IDIAP RESEARCH REPORT

Robust Speech Recognition and Feature Extraction Using HMM2

Katrin Weber[1,2] Shajith Ikbal[1,2]
Samy Bengio[1] Hervé Bourlard[1,2]

IDIAP-RR 01-42

November 2001, revised version November 2002

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 - 27 - 721 77 11
fax +41 - 27 - 721 77 12
email secretariat@idiap.ch
internet http://www.idiap.ch

1. Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP), Martigny, Switzerland
2. Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland

IDIAP-RR 01-42

Robust Speech Recognition and Feature Extraction Using HMM2

Katrin Weber, Shajith Ikbal, Samy Bengio, and Hervé Bourlard

November 2001, revised version November 2002

Abstract: This paper presents the theoretical basis and preliminary experimental results of a new HMM model, referred to as HMM2, which can be considered as a mixture of HMMs. In this new model, the emission probabilities of the temporal (primary) HMM are estimated through secondary, state specific, HMMs working in the acoustic feature space. Thus, while the primary HMM is performing the usual time warping and integration, the secondary HMMs are responsible for extracting/modeling the possible feature dependencies, while performing frequency warping and integration. Such a model has several potential advantages, such as a more flexible modeling of the time/frequency structure of the speech signal. When working with spectral features, such a system can also perform nonlinear spectral warping, effectively implementing a form of nonlinear vocal tract normalization. Furthermore, it will be shown that HMM2 can be used to extract noise robust features, supposed to be related to formant regions, which can be used as extra features for traditional HMM recognizers to improve their performance. These issues are evaluated in the present paper, and different experimental results are reported on the Numbers95 database.

Acknowledgements: This work was partly supported by grants FN 2000-059169.99/1 and FN 2100-061325.00/1 from the Swiss National Science Foundation. We would also like to thank the SPHEAR project partner Daimler-Chrysler for providing us with their car noise.

1 INTRODUCTION

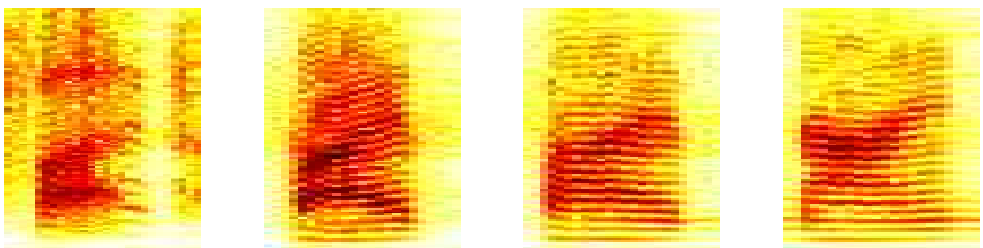
In state-of-the-art automatic speech recognition (ASR), hidden Markov models (HMMs) are widely used. While there are many suitable alternatives and design options for some parts of ASR systems such as feature extraction and phoneme probability estimation, HMMs are the uncontested model for the temporal decoding stage. The success of HMMs can (at least partly) be contributed to their ability to easily accommodate temporal variations, such as different durations of phonemes, e.g. due to varying speaking rate or speakers' accents.

However, such variations do not only occur along the time axis, but can also be observed in frequency, as shown in Figure 1. In the spectograms depicting four different pronunciations of phoneme 'ay' (including some context), inter- as well as intra-speaker variability becomes apparent (compare Figure 1a with 1b, and Figure 1b with 1c respectively). Furthermore, Figure 1d shows the same phoneme pronounced in a different context, revealing the effects of coarticulation. All sub-figures suggest that the position of spectral peaks may change significantly in the time-frequency plane during the pronunciation of a phoneme.

When using HMMs, however, it is assumed that speech segments corresponding to one phoneme or subphone unit are (1) invariant enough to be modeled by the same static distribution and (2) stationary for their duration, which clearly is not the case. In an attempt to relax these rather rigid assumptions, and encouraged by many more practical motivations (as further elaborated in Section 2.2), we recently proposed the HMM2 approach (Weber, Bengio & Bourlard, 2000). HMM2 can be understood as an HMM mixture consisting of a primary HMM, modeling the temporal properties of the speech signal, and a secondary HMM, modeling the speech signal's frequency properties. A secondary HMM is in fact inserted at the level of each state of the primary HMM, estimating local emission probabilities of acoustic feature vectors (conventionally done by Gaussian mixture models (GMM) or artificial neural networks (ANN)). Consequently, an acoustic feature vector is considered as a fixed length sequence of its components, which has supposedly been generated by the secondary HMM.

Although HMM2 was developed independently, a similar approach had already been proposed and used with some success in computer vision (Levin & Pieraccini, 1993; Kuo & Agazzi, 1993; Samaria, 1994; Eickeler, Müller & Rigoll, 1999). However, as further discussed below, our approach includes full EM training and was extended to take care of specificities of the problem at hand.

The purpose of this paper is to revise theoretical and practical aspects of the HMM2 approach with regard to its application to speech recognition. Firstly, a brief description of HMM2 is given and motivations for applying it to speech recognition are outlined. This is followed by the HMM2 theory, including training and decoding, as well as some notes on its practical implementation. A thorough analysis of HMM2, its possible drawbacks and constraints is then given. Finally, the applications of HMM2 in the domain of speech, namely as a speech decoder and feature extractor, are investigated. Encouraging results for all these cases are given.



(a) Speaker 1: 'five' *(b) Speaker 2: 'five' (1st occurrence)* *(c) Speaker 2: 'five' (2nd occurrence)* *(d) Speaker 2: 'nine'*

Figure 1: *Spectograms of different pronunciations of the phoneme 'ay' by different speakers and in different contexts. Dark regions correspond to high, light regions to low energy spectral components. The vertical axis is the frequency, the horizontal one the time evolution.*

2 HMM2 DESCRIPTION AND MOTIVATIONS

2.1 Description

HMMs are quite powerful statistical models which are used to represent sequential data, e.g. a sequence of T acoustic vectors $y_1^T = \{y_1, y_2, \ldots, y_t, \ldots, y_T\}$ in speech recognition (as shown in the upper part of Figure 2). As each acoustic vector y_t can itself be considered as a fixed length sequence of its S components $y_t = y_{t(1,S)} = \{y_{t(1)}, y_{t(2)}, \ldots, y_{t(s)}, \ldots, y_{t(S)}\}$, another HMM can be used to model this feature sequence (displayed in the lower part of the figure). By 'component' we mean a subvector of low dimension. For instance, a temporal feature vector of dimension S is split up into S 1-dimensional subvectors. For the sake of simplicity, this case is assumed for all theoretical derivations throughout this paper. However, the extension of this case to higher-dimensional subvectors (consisting, e.g., of a coefficient and its first and second time derivatives) is straightforward, and was in fact used for the practical experiments.

