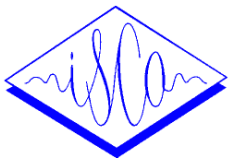

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

SPEAKER INDEPENDENT CONTINUOUS SPEECH RECOGNITION USING AN ACOUSTIC-PHONETIC ITALIAN CORPUS

B. Angelini, F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter, M. Omologo

IRST-Istituto per la Ricerca Scientifica e Tecnologica,
I-38050 Povo di Trento (Italy)

ABSTRACT

The objective of this paper is to describe the activity that is being carried out at IRST laboratories for the development of an HMM-based speaker independent continuous speech recognition system for the Italian language. The recognition system is trained and tested using the acoustic-phonetic continuous speech portion of the APASCI corpus. Acoustic modeling is based on the use of Continuous Density HMMs with gaussian mixture observation densities. As a baseline, a set of 38 Context Independent Units was evaluated using different numbers of mixture components. Then, two other classes of Context Dependent Unit sets were considered, that provide different performance and system complexity. Performance, expressed in terms of Phone loop recognition accuracy and Word loop recognition accuracy, shows an improvement using both of these classes of unit sets, with respect to the baseline.

I. INTRODUCTION

A baseline of a speaker independent continuous speech recognition system for the Italian language has been developed. The activity is oriented to provide a Hidden Markov Model based technology to design various recognition applications [1].

For large vocabulary recognition tasks, the choice of a suitable continuous speech corpus represents a critical step to obtain consistent and effective tools for training and testing the resulting technology. In order to model large sets of context dependent units a task oriented acoustic database may not be sufficient for a robust modeling due to the lack of data. To cope with this drawback a phonetically rich database can be collected and enriched, in successive steps, to provide further training material for refining models to be used in specific tasks.

Along these lines an acoustic-phonetic continuous speech database, called APASCI (Acoustic Phonetic And Spontaneous speech Corpus of IRST), was designed and collected at IRST.

The present release APASCI 2.0 includes 3900 phonetically rich utterances, segmented and labelled using the system described in [2, 3]. This procedure provided an optimal transcription of a given utterance, taking into account some phonological rules of the Italian language.

The recognition system is based on Continuous Density HMMs of acoustic-phonetic units. A preliminary baseline of this system was described in [4] with 37 context independent units. In the present work, different context dependent unit sets are considered, in order to establish a suitable compromise between system complexity and recognition performance. Two approaches were followed to choose these sets, one based on phone statistics, the other based on acoustic and phonotactic knowledge of Italian language.

This paper is organized as follows: first, a description of the speech corpus is given in Section II; then, a brief overview of the recognition system architecture is outlined in Section III; some recognition results are given in Section IV concerning the use of different unit sets; finally future work is described in Section V.

II. APASCI CORPUS

As mentioned above, the acoustic-phonetic continuous speech part of APASCI is conceived for training and testing a generic speaker independent continuous speech recognizer; similarly to TIMIT [5] and BREF [6], the acoustic-phonetic part of the corpus was designed with the purpose of collecting speech material as phonetically rich as possible. To achieve this result the procedure outlined below was followed.

First of all, a vocabulary of about 3000 words was defined, looking at the most frequent Italian words (including many functional words), and then considering other words covering a large variety of phonetic contexts. It is worth noting that overlapping with our potential tasks (robot telecontrol, radiological reporting, etc.) was not considered, while selecting the word dictionary.

Then, a grammar was used to randomly generate several sentences, which were also phonetically transcribed. All of the sentences were designed in order to be syntactically correct, even if often meaningless. Finally, a suboptimal procedure [4] was used to select a subset having good phonetic and diphonic coverage (that is a given number of occurrences for each diphone is ensured).

