A FLEXIBLE VOCABULARY RECOGNITION SYSTEM FOR ITALIAN

P. Bonaventura Δ, *L. Fissore*, *H. Leprieur* ◊ *and G. Micca*
CSELT - Centro Studi e Laboratori Telecomunicazioni
Via G. Reiss Romoli 274 - 10148 Torino, Italy
Δ CSELT Consultant ◊ PhD student

ABSTRACT

The present paper describes a flexible vocabulary speech recognition system developed at Cselt. Main modules are described and some recognition results are provided. Details on the phonetic component are given. Preliminary tests have been performed on a speaker-independent, continuous speech recognition task, using the most recent release of the recognizer.

INTRODUCTION

CSELT research in the field of automatic speech recognition is presently being carried out along two main directions: a) development of real-word applications based on speech technologies already mature, as Voice Dialling by name [1] and advanced research activities for speech understanding through natural language in man-machine interaction systems with spontaneous dialogue [2].

One key factor in the advancement of speech recognition technology was the shift from Whole Word modeling to sub-word modeling technology. This latter, adds the important feature of Vocabulary Independency to speech recognition systems and paves the way to the wide range of flexible vocabulary applications [3].

Since September 1994 CSELT is internally experimenting a voice dialling-by-name service based on FLEXUS®, a flexible vocabulary recognizer trained in speaker-independent mode for the telephone network. Currently the application supports nearly 1,000 surnames and runs in nearly real time on a Personal Computer equipped with a multichannel DSP board. The clear separation between Language-Dependent and -Independent components, as well as the close modularization of the architecture makes the system easily extensible to other languages.

SYSTEM OVERVIEW

The block diagram of the speech recognizer is depicted in Fig1.

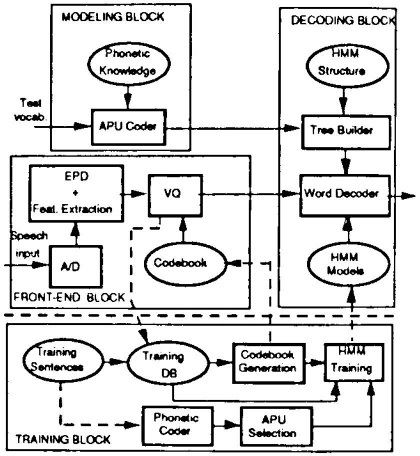

Fig.1 *Block diagram of the speech recognition system*

The training phase is carried out off-line on a SUN Workstation. The **Feature Extraction** module computes spectral features from 10 ms segments (frames) of digitized speech, then a **Vector Quantization** module computes codeword indexes for each input frame with respect to a given codebook. The **Phonetic Coder** transcribes the lexicon extracted from the whole corpus of training in terms of phonemes, then an appropriate set of **Acoustic-Phonetic Units** (APU's) is generated which suitably satisfies two fundamental but competing requirements: *precision* of the model, that is higher with a larger number of units, and *trainability*, that is greater with a reduced number of units. APU's are then associated to Hidden Markov Automata with a given topological structure. The **HMM Training** module learns a stochastic structure for each state of the model (transition and emission probabilities). The on-line decoder consists of a **Lexical Tree Builder** which represents the target lexicon in terms of a tree of sub-word units, of a **real time** version of the Feature Extractor and VQ modules, and of a real time implementation of the **Word Decoder** which computes likelihood scores for word or sentence candidates, sorts them and provides the best one or the N-best ones in output.

LMS

The LMS (Lexical Modeling System) component provides the lexical and acoustic-phonetic knowledge to the recognition system and is divided in modules communicating through simple data structures. Both declarative and procedural knowledge is included. The prominent declarative knowledge is represented by the phonetic coding rules, by the phonetic and APU tables and by coarticulations rules. Procedural knowledge consists of algorithms for sentence generation, lexicon extraction, phonetic rule instantiation, APU generation, graph representation and statistics computations. Phonetic rules and tables, prototypes and semantic classes can be easily substituted for those of a new language, giving LMS a multilingual processing capability. Presently, an English version is being experimented within the RAILTEL EEC-funded project.

Syntactic prototypes

Sentences in a given application domain can be automatically generated from syntactic prototypes represented by graphs. Arcs in a graph are associated to *terminal symbols* or to *semantic classes*. Terminal symbols are single words or sequences of words. Phonetic classes are sets of words associated to a semantic concept: f.i., <TOWNS>, <NUMBERS>, <COMPANIES>. Syntactic forms can also be grouped in classes: *"vorrei"*, *"potrei"*, *"desidererei"* are members of the class <QUERY-START>. Sentences are then generated by visiting the graph along randomly chosen different paths.

