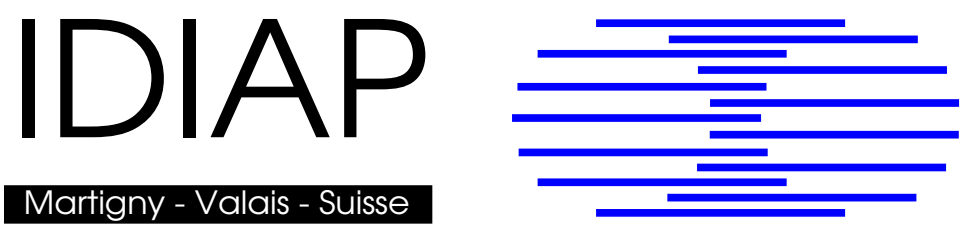
IDIAP RESEARCH REPORT



A Pragmatic View of the Application of HMM2 for ASR

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

IDIAP-RR 01-23

July 2001

REVISED VERSION

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 01-23

A Pragmatic View of the Application of HMM2 for ASR

Katrin Weber, Samy Bengio, and Hervé Bourlard

July 2001

Abstract: This report investigates the HMM2 approach recently introduced in the framework of automatic speech recognition. HMM2 can be seen as a mixture of HMMs, where a conventional primary HMM (processing a time series of speech data) is supported on a lower level by a secondary HMM, working along the frequency dimension of a temporal segment of speech. The application of HMM2 to the speech signal is motivated by numerous potential advantages. However, speech recognition results did not show the expected performance improvements. In this paper, the HMM2 approach is pragmatically analyzed and evaluated on speech data, revealing some problems and suggesting potential solutions.

Acknowledgements: This work was partly supported by grant FN 2000-059169.99/1 from the Swiss National Science Foundation.

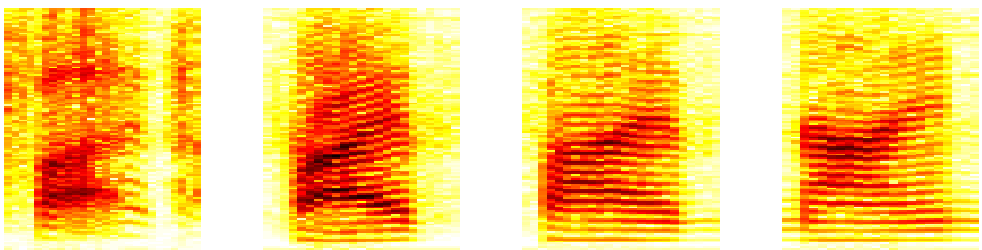
1 INTRODUCTION

In state-of-the-art automatic speech recognition (ASR), hidden Markov models (HMM) are widely used. While there are many suitable alternatives and design options for some parts of the ASR systems such as feature extraction and phoneme probability estimation, HMMs are the uncontested model for the temporal decoding part. 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 speaker's accents.

However, such variations do not only occur along the time axis, but they 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 1(a) with (b), and Figure 1(b) with (c) respectively). Furthermore, Figure 1(d) shows the same phoneme pronounced in a different context, revealing the effects of coarticulation. In all sub-figures, it is demonstrated that the position of spectral peaks may change significantly in the time-frequency plane during the pronunciation of a phoneme.

When using HMMs, we assume however that speech segments corresponding to one phoneme or subphone units are (1) invariable enough to be modeled by the same (mixture) distribution and (2) stationary for their duration, which obviously 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), we recently introduced the HMM2 approach [11]. A similar approach has previously shown some success in computer vision [3, 6, 10]. 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.

In spite of its numerous potential advantages, HMM2 has not yet shown competitive results in speech recognition. The purpose of this paper is to investigate in depth the HMM2 approach and its implications. In the following section, the HMM2 approach will be motivated. After having explained in more detail how HMM2 works and how such a system can practically be realized, we will give some speech recognition results. A thorough analysis of the drawbacks of HMM2 for this application is followed by a brief revision of alternative models in the framework of HMM2.



(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 MOTIVATIONS

In the previous section, we motivated HMM2 using real speech examples and explaining the problems encountered when conventional HMMs are applied for speech recognition. In summary, HMMs assume piecewise stationarity of the speech signal and do not truly take into account the existing variability along the feature (frequency) dimension. Using a secondary HMM for the local likelihood estimation, these assumptions are relaxed (at least to some degree), 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 frequency HMM, even though its position on the frequency axis is quite variable (as seen in Figure 1). Such a sparse frequency HMM topology also allows for efficient **parameter sharing**. The number of parameters can easily be controlled by the model topology and the probability density function associated with the frequency HMM states.

Furthermore, **correlation** between feature vector components is not ignored, but supposed to be modeled through the frequency HMM's topology. 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.

The secondary HMM performs automatically a **non-linear spectral warping**. While the conventional HMM does time warping and time integration, the frequency 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). With the same mechanism, also intra-speaker variations as well as coarticulation effects are taken care of.

Furthermore, the HMM2 topology permits implicitly 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 frequency HMM may be a good indicator for the position of a formant. Formants are assumed to carry most discriminative information in the speech signal, moreover being quite robust in the case of degraded speech. In [12] it was shown that the frequency HMM is indeed able to extract some meaningful and even discriminative formant-like structural information.

In the same line of reasoning, HMM2 can also be interpreted as a **dynamic approach to multiband processing**, where each frequency band is modeled by one frequency HMM state. By that we mean that each such state is supposed to emit a stationary sequence of spectral components belonging to a certain subband. The frequency position of the subbands would then automatically be adapted to the data, following e.g. formant-like structures.

We are now going to describe the HMM2 approach in some more detail, followed by an experimental and an error analysis section.

