8

Overcoming HMM Time and Parameter Independence Assumptions for ASR

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

1. Introduction

Understanding continuous speech uttered by a random speaker in a random language and in a variable environment is a difficult problem for a machine. To take into account the context implies a broad knowledge of the world, and this has been the main source of difficulty in speech related research. Only by simplifying the problem – 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 – has automatic speech recognition been possible.

For modeling temporal dependencies or multi-modal distributions of "real-world" tasks, Hidden Markov Models (HMMs) are one of the most commonly used statistical models. Because of this, HMMs have become the standard solution for modeling acoustic information in the speech signal and thus for most current speech recognition systems.

When putting HMMs into practice, however, there are some assumptions that make evaluation, learning and decoding feasible. Even if effective, these assumptions are known to be poor. Therefore, the development of new acoustic models that overcome traditional HMM restrictions is an active field of research in Automatic Speech Recognition (ASR).

For instance, the independence and conditional-independence assumptions encoded in the acoustic models are not correct, potentially degrading classification performance. Adding dependencies through expert knowledge and hand tuning can improve models, but it is often not clear which dependencies should be included.

Different approaches for overcoming HMM restrictions and for modeling time-domain dependencies will be presented in this chapter. For instance, an algorithm to find the beststate sequence of HSMMs (Hidden Semi-Markov Models) allows a more explicit modeling of context. Durations and trajectory modeling have also been on stage, leading to more recent work on the temporal evolution of the acoustic models. Augmented statistical models have been proposed by several authors as a systematic technique for modeling HMM additional dependencies, allowing the representation of highly complex distributions. These dependencies are thus incorporated in a systematic fashion, even if the price for this flexibility is high.

Focusing on time and parameter independence assumptions, we will explain a method for introducing N-gram based augmented statistical models in detail. Two approaches are presented: the first one consists of overcoming the parameter independence assumption by modeling the dependence between the different acoustic parameters and mapping the input

Source: Speech Recognition, Technologies and Applications, Book edited by: France Mihelič and Janez Žibert,
ISBN 978-953-7619-29-9, pp. 550, November 2008, I-Tech, Vienna, Austria

signal to the new probability space. The second proposal attempts to overcome the time independence assumption by modeling the temporal dependencies of each acoustic parameter.
The main conclusions obtained from analyzing the proposals presented will be summarized at the end, together with a brief dissertation about general guidelines for further work in this field.

2. ASR using HMM

2.1 Standard systems

Standard ASR systems rely on a set of so-called acoustic models that link the observed features of the voice signal with the expected phonetic of the hypothesis sentence. The most common implementation of this process is probabilistic, that is, Hidden Markov Models, or HMMs (Rabiner, 1989; Huang et al., 2001).
A Markov Model is a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event. This characteristic is defined as the Markov property. An HMM is a collection of states that fulfills 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 determined by the direct acoustic vector plus its dynamic features (namely, its first and second derivatives), plus the energy of the spectrum. The dynamic features are the way of representing the context in HMMs, but generally they are 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 or the Viterbi algorithm. In recent years, the discriminant training method and the criteria of Minimum Classification Error (MCE), based on the Generalized Probabilistic Descent (GPD) framework, has been successful in training HMMs for speech recognition (Juang et al., 1997). For decoding, both Viterbi and Baum-Welch algorithms have been implemented with similar results, but a better computational behavior is observed with the former.
The first implementations of HMM for ASR were based on discrete HMMs (or DHMM). DHMMs imply the need of a quantization procedure to map observation vectors from a continuous space to the discrete space of the statistical models. There is, of course, a quantization error inherent in this process that can be eliminated if continuous HMMs (or CHMMs) are used. In CHMMs, a different form of output probability functions is needed. Multivariate Gaussian mixture density functions are a clear first 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 it will need to accommodate the M-mixture observation densities used.
In many implementations, the gap between discrete and continuous mixture density HMMs has been bridged with some minor assumptions. For instance, in tied-mixture HMMs, the mixture density functions are tied together across all the models to form a set of shared kernels.
Another solution is the use of semi-continuous HMMs (or SCHMM), where a VQ codebook is used to map the continuous input feature vector $\mathbf{x}$ onto o, as in discrete HMMs. However,

