AN AUTOMATIC SPEECH RECOGNITION SYSTEM FOR THE ITALIAN LANGUAGE

Paolo D'Orta, Marco Ferretti, Alessandro Martelli, Stefano Scarci
IBM Rome Scientific Center
via Giorgione 159, ROME (Italy)

ABSTRACT

An automatic speech recognition system for Italian language has been developed at IBM Italy Scientific Center in Rome. It is able to recognize in real time natural language sentences, composed with words from a dictionary of 6500 items, dictated by a speaker with short pauses among them. The system is speaker dependent, before using it the speaker has to perform the training stage reading a predefined text 15-20 minutes long. It runs on an architecture composed by an IBM 3090 mainframe and a PC/AT based workstation with signal processing equipments.

PROBABILISTIC APPROACH

The problem of recognizing human voice is approached in a probabilistic manner. Let $\overline{W} = w_1, w_2, \ldots, w_n$ be a sequence of n words, and let $\overline{A}$ be the acoustic information extracted from the speech signal, from which the system will try to identify the pronounced words. $P(\overline{W} \mid \overline{A})$ indicates the probability that the the sequence of words $\overline{W}$ has been spoken, once we observe the acoustic string $\overline{A}$ produced at the end of the signal processing stage. The most probable sequence of word, given $\overline{A}$, is that maximizing $P(\overline{W} \mid \overline{A})$. Through Bayes' formula:

$$P(\hat{\overline{W}} \mid \overline{A}) = \max_{\overline{W}} P(\overline{W} \mid \overline{A}) = \max_{\overline{W}} \frac{P(\overline{A} \mid \overline{W}) P(\overline{W})}{P(\overline{A})}$$

$P(\overline{A} \mid \overline{W})$ denotes the probability that the sequence of words $\overline{W}$ will produce the acoustic string $\overline{A}$, $P(\overline{W})$ is the a priori probability of word string $\overline{W}$, $P(\overline{A})$ is the probability of acoustic string $\overline{A}$. To find the word sequence which maximizes the third term in the preceding equation, it is sufficient to find the sequence which maximizes the numerator; $P(\overline{A})$ is, in fact, clearly not dependent on any $\overline{W}$. Then, the recognition task can be decomposed in these problems:

1. perform an **acoustic processing** able to extract from the speech signal an information $\overline{A}$ representative of its acoustic features, and, at the same time, adequate for a statistical analysis;

2. create an **acoustic model** which makes it possible to evaluate $P(\overline{A} \mid \overline{W})$, that is the probability that the acoustic string $\overline{A}$ will be produced when the speaker pronounces the word string $\overline{W}$;

3. create a **language model** giving the probability $P(\overline{W})$ that the speaker will wish to pronounce $\overline{W}$;

4. find, among all possible sequences of words, the most probable one. Even with small vocabularies it is not feasible to conduct an exhaustive search; so, we need to identify an efficient **search strategy**.

ACOUSTIC PROCESSING

Acoustic processing is performed in the *acoustic front-end* formed by an *acquisition stage* (microphone, filter, amplifier, A/D converter) and a *processing stage*. The analog to digital converter gives a numeric representation of the signal picked up by the microphone, constituted by 20000 samples/sec., each of 12 bits. Every 10 milliseconds an *acoustic vector* of 20 parameters is computed describing, through its spectral features, the behavior of speech signal for that interval. This operation takes into account recent studies on physiology of the human ear and on psychology of sounds perception. The signal energy in several frequency bands is determined through a Fourier analysis [6]. Width of bands is not uniform; it grows with frequency. This is in accordance with the behavior of the cochlea that has a better resolution power at low frequencies. Furthermore, computation of parameters considers other features of auditory system, as dynamic adaptation to signal level.

Each acoustic vector is then compared with a set of 200 *prototype vectors* and the closest prototype is chosen to represent it; the *label* of this prototype (a number from 1 to 200), will then be substituted to the original vector. Therefore, the acoustic information $\overline{A}$ is formed by a sequence of labels $a_1, a_2, \ldots$, with a considerable reduction in the amount of data needed to represent the speech signal.

