Double Layer Architectures for Automatic Speech Recognition Using HMM

Marta Casar and José A. R. Fonollosa
Dept. of Signal Theory and Communication, Universitat Politècnica de Catalunya (UPC), Barcelona, Spain

1. Introduction

Understanding continuous speech uttered by a random speaker in a random language and in a variable environment is a challenging problem for a machine. Broad knowledge of the world is needed if context is to be taken into account, and this has been the main source of difficulty in speech-related research. Automatic speech recognition has only been possible by simplifying the problem - which involves restricting the vocabulary, the speech domain, the way sentences are constructed, the number of speakers, and the language to be used-, and controlling the environmental noise.

Current speech recognition systems are usually based on the statistical modelling of the acoustic information, generally using hidden Markov models (HMM). However, these systems are subject to some restrictions regarding the incorporation of other speech-related knowledge that might have an influence on the recognition rate.

The evolution of Automatic Speech Recognition (ASR) technology over the last few years has led to the development of applications and products that are able to operate under real conditions by acknowledging the above-mentioned limitations (or simplifications). ASR applications include dialogue systems, speech-based interfaces (such as automatic access to information services) and voice-controlled systems (like voice-driven database retrieval).

Due to the number of potential ASR applications, research efforts have been focused on developing systems that can accept spontaneous speech in a wide range of environments and from a wide range of speakers. However, in the case of spontaneous speech, large vocabularies must be considered. Moreover, language must be modelled by a non-restrictive grammar, which takes into account events that are common in natural speech, such as truncated or grammatically incorrect sentences, non-speech-events and hesitation. To deal with this, and to be able to introduce all the information available into the recognition architecture, a change of paradigm from conventional speech recognition had to be proposed.

In this chapter, we will talk about different approaches to a double layer architecture using HMM for ASR, which should allow other, non-acoustic information to be incorporated and more complex modelling of the speech signal than has been possible up to now. After analyzing different approaches, the main conclusions will be summarized and possible further work in this field will be briefly discussed.

Source: Robust Speech Recognition and Understanding, Book edited by: Michael Grimm and Kristian Kroschel,
ISBN 987-3-90213-08-0, pp.460, I-Tech, Vienna, Austria, June 2007

2. ASR Using HMM

2.1 Standard systems

A standard ASR system is based on a set of so-called acoustic models that link the observed features of the voice signal to the expected phonetics of the hypothesis sentence. The most typical implementation of this process is probabilistic, namely Hidden Markov Models (HMM) (Rabiner, 1989; Huang et al., 2001).

A Markov model is a stochastic model that describes a sequence of possible events in which the probability of each event only depends on the state attained in the previous event. This characteristic is defined as the Markov property. An HMM is a collection of states that fulfils the Markov property, with an output distribution for each state defined in terms of a mixture of Gaussian densities (Rabiner, 1993). These output distributions are generally made up of the direct acoustic vector plus its dynamic features (namely, its first and second derivatives), plus the energy of the spectrum. Dynamic features are the way of representing context in an HMM, but they are generally only limited to a few subsequent feature vectors and do not represent long-term variations. Frequency filtering parameterization (Nadeu et al., 2001) has become a successful alternative to cepstral coefficients.

Conventional HMM training is based on maximum likelihood estimation (MLE) criteria (Furui & Sandhi, 1992), via powerful training algorithms such as the Baum-Welch algorithm and the Viterbi algorithm. In recent years, the discriminative training method and the minimum classification error (MCE) criteria, which is based on the generalized probabilistic descent (GPD) framework, has been successful in training HMMs for speech recognition (Juang et al., 1997). For decoding, both the Viterbi and Baum-Welch algorithms have been implemented with similar results, but the former showed better computational behaviour.

The first implementations of HMMs for ASR were based on discrete HMMs (DHMMs). In a DHMM, a quantization procedure is needed to map observation vectors from the continuous space to the discrete space of the statistical models. Of course, there is a quantization error inherent to this process, which can be eliminated if continuous HMMs (CHMMs) are used.

For CHMMs, a different form of output probability function is needed. Multivariate Gaussian mixture density functions are an obvious choice, as they can approximate any continuous density function (Huan et al., 2001). However, computational complexity can become a major drawback in the maximization of the likelihood by way of re-estimation, as the M-mixture observation densities used must be accommodated.

In many implementations, the gap between the discrete and continuous mixture density HMM has been bridged under certain minor assumptions. For instance, in a tied-mixture HMM the mixture density functions are tied together across all the models to form a set of shared kernels.

Another solution is a semi-continuous HMM (SCHMM), in which a VQ codebook is used to map the continuous input feature vector $\mathbf{x}$ to o_k, as in a discrete HMM. However, in this case the output probabilities are no longer used directly (as they are in a DHMM), but rather combined with the VQ density functions. That is, the discrete model-dependent weighting coefficients are combined with the continuous codebook's probability density functions.

From another point of view, semi-continuous models are equivalent to M-mixture continuous HMMs, with all the continuous output probability density functions shared by all the Markov states. Hence, SCHMMs maintain the modelling ability of large-mixture probability density functions. In addition, the number of free parameters and the computational complexity can be reduced, because all the probability density functions are tied together, thus providing a good compromise between detailed acoustic modelling and trainability.

However, standard ASR systems still do not provide convincing results under changeable environmental conditions. Most current commercial speech recognition technologies still work using either a restricted lexicon (i.e. digits or a definite number of commands) or a semantically restricted task (i.e. database information retrieval, tourist information, flight information, hotel services, etc.). Extensions to more complex tasks and/or vocabulary still have a reputation for poor quality and are thus viewed with scepticism by both potential users and customers.

Because of the limitations of HMM-based speech recognition systems, research has had to progress in a number of different directions. Rather than adopting an overall approach to tackling problems, they have generally been dealt with individually. Regarding robust speech recognition, the spectral variability of speech signals has been studied using different methods, such as variable frame rate (VFR) search analysis of speech. Model adaptation has also been on scope or, more specifically, speaker adaptation and vocal tract normalization (VTN).

