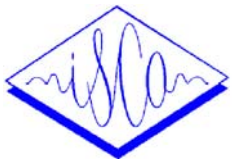

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

VITERBI BASED SPLITTING OF PHONEME HMM'S[1]

L. J. Rodríguez and M. I. Torres
Dpto. Electricidad y Electrónica. Fac. Ciencias.
Universidad del País Vasco. Apdo. 644. 48080 Bilbao (Spain)
E-mail: luisja@we.lc.ehu.es; manes@we.lc.ehu.es

ABSTRACT

Continuous Speech Recognition Systems (CSR) usually include large sets of context dependent units to model contextual variations in the pronunciation of phones. The goal of this work was to obtain adequate sets of sub-lexical models by using acoustic information but excluding any previous phonological knowledge. At each iteration of a classical Viterbi training scheme each acoustic model was split into a set of more accurate models. This approach was evaluated over a Spanish acoustic phonetic decoding task. The experimental results showed that this approach produces similar recognition rates than classical triphones.

10.21437/Eurospeech.1997-26

1. INTRODUCTION

Continuous Speech Recognition Systems (CSR) usually include large sets of context dependent units to model contextual variations in the pronunciation of phones [1-4]. Although phoneticians have defined sets of rules that explain such phonological variations, automatic methods selecting the most frequent triphones in a given training corpus, have been widely used [2]. Then the usually very high number of units is decreased by clustering contexts that produce similar effects [2] [5]. Alternatively, decision trees have been used as a tool to obtain sets of context dependent units [6]. This methodology considers both, the previous phonological knowledge of the language and the inductive knowledge automatically learned from the available training corpus. Nevertheless, the goal of all these approaches is to increase a previously established set of phone-like units by including, and then modelling, contextual information.

It is well known that acoustic diversity is mainly due to contextual effects in the pronunciation of phones. However, the phonological knowledge does not fully explain all the acoustic variability appearing in the training corpus of a CSR system. Moreover, the only use of this knowledge may condition, excessively, the finally obtained sets of units since task nor speaker dependencies are considered [1].

Our goal was to obtain adequate sets of sub-lexical models by using acoustic information but excluding any previous phonological knowledge. Such kinds of units are, as many other parts of nowadays CSR systems, task dependent. They summarise all the sources of acoustic variability appearing in a given task.

A previously established set of Spanish phone-like units [7] was used as the initial model. Then each unit was split into a set of more accurate models. In each split the set of samples that best matched the current acoustic model, under a maximum likelihood criterion, was considered to generate a new model. The remaining samples, not well enough represented, were used to train a second model. Both of them were considered for further splitting. An iterative procedure consisting of utterance segmentation / splitting / new model re-estimation was carried out until some convergence criteria were achieved [8].

Section 2 describes in detail the proposed methodology. In Section 3 the experimental environment used to evaluate this approach is described. Experimental results are presented and analysed in Section 4. Finally Section 5 presents some concluding remarks.

2. METHODOLOGY

Our baseline system included a set of phone-like units previously established for Spanish CSR [7]. Then, the training corpus was segmented by a Viterbi alignment of the sequence of codewords or acoustic feature vectors representing each utterance with the phone-like unit sequence that corresponds to the phonetic transcription of the

[1] Work supported by the Spanish CICYT under grant TIC95-0884-CO4-03.

sentence uttered. As a consequence a set of samples consisting of vector or codeword sequences, was obtained for each sub-lexical unit. Then, the objective was to split such sets of samples into several subsets representing acoustic variability of the unit. These subsets could be used to train several new acoustic models of the original phone-like unit.

The splitting was done under the assumption that the probability of each set of samples being generated by the model of the corresponding sub-lexical unit can be approximated by a normal distribution:

$$p(H_i / M_i) \approx \mathcal{N}(\mu_i, \sigma_i^2)$$

where H_i is the set of training sequences corresponding to model M_i and μ_i, σ_i^2 are the parameters of the normal distribution.

