IDIAP RESEARCH REPORT

IDIAP

Martigny - Valais - Suisse

HMM2- A NOVEL APPROACH TO HMM EMISSION PROBABILITY ESTIMATION

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

IDIAP-RR 00-30

September 2000

TO APPEAR IN

Int. Conf. on Spoken Language Processing, ICSLP 2000, Beijing

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 00-30

HMM2- A NOVEL APPROACH TO HMM EMISSION PROBABILITY ESTIMATION

Katrin Weber, Samy Bengio, and Hervé Bourlard

September 2000

TO APPEAR IN

Int. Conf. on Spoken Language Processing, ICSLP 2000, Beijing

Abstract: In this paper, we discuss and investigate a new method to estimate local emission probabilities in the framework of hidden Markov models (HMM). Each feature vector is considered to be a sequence and is supposed to be modeled by yet another HMM. Therefore, we call this approach 'HMM2'. There is a variety of possible topologies of such HMM2 systems, e.g. incorporating trellis or ergodic HMM structures. Preliminary HMM2 speech recognition experiments on cepstral and spectral features yielded worse results than state-of-the-art systems. However, we believe that HMM2 systems have a lot of potential advantages and are therefore worth investigating further.

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

1 INTRODUCTION

In automatic speech recognition (ASR), HMMs represent the state-of-the-art for phoneme and word recognition from sequences of acoustic feature vectors. In such HMMs, the computation of the likelihood of a feature vector given a certain state is conventionally performed by Gaussian mixture models (GMM) or artificial neural networks (ANN).

This paper investigates a new approach for HMM state likelihood calculation, using the modeling power of Gaussian distributions and at the same time allowing for more flexibility in the choice of features and integrated training. In fact, instead of using Gaussian distributions or ANNs, we introduce yet another HMM (denoted "internal HMM") at the level of each state of the conventional HMM (denoted in this context "external HMM").

After having described the HMM2 approach in more detail and introduced two particular examples of HMM2 system, we will explain potential advantages of the system and report initial results on its application to speech recognition as well as frequency segmentation.

2 THE HMM2 MODEL

Figure 1 shows an example of an HMM2 system. The external (temporal) HMM emits a sequence of feature vectors, exactly as a conventional HMM does. Each of these feature vectors, however, is emitted by the internal HMM as a sequence of subvectors (see Figure 2). The internal HMM thus replaces Gaussian mixture distributions of conventional HMMs. As the state likelihoods of the internal HMM are again estimated by Gaussian mixture models (GMM), the HMM2 approach is in fact a generalization of conventional GMM-based HMMs. An adapted version of the EM algorithm has already been developed to train HMM2 systems (see [1]).

As explained above, an internal HMM is introduced in each state of the external HMM. It is therefore responsible for estimating the likelihood of a feature vector, given a state of the external HMM. This feature vector is cut into subvectors (denoted in the following "internal feature vectors"). For example, a feature vector consisting of 45 coefficients can be split into a series of 3 internal feature vectors, each of which comprises 15 coefficients. A series of 15 3-dimensional internal feature vectors, each of which covering a coefficient and its first and second order derivatives, might be a straightforward choice for commonly applied ASR features. Pushing the HMM2 approach to its extremes, we obtain a series of 45 1-dimensional internal vectors, whereas a "series" of 1 45-dimensional internal feature vector is one way to capture the special case of the conventional GMM.

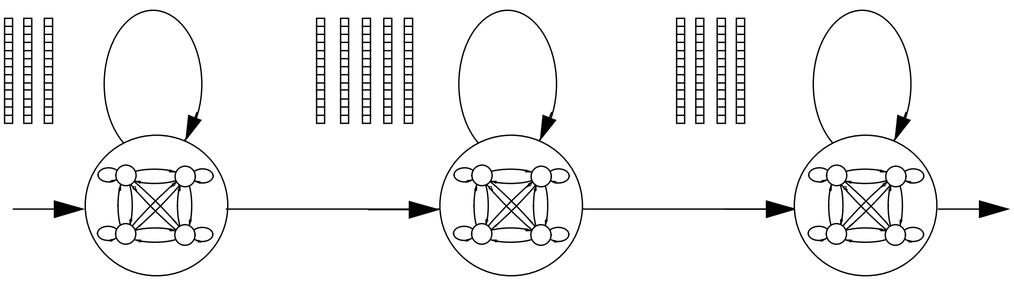

Figure 1: *HMM2 system: Integration of the internal HMMs into the states of the external HMM.*

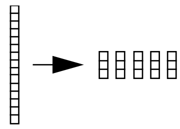

Figure 2: *Example of a feature vector of size 15 being decomposed into a sequence of 5 3-dimensional "internal feature vectors".*

Using internal HMMs implies the same assumptions as for processing time series of data with conventional HMMs, notably that the series of internal feature vectors can be modeled by a first order Markov model. However, there is some flexibility in the choice of topology for the internal HMM. Naturally, it will be influenced by the kind of features employed. An internal HMM can have fewer or more states than there are internal feature vectors. If there are more states, paths through an internal HMM will tend to encode some correlation information between the internal feature vectors. If there are fewer states, the internal HMM will segment a feature vector, possibly into subbands-equivalent units. Furthermore, the connectivity of the internal HMM has still to be explored. In the following section, we will introduce two particular architectures of the internal HMM: the hidden Markov trellis (having more states than internal feature vectors) and the frequential HMM (having fewer states).

