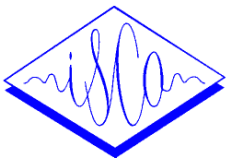

3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

A man-machine dialogue system for speech access to train timetable information

D. Clementino L. Fissore

CSELT - Centro Studi e Laboratori Telecomunicazioni
Via G. Reiss Romoli 274 - 10148 Torino, Italy
Tel.: (+39) 11 2285.111

Abstract

This paper describes the integration of the main components of a man-machine dialogue system for the Italian language, which allows information services to be accessed through the telephone line. The components are: 1) a large vocabulary speaker-independent continuous speech recogniser (AFE), 2) a natural language understanding stage (LP), 3) an intelligent dialogue manager (DM) and 4) a message generator (MG), including a high quality text-to-speech synthesizer. A prototype of an intermediate development state for E-mail access was presented in [1]. In the current implementation, a user can access a train information service through a PBX telephone line. A general overview of the system architecture will be given together with an evaluation of the real-time demonstrator performance through experienced and naive users.

10.21437/Eurospeech.1993-426

1 Introduction

A man-machine dialogue system was developed in order to provide users with an interactive access by telephone to a remote data base.

The previous implementation of the system [1] on E-Mail domain, has been improved as to hardware integration and global performance, making the system more robust to speaker variations. Greater robustness to linguistic variations has been obtained through partial parsing.

The recognition and the synthesis stage are interconnected to the PBX through a telephone interface, whilst the DM stage is connected to a Computer Information System in order to obtain the requested train information from the data base.

The whole system was implemented on a SUN workstation, equipped with some DSP accelerator boards for implementing the AFE and the final stage of the synthesizer. Some alternative real-time implementations of the AFE are available, because it can support both Discrete Density HMMs (DDHMMs) and Continuous Density HMMs (CDHMMs) [11], Viterbi or Forward decoding algorithms with or without grammars, providing different degrees of accuracy and speed, (see below).

*This work has been partially supported by CEC Esprit II project 2218 SUNDIAL "Speech UNderstanding and DIALogue"

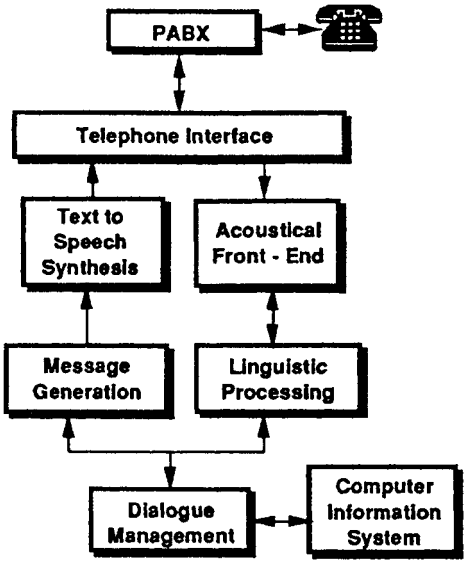

Figure 1: Overall architecture of the system

The AFE and LP are coupled so that the LP finds the most likely sentences from a first AFE recognition output (best sequence and/or lattice). The selected sentences are then sent back to the AFE for a final verification.

The LP uses a robust parsing strategy; the main feature of this stage is to extract the information content of an utterance, using the minimum amount of linguistic knowledge from a lattice of word hypotheses.

The LP finds parts of an utterance from the AFE recognition output; this process requires a further acoustic disambiguation phase. In that instance, the LP sends a set of candidates to the AFE to be acoustically verified.

The DM models the user/system interaction and contextually interprets utterances using a prediction mechanism designed to deal with ill-formed, partially parsed and disrupted input. The DM implements many recovery strategies in order to allow completion of the task, even with errors at the AFE and LP stages. Finally, it generates the answer and sends it to the text-to-speech synthesiser.

The synthesis stage contains specific prosodic rules ori-

ented to the dialogue, which account, e.g., for a number of different types of interrogative sentences [2].

The demonstrator is provided with different tools to collect on-line user dialogues, to manage speech files which give intermediate system output, and to elaborate suitable statistics: this environment is suited both to collect specific spontaneous speech phenomena and to obtain guidelines for further system improvements.

2 The Acoustical Front-End

The acoustical front-end (AFE) performs feature extraction and acoustic-phonetic decoding, providing the best decoded sequence, spanning the whole utterance and the word lattice.