The acoustic-phonetic part of the present version APASCI 2.0 consists of 3900 utterances, read by 88 male and 88 female speakers (1 calibration sentence (*ca*) , 4 *ph* sentences and 20 (or 25) *di* sentences for each speaker). Most of the speakers come from North Italy and are aged between 22 and 50. Average length duration of the sentences is approximately 5 seconds, while the average number of words per sentence, computed on *ph* and *di* sentences, is 8.5.

Recordings were performed in a quiet room. Speech was acquired at 48 kHz, with 16 bit accuracy, by means of a Digital Audio Tape-Corder Sony TCD-D10PRO and a super-cardioid microphone Sennheiser MKH 416-T. Then digital recordings were downsampled to 16

kHz.

The whole corpus was divided into a Training set (50 males and 50 females), a Development set (18 males and 18 females), and a Test set (20 males and 20 females). Leaving out the calibration sentence, which is the same for all speakers, the training set, the Development set and the Test set are composed of 2140, 900 and 660 sentences, respectively. The dictionaries of Training and Development sets include 1383 and 1034 words respectively, that are common with the Test set dictionary. Other characteristics are given in Table 1.

Set	N. of sentences	N. of words	N. of CIUs	Dictionary size
Training	2140	18820	130272	2205
Develop.	900	6744	45368	1359
Test	660	5660	38407	1552

Table 1: *Statistics over Training, Development and Test sets.*

III. RECOGNITION SYSTEM

3.1 Acoustic processing

Each signal in the database was preemphasized using a digital filter having transfer function $H(z) = 1 - 0.95 \times z^{-1}$. The signal was blocked into frames by applying a 20 ms Hamming window every 10 ms. For each frame, 8 Mel scaled Cepstral Coefficients (MCCs) were extracted, using a 24-channel filter-bank. The log-energy, normalized with respect to the maximum value in the sentence, was also computed. Mean values of MCCs, computed on the whole utterance, were subtracted frame by frame. The resulting MCCs and the log-energy, together with their first and second order derivatives, computed on windows of 50 ms and 70 ms length respectively, were arranged in a single observation vector of 27 components.

3.2 Recognition Module

The recognition system is based on Continuous Density HMMs of acoustic-phonetic units. The recognizer training, based on Maximum Likelihood Estimation, was accomplished by using the available segmentation and labeling. Also, during training, unfrequently used gaussian components of a mixture were discarded.

Recognition was performed by applying the Viterbi algorithm on networks [7] representing a Phone Loop (PL) grammar or a Word Loop (WL) grammar. For each task, a Finite State Network (FSN) was compiled, which generates the desired language in terms of model sequences. The output of the acoustic recognizer is therefore a sequence of unit labels, which has to be translated into a sequence of words. The output does not necessarily correspond to a unique sequence of words: a procedure was developed that extracts all of the sentences corresponding to this output, and selects the one having the minimum number of word errors.

Both in the PL and in the WL recognition tasks, a balance between deletion and insertion errors was obtained introducing a small constant probability penalty for each phone (or word) spotted during the Viterbi search. Tuning of these probabilities was always accomplished using the Development set material.

3.3 Acoustic-Phonetic Modeling

As discussed in the following, the baseline system utilizes 38 context-independent phone models. Preliminary results obtained using a very similar set as well as the description of the first release of our database were given in [4].

Context dependent HMMs become important when the goal is to improve recognition accuracy with large vocabularies. Two different approaches were investigated, aimed at identifying suitable unit model sets: one is intended to exploit acoustic-phonetic knowledge of the Italian language, while the other is based on statistics. Different unit sets were considered, but in the following, only the most significant ones will be described and detailed in terms of performance improvement.

Context Independent Units (CIUs) The baseline unit set is defined as follows: 7 vowels (a, e, E, i, o, O, u), 4 affricates(C, G, z, Z), 3 nasals (m, n, N), 6 occlusive bursts (p, t, k, b, d, g), 3 liquids (l, r, L), 2 closures (cl,vcl), 5 fricatives (f, v, x, X, S), 2 semivowels(j, w), 1 vowel tail (@epi), 1 glottal closure (@q), 1 breath (@res), 1 schwa (@sch), 1 silence and 1 pause (@sil and @pau).

