USE OF PARALLEL RECOGNIZERS FOR ROBUST IN-CAR SPEECH INTERACTION

Luca Cristoforetti, Marco Matassoni, Maurizio Omologo, Piergiorgio Svaizer

ITC-irst (Centro per la Ricerca Scientifica e Tecnologica) I-38050 Povo - Trento (Italy)

$[cristofo, matasso, omologo, svaizer]$@$itc.it$

ABSTRACT

This paper refers to an activity under way at the speech recognition technology level for the development of a hands-free dialogue interaction system in the car environment.

The use of a set of HMM recognizers, running in parallel, is being investigated in order to ensure low complexity, modularity, fast response, and to allow a real-time reconfiguration of the language models and grammars according to the policy indicated by natural language understanding and dialogue manager modules.

A corpus of spontaneous speech interactions was collected using the Wizard-of-Oz method in a real driving situation with a microphone placed far from the driver. The use of parallel recognition units, each specialized on a given geographical domain, was explored using the resulting real corpus. Experiments show the advantage of selecting the recognized sentence according to the maximum likelihood among the active units when compared to the use of a single language model based on a very large vocabulary.

1. INTRODUCTION

In the recent years, continuous speech recognition technology for car environment has progressively been improved in terms of robustness and flexibility with regard to the applications. At the same time, a new market demand emerged for systems allowing the driver to control devices like RDS-tuner or air-conditioner by voice, with increased reliability and safety. As a result, in the next future more complex applications like satellite navigation systems, remote information services access or Web browsing are likely to be practicable in a full hands-free modality.

To this purpose, hands-free interaction in such a challenging scenario still needs to be improved to make it more robust under variable conditions [1]. In fact, security and convenience of hands-free interaction require a non-intrusive placement of the microphone, that cannot be situated close to the driver's mouth. As a consequence the input signal is inevitably characterized by a low SNR [2]. Some noise components are unpredictable and in general unstationary (e.g. road bumps, rain, traffic noise, etc.). In addition, acoustic effects of the car enclosure and speaking style modifications (i.e. Lombard effect) contribute to lower the speech recognizer performance, especially in conjunction with the word confusability induced by large vocabularies.

The European project VICO (Virtual Intelligent CO-driver) [3], to which ITC-irst contributes, is aimed at developing an in-car system for accessing tourist information databases and obtaining driving assistance. The system is based on a complex vocal interaction between the driver and the on-board computer, often characterized by the use of words belonging to large lists.

Voice interaction can be in English, German or Italian. ITC-irst has in charge the development of the ASR engine for the Italian language, while Daimler Chrysler AG [4] will develop the corresponding engines for English and German. It is worth noting that the recognition of foreign words makes the task harder, as the typical experimental context regards a German or an English tourist who is visiting the Italian Trentino area.

In order to facilitate the development of the Natural Language Understanding module, a common interface was specified and all ASR engines are integrated into a CORBA system architecture. Due to the need of alignment among language models for the different languages as well as to the need of reducing the complexity while managing large vocabularies (e.g. streets and points of interest in a city, cities in a region, etc.), a framework was realized, based on the concept of several speech recognition units that run in parallel and use class-based statistical language models or grammars. This kind of approach has been already explored in different frameworks [5].

The objective of this paper is that of investigating on a simple selection method to choose among the output given by a set of recognition units. The work refers to the use of real spontaneous speech utterances as test corpus.

The paper is organized as follows: a first section introduces the general system architecture and the most relevant tasks which are being addressed; the front-end processing and the HMM engine are then described with more details on the set of recognition units presently used; a description of the test database collected through Wizard-of-Oz (WOZ) experiments is provided together with some preliminary recognition results. In the final section, we will draw some conclusions and describe the future work that is planned.

