Large-vocabulary speech recognition: A system for the Italian language

by Paolo D'Orta
Marco Ferretti
Alex Martelli
Sergio Melecrinis
Stefano Scarci
Giampiero Volpi

We describe a research project in automatic speech recognition which has led to the development of an experimental large-vocabulary real-time recognizer for Italian, and show how the maximum-likelihood techniques which had been employed in the development of prototype recognizers for English can be tailored to a language with substantially different characteristics.

Introduction

Existing speech-recognition technologies have proven adequate for simple tasks involving small vocabularies (tens or hundreds of words) and suitable for limited applications (typically, recognition of a set of commands uttered in an isolated fashion by an operator whose hands are busy). Systems found on the market are usually independent of the target language. Interesting applications in an office environment, such as text dictation and database query, need, on the other hand, the ability to handle natural language and pronunciation. This requires large vocabularies (thousands of words) and substantially more sophisticated techniques which take into account language-specific knowledge on phonology, syntax, and (surface) semantics.

Two approaches are currently popular in the research community. One approach, based on *a priori* expert knowledge of the acoustic properties of spoken language [1, 2], attempts to decode the phonetic information conveyed by speech, by means of Artificial Intelligence methods. Sometimes a linguistic analysis based on syntactic and semantic rules [3] is also performed. The other approach employs statistical models for both acoustic and linguistic analysis, and has already been successfully applied to develop experimental speech-recognizer prototypes. 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 implementation of the model by automatic learning from data.

The most notable example of the latter approach is found in the techniques proposed by researchers at the IBM Thomas J. Watson Research Center, which led to the development of prototype large-vocabulary real-time recognizers of spoken English [4, 5]. Our work on speech recognition of Italian was founded on these techniques. The same methodology is being applied to large-vocabulary recognition of French by a research group at the IBM Paris Scientific Center [6].

One feature distinguishing our approach is the identification of the human-determined elements of the model that could be deemed independent of the target natural language, and those that had to be changed, thereby showing that the method can be successfully tailored to different languages.

The real-time isolated-utterance speech-recognition system we developed for the Italian language handles dictionaries of

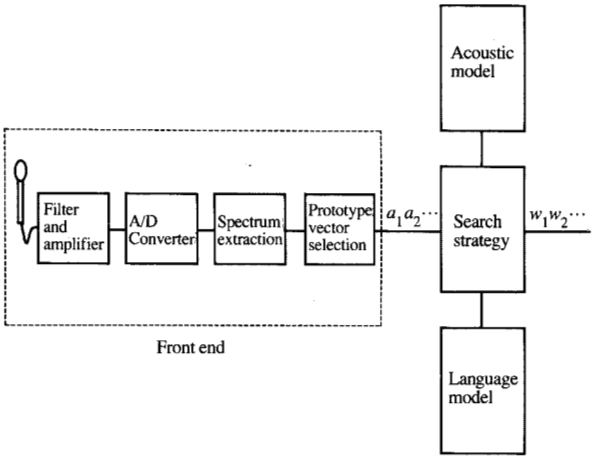

Figure 1

Structure of the probabilistic speech recognizer.

up to 6500 words. Experimental recognition accuracy is over 96% [7].

Research on large-vocabulary Italian speech recognition has come to be of widespread interest and is being pursued at several public and private institutions [8–10]. No prototype showing comparable performances has as yet been made public.

In the next section a brief description of the probabilistic approach is provided. Acoustic and language modeling of Italian, and architectural issues are discussed in the succeeding sections. The last sections offer experimental results and remarks about possible extensions to other languages.

The probabilistic approach to speech recognition: A review

This section gives, for the convenience of the reader, a brief description of the probabilistic approach to speech recognition.

Let $\bar{W} = w_1 w_2 \cdots w_N$ be a sequence of N words, and let $\bar{A}$ 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 which maximizes the conditional probability $P(\bar{W} \mid \bar{A})$, i.e., the most likely word sequence given the acoustic information. By Bayes' theorem,

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

$P(\bar{A} \mid \bar{W})$ is the probability that the sequence of words $\bar{W}$ will produce the acoustic string $\bar{A}$, that is, the probability that the speaker, pronouncing the words $\bar{W}$, will utter sounds described by $\bar{A}$. $P(\bar{W})$ is the *a priori* probability of the word string $\bar{W}$, that is, the probability that the speaker will wish to pronounce the words $\bar{W}$. $P(\bar{A})$ is the probability of the acoustic string $\bar{A}$; it is not a function of $\bar{W}$, since it is fixed once $\bar{A}$ is measured, and can thus be ignored when looking for the maximum over $\bar{W}$.

A consequence of this equation is that the recognition task can be decomposed into the following subtasks (**Figure 1**):

1. Perform *acoustic processing* to encode the speech signal into a string of values $\bar{A}$ representative of its acoustic features, and, at the same time, adequate for a statistical analysis.
2. Compute the probability $P(\bar{A} \mid \bar{W})$ (for this purpose an *acoustic model* must be created).
3. Evaluate $P(\bar{W})$ (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*.

