Analysis of HMM Temporal Evolution for Automatic Speech Recognition and Utterance Verification

Marta Casar, José A.R. Fonollosa

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

Abstract

This paper proposes a double layer speech recognition and utterance verification system based on the analysis of the temporal evolution of HMM's state scores. For the lower layer, it uses standard HMM-based acoustic modeling, followed by a Viterbi grammar-free decoding step which provides us with the state scores of the acoustic models. In the second layer, these state scores are added to the regular set of acoustic parameters, building a new set of expanded HMMs. Using this expanded set of HMMs for speech recognition a significant improvement in performance is achieved. Next, we will use this new architecture for utterance verification in a "second opinion" framework. We will consign to the second layer evaluating the reliability of decoding using the acoustic models from the first layer. An outstanding improvement in performance versus a baseline verification algorithm has been achieved.
Index Terms: speech recognition, HMM acoustic modeling, state scores, utterance verification.

1. Introduction

A widely-used type of speech recognition system is 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 most usual implementation of this process is probabilistic, namely Hidden Markov Models [1]. A HMM is a collection of states with an output distribution for each state, defined in terms of a mixture of Gaussian densities. These output distributions are generally conformed by the direct acoustic vector plus its dynamic features (namely, its first and second derivatives), plus the energy of the spectrum. These dynamic features are the way of representing the context in HMM. However, although using such augmented feature vectors significantly improves performance, current speech recognition systems still don't provide convincing results when conditions are changeable (noise, speakers, dialects, ...).

We propose a two-layer speech recognition architecture dividing the modeling process into two levels and training a set of HMM for each level. References to other layered architectures for speech recognition [2], or meta-models [3] can be found in the literature.

Recognition is not our only goal. Every time a recognized word sequence is considered there is, inherently to it, some degree of uncertainty about its correctness. Therefore, it is necessary to build up a measure of how corresponding to the input utterance is the resulting word sequence. From this measure, a decision can be taken on whether to consider the output as correct or incorrect.

The correctness of recognition results, given either alternative hypothesis models or an N-best algorithm, has been broadly studied by several authors. Moreover, verification can be implemented using a "second opinion" approach [4], the correctness of the decoded hypothesis determined by comparing the results of two recognition systems for consensus. One variant of this approach could be using an analysis of the recognition process itself as the second opinion. Using the two-layer architecture proposed it can be consigned to the second layer the evaluation of the reliability of the models from the first layer.

The paper is organized as follows: first in section 2 we introduce our proposal for modeling HMM temporal evolution using state scores, and its implementation into a double layer speech recognition system. In section 3 we deal with utterance verification, presenting a second opinion based approach using the layered architecture defined. In section 4 the experiments to test the performance of our approach are presented, together with the databases and baseline systems for each task. General conclusions about this work are presented on section 5.

2. Modeling HMM temporal evolution using state scores

In standard HMM based modeling feature vectors depend only on the states that generated them. Context is represented by the dynamic features which, generally, do not model long-term variations. We present a method to incorporate context into HMM by considering the state scores obtained by a phonetic units recognizer. These state scores are obtained from a Viterbi grammar-free decoding, and added to the original HMM, obtaining a set of "expanded" HMM. A similar approach was used in [5] integrating the state scores of a phone recognizer into the HMMs of a word recognizer, and using state-dependent weighting factors.

2.1. Mathematical Formalism

In 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 $m \epsilon M$:

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

In [5] probability density functions are considered which make it possible to integrate a large context $x_1^{t-1} = x_1, \ldots, x_{t-1}$ of feature vectors which have been observed so far, into the HMM output densities. For that purpose, a new hidden random variable l (class

label) corresponding to phone symbols is introduced, which is a discrete representation of the feature vectors x_1^{t-1}. Thus, Eq. (1) is expanded and the output probability is defined (see [5]) as:

