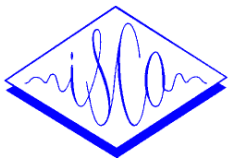

2nd European Conference on
Speech Communication and Technology
EUROSPEECH '91
Genova, Italy, September 24-26, 1991

Selection of Speech Units for a Speaker-Independent CSR Task *

L. Fissore ◇ E. Giachin ◇ P. Laface ⋆ G. Micca ◇

◇ CSELT - Centro Studi e Laboratori Telecomunicazioni
Via G. Reiss Romoli 274, I-10148 Torino, Italy
⋆ Dipartimento di Automatica e Informatica - Politecnico di Torino
Corso Duca degli Abruzzi 24, I-10129 Torino Italy

Abstract

This paper focuses on the problem of finding a set of Hidden Markov Models that can be trained to model context dependencies with good statistical accuracy, given the constraint of a fixed amount of training data.
Two aspects have been investigated in this work: clustering of intra-word context-dependent units with similar contexts on the basis of different similarity measures, and definition of inter-word coarticulation units.
A Dynamic Programming procedure is presented that allows a large set of context-dependent units to be clustered into a given number of units while optimizing a global cost measure. Inter-word units were found to provide better phonetic representations of word junctures and to increase recognition accuracy, though less than it has been reported for the English language.

1 Introduction

Recent developments of the Continuous Speech Recognizer being developed at Cselt labs [1] as a component of a Speech Understanding System [2] have been focused on the definition of context-dependent units for robust speaker-independent HMM modelling.
Acoustic variability of speech comes from many sources: speakers, phonetic context, environmental noise, speaking rate, and others. It has been demonstrated that the specific phenomena which cause the same phoneme to be pronounced in different ways according to the phonetic context can be successfully modeled by context-sensitive speech units. Increasing the number of different context-dependent models, however, does not necessarily increase their acoustical accuracy. On the contrary, it is possible that their statistical robustness decrease, due to the reduced amount of observations that are used for training each model. An interesting solution to this problem has been proposed in [3] where an agglomerate clustering approach, with some heuristic enhancements, has been introduced for detecting similar context-dependent triphones to be merged together; these merged units are referred to as "Generalized Triphones". The results reported in [3] show that, given the same database, a given number of generalized triphones perform significantly better than the same number of context-dependent units obtained through an explicit definition of the contexts.
In this work, in addition to the statistical significance of the model parameters obtained, another important constraint is also taken into account: the total number of units that a hardware implementation for a real time application is able to host and to process. The first part of this paper (Section 2) presents an approach for the automatic selection of a predefined number of Generalized Triphones. A Dynamic Programming procedure is proposed that selects the best mixture of triphones and biphones which optimizes a given cost function. Different cost functions associated to the merging of Discrete Density HMM units have been defined and the resulting unit sets compared.
The second aspect of the paper (Section 3) is related to the coarticulation phenomena which happen to appear at the adjacency points of contiguous words [7]. Boundary units model speech segments lying at the borders of words; inter-word units take into account the phonetic context. Several ways of combining these types of units into a working set have been tested. Moreover, a special architecture for a Viterbi-like decoding procedure has been devised to run a recognition system, based on inter-word unit HMMs, on a sequential machine taking also into account word-pair constraints.

2 Generalized Triphones

2.1 Baseline models

The reference unit set (REF) previously developed [4] consists of 305 units, of which 27 are the basic Italian phoneme set, 57 are function-word dependent units derived from 25 principal connectives for the Italian language (mono- and bi-syllabic conjunctions, articles, prepositions), 221 are context-dependent phones of which 108 are triphones and 113 are biphones. The context-dependent units were selected on the basis of their occurrence within the training database which consists of 8800 sentences uttered by 110 speakers [1]: each unit appears at least 500 times in the training set. The resulting coverage of the training sentences in terms of context-dependent units is 82%, distributed as follows: 4.3% function-word phones, 39.1% triphones and 32.1% biphones. The best Word Accuracy result obtained with Smoothed Discrete HMMs (two 8 bit codebooks for Mel-scaled cepstral and dif-

*This work has been partially supported by the ECC Esprit II Project 2218 – SUNDIAL:"Speech UNderstanding and DIALogue".