The feature extractor computes Energy and ΔEnergy, 12 MFCC and 12 ΔMFCC parameters in telephone bandwidth, derived from a 256-sample FFT applied at each 10 ms intervals on a speech signal sampled at 8KHz. Three separate codebooks are used in order to perform Vector Quantization when DDHMM are adopted : two for MFCCs and ΔMFCCs (8 bit respectively), and the last one (5 bit) for Energy and ΔEnergy parameters.

The acoustic-phonetic knowledge is represented by a subword speech unit-set, made up of 310 entries, including context-independent phonemes and context-sensitive units, which capture contextual effects from adjacent phonemes. The context-dependent units (function-words, triphones and right-context biphones) were selected according to their occurrence counts in the training set (over than 12000 sentences) [12]. The speech units are represented by means of a HMM (first order, left to right, 3 states without skip); both DDHMM and CDHMM are supported. The training phase was carried out on a speech corpus of more than 12000 sentences from 146 speakers, including an application-dependent (train information domain) set of utterances, which noticeably increased the word accuracy figure.

Extra-linguistic phenomena such as coughs, blows and external noises have been represented by means of specific acoustic models trained with hand-labelled segments extracted from spontaneous speech corpora. These models were considered as additional entries of the lexicon, which is structured in terms of a tree whose nodes are Hidden Markov states. The nodes of the lexical tree are expanded according to a beam-search strategy. The decoding was performed by the Viterbi or Forward algorithm within words and by the Viterbi algorithm between the words. A score normalization was applied both with Viterbi and Forward decoding to "equalise" word lattices for the parsing process [6].
In the E-mail access application, the recogniser performed only an acoustic-phonetic decoding; for the train information service, the AFE can also incorporate linguistic constraints represented by finite states grammars in an efficient time-synchronous decoding process. The recogniser can also switch to the Isolated Word or Spelling modes, according to the DM requests.

3 The Linguistic Processor

The Linguistic Module input can be either the best decoded sequence or a reduced lattice of word hypotheses generated by the recognition module. With respect to the E-mail access application [1], a larger linguistic coverage and a greater robustness was required, especially in order to experiment the real time system with naive users. To this purpose, a new parsing strategy, was designed, based on a multi-step, robust application of our previous parser. This parser, island driven and score guided, was based on the use of Dependency Grammar rules for syntax and Caseframes for semantics; these formalisms were compiled off-line, generating efficient internal structures to obtain real-time operation [7]. A global solution - covering the whole utterance time interval - was not necessarily produced due to robustness reasons. Partial solutions were accepted and just the linguistic knowledge, necessary to create the islands was provided. Each parsing step was performed at two levels: the first one led to the generation of information items by detailed syntactic and semantic knowledge; the second step accepted one or more of these information items, using semantic/pragmatic knowledge. The steps were repeated until no additional information items could be acquired from the utterance (see [8] for a description of the partial parsing strategy).

A feedback verification procedure (FVP) [9] was used in order to eliminate semantic ambiguities caused by a possible unreliable recognition of short function words, (e.g. distinction between the phrase "da Torino" (from Turin) and "a Torino" (to Turin) with preposition lacking in the lattice). At the end of the feedback verification procedure, a final internal semantic structure was generated and given to the Dialogue Module.

4 The Dialogue Manager and Message Generator

The DM aimed to ensure an acceptable interaction quality (eliminating problems at previous levels of analysis) and to interpret users utterances containing anaphoric or elliptical references [5]. The latter goal was achieved by constructing and updating a dialogue history (a context used in the interpretation of referential sentences), throughout the interactions with the users. The dialogue history was based on a realistic model of user-system interaction. From that dialogue context, the DM gets a set of predictions, i.e. pragmatic expectations about probable utterances of the subjects at a given stage of the dialogue. On the basis of these predictions, the DM was able to filter the LP output (when redundant) and to interpret it (when ambiguous). Finally, the DM activates a set of strategies to recover recognition or parsing errors in degraded user interactions. Those repair strategies allowed users to complete their enquiries by resorting to more constrained dialogue and to isolated word recognition strategies. The DM interacted with a template-based generation module which produces a message string as input to the synthesizer.

The text-to-speech synthesis converts the textual message to a phonetic prosodic representation, used by the signal

processing module; the process was executed onto a DSP board, that accesses the diphone dictionary and generates the synthetic speech signal.

