Lecture Notes in Computer Science

Edited by G. Goos and J. Hartmanis

399

V. Cantoni R. Creutzburg
S. Levialdi G. Wolf (Eds.)

Recent Issues in Pattern Analysis and Recognition

Springer-Verlag
New York Berlin Heidelberg London Paris Tokyo Hong Kong

A 20000-Word Speech Recognizer of Italian

M. Brandetti, M. Ferretti, A. Fusi, G. Maltese, S. Scarci, G. Vitillaro

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

Abstract

A real-time speech recognition system of Italian has been developed at IBM Rome Scientific Center. It handles natural language sentences from a 20000-word dictionary, dictated with words separated by short pauses. The architecture consists of a PC/AT equipped with signal processing hardware. The paper describes the system, shows results of decoding tests and includes descriptions of the topics in speech recognition being currently investigated.

1. Introduction

Existing speech recognition technologies have proven adequate for simple tasks, involving knowledge of a small vocabulary (tens or hundreds of words), suiting limited applications (typically recognition of a set of commands uttered in an isolated fashion by an operator whose hands are busy); they are usually independent of the target language.

Interesting applications in an office environment, such as text dictation and database query, on the other hand, must be capable of handling natural language and pronunciation. This requires large vocabularies (thousands of words), and necessitates substantially more sophisticated techniques, which take into account language-specific knowledge on phonology, syntax and (surface) semantics.

Rome Scientific Center has developed a real-time isolated-utterance speech recognition system for the Italian language, based on a 20000-word vocabulary. The recognizer architecture consists of a workstation based on a PC/AT equipped with signal processing hardware. Word-recognition accuracy for pre-recorded sentences ranges from 95% to 98%. The words must be uttered separated by short pauses.
The Speech Recognition Project started at IBM Rome Scientific Center from a cooperation with the IBM T.J. Watson Research Center, where advanced prototypes for the English language have been developed. The mathematical approach being applied to the Italian language is probabilistic, based on the maximum likelihood principle [1]. The role of human knowledge is limited to the design of a basic model of speech production and perception; statistics is used as a methodology for integration of the conceived model by "automatic learning" from data.

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 recognize which words were uttered. The aim is to find the particular sequence of words $\hat{W}$ which maximizes the conditional probability $P(\overline{W} \mid \overline{A})$, i.e. the most likely word sequence given the acoustic information. By Bayes' theorem,

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

$P(\overline{A} \mid \overline{W})$ is the probability that the sequence of words $\overline{W}$ will produce the acoustic string $\overline{A}$, that is, the probability that the speaker, pronouncing the words $\overline{W}$, will utter sounds described by $\overline{A}$. $P(\overline{W})$ is the a priori probability of the word string $\overline{W}$, that is, the probability that the speaker will wish to pronounce the

words $\overline{W}$. $P(\overline{A})$ is the probability of the acoustic string $\overline{A}$; it is not a function of $\overline{W}$, since it is fixed once $\overline{A}$ is measured, and can thus be ignored when looking for the maximum over $\overline{W}$.
A consequence of this equation is that the recognition task can be decomposed in the following problems:

1. perform **acoustic processing** to encode the speech signal into a string of values $\overline{A}$ representative of its acoustic features, and, at the same time, adequate for a statistical analysis;
2. compute the probability $P(\overline{A} \mid \overline{W})$ (for this purpose an **acoustic model** must be created);
3. evaluate $P(\overline{W})$ (for this a **language model** is needed);
4. look, among all possible sequences of words, for the most probable one, by means of an efficient **search strategy** (an exhaustive search is not feasible, even for small vocabularies).

A description of the system architecture is provided in the next section. In the following sections, acoustic and linguistic modeling of the Italian language are discussed and experimental recognition results are given; furthermore a description is given of topics in speech recognition being investigated, including fast speaker adaptation [2]; speech databases [3]; automatic phonetic transcription [4]; human factors of voice-activated text-editing [5].

2. System Architecture

Recognition and transcription of speech are performed by a workstation consisting of an IBM PC-AT equipped with four signal processing cards and the IBM ECD high resolution screen. [6][7] . Speech is collected by either a lip microphone (providing good noise immunity) or a table pressure zone microphone (more sensitive to background noise, but very comfortable for the speaker) [8]. The digitized acoustic signal (20K samples/sec, 12 bits/sample) is processed to extract, every 10 milliseconds, a vector of 20 parameters, which represent, essentially, the signal log energy in 20 frequency bands (spaced in accordance to the frequency sensitivity of the human ear), and transformed nonlinearly to take into account the adaptation capability to different sound levels. The vector-quantization replaces each vector with an *acoustic label* identifying the closest prototype vector belonging to a speaker-dependent pre-computed codebook of 200 elements.