In particular:

- @epi represents the tail of a vowel when characterized by low energy and fading formants.
- @sch represents schwa sound, present between an "occlusive burst" and a "r", as well as after a "r" preceding some consonants (e.g. occlusives).
- @res represents breaths that could be confused with fricative sounds.
- @q was used to label segments including phenomena like diplophonic sounds, often present at boundary between adjacent words ending and beginning with either the same or very similar phonemes.

Context Dependent Units (CDUs) Context dependent models have been widely used in the last years in order to better model phone acoustic realizations [6, 7, 8]. Generally, they are not trained as well as the context independent models (due to the lack of data), which are in this sense more robust, but they are able to better capture fine acoustic variations due to the context.

In this work, three sets of CDUs were considered and the contexts to model were selected taking into account their frequency in the training set. First, a set of 299 right-context dependent models was selected, given a frequency threshold of 50 occurrences; then, a second set was obtained adding the most frequent 512 triphone models to the right-context dependent ones. A third set of CDUs was obtained by selecting very frequent contexts included in the second set, namely 220 right-context dependent models and 220 triphone-models.

Finally, 38 CIU models were included in all the three mentioned sets. These sets had a final size of 337, 849 and 478 units, respectively. The 38 CIU models (of the system baseline) were used for bootstrapping the CDU models.

Syllable Unit HMMs (SUs) An alternative approach was investigated It consists in modeling a phone sequence in a single unit, starting from the beginning of the first phone and ending at the end of the last one. This sequence is characterized by having formant trajectories otherwise difficult to model. From the point of view of coarticulation the resulting representation is more coherent than the one provided by the usual CDUs.

The selection of phone sequences, that we called Syllable Units (SUs), was made according to the acoustic

context. Only consonant+vowel units were considered, in order to avoid ambiguities that could arise if they were used together with vowel+consonant units. Furthermore, a frequency threshold of 50 occurrences was used to select, or discard, a predefined SU.

In practice, starting from the baseline consisting of 38 units we have incrementally added:

- the occlusive+burst+vowel transitions and some occlusive+burst+liquid transitions: the resulting new set included 73 units;
- the liquid+vowel transitions, obtaining a new set of 90 units;
- the diphthongs and the nasal+vowel transitions, obtaining another set of 141 units.

This approach can lead quickly to a considerable increase of mixture components, as shown in Table 2; tying among mixtures can represent an effective method both to limit this drawback and to ensure robust training.

3.4 Topologies

A left-to-right HMM topology without skip among states was chosen for all the units, with the exception of silence, pause, and breath for which an ergodic model was adopted. In particular, a four state model was used for both CIUs and CDUs, with the exception of the "r" unit and the occlusive bursts, which were modeled by three state HMMs. Output distribution probabilities were modeled by mixtures of gaussian probability densities having diagonal covariance matrix. Transitions leaving the same model state shared the same output distribution probabilities.

Since explicit state duration is not included in these HMMs, at first the length (in terms of number of states) of the SU models had been chosen proportional to the mean duration of the elementary constituents CIUs. Also, skip among states had been introduced according to the variance of the unit duration. However, preliminary results suggested to simply use left-to-right topologies without skips and to select a model length equal to the sum of the corresponding CIU lengths.

IV. EXPERIMENTAL RESULTS

4.1 Evaluation Criteria

A set of experiments was carried out in order to assess performance of the resulting system configurations. Performance is reported in terms of both Phone Accuracy and Word Accuracy. Phone and Word Accuracy percentages are determined by $1-(Del+Ins+Sub)/N$, where N is the total number of units and *Del*, *Ins*, *Sub* represent the number of deletions, insertions and substitutions of units.

