

OVERCOMING HMM TIME INDEPENDENCE ASSUMPTION USING N-GRAM BASED MODELLING FOR CONTINUOUS SPEECH RECOGNITION

Marta Casar, José A.R. Fonollosa

TALP Research Center, Dept. of Signal Theory and Communications
Universitat Politècnica de Catalunya, Jordi Girona 1-3, 08034 Barcelona, Spain
{mcasar,adrian}@gps.tsc.upc.edu

ABSTRACT

The development of new acoustical models that overcome traditional HMM restrictions is an active field of research in automatic speech recognition. One possible approach to achieve this goal is to work with N-gram based augmented HMM. In this paper, we propose to deal with time independence assumption of HMM using N-gram based modelling. For this, the temporal dependencies of each acoustic feature are explicitly modelled. Results obtained in this work testing this approach in continuous speech recognition show the suitability of adding long span information for ASR performance. Moreover, we are improving previous results obtained using this modelling scheme in connected digit recognition experiments.

1. INTRODUCTION

From the beginnings of automatic speech recognition (ASR) statistical modelling of the acoustic information has been used. Standard speech recognition systems are based on a set of so called acoustic models that link the observed features of the voice signal with the expected phonetics of the hypothesis sentence. The acoustic model $p_r(\mathbf{O}|\mathbf{W})$ describes the speech production process which generates a feature sequence \mathbf{O} for a specific word sequence \mathbf{W} . In fact $p_r(\mathbf{O}|\mathbf{W})$ serves to calculate the acoustic likelihood of the word sequence \mathbf{W} . A wide range of acoustic models have been studied to better incorporate the variability of the speech signal. Despite of the improvements achieved in this field during the last years, recognition accuracy is still a weakness to overcome when real-world applications are considered.

In today's state-of-the-art ASR systems it is common practise to model $p_r(\mathbf{O}|\mathbf{W})$ by a first order hidden Markov process [1]. Although this model assumption is definitely a simplification [2], in practise Hidden Markov Models (HMM) have proven to perform well for this task. 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 [3]. However, HMMs are based upon some assumptions that are known to be poor [1]. In particular, successive frames of speech are assumed to be independent given the state that generated them. However, it is a proven fact that there is context dependence among frames. This dependence is usually modelled through the introduction of the first and second derivatives of the acoustic features. But different approaches can be found in literature for

a more explicit modelling of the time-domain dependencies of the acoustic models.

One interesting approach for allowing complex dependencies to be represented is the augmented statistical model [4]. Thus, we will avoid time independence assumption by modelling the temporal evolution of regular frequency-based features, using a set of augmented statistical models. This approach was previously introduced in [5] for connected digit recognition, where a new framework was introduced for dealing with temporal and feature dependencies while still working with regular HMM. In this work, we take the approach that presented best results and develop it for working with continuous speech recognition.

This paper is organized as follows: first, in section 2 some state-of-the-art approaches for dealing with time independence assumption are presented, introducing our work. In section 3 augmented statistical models are presented and in section 4 we make a short overview of the approach used for this work. The experiments performed and some practical considerations are explained in section 5, together with the results obtained. Finally, the main conclusions are summarized in section 6.

2. DEALING WITH TIME INDEPENDENCE ASSUMPTION

When modelling temporal dependencies or multi-modal distributions of 'real world' tasks, HMM are one of the most commonly used statistical models. Because of this, HMM have become the standard solution for modelling the acoustic information of the speech. In HMM there are some assumptions that make evaluation, learning and decoding feasible. One of them is the Markov assumption for the Markov chain [1], which states that the probability of a state s_t only depends on the previous state s_{t-1} .

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 achieved. But it is often not clear which dependencies should be included.

We find several approaches in literature for modelling time-domain dependencies. In [6] an algorithm to find the best state sequence of HSMM was implemented for a more explicit modelling of context. Duration [7] and trajectory modelling[8] have also been on stage, leading to more recent work on the temporal evolution of the acoustic models [9].

In the present work we are developing a framework for dealing with temporal dependencies while still working with regular HMM, by using augmented HMM.