The search strategy is based on the *stack sequential decoding* algorithm [9]. It controls the decoding process by hypothesizing the most likely sequence of words (by means of an efficient heuristic method), and requests the evaluation of linguistic and acoustic probabilities according to the hypothesized left context of the sentence. Stack decoding proceeds from left to right, and therefore is intrinsically well suited to a real-time system, which recognizes word sequences while they are being spoken.

The human interface of the speech recognizer consists of a text editor, which allows the use of both voice and keyboard for text input and editing. Commands for text insertion and deletion, word-searching, formatting (with a "what you see is what you get") interface are included. Documents can be filed, retrieved and printed. All editor commands can be given either by keyboard or by voice. A word (or any character string) not included in the vocabulary can be input by pronouncing a keyword (which sets the system to a single- character input mode and by spelling it).

3. Acoustic Modeling

The acoustic model is based on Markov models [10] of Italian phonemes as fundamental building blocks. It has been observed, both for English and Italian, that the same Markov structure can adequately be used for all the phonetic elements of the language, if it provides enough degrees of freedom. Differentiation among phonetic Markov sources is thus left entirely to the parameter estimation process [11]. Therefore, the essential problem is the design of the set of phonetic elements by which the language sounds are described. Phonemes, the classical units defined by the phonology of the language, are a good starting point, but don't adequately take into account the variability of the speech phenomena. On the other hand, a too detailed model, involving a large number of parameters, might require an unacceptably large statistical sample of the speaker's voice to be trained. The design of the phonetic alphabet should then look for the best trade-off between detail of modeling and brevity of training.

A systematic procedure to look for an optimal phonetic alphabet has not been developed yet. Our approach combines the results of traditional acoustic and phonetic research with analysis of statistical data. For this purpose, the speech signal is aligned to the Markov source by means of the Viterbi algorithm [12]. A measure of the quality of the phonetic representation may be provided by the mutual information between the phonetic alphabet and the set of speech alignments. After making experiments with various phonetic alphabets (see below) we adopted a set of 56 phonetic units [13], while Italian is usually described in terms of 30 distinct phonemes.

Recognition experiments are the most reliable way to evaluate the effectiveness of a modification to the phone alphabet, but are slow and computationally expensive. We experimented some faster measures, which proved very useful. The *Kullback divergence* (or *cross-entropy*) [14] can show whether utterances of two units have significant statistical differences. This measure is especially convenient when considering to split a set of sounds, previously described by a single phonetic unit, into two sets described by two different units (usually depending on the phonetic context).
Exact computation of divergence requires that the summation be extended to all possible sequences of acoustic labels $\overline{A}$. As this is infeasible, approximate techniques are needed. We experimented three techniques, described in [15].

A notable problem of Italian is the presence of inflections due to mispronunciations by speakers from some regions. A possible solution consists in describing mispronounced words with more than one word model; this requires that more than one source be matched to the incoming utterance during recognition. Our more efficient solution consists in introducing "ambiguous" phonetic units, which, after the parameter estimation performed by the training procedure, are flexible enough to model the inconsistencies of the speaker's pronunciation.
The system has indeed proven capable of handling speakers from different Italian regions with essentially identical performance.
We made experiments on word recognition accuracy when decoding is purely acoustic (i.e., the language model gives all words the same probability), for three phone sets, using the 6000-word vocabulary recognition system. The first one, **PH45**, consists of 45 phones, obtained by augmenting the set of 30 Italian phonemes by means of basic phonetic knowledge. The above described statistical techniques were employed to further refine the set to include 55 phones (**PH55**). Finally, some experimental data on words ending with a consonant (few in Italian, but rather frequent and confusable, because of their short duration) suggested introduction of a special unit in order to model the glottal pulse often occurring at the end of these words (**PH56**). The accuracies were 88.7%, 90.9%, 92.2% using **PH45**, **PH55**, **PH56**, respectively.
Another peculiarity of the Italian language is the high frequency of vowels. The ratio of consonants to vowels in a word, which is particularly low in all Romance languages, is only 1.12 for Italian, while for

English is 1.41 and for German is 1.71 [16]. Therefore, special care was used in modeling vowels: the seven vowel phonemes of Italian are described by eighteen distinct phonetic units.