Language modelling research has played a significant role in improving recognition performance in continuous speech recognition. However, another problem that faces standard speech recognition is the dependency of the models obtained regarding to the speakers (or database) used for training. Speaker adaptation can be used to overcome this drawback, as it performs well as a solution to certain tasks. However, its benefits are not entirely clear for speaker-independent tasks, because the adaptation costs are higher.

Of all the active fields of research in speech recognition, we will focus our attention on those closest to the approach presented in this chapter.

2.2 Modelling temporal evolution using temporal and trajectory models

At the outset of speech recognition research, the application of statistical methods, i.e. Markov Models, proved to be clearly advantageous. First-order Markov Models are sufficiently flexible to accommodate variations in probability along an utterance and, at the same time, simple enough to lend themselves to mathematically rigorous optimization and the deployment of search strategies. However, as the Markov property is an artificial constraint forced upon a model of the temporal speech utterance, it was expected that some speech characteristics would not be correctly modelled.

Several approaches for modelling temporal evolution have been proposed in the past. Temporal models have been used to optimally change the duration and temporal structure of words. Experiments showed that first-order Markov chains do not model expected local duration effectively. Thus, different approaches for a more explicit modelling of duration led to an improvement in performance.

Some approaches started by directly introducing continuously variable duration into the HMM. In (Russell & Cook, 1987) and (Bonafonte et al., 1996), each HMM state is expanded to a sub-HMM (ESHMM) that shares the same emission probability density and performs the correct state duration distribution using its own topology and transition probability. To reduce the loss of efficiency introduced by the ESHMM, a post-processor duration model can be implemented (Wu et al., 2005) using the output of a Viterbi algorithm and ranking the proposed paths through the use of better models for state duration. However, incorporating explicit duration models into the HMM also breaks up some of conventional Markov assumptions. When HMM geometric distribution is replaced with an explicitly defined one, Baum-Welch and Viterbi algorithms are no longer directly applicable.

In (Bonafonte et al., 1993), Hidden Semi-Markov models (HSMMs) are proposed as a framework for a more explicit modelling of duration. In these models, the first-order Markov hypothesis is broken in the loop transitions. The main drawback of an HSMM, however, is an increase in the computational time by a factor of D, D being the maximum time allowed in each state. Hence, the Viterbi algorithm must be modified to cope with this higher complexity and to limit the computational increase.

An alternative to this is to model the occupancy of each HMM state by means of a Markov chain (Vidal et al., 2004). This occupancy is represented using a distribution function (DF). Thus, each state of the initial HMM is expanded by the DF estimated for that state.

In another approach to overcome the limitations of standard HMM framework, alternative trajectory models have been proposed that take advantage of frame correlation. Although these models can improve the speech recognition performance, they generally require an increase in model parameters and computational complexity.

In (Tokuda et al., 2003) a trajectory model is derived by reformulating the standard HMM whose state output vector includes static and dynamic feature parameters. This involves imposing the explicit relationship between the static and dynamic features. A similar technique is based on maximizing the models’ output probability under the constraints between static and dynamic features.

A smooth speech trajectory is also generated from an HMM by maximizing the likelihood subject to the constraints that exist between static and dynamic features. A parametric trajectory can be obtained using direct relationships between the vector time series for static and dynamic features, or from mixture distribution HMMs (Minami et al., 2003). This method chooses the target sequence of Gaussian distributions by selecting the best Gaussian distribution for each state during Viterbi decoding. Thus, the relationship between the cepstrum and the dynamic coefficients is now taken into account in the recognition phase, unlike in previous approaches.

2.3 Second-order models (HMM2)

HMM-based speech modelling assumes that the input signal can be split into segments, which are modelled as states of an underlying Markov chain, and that the waveform of each segment is a stationary random process. As previously mentioned, the sequence of states in an HMM is assumed to be a first-order Markov chain. This assumption is motivated by the existence of efficient, tractable algorithms for model estimation and recognition.

To overcome the drawbacks of regular HMMs regarding segment duration modelling and trajectory (frame correlation) modelling, some authors have proposed a new class of models in which the underlying state sequence is a second-order Markov chain (HMM2) (Mari et al., 1997). These models show better state occupancy modelling, at the cost of higher computational complexity. To overcome this disadvantage, an appropriate implementation of the re-estimation formulation is needed. Algorithms that yield an HMM only N_i times slower than an HMM1 can be obtained, N_i being the average input branching factor of the model.

Another approach to a second-order HMM is a mixture of temporal and frequency models (Weber et al., 2003). This solution consists of a primary (conventional) HMM that models the temporal properties of the signal, and a secondary HMM that models the speech signal's frequency properties. That is, while the primary HMM is performing the usual time warping and integration, the secondary HMM is responsible for extracting/modelling the possible feature dependencies, while also performing time warping and integration.

In these models, the emission probabilities of the temporal (primary) HMM are estimated through a secondary, state-specific HMM that works in the acoustic feature space. Such models present a more flexible modelling of the time/frequency structure of the speech signal, which results in better performance. Moreover, when such systems are working with spectral features, they are able to perform non-linear spectral warping by implementing a form of non-linear vocal-tract normalization.

To solve the increase in computational complexity associated with this solution, the Viterbi algorithm must be modified, which leads to a considerable computational increase.

The differences in performance between an HMM and HMM2 are not particularly remarkable when a post-processor step is introduced. In this post-processor step, durational constraints based on state occupancy are incorporated into conventional HMM-based recognition. However, in this case HMM2s are computationally better, while the complexity increase is similar in both cases.

2.4 Layered speech recognition

With regard to the integration of different information into the ASR architecture, and going one step further from the HMM2, several authors have proposed using layered HMM-based architectures (Demuynck et al., 2003).

Layered ASR systems fit all the knowledge levels commonly used in automatic speech recognition (acoustic, lexical and language information) in a final model. From these architectures, a modular framework can be suggested that allows a two-step (or multi-step) search process. The usual acoustic-phonetic modelling is divided into two (or more) different layers, one of which is closer to the voice signal for modelling acoustic and physical characteristics, whilst the other is closer to the phonetics of the sentence. The modelling accuracy and the ease with which acoustic and phonetic variability can be managed are thus expected to increase.

