ON THE USE OF AUGMENTED HMM MODELS FOR OVERCOMING TIME AND PARAMETER INDEPENDENCE ASSUMPTIONS IN ASR

Marta Casar and José A. R. Fonollosa

Departament de Teoria del Senyal i Comunicacions,
Universitat Politècnica de Catalunya (UPC)

ABSTRACT

There is significant interest in developing new acoustic models for speech recognition that overcome traditional HMM restrictions. In this work, we propose to use N-gram based augmented HMMs. Two approaches are presented. The first one consists on overcoming the parameter independence assumption. This is achieved by modeling the dependence between the different acoustic parameters, using N-gram modeling. Then, the input signal is mapped to the new probability space. The second proposal tries to overcome the time independence assumption, by modeling temporal dependencies of each acoustic feature. Different configurations have been tested for connected digit and continuous speech recognition, results showing that adding long span information is beneficial for ASR performance.

1. INTRODUCTION

For modeling temporal dependencies or multi-modal distributions of 'real-world' tasks, Hidden Markov Models (HMM) 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, even if effective, are known to be poor [1], degrading classification performance. Adding dependencies through expert knowledge and hand tuning can improve models, but it is often not clear which dependencies to include. Therefore, the development of new acoustic models that overcome traditional HMM restrictions is an active field of research in Automatic Speech Recognition (ASR).

In order to overcome HMM limitations, many extensions have been proposed. One interesting approach for allowing complex dependencies to be represented are augmented statistical models [2], which are used in this paper in a new framework for dealing with temporal and parameter dependencies while still working with regular HMMs.

This work has been partially supported by the TECNOPARLA project, granted by the Catalan Government

2. MODELING TIME AND PARAMETER DEPENDENCES

In HMMs there are some assumptions that make evaluation, learning and decoding feasible. Among them, the Markov assumption for the Markov chain [1] 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. It 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. 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, 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 [3]. 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. In another approach to overcome the temporal limitations of the standard HMM framework, alternative trajectory modeling [4] has been proposed, taking advantage of frame correlation. The models obtained can improve speech recognition performance, but they require a demoralizing increase in model parameters and computational complexity. A smooth speech trajectory can be generated by HMMs through maximization of the model's output probability under the constraints between static and dynamic feature, modeling the temporal evolution of the acoustic models [5].

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 [2].

The approach presented in this chapter consists of creating an augmented set of models, modeling temporal and inter-parameter dependence.

3. N-GRAM MODELING

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 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.

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 $m \in M$ (see [1]):

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

Now 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 $p(x_t|m)$ will be modified as some independencies will no longer apply.

From this new set of features, regular SCHMM-based training will be performed, leading to a new set of augmented statistical models.

3.1. Modelling inter-parameter dependence

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 applying Bayes' rule:

$$P(f_0,f_1,f_2,f_3) = P(f_0)P(f_1|f_0)P(f_2|f_1,f_0)P(f_3|f_2,f_1,f_0) \quad (2)$$

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:

$$P(f_0,f_1,f_2,f_3) = P(f_0)P(f_1)P(f_2)P(f_3) \quad (3)$$

To overcome parameter independence, 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 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 expressed as:

$$P(f_0,f_1,f_2,f_3) = P(f_0)P(f_1|f_0)P(f_2|f_1,f_0)P(f_3|f_1)$$

In practice, not all the combinations of parameters will be used for modeling each parameter dependence 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 dependences 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 3-grams and 2-grams), 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 feature used to characterize the speech signal. Then, 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 new weights (taking advantage of 2-grams and 3-grams):

$$\begin{aligned} P_i(f_0) &= \sum_m c^0_{i,m} \cdot p(f_0|m) \\ P_i(f_1) &= \sum_m c^1_{i,m,\hat{m}_0} \cdot p(f_1|m) \\ P_i(f_2) &= \sum_m c^2_{i,m,\hat{m}_0,\hat{m}_1} \cdot p(f_2|m) \\ P_i(f_3) &= \sum_m c^3_{i,m,\hat{m}_1} \cdot p(f_3|m) \end{aligned}$$

where $$\hat{m}_k = \operatorname*{argmax}_m p(f_k|m)$$

is the likeliest class for parameter f_k at state i and time t. The new weights are defined according to N-gram based feature combinations:

- $c^1_{i,m,j} = c^1_{i,m}$ if the 2-gram "j,m" is not defined
- $c^2_{i,m,j,k} = c^2_{i,m,j}$ when the 3-gram "k,j,m" is not defined, but it is defined the 2-gram "j,m", and $c^2_{i,m,j,k} = c^2_{i,m}$ when neither the 3-gram not the 2-gram are defined
- $c^3_{i,m,j} = c^3_{i,m}$ when the 2-gram "j,m" is not defined

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

3.2. Modelling temporal dependencies

Next, we study the Markov assumption for the Markov chain. It is generally expressed as:

$$P(s_t|s_1^{t-1}) = P(s_t|s_{t-1}) \quad (4)$$

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

Considering temporal dependences, equation 4 should be reformulated. But, 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 4 can be expressed as:

$$P(s_t|s_1^{t-1}) = P(s_t|s_{t-2}, s_{t-1})$$

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

$$P(f_i) = P(f_i|f_{i_{t-2}}, f_{i_{t-1}})$$

Again, the most frequent combinations of acoustic parametrization 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 as:

$$P_i(f_k) = \sum_m c^k_{i,m,\hat{m}_{k,t-1},\hat{m}_{k,t-2}} \cdot p(f_k|m) \qquad (5)$$

with $$\hat{m}_{k,t-i} = \underset{m}{\operatorname{argmax}}\, p(f_k|m, t-i)$$

Notice that if the 3-gram "$\hat{m}_{k,t-2}, \hat{m}_{k,t-1}, m$" does not exist, the 2-gram or 1-gram case will be used.

4. EXPERIMENTS AND RESULTS

4.1. Methods and tools

For the experiments performed to test these approaches, the semi-continuous [6] HMM-based speech recognition system RAMSES [7] 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 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 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[1] was used for testing the performance of the models for continuous speech recognition.

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

4.2. Results 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 derivative and for the energy, and 2 and 3-grams for the second cepstral derivative. 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 will be 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 codewords 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.

database	configuration	SRR	WER
SpeechDat	baseline	90.51	2.65
	−/2000/2000,2000/2000	91.04	2.52
DigitVox	baseline	93.30	1.27
	−/2000/2000,2000/2000	93.71	1.17

Tabla 1. *Connected digit recognition rates modeling inter-parameter dependencies*

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. 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 7.8%. Because both datasets are independent from the training datasets, we didn't expect adaptation of the solution to the training corpus.

4.3. Results modeling temporal dependencies

When modeling temporal dependencies, each new HMM feature models the dependencies of the original acoustic features. Again, the different configurations are represented by a 4-digit string with the number of N-grams

used in equation 5 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.

A first set of experiments using connected digit recognition was used to analyze the evolution of recognition performance regarding **N**, and also to study the differences in performance when testing the system with the SpeechDat database or an independent database (DigitVox). 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 time dependencies modeling 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 2 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.

configuration	**WER**	**WER**$_{var}$
baseline	28.62	-
3240/2939/2132/6015	24.56	14.19%
7395/6089/4341/8784	27.73	24.07%
20967/18495/17055/15074	21.66	24.32%

Tabla 2. *Continuous speech recognition rates modeling time dependencies with TC-Star database*

5. CONCLUSIONS

In this paper we present two approaches for using N-gram based augmented HMMs. 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, trying to break the time independence assumption.

Experiments on connected digit recognition and continuous speech recognition have been performed. The results presented 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.

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.

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 ASR or even unrestricted vocabulary tasks.

6. BIBLIOGRAPHY

[1] X. Huang, A. Acero and H.W., *Spoken Language Processing*, Prentice Hall PTR, 1st edition, 2001.

[2] M.T. Layton and M.J.F. Gales, "Augmented statistical models for speech recognition," *Proceedings of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2006.

[3] J. Pylkkönen and M. Kurimo, "Duration modeling techniques for continuous speech recognition," *Proceedings of European Conf. on Speech Technology (EUROSPEECH)*, 2003.

[4] S. Takahashi, "Phoneme HMMs constrained by frame correlations," *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1993.

[5] M. Casar and J.A.R. Fonollosa, "Analysis of hmm temporal evolution for automatic speech recognition and utterance verification," *Proc. of IEEE Int. Conf. on Spoken Language Processing (ICSLP)*, 2006.

[6] X.D. Huang and M.A. Jack, "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)*, 1989.

[7] A. Bonafonte et alter, "Ramses: el sistema de reconocimiento del habla continua y gran vocabulario desarrollado por la UPC," *VIII Jornadas de Telecom I+D*, 1998.