VOCABULARY INDEPENDENT ACOUSTIC-PHONETIC MODELING FOR CONTINUOUS SPEECH RECOGNITION

L. Fissore ◇ P. Laface ⋆ G. Micca ◇ F. Ravera ◇

◇ CSELT - Centro Studi e Laboratori Telecomunicazioni
Via G. Reiss Romoli 274 - I-10148 Torino, Italy
E-Mail fissore@cselt.stet.it
⋆ Dipartimento di Automatica e Informatica - Politecnico di Torino
Corso Duca degli Abruzzi 24 - I-10129 Torino, Italy
E-Mail laface@polito.it

ABSTRACT

This paper investigates the problem of defining the acoustic-phonetic unit set for flexible vocabulary continuous speech recognition systems.
As an alternative to the classical modeling approach with biphones and triphones, a set of stationary/transitory state units is defined that is limited enough in number as to represent a closed set trainable once and for all. A major benefit of these units is that inter-word transitions can easily be taken into account. We show that a system employing these new units favorably compares with respect to a baseline recognizer with Continuous Density Hidden Markov Models of context-dependent biphones and triphones, selected through a minimal occurrence criterion within the training database.

1 Introduction

Subword unit modeling is mandatory for large vocabulary recognition systems, as well as for flexible vocabulary applications. It is well known that a central issue for these tasks is the selection of a set of basic units that can be accurately modeled with the available training data, but that are also robust to phonetic contexts which never appeared in the training database.
Context-sensitive phonetic models (triphones) are generally used for taking into account the coarticulation effects. The extension to larger sets of acoustic-phonetic units determines a decrese in statistical robustness due to undertraining. Moreover, even if it could be possible to accurately train the triphones appearing in a specific application vocabulary, it is not feasible to train all triphones of a language to cover any new vocabulary. Since not all contexts can be accounted for, smoothing or tying techniques are required to compensate for insufficient training data.
Several approaches have been proposed for the generation of trainable and consistent phone-based units. Relevant examples include:

- context-independent phoneme modeling [10] where coarticulation effects are taken into account augmenting the acoustic vector;
- parameter smoothing of detailed context dependent models with less detailed, but better trained models [12].
- parameter sharing [8] by tying similar units [8, 3] or similar distributions [5, 14].

This paper discusses the advantages of designing a set of units including only stationary context-independent phonemes and phone-to-phone transition units and compares them with the classical diphones or triphones. Opposite to context-dependent phones, diphone-like transition units, introduced in a previous paper [6], are limited enough in number. Thus it is possible to train once and for all, given a properly designed database, a complete set of "universal" units that are able to represent any new vocabulary.
The models of these units have been evaluated on a 751 word speaker independent spontaneous speech recognizer for a railway timetable inquiry application, managed by a dialog system. In particular the states corresponding to word junctures have been trained without distinguishing inter-word and intra-word units having the same context, and the decoding search algorithm was modified to explicitly deal with inter-word transitions.
The paper is organized as follows: the definition and modeling of the units is recalled in Section 2, the data structure and the decoding algorithm are detailed in Section 3, while the speech databases used, the experiments performed and the results obtained are described in Section 4.

2 Unit modeling

The recognition system described in this paper is based on Continuous Mixture Density Hidden Markov Models

Table 1: Context-independent phonemes and diphone-transition units

Phoneme sequence	...xpy ...								
States	x_l	x_c	x_r	p_l	p_c	p_r	y_l	y_c	y_r
Diphone-transitions	...		<xp>			<py>		...	
Context-independent phonemes	...	<x>			<p>			<y>	...

of subword units. The units include context independent phonemes (CI) and context-dependent units (CDU). The CDU set is selected according to a minimal occurrence criterion within the training database [3]. Linear 3-state left-to-right models are assigned to each unit, while a single state unit models the background noise. This is a rather standard model, where lateral states are supposed to represent the coarticulation effects occurring in the transition from the preceding phone and to the successive phone, while the central state mostly represents the stationary component of the phone.
The limitations of such a model are listed below:

- The central state of a triphone (l)p(r), which represents the stationary part of phone <p>, is trained by means of context-dependent samples only. The resulting distribution may become very detailed while loosing generalization strength.

