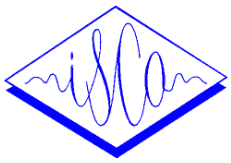

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

Automatic Generation of Words toward Flexible Vocabulary Isolated Word Recognition

P. Laface ⋆ L. Fissore ◇ F. Ravera ◇

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

Abstract

The paper deals with flexible and very large vocabulary isolated word recognition systems. In particular it discusses two main topics referring to a speaker independent isolated word recognizer: the evaluation of the recognizer performance with vocabularies of different size and with different entries, and the generation of "artificial" word utterances for fast lexical access to very large vocabularies.

1 Introduction

The recognition of directory entries is a relevant application of speech recognition technology in the telecom environment. The recognition of the surnames included in a company directory, typically uttered through the internal PABX, is feasible in real time with the current technology for a medium size vocabulary. On the contrary, it is a challenging research problem, the recognition of very large vocabularies, for an automatic directory service through the telephone line. Both applications, however, still offer open issues for their design and assessment. For of medium size vocabulary recognizer two important questions from the application point of view are:

- given a set of test utterances, is the recognizer robust to vocabulary variations?
- it is possible to assess the recognizer performance for a completely new, custom vocabulary, without collecting a new test database?

In this work we evaluated the robustness of a speaker independent recognizer with respect to different vocabularies by randomly generating a new vocabulary for each speaker and recognizing his utterances against this new set of words. Moreover, in order to avoid collecting a database for testing a custom vocabulary, we present a method for generating "artificial" utterances acoustically similar to the natural ones.

For the recognition of very large vocabularies, the main problem is, of course, the extremely large perplexity of the task and the similarity of the vocabulary words. We present an approach to very large vocabulary recognition based on a two-step strategy of fast lexical access to construct groups of acoustically similar words. This technique is similar to the one proposed in [1] with a relevant difference: in [1], to generate the word group lists, 10 speakers collected 20000 utterances, while in our approach we can easily generate and effectively recognize millions of different "synthetic" speaker independent utterances.

The rest of paper is organized as follows. Section 2 presents the main features of the recognizer and the vocabularies that have been evaluated. Section 3 describes the techniques for assessing the robustness of the recognizer when the vocabulary changes. In Section 4 a method for creating "artificial" utterances is presented, while in Section 5 we show how they have been used to construct lists of acoustically similar words for fast lexical access.

2 Features of the recognizer

The recognition system that has been evaluated and improved is based on a total of 203 Discrete Density Hidden Markov Models of subword units. The units include 27 context independent phonemes and 176 con-

Word	Occurrences	Word	Occurrences
ROSSI	40117	FERRARI	22905
RUSSO	21983	BIANCHI	18664
COLOMBO	17470	ESPOSITO	16245
ROMANO	13333	RICCI	12908
CONTI	11777	COSTA	10728

Table 1: Surnames occurring more than 10000 times in the Italian PSTN directory

Inclusion rate (%) for 600 word vocabularies		
Vocabulary	Number of hypotheses	
	1	5
Fixed	90.5	98.9
Uniform distribution	93.4	99.0
Weighted distribution	91.3	98.2
Artificial database	93.6	99.5

Table 2: Inclusion rate for 600 word vocabularies

text dependent diphones. Every unit, excluding the silence, is modeled by 3 states without skip transitions; a single state models the silence. The units were trained using both a continuous and an isolated word speech database. The first database includes 11680 sentences collected from 146 speakers. The second one contains a total of 12000 surnames collected by 173 speakers that pronounced a set of 950 entries in the phone directory of CSELT. The test database includes a total of 12720 utterances of a set of 600 surnames, pronounced by 120 speakers. The surnames in this database were selected from the 188892 entries that appear in the Italian general telephone directory at least two times. This subset of the most frequent 188892 Italian surnames - Table 1 shows those occurring more than 10000 times - will be referred to in the following as the Very Large Vocabulary (VLV). Part of the 600 surnames was selected from the top of the directory list in which the surnames appear sorted in descending order of occurrence, but many other were selected to assure a large phonetic coverage and as minimal pairs (/MILANI/ and /MELANI/ for example) leading to a rather difficult vocabulary. These databases were collected through a PABX.

3 Evaluation of the recognizer

The robustness of the recognizer with respect to different vocabularies has been assessed by using a new vocabulary for each new speaker. The new vocabulary includes the set of the words pronounced by that speaker and another set of words randomly selected from the VLV to reach the vocabulary size of 600 entries. The generation of a new vocabulary was performed by considering the entries in VLV uniformly distributed or by taking into account their frequency in the general directory.