While the primary HMM models temporal properties of the speech signal, the secondary, state-dependent HMM is working along the frequency dimension. The secondary HMM is in fact acting as a likelihood estimator for the primary HMM, a function which is accomplished by GMMs or ANNs in conventional systems. However, the state emission distributions of the secondary HMM are again modeled by GMMs. Consequently, HMM2 is a generalization of the standard HMM/GMM system, which it includes as a particular case. In fact, a standard HMM can be realized with HMM2 by choosing a particular topology, e.g. the secondary HMM consists just of one state and emits the entire temporal feature vector at once. An alternative way of realizing a conventional HMM within the HMM2 framework is described in (Weber, Bengio & Bourlard, 2001). The parameters of an HMM2 are the primary transition probabilities $P(q_t|q_{t-1})$, the secondary transition probabilities $P(r_s|r_{s-1}, q_t)$, and the emission probabilities $p(y_{t(s)}|r_s, q_t)$.

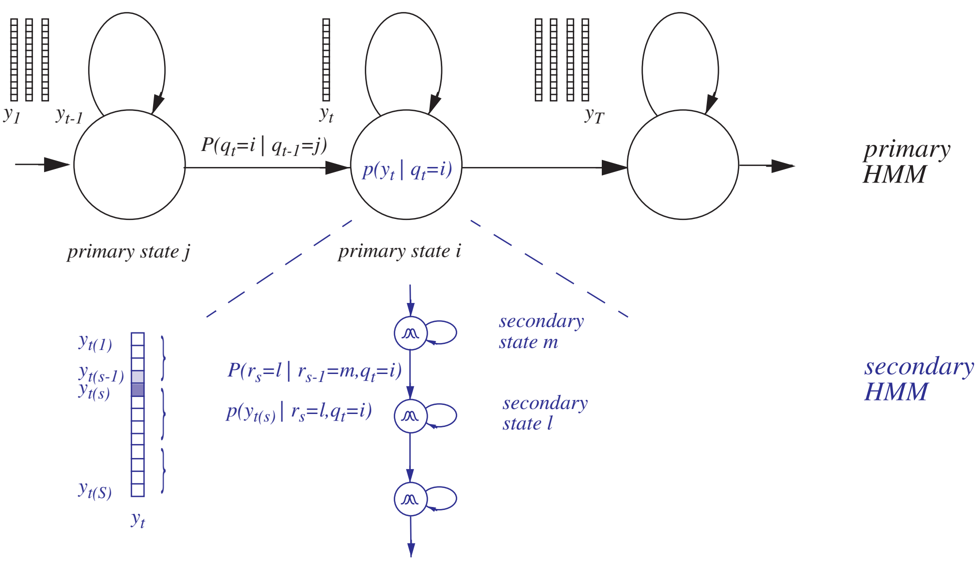


Figure 2: *HMM2 system. In the upper part, a conventional HMM, working along the temporal axis, can be seen. The local emission probability calculation is done with a secondary HMM, working along the frequency axis (depicted in the lower part of the figure).*

2.2 Motivations

a) Better and more flexible modeling

HMMs assume piecewise stationarity of the speech signal and have difficulty in modeling the dynamic properties along the feature (frequency) dimension. Using a secondary HMM for the local likelihood estimation, these assumptions are relaxed, as a more flexible modeling of the variability and dynamics inherent in the speech signal is allowed. For instance, a spectral peak could be modeled by a single state of the secondary HMM, even though its position on the frequency axis is quite variable (as seen in Figure 1). Such a sparse secondary HMM topology also allows for efficient parameter sharing. The number of parameters can be easily controlled by the model topology and the probability density function associated with the secondary HMM states.

b) Modeling of correlation through secondary HMM topology

Under the typical HMM assumptions, correlation between feature vector components is not ignored, but supposed to be modeled through the topology of the secondary HMM. Thereby, correlation of close feature vector components is emphasized in comparison to distant correlation, which corresponds to the properties of data we aim to model. In fact, HMM2 could allow a sophisticated modeling of the underlying time-frequency structures of the speech signal and model complex constraints in both the temporal and the frequency dimensions. In the same spirit, it was proposed in (Bilmes, 1999) to model time/frequency correlation in the framework of buried Markov models by using Bayesian networks to compute emission probabilities, where the connectivity of the Bayesian network was determined by the degree of mutual information between coefficients.

c) Non-linear, state dependent spectral warping

The secondary HMM automatically performs a non-linear, state dependent spectral warping. While the conventional HMM does time warping and time integration, the secondary HMM performs warping and integration along the frequency axis. This frequency warping has the effect of automatic non-linear vocal tract normalization, providing a kind of unsupervised and implicit speaker adaptation (therefore tackling the problem of inter-speaker variations). Applying HMM2 in this field is also encouraged by the work of Lee & Rose (1998), who use a related frequency warping approach to speaker normalization. With the same mechanism, intra-speaker variations as well as coarticulation effects are also taken care of.

Furthermore, it can be expected that HMM2 will perform a dynamic formant trajectories tracking. As a spectral peak (formant) can be modeled by an HMM state and a spectral valley by another, the segmentation performed by the secondary HMM may be a good indicator for the position of a formant. Formants are assumed to carry discriminant information in the speech signal, moreover being especially robust in the case of degraded speech (Garner & Holmes, 1998, Welling & Ney, 1998).

d) Extension of multi-band processing

Currently, much research in speech recognition is being devoted to multi-band speech recognition (Morris, Hagen, Glotin & Bourlard, 2001). In this case, the full frequency band is split into multiple subbands which are processed independently (to a certain extent) by different classifiers before recombining the resulting probabilities to yield the fullband phonetic probabilities. More recently, this multi-band ASR approach was extended by using the so called 'full combination approach' in which subband probability combination is performed by integrating over all possible reliable subband combinations. HMM2 can be seen as a further extension to this approach. Indeed, all possible paths through the secondary HMM will correspond to different subband segmentations and recombinations. The frequency position of the subbands is then automatically adapted to the data, following for example formant-like structures.

