Information Theoretic Factorization of Speaker and Language in Hidden Markov Models, with Application to Speaker Recognition

Naftali Tishby

AT&T Bell Laboratories

Murray Hill, New Jersey 07974

Abstract

An information theoretic approach to speech modeling with prior statistical knowledge is proposed. Using the concept of minimum discrimination information (MDI), a model of speech can be factored into a prior distribution and an exponential correction term, depending on the specific training data. The discrimination information measures the statistical deviations of the training data from a prior model, in a way that is known to be optimal in a well defined sense. The minimization of the discrimination information, subject to the given training data as constraints, yields a set of Lagrange multipliers. These multipliers serve to characterize the part of the training data which is not described by the prior model. The problem of separating the speaker dependent part from a "universal" speaker independent prior in hidden Markov models is studied in this framework and a practical method for achieving this separation is derived. As an example, universal hidden Markov priors for isolated English digits are trained for male and female speakers, using a database of 100 speakers and 20000 spoken digits. The speaker specific part is modeled by the individual Lagrange multipliers obtained by minimizing the discrimination information between the training data and the corresponding prior language model.

1. Introduction

One of the most persistent problems in automatic speech recognition is the inability to separate the speaker dependent from the 'universal' speaker independent properties of the speech signal. When training a statistical speech model, these different characteristics of the signal are mixed in the estimation of the model's parameters. This is a special case a the more general problem: speech modeling with prior information. By prior information we consider any knowledge about the speech signal available before the specific training data is observed. In the above problem, the prior information can be the knowledge of the spoken language, the vocabulary used, or even the sex and age of the speaker.

One possible approach to this problem is separating the modeling into two phases. The first is a statistical modeling of the prior information, while the second models the deviations, or residue, of the specific training data from this prior model. The prior model is trained to capture the common features of different speakers, of the same sex, using the same language and vocabulary. Such a model is, however, vague about the speaker specific information in the training data. Using this prior statistical model, our knowledge is expressed in terms of a prior probability distribution function. Once such a prior distribution is available, a general information theoretic method can be used for the second modeling phase,[1] by minimizing the discrimination information (MDI) between the observed data and the prior distribution. This constrained minimization procedure yields a set of Lagrange multipliers which serve as the speaker specific parameters in this study. The speech probability distribution function is thus factored in the desired form

$$P('speech') = p^0('language') \cdot p('speaker') \qquad (1.1)$$

The discrimination information itself provides the natural distance measure in this approach.

A general procedure for training hidden Markov models (HMM) for speech recognition, using MDI, was recently proposed by Ephraim, Dembo and Rabiner.[2] In the present study similar concepts are used, but the prior distribution is given an entirely different meaning, while the Lagrange multipliers serve as part of the model's parameters. Though the focus of this work is on the problem of speaker recognition, similar modeling methods can be used for other problems, such as speaker normalization in speech recognition.

2. Prior knowledge and constraints

Let $\mathbf{Y}_T \equiv (\mathbf{y}_0, \mathbf{y}_2, \ldots, \mathbf{y}_T)$ be a sequence of stochastic observation vectors, where $\mathbf{y}_t \in \mathbf{R}^n$. In the case of speech $\mathbf{y}_t$ is just a frame of speech samples. All the prior knowledge on these observations can be given in terms of a probability distribution function $p^0(\mathbf{Y}_T)$, which characterize our prior probability to observe a certain sequence. Without losing any generality, we assume that any additional measured information is given by sample averages of observables, which are real functions of the sample space $\mathbf{Y}_T$: $F(\mathbf{Y}_T)$. The averages of these functions play the role of relevant coordinates in this approach. Choosing the 'right' observables for the given problem is crucial, and is done by using independent knowledge or considerations which are beyond the general formulation. In some cases these observables are the sufficient statistics for the parameters of the model.[3]

Sample averages are (linear) equations in the distribution and therefore restrict its possible values. We thus refer to these averages of the 'relevant' observables, as the *constraints* of the problem. By identifying the sample averages as the expected values of the relevant observables with a probability distribution q, these constraints can be written as:

$$f_k \equiv E_q\{F_k(\mathbf{Y}_T)\}, \; 1 \le k \le K \; . \qquad (2.1)$$

This additional information, i.e. the constraints, is generally not sufficient to determine a single probability distribution. There is always a *convex* set of distributions which satisfy those constraints

$$\mathbf{Q}_M \equiv \{q(\mathbf{Y}_T) \mid E_q\{F_k\} = f_k \; , \; 1 \le k \le K\}. \qquad (2.2)$$

This set can be empty if the constraints happen to be inconsistent.