in this case, the output probabilities are no longer directly used (as in DHMM), but are rather combined with the VQ density functions. That is, the discrete model-dependent weighting coefficients are combined with the continuous codebook probability density functions. Semi-continuous models can also be seen as equivalent to M-mixture continuous HMMs with all of the continuous output probability density functions shared among all Markov states. Therefore, SCHMMs do maintain the modeling ability of large-mixture density functions. In addition, the number of free parameters and the computational complexity can be reduced, because all of the probability density functions are tied together, thus providing a good compromise between detailed acoustic modeling and trainability.
However, standard ASR systems still don't provide convincing results when environmental conditions are changeable. Most of the actual 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 bad reputation in terms of quality, which entails the mistrust of both potential users and customers.
Due to the limitations found in HMM-based speech recognition systems, research has progressed in numerous directions. Among all the active fields of research in speech recognition, we will point out only those similar to the approach presented in this chapter.

2.2 Semi-continuous HMM

Semi-continuous hidden Markov models can be considered as a special form of continuous mixture HMMs, with the continuous output of probability density functions sharing a mixture Gaussian density codebook (see (Huang & Jack, 1989)). The semi-continuous output probability density function is represented by a combination of the discrete output probabilities of the model and the continuous Gaussian density functions of the codebook. Thus, the amount of training data required, as well as the computational complexity of the SCHMM, can be largely reduced in comparison to continuous mixture HMM. Thus, SCHMMs become the perfect choice for training small vocabulary and/or for low resource applications.
Moreover, the ease of combining and mutually optimizing the parameters of the codebook and HMM leads to a unified modeling approach. In addition, the recognition accuracy of semi-continuous HMMs is comparable to that of both discrete HMMs and continuous HMMs under some conditions, which include considering the same number of Gaussian mixtures for all techniques and keeping this number low. It is not a coincidence that these conditions apply to the applications in which we are interested: real-time and/or low-resource scenarios.

2.3 Time independence and parameter independence assumptions

In HMMs, there are some assumptions that make evaluation, learning and decoding feasible. One of them is the Markov assumption for the Markov chain (Huang et al., 2001), which states that the probability of a state s_t depends only on the previous state s_{t-1}. Also, when working with different parameters to represent the speech signal, we rely on the parameter independence assumption. This assumption states that the acoustical parameters modeled by HMMs are independent, and so are the output-symbol probabilities emitted.

However, in many cases, the independence and conditional-independence assumptions encoded in these latent-variable models are not correct, potentially degrading classification performance. Adding dependencies through expert-knowledge and hand-tuning improved models can be done, but it is often not clear which dependencies should be included.

For modeling dependencies between features, Gaussian mixture distribution-based techniques are very common. The parametric modeling of cepstral features with full covariance Gaussians using the ML principle is well-known and has led to good performance. However, although standard cepstral features augmented with dynamics information performs well in practice, some authors have questioned its theoretical basis from a discriminant analysis point of view (Saon et al., 2000). Thus, work has been done to extend LDA methods to HDA (Heteroscedastic Discriminant Analysis) (Saon et al., 2000) or maximum likelihood rotations such as LDA+MLLT. However, these techniques are expensive with real-time and/or low resource applications.

For modeling time-domain dependencies, several approaches have focused on studying the temporal evolution of the speech signal to optimally change the duration and temporal structure of words, known as duration modeling (Pylkkönen & Kurimo, 2003). However, incorporating explicit duration models into the HMM structure also breaks some of conventional Markov assumptions: when the HMM geometric distribution is replaced with an explicitly defined one, Baum-Welch and Viterbi algorithms are no longer directly applicable.

Thus, in Bonafonte et al. (1993), Hidden Semi-Markov Models (or HSMMs) were proposed as a framework for a more explicit modeling of duration. In these models, the first order Markov hypothesis is broken in the loop transitions. Then, an algorithm to find the best state sequence in the HSMM was defined, aiming for a more explicit modeling of context.

In another approach to overcome the temporal limitations of the standard HMM framework, alternative trajectory modeling (Takahashi, 1993) has been proposed, taking advantage of frame correlation. The models obtained can improve speech recognition performance, but they generally require a demoralizing increase in model parameters and computational complexity.