3 HMM2 THEORY AND IMPLEMENTATION

This section gives a detailed description of the HMM2 approach, including training, decoding and implementation. As stated before, although HMM2 was proposed independently and with an entirely different motivation, it is related to similar approaches used previously for computer vision, such as Planar HMMs (Levin & Pieraccini, 1993) and Pseudo 2D HMMs (Kuo & Agazzi, 1993; Samaria, 1994; Eickeler, Müller & Rigoll, 1999). However, while these models are trained using either a planar segmentation algorithm based on Viterbi (Levin & Pieraccini, 1993), a segmental k-means algorithm (Kuo & Agazzi, 1993), or (after the two-dimensional model has been converted to a similar one-dimensional HMM) with conventional EM training (Samaria, 1994; Eickeler, Müller & Rigoll, 1999), we here develop an EM algorithm which is especially adapted to HMM2.

3.1 Notation

Basic notations used throughout this paper are explained in Figure 2, with the following definitions:

- y_t the observed vector at time step t, and $y_{t(s)}$ its observed component at frequency step s,
- q_t the primary HMM state at time t, where Q is the set of all possible paths through the primary HMM, and $r_s(q_t)$ the secondary state associated with primary state q_t at frequency step s, where $R(Q)$ is the set of all possible paths through the secondary HMMs,
- $p(y_t|q_t)$ the emission probability in the primary HMM, where the instantiation $p(y_t|q_t = i)$ is the probability to emit y_t in state i, and $p(y_{t(s)}|r_s, q_t)$ the emission probability in the secondary HMM, where the instantiation $p(y_{t(s)}|r_s = l, q_t = i)$ is the probability to emit component $y_{t(s)}$ in secondary state l of primary state i,
- $P(q_0)$ the initial state probability of the primary HMM, and $P(r_0|q_t)$ the initial state probability of the secondary HMM in primary HMM state q_t,
- $P(q_t|q_{t-1})$ the state transition probabilities in the primary HMM, where the instantiation $P(q_t = i|q_{t-1} = j)$ is the probability to go from primary state j at time $t-1$ to state i at time t, and $P(r_s|r_{s-1}, q_t)$ the state transition probabilities in the secondary HMM in primary state q_t, where the instantiation $P(r_s = l|r_{s-1} = m, q_t = i)$ is the probability to go from secondary state m at the frequency step $s-1$ to secondary state l at frequency step s while in primary state i at time t,
- N the number of states in the primary HMM, and N_i the number of states of the secondary HMM in the primary HMM state i,
- T the size of the sequence $y_1^T = \{y_1, y_2, \ldots, y_T\}$, and S the size of the sequence of components $y_{t(1,S)} = \{y_{t(1)}, y_{t(2)}, \ldots, y_{t(S)}\}$.

The likelihood of the data sequence Y given the model parameters θ at training step k is then

$$L(Y|\theta) = p\left(y_1^T \middle| \theta^k\right). \tag{1}$$

3.2 Training

Since an HMM is a special kind of mixture of distributions, an HMM2, being a mixture of HMMs, can therefore also be considered as a more general mixture of distributions. It should hence be natural that an Expectation-Maximization (EM) algorithm could be derived when the emission and transition probabilities of the secondary (feature-based) HMMs are represented by mixtures of Gaussians and multinomials respectively. In this section, a sketch of such a derivation is given, which is detailed in (Bengio, Bourlard & Weber, 2000).

The general idea of EM is to select a set of hidden variables such that the knowledge of these variables would simplify the learning problem. Hence, in the estimation step, the value of the hidden variables is estimated, while in the maximization step, the expectation of the log likelihood of the observations and the hidden variables is maximized, given the previous values of the parameters. This two-step process is repeated iteratively and is proved to converge to a local optimum of the likelihood of the observation (Dempster, Laird & Rubin, 1977).

In the case of HMM2, two sets of indicator variables $Z = \{z_{i,t}\}$ and $W = \{w_{i,t(l,s)}\}$ are defined such that $z_{i,t}$ is defined to be 1 when $q_t = i$ and 0 otherwise, and $w_{i,t(l,s)}$ is defined to be 1 when $q_t = i$ and $r_s = l$, and 0 otherwise. The joint likelihood of the observations and the hidden variables is then defined as:

$$L(Y, Q, R(Q)) = P(q_0)\prod_{t=1}^{T}\prod_{i=1}^{N}\left[p(y_t|q_t = i)^{z_{i,t}}\prod_{j=1}^{N}P(q_t = i|q_{t-1} = j)^{z_{i,t}\cdot z_{j,t-1}}\right] \tag{2}$$

which has the same form as the joint likelihood in standard HMMs, where the emission probability is expressed as:

$$p(y_t|q_t = i) = P(r_0|q_t = i)\prod_{s=1}^{S}\prod_{l=1}^{N_i}\left[p(y_{t(s)}|r_s = l, q_t = i)^{w_{i,t(l,s)}}\right. \tag{3}$$
$$\left.\prod_{m=1}^{N_i}P(r_s = l|r_{s-1} = m, q_t = i)^{w_{i,t(l,s)}\cdot w_{i,t(m,s-1)}}\right]$$

Having stated the problem, it is straightforward to derive both the E-step and the M-step of EM, which are very similar to the general EM formulation for HMMs. During the E-step, the expected value of the following variables are estimated:

$$\hat{\gamma}(i,t) = E[z_{i,t}|y_1^T;\theta^k] = P(q_t = i|y_1^T, \theta^k) \tag{4}$$

$$\hat{\xi}(i,j,t) = E[z_{i,t}, z_{j,t-1}|y_1^T;\theta^k] = P(q_t = i, q_{t-1} = j|y_1^T, \theta^k) \tag{5}$$

$$\hat{\gamma}_{i,t}(l,s) = E[w_{i,t(l,s)}|y_1^T;\theta^k] = P(r_s = l|q_t = i, y_1^T, \theta^k) \tag{6}$$

$$\hat{\xi}_{i,t}(l,m,s) = E[w_{i,t(l,s)}, w_{i,t(m,s-1)}|y_1^T;\theta^k] = P(r_s = l, r_{s-1} = m|q_t = i, y_1^T, \theta^k) \tag{7}$$

given the model parameters θ^k at the k-th EM iteration.

