USE OF REAL AND CONTAMINATED SPEECH FOR TRAINING OF A HANDS-FREE IN-CAR SPEECH RECOGNIZER

M. Matassoni, M. Omologo and P. Svaizer

ITC-irst - Centro per la Ricerca Scientifica e Tecnologica
I-38050 Povo - Trento (Italy)
$[matasso, omologo, svaizer]@itc.it$

Abstract

A database of in-car speech for the Italian language was collected under the European projects SpeechDatCar and VODIS II. It consists of 600 sessions recorded under various noise and driving conditions and includes close-talk signals and far microphone signals for hands-free interaction.

This paper describes some recognition experiments on two tasks conceived on a portion of this database: connected digit sequences and isolated command words. Recognition rate achieved by means of HMMs trained on real in-car speech is compared with that accomplished by a speech contamination approach, which aims at simulating in-car data starting from a clean speech corpus.

Recognition performance is also analyzed as a function of the different noise conditions and of the consequent SNR at the far microphones. Finally, the effect of HMM adaptation is investigated in order to tune the recognizer on the conditions of the various sessions.

1. Introduction

Reliable hands-free speech interaction inside the car is still a challenging scenario. An essential requirement is robustness of speech recognition against the various kinds of noise typical of the car environment.

Several new applications in this context are envisaged in the next future, allowing the driver to control by voice devices such as RDS-tuner, CD and cassette player, air conditioner, etc. Also more complex interactions like mobile telephone dialing and access to a navigation system or to remote information services [1] will be practicable in a full hands-free modality, with increased flexibility and safety for the driver who can concentrate his attention on the road.

Security and convenience of hands-free interaction require that the microphone must not encumber the user and therefore can not be put close to his/her mouth. As a consequence the input signal is characterized by a low SNR, being affected by several noise components [2]. Engine and tyres contribute mainly low frequency noise, while aerodynamic turbulence, predominant at high speed, has a broader spectral content [3]. Moreover, other much more unpredictable noise events (road bumps, rain, traffic noise...) characterize the car environment.

As a result, car speech recognition is a notably hard task, due to the resulting disturbance, mainly additive, generally non-stationary and almost incoherent, together with the low SNR, the car-enclosure acoustic effect and the Lombard speech effect. In this context having large corpora of data acquired on the field and representative of the various situations is a fundamental starting point for the development and assessment of an applicative technology.

The target of the European projects SpeechDatCar and VODIS II[1] was the collection of speech databases in the car environment, with homogeneous characteristics in 9 different languages[2], to develop robust multi-lingual applications [4]. Under these projects we collected an Italian database consisting of 300 speakers ($\times$ 125 items $\times$ 2 driving conditions) which can be used to investigate many applicative aspects: from the study of a speech recognizer under different environmental conditions, to the influence of hands-free interaction with different microphone types and positions, to the impact of a GSM channel, to scenarios based on Distributed Speech Recognition (DSR), etc.

This paper addresses some of these aspects, presenting results of speech recognition on two standard tasks: connected digits and isolated command words. Performance obtained with HMMs trained on real data is compared with that achievable when training HMMs on data artificially produced by contaminating a clean speech corpus [5]. The effect of batch adaptation is also examined.

In the following we describe the speech corpus collected in the car environment, the hands-free speech recognition system under development and the early stage of experiments performed on a portion of the whole database.

2. The speech corpus

The SpeechDatCar/VODIS II Italian corpus consists of 600 sessions (300 speakers $\times$ 2 sessions) recorded under one of the following conditions: car stopped with motor running (*stop*), driving in the town-traffic (*town*), driving at low speed on rough road (*low*), driving at high speed on good road (*high*). The recordings are made either with or without additional environmental noise due to air-conditioning, open windows, etc (*noisy*) and with or without the car radio on (*radio*). Table 1 reports on the distribution of the acquired sessions according to the noise conditions as well as on the distribution of the 300 speakers according to their geographic origin and their age.