ACOUSTIC MODEL

The acoustic model must compute the probability $P(\overline{A} \mid \overline{W})$ that the pronunciation of word string $\overline{W}$ will produce the label string $\overline{A}$. To design the acoustic model it is essential to understand the relationship between words and *sounds* of a language. With sounds of a language we mean those particular sounds usually generated during speaking. Phonetics is helpful in this task. Experts in linguistics usually classify sounds in classes, called *phonemes* [2]. The same phoneme can be representative of many different sounds, but they are completely equivalent from a linguistic point of view. The Italian language is usually described with 31 phonemes; in our system we use an extended set composed of 56 phonetic elements, to take into account particular aspects of the process of pronunciation not considered by the usual classification: coarticulation, different behavior in stressed and non-stressed vowels, pronunciation of vowels and fricatives by people from different regions. Each word in the language can be phonetically described by a sequence of phonemes,

representing the sequence of basic sounds that compose it. So, it is very useful to build up the acoustic model starting from phonemes.

For each phoneme, a Markov source [5] is defined, which is a model representing the phenomenon of producing acoustic labels during pronunciation of the phoneme itself. Markov sources can be represented by a set of states and a set of transitions among them. Every 10 milliseconds a transition takes place and an acoustic label is generated by the source. Transitions and labels are not predetermined, but are chosen randomly on the basis of a probability distribution. Starting from phoneme models, we can build models for words, or for word strings, simply by concatenating the Markov sources of the corresponding phonemes. Figure 1 shows a typical structure for Markov model of a phonetic unit and figure 2 the structure of the Markov model for a word.

The structure of Markov models is completely defined by the number of states and by interconnections among them. It is unique for all the phonemes and for all the speakers and has been determined on the basis of intuitive considerations and experimental results, because no algorithm is known to find the best structure to describe such a phenomenon. The different behavior in different phonemes and in the voice of different speakers is taken into account in the evaluation of the model parameters: probability of transition between pair of states and probability of emission of labels. This evaluation, executed in the training stage, is performed, given the word sequence $\overline{W}$ of training text and collected the acoustic label string $\overline{A}$ from the front-end, accordingly to the *maximum likelihood* criterion [1], maximizing the probability $P(\overline{A} \mid \overline{W})$. A speaker, during training, does not have to pronounce all the words in the dictionary; on the other hand, it is necessary that the text to be read contains all the phonemes of the language, each of them well represented in a great variety of phonetic contexts.

In the recognition stage the term $P(\overline{A} \mid \overline{W})$ is computed on the basis of statistical parameters determined during the training; then it is necessary to evaluate the probability that the Markov source for the word string $\overline{W}$ will emit the label string $\overline{A}$, going from its initial state to its final one. This must be done summing the probability of all the paths of this kind, but it could be computationally very heavy and impractical to count them all because their number depends exponentially on the length of $\overline{A}$. Using dynamic programming techniques, it is possible to reach this goal limiting the amount of calculation to be done. The *forward pass* algorithm [5], is, in fact, computationally linearly dependent on the length of $\overline{A}$.

LANGUAGE MODEL

The language model is used to evaluate the probability $P(\overline{W})$ of the word sequence $\overline{W}$. Let $\overline{W} = w_1, w_2, \dots, w_n$; $P(\overline{W})$ can be computed as:

$$P(\overline{W}) = \prod_{k=1}^{n} P(w_k \mid w_{k-1}, \dots, w_1)$$

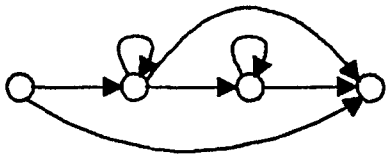

Figure 1. Typical structure for Markov model of a phonetic unit.

So, the task of the language model is to calculate $P(w_k \mid w_{k-1}, \dots, w_1)$, that is, given the beginning of a sentence $w_1, w_2, \dots, w_{k-1}$, to evaluate the probability of words in the vocabulary to be at place k in the sentence, or, in other terms, to estimate the probability of the word to appear in that context.

If we ignore the language model (that means considering words as equiprobable), it would be impossible to distinguish omophones, (acoustically equivalent words), and it would be very hard to recognize correctly very similar words on the basis of the acoustic information only. The estimation of probabilities could be based on grammatical and semantic information, but a practical and easy way to use this approach has not been found yet. For this reason, in our approach the language model is built up from the analysis of statistical data. They have been collected from a huge set (*corpus*) of Italian sentences (in all, about 6 millions of words). Even using a small dictionary, no corpus can contain all the possible contexts $w_{i-1}, w_{i-2}, \dots, w_1$. The evaluation of the term

$$P(\overline{W}) = \prod_{k=1}^{n} P(w_k \mid w_{k-1}, \dots, w_1)$$

is then based on the intuitive consideration that recently spoken words in a sentence have more influence than old ones on the continuation of the sentence itself. In particular, we consider the probability of a word in a context depending only on the two preceding words in the sentence:

$$P(w_k \mid w_{k-1}, w_{k-2}, \dots, w_1) = P(w_k \mid w_{k-1}, w_{k-2})$$

Such a model is called trigram language model. It is based on a very simple idea and, for this reason, its statistics can be built very easily only counting all the sequences of three consecutive words present in the corpus. On the other hand, its predictive power is very high. If the information given by the language model were not available, in every context there would be uncertainty about the next word among all the 6500 words in the dictionary. Using the trigram model, uncertainty is, on the average, reduced to the