Our first problem is to select a unique distribution out of this set which 'best' represents our complete knowledge, both in the prior and in the constraints. One possible way of doing this is to select the distribution which is 'least biased',[4] i.e. contains no other unknown information, out of all the possible $q \in \mathbf{Q}_M$. This is done by maximizing the entropy of the distribution given the prior and the constraints, or equivalently, by selecting the distribution $\hat{q}$ which *minimizes the discrimination information* between the prior p^0 and the set $\mathbf{Q}_M$.

CH2561-9/88/0000-0087 $1.00 © 1988 IEEE

3. Minimum Discrimination Information (MDI)

Information theory provides us with a natural distance measure between probability density functions: the *discrimination information* (also known as cross-entropy, directed divergence, Kullback-Liebler number, etc.). The discrimination information between the distributions $q(\mathbf{x})$ and $p(\mathbf{x})$ is defined by:

$$D(q \mid p) = \int d\mathbf{x}\, q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})} \tag{3.1}$$

This is a non-negative asymmetric measure, with many important properties, which are discussed in detail in the literature.[5] In particular, the measure is additive over statistically independent events and is a convex function of its arguments. This guarantees the existence of a unique minimum of $D(q \mid p)$ inside any convex and compact set of distributions, in particular in the convex set $\mathbf{Q}_M$, if this set is compact.

Minimaizing D subject to linearly independent constraints of the form (2.2), using the method of Lagrange undetermined multipliers, yields an exponential form for the posterior distribution

$$\hat{q}(\mathbf{x}) = p(\mathbf{x}) \exp\left[-\lambda^0 - \sum_{k=1}^{K} \lambda^k F_k(\mathbf{x})\right] . \tag{3.2}$$

The λ^k is the unique vector of Lagrange multipliers attached to the vector of constraints $E_q\{F_k\} = f_k$, and λ^0 is a normalization function, defined as

$$\lambda^0(\lambda^1, \ldots, \lambda^K) = \log \int d\mathbf{x} p(\mathbf{x}) \exp\left[-\sum_{k=1}^{K} \lambda^k F_k(\mathbf{x})\right] . \tag{3.3}$$

The minimal discrimination information is expressed in terms of the Lagrange multipliers and the constraints values as

$$D_{\min} = D(\hat{q} \mid p) = -\lambda^0 - \sum_{k=1}^{K} \lambda^k f_k \equiv -\lambda^0 - \boldsymbol{\lambda} \cdot \boldsymbol{f} . \tag{3.4}$$

These Lagrange multipliers play an important role in this approach. The reason for this is that a Lagrange multiplier, λ^k, is zero, if and only if, the corresponding constraint $E_q\{F_k\} = f_k$ contains no new information with respect to the prior distribution. Thus the Lagrange multipliers characterize the *mismatch* between the prior and the data, and as such are the natural parameters for this modeling problem.

The minimal discrimination information, (3.4), can be used for clustering points in the feature space described by the observations: $\boldsymbol{f} \equiv (f_1 = E\{F_1\}, \ldots, f_K = E\{F_K\})$. Given the prior distribution p^0, equations (3.2-3) can be viewed as a one-to-one mapping from the observable feature space onto the space of probability distributions with minimal discrimination information of the form (3.2). Equivalently, this is a one-one mapping from the observations vector space $\boldsymbol{f} \equiv (f_1, \ldots, f_K)$ onto the *dual* space of Lagrange multipliers $\boldsymbol{\lambda} \equiv (\lambda^1, \ldots, \lambda^K)$, which is the parameter space of these distributions. The distance between points in the observations space is thus naturally defined as the discrimination information between the corresponding distributions

