Speech Recognition and Understanding

Recent Advances, Trends and Applications

Edited by

Pietro Laface

Dipartimento di Automatica e Informatica
Politecnico di Torino
Corso Duca degli Abruzzi 24, 10129 Torino, Italy

Renato De Mori

School of Computer Science
3480 University St., Montreal, Quebec H3A 2A7, Canada

Springer-Verlag
Berlin Heidelberg New York London Paris Tokyo
Hong Kong Barcelona Budapest
Published in cooperation with NATO Scientific Affairs Division

ıble Speech Recognition
1989)

ystem: Preliminary De-
New Jersey : IEEE, 1990
.990)

Experimenting Text Creation by Natural-Language, Large-Vocabulary Speech Recognition

P. Alto, M. Brandetti, M. Ferretti, G. Maltese, F. Mancini, A. Mazza, S. Scarci, G. Vitillaro

IBM Italy Rome Scientific Center
via Giorgione 159, 00147 ROME (Italy)

1. Introduction

In the last years the probabilistic approach to speech recognition has allowed the development of high-performances large-vocabulary speech recognition systems [1] [2]. At the IBM Rome Scientific Center a speech-recognition prototype for the Italian language, based on this approach, has been built. The prototype is able to recognize in real time natural-language sentences built using a vocabulary containing up to 20000 words. [4]. Once and for all the user has to perform an acoustic training phase (about 20 minutes long), during which he is required to utter a predefined text. Words must be uttered inserting small pauses (a few centiseconds), between them. The prototype architecture is based on a personal computer equipped with special hardware. The first system we developed was aimed at a business and finance lexicon. Many laboratory tests have shown the effectiveness of the prototype as a tool to create texts by voice. After a first phase during which in-house experiments were carried on [5], the need arose to test the system in real work enviroments and for different applications. Two applications were considered: the dictation of radiological reports and of insurance company documents. Due to their characteristics, these applications seemed to be very well suited for our purposes. Since the vocabulary of the recognizer must be predefined, we had to adapt the system to the lexicon required by the new applications. The paper describes the techniques developed to efficiently adapt the basic component of the recognizer: the acoustic and language models. The results obtained experimenting automatic text dictation during real work are also presented.

2. System structure

To understand the steps necessary to perform the system adaptation a brief outline of its basic structure is needed. In the probabilistic approach to speech recognition we look for the sequence of words $\overline{W}$ which has the highest probability given the acoustic information $\overline{A}$ extracted from the observed signal [1]. In our case the acoustic signal is a sequence of acoustic labels extracted from the signal every centisecond and representing the energy content of the signal in 20 frequency bands.
Applying the Bayes theorem we can write:

$$P(\overline{W}\,|\,\overline{A}) = \frac{P(\overline{A}\,|\,\overline{W})P(\overline{W})}{P(\overline{A})} \qquad (1)$$

where $P(\overline{A}\,|\,\overline{W})$ is the probability that the sequence of words $\overline{W}$ will produce the sequence of acoustic information $\overline{A}$. $P(\overline{W})$ is the *a priori* probability of the sequence of words $\overline{W}$. $P(\overline{A})$ is the probability of the sequence of acoustic information $\overline{A}$. We seek the maximum of the above expression with respect to $\overline{W}$. We can ignore $P(\overline{A})$ because it does not depend on $\overline{W}$. Therefore we need to maximize the numerator of the expression (1).
As a consequence of equation (1), the recognition task can be viewed as composed by four components:

1. an acoustic processor to extract the acoustic information $\overline{A}$ from the speech signal;

NATO ASI Series, Vol. F 75
Speech Recognition and Understanding. Recent Advances
Edited by P. Laface and R. De Mori

2. an **acoustic model** to compute the probability $P(\overline{A} \mid \overline{W})$;
3. a **language model** to compute the a-priori probability $P(\overline{W})$;
4. an efficient **search strategy**, to find, among all possible sequences of words $\overline{W}$, the most probable one.

While the signal processing stage and the search strategy can be considered as vocabulary-independent, both the acoustic and the language model depend on the application which the system is aimed for, so they must be adapted to the lexicon of the new application. In the next paragraphs we will present the structure of the models and the techniques employed to adapt them.