Thus we considered that samples with values of $p(x_j / M_i), x_j \in H_i$, greater than a given threshold were well represented by model M_i (*good* segments) and corresponded to the dominant variety, whereas samples with low probability values (*bad* segments) corresponded to marginal varieties. Thus, the set of segments corresponding to each sub-lexical unit was split into two subsets: *good* and *bad* segments and the initial model was then replaced by two new models, both of them trained with the corresponding samples. In this procedure an evaluation function was also considered to assess the *goodness of the split* (*gos* function). This function was proposed in [6] and is given by:

$$gos(i) = \log\left(\frac{P(H_{ig} / M_{ig})P(H_{ib} / M_{ib})}{P(H_i / M_i)}\right)$$

where H_{ig} and H_{ib} are the good and bad subsets of sample H_i used to train models M_{ig} and M_{ib} respectively.

Each new split required a minimum threshold for the *gos* function and a minimum threshold for the number of samples assigned to new models.

Thus, the whole training scheme consisted of two phases:

1. The initialisation phase, where an initial HMM model for each previously selected phone-like unit was learnt from a small subset of the training data which was segmented and labelled by hand.
2. The iteration phase which consisted of three parts:

- The entire training set was segmented by the Viterbi alignment procedure according to the current set of units. Then, $p(x_j / M_i), x_j \in H_i$ as well as the parameters of each distribution, μ_i and σ_i^2, were calculated.
- For each current unit we considered the split of the samples into *good* and *bad* segments according to an heuristically established threshold for the probability. However the split was only considered when the following conditions were verified:

  - the distribution was wider enough since sharp distributions were supposed to be homogeneous.

  - the number of samples in both subsets was high enough to guarantee a robust estimation of the new models.

  - the *gos* function exceeded a minimum threshold.
- The parameters of the new models were obtained by using the Baum-Welch re-estimation procedure.

The iteration phase finished when no splitting was made for any unit, i.e., when one of above mentioned stopping criteria (sharp distribution, small number of samples or low value of the *gos* function) was achieved for all the units.

3. EXPERIMENTAL ENVIRONMENT

The proposed formalism was experimentally evaluated on a Spanish acoustic-phonetic decoding task. A training corpus consisting of 842 sentences uttered by 43 speakers resulting in a total of 37,921 phones was considered. Forty-four sentences were previously segmented by hand to initialise the phone-like models. For testing purposes a test set consisting of 225 new sentences uttered by 17 new speakers resulting in a total of 12,800 phones were selected.

This corpus was acquired at 16 kHz and parametrised, resulting in vectors of dimension 11, i.e. 10 Cepstrum Coefficients (CC) plus energy (EN). From these parameters, we got their respective first derivatives (ΔCC and ΔEN).

All the sentences were automatically transcribed into sequences of sub-lexical units. The initial set of sub-lexical units used in the experiments to be presented was composed by 23 units that roughly

corresponded to the 24 Spanish phonemes. Each acoustic unit was represented by a left-to-right discrete HMM of three states without skips and only a self-loop transition in the second state.

The recognition model consisted of a simple finite state network since any phonological constraints were imposed (nor pair-gram, bigram, etc.). Thus, all models were placed in parallel, sharing a common initial state and a common final state which were connected trough a feed back-nil transition.

For each experiment, the following values were computed: 1) the number of sub-lexical units that were recognised correctly (*corr*); 2) the number of sub-lexical units that were inserted (*ins*), deleted (*del*) and substituted (*subs*). These values were obtained by an editing comparison between the output of the acoustic phonetic decoder and the correct sub-lexical transcription of each test utterance. From these values, the following parameter was obtained:

$$\%recognition = \frac{c}{i + d + s + c} \times 100$$

4. EXPERIMENTAL RESULTS

Two series of experiments, single and multiple codebook, were carried out.

In the first series of experiments two single codebook experiments were carried out. Two codebooks of 128 and 256 codewords respectively were obtained from vectors of dimension 11 (CC+EN). Table 1 (128 codewords) and Table 2 (256 codewords) show the experimental results obtained for these experiments.

A minimum of 50 samples per unit was required in both experiments. The number of units and recognition rates at each iteration of the first experiment are shown in Table 1. The previously established set of phone-like units was used in iteration zero. One of the three stopping criteria was achieved at iteration number 8 for all the current sub-lexical models.

In the second experiment a single codebook of 256 codewords was used. Table 2 shows the same results for this experiment.