2. SYSTEM ARCHITECTURE

The general architecture of the system is shown in Figure 1. The early stage consists in a front-end processing based on robust speech activity detection, noise reduction and feature extraction modules, and in a recognition engine, conceived as a set of single recognition units in parallel and an output selection module. The latter one has the role of providing an output to the Natural Language Understanding (NLU) module, which is more reliable than what would be produced by using a single comprehensive Language Model (LM) and a related very large vocabulary. As shown in the Figure, we assume that the Dialogue Manager (DM) module decides which units have to be active for each dialogue step (i.e. recognition process). If necessary DM loads a new LM in a Speech Recognition Unit (SRU) and runs it as a second pass on the given input utterance, if no one of the selected SRUs has provided

This work was partially funded by the Commission of the EC, Information Society Technologies (IST), 2000-25426, under VICO.

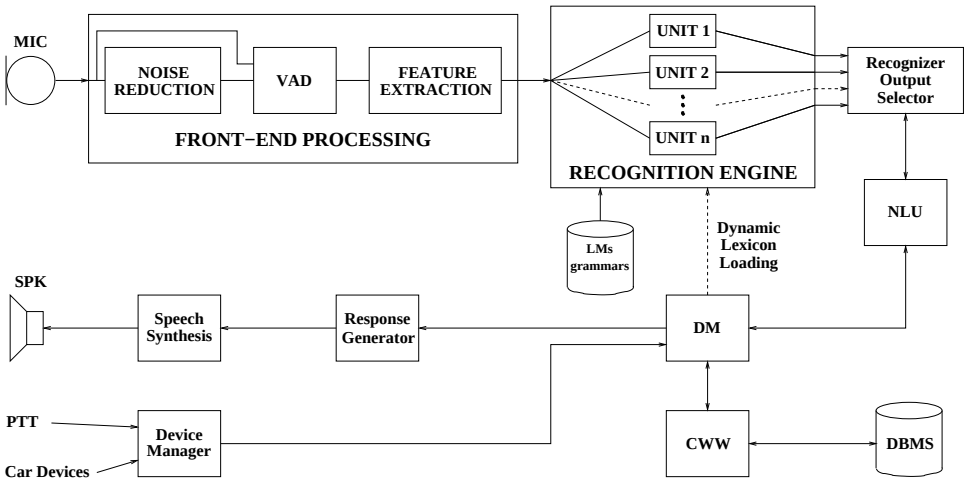

Fig. 1. System Architecture

a reliable output. Note that the SRUs, once loaded, can be selected to be running at the same time, which means that a user utterance is being processed in parallel by all active SRUs in a very efficient manner, this way avoiding the delay that would be introduced by any equivalent sequential recognition approach. The diagram also shows the other modules of the VICO system. It is worth noting that the Car Wide Web (CWW) module is an interface (also realized at ITC-irst) to a set of databases using an XML protocol to communicate with DM. Presently, it allows a fast access to a tourist database that includes most of the relevant information about Trentino (produced by Azienda per la Promozione Turistica del Trentino) and to a geographical-topographical database (produced by TeleAtlas) for any query concerning navigation.

2.1. Front-end Processing

2.1.1. Acoustic Front-end

The present front-end processing is a traditional one. The feature extraction module processes the input signal pre-emphasizing and blocking it into frames of 20 ms duration from which 12 Mel scaled Cepstral Coefficients (MCCs) and the log-energy are extracted. MCCs are normalized by subtracting the current MCC means. The log-energy is also normalized with respect to the current maximum energy value. The resulting MCCs and the normalized log-energy, together with their first and second order time derivatives, are arranged into a single observation vector of 39 components. Note that a real speech activity detection module is included in the front-end processing. It is based on the energy information in the case of close-talk input and on a spectral variation function technique applied to the output of the Mel-based filter bank in the case of far-microphone signal. According to preliminary experiments on SpeechDatCar material [6], both techniques allow recognition performance equivalent to that determined by using manually segmented utterances, except for cases of unstationary noise events.

2.1.2. Noise Reduction

Optionally, a noise subtraction module can be added to the above described processing for far-microphone input processing.

Basic noise reduction algorithms are an easy and effective way to reduce mismatch between noisy conditions and clean HMMs, and can also be used with some benefits in matched conditions, as was shown in [7]. On the basis of that work, magnitude spectral subtraction and log-MMSE estimation were adopted for background noise reduction, together with quantile noise estimation.