The results of the experiments with a fixed set of models and parameters, and with different vocabularies of size 600, given in Table 2, show that the performance of the recognizer does not decrease using different vocabularies, on the contrary, slight better results are achieved because it is unlike to select randomly words from a very large vocabulary and to find them acoustically similar. Moreover, long, composite, or foreign surnames, can be included by a random selection, especially if it is not weighted by the occurrence frequency. These experiments were repeated 10 times with different seeds for the generation of the vocabularies: since the standard deviation of the obtained results is very low (of the order of 0.1%), we conclude that this set of subword models has good generalization capabilities.

4 Artificial utterances

Another important evaluation for a flexible vocabulary recognizer is its performance with respect to a custom vocabulary. This assessment would require, however, a new test database for each vocabulary. Of course it would be expensive or even impossible to collect such databases. Our solution to this problem is to generate "artificial" utterances for the test database by concatenating subword unit "templates". Stated more precisely, the problem is to produce several different spectral instances (in terms of frames) of a word, given its orthographic form. The task is similar to that of diphone-based text-to- speech systems, but it is much easier because rather than producing good quality synthetic speech, our goal is to obtain synthetic utterances that are similar to the natural ones according to an objective measure of similarity between words: the same used by the recognizer.

To generate automatically several "utterances" of a word, an archive of subword spectral "templates" has been created by using a Viterbi procedure to segment the *test* database. The archive includes more than 1 million frames segmented into 85282 templates of the 203 subword units. These templates are the basic elements for the generation of synthetic utterances.

Given the phonetic transcription of a word, in terms of a sequence of subword units, an utterance of this word

is produced by randomly selecting from the archive a template of each unit in the sequence. The spectral representation of the utterance is simply obtained by appending the frames of each selected template. It is worth noting that:

- the subword unit templates selected to generate a single utterance of a word are segments of different utterances of a large variety of speakers (both male and female).
- the coarticulation effects are partially take into account since the unit templates are context dependent.
- the frames included in this synthetic database did not contribute to train the parameters of the subword unit models.

To compare the effectiveness or similarity of the artificial utterances with the real ones, an artificial database has been created that includes the same 12720 word utterances included in the real test database. The recognition results for the artificial database tested against the 600 word vocabulary, reported in Table 2 are slightly *better* than the results obtained for the real database. This quite surprising result demonstrates that the unit templates are "good" acoustic examples because they have been segmented by their corresponding HMM, but it can also be explained by considering that the probability of a bad pronunciation of a word is greater than the probability of random selecting a set of acoustically bad templates to generate an artificial word.

5 Fast lexical access

The similarity of the artificial utterances with the real ones has been exploited for a two-step strategy of fast lexical access. In the first step, a fast matching module generates, by means of a small beam search threshold, a short list of word candidates. In the second step, another module adds to this list the words that are acoustically similar to each candidate, and rescores the resulting set with a larger beam search threshold or with more accurate subword models. In [1] it is proposed to derive off-line groups of acoustically similar words using training utterance recorded by several speakers. We followed a similar approach, but with a relevant difference: in [1], to generate the word group lists, 10 speakers collected 20000 utterances, while in our approach we can easily generate and effectively recognize millions of different "synthetic" speaker independent utterances. To this aim, we used "artificial" utterances generated from templates derived from the 950 word *training* database.

To construct these lists of acoustically similar words we need a set of "speaker independent utterances" of each word. Then, we perform the acoustic decoding of each utterance in order to evaluate the likelihood of each entry in the given vocabulary. Since the complexity of the decoding process is related both to the number of utterances and to the size of the lexicon, the generation process described in the previous Section is effective only if the artificial utterances do not belong to a very large vocabulary. Since the average duration of a word in our databases is about 0.9 sec, assuming the same average duration for the synthetic utterances, it would require more than 45 hours for a *real-time* recognizer to decode a single utterance of our 188892 word vocabulary. Since the generation of the confusion lists requires several utterances of the same word, for the sake of effectiveness, a different strategy is mandatory: we propose that the generation of an utterance and the related decoding process are carried out in parallel using a lexical tree.

Every vocabulary word is, thus, transcribed as a sequence of units, and the transcribed lexical entries are merged into a tree in which initial common sequences of subword units are shared [2]. Since each arc of the tree is associated with a subword unit, the internal and terminal nodes of the tree represent partial or complete transcriptions of a vocabulary word respectively. The lexical tree includes 775874 arcs (versus 1628639 units of the linear lexicon) with a potential search space of more than 2 million states.