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

$$\approx \Big[\sum_m c_{i,m} \cdot p(x_t|m)\Big] \cdot \Big[\sum_l P(l|i)P(l|x_1^{t-1})p(x_t|l)\Big] \quad (2)$$

In our case, we don't want to introduce the modeling of the context for each feature vector into the HMM output densities, but to create a new feature modeling the context. So, a new probability term is defined:

$$b'_i(x_t) = \sum_l P(l|i)P(l|x_1^{t-1})p(x_t|l) \propto \sum_l P(l|i)P(l|x_1^t) \quad (3)$$

This is obtained by applying Bayes' rule to $P(l|x_1^t)$:

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

And, given that class l is itself a discrete representation of feature vectors x_1^{t-1}, we can approximate $p(x_t|l,x_1^{t-1}) \approx p(x_t|l)$. Also, $p(x_t,x_1^{t-1})$ is a constant in its evaluation across the different phonetic units, so $P(l|x_1^t) \propto p(x_t|l)P(l|x_1^{t-1})$.

$P(l|i)$ from Eq.(3) is estimated during the Baum-Welch training of the expanded set of models, and $P(l|x_1^t)$ corresponds to the state scores output obtained by the Viterbi grammar-free decoding.

We can see that, when combining $b_i(x_t)$ for each spectral feature and $b'_i(x_t)$ for the phonetic unit feature, the joint output densities are equivalent to Eq.(2).

2.2. Implementation

Figure 1 illustrates the double layer architecture implemented. We use a standard HMM-based scheme for the lower layer. From the acoustic models obtained, the phonetic units recognizer performs a grammar-free decoding, providing us with the current most likely last state score for each unit. This process can also bee seen as a probabilistic segmentation of the speech signal, keeping only the last state scores associated to the unit with the highest accumulated probability.

Different units were tested with the phonetic units recognizer, obtaining best performance (for digits recognition) working with semidigits as acoustic units. Consequently, labels l will represent last states of semidigit models and the density value can be computed as the probability that the current state s_t of a semidigit model is equal to l. Thus, semidigit last state scores output will be our new parameter to be added to the original parameter set.

In the upper layer, the new set of expanded HMM is built, adding the new parameter (state scores probability) to the original features (spectral parameters). This way, five parameters are considered henceforth for further training and decoding. As in [5], we will introduce a weighting factor w to control the influence of the state scores information, regarding the spectral parameters. However, we will work with a global weighting factor for the new parameter (not state-dependent), testing different values in the search of an optimal empirical weight.

Results obtained with the new recognition architecture are summarized in table 1, achieving a slight improvement in performance. However, this approach is regarded keeping in mind our long-term target: obtaining a reliable second opinion for utterance verification based on the analysis of HMM temporal evolution.

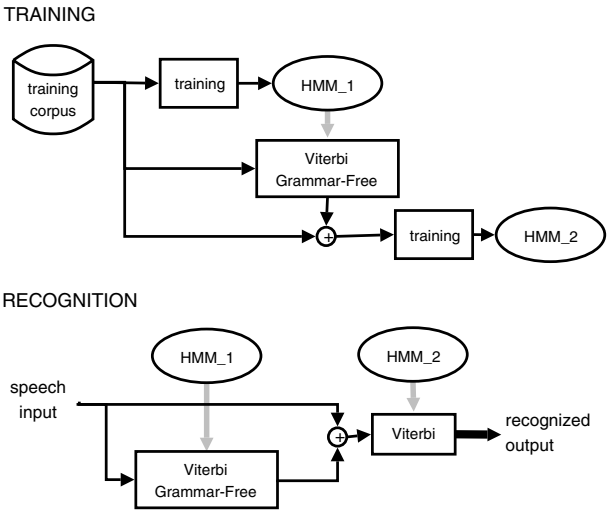

Figure 1: *Training and recognition schemes used for the double layer recognition system.*

3. Analysis of HMM temporal evolution for utterance verification

To be able to consider a second opinion, it is mandatory to have a certain confidence about its reliability. In our approach we are not using a second independent opinion to validate the output of the system, but analyzing the coherence of the recognition by means of a second decoding. The expanded HMM set built contains both the original spectral parameters plus the state scores parameter that can be seen as a model of the HMM temporal evolution. Therefore, we can compare the output of decoding using the expanded HMMs to that of a first decoding using regular acoustic models. If the two outputs differ, it means the temporal evolution of the models used presents some incoherence and, thus, the outputs are probably wrong (and therefore refused). If otherwise equivalent, they are supposed to be right (and accepted).