In [8] an optimal set of parameters was determined for the use of spectral subtraction and log-MMSE on a connected digit recognition task; the same set is used here.

2.2. Recognition Engine

The recognition engine is as a set of standard HMM recognition units, running independently.

2.2.1. HMM

For complexity reasons as well as for availability of training real material at this moment each HMM recognition unit is based on a set of 34 phone-like speech units. During the second phase of the project context-dependent and word-based HMMs will be explored also using the contaminated speech material based technique discussed in [9].

Each acoustic-phonetic unit is modeled with left-to-right Continuous Density HMMs with output probability distributions represented by means of mixtures having 16 Gaussian components with diagonal covariance matrices. HMM training is accomplished through the standard Baum-Welch training procedure. Phone units were trained by using far-microphone (and close-talk) signals available in the Italian portion of SpeechDatCar corpus. The training portion of this corpus consists in about 3000 phonetically rich sentences pronounced by 150 speakers.

2.2.2. Parallel Recognition Units

The SRUs can be viewed as imperfect noisy decoders with adjustable parameters to give more weight to some grammars, language models, or word lists according to the policy and the beliefs of the Natural Language Understanding and the Dialogue Manager modules.

A crucial aspect is the selection of the output to feed NLU. For a given input utterance, the outputs provided by the different active units have to be compared each other in a reliable way. A specific work is being conducted with the scope of using confidence measures at sentence or word level and, in a next phase, this will be extended to the use of word graph hypotheses.

At the moment, the simplest decision policy is adopted, which is based on the selection of the output having the maximum likelihood. Although this is the most immediate approach, it allows to understand if the parallel recognition unit based approach may offer some advantages in terms of performance and complexity, besides the advantage in flexibility as pointed out above.

3. TASKS

The present VICO architecture is based on parallel recognizers covering distinct domains or geographical clusters. In a first stage, the following subdivision and related grammar/language model development was defined: navigation and tourism information in Trentino (SR0); navigation in Alto-Adige (SR1); hotel and restaurant booking in Trentino (SR2); confirmation, refusal and other general commands (SR3); handling the big list of cities and places of Trentino (SR4); handling the big list of streets of Trentino (SR5); handling the spelling mode (SR6).

Due to the characteristics of the presently available test database, in this work we refer to a subset of recognition units defined as follows:

- SR_ge: navigation, tourist information, hotel and restaurant reservation, without any geographical reference or constraint (dictionary size equal to 2000 words)
- SR_c1: same as SR_ge with class contents related to the city of Trento (dictionary size equal to 3000 words)
- SR_cc1: same as SR_ge with class contents related to Trentino except for the city of Trento (dictionary size equal to 11000 words)
- SR_cmd(= SR3): confirmation, refusal and other general commands (dictionary size equal to 40 words)

This alternative set was conceived to deal with the situation of a tourist who is visiting the city of Trento and has the need to navigate or ask information about either the city or the surrounding region. Note that Trentino comprises a very large number of cities, streets and Points Of Interest (POIs). In particular, there are about 5000 streets and 2000 POIs, often having quite similar names. As a result, SR_cc1 represents the most complex and ambiguous domains. A reference model SR_cg was also introduced with class contents related to the whole Trentino area (SR_cg = SR0 + SR2) with a dictionary size equal to 12500 words.

An effort was devoted to the statistical language models, basically because of the relative lack of appropriate training material. As a first step, a corpus of handwritten questions and requests and another corpus of texts regarding the tourist domain were employed to train a class-based statistical language model (SR_ge) using trigram statistics. The classes are kept empty as the corresponding SRU has to deal with generic queries, in principle not depending on any geographical area constraint. The other LMs (SR_cg, SR_c1, SR_cc1), were derived from the latter one, by filling each class with the list of words pertaining the corresponding domain (Trentino, city of Trento, Trentino except the city of Trento). Each language model is based on four classes which include lists of cities, streets, hotels and POIs. The SR_cmd unit is based on a regular grammar, as it deals with a restricted domain of typical expressions of confirmation and refusal and it is based on a very small vocabulary.