Acoustic processing

In our system, as in the prototypes for the English language, acoustic processing is implemented (in an *acoustic front-end*) by a model of the human auditory system [11] and a vector quantizer. The digitized acoustic signal (20K samples per second, 12 bits per 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 with the frequency sensitivity of the human ear), and transformed nonlinearly to take into account 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 precomputed codebook of 200 elements, as shown in **Figure 2**.

Search strategy

The search strategy is based on the *stack sequential decoding* algorithm [12]. 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.

Acoustic model

The acoustic model is based on Markov sources [13]. A Markov source of acoustic labels is essentially a probabilistic finite-state machine. At fixed intervals of time a random transition is taken, which may or may not cause a change of state, and a random acoustic label is emitted. Transitions

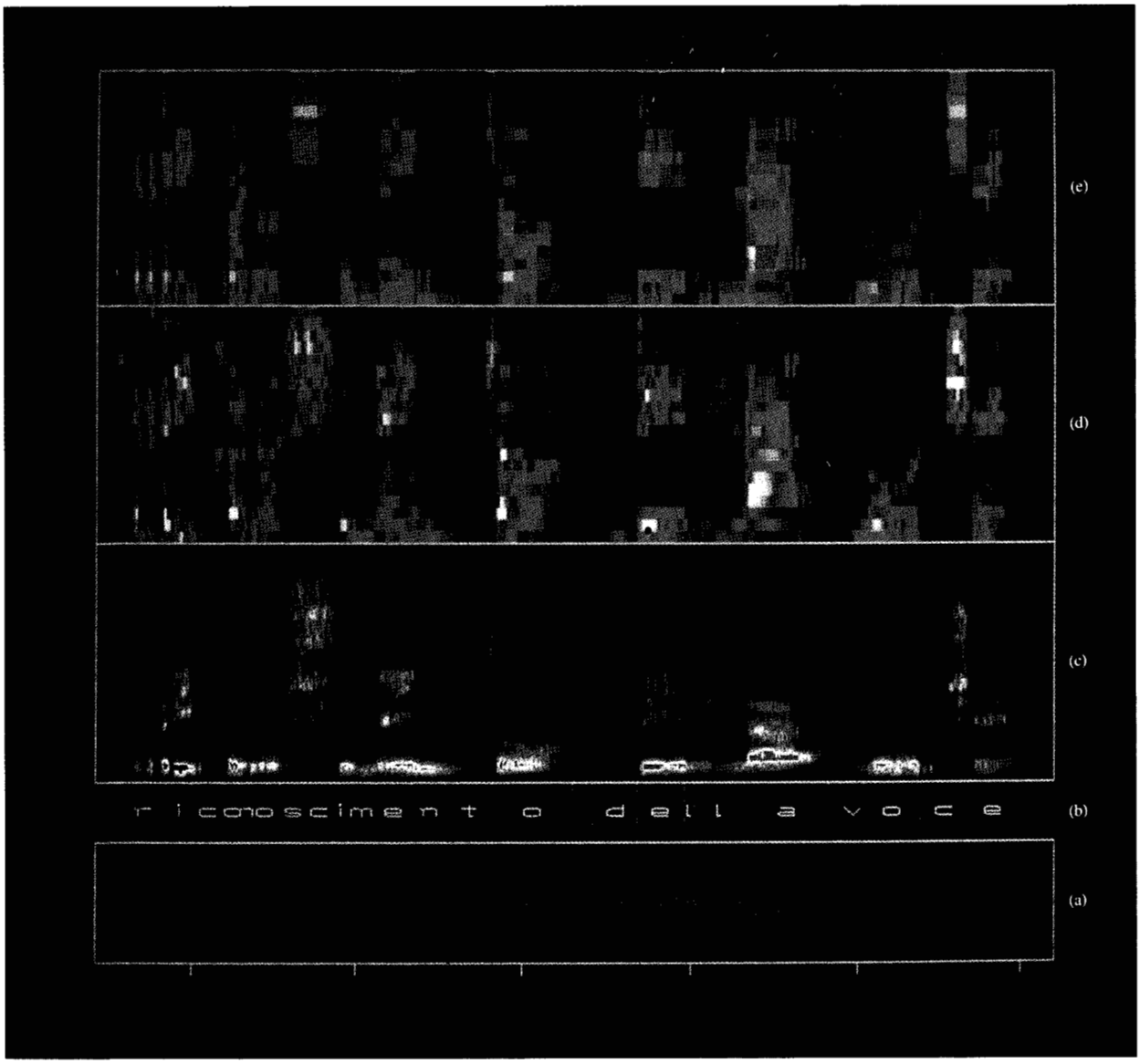

Figure 2

Acoustic representation of the Italian phrase *riconoscimento della voce* ('voice recognition'): (a) speech waveform; (b) spelling segmented to align with (c) wideband energy spectrum; (d) centisecond acoustic vectors; (e) vector prototypes corresponding to acoustic labels.

and label emissions occur according to probability distributions depending only on the source state, not on previous history (**Figure 3**). While it is possible to observe the string of labels produced by the source, the sequence of states it visits remains hidden. These models are therefore named *hidden Markov models*. For the purpose of speech recognition, a *phonetic unit*, modeled by a Markov source, is associated with each of the basic sounds of the language. A word is described by the concatenation of the Markov sources corresponding to the string of phonetic units forming its pronunciation (**Figure 4**). Estimation of the probability parameters of the Markov models (*acoustic training*) is accomplished by the Baum–Welch algorithm [14], which attempts to maximize $P(\bar{A} \mid \bar{W})$ for a known training text uttered by the speaker. Algorithms are also available for the task of *acoustic matching*, i.e., the evaluation of $P(\bar{A} \mid \bar{W})$ when performing speech decoding, according to the model parameters computed during training.