3 HMM2

HMMs are quite powerful statistical models which are used to represent sequential data, e.g. a sequence of acoustic vectors y_1^T in speech recognition[1] (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 components[2] $y_t = y_{t,1}^{t,S}$, another HMM can be used to model this feature dimension (displayed in the lower part of the figure).While the primary HMM mod-

1. All notations used in this report are explained in Appendix A.
2. 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 (i.e., a subvector is a coefficient). However, the extension of this approach to higher-dimensional subvectors (consisting, e.g., of a coefficient and its first and second time derivatives) is straightforward.

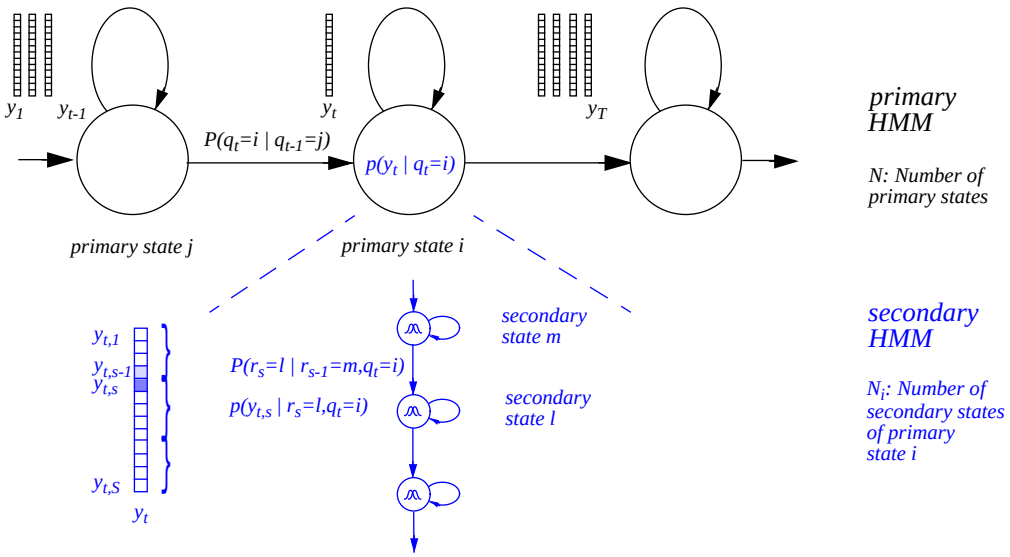


Figure 2: *HMM2 system. In the upper part, a conventional HMM, working along the temporal axis, can be seen. The local emission probability calculation is done with a secondary HMM, working along the frequency axis (depicted in the lower part of the figure).*

els 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).

HMM training is typically based on the expectation maximization (EM) algorithm. A generalization of the standard EM algorithm for HMM2 has been introduced in [1]. In the framework of this paper, we will investigate in more detail the estimation of $p(y_t|q_t)$ in the primary HMM states. Under the typical HMM assumptions (i.e. piecewise stationarity and data independence assumptions), the likelihood of an acoustic feature vector (i.e., a sequence of its components) given the primary HMM state can be expressed as:

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

or, using the Viterbi approximation:

$$p(y_t|q_t) \cong \arg\max_R \left[P(r_0|q_t) \prod_{s=1}^{S} [p(y_{t,s}|r_s, q_t) P(r_s|r_{s-1}, q_t)] \right] \tag{2}$$

where $P(r_0|q_t)$ is the initial state probability of the secondary HMM, $P(r_s|r_{s-1}, q_t)$ the state transition probabilities of the secondary HMM, and $p(y_{t,s}|r_s, q_t)$ the local likelihoods of the data. Naturally, every term of this equation is conditioned on the state of the primary HMM. As we use GMMs with diagonal covariance matrices for the likelihood estimation in the states of the secondary HMM, the corresponding local probability density functions (PDF) are defined as follows:

$$p(y_{t,s}|r_s = l, q_t = i) = \sum_{k=1}^{K} w_{ilk} \frac{1}{\sqrt{2\Pi\sigma_{ilk}^2}} e^{-\frac{1}{2}\left(\frac{y_{t,s}-\mu_{ilk}}{\sigma_{ilk}}\right)^2} \tag{3}$$

where K is the number of Gaussian mixtures.

After having described some practical realizations and experimental results of HMM2, we will come back to these mathematical derivations and investigate in detail their impact on practical implementations of the HMM2 approach.

4 PRACTICAL REALIZATION AND EXPERIMENTAL RESULTS

There are different ways to realize an HMM2 systems. Figure 3 shows two possibilities. The first realization (see Figure 3a) is based on the implementation of a generalized form of the standard EM algorithm, as described in [1]. This is the straight-forward way of realizing HMM2, implementing eq. 1 for the local likelihood estimation.

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 [10, 3], see Figure 3b). State likelihoods of the primary HMM are estimated using eq. 2. For this implementation, synchronization constraints have to be introduced to insure that exactly one feature vector is emitted between each two transitions in the primary HMM. This requires (1) additional synchronization states (grey in the figure) and (2) a re-arrangement of the data (as shown in the lower part of Figure 3b). Out-of-range synchronization components (modeled exclusively by the synchronization states) are introduced between the original feature vectors. The transitions between primary HMM states correspond to transitions between the synchronization states. Standard EM training algorithms (and therefore well-established tools such as HTK [14], which moreover offers sophisticated functionality especially adapted to speech recognition problems) can easily be used.

We did preliminary tests with both of the HMM2 systems described above. It was found that they yield a similar performance on small problems. For practical reasons, all further experiments used the implementation shown in Figure 3b, realized with the HTK system.