A smooth speech trajectory is also generated by HMMs through maximization of the models' output probability under the constraints between static and dynamic features, leading to more recent work on the temporal evolution of the acoustic models (Casar & Fonollosa, 2006b).

Therefore, a natural next step, given this previous research, is to work on a framework for dealing with temporal and parameter dependencies while still working with regular HMMs, which can be done by using augmented HMMs.

Augmented statistical models have been proposed previously as a systematic technique for modeling additional dependencies in HMMs, allowing the representation of highly complex distributions. Additional dependencies are thus incorporated in a systematic fashion. However, the price for flexibility is high, even when working with more computationally-friendly purposes (Layton & Gales, 2006).

In an effort to model the temporal properties of the speech signal, class labels modeling (Stemmer et al., 2003) has been studied in a double layer speech recognition framework (Casar & Fonollosa, 2006a). The main idea was to deal with acoustic and temporal information in two different steps. However, the complexity of a double decoding procedure was not offset by the results obtained. But temporal dependence modeling is still a challenge, and a less complex scheme needed to be developed.

The approach presented in this chapter consists of creating an augmented set of models. However, instead of modeling utterance likelihoods or the posterior probabilities of class labels, we focus on temporal and inter-parameter dependence.

3. Using N-grams for modeling dependencies

To better analyze the influence of temporal and parameter dependencies in recognition performance, both dependencies can be modeled in an independent fashion. Thus, a new set of acoustic models will be built for each case without losing the scope of regular HMMs.
For both cases, the most frequent combinations of features from the MFCC-based parameterized signal will be selected following either temporal or parameter dependence criteria. Language modeling techniques (i.e. by means of the CMU Statistical Language Modeling (SLM) Toolkit[1]) should be used for performing this selection. In this way, a new probability space can be defined, to which the input signal will be mapped, defining a new set of features.
Going into the mathematical formalism, we start by recalling how, in standard semi-continuous HMMs (SCHMMs), 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 $m \in M$ (Huang et al., 2001):

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

In our case, new weights should be estimated, as there are more features (inter-parameter dependencies or temporal dependencies) to cover the new probability space. Also, the posterior probabilities $p(x_t \mid m)$ will be modified, as some independence restrictions will no longer apply.
From this new set of features, a regular SCHMM-based training will be performed, leading to a new set of augmented statistical models.

3.1 Modeling inter-parameter dependence

In most HMM-based ASR systems, acoustic parameters are supposed to be independent from each other. However, this is no more than a practical assumption, as successive derivatives are by definition related to the parameters from which they are derived. Therefore, we can model the dependence between the different acoustic parameters, and thus overcome the parameter independence assumption.
Let us assume that we work with four MFCC features: cepstrum (f_0), its first and second derivatives (f_1,f_2) and the first derivative of the energy (f_3). We can express the joint output probability of these four features by applying Bayes' rule:

$$P(f_0, f_1, f_2, f_3) = P(f_0)P(f_1 \mid f_0)P(f_2 \mid f_1, f_0)P(f_3 \mid f_2, f_1, f_0) \tag{2}$$

where f_i corresponds to each of the acoustic features used to characterize the speech signal. Assuming parameter independence, HMM theory expresses equation (2) as:

$$P(f_0, f_1, f_2, f_3) = P(f_0)P(f_1)P(f_2)P(f_3) \tag{3}$$

[1] See http://www.speech.cs.cmu.edu

If parameter independence is to be overcome, some middle ground has to be found between equations (2) and (3). Thus, instead of using all dependencies to express the joint output probability, only the most relevant dependence relations between features are kept. For the spectral features, we take into account the implicit temporal relations between the features. For the energy, experimental results show in a more relevant dependence on the first spectral derivative than to the rest.
Thus, equation (2) is finally expressed as:

$$P(f_0, f_1, f_2, f_3) = P(f_0)P(f_1 \mid f_0)P(f_2 \mid f_1, f_0)P(f_3 \mid f_1) \tag{4}$$