- The distribution of the final state of a left context-dependent diphone (l)p is forced to merge the coarticulation effects of the following contexts. It is, therefore, often too smoothed unless it is modeled with a large number of emission densities. The same observation applies to the first state of a right context-dependent diphone p(r).

- A context-independent phone <p> is trained by only those samples which have not been used for training the CDUs. If many different CDUs have been defined, the resulting CI models can result to be undertrained.

To avoid these drawbacks, a new type of units was proposed in [6], where the central states of all CDUs of the same phoneme are tied, i.e. the stationary state of phone <p> is trained in context-independent mode. A two state transition unit <pq> is created by concatenating the final state of <p> and the first state of the next phone <q>. A phonetic transcription is, thus, represented as a sequence of stationary context-independent phones and diphone-like transition units as shown in Table 1. Since each unit is modeled by a large number of observation frames of the same context, it is robust and has good generalization capability. This set of units corresponds to the pseudo-diphones presented, in another context and for a limited domain, in [4] and to phonicles [13].

2.1 Mixture selection

The emission densities in our systems are modeled by mixtures of Gaussians. In many systems the number of densities per state is fixed and selected a priori according to the size of the training database. Increasing the size of the mixtures generally leads to more detailed models and better recognition results.
However, the number of densities for each state should be carefully selected to fit the actual distribution of the training data in order to avoid some states to be overtrained.
Therefore, in our procedure the optimization of the number of gaussian mixtures with respect to the amount of training data is performed according to the following steps:
we arbitrarily divide the training database into a clustering and an evaluation subset (2/3 and 1/3 in our experiments). Segmental Viterbi training and alignment of the training observations to each state is then performed using an available set of models and the *complete* database. For each state, we use the segmentation of the *clustering* subset obtained from the previous step, to perform the K-Means clustering. The number of densities is increased until it reaches a preset maximum value or the average likelihood of the observations in the *evaluation* subset does not decrease (a clear cue of overtraining). The number of densities associated to the state of each model is, thus, adapted to a portion of the training data and evaluated on an independent subset. We then iterate this procedure starting from the Segmental Viterbi training and alignment of the *complete* training database using the new models.
By properly selecting the number of densities per state we obtain better accuracy of the models (both diphone-transition units or classical diphones and triphones) and simultaneously halve the total number of densities.

3 Decoding

We tested the effectiveness of these units in an isolated word 600 surname recognition task, where the word error rate was nearly halved (from 8.0% to 4.1%).
In order to extend this approach to continuous speech, the states corresponding to word junctures have been trained without distinguishing inter-word and intra-word units having the same context, and the decoding search algorithm was modified to deal with inter-word transitions.