Phone Accuracy evaluation was always accomplished by reconducting transcriptions to the 38 CIU set. Actually, performance was computed after removing labels beginning with @: in this way, we believe that a more consistent result is obtained, since these labels do not provide any phonetic or linguistic information and should not affect word recognition. These labels are also present in the training documentation, since they can be included by the segmentation and labeling system. We observed that removing these labels both in the recognition output sequence and in the corresponding reference transcription did increase Phone Accuracy of approximately 4% (from 71.60% to 75.37% for the baseline using 16 mixture components).

A second aspect that deserves to be mentioned refers to the contribution due to the use of optimal reference transcriptions. They were generated taking into account both insertion-deletion phenomena (e.g. elision at word boundary, breaths, etc.) and other phonotactical phenomena. One could argue that using these transcriptions, for both training and testing of the recognizer, discrepancies and misleading results could be obtained. For the sake of clearness, a recognition experiment was conducted to compare output phone sequences either with optimal transcriptions of the test material, or with canonical transcriptions that were obtained joining the word phonetic transcriptions. In this case Phone Accuracy decrease was approximately 3% (from 75.37% to 72.56% for the 16 mixture component baseline). Nevertheless, reference transcriptions are quite correct (as reported in [3]) and are used in a consistent way for comparison purposes of the different recognition experiments.

4.2 Phone Recognition

Phone Accuracy for each unit set was evaluated using a FSN that, in principle, allows units to follow one each other, without using any phone statistics.

When using CDUs, phonotactic constraints were added to the FSN, in order to inhibit the recognition of unit sequences having incompatible contexts.

These constraints improved Phone Accuracy from 2% to 3%, depending on the unit set.

4.3 Word Recognition

Word Loop recognition, with CDUs, was carried out with a similar approach to ensure consistency at word boundaries. In this case, applying correct CDUs to a given word boundaries requires the knowledge of the preceding and following word.

This problem was addressed by using a two stage approach. In the first stage, for each sentence a Word Network (WN) was generated [9] using the 38 CIUs. In the second stage, the WN was modified by substituting each CIU with the corresponding CDU, according to its context.

In this way, the average number of successors for each word, dropped from the dictionary size to a more manageable number. There is a trade-off between the size of the WN (defined as the number of different words contained in the WN over the number of different words in the sentence) and the accuracy of the best word string contained in it (Word Error Rate lowerbound): we choose WNs having size 15.4 and Word Error Rate lowerbound 5.5.

4.4 Recognition Performance

For the various unit sets, results are summarized in Table 2. For each set, the total number of mixture components is reported to provide an indication of the corresponding system complexity. Using the set of 38 CIU models, performance was investigated for different numbers of mixture components. Increasing this quantity from 4 to 128, Phone Accuracy (and Word Accuracy) improved from 71.34%(64.82%) to 79.04%(73.14%). This trend suggests that further improvement could be attained increasing the number of mixture components.

CDU and SU sets were tested using 16 mixture components. Using CDU models, higher performance are immediately obtained with the CD337 set, but the trend of further improvement deserves next investigation. Due to the implicit complexity of the resulting recognizer, CD478 and CD849 sets seem to be less convenient. These sets include many triphones that

could not be modeled and used in an optimal way: as mentioned above, the APASCI corpus was conceived to ensure a high number and variety of phonetic and diphonic contexts, while triphonic contexts were not considered during its design. In conclusion, the CD337 set provides significant performance, definitely higher than the best obtained using context independent unit set CIU38-128, but with a comparable complexity.

Concerning the use of SU models, performance, especially in terms of Word Accuracy, shows the advantage of these types of units with respect to the use of CIU models. However, in this case, phone accuracy does not show a corresponding improvement, because insertion errors are more frequent than in the other cases. In fact longer unit models are penalized in the PL network. Probably we should take into account the different model lengths by imposing different weights in the PL network arcs.