A session consists of 125 items, including isolated words, spelled words, connected digit sequences, phonetically rich utterances, continuous speech, etc. Recordings were accom-

[1]The collection of the Italian database under SpeechDatCar and Vodis II projects was partially funded by the Commission of the EC, Telematics Applications Programme, Language Engineering, Contracts LE4-8334 and LE4-8336.

[2]For more details on the whole database design and collection see the website http://www.speechdat.org/SP-CAR/

plished by using a PC equipped with a set of preamplifiers and a multichannel acquisition board. The inputs included a SHURE SM10A close-talk and three far electret condenser microphones, namely: a AKG Q400Mk3T placed near the A-pillar (*Mic1*), a Peiker ME15/V520-1 placed in front of the driver behind the sunvisor (*Mic2*), another AKG microphone placed over the midconsole near the rear-view mirror (*Mic3*).

$Condition$ $(600 sessions)$	$stop$	62
	$town$	85
	$town + noisy$	101
	low	103
	$low + noisy$	100
	$high$	89
	$high + radio$	60
$Region$ $(300 speakers)$	$north - east$	141
	$north - west$	52
	$center$	50
	$south$	57
Age $(300 speakers)$	$18 - 30$	134
	$31 - 45$	117
	$46 - 60$	46
	> 60	3

Table 1: *Characteristics of the database: distribution of the sessions according to the noise conditions and partitioning of the speakers according to geographic origin and age.*

For all of these input channels the recordings were realized with 16 kHz sampling frequency and 16 bit accuracy. At the same time an additional AKG far-microphone, connected to a mobile telephone equipment, allowed remote recording (at 8 kHz/8bit A-law compression) of speech signals transmitted through the GSM telephone network.

The acquired data were then annotated channel by channel (close-talk, three far microphones, GSM channel). Each item was documented by means of specific labels to detail what the speaker really uttered and to account for the various background noise as well as for the acoustic events occurred in the recording. This operation was carried out by means of a specific software tool (*JavaSgram* [6]), conceived for the annotation of multichannel corpora. Thanks to this software and to the application of a segmentation tool, utterance boundaries were derived automatically and then checked manually. Figure 1 shows an example of noisy signals acquired from a close-talk and a far microphone.

2.1. Recognition tasks

As annotation and validation were still in progress, for the experiments described in the following only 400 out of 600 session were used. In particular a set of 100 speakers (60 males and 40 females) was considered as training set to produce phone HMMs, based on a total of 2410 phonetically rich utterances. Other 100 speakers (59 males, 41 females) were designated as test set. Thanks to the variety of speech material available, several recognition tasks may be conceived: in this work two of them are considered, namely connected digit and isolated command recognition. The first one concerns connected digit sequences of unknown length (*cdigits*) and consists of 1091 sequences, for a total of 9946 digits. The second task (*Vodis81*) involves utterances from a vocabulary of 81 isolated commands

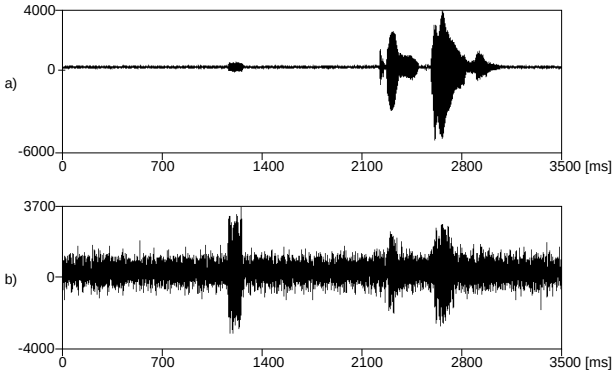

Figure 1: *Example of an utterance acquired in the noisy car environment with the close-talk (a) and with a far-talk microphone (b), respectively.*

(specifically designed under VODIS project), aimed at the control of in-car devices, and consists of 6832 utterances.