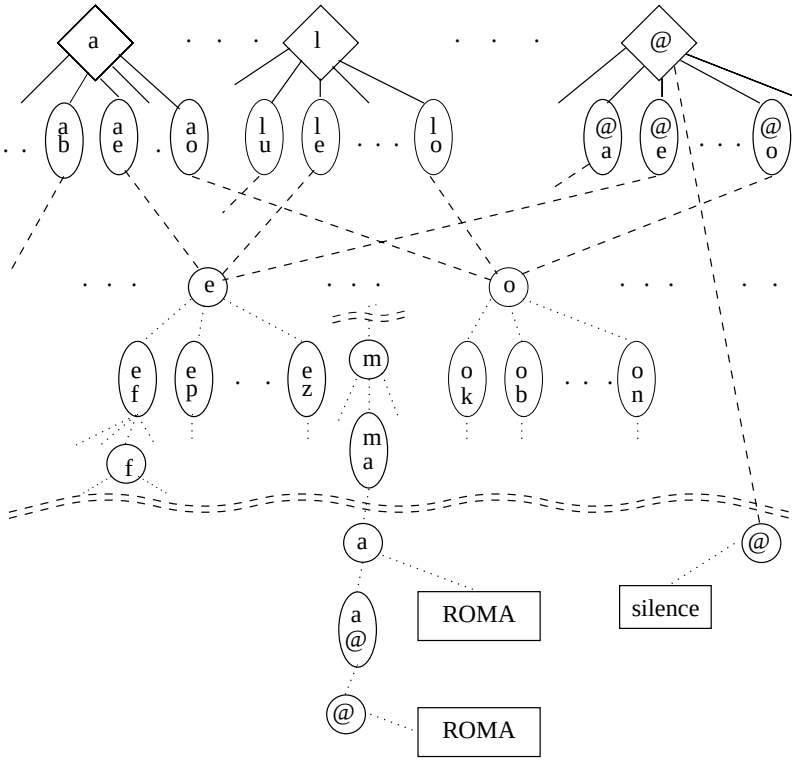

Figure 1: Lexical tree with multiple roots

The use of inter-word transition units does not affect the training algorithm, since transcription of the training sentence automatically places the correct units at word boundaries. In particular, the word junctures are represented with a twofold path. One path directly models the transition at the boundary of the two words. The second path includes an optional silence linking the corresponding positional phone-to-silence ending unit of the preceding word, and the silence-to-phone starting unit of the successive word.

Since the next word is not known during the recognition, word endings must be represented in the lexical tree with all the transitions toward the beginning of the next word. The explicit representation of all possible starting and ending coarticulation units for each word in the lexicon is not feasible, because it would imply excessive memory and computational costs

Our decoding process exploits a tree based Viterbi decoding algorithm, but the search is performed on a tree with multiple roots that explicitly represents word transitions as described in the following (Fig. 1).

Each word is represented in the lexical tree by a transcription beginning and ending with stationary units. A second transcription accounts for possible silences between words. For example, in Fig. 1 word ROMA ends with the stationary unit /a/, while all words beginning by unit /e/ stem from the same tree node. Stationary units are shown, in the figure, as circles, transition units nodes are depicted as ellipses, word identifiers as rectangles, and the dummy reactivation root nodes by diamonds. Every node, excluding dummy nodes, is associated to its Markov Model including one or two states.

A standard Viterbi decoding algorithm activates the dummy silence node at the beginning of the sentence. It expands the successor nodes of the lexical tree according to the beam-search strategy, until a terminal node associated to a word identifier is reached. A word is inserted in the lattice of word hypotheses if its likelihood is greater than a predefined threshold. The tree is then reactivated by expanding the successor nodes of the dummy reactivation node corresponding to the ending phoneme of the word. These nodes are the transition nodes toward the beginning of all the word in the dictionary, including function words and extra linguistic phenomena that are not shown in the figure. Notice that the multi rooted tree merges the initial transition nodes into the stationary unit nodes that model the beginning of the words.

The training database that has been used for these experiments does not include all possible transition units. Therefore, the transition unit that may appear as junctures of two vocabulary words have been generated by tying the states of similar units.

It is worth noting that, opposite to other approaches, this lexical tree organization is also well suited to decode continuous speech using a bigram language model (LM). To achieve this goal, the word bigrams are represented in terms of a network that merges the general lexical tree

and a set of bigram subtrees [1, 9]. The bigram probabilities are distributed along the tree branches. In the network there is one multi rooted subtree for each word appearing in the corpus used for training the language model. Bigram subtrees offer the attractive properties of the lexical tree organization while also allowing to keep separate theories that differ for the last word and to anticipate the application of word bigram LM probability.
Using this data structures, the overhead due to the inter-word transitions is minimal because the LM decoding algorithm simply expands the path reaching a terminal node by reactivating the successors of two dummy nodes only: those identified by the final phoneme of the ending word in the general lexical tree, and in the word bigram subtree.

4 Results

The training database [2] includes about 6,000 spontaneous sentences, collected from naive users through a PBX; the test set consists of 858 sentences.
Three different unit sets were compared:

- a set of 313 intra-word biphones and triphones;
- a set of 459 stationary/transition units without word junctures;
- a set of 537 stationary/transition units with word junctures.