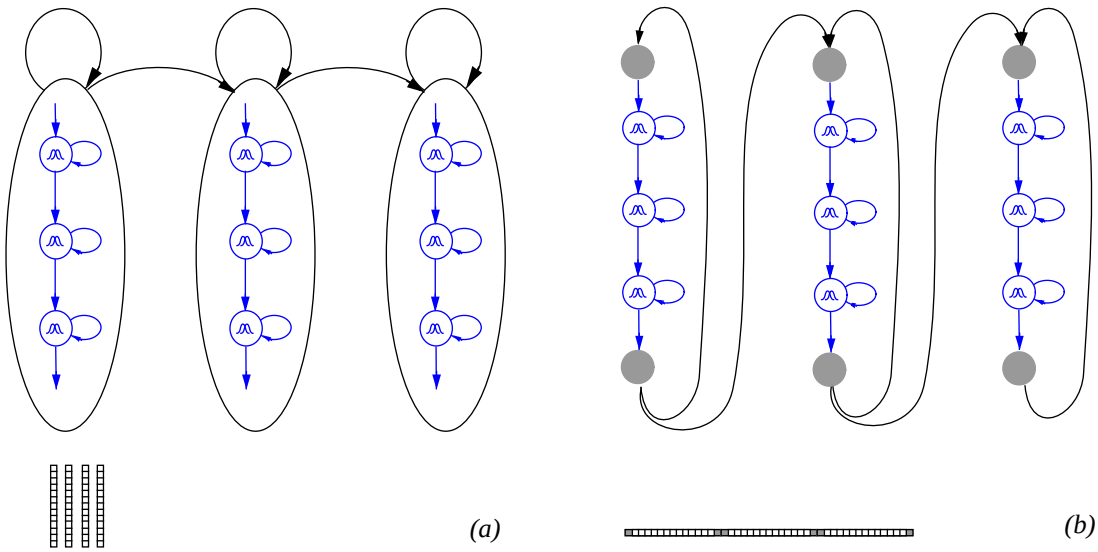


Figure 3: *HMM2 realizations: (a) direct implementation and (b) implementation with synchronization constraints. While the model in (a) is emitting a sequence of feature vectors (as usual), the model in (b) is emitting a sequence of (low-dimensional) components, intermitted by synchronization components at regular intervals.*

Another major concern when working with HMM2 is the choice of the features. We investigated different representations such as filterbanks, Rasta, and MFCCs. Obviously, for the motivations outlined in section 2 to hold, features in the spectral domain should be employed (although HMM2 might even show some advantages with different features). For most of our experiments, we used so-called FF2 features [8], which are frequency filtered filterbank coefficients. Compared to MFCCs, these features show only slightly worse speech recognition results on our HTK-based system (this result applies to clean data; however, performance degrades significantly in noisy conditions[1]). In addition to staying in the spectral domain (which also offers some benefits not further discussed here), FF2 features offer the advantage of being normalized to some degree (possibly large signal level variations are in fact smoothed out through the differencing).

The goal of preliminary experiments was to evaluate the HMM2 approach. To be able to directly compare HMM2 with the conventional HMM/GMM system, the topology of the primary HMM was left constant throughout the tests. Only the likelihood estimation in each primary HMM state was changed.

The Numbers95 database (a telephone-quality, small vocabulary, multi-speaker database containing continuously spoken digits, see [2]) was used throughout the tests. Each phoneme (triphone) present in this database was modeled with a primary HMM containing 3 emitting states. In the baseline system, the local likelihoods were estimated using a GMM with 10 Gaussian mixtures. For HMM2, several topologies for the secondary HMM were tested.

In all our experiments, a significant performance drop was observed when using HMM2 (with any secondary HMM topology). Speech recognition accuracy decreased significantly as compared to the conventional HMM/GMM system. This result is consistent for the two different HMM2 realizations described above, and holds for all kinds of features tested. In the following, we are investigating possible reasons for the observed degradation.

5 DIAGNOSTICS

The performance drops encountered in HMM2 require some careful, step-by-step error analysis. Consequently, we started from a simple secondary Markov model topology simulating a Gaussian distribution (i.e., here the Markov model is not hidden), gradually adding complexity. Again, the primary model topology was left constant. Results are compared to the conventional HMM/GMM baseline system. The experiments described in the following give some important cues about drawbacks of the HMM2 approach. Representative results can be found in Appendix B.

- **Experiment 1:** Simulation of an HMM/GMM with a single Gaussian distribution. The secondary Markov model (MM) has a strictly top-down topology without loops. The number of states is equal to the length of the sequence to be emitted (see Figure 4a). As there is only one possible state sequence R, the Markov model is not hidden anymore. The local likelihoods of the secondary MM states are estimated with single Gaussian distributions. As expected, recognition results are equivalent to those obtained with conventional HMMs employing a single Gaussian probability density function in each state.
- **Experiment 2:** Introduction of Gaussian mixtures (instead of single Gaussians) for local likelihood estimation (see Figure 4b). Here, the same not-hidden Markov model topology as in experiment 1 is used,

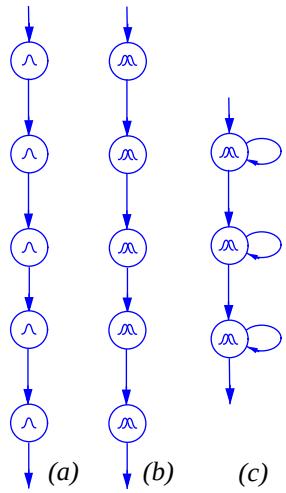


Figure 4: *Different frequency HMM topologies tested for error analysis.*

1. Unfortunately, in the framework of HMM/GMM, spectral features are usually not competitive with cepstral features such as MFCCs.

but at the level of the secondary HMM states, the single Gaussians are replaced by Gaussian mixtures. Speech recognition results improve as compared to experiment 1. However, in comparison to HMM/GMM incorporating an equivalent number of Gaussian mixtures, this model performs much worse.

- **Experiment 3:** Real secondary HMM. Compared to experiment 2, the number of states in the secondary HMM is reduced and self-transitions (loops) are added at each state. As there are fewer states than emitted components, the secondary Markov model becomes hidden (see Figure 4c). Speech recognition accuracy decreases as compared to the other systems tested.