2.2. Signal to noise ratio

Variable noise level, driving conditions and speaker characteristics induce heterogeneous Signal to Noise Ratios (SNRs) in the utterances acquired by the microphones. Table 2 reports on the average SNRs computed on the training set and on the test sets acquired from the close-talk and the three far microphones.

Here SNR is calculated as $10 \log_{10}((P_s - P_n)/P_n)$, where P_s and P_n represent the average power of the speech segment and the average power of the preceding and following background noise segments, respectively. In the following this aspect will be better detailed with regard to SNR distribution among different utterances.

	ClTalk	*Mic1*	*Mic2*	*Mic3*
Training	25.9	8.5	9.3	10.0
test cdigits	25.9	8.2	9.2	9.9
test Vodis81	25.7	8.4	8.9	9.7

Table 2: *Average SNR (in dB) measured on training and test sets.*

3. System description

The in-car hands-free recognition system being developed at ITC-irst consists of the acquisition system used for the database collection, a feature extraction module and a Hidden Markov Model (HMM)-based recognizer.

3.1. Feature extraction

The feature extraction module processes the input signal preemphasizing it and blocking it into frames of 20 ms duration (with 50% frame overlapping). For each frame, 8 Mel scaled Cepstral Coefficients (MCCs) and the log-energy are extracted. MCCs are normalized by subtracting the MCC means computed on the whole utterance. The log-energy is also normalized with respect to the maximum value in the utterance. 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 27 components.

Note that here end-point detection is not considered, as manually segmented speech items were used both in training and in test. Nevertheless this is a critical issue for the development of real application systems, and is being investigated at our labs.

3.2. Recognition System

The HMM module is based on a set of 34 phone-like speech units. 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.

3.2.1. HMM training

HMM training was accomplished through the standard Baum-Welch training procedure and was carried out exploiting the 2410 phonetically rich sentences of the training set.

For comparison purposes, another set of HMMs was trained on data artificially derived [7] from a clean corpus [8] to simulate the car environment. The effect of additive noise was accounted for by summing clean speech data and real noise sequences recorded inside a car (different from that of database collection), with properly scaled amplitudes to reproduce various SNRs in the range $0 \div 12$ dB [5].

3.2.2. HMM adaptation

HMM adaptation is used to reduce the mismatch in acoustic-phonetic modeling between training and testing conditions. While using HMMs trained on contaminated speech there is an actual mismatch with the test, in the case of training on real data HMM adaptation [7, 9] can be used to comply with the speaker characteristics and with the specific noise condition.

Maximum Likelihood Linear Regression (MLLR) approach [10] was adopted for batch adaptation of the initial set of Gaussian mixtures to the speaker and to the actual operating acoustic conditions. Each adaptation data set consisted in four phonetically rich sentences which were uttered in the same session of test data collection. Due to the small amount of adaptation data, only means were adapted using a global transformation matrix.

4. Experiments

A first recognition experiment investigates the performance obtained when using three different sets of HMMs, namely: models trained under matched conditions (*Matched*), models trained on contaminated speech (*Contam*) as described in Section 3.2.1, and models trained on clean speech (*Clean*) as a reference case. The resulting Word Recognition Rates (WRRs) are reported in Table 3.

A relevant improvement is observed on *ClTalk* case when using HMMs trained under *Matched* conditions: this fact may be related to the very different interaction style as well as to the background noise and to the acquisition systems characterizing the clean and the in-car databases. Table 3 shows also a progressive relevant improvement obtained with far microphones, firstly using HMMs trained on contaminated speech and finally using HMMs trained on real data.

Recognition performance are then investigated as a function of the SNR at the input, with no matter about which environmental condition corresponded to each utterance. Both test sets were split into three subsets, according to the SNR estimated for each utterance ($SNR < 5dB, 5dB \leq SNR \leq 15dB$ and $SNR > 15dB$). Only *Matched* HMMs are used here and only the channel *Mic2* is considered, as it provides the best results (it was verified that the better performance of *Mic2* is mainly due to its position and not to its type). Table 4 reports on the WRRs corresponding to the three SNR-based subsets.