Phonetic rules description

The Phonetic Coder includes grapheme-to-phoneme conversion rules for transcription of standard italian and of major regional pronunciations. The regional rules describe the most common linguistic phenomena (assimilation, insertions and deletions) that modify standard Italian when pronounced according to a regional variety [6]. The regional rules have been subdivided into three categories, according to their distribution on the national territory: category 1 includes pronunciations present at least in 5 regions, category 2 describes variants present in three to two regions, and category 3 includes pronunciations typical of one region.

Thirteen regional variants [6] have been selected as representative of the twenty official Italian dialects: Piedmont (PI), Liguria (LI), Lombardy (LO), Veneto (VE), Giulia (GI), Emilia (EM), Sardinia (SA), Tuscany (TU), Umbria-Marche (U-M), Lazio (LA), Apulia (PU), Campania (CA) and Sicily (SI). The regional rules have been written and implemented in the SCYLA language [5], whose syntax is very similar to the generative one; the alphabet used for the formulation of the regional rules includes the following elements:

{ }=	'aut' (exclusive 'or')
C=	any consonant
V=	any vowel
##=	word boundary
+=	morpheme boundary

An example of rewriting rule is the following:

s -> ʃ/ ___ C
[-voiced, +stop]
(SICILIA, Ex. 'aspetto')

The features adopted (e.g. 'voiced', 'stop') define the phones on the basis of phonetic categories (place and mode of articulation), and do not have phonological value. The SCYLA grapheme-to-phoneme transcription component converts an input text string into an output phonetic transcription. Rules for standard Italian [10] have already been written in SCYLA and implemented in a text-to-speech synthesis.

The transcription rules actually convert graphemes into phones and not phonemes: however, this linguistically inexact terminology has been used throughout the paper, because it is conventionally accepted in the field of speech synthesis and recognition.

Linguistic phenomena described by the rules

The rules have been subdivided into three categories: Category 1 rules describe phenomena which occur in 7 regions (usually the northern ones vs. center-southern ones). Linguistic phenomena common to these areas are: a) voicing of intervocalic /s/ (e.g. 'casa', 'house': central-south. ['kasa] vs. north. ['kaza]). b) pronunciation of [s] as a dental affricate [ts], after a nasal and a liquid (e.g. 'arso', 'ansa' and 'salsa': stand. ['arso], ['ansa], ['salsa], regional ['artso], ['antsa] and ['saltsa]); c) pronunciation of velar allophone [ŋ] before consonant in the north of Italy (e. g. 'cantare': standard [kan'tare], regional [kaŋ'tare]); d) insertion of glide [j] after palatal affricate, liquid and nasal consonants [tʃ], [ʃ], [ʎ], [ɲ] if corresponding graphemic groups 'ci , 'sci', 'gli', 'gni' precede a vowel (e.g. 'cielo': standard ['tʃɛlo], regional ['tʃjɛlo]; 'scienza': standard ['ʃɛntsa], regional ['ʃjɛntsa]; 'gliene': standard ['ʎene], regional ['ʎjene]; 'sogniamo': standard [soɲ'ɲamo], regional [so'ɲjamo]).

Category 2 rules are present in 3 to 2 regions and describe the following phenomena: a) voicing of the affricate [ts] before vowel and after [n], in TO, PU, CA (e.g. 'fidanzato': stand. [fidan'tsato], region. [fidan'dzato]; b) doubling of intervocalic [b] and [dʒ] in SI, CA, LA (e.g. 'cabina': stand. [ka'bina], region. [kab'bina]; 'agiato': stand. [a'dʒato], region. [ad'dʒato]; c) progressive assmilation of consonant clusters [tm], [kn], [nr] in TO, U-M, LA (e. g. 'atmosfera': stand. [atmos'fɛra], region. [ammos'fɛra]; 'tecnico': stand. ['tɛkniko], region. ['tɛnniko]; 'Enrico': stand. [en'ri:ko], region. [er'riko]; d) simplification of the intervocalic palatal lateral [ʎ] as the glide [j] in U-M, TO (e.g. 'moglie': stand. ['moʎ:e], region. ['mojje].

Linguistic phenomena present in one region (category 3) are the following: a) palatalization of [s] in EM in every context (e.g. 'sposare': stand. [spo'sare], region. [ʃpo'ʃare]; b) realization of [l] as [r] before consonant in Lazio (e.g. 'alzare': stand. [al'tsare], region. [ar'tsare]).

APU generation

The set of APU's is selected as a viable trade-off between precision of the model and statistical robustness. Acoustic variability can be better captured by differentiating phones with respect to their context: triphones specify a left and a right context, biphones a left or a right context. *(t) r (e)*, f.i., is phone [r] in the word 'treno'.