Finally, during the M-step, the values of the parameters θ^{k+1} that maximize the expectation of $\log L(Y, Q, R(Q))$ are found. The final HMM2 update equations are similar to the update equations used in normal HMMs except that all the posteriors computed in the secondary HMM are weighted by the posterior probability of being in the given state of the primary HMM. For instance, if the emission probability of primary state i and secondary state l is defined as a diagonal Gaussian with mean μ_{il} and standard deviation σ_{il}, the update equations are:

$$\mu_{il}^{k+1} = \frac{\sum_t\left[\hat{\gamma}(i,t)\sum_s y_{t,s}\cdot\hat{\gamma}_{i,t}(l,s)\right]}{\sum_t\left[\hat{\gamma}(i,t)\sum_s\hat{\gamma}_{i,t}(l,s)\right]} \tag{8}$$

and

$$\sigma_{il}^{k+1} = \sqrt{\frac{\sum_t\left[\hat{\gamma}(i,t)\sum_s\hat{\gamma}_{i,t}(l,s)\cdot(y_{t,s}-\mu_{il})^2\right]}{\sum_t\left[\hat{\gamma}(i,t)\sum_s\hat{\gamma}_{i,t}(l,s)\right]}}. \tag{9}$$

3.3 Decoding

The aim of HMM decoding is to find the sequence of HMM states which best explains the input data, while at the same time taking account of phonological, lexical and syntactical constraints. Therefore, under the typical HMM assumptions (i.e. piecewise stationarity and data independence), the recognized word sequence is defined by the path Q^* (from the set of all possible paths Q) which maximizes the joint likelihood of the data and the hidden variables, given the model parameters:

$$Q^* = \underset{Q}{\operatorname{argmax}}\left[P(q_0)\prod_{t=1}^{T}[p(y_t|q_t)P(q_t|q_{t-1})]\right]. \tag{10}$$

The likelihood of an acoustic feature vector (i.e., a sequence of its components) given the primary HMM state $p(y_t|q_t)$ may be calculated in two different ways:

$$p(y_t|q_t) = \sum_{R}\left[P(r_0|q_t)\prod_{s=1}^{S}[p(y_{t(s)}|r_s, q_t)P(r_s|r_{s-1}, q_t)]\right] \tag{11}$$

or, using the Viterbi approximation:

$$p(y_t|q_t) \cong \max_{R}\left[P(r_0|q_t)\prod_{s=1}^{S}[p(y_{t(s)}|r_s, q_t)P(r_s|r_{s-1}, q_t)]\right] \tag{12}$$

where R is the set of all possible paths through the model. Naturally, every term of this equation is conditioned on the state of the primary HMM. As GMMs with diagonal covariance matrices are used for the likelihood estimation in the states of the secondary HMM, the corresponding local probability density functions (PDF) are defined as follows:

$$p(y_{t(s)}|r_s = l, q_t = i) = \sum_{g=1}^{G} w_{ilg}\frac{1}{\sqrt{2\pi\sigma_{ilg}^2}}e^{-\frac{1}{2}\left(\frac{y_{t(s)}-\mu_{ilg}}{\sigma_{ilg}}\right)^2} \tag{13}$$

where G is the number of Gaussian mixtures.

3.4 Implementation

There are different ways to implement HMM2 systems. A straightforward realization is based on the implementation of a generalized form of the standard EM algorithm, as described in the previous section. A second way is to unfold the HMM2 (which, as previously stated, is a kind of HMM mixture) into one large HMM, as described before in (Samaria, 1994; Eickeler, Müller & Rigoll, 1999; Weber, Bengio & Bourlard, 2001a). For this implementation, synchronization constraints have to be introduced to ensure that exactly one feature vector is emitted between each two transitions in the primary HMM. Standard EM training algorithms can be used to implement this unfolded HMM2, and Viterbi decoding has to be used at the level of both the primary and the secondary HMM.

4 ANALYSIS

Hidden Markov models are a generalization of Gaussian mixture models (suitable for sequential data). Given a sufficiently large number of appropriately chosen parameters, these mixture models can approximate any continuous density to arbitrary accuracy (Bishop, 1995). Practically, however, there are limitations. The number of parameters in a mixture model has to be small enough to be reliably estimated from the given amount of training data. Furthermore, for the case of sequential data modeled by an HMM, there are additional assumptions, notably that of data independence (conditioned on the state) and of piecewise stationarity (Rabiner & Juang, 1993). Moreover, there may be constraints imposed by the HMM topology.

Naturally, in the case of HMM2, these assumptions and constraints not only apply to the primary, but also to the secondary HMM. As is generally the case for a temporal sequence of speech data, the assumptions of conditional data independence and of piecewise stationarity are also not entirely satisfied for the HMM2 feature vector sequence, which may result in a mismatch between the data and the model's capacity for data representation and discrimination (Weber, Bengio & Bourlard, 2001a). In the following, the implications of the stationarity assumption and of the topology chosen for the secondary HMM are investigated.

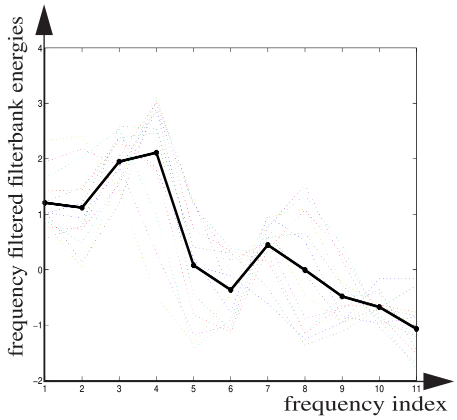


Figure 3: *Energy spectrum of a pronunciation of phoneme 'ay'. Each line in the figure corresponds to one time step, and thus to one feature vector (the thick black line is the mean).*

Figure 3 shows an energy spectrum of phoneme 'ay'. Although the assumption of piecewise stationarity is not entirely correct, it is possible to segment this representation along the (horizontal) frequency axis into a few quasi-stationary sectors, which could subsequently be represented by the same PDF. Consequently, several frequency components could be modeled by one secondary HMM state.