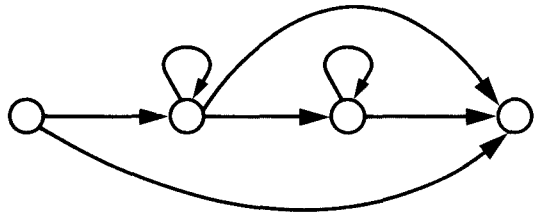

Figure 3

Example of Markov model of a phonetic unit. Every centisecond a transition is taken and an acoustic label is emitted.

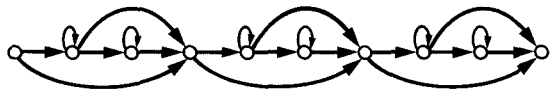

Figure 4

Markov model of a word, obtained by concatenation of the models of its phonetic units.

Language model

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

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

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 [15]):

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

The predictive power of a probabilistic language model is measured by *perplexity* [16], defined as

$$p = 2^{\tilde{H}},$$

where $\tilde{H}$ is an estimate of the entropy (according to the language-model probability $\tilde{P}$) computed on a text $w_1 \cdots w_L$ generated by the source which is being modeled:

$$\tilde{H} = -\frac{1}{L} \times \sum_{i=1}^{L} \log_2 \tilde{P}(w_i \mid w_{i-N+1} \cdots w_{i-1}).$$

Perplexity is the average uncertainty (the *branching factor*) of the model expressed by the equivalent number of equiprobable words.

Unlike the acoustic front-end and the search strategy, the acoustic model and the language model take into account specific properties of the Italian language. Only a general description is given here, while language-specific issues are discussed in greater detail in the following sections.

Acoustic modeling of Italian

To achieve good recognition accuracy, it is necessary to design a specific set of Markovian phonetic units (the *phonetic alphabet*) to describe the pronunciation of the words of the language. This set, in order to preserve the linguistic information conveyed by the utterance of a word, should not be simpler than the set of *phonemes*, the classical units defined by the phonology of the language as classes of sounds carrying the same linguistic information. The phonetic alphabet should also describe the most relevant of systematic speech-variability phenomena (such as stress and coarticulation) not reflected by phonemes. A too-detailed model, involving a large number of parameters, might require an unacceptably large statistical sample of the speaker's voice for training. The design of the phonetic alphabet should, then, look for the best trade-off between detail of modeling and brevity of training. Some researchers use units based on structural elements more complex than phonemes, such as diphones, demisyllables, or syllables. In order to keep the number of parameters low, we based our units on an augmented set of phonemes (context-dependent phonemes are also proposed in [17]). This allowed the choice of a single topological structure—designed to provide enough degrees of freedom—for all the Markov sources associated with the phonetic units. Differentiation among phonetic Markov sources is thus left entirely to the parameter-estimation process (acoustic training).

A systematic procedure for finding an optimal phonetic alphabet has not yet been developed. Our approach combines the results of traditional acoustic and phonetic research with analysis of statistical data. The procedure is largely a trial-and-error process. We introduce modifications to the phonetic alphabet (initially composed of the original 30 Italian phonemes) and then verify whether an improvement has occurred. For the purpose of data analysis and performance evaluation, a large multispeaker speech database (more than 50000 utterances of individual words) was built. The speech signal is aligned to the Markov sources describing the spoken sentences by means of the Viterbi algorithm [18], thus finding the segments of the utterance corresponding to each phonetic unit.

Possible modifications to the phonetic alphabet are often suggested by phonological considerations. An example is the sound /n/, normally an alveolar nasal, which becomes velar when followed by a /g/ or /k/: This leads to the definition of a separate phonetic unit for the latter case. Potentially weak phonetic units are also found by performing recognition tests on utterances from the database, without using the information provided by the language model (which may mask acoustic-model inaccuracies), and by analyzing decoding errors. The phonetic description of words most frequently unrecognized is studied for possible improvements.

In order to verify whether a modification to the phonetic alphabet produces an improvement, the most conclusive measure consists in performing recognition tests on several speakers. We developed some faster measures which proved very helpful. A modification frequently introduced is the modeling of a sound, previously described by a single phonetic unit M, by two new units M_1 and M_2, chosen on the basis of the phonetic context (as for the above-mentioned case of the sound /n/). To measure the value of the modification, we estimate whether the utterances of the new units show systematic, statistically significant differences, by computing their *Kullback divergence* (or *cross-entropy*), defined as

$$d(M_1, M_2) = \sum_{\bar{A}} P(\bar{A} \mid M_1) \log \frac{P(\bar{A} \mid M_1)}{P(\bar{A} \mid M_2)} + \sum_{\bar{A}} P(\bar{A} \mid M_2) \log \frac{P(\bar{A} \mid M_2)}{P(\bar{A} \mid M_1)},$$

where the summation should include all possible strings $\bar{A}$ of acoustic labels. A global measure of the quality of the phonetic representation is provided by the mutual information between the phonetic alphabet **M** and the set of speech alignments **A**:

$$m(\mathbf{M}, \mathbf{A}) = \sum_{\bar{A} \in \mathbf{A}} \sum_{M \in \mathbf{M}} P(\bar{A}, M) \log \frac{P(\bar{A} \mid M)}{P(\bar{A})}.$$

A significant increase of mutual information is a good index of an improvement of the phonetic alphabet. Practical methods for estimating divergence and mutual information are described in [19].