Acoustic Model

The acoustic model task is to compute $P(\overline{A} \mid \overline{W})$. The approach currently employed for the acoustic model is based on **hidden Markov models** that are finite state automata [3]. For every time slice the model takes a transition from the current state to one of the allowed states (the transition can also produce no state change). For each transition an acoustic label is produced [1]. Both the transitions and the label emission occur according two probability distributions which depend on the current state only. These models are called *hidden* because it is only possible to observe the sequence of acoustic symbols produced, while the sequence of states is unknown.

Each word belonging to the vocabulary is represented by a different model. To allow speech recognition for a very large vocabulary, the basic speech units modeled by the Markov sources are associated to the basic sounds of the language. So an alphabet of acoustic units must be defined to represent the basic sounds of the language. Examples of acoustic units used for speech recognition are: syllables, diphones, phones. We choose the phone as phonetic unit. The basic sounds of the Italian language were described by a set of 56 phonetic units [4]. For each phonetic unit a Markov model representing its pronunciation has been created. In our system all the phonetic units have the same topological structure. The distinction between different sounds is left entirely to the probability distributions, namely, to the *parameters* of the model. The computation of the parameters is accomplished during the acoustic training phase employing the predefined text uttered by the user. The model for each word is built by the concatenation of the Markov sources corresponding to the string of phonetic units forming its pronunciation.

Language Model

The task of the language model is to compute $P(\overline{W})$. The computation is performed in the following way:

$$P(\overline{W}) = \prod_{i=1}^{N} P(w_i \mid w_{i-1}, w_{i-2}) \tag{4}$$

This is the so-called **trigram language model**[2] : it contains the approximation of considering equivalent all the sentences ending with the same couple of words $w_{i-1}w_{i-2}$. The number of possible trigrams is so large that is practically impossible to collect the amount of data needed to estimate the probability of each of them. To overcome the problem, an interpolation between different probability distributions is performed. Three different distributions are computed for trigrams, bigrams and unigrams. The probability of word w_3 given the words w_1 and w_2 is estimated as follows:

$$P(w_3 \mid w_1 w_2) =$$
$$\lambda_3 \frac{C(w_1 w_2 w_3)}{C(w_1 w_2)} + \lambda_2 \frac{C(w_2 w_3)}{C(w_2)} + \lambda_1 \frac{C(w_3)}{N} + \lambda_0 \frac{1}{V}$$

where $C(w_1, \ldots, w_n)$ is the number of times the word string $w_1, \ldots, w_n$ was observed in the training data and V is the number of words in the vocabulary. The λ coefficients are estimated using the *expectation-*

maximization algorithm
straints for speech reco
larger is the training co
Perplexity is a typical
number of words that
used, the perplexity val

Vocabulary definit

In our system the voca
the system performanc
In the first version of
at the dictation of eco
ones in a **corpus** cont
economy and finance
Mondo), and press age

3. System adapta

To build a speech reco

- to define the vocab
- to adapt the acous

While the adaptation o
on factors like the lex
it will be described the

Acoustic model ac

We have just seen that
forming its pronunciat
application, is to perfo
number of words that
netic transcriptions us
for which a new phone
Usually, the phonetic
vocabularies the transc
automatic as possible.
transcription are based
systems cannot provid
complexity of the prob
employed a different
described by a limited
description). Given th
phonetic transcriptions
the grapheme-to-phon
of his lexical knowled
uncertainty. A set of
right part. The left p
contexts; the right par

maximization algorithm [9]. The trigram language model is an effective tool to represent the linguistic constraints for speech recognition purposes; on the other hand it requires a large amount of training data. The larger is the training corpus, the higher is the recognition rate [10].
Perplexity is a typical measure of the predictive power of a language model [11]. It estimates the average number of words that are considered equiprobable by the model. Of course, if the language model is not used, the perplexity value equals the vocabulary size.

Vocabulary definition

In our system the vocabulary is predefined; this means that the choice of the vocabulary is a crucial step for the system performances.
In the first version of our prototype a vocabulary containing more than 20000 words was used. It is aimed at the dictation of economy and finance reports. [5]. The 20000 words were chosen as the most frequent ones in a corpus containing 44 millions of words composed by articles from the most important Italian economy and finance newspaper (*Il Sole 24 Ore*), articles from an economy and finance newsmagazine (*Il Mondo*), and press agency news. The coverage of this vocabulary computed on a disjoint corpus was 96.5%.

