A ROBUST SPEECH RECOGNITION SYSTEM FOR CAR ENVIRONMENT

Francisco J. Fraga[1], André G. Chiovato [2] and Rodrigo B. Brito [3]

Abstract *— This paper presents a robust speech recognition system designed for the car environment. It was developed in an undergraduate research project and we carried it out in two steps. First, a specific database was designed for this speech recognition task. In the second step, an Automatic Speech Recognition (ASR) system was designed using HTK®, a software tool which is worldwide used in the development of HMM-based (HMM: Hidden Markov Models) ASR systems. We performed two experiences at different settings of training and test. In the first experience, the test conditions were matched exactly to the training conditions, and a recognition rate of 99.88% was achieved. In the second experience, using signals with high SNR for training and signals with low SNR for test, 99.16% of the total test utterances were correctly recognized. These results motivated the undergraduate students in continuing the research towards the implementation of a robust and low-cost ASR system for car environment.*

Index Terms — Robust speech recognition, Hidden Markov Models, car noise.

INTRODUCTION

The robustness of an automatic speech recognition system is strongly influenced by the capability to handle the presence of background additive noise and to deal with the distortion caused by the frequency response of the transmission channel (additive and convolutional noise).

Among the distortions caused by the environment we can include: limited bandwidth of the transmission channel, channel distortion caused by non-linear phase transfer functions, cross talk and echo [1]. These are some of the problems we have to face when treating with real-life scenarios in automatic speech recognition.

Furthermore, it is well known that in noisy environments the speaker increases the vocal effort to overcome the noise, causing a speech distortion called *Lombard effect* [2]. Other effect that changes the speaker's voice is the high gravitational acceleration when the pilot pronounces a word command in the cockpit of an aircraft [2].

In the recent past, several speech recognition systems achieved good performance working in laboratory environment, but failed dramatically when tested in real-life scenarios [1]. In order to improve the recognition rate, modern systems realize an appropriate extraction of robust features to represent speech patterns and/or match the training and test conditions to the same real-life situations. During the last years, an increasing number of researches were publicized on these topics, reflecting the importance of these issues.

This paper presents a robust speech recognition system for working in a real-life scenario that is the car environment, an application with good commercial acceptance and a real challenge for current available speech recognition algorithms.

Speech captured in a running car is perturbed by noise from engine and tires, from wind, rain and traffic. With a microphone placed 50 cm away from the mouth of the speaker, the resulting captured signal may exhibit a negative signal-to-noise ratio (SNR) in decibels [3].

The speech recognition for controlling the car functions (such as headlight, windows, lock and others) is an application we are currently working on in an undergraduate research project of the Digital Signal Processing Research Group from Inatel. This is a typical situation that is extremely appealing for an ASR application.

When a person is driving a car, his or her eyes and hands are entirely occupied on this task and the manual control over any car function can cause distraction in traffic. The use of the own voice for command the car functions could avoid this problem. Other interesting use of an ASR system is in long trips, especially during the night, when the voice control over the headlights (full beam, low beam) can diminish the possibility of an accident caused by sleeping.

In order to develop an ASR system for this application, we have divided the job in two steps. First of all we have done several speech recordings to build an organized speech database for training the speech models, a mandatory requirement for any ASR system. After that we have designed the system itself, using HTK® (*Hidden Markov Models Toolkit*), a computer tool worldwide used for developing HMM-based ASR systems.

THE SPEECH DATABASE

With the aim of developing an ASR system for the car environment, we built a specific speech database, which was composed by utterances from two male speakers. They recorded his voices when were driving a car under various conditions: in streets and roads at different speeds (35 and 80 km/h), with and without traffic noise, with asphalt or stone pavement and with the car windows opened or closed. Two microphones simultaneously recorded the speech signal; both were about 50 cm away from the driver's mouth. One microphone was placed on

[1] Francisco José Fraga, Inatel - Instituto Nacional of Telecomunicações, Santa Rita do Sapucaí , MG, Brazil, fraga@inatel.br
[2] André Godoi Chiovato, Inatel, Av. João de Camargo, 510, 37540-000, Santa Rita do Sapucaí , MG, Brazil, andregc@inatel.br
[3] Rodrigo Barbosa Brito, Inatel, Av. João de Camargo, 510, 37540-000, Santa Rita do Sapucaí , MG, Brazil, rbarbosa@inatel.br

the car panel (at the same level of the steering wheel) and the other was placed near the car ceiling (this second microphone had poor quality, it costs five times less than the first one). The vocabulary contained about 50 isolated words, each of them was pronounced three times by the speakers in each recording condition. The final speech database had over 10,000 utterances.

The microphones used were: *Super Directional Desktop Microphone – M60* e *Handheld Recorder Microphone – M10* (the cheaper one), from *Telex Communications*® company, which illustrations can be viewed in figures 1 and 2. The speech recording was done by means of a portable Digital Audio Tape (DAT), model *TCD-D100* from *Sony*® company, as illustrated in figure 3. Figure 4 shows the recording scenario, which is the interior of a car from *Fiat*® company, model *Pálio 1.6 16 V*. In this figure we can see the microphones connected to the DAT through two audio cables.

