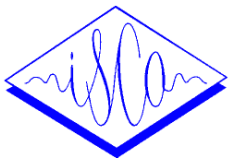

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

ACOUSTIC-PHONETIC MODELING FOR FLEXIBLE VOCABULARY SPEECH RECOGNITION

L. Fissore ◇ *and F. Ravera* ◇ *and P. Laface* ⋆

◇ 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 focuses on the definition and modeling of robust context-dependent units for flexible vocabulary recognition. It proposes a new technique for tuning the acoustic resolution of the models, and discusses the advantages of representing phonetic transcriptions in terms of a sequence of stationary context-independent phonemes and diphone-transition coarticulation units rather than with the classical diphone or triphone units. Combining these two techniques, the recognition rate of a speaker-independent recognizer with a vocabulary of 600 surnames increases from 91.2% to 96% using less than one third of the densities of the original models.

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. Since the same phoneme is pronounced in different ways according to its acoustic context, context-sensitive phonetic models (triphones) are generally used for taking into account the coarticulation effects. Increasing the number of different context dependent models, however, does not necessarily increase their acoustical accuracy. On the contrary, it is possible that some models are undertrained due to the reduced amount of observations that are assigned to them. Moreover, even if it could be possible to accurately train the triphones appearing on a specific application vocabulary, it is not feasible to train all triphones of a language to cover a new vocabulary. In principle, it is possible, instead, to train with an appropriate database a complete set of diphones that are able to represent any new vocabulary. To create trainable and consistent units several solutions have been proposed that can summarized into three main classes:

- context-independent phoneme modeling [7] 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 [9].
- parameter sharing [5] by tying similar units [5, 1] or similar distributions [3, 10].

To take advantage of the generality of the diphones without loosing the accuracy of the triphones we have exploited some good properties of the first and third class of solutions by defining stationary context independent and diphone-transition units and tying the distributions on the base of their acoustic-phonetic similarity.

2. Unit modeling

The recognition system described in this paper is based on Continuous Mixture Density Hidden Markov Models of subword units. The units include context independent phonemes and context-dependent units (CDU). The CDU set is selected according to a minimal occurrence criterion within the training database [1]. Every unit has three states without skip transitions, while silence is modeled by a single state. The first and the last state of each unit are supposed to model the coarticulation effects occurring during the transition from the previous and next phoneme respectively, while the stationary part of the phoneme should be modeled by the central state of the unit. This definition of the units

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>	...

presents several drawbacks if the issue of robust training for flexible vocabulary applications has to be taken into account:

- The central state of a triphone (l)p(r), which represents the stationary part of phoneme <p>, is trained using context dependent samples only. The resulting distribution is, thus, very detailed, but it lacks generalization capabilities.
- The distribution of the final state of a left context-dependent diphone (l)p is trained by merging observations occurring in many different right contexts. Unless this distribution is modeled with a large number of emission densities, it is often too smoothed because it merges the coarticulation effects of the next contexts. The same weakness applies to the first state of a right context-dependent diphone p(r).
- A context-independent phoneme <p> is trained only with the observation vectors that have not been used for training the CDU. If many different CDU have been defined, the resulting context-independent models can be undertrained.

To avoid these drawbacks, the definition of a new set of units is proposed where the central state of a context-dependent unit <(l)p(r)> is tied to the central state of all the other context-dependent units of the same phoneme <p>: the stationary state of phoneme <p> is, therefore, trained as much as possible as a context-independent unit. The final state of phoneme <p> is connected to the first state of the next phoneme <q> leading to a two-state diphone-transition unit <pq>. A phonetic transcription is, thus, represented as a sequence of stationary context independent phonemes <p> and diphone transition units (TU) <pq>, as shown in Table 1.
Since each unit is modeled by a sufficient number of observation frames of the same context, it is robust and has good generalization capability. Moreover, the accuracy of the triphones is not lost because a generic CDU triphone <(l)p(r)> is still modeled by the 3 states of the TU sequence <lp><p><pr>.
This set of units corresponds to the pseudo-diphones presented, in another context and for the limited domain of connected digits recognition in [2].
Of course, different topologies of the basic units can be devised and experimented to obtain better spectral or duration modeling.