By splitting the recognition scheme into an acoustic lower layer and a language-based upper layer, the introduction of new functionalities may be consigned to the second layer. The goal is to develop models that are not limited by acoustic constraints (such as left-to-right restrictions). This also provides an open field for the introduction of new (and high-level) information with no loss of efficiency. Moreover, layered architectures can increase speaker independence if the upper layer is trained with a different set of recordings to that used for the acoustic layer, which approaches conditions similar to those faced in the recognition of unknown speakers.

In the following sections, two approaches for a double-layer architecture are presented and justified. Thanks to the advantages mentioned above, layered architectures are expected to bring standard HMM-based ASR systems up to date.

3. HMM State Scores Evolution Modelling

3.1 Justification

In standard HMM-based modelling, feature vectors only depend on the states that generated them. Dynamic features (generally first and second derivatives of the cepstral coefficients and derivatives of the energy) are used to represent context in HMM. However, they only consider a few subsequent vectors and do not represent long-term variations. Moreover, first-order Markov chains do not effectively model expected local duration. Furthermore, as seen above, incorporating explicit state duration models into the HMM breaks up some of conventional Markov assumptions. An alternative way of incorporating context into an HMM lies in taking a similar approach to the evolution of state scores as that used for well-known trajectory models. However, in this case a double-layer architecture is used.

In (Casar & Fonollosa, 2006a), a method is presented for incorporating context into an HMM by considering the state scores obtained by a phonetic-unit recognizer. These state scores are obtained from a Viterbi grammar-free decoding step that is added to the original HMM, which yields a new set of "expanded" HMMs. A similar approach was used by (Stemmer et al., 2003), who integrated the state scores of a phone recognizer into the HMM of a word recognizer, using state-dependent weighting factors.

3.2 Mathematical formalism

To better understand the method for implementing HMM state scores evolution modelling presented in (Casar & Fonollosa, 2006a), the formulation on which it relies must be introduced.

In a standard SCHMM, the density function $b_i\ (x_t)$ for the output of a feature vector x_t by state i at time t is computed as a sum over all codebook classes $m \in M$ (the number of mixture components):

$$b_i(x_t) = \sum_m c_{i,m} \cdot p(x_t \mid m, i) \approx \sum_m c_{i,m} \cdot p(x_t \mid m) \qquad (1)$$

where $p(x_t \mid m) = \mathcal{N}(x_t, \mu_m, \Sigma_m)$ denotes the Gaussian density function shared across all Markov models and $c_{i,m}$ are the weights for the k^{th} codeword that satisfy $\sum c_{i,m} = 1$.

As in (Stemmer et al., 2003) probability density functions can be considered that make it possible to integrate a large context $x_1^{t-1} = x_1, \ldots .\ x_{t-1}$ of feature vectors observed thus far into the HMM output densities.

If we tried to directly integrate this context into $b_i,$ this would result in a large increase of computational effort. Therefore, a new hidden random variable l (henceforth called *class label*) is introduced, which is a discrete representation of the feature vectors x_1^{t-1}. These class labels l can correspond to the units whose context is to be modelled, for instance phone symbols.

From now on, each state i not only chooses between the codebook classes $m \in M$ but also takes an independent decision for the class label l. The integration of l into the output density makes b_i dependent on the history x_1^{t-1}.

Thus, Equation (1) is expanded by defining the new output probability and integrating the context, as in (Stemmer et al., 2003):

$$b_i(x_t \mid x_1^{t-1}) = \sum_{m,l} p(x_t \mid l,m) \cdot P(l,m \mid i, x_1^{t-1}) \tag{2}$$

Moreover, as x_1^{t-1} is the same for all states i at time t there is no increase in the computational complexity of the algorithms for training and decoding.

However, the representation of $b_i\ (x_t \mid x_1^{t-1})$ needs additional simplifications if the number of parameters to be estimated is to be reduced.

Since the decisions of l and m are independent, we can use the following approximation:

$$P(l,m \mid i, x_1^{t-1}) = P(l \mid i, x_1^{t-1}) \cdot P(m \mid i, x_1^{t-1}) \tag{3}$$

and as m does not depend on $x_1^{\ t-1}$, $P(m \mid i, x_1^{t-1}) = P(m \mid i) = c_{i,m}$.

Thus, Equation (3) can be reformulated as:

$$P(l,m \mid i, x_1^{t-1}) = c_{i,m} \cdot P(l \mid i, x_1^{t-1}) \tag{4}$$

We can also split the second term $P(l \mid i,\ x_1^{t-1})$ into two parts in which i is considered separately from $x_1^{\ t-1}$ and by applying Bayes' rule:

$$P(l \mid i, x_1^{t-1}) = C_{l,i} \cdot P(l \mid x_1^{t-1}) \tag{5}$$

where $C_{l,i}$ is related to $P(l \mid i)$ by means of a variable proportionality term.

To summarize, we can express Equation (2) by splitting the separately considered contributions of m and l, thus:

$$b_i(x_t \mid x_1^{t-1}) \approx \left[\sum_m c_{i,m} \cdot p(x_t \mid m)\right] \cdot \left[\sum_l C_{l,i} \cdot P(l \mid x_1^{t-1}) p(x_t \mid l)\right] \tag{6}$$

where the first term corresponds to Equation (1).

In this case, we do not want to introduce the modelling of the context for each feature vector into the HMM output densities, but to create a new feature by modelling the context. A new probability term is defined:

$$b_i^{'}(x_t) = \sum_l C_{l,i} \cdot P(l \mid x_1^{t-1}) p(x_t \mid l) \tag{7}$$

Thus, when a regular $b_i(x_t)$ for each spectral feature and a $b_i{}'(x_t)$ for the phonetic-unit feature are combined, the joint output densities of the expanded set of models are equivalent to Equation (6).

We can express $P(l \mid x_1^t) = P(l \mid x_t, x_1^{t-1})$, and applying Bayes' rule to the second term of this expression:

$$P(l \mid x_1^t) = \frac{p(x_t \mid l, x_1^{t-1}) P(l \mid x_1^{t-1})}{p(x_t, x_1^{t-1})} \tag{8}$$

Given that class l is itself a discrete representation of feature vectors x_1^{t-1}, we can approximate $p(x_t \mid l, x_1^{t-1}) \approx p(x_t \mid l)$.