In practice, not all of the combinations of parameters will be used for modeling each parameter dependency for each $P(f_i)$, but only the most frequent ones. Taking into account the parameter dependence restrictions proposed, a basic N-gram analysis of the dependencies in the training corpus is performed, defining those most frequent combinations of acoustic parameterization labels for each spectral feature. That is, we will consider dependence between the most frequent parameter combinations for each feature (considering trigrams and bigrams), and assume independence for the rest.
The input signal will be mapped to the new probability space. Recalling equation (1), we can redefine the output probability of state i at time t for each of the features used as $P_i(f_k)$, where f_k corresponds to each of the acoustic features used to characterize the speech signal.
Starting with the first acoustic feature, the cepstrum, the new output probability is defined as a sum over all codebook classes $m \in M$ of the new posterior probability function weighted by the original weights $c_{i,m}$ for the acoustic feature f_0. That is:

$$P_i(f_0) = \sum_m c_{i,m}^0 \cdot p(f_0 \mid m) \tag{5}$$

For the second acoustic feature, the first derivative of the cepstrum (f_1), the new output probability is defined as:

$$P_i(f_1) = \sum_m c_{i,m,\hat{m}_0}^1 \cdot p(f_1 \mid m) \tag{6}$$

The new weights in this output probability are defined according to N-gram-based feature combinations, taking advantage of the bi-gram $"\hat{m}_0, m"$, where $\hat{m}_0$ is the likeliest class for feature f_0 at each state i and time t considered in the sum of probabilities. It is defined as:

$$\hat{m}_0 = \arg\max_m p(f_0 \mid m) \tag{7}$$

When the bi-gram $"\hat{m}_0, m"$ is not defined, and $c_{i,m,\hat{m}_0}^1 = c_{i,m}^1$, which are the original weights for feature f_1.
For the third acoustic feature, the second derivative of the cepstrum (f_2), the new output probability is defined as:

$$P_i(f_2) = \sum_m c_{i,m,\hat{m}_0,\hat{m}_1}^2 \cdot p(f_2 \mid m) \tag{8}$$

Now the new weights are defined according to N-gram-based feature combinations as $c_{i,m,\hat{m}_0,\hat{m}_1}^2$. Extrapolating equation (7):

$$\hat{m}_k = \arg\max_m p\left(f_k \mid m\right) \tag{9}$$

Now, if the tri-gram "$\hat{m}_1, \hat{m}_0, m$" is not defined, $c^2_{i,m,\hat{m}_0,\hat{m}_1} = c^2_{i,m,\hat{m}_0}$ if the bi-gram "$\hat{m}_0, m$" applies, or $c^2_{i,m,\hat{m}_0,\hat{m}_1} = c^2_{i,m}$ otherwise.

Finally, for the energy:

$$P_i\left(f_3\right) = \sum_m c^3_{i,m,\hat{m}_1} \cdot p\left(f_3 \mid m\right) \tag{10}$$

where the new weights are defined according to the bi-grams "$\hat{m}_1, m$". If this bi-gram is not defined, again the original weights $c^3_{i,m}$ apply.

From these new output probabilities, a new set of SCHMMs can be obtained using a Baum-Welch training and used for decoding, following the traditional scheme.

3.2 Modeling temporal dependencies

Generally, HMM-based ASR systems model temporal dependencies between different frames by means of the successive derivatives of the acoustic features. However, a more explicit modeling of the time domain information seems relevant for improving recognition accuracy.

The observation probability distributions used in HMMs assume that successive information s_1 ... s_t within a state i can be considered independent. This is what is generally known as the Markov assumption for the Markov chain, and it is expressed as:

$$P\left(s_t \mid s_1^{t-1}\right) = P\left(s_t \mid s_{t-1}\right) \tag{11}$$

where s_1^{t-1} represents the state sequence $s_1, s_2, \ldots, s_{t-1}$.