Monosyllabic functional words (articles, prepositions), which are more prone to heavy coarticulation in continuous speech, are given specific, word-dependent units. The large number of possible triphones in Italian (a few thousands) is downsized by a) imposing a minimum occurrence threshold and b) backing off to biphones.

Graph concatenation

APU's are chained up to form whole words, and words are chained up to form sentences. Therefore, at the end of the concatenation process a single sentence is represented by a graph, which includes pronunciations variants, optional silences and schwa's, and allophonic alternatives. A two step optimization was implemented: first, linear sequences of APU's are obtained for each phonetic variant of every word, then a first optimization is carried out to obtain a word graph; successively, the word graph is chained to the graph corresponding to the sub-sentence already processed, and a second optimization is performed to minimize the number of nodes of the overall graph. The complexity of a sentence graph is ranges from the simplest level (only standard transcription) to the most complex level (all regional variations included).

APU coder

This module codes a given test vocabulary in terms of the APU's alphabet. Linear representations of word are obtained, which will be given in input to the Word Tree Builder to provide a compact tree structure for the on-line Decoder.

HMM topology

Each unit is given a topological structure, based on Bakis' linear model with loop and forward arcs. In the standard three state model, lateral states account for coarticulation effects with adjacent sounds, while the central state captures the stationary component of the phoneme.

Coverage

Statistical tools are included in LMS, to provide general information about the distribution of words, phones, APU's in lexical corpora. The coverage parameters are computed as percentage of occurrence of a given type of APU within a corpus. The coverage parameters are important in evaluating the level of precision of a given model with respect to a training corpus, and the degree of "generalization" of that model with respect to a new application lexicon.

TESTS

Experimental Set up

The speech signal is collected through a linear electret-type telephone connected to a local PBX, filtered in the telephone bandwidth of 300-3400 Kz, and sampled at 8 Khz. 12 Mel-based cepstral and ΔCepstral coefficients are computed each 10 ms time frame, plus Energy and ΔEnergy. A variable threshold, energy-based end-point detector is used. Two 256-codeword codebooks are generated for Cepstrum and ΔCepstrum, plus a 32-codeword one for Energy and DEnergy. Discrete density HMM's were trained through a few Forward-Backward training iterations. A beam-search Viterbi decoder generates the best-scored sequence of words along which we compute the Word Accuracy by alignement with the pronounced sequence [4].

Performance results

Preliminary results were obtained by training the recognizer on a 12,000 utterances data base from 146 speakers, and by testing on 600 utterances from 10 diferent speakers, in the train information query domain, with a vocabulary og 718 words. Utterances were read on a screen in a quiet room, with a linear (electret) microphone, on the local PBX. A Word Accuracy figure of 76.6% has been obtained with Discrete density HMM's, which increased to 79% by using Continuous density HMM's, and to about 90% with statistical language models (word bigrams). Present experimentation aims to improve recognition performance by adopting new types of APU's, including transitional and stationary components, to adapt the recognizer to new languages and to increase the robustness of the models to channel and line variations. The capability of the acoustic-phonetic module to match specific regional pronunciations will be assessed in future experiments.

CONCLUSIONS

A description of the Flexible Vocabulary Recognition Technology developed at Cselt has been given. The most important features of the new release of the LMS component for lexical and phonetical processing have been described. Some recognition tests have been carried out on a speaker-independent, continuous speech recognition task over the telephone, and the corresponding performance figures have been presented. Finally, future developments are mentioned.

REFERENCES

[1] Ciaramella A., Clementino D., Fissore L., Pacifici R., Sperti S.(1993) *"Voice Dialling by name in a PBX environment"*, ESCA Workshop, Bavaria, Germany, pp. 179-182.

[2] Baggia P., Fissore L., Micca G., Rullent C., Laface P.(1994) *"A Speech Understanding System for Information Retrieval"* Int. J. Patter. Rec. and A.I., Vol. 8(1), pp. 71-97.

[3] Lennig M. et al. *"Flexible Vocabulary recognition of speech"* ICSLP'92, Banff (Canada), pp. 93-96.

[4] Fissore L., Laface P., Micca G. (1991)*"Comparison of Discrete and Continuous HMMs in a CSR Task over the Telephone"*, ICASSP '91, Toronto, pp. 253-256.

[5] Nebbia L., Lazzaretto S.(1987) *"SCYLA: Speech Compiler for Your LAnguage"*, Europ. Conf. Speech Techn., Edinburgh.

[6] Canepari L. (1979), *"Introduzione alla fonetica"*; Torino, Einaudi.

[7] Salza P.G.(1990), *"Phonetic Transcription rules for Text-to-Speech Synthesis of Italian"*; Phonetica, (47), pp.66-83.