The architecture proposed for recognition and verification, represented in figure 2, relies in a double procedure. Once the two decoding outputs are generated, they are compared for consensus and classified following a sentence based criterion as accepted or refused. To evaluate the performance of this decision, sentences have been tagged in four categories: *exact* when correctly accepted by the verification step, *error* when incorrectly accepted, *detected* if correctly refused and *rejected* when incorrectly refused. In order to do this taggin both recognition outputs were previously evaluated classifying the sentences as correct or incorrectly recognized. Then, detected sentences will be those incorrectly recognized only by the first recognizer, rejected the ones incorrectly recognized only by the second, exact sentences those correctly recognized by both decodings, and error sentences the ones incorrectly decoded by both recognizers.

The state scores parameter weighting factor w will have a relevant paper in the utterance verification performance, as it will imply to give more or less importance to the temporal evolution of HMM states.

The decoded output will be in the shape of a word string conformed by a chain of recognized words. A first string level filtering stage can be performed before comparing the two outputs, making a first acceptance/rejection decision of the hypothesis made by the

recognizer. This stage consists on filtering the two decoding outputs using several task-dependent rules (i.e. sentence length, presence of out-of-vocabulary words, etc.). Sentences rejected on this previous basis will be tagged as *garbage*. In the following sections we present some results for different experiments with and without this string filtering stage.

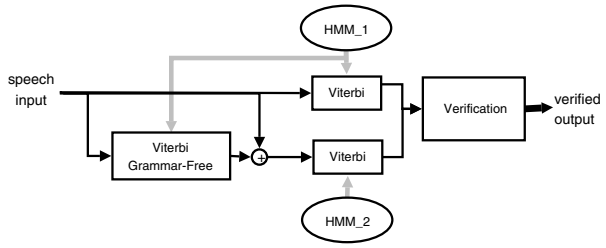

Figure 2: *Recognition and verification scheme of the utterance verification system based on the analysis of HMM temporal evolution.*

4. Evaluation experiments

4.1. Databases

Experiments have been developed using two different databases. First, the Spanish corpus of the *SpeechDat* and *SpeechDatII* projects [6] has been divided into three sets: a training dataset, a developing dataset (for training the HMM of the second layer), and a testing dataset. This 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 set of recordings.

The results from the experiments using this first testing dataset have been used for selecting the best configuration of the new system and, when necessary, tuning the parameters used. Afterwards, all the models have been tested with an independent database obtained from a real telephone voice recognition application, henceforth DigitVox. It contains 5317 sentences with identity card numbers (8 digit chains) recorded in noisy conditions. Experiments developed using this database will test the independence of our models, thus approaching to similar conditions as those faced when recognizing unknown speakers in a changeable environment.

4.2. Reference speech recognition system

Our reference speech recognition system is the semi-continuous HMM based system RAMSES [7]. The main features of this system are:

- Speech is windowed every 10ms with 30ms window length. Each frame is parametrized with the first 14 melfrequency cepstral coefficients (MFCC) and its first and second derivatives, plus the first derivative of the energy.
- Spectral parameters are quantified to 512 centroids, energy to 64 centroids.
- Semidigits are used as HMM acoustic units. 40 semidigit models are trained, plus one noisy model for each digit, modeled each with 10 states. Silence and filler models are also used, modeled each with 8 states.
- For decoding, a Viterbi algorithm is used implementing beam search to limit the number of paths. Frames are quantified to 6 centroids for spectral parameters and 2 for energy.