Taking into account a state sequence of length N, equation (11) can be reformulated to:

$$P\left(s_t \mid s_1^{t-1}\right) = P\left(s_t \mid s_{t-N} \ldots s_{t-1}\right) \tag{12}$$

For simplicity, not all of the sequence of observations is taken into account, but only the two previous ones for each observation s_t, working with the 3-gram s_{t-2}, s_{t-1}, s_t. Then, equation (12) can be expressed as:

$$P\left(s_t \mid s_1^{t-1}\right) = P\left(s_t \mid s_{t-2}, s_{t-1}\right) \tag{13}$$

Applying independence among features (recall equation (3)), the output probability of each HMM feature will be expressed as:

$$P\left(f_i\right) = P\left(f_i \mid f_{i_{t-2}}, f_{i_{t-1}}\right) \tag{14}$$

Again, the most frequent combinations of acoustic parameterization labels can be defined, and a set of augmented acoustic models can be trained. The output probability (from equation (1)) of state i at time t for each feature k will be rewritten following the same line of argument as in previous sections (see section 3.1, and equations (5)-(9)).

Now:

$$P_i(f_k) = \sum_m c^k_{i,m,\hat{m}_{k,t-1},\hat{m}_{k,t-2}} \cdot p(f_k \mid m) \tag{15}$$

with

$$\hat{m}_{k,t-\mathrm{i}} = \arg\max_m p(f_k \mid m, t-i) \tag{16}$$

Notice that if the trigram $\hat{m}_{k,t-2}, \hat{m}_{k,t-1}, m$ does not exist, the bigram or unigram case will be used.

4. Some experiments and results

4.1 Methods and tools

For the experiments performed to test these approaches, the semi-continuous HMM -based speech recognition system RAMSES (Bonafonte et al., 1998) was used as reference ASR scheme, and it is also used in this chapter as baseline for comparison purposes.

When working with connected digit recognition, 40 semidigit models were trained for the first set of acoustic models, with the addition of one noisy model for each digit, each modeled with 10 states. Silence and filler models were also used, each modeled with 8 states. When working with continuous speech recognition, demiphones models were used. For the first set of acoustic models, each phonetic unit was modeled by several 4-state left-to-right models, each of them modeling different contexts. In the second (augmented) set of HMMs, each phonetic unit was modeled by several models that modeled different temporal dependencies, also using 4-state left-to-right models.

Connected digits recognition was used as the first working task for testing speech recognition performance, as it is still a useful practical application. Next, a restricted large vocabulary task was tested in order to evaluate the utility of the approach for today's commercial systems.

Different databases were used: the Spanish corpus of the SpeechDat and SpeechDatII projects[2], and an independent database obtained from a real telephone voice recognition application, known as DigitVox, were used for the experiments related to connected digits recognition. The Spanish Parliament dataset (PARL) of the TC-STAR project[3] was used for testing the performance of the models for continuous speech recognition.

4.2 Experiments modeling parameter dependencies

In the first set of experiments, we modeled parameter dependencies. The different configurations used are defined by the number of N-grams used for modeling the dependencies between parameters for each new feature. In the present case, no dependencies are considered for the cepstral feature, 2-grams are considered for the first cepstral feature and for the energy, and 2 and 3-grams for the second cepstral derivative.

---

[2] A.Moreno, R.Winksky, 'Spanish fixed network speech corpus' *SpeechDat Project*. LRE-63314.

[3] TC-STAR: Technology and corpora for speech to speech translation, www.tc-star.org

As explained in section 3, as we cannot estimate all the theoretical acoustic parameter combinations, we define those **N** most frequent combinations of parameterization labels for each spectral feature. A low **N** means that only some combinations were modeled, maintaining a low dimension signal space for quantization. On the other hand, increasing **N** more dependencies are modeled at the risk of working with an excessive number of centroids to map the speech signal.
Different configurations were tested. Each configuration is represented by a 4-digit string with the different values of **N** used for each feature. The total number of code words to represent each feature is the original acoustic codebook dimension corresponding to this feature plus the number of N-grams used. The different combinations that result in the configurations chosen were selected after several series of experiments, defined to either optimize recognition results or to simplify the number of N-grams used.
In Table 1, we present the best results obtained for connected digit recognition experiments. Results are expressed according to SRR (Sentence Recognition Rate) and WER (Word Error Rate) to measure the performance.