Likewise, $p(x_t,x_1^{t-1})$ is a constant in its evaluation across the different phonetic units so we can simplify Equation (8) to $P(l\,|\,x_1{}^t) = K\,p(x_t\,|\,l)\,P(l\,|\,x_1{}^{t-1})$ (with K as a constant).
Finally,

$$b_i^{'}(x_t) = \sum_l C_{l,i}^{'} \cdot P(l\,|\,x_1^t) \qquad (9)$$

The new terms in Equation (9) are obtained as follows: $C'_{l,i}$ is estimated during the Baum-Welch training of the expanded set of models, whilst $P(l\,|\,x_1{}^t)$ corresponds to the state scores output obtained by the Viterbi grammar-free decoding step. Thus, the output probability distribution of the models in the second layer can be estimated through a regular second training process.

3.3 Recognition system

The double-layer architecture proposed (Figure 1) divides the modelling process into two levels and trains a set of HMMs for each level. For the lower layer, a standard HMM-based scheme is used, which yields a set of regular acoustic models. From these models, a phonetic-unit recognizer performs a Viterbi grammar-free decoding step, which provides (at each instant t) the current most likely last state score for each unit.
This process can also be seen as a probabilistic segmentation of the speech signal, for which only the last state scores associated with the unit with the highest accumulated probability are kept.

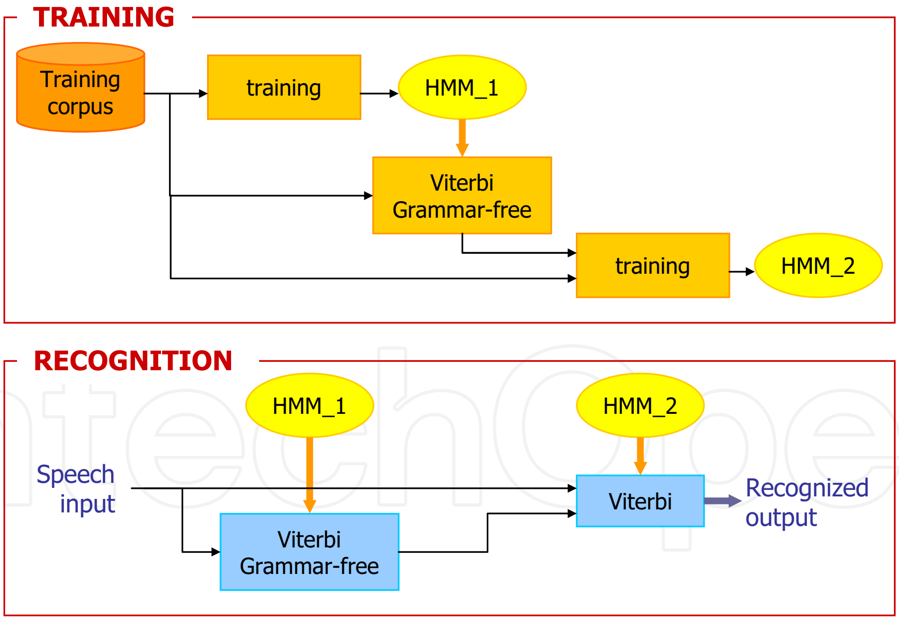


Figure 1. Double-layer ASR system with HMM state scores modelling

Let us consider, for instance, semidigit acoustic models for the first layer. In this case, labels l in the second layer represent the last states of each semidigit model. Thus, the output density value for each unit can be computed as the probability that the current state s$_t$ of a semidigit model is equal to *l*:

$$P(l\,|\,x_1^t) := P(s_t = l\,|\,x_1^t) \tag{10}$$

where $P(s_t = l\,|\,x_1^t)$ is calculated from the forward score:

$$P(s_t = l\,|\,x_1^t) = \frac{P(s_t = l, x_1^t)}{\sum_j P(s_t = j, x_1^t)} \tag{11}$$

The last state scores probability will be the new parameter to be added to the original set of features (spectral parameters). Henceforth, five features are considered for further training the joint output densities of the expanded set of HMM parameters. However, they are not independent features as the phonetic-unit feature models the evolution of the other features. A weighting factor *w* is also introduced, as in (Stemmer et al., 2003) to control the influence of the new parameter regarding the spectral features. A global, non-state-dependent weighting factor will be used.

In Table 1, digit chain recognition results obtained using this architecture are compared with the baseline results obtained using the regular RAMSES SCHMM system (Bonafonte et al., 1998). Results show a significant improvement in both sentence and word recognition rates.

Configuration		Sentence recognition rate	Word recognition rate	Relative reduction in WER
System	**w**			
Baseline	-	93.304 %	98.73 %	-
Layered	0.5	93.605 %	98.80 %	5.51 %
	0.2	93.699 %	98.81 %	6.3 %

Table 1. Recognition rates using expanded state-scores based HMM.

The relevance of choosing a suitable weighting factor *w* is reflected in the results. Different strategies can be followed for selecting *w*. In this case, as in Stemmer et al. (2003), an experimental weighting factor is selected and its performance verified using an independent database.

4. Path-based Layered Architectures

4.1 Justification

HMM-based speech recognition systems rely on the modelling of a set of states and transitions using the probability of the observations associated with each state. As these probabilities are considered independent in SCHMMs, the sequence of states leading to each recognized output remains unknown. Thus, another interesting approach for implementing the second layer of a double-layer architecture consists in training the appearance pattern instead of modelling the temporal evolution of the states scores.

In (Casar & Fonollosa, 2003b) the “path” followed by the signal is modelled; each final active state is taken as a step. Recognition is then associated with decoding the best matching path. This aids the recognition of acoustic units regardless of the fact that they may vary when they are uttered in different environments or by different speakers, or if they are affected by background noise.

Let us examine an example using phoneme HMMs, with three states each, which allows a maximum leap of 2 for intra-model state transitions, as in Figure 2.