Unit Set - N. of M.C.	Total N. of M.C.	Phone Acc. (% Correct)	Word Acc.(%)
CIU38-4	406	71.34(73.76)	64.82
CIU38-8	807	73.90(76.16)	66.78
CIU38-16	**1588**	**75.37(77.44)**	**69.63**
CIU38-32	3172	76.72(78.85)	71.01
CIU38-64	6094	77.98(80.02)	71.80
CIU38-128	11681	79.04(81.04)	73.14
CDU337-16	13040	81.36(84.00)	76.71
CDU478-16	18722	81.84(84.59)	77.01
CDU849-16	31441	82.44(85.38)	77.31
SU73-16	5161	75.89(80.30)	72.25
SU90-16	6717	75.38(80.51)	72.79
SU141-16	10613	76.60(82.52)	74.03

Table 2: *Recognition Performance on the Test set for various unit sets and number of Mixture Components.*

Investigation of the APASCI corpus labeling and of referred recognition results suggests that distinguishing between open and close vowels ("e" vs "E", "o" vs "O") of Italian language is not convenient anymore for recognition purposes. Further improvement can be obtained without this distinction. On the other hand, a more promising alternative approach seems to be the use of stressed vowel units both for Context Independent and for Context Dependent modeling.

V. FUTURE WORK

In this paper performance results of different sets of units are reported for a large vocabulary continuous speech recognition task. Complexity of the various sets, expressed in terms of total number of gaussian probability densities, are also reported. Preliminary experiments, carried out using the baseline set, on an independent recognition task consisting of dictated radiological reports [1] has provided satisfactory performance. Further experiments with this corpus will be performed in the future. Concerning acoustic modeling, a future activity will be devoted both to the individuation of other richer sets and to the reduction of mixture components in the predefined models. Also, unit sets and phonological rules will be investigated for a multiple stage recognition system in order to rescore the best hypothesis represented by a word graph [9], provided in the first stage.

References

[1] B. Angelini, G. Antoniol, F. Brugnara, M. Cettolo, M. Federico, R. Fiutem and G. Lazzari. Radiological Reporting by Speech Recognition: The A.Re.S. System. *In these Proceedings*

[2] F. Brugnara, D. Falavigna and M. Omologo. Automatic Segmentation and Labeling of Speech Based on Hidden Markov Models. *Speech Communication*, Vol. 12, no. 4, (1993):357–370.

[3] B. Angelini, F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter and M. Omologo. Automatic Segmentation and Labeling of English and Italian Speech Databases. In *Proceedings of Eurospeech-93*, Vol. 1, pp. 653–656, Berlin, September 1993.

[4] B. Angelini, F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter and M. Omologo. A Baseline of a Speaker Independent Continuous Speech Recognizer of Italian. In *Proceedings of Eurospeech-93*, Vol. 2, pp. 847–850, Berlin, September 1993.

[5] L. F. Lamel, R. H. Kassel and S. Seneff. Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus. In *Proceedings of the DARPA Speech Recognition Workshop*, pages 100–109, Palo Alto, California, USA, February 1986.

[6] L.F. Lamel and J.L. Gauvain. Cross-lingual Experiments with Phone Recognition. In *Proceedings of ICASSP-93*, Vol. 2, pp. 507–510, Minneapolis, Minnesota, 1993.

[7] X. Aubert, H. Umbach and H. Ney. Continuous Mixture Densities and Linear Discriminant Analysis for Improved Context-Dependent Acoustic Models. In *Proceedings of ICASSP-93*, Vol. 2, pp. 648–651, Minneapolis, Minnesota, 1993.

[8] K. Lee and H.W. Hon. Speaker-Independent Phone Recognition Using Hidden Markov Model. *IEEE Transactions on Acoustic, Speech, and Signal Processing*, 37(11):1641–1648, 1989.

[9] R. De Mori, D. Giuliani and R. Gretter. Phone-Based Prefiltering for Continuous Speech Recognition. *In these Proceedings.*