Estimation of Markov parameters is accomplished by the Baum-Welch algorithm [17], which attempts to maximize $P(\overline{A} \mid \overline{W})$ for the (known) training text uttered by the speaker.
In the standard training procedure, the user of the dictating-machine prototype is requested to read a text, which will be called **L** in the following, consisting of 100 meaningful sentences (1063 total words). The resulting speech sample is about 15-minute long. The text has been designed in order to provide several instances of each phone in a representative set of phonetic contexts.
During recognition, the acoustical model is used to compute the probability $P(\overline{A} \mid W)$. As it is infeasible to carry out the computation for all the words in the vocabulary in real time, the acoustical match consists of two stages. A fast, rough analysis is first performed to discriminate words displaying gross mismatches to the incoming utterance [18]. In this way a small number of words is selected, for which a detailed match computation is carried out.
Sentences are uttered with short pauses between words. However, the decoder does not rely on silence detection to identify word boundaries. A probabilistic determination of the most likely end point of each word is carried out by the acoustical matcher itself. This allows very short pauses between words, while direct silence detection would require long pauses (about half a second) to avoid confusion with silence segments inside words, due to stop consonants.

4. Language modeling

The language model estimates the probability of a word sequence $\overline{W} = w_1 w_2 \ldots w_N$ by evaluating the probability of each word, given the left context of the sentence:

$$P(w_1 \ldots w_N) = \prod_{i=1}^{N} P(w_i \mid w_1 \ldots w_{i-1}).$$

In accordance with the statistical approach, the estimator is built from relative frequencies extracted from a large corpus of sentences. To estimate the probability of a word, contexts with the same last $N-1$ words are considered equivalent (N-gram language model [20]):

$$P(w_i \mid w_1 \ldots w_{i-1}) = P(w_i \mid w_{i-N+1} \ldots w_{i-1})$$

A value $N = 3$ (trigram language model) was actually used. The predictive power of a probabilistic language model is measured by *perplexity*[19] , which can be regarded as the average uncertainty (the *branching factor*) [19] of the model expressed by the equivalent number of equiprobable words.
The language model is built on a backing-off approach [20], combining N-gram statistics (computed from a corpus of 107 million words) and the Turing's statistical technique to estimate the probability of linguistic events not observed in the corpus [20][21] . The threshold for bigram and trigram discount factors was chosen as in [20]. Turing's formula was tested on a 10 million word corpus and showed results very close to experimental data [21].
The twenty thousand words in the system's vocabulary were chosen as the most frequent ones over a subset (44 million words) of the corpus used for language model training, which was taken from magazine and daily newspaper articles and from news-agency flashes on economy and finance, provided by "*Il Mondo*" weekly magazine, the "*Sole 24 Ore*" daily newspaper and the "*Ansa*" agency, respectively. The vocabulary gives a coverage of 96.5 % on disjoint test sets taken from the same sources as the training corpus.

The language model gives perplexities of 98 and 86 on the text used for decoding tests and on a disjoint text taken from the same sources as the training corpus, respectively.

5. Decoding tests

The following table shows the word-recognition accuracy of the decoder as measured on 62 test sentences amounting to 1043 words.

Table 1. Speech recognizer performances.. Average, best and worst recognition accuracies concerning speakers with various degrees of experience in using the decoder for 20000-word vocabulary.

Speakers			Accuracy (%)		
Experience	Gender	No. subjects	A	B	W
Good	M	5	97.5	98.2	96.4
None	M	10	96.3	98.0	94.2
None	F	6	96.3	98.2	94.8

6. Current research areas

In this section a brief overview of the topics in speech recognition area currently being investigated is given.

FAST SPEAKER ADAPTATION

The 15-minute training speech sample L is normally required from each speaker to find an optimal set of prototype vectors for the codebook (via k-means clustering), and to compute HMM parameters, i.e. transition and emission probabilities.

Speaker-independent recognition experiments were performed (using the 6000-word vocabulary recognition system) by collecting speech samples by 10 speakers and computing common prototypes and probabilities; recognition rates ranging from 84% to 93% were achieved on new speakers. The techniques we are studying [2] are aimed at enhancing recognition accuracy by adapting the common prototypes and probabilities by a rapid analysis of a short (about 1-minute) speech sample S provided by the new speaker.

