14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

A PATH-BASED LAYERED ARCHITECTURE USING HMM FOR AUTOMATIC SPEECH RECOGNITION

Marta Casar, José A. R. Fonollosa and Albino Nogueiras

TALP Research Center, Universitat Politècnica de Catalunya
C. Jordi Girona 1-3, 08034, Barcelona, Spain
phone: + (34) 934010964, fax: + (34) 934016447, email: {mcasar,adrian,albino}@gps.tsc.upc.edu
web: www.talp.upc.edu

ABSTRACT

Generally, speech recognition systems are based on one layer of acoustic HMM states where the recognition process consists on selecting a sequence of those states providing the best match with the speech utterance. In this paper we propose a new approach based on two layers. The first layer implements a standard acoustic modeling. The second layer models the path followed by the speech signal along the activated states of the acoustic models, defining a set of state-probability based HMMs. This method presents two main advantages in front of conventional recognizers: a consistent pruning of the possible paths preceding and following each state in the recognition process, and the possibility of modeling high-level information in the second layer in a somewhat independent fashion from the acoustic training. A testing database from a real voice recognition application has been used to study the performance of the system in a changeable environment.

1. INTRODUCTION

A standard 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 HMM [1].

Most current speech recognition systems are based in either continuous or semicontinuous HMM, using a representation of the signal space by means of a set of parameters (normally spectrum, delta spectrum, delta-delta spectrum and energy). These systems provide fine recognition performance in concrete tasks or working in a controlled environment. But when conditions are changeable results become discouraging. In the search of a new speech recognition architecture that deals with all the information available at recognition time without being restricted by HMM acoustic constraints, we present a new layered speech recognition system. The proposed architecture suggests a modular framework allowing a two-steps search process. Some references to layered architectures for speech recognition [2], or meta-models [3] can be found in the literature (implemented for LVCSR and for confidence estimation in speech recognition, respectively).

Following the first approach [2], we will split the recognition scheme into two levels developing a set of HMMs for each level. In the first level we will perform a conventional HMM based acoustic analysis. Then, in the second level we will model the path followed by the speech utterance along the activated states of the acoustic HMMs. This way we provide an open field for the introduction of information present at recognition time that cannot be treated in existing one-layer architectures, as well as relaxing constraints for acoustic optimization. Moreover, a layered architecture where each level is defined by its own set of HMM offers another advantage: the reduction of speaker dependency by training each layer with a different set of recordings, approaching to similar conditions as those of the testing stage or when facing the recognition of unknown speakers.

This project has been partially supported by Spanish MCyT under the project TIN-2005-08852

2. LAYERED SPEECH RECOGNITION

HMM based speech recognition systems rely on the modeling of a set of states and transitions using the probability of the observations associated to each state. These probabilities being independent, the path followed by the signal is not modeled. In our approach we try to model this "path" in order to associate recognition to the best matching path. This will help with the recognition of acoustic units regardless of their variations when uttered in different environments, by different speakers or affected by different background noises.

Following the idea of breaking the recognition architecture into two layers, we will keep a conventional acoustic modeling step for the first layer and consign to the second layer modeling the evolution followed by the speech signal. This evolution will be defined as the path through the different states of the sub-word acoustic models defined in the first layer.

In semicontinuous HMM, a VQ codebook is used to map the continuous input feature vector $\mathbf{x}$ to o_k (the k_{th} codeword) so we can use a discrete output probability distribution (pdf) $b_j(k)$ for state j (see [1]):

$$b_j(\mathbf{x}) = \sum_{k=1}^{M} b_j(k) f(\mathbf{x}|o_k) \tag{1}$$

The input to the HMM of the second layer will be the vector of state probabilities given by the acoustic models of the first layer. So hence, a new set of semicontinuous output pdfs $b'_j(\mathbf{x})$ will be defined:

$$b'_j(\mathbf{x}) = \sum_{k=1}^{M'} b'_j(k)\, b_k(\mathbf{x}) \tag{2}$$

This equation can be expressed in terms of a new distribution function $f'(\mathbf{x}|b_k)$ where the output probabilities vectors b_k

play the role developed by o_k in the first level. Actually, we are defining a new codebook covering the sub-word state-probability space.

$$b'_j(\mathbf{x}) = \sum_{k=1}^{M'} b'_j(k)\, f'(\mathbf{x}|b_k)$$

Acoustic HMMs from the first layer are used for estimating the new state probability representation for the second layer HMMs. Thus, the new weights $b'(\mathbf{x})$ will be obtained through a new Baum-Welch estimation, in a second modeling step. New observation distribution for the second layer HMMs are trained using the same stochastic matrix of the original acoustic HMMs.