In both experiments a small error reduction was achieved. For comparison purposes the same decoding experiment was carried using classical triphones (Table 3). The threshold required for the number of samples per unit in the training set determined in this case the number of unit to be used. Table 3 shows the decoding rates for single codebook experiments when 128 codewords were used.

Table 1: Decoding results for single codebook experiments when 128 codewords were used.

# iteration	Number of units	% recognition
0	26	43.35
1	50	44.12
2	80	44.39
3	122	44.95
4	164	45.11
5	185	45.56
6	198	45.65
7	203	45.13
8	205	45.39

Table 2: Decoding results for single codebook experiments when 256 codewords were used.

# iteration	Number of units	% recognition
0	26	44.14
1	48	45.58
2	80	45.63
3	124	45.91
4	183	46.45
5	216	46.92
6	237	47.19
7	245	47.21
8	247	47.15

Table 3: Decoding results for single codebook experiments (128 codewords) when classical triphones were used.

Threshold	Number of units	% recognition
175	30	44.65
105	60	45.43
70	89	44.98
35	286	45.46

Table 3 also shows a small improvement of recognition rates with the number of units. Nevertheless the recognition rates are similar to those obtained when the proposed approach was used (Table 1).

A second series of experiments was carried out using three different codebooks of 256 codewords: CC, ΔCC and EN+ΔE. For these experiments a minimum of 35 samples per unit was required. Thus, a higher number of units was obtained. In this case, one of the three stopping criteria was achieved for all the sub-lexical units at iteration 9. The number of units and recognition rates at each iteration of this experiment are shown in Table 4. This table shows a slight reduction of error rate when the proposed approach was used.

Table 4: Decoding results for multiple codebook experiments.

# iteration	Number of units	% recognition
0	26	54.62
1	48	55.68
2	83	
3	135	55.26
4	209	
5	301	
6	363	54.56
7	391	
8	392	
9	393	54.71

5. CONCLUDING REMARKS

The goal of this work was to obtain adequate sets of sub-lexical units by using acoustic information but excluding any previous phonological knowledge. A procedure to split HMM´s in Viterbi training was presented in detail. This procedure was evaluated over a Spanish acoustic-phonetic decoding task. The experimental results showed that the proposed produced similar decoding rates to those obtained when classical triphones were used. However more exhaustive experimentation including alternative stopping criteria is still needed to fully evaluate this methodology.

6. REFERENCES

[1] C.H. Lee, L.R. Rabiner, R. Pieraccini and J.G. Wilpon. "Acoustic modeling for large vocabulary Speech Recognition". *Computer Speech and Language*, N 4, pp. 127-165 (1990).

[2] K.F. Lee. "Context-Dependent Phonetic Hidden Markov Models for Speaker-Independent Continuous Speech recognition". *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP - 38, pp. 599-609 (1990).

[3] H. Niemann,. E. Nöth,. E.G. Schukat-Talamazini, A. Kiessling, R. Kompe, T. Kuhn, K. Ott,. S. Rieck. "Statiscal Modeling of Segmental and Suprasegmental Information". News Advanced and Trends in Speech Recognition and Coding, A. Rubio, J.M. López (eds). Springer-Verlag, Series F: Computer and Systems Sciences (1995).

[4] A. Bonafonte, R. Estany, E. Vives. "Study of subword units for Spanish speech recognition". *Proc. of European Conference on Speech Technology*, pp. 1607-1610 (1995).

[5] R. de Mori, M. Galler, F. Brugnara: "Search and learning strategies for improving Hidden Markov Models". *Computer, Speech and Language*, N 9, pp. 107-121 (1995).

[6] L.R. Bahl, P.V. de Souza, P.S. Gopalakrishnan, D. Nahamoo, M. A. Picheny. "Decision Trees for Phonological Rules in Continuous Speech", *Proc. of IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp.185-188 (1991).

[7] I.Torres, F. Casacuberta "Spanish Phone Recognition using Semicontinuous Hidden markov Models". *Proc. of IEEE Int. Conf. Acoust., Speech, Signal Processing*, Vol II, pp. 515-518 (1993).

[8] L. J. Rodríguez and M. I. Torres: "Splitting phoneme HMM's in Viterbi training". *Proc. of VII National Symposium on Pattern Recognition and Image Analysis*, Vol. II, pp. 42-43 (1997).