All tests were performed without grammar constraints. The tests were carried out aiming at a preliminary evaluation of the method, and not to optimize absolute recognition performance figures, in fact only the spontaneous speech component of the available corpus was used for training. A large read speech corpus is being added to the training set, and an overall assessment of the methodology is under way.

units	313	459	537
WA	67.5	68.1	69.8

Table 2: Word Accuracy figures for different unit sets

We believe that a further improvement might be obtained by separately modeling inter- and intra-word transition units according to generally accepted phonetical principles [11]. With this approach, the large increase in the number of possible coarticulation units would require the adoption of suitable techniques for dealing with their prediction and interpolation.

5 Conclusions

We have proposed an efficient method to deal with inter-word coarticulation units in continuous speech recognition. The approach combines transitory/stationary state representations and explicit modeling of word junctures. Little additional effort in computational load is required.

References

[1] G. Antoniol, F. Brugnara, M. Cettolo, and M. Federico, "Language Model Representations for Beam-Search Decoding", *Proc. of the ICASSP 1995*, Detroit, pp. 560–563, 1995.

[2] P. Baggia, E. Gerbino, E. Giachin and C. Rullent, "Experiences of spontaneous speech interaction with a dialogue system", CRIM/FORWISS Workshop, Muenchen, september 1994.

[3] L. Fissore, E. Giachin, P. Laface, and G. Micca, "Selection of Speech Units for a Speaker-independent CSR Task", *Proc. EUROSPEECH 91*, pp. 1389–1392, Genova, Italy, 1991.

[4] D. Jouvet, L. Mauuary, J. Monné, "Automatic Adjustments of the Structure of Markov Models for Speech Recognition Applications", *Proc. EUROSPEECH 91*, Genova, Italy, pp. 927–930, 1991.

[5] M. Hwang, X. Huang, "Shared-Distribution Hidden Markov Models for Speech Recognition", *IEEE Transactions on Speech and Audio Processing*, Vol. 1, n. 4, Oct. 1993.

[6] L. Fissore, F. Ravera, P. Laface, "Acoustic-Phonetic Modeling for Flexible Vocabulary Speech Recognition", *Proc. EUROSPEECH 95*, pp. 799–802, 1995.

[7] P. Laface, L. Fissore, and F. Ravera, "Automatic Generation of Words toward Flexible Vocabulary Isolated Word Recognition", *International Conference on Spoken Language Processing*, Yokohama, Japan, pp.2215–2218, 1994.

[8] K. Lee, "Context-dependent Phonetic Hidden Markov Models for Speaker-independent Continuous Speech Recognition" *IEEE Trans. ASSP*, Vol.38, n.4, April 1990, pp. 599-609.

[9] P. Laface, L. Fissore, et alii, "Segmental Search for Continuous Speech Recognition", To appear in *International Conference on Spoken Language Processing*, Philadelphia, 1996.

[10] H. Ney, A. Noll, "Acoustic-Phonetic Modeling in the SPICOS System", *IEEE Transactions on Speech and Audio Processing*, Vol. 2, n. 2, Apr. 1994.

[11] G. Marotta, D. Ricca, P. L. Salza "Duration and Formant Frequencies of italian bivocalis sequences" Proc. of ICPhS 1987

[12] R. Schwartz, Y. Chow, S. Roucos, M. Krasner, J. Makhoul, "Improved Hidden Markov Modeling of Phonemes for Continuous Speech Recognition , " *Proc. of the ICASSP 1984*, pp. 35.6.1-35.6.4, 1984.

[13] L.C. Wood, D.J.B. Pearce, F. Novello, "Improved Vocabulary-Independent Sub-Word HMM Modelling", *Proc. of the ICASSP 1991*, pp. 181–184.

[14] S.J. Young, P.C. Woodland, "The Use of State Tying in Continuous Speech Recognition", *Proc. EUROSPEECH 1993*, Berlin, 1993, pp. 2207–2210.