Database	Configuration	SRR	WER
SpeechDat	Baseline	90.51	2.65
	-/2000/2000,2000/20000	91.04	2.52
DigitVox	Baseline	93.30	1.27
	-/2000/2000,2000/2000	03.71	1.17

Table 1. Connected digit recognition rates modeling inter-parameter dependencies

We can see an important improvement in speech recognition for this task using the SpeechDat dataset, with a relative WER decrease of nearly a 5%. When using the DigitVox dataset, this improvement is slightly higher, with a relative WER decrease of 7.788%. Because both datasets are independent from the training datasets, we didn't expect adaptation of the solution to the training corpus.

4.3 Experiments modeling temporal dependencies

When modeling temporal dependencies, each new HMM feature models the temporal dependencies of the original acoustic features. Again, the different configurations are represented by a 4-digit string (henceforth **N**), with the number of N-grams used in equation (15) for modeling each acoustic parameter. In contrast to inter-parameter dependence modeling, a wider range of **N** leads to an increase in recognition accuracy. Thus, this is a more flexible solution, where we can chose between optimizing the accuracy and working with reasonable codebook size (close to the state-of-the art codebooks when working with standard implementations) while still improving the recognition performance.
First, we want to focus attention on the evolution of recognition performance regarding **N** and also analyze the differences in performance when testing the system with the SpeechDat database (a different set of recordings from the training dataset) or an independent database (DigitVox).
Table 2 presents the results for connected digit recognition, according to SRR and WER, working with both databases.

Database	Configuration	SRR	WER
SpeechDat	Baseline	90.51	2.65
	14113/13440/69706/6113	92.30	1.96
DigitVox	Baseline	93.30	1.27
	14113/13440/69706/6113	93.79	1.14

Table 2. Connected digit recognition rates modeling time dependencies

Results obtained with the SpeechDat dataset show that by modeling time dependencies, we can achieve a great improvement in recognition, outperforming the inter-parameter dependencies modeling approach with a relative WER reduction of around 26% compared to baseline results. However, the improvement when using the DigitVox dataset was slightly lower, with a relative WER reduction of 10.2%. Thus, this solution seems more likely to be adapted to the training corpus for connected digit recognition.

To test whether this time dependencies modeling based solution works better using a bigger (and wider) training corpus, continuous speech recognition was used, with new sets of acoustic models based on demiphones, using the PARL dataset.

The results presented in Table 3 show a WER reduction between 14.2% and 24.3%. We observe some saturation in WER improvement when **N** is increased over certain values: after reaching optimum values, WER improvement becomes slower, and we should evaluate if the extra improvements really do justify the computational cost of working with such large values of **N** (which means working with high codebook sizes). Afterwards, additional WER improvement tends to zero, so no extra benefit is obtained by working with a very high number of N-grams. Thus a compromise between the increase in codebook size and the improvement in recognition accuracy is made when deciding upon the best configuration.

Database	Configuration	WER	WER_{var}
TC-STAR	Baseline	28.62	-
	3240/2939/2132/ 6015	24.56	14.19%
	7395/6089/4341/ 8784	21.73	24.07%
	20967/18495/17055/15074	21.66	24.32%

Table 3. Continuous speech recognition rates modeling time dependencies

The performance of the time dependencies modeling based system compared to the reference ASR system has also been analyzed in terms of the computational cost of recognition. Despite of the computational cost increase associated with the complexity of the system's training scheme, the system clearly outperforms the reference system in general terms. This good performance is due to a reduction in the computational cost of recognition of about 40% for those solutions which are a good compromise between codebook size

increase and recognition accuracy improvement (i.e. N-gram configuration "7395/6089/4341/8784" in Table 3).

6. Discussion

The future of speech-related technologies is clearly connected to the improvement of speech recognition quality. Commercial speech recognition technologies and applications still have some limitations regarding vocabulary length, speaker independence and environmental noise or acoustic events. Moreover, real-time applications still miss some improvement with the system delays.

Although the evolution of ASR needs to deal with these restrictions, they should not be addressed directly. Basic work on the core of the statistical models is still needed, which will contribute to higher level improvements.

HMM-based statistical modeling, the standard state-of-the-art for ASR, is based on some assumptions that are known to affect recognition performance. Throughout this chapter, we have addressed two of these assumptions by modeling inter-parameter dependencies and time dependencies. We noted different approaches for improving standard HMM-based ASR systems introducing some actual solutions.