4. WOZ DATABASE

An Italian WOZ-based data collection was organized in order to reconstruct a real interaction between the driver and the VICO system for tasks as “reach a POI in Trento city”, “ask for hotel/restaurant information and book a room or reserve a table”, “ask information about the car”, “ask information about a museum”, etc., all of them covered by the SRUs introduced in the previous sections.

During recordings, a co-driver was always in the car to describe each goal the driver had to pursue by voice interacting with the system. The wizard was at ITC-irst labs, connected to the mobile phone of the car. A specific setup was designed in order to simulate an interaction as realistic as possible and to allow a synchronous speech acquisition through two input channels, one connected to a close-talk head-mounted microphone and the other to a far-microphone placed on the ceiling. The audio prompts were produced by using a commercial text to speech synthesizer.

The resulting speech material and related transcriptions may be useful both to train language models and to reinforce the acoustic models with typical spontaneous speech expressions. At this moment, they are used only for test purposes. The present release includes 16 speakers (8 M + 8 F), that uttered a total of 1612 spontaneous speech utterances (equivalent to 9150 word occurrences). The total speech corpus duration is 132 minutes (mean duration of utterance is 4.9 s) and the total dictionary size is 918 words.

Note that all of the speakers were naive to the use of this type of systems and that the wizard behaviour was based on an interaction model, previously defined, that comprised the simulation of recognition errors typical of the foreseen real scenario. As a result, many sentences include typical spontaneous speech problems (e.g. hesitations, repetitions, false starts, wrong pronounciations, etc.) and often consists in many words (in a few cases the input utterance contained more than 25 words). The realism of the experiment is also shown by the fact that at the end of the experiment, after more than one hour, all the speakers declared they were not aware of the fact that a human was interacting with them.

5. SYSTEM PERFORMANCE

The first row of Table 1 summarizes the statistics for each recognition unit, in terms of number of WOZ test sentences that were associated to that unit. Note that labeling of WOZ test data was done manually and that some of the sentences might be associated to more units (i.e. the input sentence may be recognized correctly by more than one unit).

As seen in the previous sections, this work was conceived to investigate on the convenience of using sentence likelihood (ML) in order to select the most reliable recognition unit. To this purpose, recognition experiments were firstly conducted on the entire WOZ corpus by using each of the five recognition units. From the second to the fifth row of Table 1 one can see the word recognition rate of each recognition unit (SR_c1, SR_cc1, SR_ge, SR_cmd) if applied to the portion of the test material labeled as matching domain and (in the last column) if applied to the whole test corpus, without taking care of out-of-vocabulary words.

	c1	*cc1*	*ge*	*cmd*	*tot.*
N. sent.	239	155	823	395	1612
SR_c1	63.7	-	-	-	62.6
SR_cc1	-	58.0	-	-	54.5
SR_ge	-	-	67.3	-	58.8
SR_cmd	-	-	-	89.7	7.9
SR_cg	55.6	58.6	59.8	55.4	58.5
ML	61.3	59.6	61.8	73.1	62.2

Table 1. Word Recognition Rates for close-talk microphone input. Each column corresponds to the use of a portion of the test corpus associated to a given domain. The first row reports on the number of sentences for each portion.

	c1	*cc1*	*ge*	*cmd*	*tot.*
SR_c1	59.9	-	-	-	56.7
SR_cc1	-	47.9	-	-	42.5
SR_ge	-	-	59.3	-	52.9
SR_cmd	-	-	-	86.5	7.3
SR_cg	47.8	48.8	46.0	31.6	45.7
ML	49.0	48.4	47.1	53.7	48.1

Table 2. Word Recognition Rates for far-microphone input.

The sixth row gives the performance obtained using the general language model SR_cg (covering all the domains). From these results one can see the potential advantage in using specialized recognition units. However, given an input utterance, one has to decide which recognition is the most reliable. The last row (ML) of the table provides the recognition rate obtained when the output is selected from the recognition units according to the maximum among sentence likelihoods. In this way, an improvement from 58.5% to 62.2% was observed.