A 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 it is 1.41 and for German, 1.71 [20]. Therefore, special care was used in modeling vowels: The seven vowel phonemes of Italian are described by eighteen distinct phonetic units.

To achieve increased tolerance for regional accents, we introduced "ambiguous" phonetic units. An example is the vowel "e," which, according to correct Standard Italian pronunciation, should be open (/ɛ/) in some words and close (/e/) in others. The "e" of several words, though, is subject to mispronunciation (sometimes due to hypercorrection because the two vowels have merged in the native dialects of the speaker). Our first model strictly respected the correct Standard pronunciation, and included only two units for the stressed "e," *EO* and *EC*. This led to poor training and recognition of some speakers. We then introduced the unit *EX*, associated with occurrences of "e" subject to mispronunciation. For one speaker, for example, divergences computed after the modification were

$d(EO, EC) = 20.0,$

$d(EX, EC) = 7.1,$

$d(EX, EO) = 8.4,$

while before the modification it was

$d(EO, EC) = 13.2.$

These figures show that the new alphabet presents better discrimination of consistent pronunciations of "e," while the new ambiguous unit is rather well matched to both pronunciations.

Table 1 Word recognition accuracy of sentences from a 1000-word dictionary, uttered by ten speakers decoded without any language model, using three different phonetic alphabets.

Phonetic alphabet	*Recognition accuracy (%)*		
	Average	*Best*	*Worst*
PH45	88.7	91.9	84.6
PH55	90.9	93.9	85.6
PH56	92.2	95.1	89.5

Table 1 shows the recognition accuracy achieved for 50 test sentences (1025 total words, extracted from a dictionary of 1000 words), uttered by each of ten speakers and decoded without any language model, employing three different phonetic alphabets:

PH45 The 30 Italian phonemes augmented to 45 units, on the basis of simple phonological considerations.
PH55 An extended set built by means of the previously described techniques.
PH56 The previous set with the addition of a special unit to model the glottal pulse produced at the end of words with a final consonant.

The addition of the glottal-pulse unit notably increased performance, in spite of the fact that few Italian words end in a consonant, because (when PH55 was used) those words were often confused with similar words ending in a vowel.

An essential problem is the design of the training text. It should be kept as short as possible, but each phonetic unit should be represented many times in several different contexts in order to provide enough data for good estimation of the Markov parameters. Our experiments show that substantially better recognition accuracy is achieved when the training text is created from meaningful sentences rather than random sequences of words. This ensures higher consistency with the sentences uttered during recognition sessions. Therefore the training text is built manually.

In a large-dictionary real-time speech-recognition system, it is computationally very demanding to perform a detailed match of the input utterance to all the items in the vocabulary. A commonly accepted solution is to carry out recognition in more than one stage. In the first stages a fast, rough analysis is performed to eliminate items displaying gross mismatches to the incoming utterance. In this way a small number of items are selected, the most likely being identified in the last stage, through a detailed match computation.

We investigated an interesting approach to *fast acoustic matching*, consisting in grouping words into equivalence classes, in order to represent more than one word by a single acoustic model. During recognition, the utterance is initially matched against class models, and thereafter against the individual models of the words belonging to the selected classes. Let c be a generic word class, $C_F(c)$ the computational cost of a fast match against class c, and N_c the number of classes; similarly, let w be a generic word, $C_D(w)$ the computational cost of a more detailed match against word w, and N_w the average number of words selected by the fast-match stage. A small number of classes saves computation in the first matching stage:

$$C(Fast\_match) = N_c E[C_F(c)].$$

On the other hand, because the class model must represent all the words in the class, a large number of words per class leads to inaccurate models, and to low selectivity; the computational cost of the detailed match grows with N_w:

$$C(Detailed\_match) = N_w E[C_D(w)].$$

A good classification should reduce the number of classes to a minimum without losing accuracy. We studied two different methods:

1. Define a distance measure between word models and perform clustering [21].
2. Find broad phonetic categories and map the phonetic units of each word into them, so that each sequence of phonetic categories identifies a word equivalence class [19].