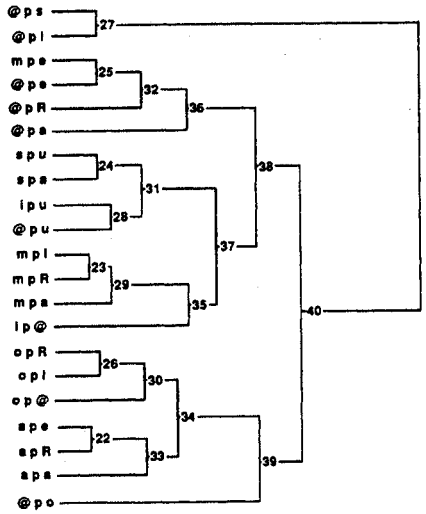

Figure 1: Dendogram of phoneme /p/

unit	occurences
CLUSTER n. 1	
@ p s	118
@ p i	530
CLUSTER n. 2	
m p e	91
@ p e	368
@ p R	677
@ p a	428
CLUSTER n. 3	
s p u	117
s p a	64
i p u	74
@ p u	311
CLUSTER n. 4	
m p l	77
m p R	227
m p a	242
i p @	58
CLUSTER n. 5	
o p R	100
o p i	112
o p @	67
a p e	100
a p R	207
a p a	79
CLUSTER n. 6	
@ p o	981

Table 1: Clusters of context-dependent units selected for phoneme /p/

ferential coefficients, and one 5 bit codebook for the energy and delta-energy parameters) is 65.7%, measured on a 1200 sentence test database pronounced by 20 speakers.

2.2 Clustering Procedure

Given the total number of units required, and a function related to the cost of merging two units, our goal is to select the identity and number of units to be merged in order to optimize a global cost function. It is worth noting that only context-dependent units of the *same phoneme* are taken into account for merging. The procedure works in two passes. In the first one, a complete-linkage hierarchical clustering analysis is performed to obtain, for each set of context-dependent units of the same phoneme, the similarity levels (cost) at which the clusters are joined. A dendogram resulting from this clustering operation is shown in Fig. 1 for the context-dependent units of phoneme /p/. By varying the setting of a similarity level threshold according to the values associated with the levels at which the clusters are joined, a number of clusters ranging from 1 to the cardinality of the set (in this case 21) is obtained. In the second pass a Dynamic Programming procedure finds, for each set of context-dependent units of the same phoneme, the similarity level value that allows the total number of required clusters to be obtained with the minimum cumulated cost. The algorithm is detailed in Fig. 2.

All clustering experiments have been carried out by using different cost functions and by setting to 305 the cardinality of the set of clustered units in order to compare their performance with respect to the REF set.

The first cost function that has been used in our experiments is the similarity measure that was found to be effective in the SPHINX system [3]: it is based on the loss of information that results by merging together two models. By weighting the Loss-of-Information by the occurrence counts computed during the training of the models, the clustering algorithm tends to merge models with fewer occurrences; therefore, these models become more robust and smoother at the same time. The Weighted Loss-of-information unit set (WL) was obtained starting from a trained set of 1157 units (S-1157) containing 929 triphones and 144 biphones, selected by setting to 50 the minimal occurrence count in the training database. A typical cluster set for phoneme /p/ is shown in Tab. 1.

	REF	*OCC*	*WL*	*LI*	*AT*	*S-1157*
WA	65.8	66.2	66.0	60.9	66.9	66.4

Table 2: Word Accuracy results for different speech unit sets.

The results, given in terms of Word Accuracy (WA) averaged on a test database of 20 speakers, 60 sentences each, are presented in Tab. 2, which includes the performance obtained by using the S-1157 set. In order to compare the contributions of occurrence counts and of the entropy of the models on the clustering, two other sets (LI and OCC) were defined using as cost function the Loss-of-Information only for the former, and the occurrence counts (OCC) only for the latter. Of course, the OCC set is generated by maximising the number of occurrences.