In practice and because M and M' are large, equations (1) and (2) will be simplified by using the I and I' most significant values (see [1]). This way, we are preventing some recognition paths to be activated, and this can result in a different decoding when I $\neq$ I'.

A second approach to path-based modeling will be to consider the context of each state in the second layer of the recognition architecture. Thus, the mapping of the models will be done using windows centered in each state and embracing one or more adjacent states: those more probable to have been visited before the actual state, and the more probable future ones.

3. EXPERIMENTS

Our new approach for building a layered architecture has been developed with the aim of overcoming some limitations of traditional HMM framework. The three main targets of this new architecture will be: improving speech units modeling regarding their variation when uttered in a changeable environment, improvement in speaker independence, and inclusion of added value information in the recognition process. At the same time, we want to keep speech recognition accuracy high, in order to improve globally our system.

Experiments have been carried out for two implementations of the second layer: one-state width and L-state width path-based modeling. In the second case, L-1 is the number of adjacent states considered as the significant context for each state of the path.

Digit chain recognition is still an application of great practical interest. Therefore, it is an usefull first target task for testing our new architecture.

3.1 Databases

Experiments have been carried out using two different databases. First, SpeechDat Spanish database [4] has been divided into three sets: a training dataset with 11443 sentences containing sequences of digits, a developing dataset (for training the HMM of the second layer) with 11535 sentences, and a testing dataset with 3405 sentences.

The results from the experiments using this first testing dataset have been used for selecting the best configuration of the new system in order to improve baseline results and, when necessary, tuning the parameters used.

Afterwards, all models have been tested with an independent database obtained from a real telephone voice recognition application, from now on referenced as DigitVox, which contains 5317 sentences with identity card numbers (8 digit chains) recorded in noisy conditions. This second set of experiments will test the independence of our models and its robustness in front of noise, thus approaching to similar conditions as those faced when recognizing unknown speakers in a changeable environment.

3.2 Baseline

The baseline speech recognition system used for evaluating the performance of our new approaches is the SCHMM based RAMSES [5]. The main features of this system are:

- Speech is windowed every 10ms with 30ms window length. Each frame is parameterised with the first 14 mel-frequency cepstral coefficients (MFCC) and its first and second derivatives, plus the first derivative of the energy.
- Spectral parameters are quantified to 512 centroids, energy is quantified to 64 centroids.
- Semidigits are used as HMM acoustic units. 40 semidigit models (two semidigits for digit, two contexts for semidigit) 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.

3.3 One-state width path-based recognition

In the way to a path-modeling based recognition, we divide the system architecture into two layers, keeping the lower layer equal to the reference scheme of the baseline recognition system, and setting target in the improvement of the upper layer.

This second layer consists of mapping the acoustic models obtained in the first layer into state-probability based HMM. Also, a new codebook covering the probability space will be defined. Therefore, we presume to have a "transparent" second layer as it is equivalent to a direct mapping of the acoustic probabilities. Performance is expected to be equivalent to that of the baseline system, if no further improvements were introduced.