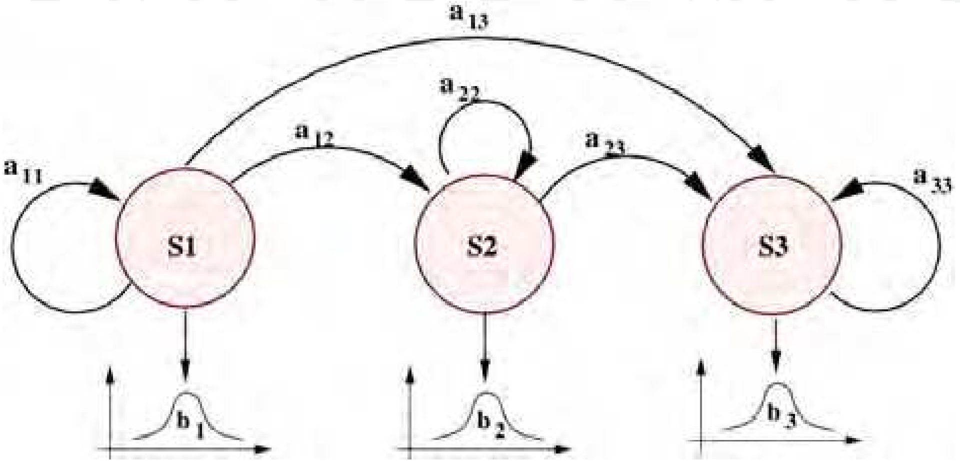


Figure 2. A three-state HMM with a maximum leap of 2 for intra-model state transitions

Thus, when a word is uttered, for example "zero", the speech signal is able to go through the different states of the models associated with each of the word's phonemes. The graph in Figure 3 represents the different "paths" that the speech signal can follow through these states at the decoding stage.

As the intra-model and inter-model state transitions allowed are also represented, by modelling this path we are also modelling the different durations of the utterance as local modifications of the path.

In a double-layer framework, the recognition architecture is broken down into two levels and performs a conventional acoustic modelling step in the first layer. The second layer is consigned to model the evolution followed by the speech signal. This evolution is defined as the path through the different states of the sub-word acoustic models defined in the first layer.

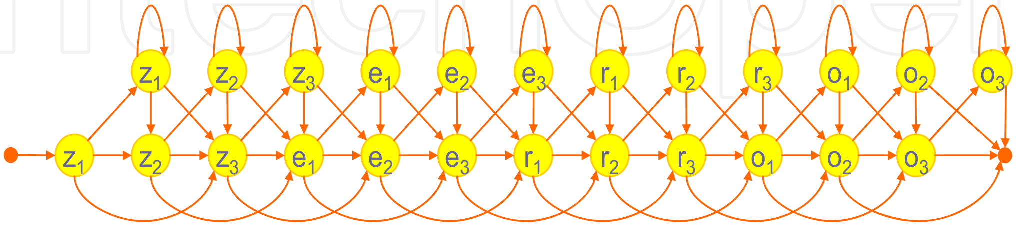


Figure 3. Example of paths that can be followed when the word "zero" is uttered

To model this path without taking into account the causal relationships between neighbour states, we can implement a transparent second layer. This is because we are doing a direct mapping of the acoustic probabilities of the activated states. However, while in traditional acoustic HMMs the Gaussian densities model the probability with which a state generates certain spectral parameters (acoustic or physical information), the new HMM generated in this second layer will give the probability of being at a certain point of the path followed by the speech signal.

If we take it one step further, the context of each state will be considered in the process of modelling the path. As shown in Figure 5, the speech signal will now be modelled by means of the different paths through the activated states, in which each path has its own associated probability. Thus, each state will be allowed to be part of different paths and to have different contexts. This increases the variability of the path and helps to model different utterances of the same speech symbol.

4.2 Mathematical formalism

In a semi-continuous HMM, a VQ codebook is used to map the continuous input feature vector x to o_k (the k^th codeword) by means of the distribution function $f(x\,|\,o_k)$. Therefore, we can use a discrete output probability distribution function (PDF) $b_j(k)$ for state j:

$$b_j(x) = \sum_{k=1}^{M} b_j(k) f(x \mid o_k) \qquad (12)$$

If Equation (12) is taken as the output density function of the models from the first layer, the input into the HMM of the second layer will be the vector of state probabilities given by the acoustic models of the first layer. Hence, a new set of semi-continuous output PDFs $b'_j\,(k)$ is defined for the second layer:

$$b_j^{'}(x) = \sum_{k=1}^{M} b_j^{'}(k) b_k(x) \qquad (13)$$

This equation can also be expressed in terms of a new distribution function $f'(x\,|\,b_k)$, where the output probability vectors b_k play the role usually carried out by o_k in the first level. In fact, by doing this we are defining a new codebook that covers the sub-word state-probability space by means of the distribution function $f'(x\,|\,b_k) = b_k\,(x)$.

The new weights $b'(x)$ will be obtained through a new Baum-Welch estimation in a second modelling step. New observation distributions for the second-layer HMM are trained using the same stochastic matrix as that of the original acoustic HMM.

In practice, as M and M' are large, Equations (12) and (13) are simplified using the most significant values of I and I'. Thus, it is possible to avoid certain recognition paths from being activated, and this can result in a different decoding when $I \neq I'$. This simplification also means that the preceding and following states to be activated for each state are pruned.

In the previous formulation, we model the path followed by the signal, taking each final active state as a step, but without studying the possible causal relationships between adjacent states. When context-dependent path-based modelling is implemented, the mapping of the models will be undertaken using windows centred in each state and that embrace one or more adjacent states, that is, the states that are most likely to have been visited before the current state and those that will most probably become future ones. Therefore, instead of taking the output probability vectors b_k of the first layer as o_k for the new distribution function $f'(x \mid b_k)$, we will work with a combination between the output probabilities of the adjacent states considered.

4.3 Path-based double-layer architectures

The main aims of the path-based double-layer architecture developed are twofold: firstly, to achieve a better modelling of speech units as regards their variation when they are uttered in a changeable environment, and secondly, to improve speaker independence by taking advantage of the double layer.