5 Real Time Implementation

The whole system deals with a speaker-independent continuous/spontaneous speech over the telephone, using a thousand words vocabulary; the system can switch, for confirmation or recovery of critical cases, to an isolated words strategy, selecting among several vocabularies according to the dialogue evolution. All modules of the system run on a Sun workstation with DSP accelerator boards - used by the AFE module - added to its internal VME bus. These boards are based upon 2 floating-point TI TMS320C30s. One DSP board was used with an analog module for the A/D and D/A conversion: the feature extraction from the input speech signal and the synthetic speech generation are executed onto this board. Another board was used, for DDHMM, to execute the decoding algorithm for the speech recognition; both Viterbi and Forward algorithms were implemented. Three other DSP boards had to be added for the likelihood computation in the CDHMM, which is very time consuming [10]. Speech recognition can also be performed by including grammatical constraints, i.e. Finite-State Networks (FSN), which improve the recognition accuracy but increase the computational requirements. Three other DSP boards needed to be added [4]: therefore real-time performance was achieved by properly distributing the speech recognition task in the multi-DSP architecture.

A shared memory area was used for exchanging commands and data between the processing modules of the system.

An external telephone interface connected the acoustical front-end and the speech synthesis module to the PBX, detecting the telephone calls t and managing the user disconnection.

6 Experimental Results

Two sets of experiments were carried on :
a) a laboratory test (*lab*) performed on a simplified system configuration (AFE + LP) in order to define the best configuration for the execution time and the accuracy;
b) experiments on naive-users and experienced-users using the integrated system.
The *lab*test-set consisted of 600 sentences uttered through the internal PBX by 10 different speakers, with a 718 words vocabulary.
The recognition with CDHMM performed in 2.7 times real-time, whilst the DDHMM scored in 1.7 times real-time. The parsing speed of the recognition output did not show any direct relation to the models used in the recognition, because it did not involve any acoustic processing. The average time for parsing the best decoded sequence (BS modality) was 0.5 sec; when the lattice of word hypotheses was added to the best decoded sequence (BS+L modality), the average parsing time was 1 sec.

The results in Table 1 show that at AFE level, (using the Forward decoding), both sentence accuracy (*fep-sa*) and word accuracy (*fep-wa*) were slightly better for CDHMM than for DDHMM experiments [11]. At LP level [8], the sentence understanding (*lp-su*) and the concept accuracy (*lp-ca*) show that the BS+L modality gave better results than the BS one, especially with DDHMM; this was due to the fact that the best sequence was more often correct in the CDHMM instance. The FVP was always included.
Further experiments with different recognition algorithms showed that the Viterbi decoding was more efficient than Forward (1.3 times real-time), but worse at the LP level with the BS+L modality.

Experiment		fep-sa	fep-wa	lp-su	lp-ca
CDHMM	BS	26.3%	74.7%	70.5%	73.5%
	BS+L	26.3%	74.7%	73.5%	76.1%
DDHMM	BS	23.0%	73.4%	67.5%	67.9%
	BS+L	23.0%	73.4%	73.3%	74.7%

Table 1: Comparison of CDHMM and DDHMM

Experiment		fep-sa	fep-wa	lp-su	lp-ca
CDHMM	no FVP	26.3%	74.7%	67.2%	70.0%
	with FVP	26.3%	74.7%	73.5%	76.1%
DDHMM	no FVP	23.0%	73.4%	61.3%	62.3%
	with FVP	23.0%	73.4%	73.3%	74.7%

Table 2: Evaluation of the FVP

Table 2 shows the improvements obtained by the FVP, due to the recovery of the short function words missing from the lattice or from the best decoded sequence; the FVP contribution was greater with DDHMM, because the short words were normally less accurately identified.
As a trade-off between the performance and the computational requirements, a system configuration was defined for the experiments shown in Table 3. The real-time prototype executes the Forward decoding algorithm, at the AFE level, on DDHMM of subword speech units without linguistic constraints. The prototype also includes the FVP and the BS modality at the LP level.
The naive-user test-set consisted of 100 dialogues for a total of 678 sentences from 20 different users (10 males and 10 females). The speakers interacted spontaneously with the system. The experienced-user corpus consisted of 66 dialogues, for a total of 464 sentences from 15 different speakers (11 males and 4 females). These speakers were familiar with the speech technology. Table 3 shows the accuracy results for both cases.