In the following, we will try to identify the reasons for the losses encountered in HMM2. As in these experiments (and compared to the baseline system) we only changed the local likelihood estimation at the level of the primary HMM, we will concentrate our theoretical investigation on this part of the system. After having elaborated the general mathematical foundations, we will investigate the suitability of the model given the (speech) data. Furthermore, some peculiarities of the speech signal plus implications on a successful discriminative model are shown.

5.1 Effects of independent modeling of components

Firstly, we will investigate the effects of independent modeling of components in the secondary HMM states, as compared to the modeling of the entire vector in a GMM. For the case of frequency HMMs (a) and (b), eq. 1 simplifies drastically: as there is only one possible state sequence R through the model, we here deal with a 'normal' Markov model and no longer with a hidden one. Therefore, $P(r_0|q_t) = 1$ and $P(r_s = l|r_{s-1} = m, q_t) = 1$ for all transition (m, l) defined through the model topology. For case (a), there is even only a single Gaussian distribution, and so we obtain from eqs. 1 and 3:

$$p(y_t|q_t = i) = p(y_t, R|q_t = i) = \prod_{s=1}^{S} \frac{1}{\sqrt{2\Pi\sigma_{il}^2}} e^{-\frac{1}{2}\left(\frac{y_{t,s}-\mu_{il}}{\sigma_{il}}\right)^2} \qquad ;r_s = l \tag{4}$$

The above equation is equivalent to the state likelihood estimation in conventional HMM systems where the distribution is modeled by a single Gaussian. This fact was confirmed by our experimental results.

For case (b) and Gaussian mixture distributions in the secondary HMM states, the simplified state likelihood equation is:

$$p(y_t|q_t = i) = p(y_t, R|q_t = i) = \prod_{s=1}^{S} \sum_{k=1}^{K} w_{ilk} \frac{1}{\sqrt{2\Pi\sigma_{ilk}^2}} e^{-\frac{1}{2}\left(\frac{y_{t,s}-\mu_{ilk}}{\sigma_{ilk}}\right)^2} \qquad ;r_s = l \tag{5}$$

This equation bears a significant difference as compared to the distribution obtained for a conventional GMM:

$$p(y_t|q_t = i) = \sum_{k=1}^{K} w_{ik} \prod_{s=1}^{S} \frac{1}{\sqrt{2\Pi\sigma_{isk}^2}} e^{-\frac{1}{2}\left(\frac{y_{t,s}-\mu_{isk}}{\sigma_{isk}}\right)^2} \tag{6}$$

It can be seen that a sum of products (in the case of a GMM) has been replaced by a product of sums (in the case of a secondary HMM). Figure 5 shows the implications of these two equations on example toy data. It can be seen that the distribution obtained by the GMM (Figure 5b) is quite irregular. In fact, the shape of the distribution obtained by a GMM is practically only limited by the number of mixtures used. For example, the resulting PDF can take an (almost) elliptical form, whose principal axes are not necessarily parallel to the coordinate system (consider any two of the Gaussians in the figure, and approach their means). On the other hand, when modeling each feature component independently in a secondary HMM state, each mixture component in each state influences linearly all mixture components in all other states. Hence, the form of any resulting distribution is very restricted, as its principal axes inevitably follow the coordinate system's orientation. This is illustrated in Figure 5c. Therefore, correlation can not be modeled as well as in GMMs[1].

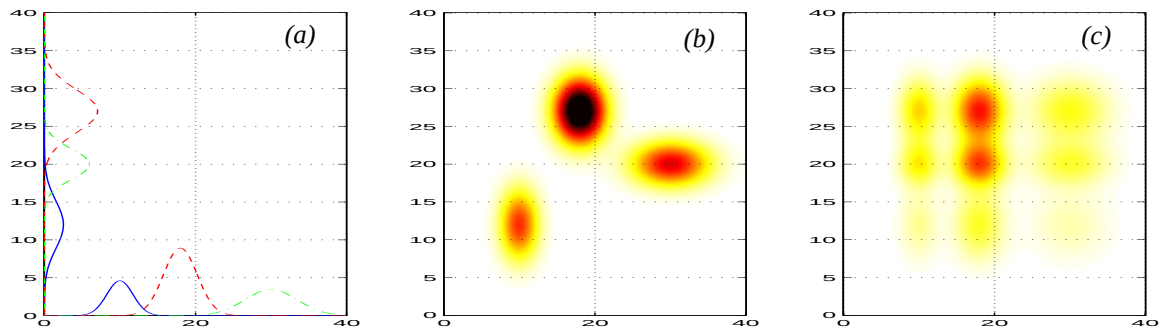


Figure 5: *Toy example: modeling power of GMM vs. HMM. In (a), a mixture of 3 2-dimensional Gaussians is defined (i.e., Gaussian means, variances and mixture weights). This GMM is visualized in (b). In (c), a distribution resulting from an HMM (also employing the parameters defined in (a)) is shown.*

5.2 Effects of the introduction of hidden states

Does this drawback generalize when moving from Markov models to hidden Markov models, or can it be compensated through some correlation modeling due to a suitable HMM topology? In the case of real HMMs (see Figure 4c), each possible path through the model corresponds to one Gaussian distribution, hence the sum over all possible paths corresponds to a Gaussian mixture (with as many mixture components as there are paths in the model):

$$\begin{aligned} p(y_t|q_t) &= \sum_R \left[P(r_0|q_t) \prod_{s=1}^{S} [p(y_{t,s}|r_s, q_t) P(r_s|r_{s-1}, q_t)] \right] \\ &= \sum_R \left[P(r_0|q_t) \prod_{s=1}^{S} P(r_s|r_{s-1}, q_t) \cdot \prod_{s=1}^{S} p(y_{t,s}|r_s, q_t) \right] \end{aligned} \tag{7}$$

where the respective products of initial and transition probabilities $P(r_0|q_t) \prod_{s=1}^{S} P(r_s|r_{s-1}, q_t)$ represent the mixture weights.