The generation of a synthetic instance of each word represented in the lexical tree is performed by visiting the tree according to a depth-first strategy, starting from the root node and traversing the current leftmost subtree first. When an arc located at a given layer of the tree is traversed, a template corresponding to the unit associated with that arc is randomly selected from the archive, and its frames are appended to the buffer that stores the partial utterance generated so far. The buffer locations of the beginning and ending frame of this segment of the utterance are also recorded. When the visit of the tree reaches a terminal node, we are ready to decode the sequence of frames corresponding to the "artificial" word associated to the node.

The decoding process is performed by means of a beam-search Viterbi search procedure using the same lexical tree to reduce the computational cost. Every arc in the tree is substituted with the Markov model of the associated subword unit. The procedure scores all word transcriptions corresponding to the terminal nodes that are not pruned after last buffer frame has been processed. The pronounced word w is then included into the confusion list of the words that occur among the top 10

Test vocabulary size	188892	188892	600	600	600
Utterances	artif.	artif.	artif.	real	real
NH	258	10	50	50	50
NW	-	5	10	10	15
Inclusion rate (%)	94.0	97.4	98.8	96.8	97.2
Candidate list average size	258	258	1315	2201	2750

Table 3: Recognition with vocabulary size 188892

hypotheses.
Since it is mandatory to minimize the search cost for obtaining the confusion lists, we exploit the first part of the trellis currently developed to decode the word that will be generated next. Due to the depth-first visit of the tree, the current "artificial" word and the next one will share their initial common sequence of frames. After the $t-th$ frame of the current artificial utterance has been processed, if t is the ending frame of a segment template, we record the list of the Markov states that survive to the beam-search pruning, and their scores. The decoding process for the next synthetic utterance initializes its list of active states from the state list corresponding to the last frame in common with the previous utterance. Thus, to minimize the decoding cost, we accept to generate words that share a set of templates: the obtained reduction of the computational cost is dramatic because it can be related to the compression factor of the lexical tree that is of the order of 1000 for the first layer. To evaluate the "goodness" and the cardinality of the set of acoustically similar words we generated 12000 synthetic utterances of a *single* word. Performing the acoustic decoding against the VLV, the correct word is included 96.1% of the time in the list of the 10 top candidates. It is also worth noting that, even after 12000 trials, the cardinality of the set of acoustically similar words has not yet reached an upper bound. Thus, the artificial utterances we generate present high acoustic variability, but they are acoustically similar to the natural ones because the recognized candidates cluster around the correct word.
For another set of experiments, we generated up to 15 artificial utterances per each word in the VLV, using the templates in the *training* database. From these utterances we constructed the word confusion groups using the 10 top hypotheses produced by decoding each utterance against the whole VLV. The results of these experiments are summarized in Table 3, in which row

	Inclusion rate (%)				
N. of top candidates	1	2	5	10	50
DDHMM	35.6	40.7	64.0	75.3	87.8
CDHMM	49.8	66.5	79.8	84.8	91.3

Table 4: One-step vs. two-steps performance

NH gives the number of the word candidates, and related word confusion groups, that are used to complete the final list of candidate to be rescored during the second step of the strategy. Row NW of Table 3, instead, shows the number of utterances per word that has been used to build the confusion lists. Testing against the VLV 188892 *artificial* utterances generated using the templates of the *test* database, an inclusion rate of 97.4% has been obtained with an average size of the candidate list of 258. As shown in the first column of Table 3, using the first step of the strategy only we achieve, instead, 94.0% of inclusion rate for the same number of word candidates.
Since the 600 words included in the test vocabulary are acoustically similar, they lead to more confusion than the words in the whole VLV. To obtain similar performance with respect to the VLV, it is necessary to use a larger the number of word confusion groups. By expanding a maximum of 50 candidates hypothesized in the first step of the strategy, the correct word is included in the list of words to be rescored 98.8% of the time, but notice that the average size of the candidate list now is 1315.
In the same experimental conditions, an inclusion rate of 97.2% is achieved for the real test database, but with an average size of the candidate list of 2750.
Finally, in the second step we rescored the resulting set with Continuous Density subword models using 16 Gaussian mixtures per state, the obtained results are given in the second row of Table 4, where they are compared with those obtained using Discrete Density models and the first step only.

References

[1] L.R Bahl, P.V. de Souza, P.S. Gopalakrishnan, D.S Kanewsky, D. Nahamoo, "Constructing Candidate Word Lists using Acoustically Similar Word Groups", IEEE Transactions on Signal Processing, Vol. 40, n. 11, pp. 2814–2816, 1992.

[2] L. Fissore, P. Laface, G. Micca, and R. Pieraccini, "Lexical Access to Very Large Vocabularies", IEEE Transactions on Acoustic, Speech and Signal Processing, Vol. 37,n. 8, pp. 1197–1213, 1989.