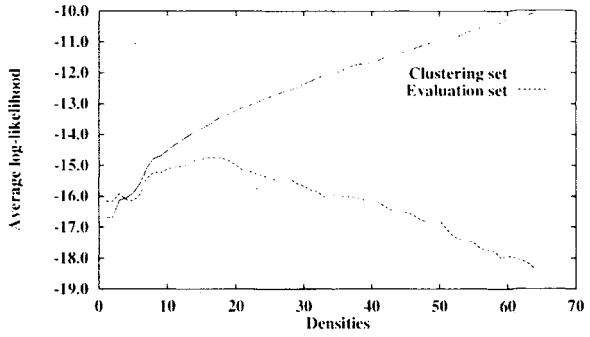

Figure 1: Average log-likelihood of the frame assigned to a state

3. Mixture selection

The second contribution of this paper is on acoustic resolution. The emission densities in our system are modeled by mixtures of Gaussian densities. In many systems, the maximum number of allowed densities per state is fixed and equal for each state. This value is a priori selected according to the size of the training database. Increasing the number of densities per state generally leads to more detailed models and better recognition results. The number of densities for each state must be, however, carefully selected to fit the actual distribution of the training data not only for reducing memory and computation costs of the recognizer, but also to avoid the overtraining of some states. This adaptation to the size and to the distribution of training data can be performed by a strategy [8] that splits a density if its average distortion is greater than the grand average distortion computed over all densities.
In our approach, instead, the adaptation of the number of densities of a state to the 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 respectively in our experiments). Segmental Viterbi training, and alignment

Table 2: Recognition rate and average number of densities per state

Set	Mixtures	Max allowed mixture size	4	8	16	32
210 CDU	Variable	Average number of densities	2.6	4.4	6.4	8.2
		% Recognition rate	90.4	91.5	91.8	92.2
625 states	Fixed	Average number of densities	4.0	7.8	14.7	25.2
		% Recognition rate	90.3	91.2	91.2	91.2
346 TU	Variable	Average number of densities	2.3	3.7	5.2	6.4
		% Recognition rate	94.7	95.2	95.6	96.0
626 states	Fixed	Average number of densities	3.8	7.2	12.8	20.9
		% Recognition rate	94.6	95.0	95.1	95.3

of the training observations to each state is then performed using an available set of models and the *complete* database.

For each state, K-means clustering is performed using the segmentation of the *clustering* subset obtained from the previous step, and increasing the number of required densities until it reaches a pre-set maximum value or the average likelihood of the observations in the *evaluation* subset decreases (a clear cue of overtraining). The number of densities associated to the state of each model is, thus, adapted using a subset of the training data and evaluated on an independent subset.

This procedure is then iterated using the automatic segmentations obtained performing Segmental Viterbi training and alignment of the *complete* training database using the new models.

Fig. 1 shows the typical behaviour of the average log-likelihood computed for the set of frames assigned to a stationary state as a function of the number of densities used to model that state. As expected, the average log-likelihood of the frames belonging to the clustering subset steadily increases with the number of densities. For the evaluation set frames, instead, it increases until it reaches its maximum value because the accuracy of the model improves with more densities, then it decreases because the model becomes too specific for the clustering subset: the model lacks generality because it uses too many densities.

By selecting a variable number of densities for each state according to the above illustrated strategy, the acoustic resolution of each state is adapted to the data. It comes out that the stationary states, although trained as context-independent units, are assigned several densities which may take into account the coarticulation effects with the neighbour transition units. The total number of densities used for a given set of models is reduced by the effect of this adaptation strategy to about 1/3, but more accurate and robust models are trained as confirmed by the results illustrated in the next Section.

4. Results

The effectiveness of the diphone-transition units and of the state mixture selection has been assessed running several experiments with a recognition system based on 210 Continuous Density HMM of subword units. The units include 27 context-independent phonemes, 176 right context-dependent diphones and 7 models for extra linguistic phenomena. The units were trained using an isolated word speech database including a total of 12000 surnames collected by 173 speakers that pronounced entries in the phone directory of CSELT. The test database includes a total of 12720 utterances of a set of 600 different surnames, pronounced by 120 speakers and collected through a PABX [4]. The results of previous experiments [6] suggest the use of diphones, rather than of triphones, for this application.