We took into consideration both the acoustic codebook and the HMM parameters estimation aspects. We rely on multi-speaker (rather than on single-speaker) references, to avoid dependency of the results on the acoustical similarity between the reference and the new speaker.

For *codebook computation*, the problem of the statistical insufficiency of the adaptation sample S is addressed according to two approaches:

1. Vector prototypes are modeled as Gaussian probability distributions. The *a priori* probability distributions of the prototypes means are estimated from sample L uttered by each of 10 speakers. Then, for each new speaker, the *a posteriori* means of the adapted prototypes, given S, are computed via Bayesian learning. For sake of computational efficiency, a diagonal covariance matrix is assumed.
2. As the recognizer performs Euclidean, rather than Gaussian, labeling of acoustic vectors, we extended the deleted-estimation technique [17] to an Euclidean framework, to find an optimal interpolation between the common prototypes C_k and the prototypes S_k obtained from S. The i-th component of the adapted prototype A_k is given by

$$A_{ki} = \lambda_{bi} C_{ki} + (1 - \lambda_{bi}) S_{ki}$$

where b indicates a *bin* dependent on the amount of data available for prototype k in S. λ_{bi} is estimated by minimizing total distortion.

Both techniques allow computation of adapted prototypes in few seconds. The following table shows recognition rates for 3 speakers, using clustered (from sample **L**), common and adapted (by technique 1 and 2 respectively) prototypes. In all cases, a complete training of the HMM parameters on sample **L** was performed.

Table 2. Different vector prototypes.. Recognition accuracies for 3 speakers using vector prototypes obtained with various techniques. Data refer to 6000-word vocabulary recognizer.

Spk	CLUS	COMM	ADP1	ADP2
SSS	98.0	95.7	98.0	97.7
STR	95.7	90.0	95.7	95.4
AFS	96.1	93.8	94.2	94.2

For fast *HMM parameters estimation*, we are applying deleted estimation to find the optimal (in the maximum likelihood sense) interpolation between common and speaker-dependent (obtained from S) statistics.

SPEECH DATABASE

An (almost completely) automatic approach to the problem of building a very large time-aligned speech database has been developed [3]. We used this approach to collect more than 30 hours of speech uttered by 10 different speakers, corresponding to over 62000 words. The data were afterwards aligned to their phonetic transcriptions.

The system architecture is composed of IBM PC-ATs equipped with attached A/D/A converters and signal processors [22]; optical devices which allow large, write-once, direct-access storage; a host mainframe; a token-ring network connecting the PCs and the host.

The speech collected according to the mentioned technique is stored in real time on the optical disk. The speech signal may then be transformed by techniques such as Fast Fourier Fransform, Linear Predictive Coding, and cepstral analysis. For the purpose of phonetic alignment, we process the signal through the acoustic front-end of the speech recognizer (see section 2) These preliminary computations are performed by the workstation; the time-alignment and checking process then takes place on the host mainframe.

We align sequences of codewords to their phonetic transcription using the Viterbi algorithm [12]. The aligned waveforms must then be analyzed in order to correct errors. These may come either from inaccuracies due the statistical nature of the Viterbi algorithm, or from problems in the recorded data, due to undesired noise or speaker mistakes. We propose a technique which overcomes the need of a complete listening of the recorded utterances [23] and produces results of comparable accuracy.

Our technique consists in performing several statistical tests to find possibly incorrect word-aligned speech segments. Gross errors are identified by the Viterbi algorithm itself. An independent likelihood measure of the obtained alignments is provided by a statistical model of the duration of the phonemes. We also compute a more detailed likelihood measure which assumes a Poisson distribution for the probability $P(C|W)$ of the codewords produced by the Markov source associated to each word [24]. We found that is much more practical to impose a likelihood threshold on $P(W|C)$ rather than on $P(C|W)$. $P(W|C)$ was estimated through the Bayes' formula.

This automatic process classified an average of 2.5% of the utterances as suspect. They were then manually examined by using an interactive system allowing high quality graphical display and replay of selected speech segments.
The whole process of database construction, consisting of recording, analysis, checking and correction of wrong utterances, took less than six weeks.

AUTOMATIC PHONETIC TRANSCRIPTION

In the development of our prototype we use Automatic Phonetic Transcription (APT) [4] for the *design* of the phonetic structure of the words of the initial vocabulary as well as for its *personalization*, i.e. adding of new words by the user. We propose an approach where phonotactical knowledge (well described by a set of formal rules) is separated from lexical knowledge (largely based on experience and not suitable to a formal description).