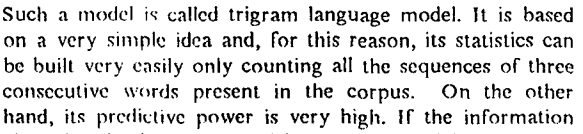

Figure 2. Typical structure for Markov model of a word

choice of a word among 100-110. In the procedure of estimating the language model statistics, a problem comes out: the probability of trigrams never observed in the corpus must be evaluated. For a 6500-word dictionary the number of different trigrams is about 270 billions; but from a corpus of 6 millions of words, only 6 millions of trigrams can be extracted, and not all of them are different. It is clearly evident that, even with the availability of a bigger corpus, it is not possible to estimate probabilities of trigrams by their relative frequencies. Trigrams never seen in the corpus must be considered allowable, although not very probable, otherwise it could be impossible to recognize a sentence containing one of them. To overcome this problem, some techniques have been developed, giving a good estimate of probability of never observed events [3].

Sentences in the corpus are taken from economy and finance magazines, and, as a consequence, the model is capable to work well on phrases about this topic, worse on other subjects. Clearly, the availability of corpus on different topics could be very useful in order to use the language model in different contexts. Nevertheless, some studies demonstrate that language model could be still fruitfully used for a matter different to the main one, if the collected data are enriched with a small corpus (about 1-2% the dimension of the main one) related to the new subject. This technique is used to allow the recognition of sentences not on finance and economy.

Figure 3 shows the coverage of the corpus on texts of economy and finance as a function of the vocabulary size.

SEARCH STRATEGY

To find the word sequence $\overline{W}$ which maximizes the term $P(\overline{W} \mid \overline{A})$, it is not feasible to consider all the sequences that can be built with words in the dictionary. For this reason an efficient *search strategy* is used that limits the investigation to a small fraction of the allowed word strings. The sequences generable with the N words in the dictionary can be represented by a *tree*. N branches, corresponding to the first word in the sentence, go out from the root, one for each word in the dictionary. Each branch ends in a new node, from which other N branches are generated for the second word in the sentence, and so on. A node in the tree defines univocally a sequence of words, constituted by words corresponding to branches in the path from the root to the node itself. During the recognition process, tree nodes are explored, and, for each of them, the probability (acoustical and linguistical) that the sentence will start with the corresponding words is computed. Nodes with a low probability are discarded; among the remaining nodes, the path that seems, so far, the more probable is extended. This choice can be modified during the process, selecting at any time the best current path. This strategy, usually called *stack sequential decoding*, leads, in general, to the requested solution: the most probable sentence [4].

The extension of a path from a node is done analyzing all the branches going out from it, that means all the words in the vocabulary. It is computationally not practical to determine the acoustic likelihood of each word through the forward pass algorithm. The problem of a fast access to a great dictionary is one of the most important topics in