$$\begin{aligned} d(\boldsymbol{f}', \boldsymbol{f}) &\equiv D(\hat{q}(\boldsymbol{f}') \mid \hat{q}(\boldsymbol{f})) = \int d\mathbf{x}\, \hat{q}'(\mathbf{x}) \log \frac{\hat{q}'(\mathbf{x})}{\hat{q}(\mathbf{x})} \\ &= (\lambda^0 - \lambda^{0\prime}) + \sum_{k=1}^{K} (\lambda^k - \lambda^{k\prime}) f_k \\ &\equiv \Delta\lambda^0 + \Delta\boldsymbol{\lambda} \cdot \boldsymbol{f} \\ &= D(\hat{q}' \mid p^0) - D(\hat{q} \mid p^0) - \sum_{k=1}^{K} \lambda^k (f_k - f'_k) \\ &\equiv -\Delta D_{\min} - \boldsymbol{\lambda} \cdot \Delta\boldsymbol{f} . \end{aligned} \tag{3.5}$$

Although this measure is asymmetric and highly non-linear, the clustering procedure is simple.[1] Finding the cluster centroid, i.e. the point $\boldsymbol{f}_c$ minimizing the average distortion to the N cluster members

$$\frac{1}{N} \sum_{i=1}^{N} d(\boldsymbol{f}_i, \boldsymbol{f}_c) = \frac{1}{N} \sum_{i=1}^{N} D(\hat{q}_i \mid \hat{q}_c) \tag{3.6}$$

is given by the (vector) arithmetic mean of the cluster members, by using equation (3.5)

$$\boldsymbol{f}_c = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{f}_i \ . \tag{3.7}$$

Hence, providing the Lagrange multipliers can be calculated in order to determine the cluster members, the clustering procedure using MDI is simple and practical. Note that averaging the observations is by no means equivalent to averaging the corresponding Lagrange multipliers, since the relation between the two is nonlinear.

4. Speech modeling with prior information

The principles described in the previous sections, reduce any statistical modeling problem to the following two questions:

1. What is the correct prior distribution?
2. What are the 'relevant' observables for the problem?

The simplest set of assumptions required for a normalizable prior distribution on the speech samples is to assume their first two moments. Assuming that the samples have zero mean and a constant variance, which in the absence of any other scale, can be taken as unity, determine (unsurprisingly) the standard white Gaussian noise as the simplest prior distribution

$$p^0(\mathbf{y}) = (2\pi)^{-n/2}\, e^{-\frac{1}{2}\tilde{\mathbf{y}} \boldsymbol{I}_n \mathbf{y}} , \tag{4.1}$$

where $\tilde{}$ denotes transpose and $\boldsymbol{I}_n$ is the unit matrix.

4.1 Memoryless models

We first consider observables that are single frame functions of the form $F^{(1)}(\mathbf{y}_t)$. Such constraints give rise to memoryless models, since there is no information which correlates different frames. Following the traditional work in speech recognition, let the single frame observables be the first $p+1$ ($p \simeq 8$) autocorrelations of the samples in each frame

$$\boldsymbol{F}_t^{(1)} \equiv \{ R_t(i) = \frac{1}{n-i} \sum_{j=1}^{n-i} y_t(j) y_t(j+i),\ \ 0 \le i \le p \ \} , \tag{4.2}$$

where $t = 1, 2, \ldots, T$ is the index of the frame and n is the number of samples in each frame. The autocorrelation vector is actually one row of the autocorrelation matrix, (or the covariance matrix) $\boldsymbol{R} \equiv E\{\mathbf{y}\,\tilde{\mathbf{y}}\}$, and is, again, a set of second moments of the signal $\{y_t(i)\}$. As just noted, the distribution with minimal discrimination information subject to the second moments as constraints is a multivariate Gaussian distribution of the form

$$\hat{b}(\mathbf{y}) = (2\pi)^{-n/2} (\det \boldsymbol{R})^{-1/2} \exp\left[-\tfrac{1}{2}\, \tilde{\mathbf{y}}\, \boldsymbol{R}^{-1}\, \mathbf{y}\right] . \tag{4.3}$$