The secondary HMM may take any topology, the most general being ergodic. However, in this paper, the topology of the secondary HMM is chosen to be strictly top-down and to have fewer states than there are components in one temporal feature vector (as also seen in Figure 2). Therefore, each secondary HMM2 state is expected to emit a number of adjacent components, i.e. all components belonging to a certain frequency band. The number of secondary HMM states determines the number of frequency bands into which the spectrum is decomposed. The cut-off frequencies and bandwidths of these frequency bands will be dynamically determined, given the data and the model parameters, during training and decoding.

It is known that HMMs cannot provide good duration modeling. This disadvantage applies similarly to the modeling of bandwidth (and subsequently frequency positions) in the secondary HMM. However, as frequency positions of different spectral regions (especially formants) represent important discriminant acoustic cues, it has to be ensured that HMM2 takes them into account in an appropriate way. This problem can be resolved with an additional coefficient of the feature vector, which indicates the frequency position of its respective component, as shown in Figure 4a (Weber, Bengio & Bourlard, 2001b). This has the effect of forcing the Viterbi algorithm to take the frequency position of each feature vector into account during the frequency segmentation.

As an example, Figure 4 illustrates the typical spectral shape of two vowel classes α and β, both consisting of 2 alternating spectral peaks (H) and valleys (L), resulting in the overall structure HLHL. These classes can be distinguished only by the position of the spectral peaks and valleys, and it is known that these positions are indeed the most important perceptual cues. Using HMM2 without frequency coefficients, the only way of modeling the differences between α and β is by the transition probabilities, which, as stated previously, do not have much influence. The two classes are therefore easily confusable. When introducing the frequency coefficients, the Viterbi segmentation of a feature vector is in some way constrained and discriminability will be maintained. In fact, the frequency coefficient is handled in the same way as the other coefficients in a feature vector, i.e. it is modeled by the GMM. The Gaussian mean will correspond to the mean frequency of the modeled frequency band, and the variance should be an indicator of the bandwidth.

While the idea of using an additional frequency coefficient may seem surprising, it is justified in the frequency warping performed by HMM2. Improved recognition results confirm the suitability of this idea (Weber, Bengio & Bourlard, 2001b). Naturally, in standard HMMs this frequency coefficient does not give any additional information, as the frequency position of each coefficient is known.

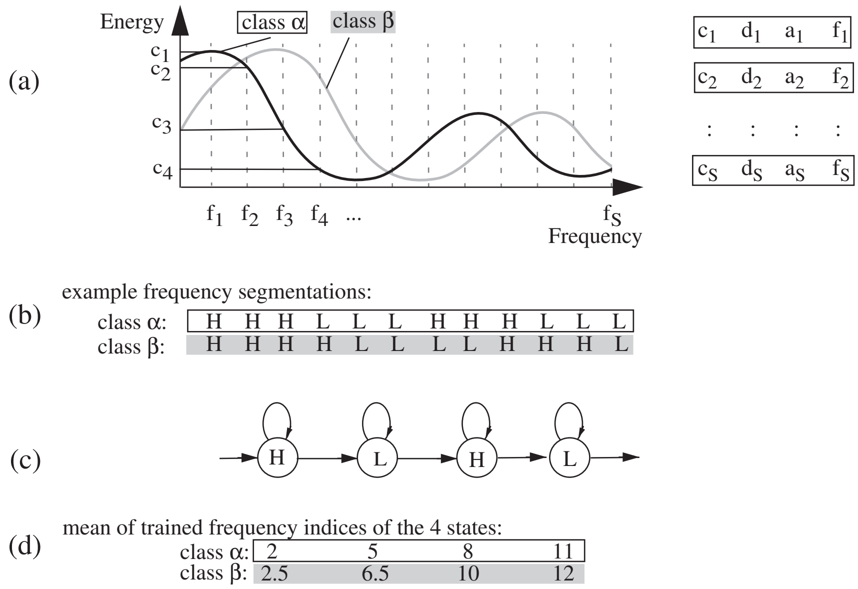


Figure 4: *The frequency index: In (a), data assumed to be typical of the classes α and β are visualized by a black and a gray curve respectively. On the right, feature vectors (corresponding to the class α curve) as used in the secondary HMM composed of coefficients c_s, their delta d_s and acceleration coefficients a_s, as well as the frequency coefficient f_s, are shown. In (b), an example frequency segmentation is shown for each class. (c) shows a structure of an HMM with alternating H and L states, which is able to model both classes. With an additional trained frequency coefficient (as shown in (d)), discriminability can be ensured.*

To investigate the meaning of such frequency information, an HMM2 system was trained, using secondary feature vectors augmented by the frequency index. In Figure 5, the corresponding Gaussian means are shown for different phonemes of the database (for comparison, those occurring in Figure 8 were chosen). It can be seen that these parameters vary across phonemes, and that, for a given phoneme, they may also vary in time.

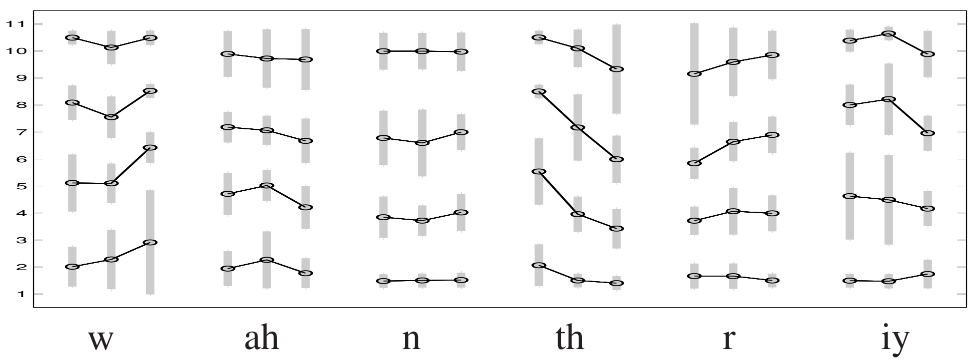


Figure 5: *Trained HMM2 parameters for different phonemes. In each column, the means of the frequency indices of the 4 secondary HMM states belonging to the same temporal state are visualized. Vertical bars show the respective variances. The 3 columns belonging to a phoneme correspond to the 3 temporal states.*

The corresponding variances are also visualized in the figure. While the trained means of the frequency index provide information about the position of the frequency bands modeled by the corresponding states, the variances model the respective bandwidths. The figure confirms that some general structural information of the phonemes is modeled, which is likely to be realteed to formant regions.