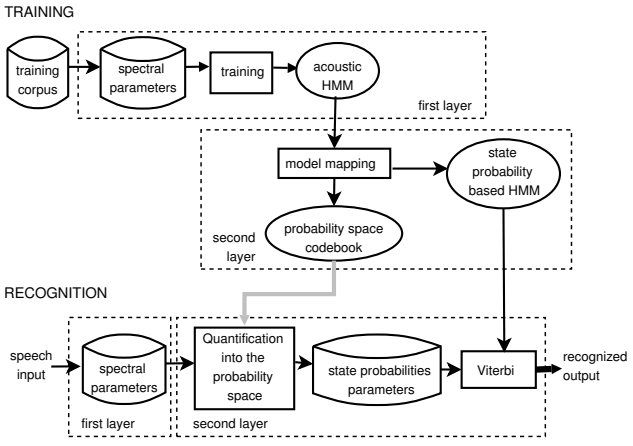

Fig. 1. Basic diagram of a one-state width path-based layered architecture. Training and recognition schemes

recognition system	Sentence recognition rate	Word recognition rate	Substitution rate	Insertion rate	Deletion rate
baseline	93.304 %	98.73 %	0.24 %	0.97 %	0.06 %
One-state width path-based recognition	94.677 %	99.10 %	0.26 %	0.46 %	0.18 %

Table 1. Recognition rates using one-state width path modeling based layered recognition architecture

We are no longer working with spectral parameters distributions but with the probabilities of the whole set of possible states: we have moved from the signal space (covered by the spectrum) to the probability space (defined by the probability values of each of the states). In traditional HMM gaussian densities model the probability with which a state generates some spectral parameters (acoustic information). New HMMs generated by the second layer will give the probability of being at a certain point of the path followed by the speech signal. Moreover, we are mapping the models to an space of dimension N (the total number of HMM states from the first layer). Then, we can implement some pruning by keeping only the $\sim$N/2 most significant values (see section 2 and [1]). This way we are constraining the possible states preceding and following each active state, preventing some recognition paths to be activated. This will increase speed and also promises an improvement in recognition performance. Actually, it can be seen as a similar strategy to CHMM gaussian mixture pruning, where each state is modeled with a mixture of private Gaussians.

A basic diagram of the architecture proposed is showed in figure 1. Recognition results obtained using this architecture are presented in table 1, showing an improvement in word recognition. This is achieved thanks to a positive weighting of the states with higher likelihood (implicit in the solution proposed) and the pruning of the preceding and following states to be activated for each state.

3.4 L-states width path-based recognition

The validity of the layered architecture using path-based recognition for implementing the second layer has been successfully tested in the previous experiments. Next step consists of introducing the "context" of each state, approaching the idea of modeling the path followed by the speech signal.

We can model the speech signal as a succession of Markov states, where transitions between states and models are restricted by the topology of the models and the grammar. The speech signal can be modeled by means of different *state successions* (or *paths*). Each path has its own associated probability, allowing one state to be part of different paths. The path with maximum likelihood will be the one conformed by the succession of states that maximizes the joint probability (defined by the product of probabilities of each state in the path).

We will limit the state context to a window of length L, in order to allow generalization and to make the implementation computationally feasible. So, we will not deal with the whole path followed by the signal along all the states succeeding one after the other and conforming a certain speech utterance.Instead, we will consider only the possible L/2 - 1 previous and following ones for each state. Also, only combinations of states corresponding with the same phonetic unit will be allowed.

Taking these simplifications to the limit, if we use windows with length L=1 we are directly mapping the acoustic models into state-probability models in the same way as in previous section. We can see L-states width path recognition as an extension of previous one-state width approach.

The architecture devised for the previous experiments will be adapted. By mapping the acoustic models obtained in the first layer, we will build a new codebook covering the probability space. Besides, a table with all the possible state combinations will be defined, taking into account restrictions aforementioned. Next, considering all state combinations, the input speech will be statistically defined obtaining a new set of parameters. These parameters will represent the probabilities of each sequence of states. In the training stage a new set of statistical models will be built. In recognition, these state-probability based HMMs will allow decoding the path followed by the speech signal represented using the states probabilities parameters. Figure 2 shows the architecture proposed.