Two implementations are possible for the second layer, namely, the state context in the definition of the path can either be taken into account or ignored. The two schemes are presented below. Firstly, in the one-state width path-based modelling scheme, the state context is not considered, that is, the path followed by the signal is considered without taking into account the causal relationships between adjacent states. This context is subsequently introduced in the L-state width path-based modelling scheme. In this case, L-1 is the number of adjacent states considered as the significant context for each state of the path.

Path-based modelling without context

In Figure 4, a basic diagram of the proposal for a double-layer architecture that implements path-based modelling without context in the second layer is shown.

The first layer of both the training and recognition schemes is equivalent to a regular acoustic HMM-based system. The second layer consists in mapping the acoustic models obtained in the first layer into a state-probability-based HMM. In addition, a new codebook that covers the probability space is defined. This means that we are no longer working with spectral parameter distributions but with the probabilities for the whole set of possible states. Thus, we have moved from the signal space (covered by the spectrum) to the probability space (defined by the probability values of each of the states).

In traditional HMMs, Gaussian densities model the probability with which a state generates certain spectral parameters (acoustic information). The new HMMs generated by this second layer will give the probability of being at a certain point of the path followed by the speech signal.

In practice, this means that in acoustic HMMs the probability of reaching a certain state s_i of model m_i depends on the parameterization value of the four spectral features, which depend on physical and acoustic characteristics. For instance, the “z” in “zero” may vary considerably when it is uttered by two different speakers (in terms of acoustic and physical parameters). It is the task of the HMM parameter to achieve a correct modelling of these variations. However, if a very flexible model is trained to accept a wide range of different utterances in the acoustic segment, the power of discrimination between units will be lost.

When we are working with path-based HMM, we are directly modelling the probability of reaching state s_i of model m_i regardless of the acoustic features’ values. We use the new codebook to map the spectral feature to the new probability space. Thus, we decode the path followed by the speech signal in terms of the probabilities of each active state.

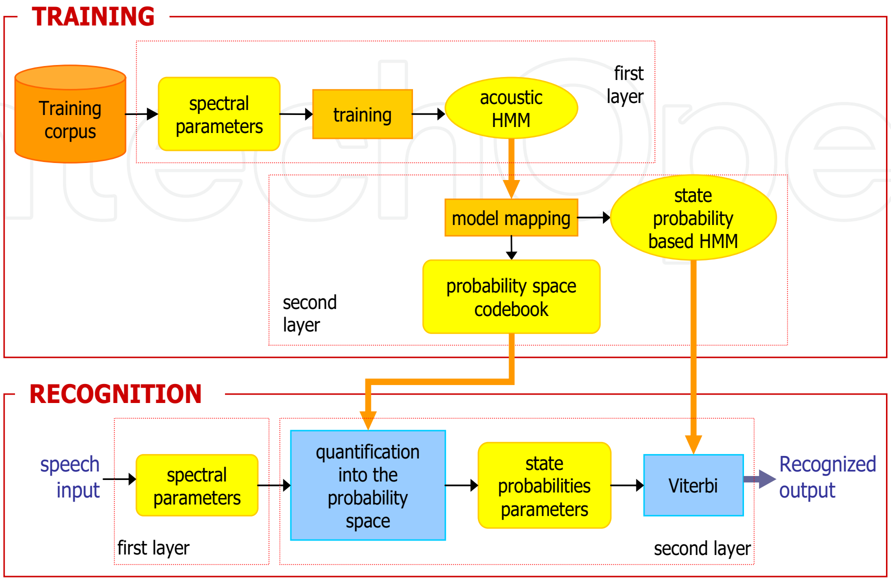


Figure 4. Basic diagram of a one-state width path-based double-layer ASR system

As the models are mapped to a space of dimension N (the total number of HMM states from the first layer), some pruning can be implemented by keeping only the N/2 most significant values (see Equation (13)). We are therefore constraining the possible states preceding and following each active state, which prevents some recognition paths from being activated. This solution will increase speed, and an improvement in recognition performance is also expected.

In fact, this can be seen as a similar strategy to CHMM Gaussian mixture pruning, in which each state is modelled using a mixture of private Gaussians and only the most likely mixtures are considered.

L-states width path-based modelling

From a mathematical point of view, speech signals can be characterized as a succession of Markov states, in which transitions between states and models are restricted by the topology of the models and the grammar. Hence, each state of every model is unique, even though the topology can allow multiple repetitions.

Thus, speech can be modelled by means of different *state successions* (or *paths*). Each path has its own associated probability, which allows one state to be part of different paths (see Figure 5). Furthermore, the context of each state becomes relevant, which brings about a higher variability in the possible paths that make up an utterance.

However, the maximum likelihood estimation criteria still apply. Thus, the path with maximum likelihood will be that configured by the succession of states that maximizes the joint probability (defined by the product of probabilities of each state in the path).

Theoretically, each path should be defined as the complete succession of states. However, as can be seen in Figure 5, the number of paths to consider can become too high (close to infinite) if the total number of previous and following states is considered for each state context.

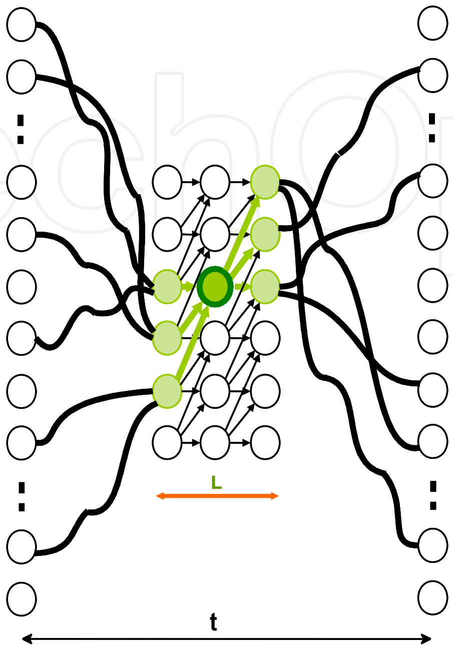


Figure 5. Different paths go into a certain state. Only those enclosed in a window of length L will be considered.