5 APPLICATIONS AND RESULTS

In the previous sections, it already has become apparent that HMM2 is not only limited in application to speech recognition decoding. It can also be used as feature extractor, e.g. to extract formant related regions. Although it has been shown in (Weber, de Wet, Cranen, Boves, Bengio & Bourlard, 2002) that, in the current HMM2 implementation, there is no one-to-one correspondence to formant positions, it is however clear that the resulting features are extracted in a principled way, optimizing a maximum likelihood criterion. The HMM2 applications are visualized in Figure 6.

Supposing that HMM2 does indeed segment the speech signal into formant-like regions, and given the fact that formants show a high robustness to noise, the HMM2 approach seems particularly promising for the recognition of degraded speech. The following gives a brief outline of the basic experimental setup, and then describes the proposed HMM2 applications, including speech recognition results in clean and noisy conditions where appropriate.

5.1 Experimental Setup

Experiments were carried out on the OGI Numbers95 corpus (Cole, Noel, Lander & Durham, 1995), on clean speech and on speech corrupted with 3 kinds of additive noises at 4 different signal-to-noise ratios (SNR). The noises were partly drawn from the Noisex database (Varga, Steeneken, Tomlinson & Jones, 1992), and partly provided by DaimlerChrysler in the framework of the SPHEAR project (SPHEAR). For the motivations described in Section 2.2, it is preferable to use features in the spectral domain. Frequency filtered filterbanks (FF2, see (Nadeau, 1999)) were chosen as they are lowly correlated spectral features which offer an acceptable baseline performance for clean speech. Twelve normalized FF2 coefficients (including one energy coefficient) were used. The 4-dimensional feature vectors consisted of a coefficient, its first and second order time derivatives and its frequency coefficient (here indices from 1 to 12). The HMM2 was implemented with HTK (Young, Odell, Ollason, Valtchev & Woodland, 1995). Eighty triphones model were used, each consisting of 3 temporal states. All secondary HMMs had 4 states connected in a looped top-down topology, similar to that in Figure 2. However, to take care of the energy, an additional state (without loops) was introduced as first state of each secondary HMM. There were 10 Gaussian mixtures in each secondary HMM state. This system was trained globally, on clean speech only, using the EM algorithm, and Viterbi-based recognition was performed under varying conditions (clean and all noises).

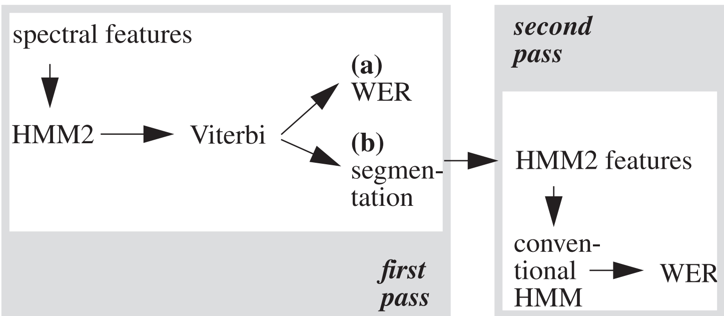


Figure 6: *HMM2 system used directly for speech recognition (a), and for features extraction (b). For (b), a second recognition pass, using a conventional HMM, is performed.*

5.2 HMM2 used as a Decoder

To realistically compare the performance of the HMM2 system (variant (a) in Figure 6) to that of a conventional HMM, tests were performed on both models given the same features (i.e., spectral FF2). Figure 7 shows results for one noise condition, with error bars indicating the 95% confidence interval. It can be seen that the differences in the performance of these 2 models are statistically significant. While HMM2 is not competitive with conventional HMMs in clean conditions or noisy speech with a high SNR, for heavily degraded noise it easily outperforms the conventional HMMs. In fact, HMM2 is better able to handle the mismatch between training and testing conditions (as training was done on clean speech only). This was confirmed on all other tested noise conditions. However, the obtained results (for both HMM and HMM2 with FF2 features) are not competitive with the state-of-the-art performance (obtained with conventional HMMs, but employing as features mel-frequency cepstral coefficients, including spectral subtraction and cepstral mean subtraction). In fact, the performance is limited due to the choice of features in the spectral domain, which were not found to be competitive with cepstral features. To further improve HMM2 performance and to evaluate whether HMM2 really has advantages over the usual HMM (using MFCCs) in adverse conditions, more research is required in the area of the robust extraction of spectral features.

5.3 HMM2 used as a Feature Extractor

One of the motivations for HMM2 is its ability to extract structural information of the speech signal, possibly corresponding to formant regions. Consequently, HMM2 could be used as a formant tracker. Although the interpretation of the segmentation of the full HMM2 as formant-like regions may not always be fully justified (as seen later), this application is motivated by HMM2 being a tool which integrates a speech decoder and a formant tracker in a unique model. This is supported by the assumption of Holmes (2000) that the “analysis of formants separately from hypotheses about what is being said will always be prone to errors”.

The segmentation between secondary HMM states, produced as a by-product of the Viterbi algorithm, can be interpreted as a separator between regions of different energy levels in the spectogram (just as the temporal segmentation separates phonetic units). If a distinct high energy region is surrounded by low energy along the frequency dimension, it may be assumed to correspond to a formant. Therefore, the HMM2 frequency segmentation could correspond to formant-like structures.

A first experiment was carried out using an HMM2 with shared parameters for all primary HMM states (i.e., the secondary HMM was the same for all phonetic units, trained on all data regardless of the labeling). In this case, no frequency index was appended to the secondary feature vectors. The frequency segmentation for an example speech unit is shown in Figure 8a. It can be seen that the 3rd secondary HMM state models a high energy region. However, in the case of less distinct or absent formants (as for the case of unvoiced phonemes), irregularities and discontinuities can be observed.