3. AUGMENTED HMM

Augmented statistical models have been previously proposed as a systematic technique for modelling HMM additional dependencies, allowing the representation of highly complex distributions. These dependencies are thus incorporated in a systematic fashion. However, the price for flexibility is high, even when working with more computationally-friendly purposes [4].

In an effort to model the temporal properties of the speech signal, class labels modelling [10] has been studied in a double layer speech recognition framework [11]. The main idea was to deal with acoustic and temporal information in two different steps. However, the complexity of a double decoding procedure claims for a stronger justification than temporal dependence modelling.

A less complex scheme was to be studied. The approach presented next starts from the same idea in [9, 11] of creating an augmented set of models. But instead of modelling utterance likelihoods or posterior probabilities of class labels, it is temporal dependence that will be on stage in this work.

4. N-GRAM MODELLING

For better analyzing the influence of temporal dependencies in the recognition performance we will build a new set of acoustic models without loosing the scope of regular HMM. A similar procedure as in [5] will be followed. From the MFCC based parametrized signal those most frequent combinations of features will be selected, following a temporal dependence criteria. Language modelling techniques will be used for performing this selection. This way, a new probability space will be defined, to which the input signal should be mapped defining a new set of features.

In standard semi-continuous HMM (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 $m \in M$:

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

Now, new weights should be estimated as there are new parameters (modelling of temporal dependencies) to cover the new probability space. Also, the posterior probabilities $p(x_t|m)$ will be modified as some independence assumptions will no longer apply.

Regular SCHMM-based training will be performed, leading to the new sets of augmented statistical models.

4.1 Modelling temporal dependencies

In most HMM based ASR systems temporal dependencies between different frames are modelled by means of the successive derivatives of the acoustic features. However, a more explicit modelling of the time domain information seems relevant for improving recognition accuracy.

For these experiments we will be working with four MFCC features: frequency (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 this 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.

The observation probability distributions used in HMM assume that successive informations $O_1 \dots O_t$ within a state i can be considered independent. Then:

$$P_i(O_t|O_1^{t-1}) = P_i(O_t) \quad (3)$$

Our proposal consists on overcoming this assumption. For simplicity reasons, not all the sequence of observations is taken into account but only the previous two, working with the 3-gram O_{t-2}, O_{t-1}, O_t . Then, Eq. (3) will be expressed as:

$$P_i(O_t|O_1^{t-1}) = P_i(O_t|O_{t-2}, O_{t-1}) \quad (4)$$

Therefore, we can apply independency among features to the joint output probability of the four MFCC features transforming Eq. (2) to:

$$P(f_0, f_1, f_2, f_3) = P(f_0)P(f_1)P(f_2)P(f_3) \quad (5)$$

expressing the output probability of each HMM feature as:

$$P(f_i) = P(f_i|f_{i-2}, f_{i-1}) \quad (6)$$

In practice, not all the combinations of parametrization labels will be used for modelling each $P(f_i)$, but only the most frequent ones. A basic N-gram analysis of the dependencies in the training corpus is performed, keeping only the most frequent combinations of trigrams and bigrams.

Now, the output probability of state i at time t for each HMM feature k will be rewritten as:

$$P_i(f_k) = \sum_m c_{i, \hat{m}_{k,t-2}, \hat{m}_{k,t-1}, m}^k \cdot p(f_k|m) \quad (7)$$

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

If the trigram " $\hat{m}_{k,t-2}, \hat{m}_{k,t-1}, m$ " does not exist, the bigram (" $\hat{m}_{k,t-1}, m$ ") or unigram (" m ") case will apply.

Once defined the new set of augmented acoustic models it is Baum-Welch trained.

5. EXPERIMENTS

5.1 Databases

Different databases have been used for training and testing our system. First, the Spanish corpus of the *SpeechDat* and *SpeechDatII* projects [12] has been divided into three sets: a training dataset (for training an initial set of HMM), a developing dataset (for training the new set of augmented HMM), and a testing dataset. *SpeechDat* database consists on recordings performed over both fixed and mobile telephone networks, with a total of 4000 speakers for the fixed corpus, and 1066 speakers for the mobile corpus.