The HMMs trained by this system will be used as acoustic models in the first layer of our layered architecture. Moreover, recognition results using this original set of HMM will be considered as our baseline for speech recognition experiments.

4.3. Speech recognition experiments

Digit chain recognition has been taken as our working task for testing speech recognition performance, as this will be the target task for utterance verification experiments. With the aim of selecting the best configuration, we have performed an analysis of the contribution of the new parameter by building several acoustic models with different state scores' weighting factor w. Table 1 summarizes the results obtained with the DigitVox database.

configuration		Sentence	Word
system	w	recognition rate	recognition rate
baseline	-	93.304 %	98.73 %
layered	1	93.191 %	98.71 %
	0.5	93.605 %	98.80 %
	0.2	93.699 %	98.80 %

Table 1: *Recognition rates using expanded HMMs.*

Performance obtained by our system slightly overcomes the baseline, specially when weighting the probabilistic state scores contributions for the new models. Still, this new recognition approach should be regarded keeping in mind our main target: obtaining a reliable second opinion for utterance verification.

4.4. Utterance verification baseline system

Our system has been compared to a standard verification algorithm relying on phone-based filler models. This method [8] is based in the normalization of the scores output by the recognizer by means of a phone-based decoding search. The phone decoding uses a network of unconnected phonemes constrained only by phone sequences characteristic of the language without respect to the current lexicon or language model. Once normalized, these scores can become a measure of an overall goodness of recognition by providing an estimate of the acoustic match of the phone models to the input unconstrained word or word-sequences models.

Phone-based filler models have proved to perform better than other vocabulary independent approaches, as word-based filler models, or anti-models (see [9]). They can be outperformed by more complex solutions like feature transformation models or lattice-based combination models, but at the cost of being optimized for each specific recognition task and environment. Our goal, however, is to find a verification solution that doesn't need additional tuning.

4.5. Utterance Verification experiments

Experiments have been carried out using DigitVox testing database, which is completely independent from the one used for training the models and tuning the parameters. This promises verification results not conditioned by an over-training of the parameters neither by adaptation to the speech recordings.

Let us define (as in [10]) the *TRR (True Rejection Rate)* as the rate between the number of incorrect hypothesis detected by the verification system (correctly refused), and the total number of incorrect hypothesis: $TRR = D/I$. Then, the *FRR (False Rejection Rate)* is the rate between the number of correct hypothesis

Baseline, with different rejection thresholds (r)						
r	Exact	Error	Detec.	Rejec.	TRR	FRR
15	92.55	4.78	1.66	1.02	25.78	1.09
30	89.62	3.10	3.33	3.95	51.79	4.22
80	75.19	1.78	4.69	18.34	72.49	19.61
Layered system, with different weighting factors (w)						
w	Exact	Error	Detec.	Rejec.	TRR	FRR
0.2	92.46	4.14	2.78	0.62	40.17	0.67
1	91.84	3.59	3.33	1.24	48.12	1.33
5	84.03	1.69	5.23	9.05	75.58	9.72

Table 2: *Sentence verification results without string filtering (in %)*

rejected by the system (incorrectly refused) and the total number of correct hypothesis: $FRR = R/C$.

In terms of the *TRR* and *FRR* measures, a ROC curve (Receiver Operating Characteristic) [11] is a curve that shows the TRR versus the FRR for every threshold level used, expressing the latest in the x axis. Depending on the curve obtained we can evaluate the performance of the verification system.

By modifying the weighting factor w given to the new parameter in the expanded HMM set we will be modifying the performance of the second decoding. This will be used to obtain different behaviors of the verification system. Table 2 shows the results obtained for sentence verification using different values of w with our layered architecture, compared to results from the baseline verification algorithm. A very performant correct recognition rate is obtained for a relatively low error rate ($w = 5$), while keeping a reasonable rejection rate. On the other hand, baseline verification performance directly relies on the rejection rate allowed. In our approach, a weighting factor ≥ 1 implies giving more confidence to the temporal evolution of HMM states, at the cost of increasing the error probability. Figures 3 and 4 show the ROC curve representing TRR vs. FRR values for both verification systems, with and without the string filtering step.