In the first experiment we compared the performance of the new set of transition unit (TU) including 346 units and the old right context-dependent CDU set using the same number of densities per state. It is worth noting that the total number of states for the new and the old sets are almost equal (626 and 625 states respectively). We stored for each state the automatic segmentation derived from the last iteration of the Segmental Viterbi algorithm using the transcription of the training database in terms of the original CDU set with 32 Gaussian densities per state. Using this segmentation and just tying the states according to the above defined strategy to generate stationary and diphone-transition units, the recognition rate of the system raises from 91.4% to 94.1%. This result clearly shows that this definition of a transition units is able to reduce dramatically the problem of a blurred distribution in the first state of a right context-dependent diphone.

The results of the remaining experiments are summarized in Table 2 and compared in Fig. 2 that shows the

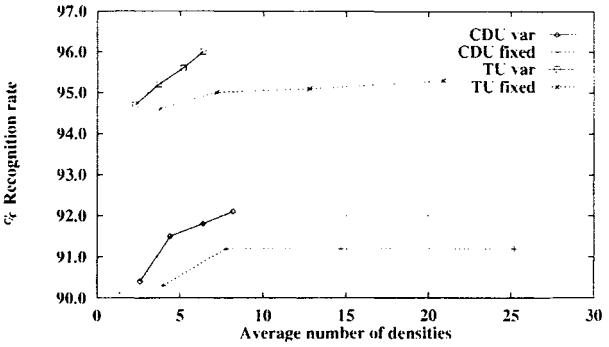

Figure 2: Recognition rate

recognition rate obtained using four set of units: the new set of transition unit (TU) and the old CDU set with fixed and variable number of densities per state respectively. Notice that even for the fixed mixture case, the mixture size of a state is almost always less than the maximum allowed value because some mixtures are eliminated during training because their weight does not exceed a pre-set threshold.

For every value of the maximum allowed number of densities, and using a fixed number of densities per state, the TU set achieves an absolute increase of about 4% in recognition rate with respect to the old CDU set (compare the rows referring to fixed mixtures in Table 2). Furthermore, the average number of densities per state for the TU set is always less than the corresponding value of the CDU set.

The second experiment gives some insights on the effect of the acoustic resolution of the models because the same set of units are used (the CDU set), but with fixed and variable number of densities per state. The set with a variable number of densities per state (CDU Variable) always achieves better results using much less densities. In particular, for a maximum allowed number of 32 densities, 1% increase in recognition rate is observed using less than 1/3 of the densities.

The same behaviour is exhibited by the new TU set whose recognition rate steadily increases with the size of the mixtures and outperforms the old CDU set starting from 94.7% and reaching 96.0% using an average of 6.4 (rather than 20.9) densities per state. Notice that using the same total amount of Gaussian densities (6.4 * 625), the performance of the CDU set is 91.8% only, the densities are, thus, better distributed among the TU model states.

Finally, the quality of the TU set with respect to the CDU set has also been assessed comparing the recognition results using Discrete Density HMMs that have been trained bootstrapping the models with the segmentations obtained by the corresponding Continuous Density HMMs. The Discrete Density recognizer using the new models achieves 91.6% recognition rate, while the old CDU set best performance is 87.7%.

5. Conclusions

Two techniques for the definition and modeling of robust context-dependent units for flexible vocabulary recognition have been proposed that allow to sensibly increase the performance of a speaker-independent recognizer. These new units represent a step toward the definition of an "universal" set of models useful for modeling flexible vocabularies.

6. References

[1] 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.

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

[3] 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.

[4] P. Laface, L. Fissore, and F. Ravera, "Automatic Generation of Words toward Flexible Vocabulary Isolated Word Recognition", Proc. ICSLP 94, Yokohama, Japan, pp.2215–2218, 1994.

[5] 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.

[6] L. Fissore, G. Micca, F. Ravera, "Incremental Training of a Speech Recognizer for Voice Dialling-by-Name", Proc. ICSLP 94, Yokohama, Japan, pp.447–450, 1994.

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

[8] H. Ney, "Experiments on Mixture Density Phoneme Modelling for the Speaker Independent 1000-Word Speech Recognition DARPA Task", Proc. of the ICASSP 1990, pp. 713–716, 1990.

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

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