Experiment	fep-sa	fep-wa	lp-su	lp-ca
naive-users	32.7%	52.1%	50.9%	35.7%
experienced-users	24.1%	60.2%	59.1%	52.6%

Table 3: Results with spontaneous speech

The DM provided an important contribution in driving the users to the achievements of their goals; the Transac-

Experiment		fep-sa	fep-wa	lp-su	lp-ca
naive-users	G_1	39.4%	58.3%	54.9%	40.9%
	G_2	46.2%	64.2%	62.2%	51.6%
exper-users	G_1	57.1%	78.5%	74.6%	69.1%
	G_2	56.2%	78.4%	73.7%	66.4%

Table 4: AFE integrating FSN grammars

tion Success rate (i.e. a measure of the system proficiency in providing users with the required information [5]) was good for both the naive-users (77.6%) and the experienced-users trails (96.6%). The better performance obtained with experienced-users was due to an improvement of the *lp-ca*and to a more efficient use of the DM repair strategies. Furthermore, off-line experiments have been performed on the *lab*corpus by integrating grammatical constraints, with a BS modality and without FVP, within the AFE. The FVP turned out not to be essential, because implicitly included in the use of the grammar. Although the computational requirements increased, the decoding accuracy was significantly improved, as shown in Table 4. The first FSN (G_1) was developed taking into account only the 600 read sentences of the *lab*corpus; its perplexity score was 195. The second FSN (G_2) also considered spontaneous sentences, resulting in a perplexity score of 285 [4]. The test sentences derived from the naive users and the experienced users were not overlapping with the corpus used to define the grammars.

Acknowledgments

The authors wish to thank all the project team : P. Baggia, M. Balestri, R. Billi, A. Ciaramella, M. Danieli, E. Gerbino, E. Giachin, G. Micca, L. Nebbia, R. Pacifici, C. Rullent.

7 Conclusions

The integration of several speech technologies into a man-machine dialogue system for the Italian language on the train-timetable domain has been presented. First results with real users have been provided. A degradation of performance was observed moving from read sentences to spontaneous speech. However, the availability of reliable repair strategies in DM, as well as the robust parsing in LP, allowed to get an acceptable transaction rate in the real-time prototype.

Further experiments with naive users are currently underway in order to increase robustness of acoustic-phonetic units set and collect a spontaneous speech corpus.

References

[1] P. Baggia, A. Ciaramella, D. Clementino, L. Fissore, E. Gerbino, E. Giachin, G. Micca, L. Nebbia, R. Pacifici, G. Pirani, C. Rullent, "A man-machine dialogue system for speech access to E-mail information using the telephone: implementation and first results", Eurospeech '91, Genova, Italy, 1991

[2] S. Quazza, P. Salza, S. Sandri, A. Spini, "Prosodic control of a text-to-speech system for Italian", submitted to ESCA Workshop on Prosody, Lund (september 1993)

[3] A. Ciaramella, D. Clementino, R. Pacifici, "Real-time speaker-independent large-vocabulary CDHMM-based continuous telephonic speech recogniser", ICSLP '92, Banff, Canada, pp.89-92.

[4] D. Clementino, E. Giachin, "Real-time continuous speech recognition integrating extensive grammars for spoken language", ICSPAT '93, Santa Clara, CA (USA)

[5] E.Gerbino and M.Danieli, "Managing Dialogue in a Continuous Speech Understanding System", Eurospeech 93, Berlin.

[6] L. Fissore, P. Laface, G. Micca, R. Pieraccini, "A Word Hypothesiser for a Continuous Speech Recogniser", ICASSP '89, Glasgow, pp. 453-456, Scotland, 1989

[7] P. Baggia, E. Gerbino, E. Giachin, C. Rullent, "Efficient Representation of Linguistic Knowledge in Continuous Speech Understanding", Proc. IJCAI 91, pp. 979-984, Sidney, 1991.

[8] P. Baggia, C. Rullent, "Partial parsing as a Robust Parsing Strategy", ICASSP '93, Minneapolis (USA), 1993

[9] P. Baggia, L. Fissore, E. Gerbino, E. Giachin, C. Rullent, "Improving speech understanding performance through feedback verification", Speech Communication 11, n. 2-3, June 1992.

[10] A. Ciaramella, D. Clementino, R. Pacifici, "Real-time speaker-independent large-vocabulary CDHMM-based continuous telephonic speech recogniser", ICSLP '92, Banff, Canada, pp.89-92.

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

[12] L. Fissore, E. Giachin, P. Laface, G. Micca, "Selection of Speech Units for a Speaker-Independent CSR Task, Proc. Eurospeech '91, Genova, Italy, 1991