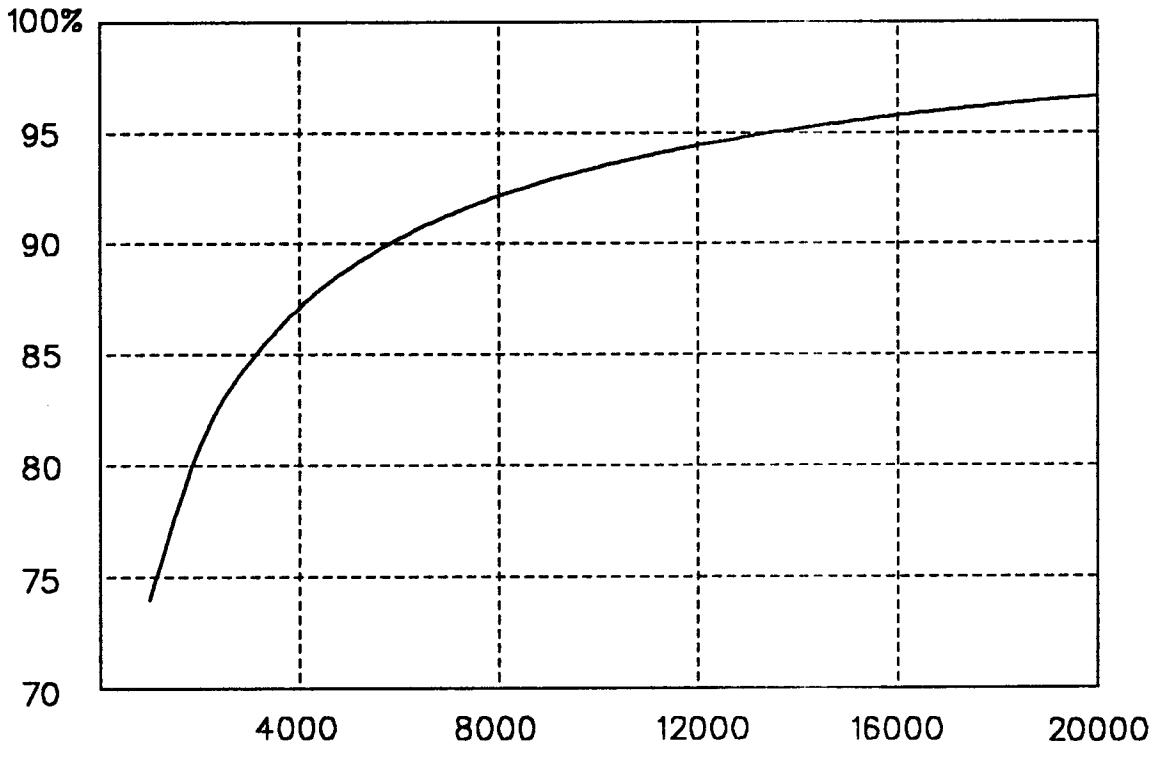

Figure 3. Coverage of the corpus as a function of vocabulary size.

speech recognition. Studies are conducted to find good strategies. In our system, first a rough match is rapidly conducted on the whole dictionary to select a subset of words. Then, a more precise search is performed on this subset with forward pass. It has been seen that this procedure assures most of the times the identification of the most acoustically likely word.

The stack decoding algorithm conducts a left to right search from the beginning to the end of the sentence, examining labels in the order they are produced by the acoustic front-end and it does not require in advance the knowledge of the whole label string. Therefore, it is well suited to operate in real time.

The search in the tree of all the possible solutions, along with the computation of acoustical and linguistical probabilities, is performed in the IBM 3090 mainframe. This task is, in fact, computationally so heavy that only this powerful system can avoid the use of specialized processors.

dictionary size	average	best	worst
1000	92.2	95.1	89.5
3000	86.1	89.6	83.3
6500	82.0	86.4	78.0

Table 1. Recognition accuracy without language model.

dictionary size	average	best	worst
1000	97.9	98.5	96.4
3000	97.1	97.9	95.9
6500	96.3	97.4	94.9

Table 2. Recognition accuracy with language model

RESULTS

Several experiments were conducted on the recognition system with ten different speakers who had previously trained the system. Each speaker dictated a text composed by natural language sentences about finance and economy. Recognition accuracy is always over 94%, and, on the average is 96%. It has been seen that the language model is capable to avoid about 10% of the errors made using only the acoustic model. This shows the importance of using of linguistic information.

Table 1 shows the recognition accuracy obtained considering all the words equiprobable for three dictionaries of different size, table 2 shows the results obtained for the same test with the language model.

REFERENCES

[1] Bahl L.R., Jelinek F., Mercer R.L. **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] Flanagan, J.L., **Speech Analysis, Synthesis and Perception,** *Springer, New York,* 1972.

[3] Nadas A. **Estimation of Probabilities in the Language Model of the IBM Speech Recognition System,** *IEEE Trans. on Acoustics, Speech, and Signal Processing,* no. 4, ASSP-32 (1984), pp. 859-861.

[4] Nilsson N.J. **Problem-Solving Methods in Artificial Intelligence** *McGraw-Hill,* New York, 1971, pp. 43-79.

[5] Rabiner L.R., Juang B.H. **An Introduction to Hidden Markov Models,** *IEEE ASSP Magazine,* no. 1, vol. 3, January 1986, pp. 4-16.

[6] Rabiner, L.R., R.W. Schafer, **Digital Processing of Speech Signals,** *Prentice Hall,* Englewood Cliffs, 1978.