An automatic method of selecting phonetic categories consists in looking for the partition $\hat{\mathbf{P}}$ of the set of phonetic units which has the highest mutual information with respect to the acoustic labels, given a target number of classes k:

$$m(\hat{\mathbf{P}}, \mathbf{A}) = \max_{\mathbf{P}} m(\mathbf{P}, \mathbf{A}).$$

Due to the combinatorial explosion of the number of partitions, an exhaustive search is infeasible, and some heuristic method must be used. Our approach consists in looking for the best partition, starting with one unit per category and reducing the number of categories by successive merges. The traditional greedy technique, which iteratively carries out the best merge of two categories, ensures optimization on a local scale only. An improvement to this technique consists in performing, each time two classes are merged, all those movements of a single element which increase mutual information. We found that still better results can be obtained by applying a more advanced heuristic method of state search, formally identical to the tree ordered-search algorithm [22]. We associate to a partition $\mathbf{P}$ into n classes a *cost* expressed by

$$C(\mathbf{P}) = \eta(n) - m(\mathbf{P}, \mathbf{A}),$$

where m is the mutual information for $\mathbf{P}$ and η is the highest expected mutual information for a partition of cardinality n. This enables us to compare the cost of partitions of different cardinalities. The algorithm keeps an ordered list of lowest-cost partitions, initialized to contain only the trivial partition with one unit per class. Iteratively, it computes new partitions (by performing merges of two classes belonging to the best partition so far) and inserts them into the ordered list.

This is the best partition into six classes found by the algorithm for one speaker:

- /a/ sounds.
- Back vowels.
- Most front vowels, one liquid, and one nasal.
- Most liquids and nasals, one /e/ sound, and one voiced plosive.
- Plosives.
- Fricatives.

This classification yields a fast match with low computational cost, but selectivity remains unsatisfactory. We are currently studying methods to identify classification techniques which jointly optimize the cost of fast match and of detailed match. The word-classification techniques were applied to Italian, but are immediately extendible to other languages.

Language modeling of Italian

Our corpus was formed from a set of magazine articles and news-agency flashes on economy and finance, amounting to about ten million words. **Figure 5** is a graph of the frequency of the words, ordered by decreasing number of occurrences.

To evaluate the practical usability of our recognizer, coverage by dictionaries of increasing size has been measured (**Figure 6**). Coverage is, as expected, not as complete as for a language like English (which has far fewer inflected forms), but is nevertheless encouraging. The curves represent values for the same dictionaries applied to three corpora on economy and finance:

- Corpus A1 (constructed from news-agency flashes), from which the dictionaries were generated by taking the most frequently occurring words.
- Corpus A2, disjoint from A1 but produced from the same source as A1.
- Corpus M1, produced from a different source (magazine articles).

The three-gram model displayed a performance comparable to that of the English one, although Italian, like the other Romance languages, is significantly different from English on the morphological level (higher number of inflected forms) as well as on the syntactic one (weaker constraints on word order in the sentence; strict gender and number concordance). The experimental perplexity of the three-gram model for the 6500-word dictionary was 110.65 (the two-gram model gave $p = 150.73$, the one-gram model $p = 780.99$).

The choice of $N = 3$ for the N-gram language model is suggested by the size of the corpora available in practice (tens of millions of words), which do not contain enough statistical data for an adequate estimation of probabilities of longer sequences of words. We verified that N-gram language models with $N > 3$, based on statistics collected from the same corpus, display perplexities not significantly lower than the three-gram model.

Perplexity is an intrinsic measure of the predictive power of the language model, which does not take into account its interaction with the acoustic model. A study on coupling effects between acoustic and linguistic models showed that they provide essentially independent information. We first measured how the predictive power of the three-gram language model (expressed by perplexity) changed when its choice was limited to a subset of m words of the vocabulary (including the right word) chosen randomly. The same experiment, performed on subsets selected according to acoustic similarity to the right word, showed no significant differences in the behavior of the perplexity as a function of m.

However, at least for a strongly inflected language like Italian, it should be possible to do even better than such independence. This remark is prompted by experimental results [23] obtained in comparing a three-gram language model with one based on grammatical categories [24]: While the former exhibited lower perplexity, the latter was found to perform better when the acoustical information was taken into account—i.e., when perplexity was measured only on a subset of words selected according to acoustical similarity to the right word.

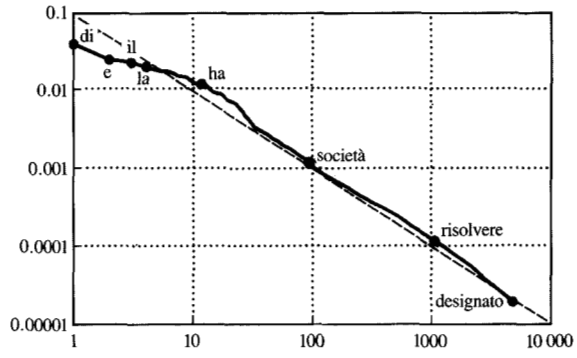

Figure 5

Frequency of words in a corpus of economic and financial news, ordered by decreasing number of occurrences. The curve is well approximated by the function $f(n) = 0.1/n$.