		ClTalk	*Mic1*	*Mic2*	*Mic3*
cdigits	*Matched*	99.1	92.6	95.3	94.2
	Contam	–	69.6	80.7	73.6
	Clean	95.5	32.8	30.5	38.1
Vodis81	*Matched*	99.1	95.8	97.5	96.3
	Contam	–	83.2	91.2	85.3
	Clean	95.7	27.7	31.9	31.3

Table 3: *WRRs (in %) obtained on the two tasks* cdigits *and* Vodis81*, when using* Matched*,* Contam *and* Clean *HMMs.*

	$< 5dB$	$5dB \div 15dB$	$> 15dB$
cdigits	90.9	96.2	98.4
Vodis81	96.8	97.3	98.7

Table 4: *WRRs (in %) obtained with* Matched *HMMs on* cdigits *and* Vodis81 *tasks, with* Mic2 *and three SNR-based subsets.*

It can be noted that performance on *Vodis81* task is less sensitive to SNR than in the case of *cdigits* task. The reason is that digit sequences of unknown length are more prone to insertions of extra digits when the noise level is high, while this cannot happen with commands, because of the grammar constraint to recognize a single word for each utterance.

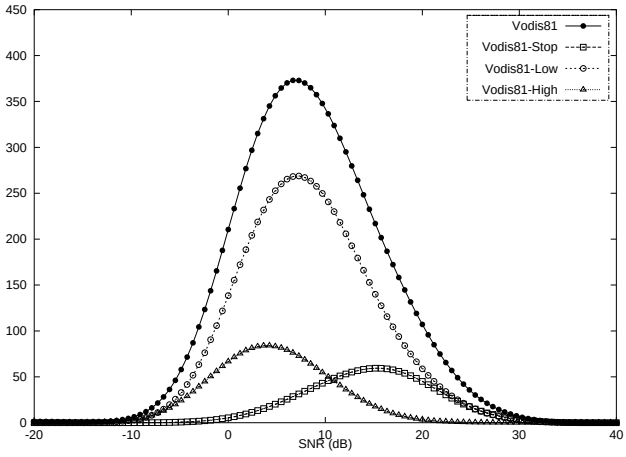

Figure 2: *Distributions of the utterances of* Vodis81 *task as a function of the SNR quantized at 1 dB step (channel* Mic2*).*

However, there is not a univocal relationship between the recording conditions listed in Section 2 and the correspondingly obtained SNR in the recorded signals. This is due to the large variability of driving situations and loudness levels of the speakers. Hence a rough partition has been determined into three levels of car speed: car stopped with motor running (*Stop*), low speed (*Low*) and high speed (*High*). It was found that a finer

partition according to all the possible noise recording conditions is not worth for a deeper insight. Figure 2 depicts the distributions of the items of *Vodis81* task as a function of the SNR. The four curves correspond to the overall test set and to the three speed levels *Stop*, *Low* and *High* respectively.

Table 5 reports WRRs with *Matched* HMMs when the test sets are split according to the three speed-based conditions during data collection. Again a lower immunity to noise (in particular at high speed) can be observed in *cdigits* task.

		Stop	*Low*	*High*	Total
cdigits	*Matched*	96.2	95.8	93.1	95.3
	Contam	86.4	82.5	70.0	80.7
Vodis81	*Matched*	97.6	97.6	97.0	97.5
	Contam	92.3	92.5	85.7	91.2

Table 5: *WRRs (in %) obtained with* Matched *and* Contam *HMMs on* cdigits *and* Vodis81 *tasks, with* Mic2 *and three speed-based subsets. The total WRRs are also reported.*

A final experiment was carried out to assess the effect of batch MLLR adaptation of HMMs to each acquisition session. Signals of each session included in the test set were recognized by using models obtained after adaptation on four phonetically rich utterances acquired from the same speaker and under the same driving conditions. Table 6 reports on the consequent WRRs.