However, if one state emits several components ($r_s = r_{s+1} = \ldots = l$), the underlying PDF for their data likelihood estimation is constant (i.e., the Gaussian parameters are shared for the likelihood calculation of all those components). Hence, the distributions which can be modeled by such a secondary HMM are again very restricted. This fact is depicted graphically on a toy example in Figure 6. It can be seen that the resulting distribution obeys the same restrictions as the one shown in Figure 5: it is not possible to model distributions whose principal axes do not follow the coordinate system's orientation. For the kind of secondary HMM we are investigating here (i.e. top-down topology with fewer states than emitted components), this conclusion generalizes to higher-dimensional data and a higher number of Gaussian mixes.

In conclusion, Figures 5 and 6 both show that feature correlation can be modeled quite well by Gaussian mixture distributions, because they allow any orientation of the principal axes of the data distributions in a given coordinate system. This is not possible with our secondary HMM, because (1) the independent modeling of components in individual HMM states and (2) the parameter sharing (allowed by the stationarity assumption and enforced through looped HMM states) both constrain the resulting distribution to follow the orientation of the coordinate system. However, if the data were conform with both the independence and the

1. In fact, the traditional multiband approach suffers from a similar handicap, for which the full-combination approach [7] offers a remedy.

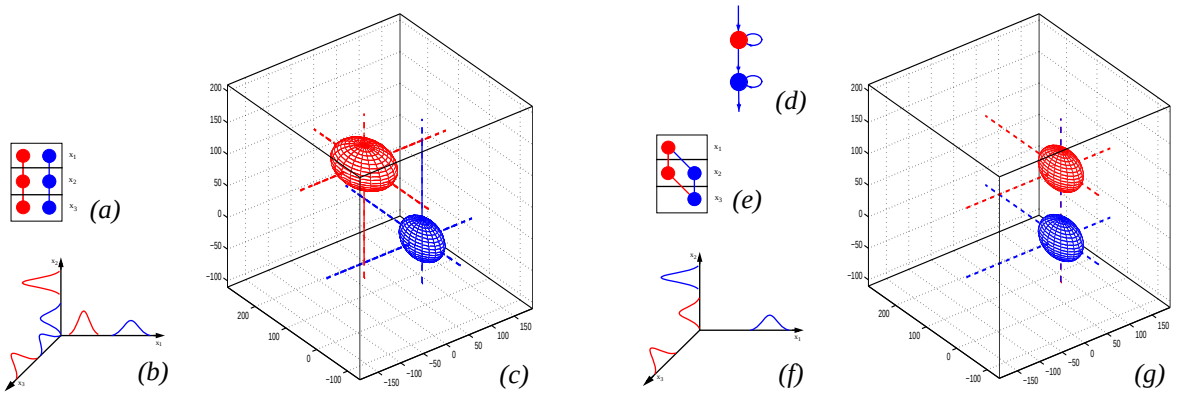


Figure 6: *Toy example: demonstration of the modeling capacity of a GMM (left part of the figure) and a secondary HMM (right part) for the case of 3-dimensional data. The GMM consists of a mixture of 2 Gaussians with diagonal covariance matrices. The secondary HMM has 2 states as shown in (d), thus there are 2 possible paths through the model (see (e), which compares to (a) for the GMM case). In (f), the Gaussian components contributing to the resulting distribution are depicted (compare to (b) for GMM). It can be seen that, for the case of the secondary HMM, only one dimension is expanded, resulting in the distribution depicted in (g). The principal axes of this distribution are constrained to follow the axes of the coordinate system, which is not the case for the distribution resulting from the GMM (depicted in (c)).*

stationarity assumptions, HMM2 could still be an appropriate model. In the next section, we will adopt a more data-driven point of view towards HMM2 and investigate the peculiarities of the speech data in respect to the above assumptions.

6 EVALUATION ON SPEECH DATA

6.1 Data representation by HMM2

In the previous section, we have collected theoretical evidence of the problems encountered in HMM2. In the following, we will investigate the implications of our findings on the application of HMM2 to speech data. Naturally, the HMM2 topology imposes similar assumptions on the data as HMMs conventionally used for time series. As described above, the data used in an HMM2 system is assumed to be conditionally independent (i.e., each data component is independent of all other components, given the primary and secondary HMM states) as well as piecewise stationary along both the time and the frequency axes (i.e., a few subsequent components are supposed to have been generated by the same probability density function). We now investigate whether these two assumptions are satisfied and their significance for the speech data representation in HMM2.

In Figure 7, correlation coefficients of FF2 features are visualized. It can be seen that the data are correlated, especially neighboring components in a feature vector (indicated in the figure by darker colors near the diagonal). Figure 8 shows how these correlated data are represented by a GMM and by a secondary HMM. The models are both trained on real FF2 speech data, and their respective parameters are visualized (in the same way as for the toy example in Figure 5). In the left part of the figure, it can be seen how the GMM parameters represent the existing data correlation. However, the HMM, shown in the right part, is not able to reproduce an appropriate data distribution. Although there are many suitable methods which orthogonalize data to some extent, completely uncorrelated features do not (yet) exist in the domain of ASR[1]. This fact alone does not favor HMM2 in the domain of speech.