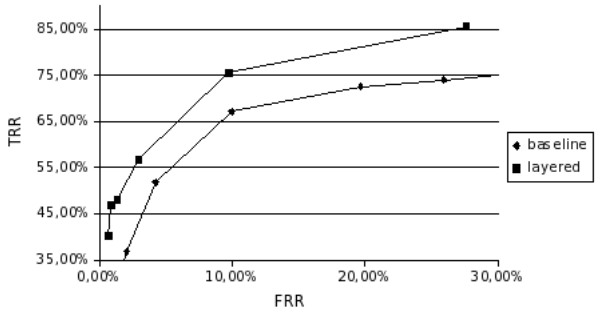

Figure 3: *TRR vs. FRR without string filtering for both systems.*

5. Conclusions

Throughout this paper we present some experiments carried out using a double layer speech recognition and utterance verification approach based on the analysis of the temporal evolution of HMM's state scores.

Speech recognition performance using the layered architecture is slightly better than with our baseline system, although computational cost increase becomes a drawback. However, when using this layered architecture for utterance verification following a "second opinion" approach, results become high-flying. Appart from

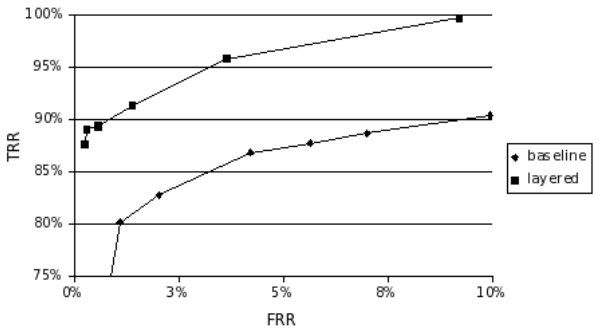

Figure 4: *TRR vs. FRR with string filtering for both systems.*

a better TRR vs FRR behavior, our verification approach offers a very performant correct recognition rate quite a low error rate

6. Acknowledgments

This work has been partially supported by Spanish MCyT under the projects TIC2002-04447-C02 and TIN-2005-08852

7. References

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

[2] Demuynck,K., Laureys,T., Van Compernolle,D., and Van Hamme,H., "Flavor: a flexible architecture for LVCSR," *Proc. Eurospeech*, pp. 1973–1976, 2003.

[3] Cox,S., and Dasmahapatra,S., "High-level approaches to confidence estimation in speech recognition," *IEEE Transactions on Speech and Audio processing*, vol. 10, no. 7, pp. 460–471, October 2002.

[4] Hernández-Ábrego,G., and Mariño,J.B., "A second opinion approach for speech recognition verification," *Proceedings of the VIII SNRFAI*, vol. I, pp. 85–92, 1999.

[5] 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.

[6] Moreno,A., and Winksky,R., "Spanish fixed network speech corpus," *SpeechDat Project. LRE-63314*.

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

[8] Young,S.R., "Detecting misrecognitions and out-of-vocabulary words," *Proceedings of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. I, pp. 21–24, 1994.

[9] Jiang,L., and Huang,X.D., "Vocabulary-independent word confidence measure using subword features," *Proceedings of IEEE Int. Conf. on Spoken Language Processing (ICSLP)*, 1998.

[10] Sanchis,A., "Phd thesis. Estimation and application of confidence measures for speech recognition (in spanish)," *Universidad Politécnica de Valencia. Departamento de Sistemas Informáticos y Computación*, 2004.

[11] Egan,J.P., *Signal detection theory and* ROC *analysis*, Academic Press, 1975.