By comparing (4.3) with the general form (3.2), and factoring out the prior distribution (4.1), we identify the Lagrange multipliers for this special case as

$$\begin{aligned} \lambda^{ij} &= \tfrac{1}{2}(\boldsymbol{R}^{-1} - \boldsymbol{I})_{ij} \\ \lambda^0 &= \tfrac{1}{2} \log \det \boldsymbol{R} . \end{aligned} \tag{4.4}$$

The matrix λ^{ij} is the matrix of Lagrange multipliers attached to the constraints matrix $\boldsymbol{R}_{ij}$, and obviously has the same symmetry as the matrix $\boldsymbol{R}^{ij}$ (Toepliz, circulant, etc.).[1] The additional (symmetry) assumption of time translational invariance (or stationarity), restricts the general Gaussian form to the commonly used Gaussian Autoregressive

(AR) source, and we are back to the standard type of speech modeling.

The Lagrange multipliers given by (4.4) are attached to every frame of the training data. By first clustering the data, using VQ or any similar clustering method, into N classes or "states" denoted by $s_1,...,s_N$, we can reduce the number of such multipliers to be proportional to the number of states N and not the size of the training data, T, which is normally much greater than N. Using such state quantization, the likelihood of the data is evaluated by summing over all possible state sequences, assuming that the states are unknown or "hidden"

$$P(\mathbf{y}_1, \ldots, \mathbf{y}_T) = \sum_{\{s(1),\ldots,s(T)\}} \prod_{t=0}^{T} \hat{b}_{s(t)}(\mathbf{y}_t) \ . \tag{4.5}$$

It turns out that in many applications the most probable state sequence gives the main contribution to the above summation, and we can replace the summation by a single term using the most likely state sequence.[9]

4.2 Hidden Markov models

The modeling described above is memoryless, i.e. there is no time order among the frames and no link between the frames. We can refer to it as a 'vector quantization' approach. This picture changes however, if two-frame observables are also observed. Let $F^{(2)}(\mathbf{y}_t, \mathbf{y}_{t'})$ be such a two frames observable, with its expectation value as constraints

$$f(t,t') = E_q\{F^{(2)}(\mathbf{y}_t, \mathbf{y}_{t'})\} \ . \tag{4.6}$$

The general prescription described in section 1, gives an additional factor to the distribution function, of the form

$$a'_{\pi'} \equiv \exp[-\eta^{\pi'} F^{(2)}(\mathbf{y}_t, \mathbf{y}_{t'})] \ , \tag{4.7}$$

where $\eta^{\pi'}$ is a Lagrange multiplier. With these additional terms, the likelihood of the data is written as (denoting by η^0 the normalization multiplier)

$$P(\mathbf{y}_1, \ldots, \mathbf{y}_T) = e^{-\eta^0} \prod_{t=0}^{T} b_t(\mathbf{y}_t) \prod_{t,t'=1}^{T} a'_{tt'} \ . \tag{4.8}$$

If we restrict the two-frame observations to adjacent frames, remembering that absence of a constraint yields a vanishing multiplier, we have : $\eta^{\pi'} = \eta^{\pi'} \delta_{t',t+1}$. We thus obtain a first order Markov process, where the $\{a_{t,t+1} \equiv e^{-\eta^0/T} a'_{t,t+1}\}$ are the *frame* transition probabilities.

If the frames are quantized into states, there is an additional variable (which may be hidden) for each frame: its state. The observables may, or may not, depend on the state of the frame. If we choose the adjacent-frames observables to depend only on the states and not on the actual data, we get the conventional 'hidden Markov model' scheme, and the $a_{t,t+1}$ are simply the state transition probabilities.

We can now consider the more general case, where the prior distribution is given by a Gaussian mixture AR hidden Markov model.[6] The rigorous formulation of this case is discussed by Ephraim *et al.*[2] The model is specified by the standard set of parameters: $\Lambda^0 \equiv (\boldsymbol{\pi}^0, \mathbf{A} \equiv \{a^0_{s_t,s_{t+1}}\}, \mathbf{B} \equiv \{b^0_{s_t}(\mathbf{y}_t)\})$,[7]

$$p^0 \equiv p_{\Lambda_0}(\mathbf{Y}_T) = \sum_{\{s_t\}} \pi^0_{s_1} \prod_{t=0}^{T-1} a^0_{s_t,s_{t+1}} \, b^0_{s_t}(\mathbf{y}_t) \ . \tag{4.9}$$