3. System adaptation

To build a speech recognition system for a new application, we need:

- to define the vocabulary;
- to adapt the acoustic and language models.

While the adaptation of acoustic and language models can be automatized, the choice of vocabulary depends on factors like the lexicon of the application and the size of the available text. Thus in the next paragraphs, it will be described the tools that have been created to make the required adaptations easy and quick.

Acoustic model adaptation

We have just seen that each word is described by the concatenation of the Markov models of phonetic units forming its pronunciation. According to the chosen technique, the first step in adapting the system to new application, is to perform the phonetic transcription of all the words in the new vocabulary. To limit the number of words that must be transcribed, a large database was built containing all the words and the phonetic transcriptions used in previous vocabularies. By using the database it is possible to find all the words for which a new phonetic transcription must be supplied.
Usually, the phonetic transcription is performed manually. It is a very expensive process and for large vocabularies the transcriptions could contain errors. We tried to make the phonetic transcription process as automatic as possible. The systems that have been proposed to solve the problem of automatic phonetic transcription are based on rules [6] [7] or on automatic learning from training data [8]. Actually, these systems cannot provide the accuracy required for automatic speech recognition. This is due both to the complexity of the problem and to the difficulty to describe all the possibilities with a limited set of rules. We employed a different technique from the mentioned ones. We separated phonotactical knowledge (well described by a limited set of rules) from lexical knowledge (based on experience and not suitable for a formal description). Given the string representing the orthographic form of the word our system produces a set of phonetic transcriptions for that word, which are the ones that can be obtained applying our set of rules for the grapheme-to-phoneme translation. The user can choose manually the correct transcription on the basis of his lexical knowledge. Grapheme-to-phoneme translation for the Italian language has a relatively low uncertainty. A set of 78 rules allows to describe all the ambiguities. Each rule consists of a left part and a right part. The left part consists of a grapheme string and its (possibly empty) left and right graphemic contexts; the right part consists of the set of possible phonetic transcriptions for the grapheme string. The

set of transcriptions produced applying this set of rules is then pruned by means of a set of global rules (which, for example, reject all the transcriptions which do not have one and just one stressed vowel). The right phonetic transcription always belongs to the resulting set. The average number of phonetic transcriptions per word is 5. Using this method it was possible to adapt rapidly the recognizer to the new application. The quality of the produced transcriptions was at least equal to a completely manual phonetic transcription.

Language model adaptation

The language model is built following the previously described technique. We have created a software structure that takes a dictionary and a *corpus* as a starting point and give the probability estimations as output:

$$P(w_3 | w_1 w_2)$$

for any word w_3, given the context w_1, w_2.

4. In-house test for the new application

Before experimenting the speech-recognizer in a real environment we needed to perform in-house tests to assess the recognition rate of the system when used to dictate pre-defined texts. To perform the experiment a text containing phrases peculiar to the application must be created; it must be dictated by several different speakers. To make a meaningful test it is important that the text contains all the phonetic units used to build the acoustic model in phonetic contexts typical of the application.
Usually, the text is built manually trying to represent a large number of different contexts using the smallest number of sentences. To avoid this manual process a procedure has been built to prepare the test automatically, so it is possible to select efficiently a set of sentences containing phonetic contexts typical of the application and suitable to assess the accuracy of the recognizer.

This preliminary study on the usage of a voice-activated text editor indicated that large-vocabulary speech recognition can offer a very competitive alternative to traditional text entry. It is likely to be well accepted even by users who have a large experience in keyboard text entry. The results (error rate 2-3%) show that the system could be a revolutionary tool, for example, for the office of the future, but poses unprecedent human factors problems. These experiments have been very useful for us, because they have suggested, for example, improvements to the system interface and the need of some special keys to make easy the revision and the correction phases.

5. Real-work applications

Two distinct application areas were considered: medical report dictation and office (insurance company) documents and memos creation. We will describe separately how the vocabulary of each application was built.
The economy and finance lexicon is quite different from the lexicon required to dictate radiological reports, while it has some similarities with respect to the lexicon used in the insurance company reports.
In the first case the vocabulary was selected by using a corpus containing only radiological reports; in the second case the vocabulary was built taking the economy-finance vocabulary as a starting point.

Radiological Reports Vocabulary

The available corpus contained about 5 million words collected in four different hospitals. The hospital where the experiment was performed provided us with a corpus containing 50000 words
The first problem was the choice of the vocabulary size. We adopted the criterion of analyzing the variation of the coverage with respect to the vocabulary size (figure 1).