The results obtained using this first testing dataset have been used for trying out different configurations of the speech recognition system. These configurations will be defined by the number of N-grams to consider in Eq. (7) (that is, the number of combinations of parametrization labels to be used, adding together trigrams, bigrams and unigrams) for each new feature.

Afterwards, the new models have been tested with the Spanish Parliament dataset of the TC-STAR project (Technology and Corpora for Speech to Speech Translation) [13]. The TC-STAR project was envisioned as a long term effort focused on advanced research in all core technologies for speech-to-speech translation (SST). The project targeted a

selection of unconstrained conversational speech domains - i.e. broadcast news, political speeches, and discussion forums - and a few languages relevant for Europe's economy and society: European English, European Spanish and Chinese. One of the project goals was the implementation of an evaluation infrastructure based on competitive evaluation, in order to achieve the desired breakthroughs in ASR and SST. The PARL database, consisting on recordings from the Spanish Parliament Plenary Sessions, is an attractive domain for the development of such evaluation.

By using this PARL database for testing the performance of the models obtained we are testing the independence of the models regarding the training database. Furthermore and thanks to the characteristics of TC-Star database, we will be approaching similar conditions as those faced when recognizing unknown speakers in a changeable environment.

Large vocabulary continuous speech recognition (LVCSR) will be used as working task for testing the performance of our proposal. The lexicon used in this task is the one defined by the TC-STAR project for the Spanish Parliament (PARL database), composed by approximately 40.000 words. The language model defined for recognition using this database relies in trigrams comprising this 40.000 words plus filler and pause models.

Thus, experiments using this database will allow us to extract general conclusions on recognition over a restricted large vocabulary lexicon.

5.2 Reference speech recognition system

Our reference speech recognition system is the semi-continuous HMM based system RAMSES [14]. Semi-continuous hidden Markov models can be considered as a special form of continuous mixture HMM with the continuous output of probability density functions sharing in a mixture Gaussian density codebook (see [15]). 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. Then, SCHMM become the perfect choice for training restricted vocabulary and/or low resource applications. Moreover, the ease to combine and mutually optimize the codebook and the HMM parameters leads to a unified modelling approach. Also, recognition accuracy of the semi-continuous HMM is comparable with that of both discrete HMM and continuous HMM (if keeping the number of Gaussian mixtures at a reasonable number).

The main features of RAMSES are:

- Speech is windowed every 10ms with 30ms window length. Each frame is parametrized with the first 14 mel-frequency cepstral coefficients (MFCC) and its first and second derivatives, plus the first derivative of the energy.
- Spectral features are quantified to 512 centroids, energy to 64 centroids.
- Semidigits and demiphones [16] can be used as HMM acoustic units. When working with semidigits, 40 semidigit models are trained for the first set of acoustic models, plus one noisy model for each digit, modelled each with 10 states. Silence and filler models are also used, modelled each with 8 states. When working with demiphones, for the first set of acoustic models each phonetic

unit is modeled by several 4 states left to right models, each of them modelling different contexts. In the second (augmented) set of HMMs, each phonetic unit is modelled by several models, each of them modelling different temporal dependencies, using also 4 states left to right models.

- For decoding, a Viterbi algorithm is used implementing beam search to limit the number of paths. Frames are quantified to 6 centroids for spectral features and 2 for energy.

RAMSES will be used as the core of our ASR system. Also, recognition results using a set of acoustic HMM obtained by the regular implementation of RAMSES will be considered as our baseline for speech recognition experiments.

5.3 Baseline results

We want to compare the results obtained for continuous speech recognition against those obtained for connected digit recognition, using semidigit instead of demiphones as acoustic units (see [5]). It is important to recall that in this and the following experiments, each new feature models the temporal dependencies of each of the original acoustic features. This temporal dependency is represented by the most frequent combination of parametrization labels. The configurations used are represented by a four numbers string (henceforth \mathbf{N}) expressing the number of N-grams used in in Eq. (7) for representing each new feature (adding together trigrams, bigrams and unigrams). Low values for \mathbf{N} will mean that only some combinations will be modelled, keeping a low dimension signal space for quantization. On the other side, increasing the values in \mathbf{N} more dependencies will be modelled, but at the risk of working with an excessive number of centroids to map the speech signal.