2.1 The hidden Markov trellis

The hidden Markov trellis was inspired by earlier work on the wavelet-domain hidden Markov tree [3]. It models a feature vector by a chain of nodes, where each internal feature vector is assigned to one distinct node (see Figure 3). Each node comprises several HMM states, and all states from one node are connected to all states in the subsequent node using transition probabilities. We can view the resulting trellis as a generalization of diagonal GMMs, where correlations between coefficients are modeled in a more tractable and less parameter-intensive way compared to full covariance matrices.

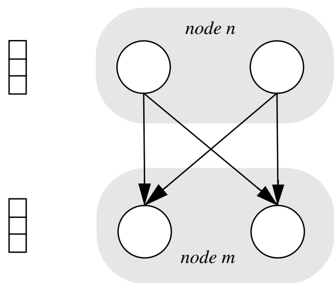

Figure 3: *The hidden Markov trellis: extract showing two nodes emitting two internal feature vector.*

2.2 The frequential HMM

While the trellis model introduced above is suitable for a wide range of features, the frequential HMM is especially adapted to features in the spectral domain. Typically, coefficients derived from neighboring frequency bands are not only strongly correlated, but might as well share the same characteristics. Therefore, several coefficients of a feature vector can be modeled by just one state in the internal HMM. We here chose an ergodic model as internal HMM, as this is the most general topology (as in Figure 1).

As spectral features can be seen as sequences of energies in subsequent frequency bands, the frequential HMM might perform an implicit segmentation of the feature vectors into subband-like units. This particular property might not only be directly useful for speech recognition, but also opens up new ways in research directions such as formant tracking and vocal tract normalization for speaker adaptation. When looking at a spectogram of a vowel, it can be seen that frequency bands of high energy, which correspond to formants, change with time during the pronunciation of this phoneme. Ideally, these formants would be modeled by one state of the frequential HMM, whereas the neighboring regions of lower energy would be modeled by others. Subsequent feature vectors could be modeled by the same sequence of internal states, while the actual alignment might be slightly different and represent the dynamics of speech.

In the same spirit, the frequential HMM could also perform some sort of speaker adaptation. For different speakers, the spectograms of the same phoneme should show an equivalent structure, but the location of the formants on the frequency axis might vary and reflect differences in the vocal tracts. Given a trained frequential HMM, we might perform vocal tract normalization just by changing the transition probabilities according to the particular speaker characteristics.

2.3 Motivation

We believe the HMM2 system to have several potential advantages, such as:

- Better modeling of the correlation across feature vector components: Contrarily to GMM-based systems, where we usually make the assumption that the feature vector components are uncorrelated, we here only assume that a sequence of subvectors can be modeled by a 1st order HMM.

- Modeling of the dynamics of the signal by implicit nonlinear frequency warping and better modeling of the underlying time/frequency structure: As described above, regions of high energy are not stationary in one frequency band, and furthermore depend on the vocal tract characteristics of a speaker. While both GMM and ANN based systems disregard these dynamics, they can be captured by HMM2. The assignment of subvectors to internal HMM states by the Viterbi algorithm could produce a frequency segmentation corresponding to formant regions. The HMM2 system could therefore possibly be used for implicit or explicit formant tracking as well as vocal tract normalization.

- More modeling capabilities with a parsimonious number of parameters: As there are generally many paths through the internal HMM, parameters may be shared by subvectors in a flexible way. This parameter sharing is data-driven and governed by the internal HMM's transition probabilities.

3 PRELIMINARY RECOGNITION EXPERIMENTS

3.1 Database and feature extraction

The OGI Numbers95 corpus [2] was used throughout. Its vocabulary comprises 30 words, and there are 27 phonemes. Experiments were carried out on two different kinds of features: MFCC and log Rasta-PLP spectra. MFCC feature vectors consisted of 13 coefficients (including energy). Spectral subtraction and cepstral mean subtraction were applied. Rasta-PLP feature vectors had 15 coefficients. In both cases, first and second order derivatives of the features were used too, tripling the number of coefficients as given above.