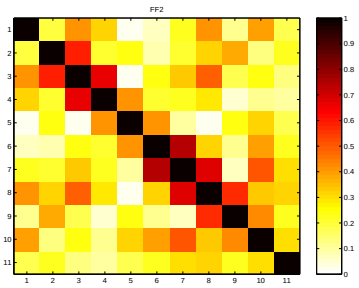


Figure 7: *Correlation coefficients of FF2 features. Dark colors correspond to high correlation coefficients.*

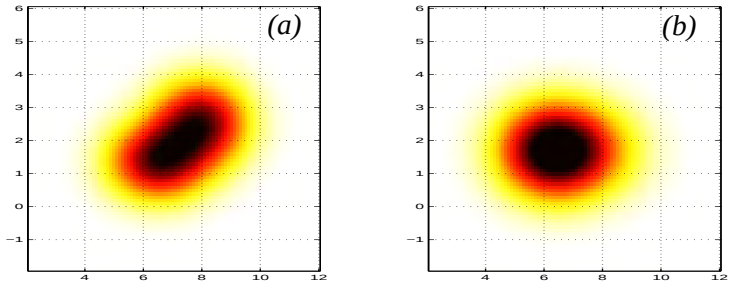


Figure 8: *Illustration of the modeling power of GMM and Markov model using real FF2 speech data. Figure (a) shows a part of a trained GMM, (b) the equivalent trained Markov model (only two dimensions are displayed). In either case, there are mixtures of 3 Gaussians. While in (a) data correlation becomes obvious, it cannot be seen in (b).*

The validity of the stationarity assumption is harder to fully prove or reject. Figure 9 shows an example pronunciation of phoneme 'ay'. It can be seen that the piecewise stationary assumption is not entirely satisfied. Nevertheless, it is intuitively (and practically, using a clustering algorithm) possible to segment this representation along the (horizontal) frequency axis in a few quasi-stationary sectors, which could subsequently be represented by the same PDF.

Even if the assumption of piecewise stationarity is to some degree satisfied, there is another implication of this assumption. Up to this point, we have investigated the ability of HMM2 to represent speech data, and we have stated some deficiencies of this approach in this respect. However, the goal in speech recognition is discrimination between phonetical units. In the following, we will examine the ability of HMM2 for discrimination.

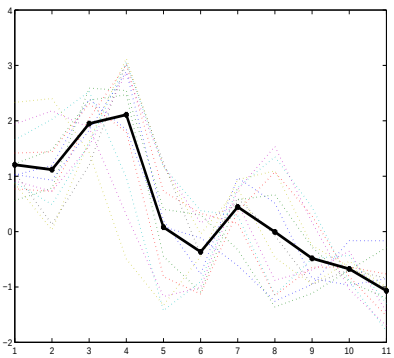


Figure 9: *A pronunciation of phoneme 'ay'. Each line in the figure corresponds to one time step, and thus to one feature vector (the thick black line is the mean). The horizontal axis shows frequency evolution, and the vertical axis shows the feature value (delta-energies in the case of FF2 features).*

6.2 Data discrimination by HMM2

It is widely acknowledged that spectral peaks (formants) contain important discriminant information [4, 13]. On the other hand, HMMs have already been applied to formant tracking [5]. If, as elaborated in section 2, the secondary HMM's frequency segmentation somehow reflects formant positions, this segmentation alone might represent rather discriminative information.

We conducted some experiments in order to find out the significance of the frequency HMM's segmentation, using a variant of the HMM2 approach: the secondary HMM is not used as a state likelihood estimator for the primary HMM, but instead as a feature extractor [12]. One secondary HMM with top-down topology and four looped states was trained on all the training data of our database (regardless of their labeling). Then, the Viterbi algorithm was used to segment each original feature vector along the frequency axis. The resulting segmentation consisted simply of 3 values, indicating the index of the feature component (in the original acoustic vector) after which a transition from one state of the secondary

1. Even the correlation coefficients of (the supposedly decorrelated) MFCC are quite comparable to those of FF2 (shown in Figure 7), with the difference of a lower correlation near the diagonal.

HMM to the next took place. In Figure 10, the means of 2 of these segmentation values (the ones corresponding to spectral peaks) are displayed for a number of voiced phonemes from our database. This figure is related to the F1-F2 plane, where vowels are positioned according to their formant frequencies (as, e.g., described in [9]). It was shown that the segmentation values indeed contain discriminative information: a conventional HMM was trained on low-dimensional vectors obtained from these rather crude 'formant features', and word recognition rates of over 56% were reached.

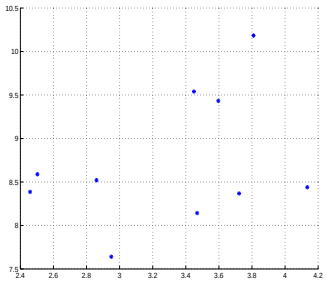


Figure 10: *Average segmentation values of the secondary HMM for different phonemes. These values correspond to spectral peaks (formants) in the signal. The figure is related to the formant-space representation of phonemes in the F1-F2 plane.*

This remarkable result proves that the secondary HMM's segmentation has a certain potential for discrimination. On the other hand, the results obtained with the original HMM2 system (where the secondary HMM was used as a likelihood estimator) show that in this approach, we cannot make use of this discriminative property, and important information seems to be lost. In that respect, HMM2 seems to suffer from the same problem as encountered in conventional HMMs: an imbalance between the contributions of HMM state likelihoods and transition probabilities to the estimation of the overall likelihood[1] (even though this effect is somewhat diminished due to the lower feature dimension in the secondary HMM). Consequently, the primary HMM state likelihoods do only insignificantly (if at all) reflect the segmentation produced by the secondary HMM. The improved flexibility of the model due to the high number of paths through the frequency HMM leads to a loss of discriminability (because of the loss of information concerning formant positions), which may rule out the potential gain through the frequency warping.

7 ALTERNATIVE MODELS

Given the problems identified in the previous sections, is there still hope for the HMM2 approach? Concluding from our findings, a successful HMM2 system would have to

- better consider data dependencies (as long as truly decorrelated features are not available) and
- assure that discriminability is maintained, e.g. that information about the position of spectral peaks in the speech signal is not lost.