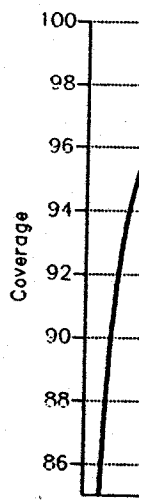

Figure 1. Coverage

A vocabulary con
coverage and the n
One of the main c
the report dictation
imenter all the 32
performed, were in
the most frequent
vocabulary had a
reports.

Insurance Com

The selection of the
million words prov
the economy and fi
insurance corpus w
selected and the 31
showed a 99% cove

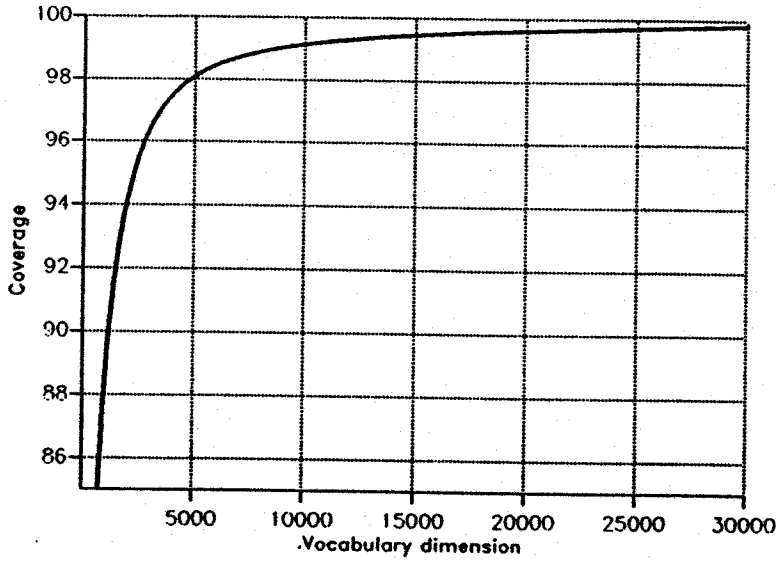

Figure 1. Coverage of corpus as function of vocabulary size

A vocabulary containing 5000 words seemed to us a reasonable trade-off between the need to reach high coverage and the need to have enough data to estimate the language model parameters.
One of the main characteristics found in this kind of lexicon was the presence of a set of words peculiar to the report dictation process at each location. To make the recognizer well suited to the needs of the experimenter all the 3200 different words found in corpus provided by the hospital where the experiment was performed, were included in the vocabulary. The vocabulary was completed adding 1900 words which were the most frequent ones in the other corpora not included in the previous list of 3200 words. The resulting vocabulary had a 100% coverage on reports from the experiment location and 97.5% coverage on other reports.

Insurance Company Reports Vocabulary

The selection of the words to be used in this vocabulary was performed by analyzing a corpus containing 1.5 million words provided by the insurance company. There was a certain similarity between this lexicon and the economy and finance one: the coverage obtained by economy and finance vocabulary (20000 words) on insurance corpus was 95.2%. The 15000 most frequent words of economy and finance vocabulary were selected and the 3100 most frequent words found in insurance corpus were added. The resulting vocabulary showed a 99% coverage on insurance texts and a 95.7% coverage on economy and finance ones.

Language models informations

Table 1. Language models characteristics		
Parameter	**Radiology Reports**	**Insurance Reports**
Vocabulary size	5100	18100
Training data	4.8	1.5 + 40
Millions of distinct trigrams	0.62	3.4
Millions of distinct bigrams	0.19	2.3
Perplexity	38	18

The radiology reports language model was trained using 4.8 million words. For the insurance company reports a corpus containing 1.5 million words typical of the application was merged with 40 million words extracted from the economy and finance corpus. In the table the most significant parameters of the language models are reported.

6. Results

In the following paragraph the results obtained experimenting the two prototypes during real work are reported.

Radiological Reports Dictation

Four doctors have used the recognizer during their every-day work to prepare the reports to be delivered to the patients. No one had any difficulty in inserting short pauses between words. The doctors dictated 150 reports containing more than 12000 words. The vocabulary coverage for the dictated reports was 98.6%. Table 2 reports the results of the experiment.