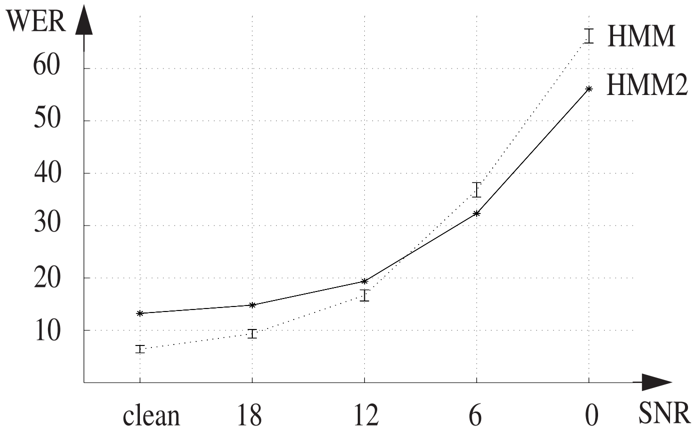
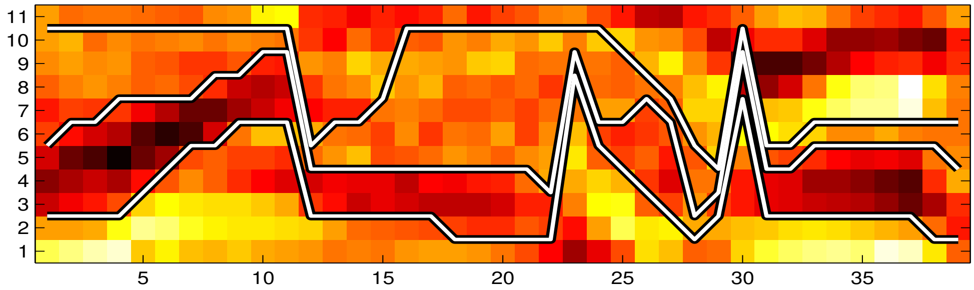
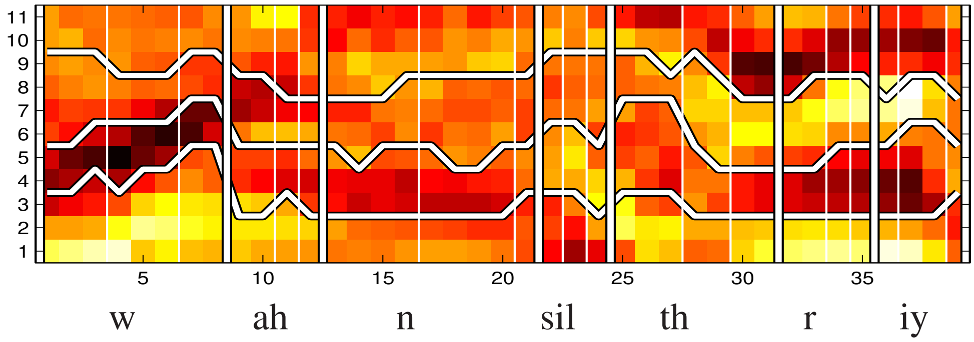


Figure 7: *HMM vs. HMM2 performance for frequency filtered filterbank features, illustrated by the dotted and solid lines respectively. Errorbars for HMM WER show the 95% confidence intervals. The results are for clean speech and car noise at different SNR.*



(a) segmentation obtained from a single secondary HMM



(b) segmentation (and labeling) obtained from a full HMM2

Figure 8: *Segmentations obtained (on unseen data) from (a) a single secondary HMM and (b) a full HMM2 system. In both figures, the horizontal lines correspond to the frequency segmentation. In (b), the vertical lines show the temporal segmentation obtained from the full HMM2 system, where phoneme boundaries are displayed as thick lines, and transitions between temporal states of the same phonemes as thin ones.*

When using a full HMM2 with class-dependent secondary HMMs and including the frequency index in the feature vector (as described in Section 5.1), the segmentation is smoother (see Figure 8b). However, high and low energy regions are not necessarily modelled by equivalent secondary HMM states (e.g., the third secondary HMM state may model a high energy region for one phoneme and a low energy region for another). Nevertheless, a certain structure of the speech signal becomes apparent from the segmentation. Furthermore, for the case of noisy speech, it has been shown that the HMM2 features exhibit a comparatively high robustness: the separation into regions of different energy levels is largely maintained.

Based on the above, HMM2 can be used as a feature extractor. This application is motivated by (1) the assumption that HMM2 extracts formant-like structures and (2) the discriminant properties of formant positions, and thus their usefulness as features for ASR. However, the formant-like features extracted by HMM2 are rather crude: they typically correspond to the frequency index of the secondary feature vector after which a transition from one secondary HMM state to the next occurred. Four secondary HMM states are used here for each phoneme, and a feature vector of 3 components is obtained, e.g. [3 7 9].

The features extracted by the simplified HMM2 (sharing parameters for all primary states) and the full HMM2 were tested. Word error rates (WER) of 37.0% and 18.6% respectively were achieved. From these results, it also appears that the smoother segmentation obtained from the full HMM2 contains more discriminant information. In addition to using the full HMM2 frequency segmentation as new features, the temporal segmentation was converted into a time index. Using the augmented 4-dimensional feature vectors, the WER could further be decreased to 15.0%. We consider this a good result given the low dimension and the crudeness of the feature vectors. However, before drawing any conclusions, further comparisons should be performed with other low-dimensional features, such as MFCCs projected down with linear discriminant analysis, and other formant-like features or even hand-labeled formants. Positive results in that sense were recently obtained on a different database and are reported in (Weber, de Wet, Cranen, Boves, Bengio & Bourlard, 2002).

When either of the different segmentation features (of the simplified or full HMM2) are combined in a multi-stream approach (at the level of the local likelihoods) with noise-robust MFCCs (already including spectral subtraction (SS) and cepstral mean subtraction), an improved robustness in noisy speech was observed, as shown in Table 1. Although equally good results might be obtained when using additional features other than those extracted with HMM2, it can be stated that a widely employed state-of-the-art noise-robust ASR system could be significantly improved (with more than 98% confidence).