We built a rule-based phonotactical APT system which, for each input word, outputs a set of possible transcriptions (5.1 on average for our Italian vocabulary) which always includes the correct one.
In the *design* process, the choice of the correct transcription is currently performed manually, by means of an efficient interactive system; for *personalization*, the user is asked to provide the spelling and a sample utterance of the new word and the most likely transcription is automatically selected, by means of a statistical algorithm.

VOICE RECOGNIZER USER ACCEPTANCE

We performed some preliminary experiments in order to assess the usability, efficiency and user acceptance of the system, and to obtain hints about possible enhancements.
Our experiments studied the task of dictating to the machine by reading a printed text. We selected an article from "*Il Sole 24 Ore*," the major Italian business newspaper, and asked several users to input it into the workstation twice: once they used the voice recognition capability of the system, and the other time they used the keyboard only. The two sessions took place in different days and in varying order. The text to be dictated was statistically representative of the texts to which the prototype is aimed.
During the experiments, the workstation recorded the behavior of the user, by keeping trace of: duration of the session; words uttered to the system in normal and in single-character mode; commands given by voice; keys pressed for character input, text manipulation, cursor movement; number of times the microphone was switched on and off.
A questionnaire was submitted to all participants to the experiment, in order to record their background in the use of keyboard and of voice recognition, their habits and wishes regarding text input, and their impressions and opinions about the usage of the system.

Participants to the experiments were 10 employees of IBM Rome Scientific Center. All of them had several years of experience of electronic text editors and used heavily the keyboard in their everyday work. Such a group of users represents an especially severe test for speech input, because of its out-of-average skills with typing.
The users can be divided into three groups according to their previous experience with voice input and to their knowledge of professional typing:

A users who have some previous experience of voice input and who need to look at the keyboard when typing (three persons);

B users who have no previous experience of voice input and who need to look at the keyboard when typing (five persons);

C users who have no previous experience of voice input and who don't need to look at the keyboard when typing (two persons).

All users preferred to input the text in a raw way first, and then revised it and made corrections. We measured the following values:

Tag	Meaning
IT	Input Time, taken by first raw input of text;
RT	Revision Time, taken by revision and correction of text;
TT	Total Time for input and correction of text;
IE	Input Errors (percent fraction of wrong words after first input);
NE	Net Input Errors, i.e. percentage of wrong words due to speaking, typing or recognition errors, and not due to the absence of the dictated word from the recognizer vocabulary;
FE	Final Errors, i.e. percentage of wrong words due not to correcting.

The following table shows the above listed average values for the three groups, for voice and keyboard input (times are in minutes):

Table 3. Voice and keyboard input. The table shows the average values for the three groups (time in minutes). See text for tag description.

Group	Mode	IT	RT	TT	IE	NE	FE
A	VOICE	13.0	9.0	22.0	6.5	3.3	0.5
A	KEYB.	21.3	6.7	28.0	2.5	2.5	1.2
B	VOICE	17.0	17.3	34.3	8.5	5.8	1.5
B	KEYB.	23.0	6.0	29.0	1.3	1.3	0.7
C	VOICE	20.5	19.5	40.0	8.8	6.1	1.5
C	KEYB.	16.5	5.5	22.0	0.5	0.5	0.1

For all speakers, except professionally trained typists (group C), text input is faster by voice than by keyboard, even if they are using a speech recognizer for the first time. The word input rate achieved in the experiments by speakers of group A by dictation was anyhow higher than that achieved by professionally trained typists when using the keyboard.

The number of errors after the first input of the text was higher for voice input than for keyboard input. This is reflected by the longer time taken by revision and correction. Users of group A were more efficient in the revision task, because users of groups B and C were experiencing voice editing commands for the first time and were brought to over-experiment with them.

Text revision seems the task which can benefit more from user experience and from improvements to the user interface (as well as from higher recognition accuracy). Errors found in a text input by voice are of a different kind than those produced using the keyboard: all the words transcribed by the system belong to the vocabulary. A spelling checker would be of little help. The system could provide instead, for each recognized word, upon request, a list of words very likely to be confused with it.

The indication that voice input is easier to learn and less tiring than traditional keyboard input is suggested by the answers to the questionnaire. 60% of the subjects said that voice editing commands are more natural

and easier to learn than keyboard commands. while 20% found no difference. All users learned in few minutes to insert pauses between words.