Table 2. Radiological reports dictation			
Speaker	Number of reports	Error rate	Speaker's errors
sp1	25	1.7%	1.4%
sp2	14	2.0%	1.2%
sp3	76	3.5%	0.9%
sp4	35	5.0%	3.2%

The error rate is referred to the number of errors done by the recognizer without taking into account errors due to words not included in the vocabulary. The speaker's error rate is referred to the errors due to misuse of the recognizer (wrong commands, absence of pause between words, etc.). The global error rate can be computed summing the numbers in the third and fourth column of the table. We can see that the recognizer's performances range from 90.5% to 95.6% of accuracy.

Insurance Co

The experimentat
lary coverage on t

Table 3. Insuran	
Speaker	
sp1	
sp2	
sp3	
sp4	
sp5	

References

[1] L.R. Bahl,
tion, *IEEE*
179-190.

[2] F. Jelinck,
73, no. 11,

[3] L.R. Bahl,
in the Tan
and Signal

[4] P. D'Orta,
Recognitio
No. 2, Ma

[5] P. Alto, M
with a 200
2-81.

[6] R. Carlson
tional Conf

[7] D. H. Kla
Trans. on A

[8] T. J. Sejno
Systems, 1

[9] F. Jelinck,
"Pattern Re
pp. 381-38

[10] M. Ferretti
listic Speec

[11] F. Jelinek,
nition Task

Insurance Company Reports Dictation

The experimentation was carried on by five different users who dictated more than 8000 words. The vocabulary coverage on the dictated text was about 99%. Table 2 reports the results obtained in this case.

Table 3. Insurance company reports dictation

Speaker	Error rate	Speaker's error rate
sp1	1.9%	1.0%
sp2	1.4%	0.5%
sp3	14.0%	2.0%
sp4	7.4%	1.5%
sp5	2.4%	0.8%

References

[1] L.R. Bahl, F. Jelinek, R.L. Mercer, **A Maximum Likelihood Approach to Continuous Speech Recognition,** *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. PAMI-5, no. 2, 1983, pp. 179-190.

[2] F. Jelinek, **The development of an experimental discrete dictation recognizer,** *Proceedings IEEE,* vol. 73, no. 11, November 1985, pp. 1616-1624.

[3] L.R. Bahl, P.F. Brown, P.V. De Souza, R.L. Mercer, M.A. Picheny, **Acoustic Markov Models Used in the Tangora Speech Recognition System,** *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing,*

[4] P. D'Orta, M. Ferretti, A. Martelli, S. Melecrinis, S. Scarci, G. Volpi, **Large-Vocabulary Speech Recognition: a System for the Italian Language,** *IBM Journal of Research and Development,* Vol. 32, No. 2, March 1988, pp.217-226.

[5] P. Alto, M. Brandetti, M. Ferretti, G. Maltese, S. Scarci, **Experimenting Natural-Language Dictation with a 20000-Word Speech Recognizer,** *IEEE CompEuro 89,* Hamburg, May 8-12, 1989, pp. 2-78 - 2-81.

[6] R. Carlson, B. Granstroem, **A Text-to-Speech System Based Entirely on Rules,** *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing,* Philadelphia, PA, April 1976.

[7] D. H. Klatt, **Structure of a Phonological Rule Component for Synthesis-by-Rule Program** *IEEE Trans. on Acoust., Speech and Sig. Proc.,* vol. ASSP-24, no. 5, 1976, pp. 391-398.

[8] T. J. Sejnowski, C. R. Rosenberg, **Parallel Networks that Learn to Pronounce English Text,** *Complex Systems,* 1 (1987), pp. 145-168.

[9] F. Jelinek, R.L. Mercer, **Interpolated Estimation of Markov Source Parameters from Sparse Data,** *in "Pattern Recognition in Practice", E.L.Gelsema and L.N. Kanal, Ed.,* North-Holland, New York, 1980, pp. 381-387.

[10] M. Ferretti, G. Maltese, S. Scarci, **Measures of Language Model and Acoustic Information in Probabilistic Speech Recognition,** *Eurospeech 89,* Paris, September 1989, pp. 473-476.

[11] F. Jelinek, R.L. Mercer, L.R. Bahl, J.K. Baker, **Perplexity - a Measure of Difficulty of Speech Recognition Tasks,** *94th Meeting Acoustical Society of America,* Miami Beach, December 1977.