		Stop	*Low*	*High*	Total
cdigits	*Matched*	96.9	96.6	94.8	96.3
	Contam	95.5	93.0	87.0	92.2
Vodis81	*Matched*	98.7	97.7	97.5	97.8
	Contam	98.6	97.2	95.4	97.0

Table 6: *WRRs (in %) obtained with HMM adaptation starting from either* Matched *or* Contam *models.*

Results show the benefits of adaptation (compare with Table 5), particularly evident in some experiments using *Contam* HMMs, even if only four utterances were used. The potential for further improvement, when using more utterances, more iterations and different adaptation techniques may be investigated. However, in a car application a more practical solution is envisaged by adopting an on-line adaptation scheme, in order to adjust the system to rapidly changing conditions.

5. Conclusions

This work has presented some baseline results for what concerns the use of an in-car hands-free speech recognizer for the Italian language.

As expected results showed an improvement with respect to our previous baseline system trained on contaminated speech, thanks to the use of a large portion of the SpeechDat-Car/VODIS II database. As a result, the most significant experiments provided 96.3% and 97.8% WRR, in connected digit and isolated command recognition tasks, respectively.

Many issues still deserve further investigation, among which: the application of a reliable end-point detector, currently under development; the use of online adaptation techniques to further reduce mismatch between training and actual test conditions; improvement in robustness for lower SNRs, for instance using new acoustic features, compensation techniques and a more complex acoustic modeling; the use of special microphones [11] or microphone arrays to further enhance input signal quality.

The next work will also address recognition tasks as spelled words and use of large vocabularies, with the final aim of integrating the resulting technology in an advanced system performing driver-machine dialogue interaction. Most of these activities will be conducted under the European project VICO (Virtual Intelligent COdriver - Information Society Technologies 2000-25426) started in March 2001.

6. References

[1] Y. Muthusamy, R. Agarwal, Y. Gong, and V. Viswanathan, "Speech-enabled information retrieval in the automobile environment", in *Proc. of ICASSP*, Phoenix (AZ), 1999, vol. 4, pp. 2259–2262.

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

[3] P. Lockwood and J. Boudy, "Experiments with NSS, HMMs and the projection, for robust speech recognition in cars", *Speech Communication*, vol. 11, pp. 215–228, 1992.

[4] Asunción Moreno, Børge Lindberg, Christoph Draxler, Gaël Richard, Khalid Choukri, Stephan Euler, and Jeffrey Allen, "SPEECHDAT-CAR. A large speech database for automotive environments", in *Proc. of LREC*, Athens (Greece), 2000.

[5] M. Matassoni, M. Omologo, L. Cristoforetti, D. Giuliani, P. Svaizer, E. Trentin, and E. Zovato, "Some results on the development of a hands-free speech recognizer for car-environment", in *Proc. of ASRU*, Keystone (CO), 1999.

[6] L. Cristoforetti, M. Matassoni, M. Omologo, P. Svaizer, and E. Zovato, "Annotation of a multichannel noisy speech corpus", in *Proc. of LREC*, Athens (Greece), 2000.

[7] D. Giuliani, M. Matassoni, M. Omologo, and P. Svaizer, "Training of HMM with filtered speech material for hands-free recognition", in *Proc. of ICASSP*, Phoenix (AZ), 1999, vol. 1, pp. 445–448.

[8] B. Angelini, F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter, and M. Omologo, "Speaker independent continuous speech recognition using an acoustic-phonetic italian corpus", in *Proc. of ICSLP*, Yokohama (Japan), 1994, vol. III, pp. 1391–1394.

[9] A. Fischer and V. Stahl, "Database and online adaptation for improved speech recognition in car environments", in *Proc. of ICASSP*, Phoenix (AZ), 1999, vol. 1, pp. 449–452.

[10] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[11] R. Aubauer, R. Kern, and D. Leckschat, "Optimized second order gradient microphone for hands-free speech recordings in cars", in *Proc. of the Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, Tampere (Finland), 1999, pp. 191–194.