As a result of using MDI and autocorrelations, the distributions in **B** are of the form (4.3) and are parameterized by inverse covariance matrices, per mixture, denoted by $\{\mathbf{S}^{-1}_{s_t,m}\}$. The a-posteriori distribution is determined by single and double frame observations, with the single frame observations being the autocorrelations given by (4.2), and having the form:

1. The matrix notation for the Lagrange multipliers is appropriate, since the the constraints are given in a matrix form. For simplicity, the superscript indices are used for inverse matrices.

$$q_\Lambda(\mathbf{Y}_T) = e^{-\lambda^0} [\sum_{\{s_t\}} \pi^0_{s_1} \prod_{t=1}^{T} a^0_{s_{t-1},s_t} \, b^0_{s_t}(\mathbf{y}_t)] \, e^{-\sum_{t=0}^{T} \boldsymbol{\lambda}^{s_t} \mathbf{F}^{(1)}(\mathbf{y}_t) - \sum_{t=1}^{T} \boldsymbol{\eta}^{s_{t-1},s_t} \mathbf{F}^{(2)}(\mathbf{y}_{t-1},\mathbf{y}_t)}$$

$$= e^{-\lambda^0} \sum_{\{s_t\}} \pi^0_{s_1} \prod_{t=1}^{T} a^0_{s_{t-1},s_t} \, e^{-\boldsymbol{\eta}^{s_{t-1},s_t} \mathbf{F}^{(2)}(\mathbf{y}_{t-1},\mathbf{y}_t)} \, b^0_{s_t}(\mathbf{y}_t) \, e^{-\tilde{\mathbf{y}}_t \boldsymbol{\lambda}^{s_t} \mathbf{y}_t} \tag{4.10}$$

$$= \sum_{\{s_t\}} \pi^0_{s_1} \prod_{t=1}^{T} a^0_{s_{t-1},s_t} \, e^{-\boldsymbol{\eta}^{s_{t-1},s_t} \mathbf{F}^{(2)}(\mathbf{y}_{t-1},\mathbf{y}_t)} \, e^{-\lambda^0_{s_t}} \, e^{-\frac{1}{2}\tilde{\mathbf{y}}_t (\mathbf{S}^{-1}_{s_t} + 2\boldsymbol{\lambda}^{s_t}) \mathbf{y}_t} \ .$$

We used the fact that the Lagrange multipliers, defined this way, are independent of the state or mixture of the prior model and instead are functions of the frame.

4.3 Averaging the Lagrange multipliers using the prior model

In order to eliminate the frame dependency of the Lagrange multipliers in (4.10), two approximations are needed. To avoid mixed states terms in (4.10), we use the most probable state sequence via the Viterbi decoding[8] of the data, *with the prior model*. As mentioned earlier, in many practical cases the Viterbi decoding gives the same results as the full Baum-Welch[7] calculation. By doing this, we simply replace the index t of the Lagrange multipliers in the last line of (4.10) with the state (and mixture) index s_t, yielding

$$q_\Lambda \approx \pi^0_{s_1} \prod_{t=1}^{T} a^0_{s_{t-1},s_t} \, e^{-\boldsymbol{\eta}^{s_{t-1},s_t} \mathbf{F}^{(2)}(\mathbf{y}_{t-1},\mathbf{y}_t)} \, e^{-\lambda^0_{s_t}} \, e^{-\frac{1}{2}\tilde{\mathbf{y}}_t (\mathbf{S}^{-1}_{s_t} + 2\boldsymbol{\lambda}^{s_t}) \mathbf{y}_t} \ . \tag{4.11}$$

This is a modified a-posteriori model, where the modified parameters are determined by the Lagrange multipliers

$$\begin{cases} \hat{\mathbf{S}}^{-1}_{s_i} = \mathbf{S}^{-1}_{s_i} + 2\boldsymbol{\lambda}^{s_i} \\ \hat{a}_{s_i,s_j} = a^0_{s_i,s_j} \, e^{-\boldsymbol{\eta}^{s_i,s_j} \bar{\mathbf{F}}^{(2)}(s_i,s_j)} \end{cases} \ . \tag{4.12}$$