In the following, we briefly propose some alternatives in the framework of HMM2, offering partial solutions regarding the requirements outlined above.

Remembering that the conventional HMM/GMM systems are a special case of HMM2, we could realize the following scenario: a GMM is modeled with a frequency HMM, as shown in Figure 11a. Then, additional transitions can be added (see Figure 11b). This would increase the model flexibility, but at the same time maintain some information about formant positions. Furthermore, as in a GMM, some data correlation could be modeled (the model should be at least as 'good' as a GMM, because, in the case that the newly added transitions do not help, their assigned probabilities after training should be 0). Experiments with such a 'trellis' model have however shown worse performances as compared to GMM. This is likely to be the effect of a

1. Together with the effects of the HMM's inherent exponential duration probability distribution, this leads in conventional HMMs (as well as in our primary HMM) to a poor duration modeling. However, these problems play in the conventional case only a subordinate role. On the one hand, the poor duration modeling can be compensated for, e.g. through lexical and grammatical restrictions in combination with word entrance penalties. On the other hand, the duration of a phoneme might not be an essential cue for discrimination, as this parameter varies considerably (depending on non-discriminant features such as the speaking rate).

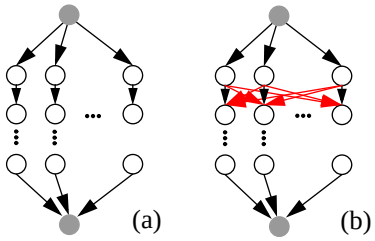


Figure 11: *In (a), a frequency HMM simulating a GMM is shown. Each vertical branch corresponds to one Gaussian mixture. (b) shows the extension of this model to a trellis topology.*

reduced discrimination capacity (in spite of the potentially increased descripitve power of the model) due to the improved model flexibility: not only the phonemes might be better represented by their respective models, but also (possibly to a greater extent) all other data. Here, discriminant training may offer a solution. However, most of the motivations given in section 2 do not hold for these systems. Having many states does not allow for an efficient parameter sharing. Frequency warping (and therefore non-linear vocal tract normalization and dynamic formant trajectory tracking) would only indirectly, if at all, be realized. Furthermore, the high number of states leads to an increased model complexity, and thus computation time quickly becomes an issue. This reasoning also applies to a similar system incorporating a lot of states in an ergodic secondary HMM topology.

If staying with the top-down and looped frequency HMM topology as originally introduced (and depicted in Figure 2), alternative design options which possibly improve the performance of an HMM2 system include

- **Emitting frequency context.** Instead of emitting just one coefficient at each frequency step, the secondary HMM could emit a vector consisting of this coefficient and its neighbors. Thereby, some correlation (near the diagonal) could be model through the GMMs in the secondary HMM states. Such a system has already been tested without much success.
- **Improve the influence of transition probabilities**. This could be done by reducing the influence of the secondary HMM state likelihoods during the estimation of the likelihood of a feature vector. One could even go as far as to, once a ‘best path’ through the secondary HMM is calculated using the Viterbi algorithm, discard these local likelihoods, just using the transition probabilities in the further computation. This approach is likely to ameliorate recognition results, as the position of formant regions would have somewhat more influence on the primary state likelihood, possibly resulting in an improved discrimination. However, the problem of the insufficient correlation modeling persists and is likely to limit the performance of the model.
- **Emit additional frequency information.** This is another way to make the frequency HMM model the positions of the spectral peaks. Each vector emitted by the secondary HMM is augmented by a coefficient indicating the position of this vector on the frequency axis. As in the previous option, correlation is still not thoroughly considered. Recognition performance was improved but is still limited by this deficiency.

All of the alternative models proposed above offer a partial rectification to the HMM2 problems stated in the previous sections. Even so, their effectiveness has yet to be shown. The possibly most promising variant of HMM2 is however the one already introduced in section 6.2, where the frequency HMM is used as a feature extractor. The resulting features represent formant-like structures, and when they are combined with state-of-the-art features such as MFCCs, speech recognition robustness has shown to improve significantly [12].

8 CONCLUSION

This paper was concerned with the HMM2 approach, where a secondary HMM is used to estimate local state likelihoods of a primary HMM, hence replacing Gaussian mixture models used in conventional HMMs. In spite of numerous strong motivations in favour of HMM2, experiments (using two different HMM2 realizations) did not show the expected results. The purpose of this paper was to outline theoretical and practical problems occurring when using HMM2 for speech recognition. Two major handicaps could be stated, con-

cerning the representative and discriminative abilities of the model respectively. It was found that the HMM2 approach suffers notably from

- the mismatch between the model capacity and the real distribution of the data, due to the unsatisfied independence assumption and
- a reduction of discriminability due to its (in some respect) higher flexibility and the ignorance of important information such as formant positions.

Consequently, present data correlation cannot be modeled, and possibly important information about positions of spectral peaks is basically lost. Some variants of the HMM2 approach offer partial solutions to the above problems, but non of them has as yet shown to be really successful in speech recognition. Although the secondary HMM's topology (and the values of the transition probabilities) might reflect some correlation as well as formant structure information, GMMs seem to be the more suitable model for phoneme discrimination (compared to our present HMM2 system).

9 REFERENCES

[1] S. Bengio, H. Bourlard, and K. Weber, "An EM Algorithm for HMMs with Emission Distributions Represented by HMMs," *IDIAP-RR 00-11*, 2000. *ftp://ftp.idiap.ch/pub/reports/2000/rr00-11.ps.gz.*

[2] R. A. Cole, M. Noel, T. Lander, and T. Durham, "New Telephone Speech Corpora at CSLU," *Proc. Eurospeech*, vol. I, pp. 821-824, Sep. 1995

[3] S. Eickeler, S. Müller, and G. Rigoll, "High Performance Face Recognition Using Pseudo 2D-Hidden Markov Models," *European Control Conference (ECC)*, Aug. 1999.