3.2 Reference system

The experiments described in the following sections cannot be compared directly to our GMM baseline systems (yielding a word error rate (WER) of 5.4% on the development test set) as, for practical reasons, some modifications had to be made. Firstly, we did not employ integrated EM training. Previously segmented data (obtained by using the baseline system) was used throughout, allowing the external and internal HMMs to be trained separately. Secondly, we only used monophone models although triphone models show much better results in our baseline system. However, as in the baseline system, the 3 emitting states of each monophone model do not share parameters. Thirdly, the number of Gaussians per mixture was not optimized but chosen in order to obtain a comparable number of parameters of reference and HMM2 systems.

The reference system for MFCC features uses mixtures of 6 multivariate Gaussians in each state. Taking into account the above described restrictions, our MFCC reference system yields frame error rate (FER) of 30.1% and a word error rate (WER) of 11.5% (both on the development test set). The reference system for Rasta-PLP spectral features performs, even with a higher number of Gaussians (we used 24), worse, yielding a FER of 44.9% and a WER of 19.6%.

3.3 HMM2 realization

As described above, we have 27 (external) monophone models, each of which comprises 3 emitting states, i.e. 3 internal HMMs. Therefore, we have a total of 81 internal HMMs. For cepstral and spectral features, different topologies for the internal HMM were tested. The models were trained using the state-segmented (at the level of the external HMM) Numbers95 training data. For the training of the internal models, the EM algorithm was used.

3.4 Experiments with cepstral features

Mel frequency cepstral coefficients (MFCC) have shown very competitive performance in speech recognition systems based on GMMs and can be seen as state-of-the-art ASR features. Therefore, they were our first choice for initial experiments. However, MFCCs of one feature vector are typically not comparable; in fact their means and variances might vary by several orders of magnitude. Therefore, modeling different cepstral coefficients in a single state does not seem to be a sensible choice. The trellis topology, to the contrary, seems very appropriate for these features.

Our hidden Markov trellis model contains as many nodes as there are coefficients (i.e. 13). Each node consists of two states. An internal HMM state emits a three-dimensional vector: a coefficient as well as its first and second order derivative. The implemented trellis system can be seen as a generalization of our reference GMM system. The reference system uses mixtures of 6 Gaussians, whereas the trellis system implements 2 mixtures (states) of 3 Gaussians in each node. There are 476 parameters in one internal HMM of the reference system and 618 in a trellis, the difference being due to the transition probabilities and Gaussian weights.

Results for both the reference and the trellis system are shown in Table 1. The results on the trellis are much worse than those of the reference system. Different statistical analyses have been carried out for both the train as well as the development test sets. Comparing per-phoneme recognition rates, most phonemes were better recognized by the reference system. The confusion matrices of the trellis system show a rather similar pattern compared to those of the reference system. In terms of likelihood mean, likelihood ratio and relative entropy calculated over all phonemes, the trellis system seems to highly outperform the reference system. However, these measures calculated on each phoneme separately show that only very few phonemes of the trellis system perform better compared to the reference system. This difference is due to large variations (of orders of magnitude) of the likelihoods. In fact, the median of the likelihoods shows that the reference system is the better one in any case.

	FER-train	WER-train	FER-devt	WER-devt
reference system	29.4	9.4	30.1	11.6
HMM2 (trellis system)	37.0	17.9	37.5	19.5

Table 1: *Comparison of HMM2 trellis system to reference system: Frame error rate (FER) and word error rate (WER) on Numbers95 full train and development test sets (denoted "train" and "devt" respectively).*

3.5 Experiments with spectral features

When using features from the spectral domain in GMM-based automatic ASR, recognition performance is generally much worse than with orthogonalized features. However, in certain cases spectral features are still preferable. If, e.g., the signal is distorted by band-limited noise, this noise will be spread over all cepstral coefficients, whereas most of the spectral coefficients remain clear. For the experiments described below, we used log Rasta-PLP spectra throughout. The frequential HMM seems to be the obvious choice for the topology for the internal model.

Different topologies for the frequential HMM were tested. The best tested internal HMM comprises seven states and has an ergodic topology. There is a mixture of 10 Gaussians in each state. The internal HMM emits a sequence of 3 15-dimensional internal feature vectors. Table 2 shows results for different topologies of the internal HMM on the Numbers95 development test set. Analyses of these results again show generally better

	FER-devt	WER-devt
reference system	44.9	19.6
HMM2 (frequential)	50.2	32.8

Table 2: *Comparison of reference and HMM2 frequential systems for Rasta-PLP spectra: Frame error rate (FER) and word error rate (WER) on Numbers95 full development test sets.*

per-phoneme recognition rates for the reference system. Likelihood mean, median, ratio and relative entropy calculated over all phonemes favor all HMM2, but the difference between the overall medians is rather small. Looking at each phoneme separately, it becomes obvious that the reference system is the better one.