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. Future studies on the usage of the voice-activated text editor will address the behavior of users who gained more experience in the tool, and of users who are not accustomed to word processing. Dictation for text creation will also be investigated.

References

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

[2] P. D'Orta, M. Ferretti, S. Scarci, **Fast Speaker Adaptation for Large-Dictionary Real-Time Speech Recognition,** *IEEE Workshop on Speech Recognition,* Arden House, Harriman, NY, May 31-June 3, 1988.

[3] M. Brandetti, P. D'Orta, M. Ferretti, S. Scarci, **Building Reliable Large Speech Databases: an Automated Approach,** *EUSIPCO-88,* Grenoble, September 5-8, 1988.

[4] S. Scarci, S. Taraglio, **Automatic Phonetic Transcription for Large-Vocabulary Speech Recognition,** *Speech 88, Seventh FASE Symposium,* Edinburgh, 22-26 August 1988.

[5] M. Brandetti, P. D'Orta, M. Ferretti, S. Scarci, **Experiments on the Usage of a Voice-Activated Text Editor,** *Speech 88, Seventh FASE Symposium,* Edinburgh, 22-26 August 1988.

[6] A. Averbuch et al., **Experiments with the Tangora 20000 Word Speech Recognizer,** *IEEE Int. Conf. on Acoust., Speech and Sig. Proc.,* Dallas, TX, April 1987, pp. 701-704.

[7] G. Shichman et al., **An IBM PC Based Large-Vocabulary Isolated-Utterance Speech Recognizer,** *IEEE Int. Conf. on Acoust., Speech and Sig. Proc.,* Tokyo, April 1986, pp. 53-56.

[8] P. D'Orta, M. Ferretti, A. Martelli, S. Melecrinis, S. Scarci, G. Volpi, **A Speech Recognition System for the Italian Language,** *ICASSP 1987,* Dallas, pp. 841-843.

[9] F. Jelinek, **A Fast Sequential Decoding Algorithm Using a Stack,** *IBM Journal of Research and Development,* vol. 13, November 1969, pp. 675-685.

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

[11] 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.

[12] G.D. Forney, **The Viterbi Algorithm,** *Proceedings of the IEEE,* vol. 61, no. 3, March 1973, pp. 268-278.

[13] P. D'Orta, M. Ferretti, S. Scarci, **Language-Specific Knowledge in the Probabilistic Approach to Speech Recognition,** *EUSIPCO-88,* Grenoble, September 5-8, 1988.

[14] S. Kullback, **Information Theory and Statistics,** *New York, Dover,* 1969.

[15] P. D'Orta, M. Ferretti, S. Scarci, **Phoneme Classification for Real Time Speech Recognition of Italian,** *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing,* Dallas, TX, April 1987, pp. 81-84.

[16] R. Carlson et al., **Phonetic and Orthographic Properties of the Basic Vocabulary of Five European Languages,** *STL-QPSR, KTH,* Stockholm, 1985, pp. 63-94.

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

[18] P. D'Orta, **Acoustic Discrimination among Words Based on Distance Measures,** *European Conference on Speech Technology,* Edinburgh, Sep. 1987, vol. 2, pp. 329-332.

[19] 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.

[20] S. Katz, **Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer,** *IEEE Trans. on Acoust., Speech and Sig. Proc.,* vol. ASSP-34, no. 3, March 1987, pp. 400-401.

[21] P. D'Orta, M. Ferretti, G. Maltese, S. Scarci, **Analisi automatica di testi per la costruzione di modelli della lingua italiana con applicazione al riconoscimento della voce,** *Atti del Convegno AICA,* Cagliari, Settembre 1988.

[22] G. Shichman, **Personal Instrument (PI) - A PC-based signal processing system,** *IBM Journal of Research and Development,* vol. 29, no.2, March 1985, pp. 158-169.

[23] R. Leonard, **A Database for Speaker-Independent Digit Recognition,** *IEEE International Conference on Acoustics, Speech and Signal Processing.* San Diego, CA, April 1984, 4.7.

[24] L.R. Bahl, R. Bakis, P.V. de Souza, R.L. Mercer, **Polling: A Quick Way to Obtain a Short List of Candidate Words in Speech Recognition,** *IEEE International Conference on Acoustics, Speech and Signal Processing.* New York, April 1988, 36.S11.