Table 2 shows that the OCC set gives results comparable with those of the WL and of the REF set, while a substantial decrease of performance derives from the use of the LI set. As the occurence counts seem to be relevant for the robustness of the models, another set of units (AT) has been defined starting from the complete set of triphones (1671) of the training database. The clustering procedure was again applied to generate a set of 305 units using as cost function the *number of occurrences of the units in the database* rather than the occurence counts. This experiment did not require, therefore, any preliminary training step. This set gave the best results, supporting the hypothesis that the amount of training data is a major parameter for robust modeling.

```
INITIALIZATION
for f = 1,...,nf
   for i = 1,...,nu
   D[f,i] = infinite
for e = 1,...,ne[1]
   D[1,e] = lev[1,e]
   bpi[1,e] = 0
   bpe[1,e] = e

ITERATE
for i = 1,...,nu
   for f = 2,...,nf
      for e = 1,...,ne[f]
         d = D[f-1,i] + lev[f,e]
         if d < D[f,i+e] then
            D[f,i+e] = d
            bpi = i
            bpe = e

BACKTRACK
f = nf
i = nu
e = bpe[nf,i]
while i >= 1 do
   tl[f] = lev[f,e]
   i = bpi[f,i]
   e = bpe[f-1,i]
   f = f - 1

NOTATION:

nu       : number of clusters required
nf       : number of phonemes
ne[f]    : number of units for phoneme f
lev[f,e] : similarity level at which e clusters
           are obtained for phoneme f
D[f,i]   : minimum cost obtained by clustering
           the first f phonemes into i units
bpi[f,i] : backpointer through the best path
bpe[f,i] : backpointer for retrieving e
           (the number of clusters for phoneme f)
tl[f]    : the similarity level value that allows
           e clusters to be obtained for phoneme f
```

Figure 2: The cluster selection algorithm

Type	Code	Contexts	Description
$[x]y(z)$	1	Two-cont. (triphone)	y at left border of a word
$(x)y[z]$	2	Two-cont. (triphone)	y at right border of a word
$[x]y[z]$	3	Two-cont. (triphone)	y belongs to a single-phone word
$[0]y(z)$	4	Right-cont. (biphone)	y at left border, any context at left
$(z)y[0]$	5	Left-cont. (biphone)	y at right border, any context at right

Table 3: Types of inter-word units

3 Inter-word phone modeling

3.1 Choice of units

Recent research has shown that the use of speech units specifically designed to model context-dependent phones lying at boundary of words is helpful for increasing recognition accuracy. The reason is that these units, called *inter-word* phones, permit to represent in greater acoustic detail the coarticulation of neighboring words.

The minimal occurrence criterion has been used to define inter-word units, starting with triphones and clustering them to biphones if no enough examples are present in the training corpus. Inter-word units are considered different from intra-word units, even if they have the same contexts. In conclusion, there are five types of inter-word units, as shown in Table 3. Square brackets indicate a phonetic context given by an external word. Round brackets refer to a context given by the inner part of the same word the central phone belongs to. Zeros indicate clustered contexts.

Since inter- and intra-word unit sets are disjoint, they are separately modeled. A new set of intra-word units has been computed, with the purpose of eliminating units trained with segments lying at boundaries of words (these units are superseded by the more specific inter-word units).

3.2 Training and recognition algorithms

The training algorithm has been modified so that the correct units are placed at word boundaries. Optional silences are left between words and at sentence boundaries. In principle, both inter-word triphones and biphones might be used at word boundaries to allow different competitive descriptions of the word juncture. However, best results are obtained using only the most precise units, i.e. a triphone if it exists for the given contexts, a biphone if that is not the case, or a monophone if no biphones are suitable.