4 PRELIMINARY ANALYSIS OF FREQUENCY SEGMENTATION BY HMM2

As for the experiments described in the previous section, we used the Numbers95 database and log Rasta-PLP features. We trained just one internal HMM on all the data assigned to the phoneme 'iy'. This internal HMM comprises 4 states and has and ergodic topology, exactly as shown in Figure 1. A sequence of 15 3-dimensional internal feature vectors (each consisting of one coefficient as well as its first and second order derivatives) is emitted. Given a trained model, we apply the Viterbi algorithm in order to obtain an alignment at the

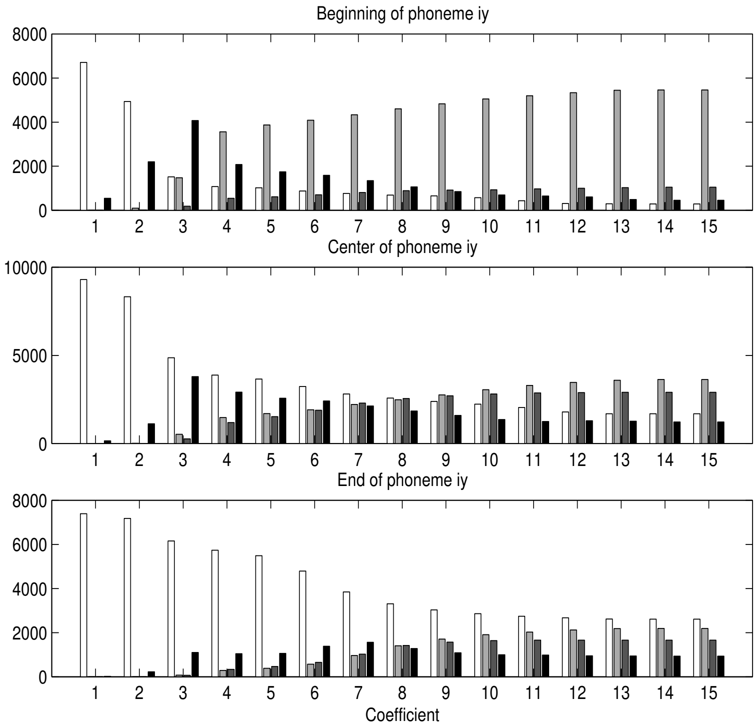

Figure 4: *State occupation of the internal HMM for each coefficient, calculated on three data subsets of phoneme iy, corresponding to the beginning, center, and end of the phoneme. The 4 different shades of the bars correspond to the 4 states of the model.*

level of the internal HMM states. For statistical analysis, the data of all examples of phoneme 'iy' were again segmented into 3 subsets (corresponding to the 3 states of the external HMM). So there are three data sets representing the beginning, the center, and the end of the phoneme respectively.

Figure 4 shows the histograms of state occupations for each coefficient and each data subset. It can be seen that the first few coefficients are mostly emitted by one state (white bar in the figure). The number of coefficients emitted by that state seems to increase with time (for the beginning of the phoneme, the white state emits only the first two coefficients, while towards the end it makes a considerable contribution for all coefficients). At the beginning of the phoneme, another state (displayed in black) seems to be responsible for the third coefficient. The importance of this state decreases with time, and its contribution at the end of the phoneme is negligible. In summary, we can see that, although the internal HMM employed here has an ergodic topology, some structural information is extracted by the frequential segmentation. This result still needs to be further analyzed.

5 CONCLUSIONS

In this article, we introduced HMM2 as a novel way to estimate HMM emission probability. We integrated internal HMMs in each state of temporal HMMs. Different topologies of the internal HMM were investigated, and experiments were run with cepstral and spectral features. A preliminary analysis of the frequency segmentation performed by HMM2 was done. While initial speech recognition results were found not to be competitive with conventional state-of-the-art HMMs, the HMM2 systems could extract some structural information from the data. Therefore, we believe that this approach provides us with a new framework with a lot of potential advantages, as described in section 2.3.

In future, we plan to further investigate the HMM2 approach for speech recognition. An optimal parameter set has not been found yet, and there might be topologies more suitable for ASR features than those tested so far. Furthermore, the relationship of the HMM2 system with formant tracking and vocal tract normalization should be explored. It has to be investigated whether an internal HMM is indeed able to represent formant-like structures and what the topology of such an HMM would be. Finally, ways of integrating such information into a (multi-stream) speech recognition system should be explored.

6 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.

[2] R. A. Cole, M. Noel, T. Lander, and T. Durham. New Telephone Speech Corpora at CSLU. *Proceedings of the European Conference on Speech Communication and Technology*, 1:821-824, September 1995.

[3] K. Keller, S. Ben-Yacoub, and C. Mokbel. Combining Wavelet-domain Hidden Markov Trees with Hidden Markov Models. *IDIAP-RR 99-14,* 1999.