The above results refer to the use of close-talk microphone input. A similar trend, with a word recognition rate of about 48%, was observed in the case of far-microphone input (see Table 2).

Finally, Table 3 shows the String Recognition Rate obtained from the same experiments. Broadly speaking, one can observe that one out of three sentences (generally, shortest ones and commands) are correctly recognized, which is an encouraging behavior for a complex voice interaction scenario as that being investigated. Again, using ML method allows to obtain a better performance than using the general language model SR_cg. This fact shows the convenience of using parallel recognition units in the given experimental framework.

6. FUTURE WORK AND CONCLUSIONS

This work represents a preliminary step in the development of a dialogue system for in-car voice interaction with advanced services of navigation assistance and tourist information access.

Next steps of this work will include the development of more robust solutions for what concerns the preprocessing and feature extraction stage as well as the acoustic modeling, presently based on context independent HMMs. On the other hand system complexity in the given framework is a crucial aspect and only a limited increase in complexity will be possible at the acoustic processing level. Hence a deeper comprehension of the most suitable strategies of the NLU and DM modules is felt necessary to fully exploit the potentiality provided by the parallel unit architecture of the recognition engine.

	$Close-talk$				
	$c1$	$cc1$	ge	cmd	$tot.$
SR_cg	19.3	16.1	24.3	51.4	29.4
ML	20.9	14.8	24.9	72.4	35.0
	Far				
	$c1$	$cc1$	ge	cmd	$tot.$
SR_cg	16.3	8.4	19.1	38.0	22.3
ML	18.4	9.0	19.0	65.1	29.2

Table 3. String Recognition Rates for both close-talk and far-microphone input.

As a first step, the application of a maximum likelihood criterion to select the recognition output represents the simplest choice, and work is under way for what regards the selection of more reliable outputs from these recognition units, on the basis of confidence measures and word graph hypotheses.

7. REFERENCES

[1] *Proceedings of the Hands-Free Speech Communication Workshop* (HSC), Kyoto (Japan), 2001.

[2] M. Omologo, P. Svaizer, M. Matassoni, "Environmental conditions and acoustic transduction in hands-free speech recognition", *Speech Communication*, vol.25, pp. 75-95, 1998.

[3] P. Geutner, F. Steffens, D. Manstetten, "Design of the VICO spoken dialogue system: evaluation of user expectations by Wizard-of-Oz experiments", *Proc. of LREC*, Las Palmas (Spain), 2002.

[4] H. Hüning, A. Berton, U. Haiber, F. Class, "Speech Recognition Methods and their Potential for Dialogue Systems in Mobile Environments", ISCA Workshop, Kloster Irsee (Germany), June 2002.

[5] H. Schwenk, J.L. Gauvain, "Combining Multiple Speech Recognizers using Voting and Language Model Information", *Proc. of ICSLP*, pp. 915-918, Beijing (China), 2000.

[6] A. Moreno, B. Lindberg, C. Draxler, G. Richard, K. Choukri, S. Euler, J. Allen, "A Large Speech Database for Automotive Environments", *Proc. of LREC*, Athens (Greece), 2000.

[7] M. Matassoni, G. A. Mian, M. Omologo, A. Santarelli, P. Svaizer, "Some experiments on the use of one-channel noise reduction techniques with the Italian SpeechDat Car database", *Proc. of ASRU*, Madonna di Campiglio (Italy), 2001.

[8] M. Matassoni, M. Omologo, A. Santarelli, P. Svaizer, "On the joint use of noise reduction and MLLR adaptation for in-car hands-free speech recognition", *Proc. of ICASSP*, Orlando (FL), 2002.

[9] M. Matassoni, M. Omologo, D. Giuliani, P. Svaizer, "HMM-Training with Contaminated Speech Material for Distant-talking Speech Recognition", *Computer Speech and Language*, 16, pp. 205-223, 2002.