In practice, each state context is limited to a window of length L in order to allow generalization and to make implementation computationally feasible. That is, we will not deal with the whole path followed by the signal in the successive states defining a certain speech utterance. Instead, for each state that defines the path only the possible previous and subsequent L/2-1 states will be considered. Grammatical (phonetic) concordance along the path will likewise be verified.

By stretching this simplification to its limit, if a window of length L=1 were used we would be directly mapping the acoustic models into state-probability models in the same way as in the previous proposal (path-based modelling without context). This means that this second proposal can be seen as an extension of the previous one.

The training and recognition schemes for this architecture, in which L-states width path-based modelling in the second layer are implemented, is very similar to the previous one (for path-based modelling without context, see Figure 4). Only some modifications to the second layer of the training scheme are needed to take the L-states with context into account.

In this case, by mapping the acoustic models obtained in the first layer, a new codebook that covers the probability space is built. Furthermore, a table with all the possible state combinations is defined, which takes the aforementioned restrictions into account. If all state combinations are considered, the input speech is statistically defined and a new set of parameters is obtained. These parameters represent the probabilities of each sequence of states. A new set of state-probability-based models is built, which makes it possible to decode the path followed by the speech signal that uses the state probability parameters.

4.4 Results

If working with the first implementation, without considering the context for the path-based modelling in the second layer, it is assumed that there is a "transparent" second layer, as it is equivalent to a direct mapping of the acoustic probabilities. Performance would be expected to be equivalent to that of a baseline system (without the second layer being implemented). However, due to (and thanks to) the pruning implemented by keeping only the most significant N/2 values of the total number of HMM states (N), it is possible to prevent some recognition paths from being activated.

For the L-states width path-based modelling, the total number of possible state combinations represented in the new codebook is a result of considering the characteristics of the acoustic models in the lower layer (number of models and number of states) and the window length. Again, for the new representation of the input signal that uses the probability space codebook, only the most significant N/2 values will be kept.

In Table 2, digit recognition results obtained with these two architectures are compared against the baseline results obtained with a regular RAMSES SCHMM system (Bonafonte et al., 1998). Using one-width path-based modelling for the second layer, there is an improvement in the sentence and word recognition rate. This is achieved thanks to a positive weighting of the states with higher likelihood (implicit in the solution proposed) and the pruning of the preceding and following states to be activated for each state.

The results for the L-states width path-based modelling show a noticeable improvement in sentence and word recognition, but lower than that resulting from the first approach. This responds to the growth of the information to be modelled and the pruning performed, which induces a loss of information.

However, the general performance of the L-states width path-based implementation for the second layer is good and the flexibility of this approach would allow added value information to be introduced into the recognition. Recognition speed, which is slightly higher than with the first implementation, is also a point in its favour.

The gain obtained by these two approaches is also shown by means of the word error rate (WER). As the original recognition error rate of the baseline is low for the task under consideration, the perceptual relative reduction of the WER achieved is a good measure of the goodness of these solutions. Therefore, what would initially seem to be just a slight improvement in the (word/sentence) recognition rates can actually be considered a substantial gain in terms of perceptual error rate reduction.

Recognition system	**Sentence recognition rate**	**Word recognition rate**	**Relative reduction in WER**
Baseline	93.304 %	98.73 %	-
One-state width path-based double-layer	94.677 %	99.10 %	29.1%
L-states width path-based double-layer	93.717 %	98.98 %	19.7%

Table 2. Recognition rates using path-based double-layer recognition architectures

5. Discussion

The future of speech-related technologies is connected to the improvement of speech recognition quality. Until recently, speech recognition technologies and applications had assumed that there were certain limitations regarding vocabulary length, speaker independence, and environmental noise or acoustic events. In the future, however, ASR must deal with these restrictions and it must also be able to introduce other speech-related non-acoustic information that is available in speech signals.

Furthermore, HMM-based statistical modelling—the standard state-of-the-art ASR—has several time-domain limitations that are known to affect recognition performance. Context is usually represented by means of spectral dynamic features (namely, its first and second derivatives). However, they are generally limited to a few subsequent feature vectors and do not represent long-term variations.

To overcome all these drawbacks and to achieve a qualitative improvement in speech recognition, a change of paradigm from conventional speech recognizers has been proposed by several authors. Although some authors propose a move away from HMM-based recognition (or, at the very least, introducing hybrid solutions), we are adhering to Markov-based acoustic modelling as we believe its approach is still unbeatable. However, to overcome HMM-related limitations certain innovative solutions are required.

Throughout this chapter we have pointed out different approaches for improving standard HMM-based ASR systems. The main solutions for modelling temporal evolution and speech trajectory have been introduced, together with some ideas on how second-order HMMs deal with the same problems. These models provide an improvement in most cases, but they also require major modifications in the decoding of algorithms. Generally, there is also a considerable increase in complexity, even if this is compensated for by a moderate gain.

Layered architectures have been presented, and special attention has been paid to the implementation of the second layer using extended HMMs. Two implementations for this second layer have been described in detail. The first relies on modelling the temporal evolution of acoustic HMM state scores. In the second one, the evolution of the acoustic HMM is modelled by the speech utterance as a new way of modelling state transitions. This can be done in two ways, namely, by taking into account or ignoring the context of each active state while the "path" followed by the speech signal through the HMM states is being modelled. Again, speech recognition performance improves that of a conventional HMM-based speech recognition system, but at the cost of increased complexity.

Although current research solutions should not be unduly concerned by the computational cost (due to the constant increase in the processing capacity of computers), it is important to keep their implementation in commercial applications in mind. Therefore, a great deal of work remains if layered architectures are to be generalized for large vocabulary applications that keep complexity down to a moderate level.

Efforts should be made in the field of research for defining and testing innovative approaches to implementing layered architectures. Although keeping an HMM-based scheme for the different layers reduces the overall complexity, a change in paradigm may help to bring about significant improvements.

6. References

Bonafonte, A.; Ros, X. & Mariño, J.B. (1993). An efficient algorithm to find the best state sequence in HSMM. *Proceedings of the 3th European Conference on Speech Communication and Technology (EUROSPEECH93)*, 1993.