SNR	HMM2 features	MFCC-SS	MFCC-SS + HMM2 features
clean	15.0	5.7	**5.7**
18	16.1	6.7	**6.6**
12	20.4	9.3	**9.0**
6	32.8	16.7	**16.1**
0	56.0	35.4	**34.3**

Table 1: *Performance of MFCC-SS and full HMM2 features, and their multi-stream combination: WER on Numbers95 at different signal-to-noise ratios: means over 3 different noise types.*

6 CONCLUSION

This paper has presented the motivations and foundations underlying the use of HMM2, a particular form of HMM in which emission probabilities are estimated through secondary, state-dependent, HMMs working along the acoustic feature dimension. It was shown that the parameters of this new model can also be trained using the Expectation-Maximization (EM) algorithm.

Including standard multi-Gaussian HMMs as a particular case, HMM2 provides additional modeling capabilities, allowing a principled approach towards flexible modeling of the time/frequency structure of speech through warping along the temporal and frequency dimensions. In particular, this paper has investigated the frequency warping aspect of HMM2. It was shown that, working in the spectral domain, HMM2 was able to automatically extract pertinent state/phone-specific formant-like structures during training and recognition. In fact, these formant-like structures can be used as low-dimensional features, which were shown to yield impressive speech recognition results. Furthermore, when using these features in conjunction with noise-robust MFCCs in standard speech recognition systems, an improved noise robustness was observed. Finally, HMM2 was also used directly as a decoder, achieving a good (although not yet fully competitive) recognition performance. These results indicate the advantages of HMM2 as an acoustic model, motivating research towards, for example, non-linear vocal tract normalization, new sub-band speech recognition approaches, and improved noise robustness.

7 ACKNOWLEDGEMENTS

This work was partly supported by grants FN 2000-059169.99/1 and FN 2100-061325.00/1 from the Swiss National Science Foundation. We would also like to thank the SPHEAR project partner DaimlerChrysler for providing us with their car noise.

8 REFERENCES

Bengio, S., Bourlard, H. & Weber, K. (2000). An EM Algorithm for HMMs with Emission Distributions Represented by HMMs. ***Technical Report IDIAP-RR 00-11***.

Bilmes, J.A. (1999). Buried Markov Models for Speech Recognition. ***Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing***, vol. II, pp. 713-716.

Bishop, C. M. (1995). ***Neural Networks for Pattern Recognition***. Clarendon Press, Oxford.

Cole, R. A., Noel, M., Lander T. & Durham, T. (1995). New Telephone Speech Corpora at CSLU. ***Proc. Eurospeech***, vol. I, pp. 821-824.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum-likelihood from incomplete data via the EM algorithm. ***Journal of the Royal Statistical Society B***, vol. 39, pp. 1-38.

Eickeler, S., Müller, S. & Rigoll, G. (1999). High Performance Face Recognition Using Pseudo 2D-Hidden Markov Models. ***European Control Conference (ECC)***.

Garner, P. & Holmes, W. (1998). On the Robust Incorporation of Formant Features into Hidden Markov Models for Automatic Speech Recognition. ***Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing***, vol. I, pp. 1-4.

Holmes, W. (2000). Segmental HMMs: Modelling Dynamics and Underlying Structure for Automatic Speech Recognition. ***IMA Workshop on Mathematical Foundations of Speech Processing and Recognition***.

Kuo, S. & Agazzi, O. (1993). Machine Vision for Keyword Spotting Using Pseudo 2D Hidden Markov Models. ***Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing***, vol. V, pp. 81-84.

Lee, L. & Rose, R. (1998). A Frequency Warping Approach to Speaker Normalization. ***IEEE Trans. on Speech and Audio Processing***, vol. 6, no. 1, pp. 49-59.

Levin, E. & Pieraccini, R. (1993). Planar Hidden Markov Modeling: from Speech to Optical Character Recognition. In: Hanson, S., Cowan, J., Giles, L., Editors, ***Advances in Neural Information Processing Systems, NIPS 5***, pp. 731-738, Morgan-Kaufmann.

Morris, A., Hagen, A., Glotin, H. & Bourlard, H. (2001). Multi-stream adaptive evidence combination for noise robust ASR. ***Speech Communication***.

Nadeu, C. (1999). On the Filter-bank-based Parameterization Front-End for Robust HMM Speech Recognition. ***Proc. Workshop on Robust Methods for Speech Recognition in Adverse Conditions***, pp. 235-238.

Rabiner L. & Juang, B. (1993). ***Fundamentals of Speech Recognition***. Prentice Hall Signal Processing Series.

Samaria, F. (1994). ***Face Recognition Using Hidden Markov Models.*** Ph.D. thesis, Engineering Department, Cambridge University.

SPHEAR TMR NETWORK project homepage, http://www.dcs.shef.ac.uk/~pdg/sphear/sphear.htm.

Varga, A., Steeneken, H. J. M., Tomlinson, M. & Jones, D. (1992). The Noisex-92 Study on the Effect of Additive Noise on Automatic Speech Recognition. ***Tech. Rep. DRA Speech Research Unit***.

Weber, K., Bengio, S. & Bourlard, H. (2000). HMM2- a novel approach to HMM emission probability estimation. ***Proc. Int. Conf. on Spoken Language Processing****, vol. III, pp. 147-150*.

Weber, K., Bengio, S. & Bourlard, H. (2001a). A Pragmatic View of the Application of HMM2 for ASR. ***Technical Report IDIAP-RR 01-23***.

Weber, K., Bengio, S. & Bourlard, H. (2001b). Speech Recognition Using Advanced HMM2 Features. ***IEEE Workshop on Automatic Speech Recognition and Understanding.***

Weber, K., de Wet, F., Cranen, B., Boves, L., Bengio, S. & Bourlard, H. (2002). Evaluation of Formant-Like Features for ASR, ***Proc. Int. Conf. on Spoken Language Processing,*** pp. 2101-2104.

Welling, L. & Ney, H. (1998). Formant Estimation for Speech Recognition. ***IEEE Trans. on Speech and Audio Processing***, vol. 6, no. 1, pp. 36-48.

Young, S., Odell, J., Ollason, D., Valtchev, V. & Woodland, P. (1995). ***The HTK Book.*** Cambridge University.