Two proposals for using N-gram-based augmented HMMs were also presented. The first solution consists of modeling the dependence between the different acoustic parameters, thus overcoming the parameter independence assumption. The second approach relies on modeling the temporal evolution of the regular frequency-based features in an attempt to break the time independence assumption.

Experiments on connected digit recognition and continuous speech recognition have also been explained. The results presented here show an improvement in recognition accuracy, especially for the time dependencies modeling based proposal. Therefore, it seems that time-independence is a restriction for an accurate ASR system. Also, temporal evolution seems to need to be modeled in a more detailed way than the mere use of the spectral parameter's derivatives.

It is important to note than a more relevant improvement is achieved for continuous speech recognition than for connected digit recognition. For both tasks, independent testing datasets were used in last instance. Hence, this improvement does not seem to be related to an adaptation of the solution to the training corpus, but to better modeling of the dependencies for demiphone-based models. Thus, more general augmented models were obtained when using demiphones as HMM acoustic models.

Moreover, although the present research solutions should not be especially concerned with computational cost (due to the constant increase in processing capacity of computers), it is important to keep in mind implementation for commercial applications and devices. Taking computational cost into consideration, we find that the training computational cost increase of this modeling scheme clearly pays off by reducing the computational cost of recognition by about 40%.

Further work will be needed to extend this method to more complex units and tasks, i.e. using other state-of-the-art acoustic units and addressing very large vocabulary ASRs or even unrestricted vocabulary tasks.

8. References

Bonafonte, A.; Ros, X. & Mariño, J.B. (1993). An efficient algorithm to find the best state sequence in HSMM, *Proceedings of European Conf. On Speech Technology (EUROSPEECH)*

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*

Casar, M.; *Fonollosa,* J.A.R. & Nogueiras,A. (2006a). A path based layered architecture using HMM for automatic speech recognition, *Proceedings of ISCA European Signal Processing Conference (EUSIPCO)*

Casar, M. & Fonollosa, J.A.R. (2006b). Analysis of HMM temporal evolution for automatic speech recognition and utterance verification, *Proceedings of IEEE Int. Conf. On Spoken Language Processing (ICSLP)*

Furui, S. & Sandhi, M. (1992). *Advances in Speech Signal Processing,* Marcel Dekker, Inc., ISBN:0-8247-8540, New Jork, USA

Huang, X.D. & Jack, M.A. (1998). Unified techniques for vector quantisation and Hidden Markov modeling using semi-continuous models, *Proceedings of IEEE Int. Conf. On Acoustics, Speech and Signal Processing (ICASSP)*

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

Layton, M.I. & Gales, M.J.F. (2006). Augmented Statistical Models for Speech Recognition, *Proceedings of IEEE Int. Conf. On Acoustics, Speech and Signal Processing (ICASSP)*

Mariño, J.B.; Nogueiras, A.; Paches-Leal, P. & Bonafonte, A. (2000). The demiphone: An efficient contextual subword unit for continuous speech recognition. *Speech Communication,* Vol.32, pp:187-187, ISSN:0167-6393

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

Pylkkönen, J. & Kurimo, M. (2003). Duration modeling techniques for continuous speech recognition, *Proceedings of European Conf. On Speech Technology (EUROSPEECH)*

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

Rabiner, L. (1993). *Fundamentals of Speech Recognition,* Prentice Hall PTR, ISBN:0-13-015157-2, New Jersey, USA

Saon, G.; Padmanabhan, M.; Goinath, R. & Chen, S. (2000). Maximum likelihood discriminant feature spaces, *Proceedings of IEEE Int. Conf. On Acoustics, Speech and Signal Processing (ICASSP)*

Stemmer, G.; Zeissler, V.; Hacker, C.; Nöth, E. & Niemann, H. (2003). Context-dependent output densities for Hidden Markov Models in Speech recognition, *Proceedings of European Conf. On Speech Technology (EUROSPEECH)*

Takahashi, S. (1993). Phoneme HMMs constrained by frame correlations, *Proceedings of IEEE Int. Conf. On Acoustics, Speech and Signal Processing (ICASP)*