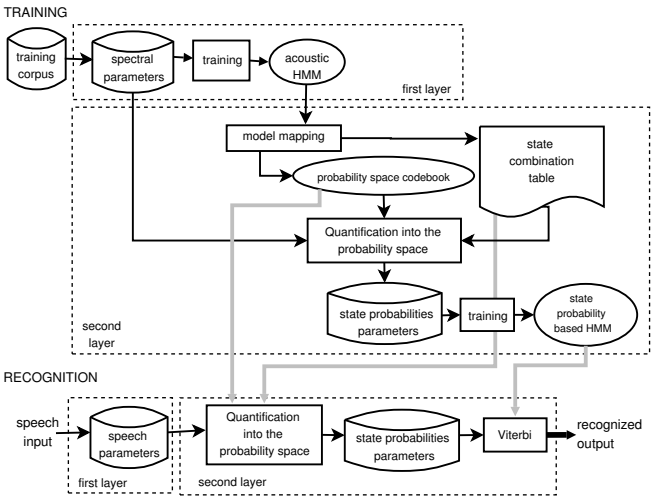

Fig. 2. Basic diagram of an L-states width path-based layered architecture. Training and recognition schemes

In this first approach to L-states width path-based recognition, we will consider the following:

- 40 semidigit acoustic models (two semidigits for digit, two contexts for semidigit) modeled each with 10 states and allowing a maximum leap of 2 for intra-model state transitions. Silence and filler models are also used, modeled with 8 states and maximum leap of 1.
- Transitions between two models are only allowed between final and initial states of each model.
- L = 3 (window length)

In the present case we will have a set of 4006 possible state combinations represented in the new codebook. This

recognition system	Sentence recognition rate	Word recognition rate	Substitution rate	Insertion rate	Deletion rate
baseline	93.304 %	98.73 %	0.24 %	0.97 %	0.06 %
L-states width path-based recognition	93.417 %	98.89 %	0.36 %	0.35 %	0.40 %

Table 2. Recognition rates using L-states width path modeling based layered recognition architecture

results from considering the characteristics of the acoustic models from the lower layer and the window length. In the new representation of the input signal using the probability space codebook we will keep only those 256 (approximately N/2) most significant values, being N the number of total HMM states from the first layer.

Results presented in table 2 show a noticeable improvement in word recognition, even if lower than using the one-state width approach. The reason for this can be the growth of the information to be modeled (the total number of state combinations, N). As only the most N/2 significant values are used for quantifying the training data, in some cases there could be a lost of information even if the general performance (and the improvement in recognition results) show the benefits of this decision.

This second approach remains justified by the higher flexibility of the second layer. Further experimentation needs to be carried out using more complex phonetic units to take profit of this flexibility for modeling high-level information.

4. CONCLUSIONS AND FURTHER WORK

A state-probability based modeling approach has been studied for the second level of a layered architecture for ASR. Keeping a conventional acoustic modeling scheme for the first layer, the second layer will model the evolution of the acoustic HMM (as a new way of modeling state transitions) along the speech utterance. This proposal aims to improve recognition accuracy while allowing the inclusion of added value information and a better adaptation to different recognition environments and speakers.

A first implementation consisted on building the second layer by directly mapping the acoustic models into state-probability based models, achieving a relevant improvement in performance. This results from a positive weighting of the states with a higher likelihood, as well as the pruning of the possible paths that can follow and precede each state, preventing some recognition paths to be activated. Actually, it can be seen as a similar strategy to CHMM Gaussian mixture pruning.

We can also see this approach as modeling the path followed by the speech through the different states of the acoustic HMM used for its representation, taking just the present state for each moment. Going one step further, we have introduced the context of the states. Experiments show this second implementation also improves baseline performance, even if results are slightly lower than those from the one-state width path-based modeling implementation. However, this second proposal provides higher flexibility, which would allow the introduction of added value information into the recognition. Further work in this direction will be undertaken, testing the performance of the layered architecture modeling other acoustic units (like phones or semiphones) and working with more complex tasks to compare performances with state of the art ASR systems.

But recognition is not our only goal. Some research is being done in utterance verification working with a "second opinion" approach [6], using a second speech recognizer to verify the output of a main recognizer. The layered architecture presented in this document can also be used for verification, using the output of the second layer as a second opinion. Previous related experiments working in a second opinion framework for utterance verification using a layered architecture make us expect positive results.

REFERENCES

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

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

[3] S. Dasmahapatra and S. Cox, "Meta-models for confidence estimation in speech recognition," *Proc. IEEE Int. Conference On Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1815–1818, 2002.

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

[5] 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.

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