The hardest computational problems arise in the recognition algorithm. In addition to the overload due to the increased number of phones, inter-word units greatly complicate the search performed during Viterbi decoding, especially if a word pair grammar is used. The network which describes the lexicon has been organised as described in [7]. Besides, a special compilation procedure has been designed. It is based on the observation that, though the number of the network

Unit type		CI	CD	CD+IW	CD+IW
No. of units		27	312	312+54	312+185
No gram	WA	62.6	75.9	76.5	77.2
	SA	17.0	29.2	29.7	32.3
WP gram	WA	70.5	85.6	86.2	87.1
	SA	26.3	52.3	53.7	56.7
Bound. coverage		-	-	28.2	54.5

Table 4: Recognition results with different IW set sizes

nodes is very high, the number of nodes having *different sets of preceding nodes* may be far smaller. Hence, nodes with the same set of preceding nodes are grouped into classes. During Viterbi decoding, the determination of the best-scored incoming path is done once per class instead of once per node. This considerably reduces processing time. Also, a more focused beam search is performed: the presence of similar paths at word boundaries increases the number of paths having similar score, therefore the beam width may be let smaller than in the case of intra-word units.

3.3 Experimental results

Preliminary experiments have shown that inter-word biphones (types 4 and 5) are slightly detrimental with respect to the case in which their *intra*-word counterparts are used at word boundaries. Inter-word biphones are probably "fragile" because they are trained on less data. (This also suggests that inter-word coarticulation is not so specific, and that more robust models may be obtained without distinguishing inter- and intra-word units when they have the same contexts). In the following experiments, only inter-word triphones (types 1, 2, and 3) have been used.

Results are shown in Table 4 and refer to *continuous density models*. The performance measures are the word accuracy, WA, and the sentence accuracy, SA, defined as the fraction of sentences for which there is no word error whatsoever. Four unit sets are evaluated: a set of 27 context-independent unit, the set of 312 context-dependent intra-word unit described above, a set of 366 (312 + 54 IW) units (count threshold 350), and a set of 497 (312 + 185 IW) units (count threshold 150). Tests have been performed with no grammar and with a word pair grammar of perplexity 145. All units have 3 states per model, 15 mixture components per state.

Both sets containing inter-word units exhibit beneficial effects on recognition. The latter set performs clearly better than the former one, because many more word junctures are covered by at least one inter-word unit and hence are described more accurately (last row of the table). This suggests that, for inter-word units, further lowering the threshold count does not result in sensible degradation of statistical reliability. Also, many short function words (like articles, prepositions, etc.) are recognized more reliably. This is because short words consist mostly or exclusively of inter-word units, thus being sensitive to their quality. The relative word error reduction is about 4% and 10% for the no grammar and the WP grammar respectively. These figures are about half of those reported for American English [7], which may be due to the higher phonological stability of spoken Italian language.

4 Conclusions

Results show that a simple cost function based on the numbe of occurrence of the context-dependent units in the trainin set is effective for automatically selecting a good set of mod els and that inter-word phones are particularly useful to giv accurate phonetic description of word junctures, though les than it has been reported for the English language.

References

[1] L. Fissore, P. Laface, G. Micca, R. Pieraccini, "Perfo mance of a Speaker Independent Continuous Speech Re ognizer, " To appear in Proc. NATO ASI on Speec Recognition and Understanding, Cetraro (Italy), 1-1 July 1990.

[2] P. Baggia et al., "A Man-Machine Dialogue System fc Speech Access to E-mail Information Using the Tel phone, " This Conference.

[3] K.F. Lee, "Context Dependent Phonetic Hidden Marko Models for Speaker Independent Continuous Speec Recognition, " *IEEE Trans. ASSP*, Vol.38, n.4, Apr 1990, pp. 599-609.

[4] L. Fissore, P. Laface, G. Micca, "Comparison of Discret and Continuous HMMs in a CSR Task over the Tel phone, " ICASSP '91, Toronto, Canada, 14-17 May, pp 253-256.

[5] E. Giachin, C. Rullent, "Linguistic Processing in Speech Understanding System, " To appear in Pro NATO ASI on Speech Recognition and Understandin Cetraro (Italy), 1-13 July 1990.

[6] P. Baggia, E. Gerbino, E. Giachin, C. Rullent, "Efficier Representation of Linguistic Knowledge in Continuou Speech Understanding, " To appear in Proc. IJCAI-9 Sydney, August 1991.

[7] E.Giachin, C.H.Lee, L.R.Rabiner, A.E.Rosenberg, an R.Pieraccini, "Word juncture modeling using inter-wor context-dependent phone-like units, " This Conference