therefore check the output from the recogniser for different spread factor values. The idea is that the more a speaker differs from the model, the less likely the recognition will be.

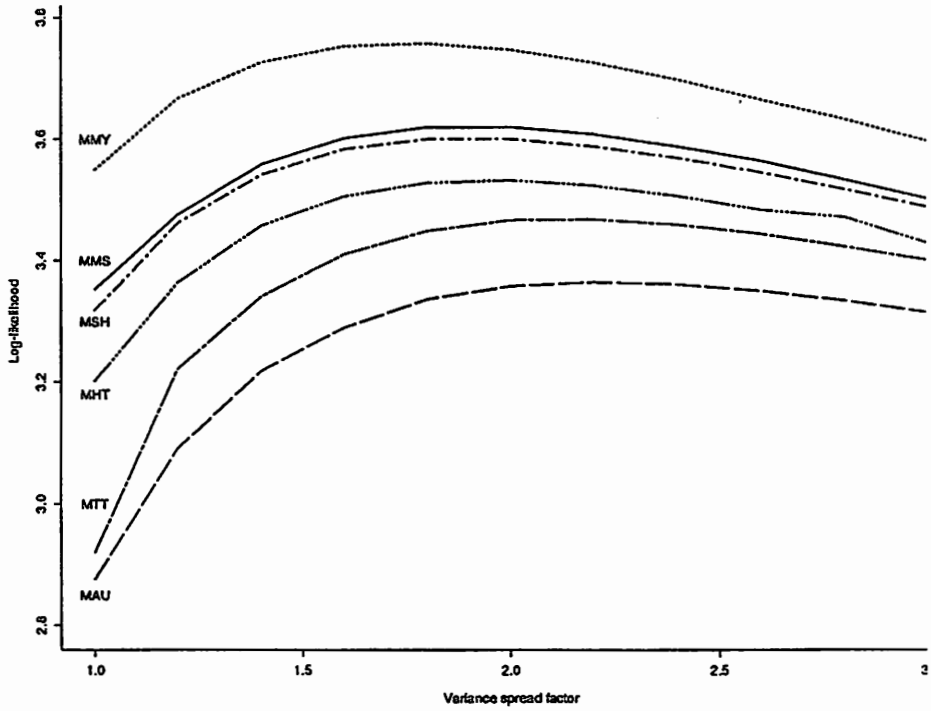

Figure 11. Log-likelihood of the first ten phrases for each speaker, for different spread factor values

6.2 Experimental results

The log-likelihood for each phrase is directly computed by the recognition algorithm, the *SSS-LR* program, and we calculated the average log-likelihood of the first ten phrases for each speaker. Figure 11 shows the resulting curves.

The results show that no correlation could be observed.

7 Conclusion

The aim of the study was to investigate the characteristics of speaker and model distributions which would explain the excellent performance achieved by variance spread adaptation. We tested four different to compute the optimal spread factor value using the available speech data :

- direct calculation of the variance factor
- correlation with the average Mahalanobis distance between distributions
- correlation with the average Bhattacharyya distance between distributions
- correlation with the log-likelihood of recognition

None of these produced satisfactory results. In all four cases we calculated a general average value, but this is justified by the fact that the spread factor is applied to all distributions, so its performance is the result of the overall accuracy of this procedure. This general method is however questionable, since it is unlikely that all distributions need to be changed in a similar manner.

It appeared from the study that the optimal spread factor is not a simple function of distribution separability, but depends on extra parameters which still need to be determined. The results obtained using the Bhattacharyya distance showed that the means and variances of speaker data differ substantially from model distributions, so the simple spreading of the variances may not lead to the most efficient results. In any case, the study used 6 speakers to test the results, which is insufficient. It would be interesting to investigate the Bhattacharyya distance using many more speakers, and to check the overall correlation with the optimal spread factor value. Another useful further study would be to actually adapt the model using a method based on the Bhattacharyya distance, to observe how efficient this technique could be.

It is quite clear that the heuristic nature of the variance spread adaptation method and the coarse manner in which the adaptation is carried out are both unsatisfactory. Further study should enable a more sophisticated method of adaptation, based on distribution separability, to be derived.

REFERENCES

[1] T. W. Anderson. *An introduction to multivariate statistical analysis.* John Wiley and Sons, New York, 1984.

[2] R. Duda and P. Hart. *Pattern classification and scene analysis.* John Wiley and Sons, New York, 1973.

[3] K. Fukunaga. *Introduction to statistical pattern recognition.* Academic Press, Inc, San Diego, 1990.

[4] X.D. Huang, Y. Ariki, and M.A. Jack. *Hidden markov models for speech recognition.* Edinburgh University Press, Edinburgh, 1990.

[5] T. Kosaka, J. Takami, and S. Sagayama. Rapid speaker adaptation using speaker-mixture allophone models applied to speaker-independent speech recognition. In *Proc. IEEE ICASSP–93*, 1993. (to appear).

[6] K. Ohkura, M. Sugiyama, and S. Sagayama. Speaker adaptation based on transfer vector field smoothing with continuous mixture density hmms. In *Proc. International Conference on Spoken Language Processing*, Banff, Canada, 1992. (in press).

[7] J. Takami and S. Sagayama. A successive state splitting algorithm for efficient allophone modeling. In *Proc. IEEE ICASSP–92*, volume 1, pages 573–, San Francisco, USA, 1992.