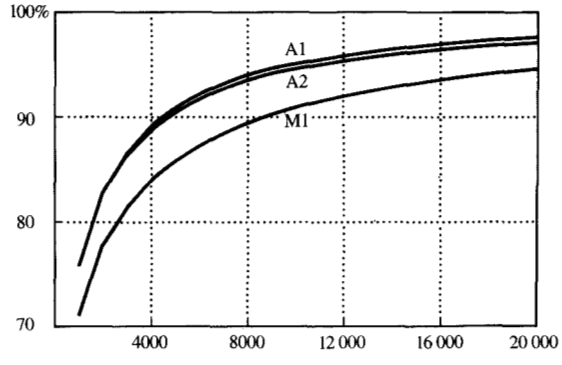

Figure 6

Coverage of the corpus as a function of the size of the dictionary.

The design of the system took into account language-specific phenomena, such as elision and what is known in Italian as apostrophe. For example, to dictate the word *all'*, in phrases like *all'obiettivo*, the user may choose among three pronunciations:

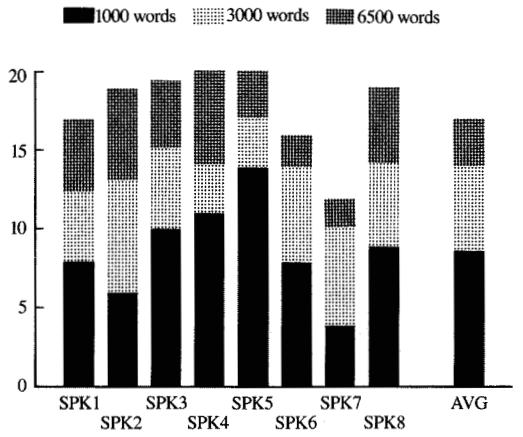

Figure 7

Percent word error rates of recognition experiments for three dictionaries of increasing size. No language model is used.

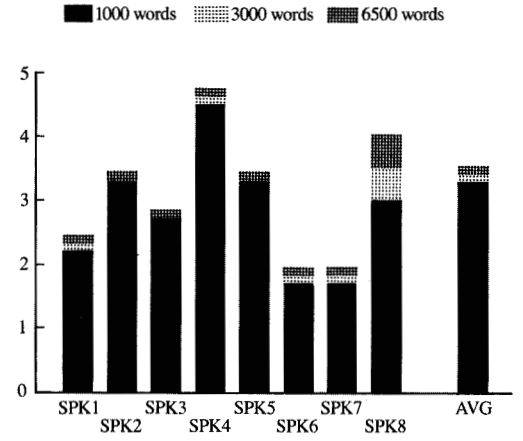

Figure 8

Percent word error rates of recognition experiments for three dictionaries of increasing size. The three-gram language model is used.

1. /al/
2. /allo/
3. /allapostrofo/

Pronunciation 1 is common to the word *al*, while pronunciation 2 is common to the word *allo*. The language model is able to disambiguate properly.

Architecture

The development of the recognizer has taken place on an experimental system consisting of a workstation based on an IBM Personal Computer AT® [25] equipped with a special signal-processing card [26] (the acoustic front-end) and a host running VM/SP (which handles the decoding stage—search coordination, acoustic and language modeling). The recognized text is displayed on a 3270 terminal. The user can edit the text using the keyboard and produce a hard copy on the attached printer.

An advantage of this architecture is that the decoding stage relies on general-purpose hardware only. This has allowed us to run the speech recognizer on several System/370 mainframes, connected through 3270 cable to the acoustic-processing workstation. An IBM 3090 CPU has provided enough computing power to achieve real-time recognition for the largest dictionary we have developed so far (6500 words).

The recognizer has also been implemented on a single workstation, consisting of a PC AT equipped with two to five special cards, in accordance with the *Tangora* architecture [5, 27].

Experimental results

Several speakers trained the system by reading a 20-minute text. The personalized parameters were used to perform recognition tests on sets of 50 meaningful sentences (1025 total words) uttered by the speaker. **Figure 7** shows the percent error rates achieved when no language model is employed (that is, the words in the dictionary are considered equiprobable), for recognizers based on three dictionaries of different sizes (1000, 3000, and 6500 words). **Figure 8** refers to the same recordings decoded using the language model.

The amount of computation performed in the decoding stage, C, displayed a sublinear increase with respect to the dictionary size: $C(3000)$ was $2.3 \times C(1000)$, while $C(6500)$ was $4.0 \times C(1000)$.

Some recognition experiments were performed on speakers who had not previously trained the system. The Markov parameters had been trained on a mixture of voices of ten speakers. Recognition accuracy, for a 1000-word dictionary, ranged from 89% to 94%.

Conclusions

Recent research on probabilistic methodologies has provided a powerful set of techniques for modeling the complex phenomenon of natural speech. These techniques have the advantage of allowing the model designers to concentrate on the specific structural properties of the language, while leaving the task of a detailed quantitative description to

automatic statistical methods. Thus, in building a structural description of the language under study, statistics and information theory can usefully integrate the knowledge provided by disciplines such as acoustics and linguistics.

The results achieved in prior studies confirm the excellent performance displayed by this approach for the English language. We believe that it can be successfully applied to many other languages. Some potential problems are the following:

- The acoustic front-end may require different analysis of the signal to take into greater account features (such as pitch) which in some languages carry more relevant linguistic information than in English and Italian.
- Languages with many forms for each lexeme and/or many compound words may need substantially larger dictionaries to achieve acceptable coverage.
- Languages with a large variety of sounds may require a larger acoustic vector codebook or may need continuous modeling.
- It might be unacceptably unnatural to leave pauses between words (we found that Italian speakers become accustomed to it very quickly).