[4] P. Garner and W. Holmes, "On the Robust Incorporation of Formant Features into Hidden Markov Models for Automatic Speech Recognition," *Proc. ICASSP*, vol. 1, pp. 1-4, 1998.

[5] G. Kopec, "Formant Tracking using Hidden Markov Models and Vector Quantization," *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. 34, pp. 709-729, Aug. 1986.

[6] S. Kuo and O. Agazzi, "Machine Vision for Keyword Spotting Using Pseudo 2D Hidden Markov Models," *Proc. ICASSP*, vol. V, pp. 81-84, Apr. 1993.

[7] A. Morris, A. Hagen, H. Glotin, and H. Bourlard, "Multi-stream adaptive evidence combination for noise robust ASR", *Speech Communication*, 2001.

[8] C. Nadeu, "On the Filter-bank-based Parameterization Front-End for Robust HMM Speech Recognition," *Proc. Robust'99*, pp. 235-238, May 1999.

[9] L. Rabiner and B. Juang. "*Fundamentals of Speech Recognition*", Prentice Hall Signal Processing Series, 1993.

[10] F. Samaria, "*Face Recognition Using Hidden Markov Models*," Ph.D. thesis, Engineering Department, Cambridge University, Oct. 1994.

[11] K. Weber, S. Bengio, and H. Bourlard, "HMM2- a novel approach to HMM emission probability estimation," *Proc. ICSLP, vol. III, pp. 147-150*, Oct. 2000. *ftp://ftp.idiap.ch/pub/reports/2000/rr00-30.ps.gz.*

[12] K. Weber, S. Bengio, and H. Bourlard, "HMM2- Extraction of Formant Structures and their Use for Robust ASR," *to be published in Proc. Eurospeech*, Sept. 2001. *ftp://ftp.idiap.ch/pub/reports/2000/rr00-42.ps.gz.*

[13] L. Welling and H. Ney, "Formant Estimation for Speech Recognition", *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 1, pp. 36-48, 1998.

[14] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "*The HTK Book*," Cambridge University, 1995.

Appendix A: Notations and Abbreviations

General Notations

i, j	designate a temporal HMM state
K	number of Gaussian mixtures
k	k-th mixture component
l, m	designate a frequency HMM state
N	Number of temporal states
N_i	Number of frequency states in temporal state i
P	probability
p	probability density function
Q	set of all possible paths in primary HMM
q_t	temporal HMM state at time step t
R	set of all possible paths in secondary HMM
r_s	frequency HMM state at frequency step s
S	feature vector dimension (or number of components in each feature vector respectively)
s	frequency step
T	length of acoustic feature vector sequence
t	time step
w_k	weight of k-th Gaussian mixture
y_t	observed feature vector at time step t
y_1^T	observed feature vector sequence from time step 1 to T
$y_{t,s}$	observed feature component at frequency step s of time step t
μ_{ilk}	mean of k-th Gaussian mixture of the i-th temporal and the l-th frequency HMM state
σ_{ilk}	variance of k-th Gaussian mixture of the i-th temporal and the l-th frequency HMM state

Abbreviations

ANN	artificial neural network
ASR	automatic speech recognition
FF2	second order frequency filtered filterbanks
GMM	Gaussian mixture model
HMM	hidden Markov model
HMM/GMM	HMM employing GMM for local likelihood estimation
HMM/ANN	HMM employing ANN for phoneme emission probability estimation
MFCC	mel frequency cepstral coefficient
MM	Markov model
PDF	probability density function

Appendix B: Experimental Results

In the following, some representative experimental results are given. They were obtained using an HMM2 realization with synchronization states, implemented in the HTK system. As database, Numbers95 was used throughout. Spectral features with 11 FF2 (delta-frequency) coefficients and one filterbank energy, plus their first and second order time derivatives, were used. The experimental settings were such as to keep a maximum conformance to the baseline system, in order to directly compare performances. Each primary HMM had 3 emitting states and a left-right topology. For the benefit of higher recognition rates, the energy coefficient and time derivatives have been kept, although they caused some practical inconvenience in the HMM2 system. Each coefficient was grouped with its time derivatives into a 3-dimensional feature vector, supposedly emitted by a secondary HMM state. The energy subvector was treated separately in an independent state without loops.

The table below shows word error rates (and in brackets the number of parameters used to model the data distribution in each primary HMM state) of the different systems, for different training steps. Training was started on 27 monophone models with single Gaussian distributions (first line in the table below). These were subsequently split up to mixtures of 10 Gaussians (second line), and finally 80 triphone models were created (last line). In the first column of the table, the FF2 baseline performance is shown. The overall word error rate on an independent test set is 6.7% (which compares to 5.7% on MFCC features). The second column shows the results for systems such as described in section 5, experiments 1 and 2: the secondary Markov model (MM) is not hidden. Comparing lines 1 and 2 of the first two columns, it can be seen that the relative improvement when introducing Gaussian mixtures is not as significant as in the HMM/GMM case. The third column shows the HMM2 performance. In this case, the secondary HMM is composed of 4 emitting states in a looped top-down topology, one additional state exclusively for the energy subvector, and 2 synchronization states. Emitting additional frequency information (as suggested in section 7) yields a word error rate of 15.9%. It can be stated that HMM2 generally has the highest word error rates.

	HMM/GMM	HMM/MM	HMM2
1 Gaussian, monophones	22.2 (66)	21.8 (66)	41.9 (30)
10 Gaussians, monophones	12.5 (670)	18.3 (760)	31.6 (358)
10 Gaussians, triphones	**6.7**	**11.4**	**20.5**

Table 1: *Word error rates (and number of parameters in a primary HMM state) on baseline HMM/GMM, HMM/MM and HMM2 systems.*