Bonafonte, A.; Vidal, J. & Nogueiras, A. (1996). Duration modelling with Expanded HMM Applied to Speech Recognition. *Proceedings of International Conference in Spoken Language Processing (ICSLP96)*, Volume2, pp:1097-1100, ISBN:0-7803-3555-4. Philadelphia (USA), October, 1996.

Bonafonte, A.; Mariño, J.B.; Nogueiras, A. & Fonollosa, J.A.R. (1998). RAMSES: el sistema de reconocimiento del habla continua y gran vocabulario desarrollado por la UPC. *VIII Jornadas de Telecom I+D (TELECOM I+D'98),* Madrid, Spain, 1998.

Casar, M. & Fonollosa, J.A.R. (2006a). Analysis of HMM temporal evolution for ASR Utterance verification. *Proceedings of the Ninth International Conference on Spoken Language Processing (Interspeech2006-ICSLP),* pp:613-616, ISSN:1990-9772. Pittsburgh, USA, September 2006.

Casar, M. & Fonollosa, J.A.R. (2006b). A path-based layered architecture using HMM for automatic speech recognition. *Proceedings of the 14th European Signal Processing Conference (EUSIPCO2006)*. Firenze, Italia. September 2006.

Demuynck, K.; Kaureys, T., Van Compernolle, D. & Van Hamme, H. (2003). FLAVOR: a flexible architecture for LVCSR. *Proceedings of the 8th European Conference on Speech Communication and Technology (EUROSPEECH2003),* pp:1973-1976. Genova, 2003.

Furui, S. & Sandhi, M. (1992). A*dvances in Speech Signal Processing,* Marcel Dekker, Inc. ISBN 0-8247-8540-1, 1st edition, 1992, New York (USA).

Huang, X.; Acero, A. & Hon, H.W. (2001). *Spoken Language Processing,* Prentice Hall PTR, ISBN 0-13-022616-5, 1st edition, 2001, New Jersey (USA).

Juang, B.H.; Chou, W. & Lee, C.H. (1997). Minimum Classification Error rate methods for speech recognition*, IEEE Transaction on Speech and Audio Processing,* Vol. 5, No. 3, (May, 1997) pp: 257-265, ISSN: 1063-6676.

Mari, J.-F.; Haton, J.-P. & Kriouile, A. (1997). Automatic word recognition based on Second-Order Hidden Markov Models*, IEEE Transaction on Speech and Audio Processing,* Vol. 5, No. 1, (January, 1997) pp: 22-25, ISSN:1063-6676.

Nadeu, C.; Macho, D. & Hernando, J. (2001). Time and frequency filtering of filter-bank energies for robust HMM speech recognition. *Speech Communication,* Vol. 34, Issues 1-2 (April, 2001) pp: 93-114, ISSN:0167-6393.

Rabiner, L.R. (1989). A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE,* No. 2, Vol. 77, (March, 1989), pp: 257-289, ISSN:0018-9219.

Rabiner, L. & Juang, B.H. (1993). *Fundamentals of Speech Recognition,* Prentice Hall PTR, ISBN:0-13-015157-2. NY, USA, 1993.

Russell, M.J. & Cook, A.E. (1987). Explicit modelling of state occupancy in Hidden Markov Models for automatic speech recognition. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'85),* Volume 10, pp:5-8. April, 1987.

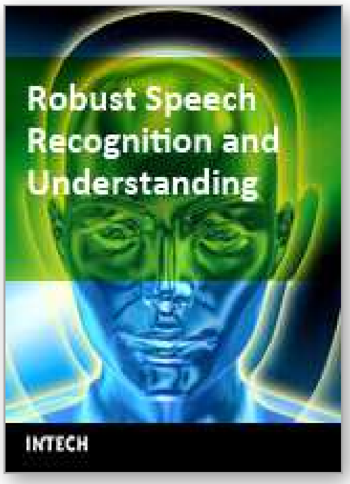
Stemmer, G.; Zeissler, V.; Hacker,C.; Nöth, E. & Niemann,H. (2003). Context-dependent output densities for Hidden Markov Models in speech recognition. *Proceedings of the 8th European Conference on Speech Communication and Technology (EUROSPEECH2003),* pp:969-972. Genova, 2003.

Tokuda, K.; Zen, H. & Kitamura,T. (2003). Trajectory modelling based on HMMs with the explicit relationship between static and dynamic features. *Proceedings of the 8th European Conference on Speech Communication and Technology (EUROSPEECH2003),* pp:865-868. Genova, 2003.

Vidal,J.; Bonafonte, A. & Fernández,N. (2004) Rational characteristic functions and Markov Chains: application to modelling probability density functions, *Signal Processing*, No. 12, Vol. 84 (December, 2004) pp: 2287-2296, ISSN: 0165-1684.

Weber, K.; Ikbal, S.; Bengio, S. & Bourlard, H. (2003). Robust speech recognition and feature extraction using HMM2. *Computer, Speech and Language,* Vol. 17, Issues 2-3 (April-July 2003) pp: 195-211, ISSN:0885-2308.

Wu,Y.-J.; Kawai,H.; Ni,J. & Wang,R.-H. (2005) Discriminative training and explicit duration modelling for HMM-based automatic segmentation, *Speech Communication*, Vol. 47, Issue 4 (December, 2005) pp: 397-410, ISSN: 0167-6393.



Robust Speech Recognition and Understanding
Edited by Michael Grimm and Kristian Kroschel

ISBN 978-3-902613-08-0
Hard cover, 460 pages
Publisher I-Tech Education and Publishing
Published online 01, June, 2007
Published in print edition June, 2007

This book on Robust Speech Recognition and Understanding brings together many different aspects of the current research on automatic speech recognition and language understanding. The first four chapters address the task of voice activity detection which is considered an important issue for all speech recognition systems. The next chapters give several extensions to state-of-the-art HMM methods. Furthermore, a number of chapters particularly address the task of robust ASR under noisy conditions. Two chapters on the automatic recognition of a speaker's emotional state highlight the importance of natural speech understanding and interpretation in voice-driven systems. The last chapters of the book address the application of conversational systems on robots, as well as the autonomous acquisition of vocalization skills.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821