Nevertheless, these methodologies seem a promising basis for the development of continuous-speech recognizers.

Short-term goals of this project will include further extension of the vocabulary and studies on human-factors aspects of the man–machine interface.

Acknowledgments

We would like to thank F. Jelinek and the Speech Research Group of the IBM Thomas J. Watson Research Center for their support, which has made this project possible. We also thank our colleagues of the IBM Rome Scientific Center, M. Brandetti, A. Fusi, and G. Maltese, for their contributions.

References and note

1. V. W. Zue, "The Use of Speech Knowledge in Automatic Speech Recognition," *Proc. IEEE* **73,** No. 11, 1602–1615 (November 1985).
2. R. de Mori and L. Lam, "Plan Refinement in a Knowledge-Based System for Automatic Speech Recognition," *Proceedings, IEEE International Conference on Acoustics, Speech and Signal Processing*, Tokyo, April 1986, pp. 1217–1220.
3. Lee D. Erman, Frederick Hayes-Roth, Victor R. Lesser, and D. Raj Reddy, "The Hearsay-II Speech-Understanding System: Integrating Knowledge to Resolve Uncertainty," *ACM Comput. Surv.* **12,** No. 2, 213–253 (June 1980).
4. F. Jelinek, "The Development of an Experimental Discrete Dictation Recognizer," *Proc. IEEE* **73,** No. 11, 1616–1624 (November 1985).
5. A. Averbuch et al., "An IBM PC Based Large-Vocabulary Isolated-Utterance Speech Recognizer," *Proceedings, IEEE International Conference on Acoustics, Speech and Signal Processing*, Tokyo, April 1986, pp. 53–56.
6. H. Cerf-Danon, A. Derouault, M. El-Bèze, B. Mérialdo, and S. Soudoplatoff, "Speech Recognition Experiment with 10,000 Word Vocabulary," presented at the Nato Advanced Institute on Pattern Recognition, Brussels, June 18–20, 1986, pp. 203–208.
7. P. D'Orta, M. Ferretti, A. Martelli, S. Melecrinis, S. Scarci, and G. Volpi, "A Speech Recognition System for the Italian Language," *Proceedings, IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, TX, April 1987, pp. 841–843.
8. R. Billi, G. Massia, and F. Nesti, "Word Preselection for Large Vocabulary Speech Recognition," *Proceedings, IEEE International Conference on Acoustics, Speech and Signal Processing*, Tokyo, April 1986, pp. 65–68.
9. M. Cravero, R. Pieraccini, and F. Raineri, "Definition and Evaluation of Phonetic Units for Speech Recognition by Hidden Markov Models," *Proceedings, IEEE International Conference on Acoustics, Speech and Signal Processing*, Tokyo, April 1986, pp. 2235–2238.
10. C. Scagliola and D. Sciarra, "Two Novel Algorithms for Variable Frame Analysis and Word Matching for Connected Word Recognition," *Proceedings, IEEE International Conference on Acoustics, Speech and Signal Processing*, Tokyo, April 1986, pp. 1105–1108.
11. Jordan R. Cohen, "Application of an Adaptive Auditory Model to Speech Recognition," presented at the 110th Meeting, Acoustical Society of America, Nashville, TN, November 4–8, 1985; abstract in *J. Acoust. Soc. Amer.* **78,** Suppl. 1, S50 (Fall 1985).
12. F. Jelinek, "Fast Sequential Decoding Algorithm Using a Stack," *IBM J. Res. Develop.* **13,** 675–685 (November 1969).
13. L. R. Rabiner and B. H. Huang, "An Introduction to Hidden Markov Models," *IEEE ASSP Magazine* **3,** No. 1, 4–16 (January 1986).
14. L. R. Bahl, F. Jelinek, and R. L. Mercer, "A Maximum Likelihood Approach to Continuous Speech Recognition," *IEEE Trans. Pattern Anal. & Machine Intell.* **PAMI-5,** No. 2, 179–190 (1983).
15. S. Katz, "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer," *IEEE Trans. Acoust., Speech & Sig. Proc.* **ASSP-34,** No. 3, 400–401 (March 1987).
16. F. Jelinek, R. L. Mercer, L. R. Bahl, and J. K. Baker, "Perplexity—A Measure of the Difficulty of Speech Recognition Tasks," presented at the 94th Meeting, Acoustical Society of America, Miami Beach, December 12–16, 1977; abstract in *J. Acoust. Soc. Amer.* **62,** Suppl. 1, S63 (Fall 1977).
17. Y. Chow, R. Schwartz, S. Roucos, O. Kimball, P. Price, F. Kubala, M. Dunham, M. Krasner, and J. Makhoul, "The Role of Word-Dependent Coarticulatory Effects in a Phoneme-Based Speech Recognition System," *Proceedings, IEEE International Conference on Acoustics, Speech and Signal Processing*, Tokyo, April 1986, pp. 1593–1596.
18. G. David Forney, Jr., "The Viterbi Algorithm," *Proc. IEEE* **61,** No. 3, 268–278 (March 1973).
19. P. D'Orta, M. Ferretti, and S. Scarci, "Phoneme Classification for Real Time Speech Recognition of Italian," *Proceedings, IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, TX, April 1987, pp. 81–84.
20. R. Carlson, K. Elenius, B. Granstrom, and S. Hunnicutt, "Phonetic and Orthographic Properties of the Basic Vocabulary of Five European Languages," *Speech Technology Laboratory—Quarterly Progress and Status Report, KTH* (Royal Institute of Technology), Stockholm, 1985, pp. 63–94.
21. P. D'Orta, "Acoustic Discrimination Among Words Based on Distance Measures," *European Conference on Speech Technology*, Edinburgh, September 1987, Vol. 2, pp. 329–332.
22. N. J. Nilsson, *Problem-Solving Methods in Artificial Intelligence*, McGraw-Hill Book Co., Inc., New York, 1971, pp. 43–79.
23. M. Codogno, L. Fissore, A. Martelli, G. Pirani, and G. Volpi, "Experimental Evaluation of Italian Language Models for Large-Dictionary Speech Recognition," *European Conference on Speech Technology*, Edinburgh, September 1987, Vol. 1, pp. 159–162.