Using the MDI clustering procedure, the Lagrange multipliers are determined by the (state) average constraints:

$$\begin{aligned} \bar{\mathbf{F}}^{(1)}(s_i) &\equiv \sum_{t=0} p^0_\Lambda(s_t = s_i) \mathbf{F}^{(1)}(\mathbf{y}_t) \\ \bar{\mathbf{F}}^{(2)}(s_i,s_j) &\equiv \sum_{t=1} p^0_\Lambda(s_{t-1} = s_i, s_t = s_j) \mathbf{F}^{(2)}(\mathbf{y}_{t-1},\mathbf{y}_t) \ . \end{aligned} \tag{4.13}$$

Note that the last relation can be made into an iterative equation by replacing p_Λ with the posterior distribution q_Λ, and iterate alternately q_Λ and the Lagrange multipliers, until convergence.

Finally, we can use these relations to evaluate explicitly the MDI distance for the simple Gaussian case with only single frame observations. If the experimental correlation matrix for the frame at time t is denoted by R_t, we first average the correlation matrices to get the state centroid $\bar{R}_{ij}$. The equation for the Lagrange multipliers, for each state, are then simply

$$\bar{R}_{ij} = (\mathbf{S}^{ij} + 2\boldsymbol{\lambda}^{ij})^{-1} \ . \tag{4.14}$$

Thus, in this simple case, the single frame Lagrange multipliers matrix is just the difference between the observed state (mixture) inverse autocorrelation matrices and the prior corresponding autocorrelation matrix

$$\boldsymbol{\lambda}^{ij} = \tfrac{1}{2}(\bar{R}^{ij} - \mathbf{S}^{ij}) \ . \tag{4.14a}$$

The discrimination information between two sets of observations (2 speakers in our case) is given by

$$D(q^1 \mid q^2) = \int q^1 \log \frac{q^1}{q^2} = \int q_1 (\log \frac{q_1}{p^0} - \log \frac{q_2}{p^0})$$

$$= (\lambda_2^0 - \lambda_1^0) + \int q_1 \tilde{\mathbf{Y}}_T (\boldsymbol{\lambda}^2 - \boldsymbol{\lambda}^1) \mathbf{Y}_T \qquad (4.15)$$

$$= \Delta\lambda^0 + \sum_{t=0}^{T} Tr(\Delta\boldsymbol{\lambda}^t R_t^1) ,$$

with

$$\Delta\lambda^0 = \sum_{t=0}^{T} \log \frac{\det \boldsymbol{R}_t^2}{\det \boldsymbol{R}_t^1} .$$

This results reduces to the expression obtained in Ref. 4 for the MDI, if $q_2 = p^0$, as it should. Note that in this simple case the distance is only weakly dependent on the prior distribution p_Λ^0 and is close to the result obtained with the standard likelihood ratio distortion measure. The prior model enters, however, through the initial clustering both in the training and in the recognition. The main objective of separating the prior dependent parameters was thus achieved. Much more is gained, if the spectral resolution of the a-posteriori model is different than that of the prior, or if the observables used for the a-posteriori model are different.

5. Application to speaker verification

As a test of the above ideas, a simple speaker verification experiment was carried out. We used a data base consisting of 20,000 isolated digits utterances, spoken by 100 speakers, 50 male and 50 female, over dialed-up local telephone lines. The experimental setup is the same used in previous speaker verification experiments.[9] [10] Since it has been shown that for text independent speaker recognition the memory terms (state transitions) are not very important,[10] only single frame autocorrelations are used as observables for this simple experiment.

5.1 The prior models

Using previously trained individual AR hidden Markov models,[10] with 8 states and 8 mixtures each, two prior models were created. The first was constructed by clustering the spectral density vectors for 30 male speakers, using the data base of isolated English digits. The second was a similar model for 30 female speakers. In order to check the validity of these priors, a simple 'sex identification' experiment on the (40) speakers outside of the training set, was performed. with equal error rates that are given in table I. The results show that the sex of the speaker was identified correctly for about 90% of the speakers based on 5 spoken digits.