For this baseline results we focused our attention in the evolution of recognition performance regarding \mathbf{N} , testing different configurations. We also analyzed the difference in performance when testing the system using the SpeechDat database or an independent database (DigitVox) obtained from a real telephone voice recognition application and containing 5317 sentences with identity card numbers (8 digit chains) recorded in noisy conditions. Results in Table 1 are expressed according to SRR (Sentence Recognition Rate) and WER (Word Error Rate) to measure the performance.

database	configuration	SRR	WER
SpeechDat	RAMSES	90.514	2.65
	14113/13440/6970/6113	92.305	1.96
DigitVox	RAMSES	93.304	1.27
	14113/13440/6970/6113	93.794	1.14

Table 1: Connected digit recognition rates modelling time dependencies.

Therefore, when using SpeechDat testing dataset a relative WER reduction around 26% was achieved, outperforming recognition results using the baseline system RAMSES. But the improvement when using DigitVox dataset was slightly lower, with a relative WER reduction of 10.2%. Thus, our solution seemed likely to be adapted to the training corpus for connected digit recognition.

5.4 Continuous speech recognition results

For testing our proposal using continuous speech recognition we have developed new sets of acoustic models based on

demiphones. However, we have first test the resulting augmented demiphone-based HMMs for connected digit recognition using SpeechDat testing dataset. These first tests were developed in order to analyze whether the performance of the new modelling scheme using demiphone as HMM acoustic units improved baseline results using the reference recognition system RAMSES. This was thought to be necessary as the computational complexity of the proposal (modelling temporal dependencies of demiphones instead of semidigits) actually increases, regarding the reference system, due to the increase in parametrization labels combinations. However, working with more complex acoustic units was necessary if continuous speech recognition was to be addressed.

The results obtained are represented in Table 2, showing a noticeable improvement in performance for connected digit recognition using demiphones, with SpeechDat testing dataset. Results are expressed, as before, according to SRR and WER. Again, each of the new features models the temporal dependencies of each of the original acoustic features. In these experiments, again, a wider range of N proved out to provide an increase in recognition accuracy, with a WER reduction between 11.5% and 16%. Thus, we can chose between optimize the accuracy, or working with reasonable codebook size (close to state-of-the art codebooks when working with standard implementations) while still improving the recognition performance.

database	configuration	SRR	WER
SpeechDat	RAMSES	88.414	3.11
	3240/2939/2132/6015	89.748	2.75
	20967/18495/17055/15074	90.116	2.61

Table 2: *Connected digit recognition rates using demiphones and modelling time dependencies.*

A reduction in performance using demiphones was expected, as semidigit acoustic units have proved to be the best acoustic units for connected digit recognition, which is the recognition task addressed in this second set of experiments. But demiphones allow us to model the vocabulary needed for working with more complex tasks, and this is the next step in our work.

The new augmented HMMs based on demiphones have been tested using the testing set of PARL database. This new set of experiments finally addresses continuous speech recognition. First, we tested the best-performing configurations of the newly developed augmented HMMs regarding connected digit recognition. However, the complexity of the task provides us with “lower” results (in absolute numbers) for both the baseline (RAMSES based) system and our system. This provides more room for improvement, which we make profit of by training new configurations of the augmented HMMs.

database	configuration	WER	WER _{var}
TC-STAR	RAMSES	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 using demiphones and modelling time dependencies.*

Table 3 presents a summary of the results obtained with the PARL testing dataset. Results are expressed according

to WER and showing a WER reduction between 14.2% and 24.3%. In this case, we can appreciate a saturation in WER improvement when increasing N over certain values: at first, WER improvement becomes slower and we should evaluate if the extra improvement achieved does really pay off for the computational cost of working with such great values of N (which means working with high codebook sizes). Afterwards, additional WER improvement tends to zero, so no extra benefit is obtained from working with very high N configurations. From the results in Table 3 we chose configuration “7395/6089/4341/8784” as a good compromise between codebook size increase and recognition accuracy improvement.