24. R. Campo, L. Fissore, A. Martelli, G. Micca, and G. Volpi, "Probabilistic Models of the Italian Language for Speech Recognition," *Proceedings of the International Workshop on Automatic Speech Recognition*, Rome, May 1986, pp. 49–56.
25. Personal Computer AT is a registered trademark of International Business Machines Corporation.
26. G. Shichman, "Personal Instrument (PI)—A PC-Based Signal Processing System," *IBM J. Res. Develop.* **29,** No. 2, 158–169 (March 1985).
27. A. Averbuch et al., "Experiments with the Tangora 20,000 Word Speech Recognizer," *Proceedings, IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, April 1987, pp. 701–704.

Received March 23, 1987; accepted for publication October 20, 1987

Paolo D'Orta *IBM Italy, Scientific Center, Via Giorgione 159, 00147 Rome, Italy.* Dr. D'Orta graduated in electrical engineering from Rome University in 1983. He joined the IBM Rome Scientific Center in 1985. Since that time, he has been involved in speech-recognition research, with special attention to the acoustic modeling of the Italian language. At the IBM Thomas J. Watson Research Center in Yorktown Heights, Dr. D'Orta worked in cooperation with the projects on speech recognition for English. In 1986, he received an Outstanding Technical Achievement Award for the development of a voice-recognition system (for the Italian language).

Marco Ferretti *IBM Italy, Scientific Center, Via Giorgione 159, 00147 Rome, Italy.* Dr. Ferretti graduated in electrical engineering from Rome University in 1984. He joined IBM at the Rome Scientific Center in 1985. He is currently involved in the development of acoustic models for speech recognition of the Italian language.

Alessandro Martelli *IBM Italy, Scientific Center, Via Giorgione 159, 00147 Rome, Italy.* Dr. Martelli graduated in electrical engineering from Bologna University in 1980. He joined IBM the following year at the Rome Scientific Center, where he has worked on Image Processing and, since 1984, on stochastic modeling of the Italian language for speech recognition. At the IBM Thomas J. Watson Research Center in Yorktown Heights, Dr. Martelli worked in cooperation with the projects on speech recognition for English. In 1986, he received an Outstanding Technical Achievement Award for the development of a voice-recognition system (for the Italian language). He is currently working on a project for high-quality text-to-speech synthesis of Italian.

Sergio Melecrinis *IBM Italy, Scientific Center, Via Giorgione 159, 00147 Rome, Italy.* Mr. Melecrinis received his degree in electrical engineering in 1967. He joined IBM in 1970 and was involved in telecommunications, distributed systems, and image processing. Since 1982, he has worked at the Rome Scientific Center in speech synthesis and recognition for the Italian language. In 1986, Mr. Melecrinis received an Outstanding Technical Achievement Award for the development of a voice-recognition system (for the Italian language).

Stefano Scarci *IBM Italy, Scientific Center, Via Giorgione 159, 00147 Rome, Italy.* Dr. Scarci graduated in electrical engineering from Naples University in 1984. Since joining the IBM Rome Scientific Center in 1985, he has been working on projects dealing with speech recognition, especially in the area of acoustic modeling of the Italian language. At the IBM Thomas J. Watson Research Center in Yorktown Heights, Dr. Scarci worked in cooperation with the projects on speech recognition for English, receiving a Research Division Award in 1985 for his contributions to those projects. The following year he received an Outstanding Technical Achievement Award for the development of a voice-recognition system (for the Italian language).

Giampiero Volpi *IBM Italy, Scientific Center, Via Giorgione 159, 00147 Rome, Italy.* Dr. Volpi graduated in mathematics from Pavia University in 1971. From 1971 to 1974, he worked at the Numerical Analysis Laboratory of the National Research Council in Pavia. In 1974 he joined IBM and was involved in research on numerical fluid dynamics, stochastic interpolation, and groundwater modeling. From 1984 to 1987, Dr. Volpi was responsible for speech-processing projects at the Rome Scientific Center. He is currently manager of the Image Processing Group there.