Table I: The prior's sex identification error rates (%).

test length	'male' prior	'female' prior
1	26.	22.
2	20.	17.
3	16.	14.
4	12.	11.
5	11.	9.

In order to train the speaker dependent models, the simplest method of clustering the individual speaker data by using the Viterbi decoding with the prior, was used. The frames of each speaker were grouped according to the nearest prior state, with the likelihood ratio (which is just the MDI distance in this case) distortion measure. The autocorrelation matrices were averaged within each state, and a single Lagrange multiplier matrix per mixture was calculated, using eq (4.14). This set of Lagrange multipliers, together with the corresponding prior model, now serves as the speaker trained model.

A speaker verification experiment, similar to the experiment reported previously,[10] on 20 speakers (10 males, 10 females) was carried out, using the distance (4.15). Table II gives the verification equal error rates (mean, median, 10 and 90 percentiles) for this experiment. The results are comparable or better than those obtained by using fully speaker trained models, which we find encouraging.

Table II: Verification equal error rates (%):

test length	mean	10%	median	90%
1	8.2	2.1	7.2	11.5
2	3.9	0.6	2.4	6.3
3	2.6	0.0	1.5	4.5
4	2.1	0.0	0.2	2.6
5	1.8	0.0	0.0	3.1
6	1.2	0.0	0.0	2.4
7	0.8	0.0	0.0	1.7
8	0.6	0.0	0.0	1.2
9	0.4	0.0	0.0	1.1
10	0.4	0.0	0.0	0.3

6. Conclusions

The separation of speaker and language in hidden Markov models is discussed as an example of speech modeling with prior information. The minimal discrimination information serves as a distance between speakers, with a common 'language' prior. The Lagrange multipliers, obtained by the minimization of the discrimination information, play the role of the models parameters, and measure the statistical deviations of the data from the prior distribution. These parameters can be clustered or averaged based on the prior's decoding. The method may have various applications in speech recognition and modeling problems, and can be used with other possible constraints e.g. cepstral coefficients. Simple speaker verification experiment demonstrates the possible utility of this approach.

Acknowledgement

The author is grateful to Larry Rabiner for his encouragement, and wishs to acknowledge useful discussions with Amir Dembo and Yariv Ephraim.

REFERENCES

1. J.E. Shore and R.M. Gray, *Minimum Cross-entropy Pattern Classification and Cluster Analysis*, ITE Trans. Pattern Analysis and Machine Intelligence, Vol. PAMI-4, No. 1, 11,17 (1982).
2. Y. Ephraim, A. Dembo, and L. R. Rabiner, *A minimum discrimination information approach for hidden Markov modeling*, ICASSP 87, IEEE Intl. Conf. on Acoust., Speech, and Signal Processing, (1987) pp. 25-28.
3. Y. Tikochinsky, N. Tishby and R.D. Levine, *Alternative approach to maximum-entropy inference*, Phys. Rev. A, Vol 30, pp. 2638-2644 (1984).
4. E. T. Jaynes, Phys. Rev. **106**, pp. 620 (1957).
5. S. Kullback, *Information Theory and Statistics*. New York: Dover, (1969) Ch. 1-2.
6. B.-H. Juang and L. R. Rabiner, *Mixture Autoregressive Hidden Markov Models for Speech Signals* IEEE Trans. Acoust. Speech, and Signal Processing, **ASSP-33** 6,pp. 1404,(1985).
7. L. R. Rabiner, *"The Theory of Hidden Markov Models and its Application to Speech Recognition"*, Submitted for publication.
8. G. D. Fomey, *The Viterbi Algorithm*, Proc. IEEE, Vol. 16, pp. 268-278,(1973).
9. F. K. Soong, A. E. Rosenberg, L. R. Rabiner, and B. H. Juang, *"A vector quantization approach to speaker recognition"*, Proc. ICASSP 85, IEEE Intl. Conf. on Acoust., Speech, and Signal Processing, (1985) pp. 387-390. AT&T Technical Journal, Vol. 66, 2, (1987) pp. 14-26 .
10. N. Tishby, *"On the application of AR hidden Markov models to text independent speaker recognition."* Submitted for publication.