Going one step further, we have analyzed the performance of our architecture compared to the reference system in terms of recognition computational cost for the selected configurations. And, despite of the training computational cost increase of this modelling scheme (due to the complexity associated to it), the proposed system clearly outperforms the reference system RAMSES thanks to a reduction in recognition computational cost of about 40% for the “compromise” configuration (N -gram configuration “7395/6089/4341/8784”).

6. CONCLUSIONS

In this paper we present the results of using N -gram based augmented HMM for modelling the temporal evolution of the regular frequency-based features, trying to break the time independence assumption, for continuous speech recognition.

Results obtained show a noticeable improvement in recognition accuracy. Therefore, it seems that time independence assumption is a restriction for an accurate ASR system. Temporal evolution seems to need to be modelled in a more detailed way than the mere use of the spectral feature’s derivatives.

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

Moreover, the training computational cost increase of this modelling scheme clearly pays off by reducing the recognition computational cost in about 40%.

7. ACKNOWLEDGEMENTS

This work has been partially supported by the AVIVAVOZ project (TEC2006-13694-C03), granted by the Spanish Department of Science and Education.

REFERENCES

- [1] Huang, X., Acero, A. and Hon, H. W., *Spoken Language Processing*, 1st ed. Prentice Hall PTR, 2001.
- [2] S. Young, “Statistical modelling in continuous speech recognition,” in *Proceedings of the Int. Conf. on Uncertainty in Artificial Intelligence*, August 2001.

- [3] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*, 1993.
- [4] Layton, M.I. and Gales, M.J.F., "Augmented statistical models for speech recognition," *Proceedings of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2006.
- [5] Casar, M. and Fonollosa, J.A.R., "A n-gram approach to overcome time and parameter independence assumptions of hmm for speech recognition," *Proceedings of ISCA European Signal Processing Conference (EUSIPCO)*, 2007.
- [6] Bonafonte, A., Ros, X. and Mariño, J.B., "An efficient algorithm to find the best state sequence in HSMM," *Proceedings of European Conf. on Speech Technology (EUROSPEECH)*, 1993.
- [7] Pylkkönen, J. and Kurimo, M., "Duration modeling techniques for continuous speech recognition," *Proceedings of European Conf. on Speech Technology (EUROSPEECH)*, 2003.
- [8] Takahashi, S., "Phoneme HMMs constrained by frame correlations," *Proceedings of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1993.
- [9] Casar, M. and Fonollosa, J.A.R., "Analysis of hmm temporal evolution for automatic speech recognition and utterance verification," *Proceedings of IEEE Int. Conf. on Spoken Language Processing (ICSLP)*, 2006.
- [10] Stemmer, G., Zeissler, V., Hacker, C., Nöth, E. and Niemann, H., "Context-dependent output densities for Hidden Markov Models in speech recognition," *Proceedings of European Conf. on Speech Technology (EUROSPEECH)*, 2003.
- [11] Casar, M., Fonollosa, J.A.R. and Nogueiras, A., "A path based layered architecture using HMM for automatic speech recognition," *Proceedings of ISCA European Signal Processing Conference (EUSIPCO)*, 2006.
- [12] A. Moreno, R. Winksky, "Spanish fixed network speech corpus," *SpeechDat Project. LRE-63314*.
- [13] "TC-STAR: Technology and corpora for speech to speech translation," <http://www.tc-star.org>.
- [14] Bonafonte, A. et al., "Ramses: el sistema de reconocimiento del habla continua y gran vocabulario desarrollado por la UPC," *VIII Jornadas de Telecom I+D*, 1998.
- [15] Huang, X.D., and Jack, M.A., "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.
- [16] Marino, J.B., Nogueiras, A., Paches-Leal, P., and Bonafonte, A., "The demiphone: An efficient contextual subword unit for continuous speech recognition," *Speech Communications*, 2000.