After recorded by the DAT, the speech signals were transferred to a computer where they were separated and titled according to the vocabulary word, the utterance (if it was the first, second or third repetition of each word), the speaker and the recording conditions, as the example showed below:

TemperaturaInterna2_André_Av_Ab.wav

It means that the speaker *André* recorded the Portuguese word command *TemperaturaInterna* for the *2nd* time at the main *Av*enue with opened windows (the car windows were "*Ab*ertos" as we say in Portuguese)

FIGURE 1
M10 MICROPHONE

FIGURE 2
M60 MICROPHONE

FIGURE 3
DAT TCD-D100

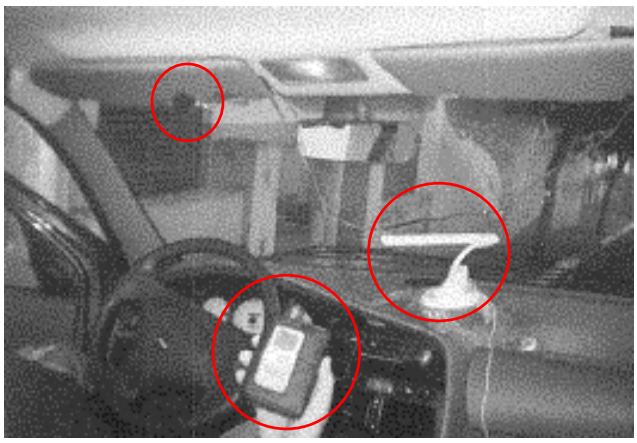

FIGURE 4
LAY-OUT OF THE RECORDING ENVIRONMENT

Table I shows the vocabulary, where we can see all the Portuguese words that the driver can say to command the car and the cellular phone functions.

TABLE I
COMPLETE VOCABULARY

Word commands for the cellular phone functions	Word commands for the car functions
1. Bloquear 2. Desbloquear Menu 3.1. Agenda 3.1.1. Nomes 3.1.2. Número 3.1.3. Polícia 3.1.4. Bombeiros 3.1.5. Emergência 3.2. Chamadas 3.2.1. Atender 3.2.2. Receber Mensagem 3.3. Personalizar 3.3.1. Nome (do usuário) 3.3.2. Perfil original 3.3.3. Campainha 3.3.3.1. Aumentar 3.3.3.2. Diminuir 3.4. Assistente Pessoal 3.4.1. Calendário 3.4.2. Atividades 3.4.3. Despertador 3.4.4. Calculadora 3.4.5. Administrador de contatos 3. Funções Avançadas 3.1. Memória (tamanho) 3.2.Conferência Telefônica 3.3. Rediscagem 4. Serviços Internet 4.1. Lazer 4.1.1. Teatro 4.1.2. Restaurante 4.1.3. Shows 4.1.4. Cinema 4.2. Trânsito 4.2.1. Livre 4.2.2. Congestionado 5. Desligar	1. Farol Alto 2. Farol Baixo 3. Pisca – Alerta 4. Travas 5. Capo 6. Porta – Malas 7. Ventilação 8. Alarme 9. Tanque 10. Combustível 11. Óleo 12. Abaixar Vidros 13. Levantar Vidros 14. Temperatura Interna 14.2. Ambiente 14.3. Motor

The list of some speech recording situations is presented in table II.

TABLE II
EXAMPLES OF SOME SPEECH RECORDING SITUATIONS

Speaker	Place	Pavement	Windows	Speed
André	BR 459	Bad asphalt	Closed	80 km/h
André	BR 459	Bad asphalt	Partially opened	80 km/h
André	BR 459	Bad asphalt	Opened	80 km/h
André	Pouso Alegre	Good asphalt	Opened	35 km/h
Rodrigo	Pouso Alegre	Good asphalt	Opened	35 km/h
Rodrigo	Pouso Alegre	Good asphalt	Partially opened	35 km/h
André	Pouso Alegre	Good asphalt	Partially opened	35 km/h
André	BR 381/km 794	Acceptable asphalt	Partially opened	80 km/h
Rodrigo	BR 381/km 775	Acceptable asphalt	Opened	80 km/h
Rodrigo	BR 381/km 794	Acceptable asphalt	Closed	80 km/h
Rodrigo	BR 459	Bad asphalt	Opened	80 km/h
André	BR 459	Bad asphalt	Opened	80 km/h
André	Ave. João de Camargo	Good asphalt	Closed	35 km/h
André	Ave. João de Camargo	Good asphalt	Opened	35 km/h
Rodrigo	Marques Street	Stone	Opened	35 km/h
Rodrigo	Marques Street	Stone	Partially opened	35 km/h
Rodrigo	Marques Street	Stone	Closed	35 km/h

The final recorded speech database consists of 5.344 files (.wav). In order to perform this task, we have to pronounce the word commands while driving the car in roads and streets for 266 km .

THE ASR SYSTEM

In the second step of our undergraduate research project we have designed an ASR system using some tools of the HTK® software.

In the front-end stage, we have extracted 12 mel-cepstral coefficients from speech frames with 30 ms duration and an overlapp of 33%, which leads us to a frame rate of 100 Hz. For each utterance, the mean of the mel-cepstral coefficients were computed over all frames and then were subtracted from the coefficients vector with the aim of neutralize the channel distortion [2].

In order to form the observation vector, the first and second time derivatives were added to the end of the 12 mel-cepstral coefficients vector.

A single left-right HMM was used for modeling each vocabulary word. The number of states of each model was variable according to the number of phonemes of the vocabulary words.

Five gaussian mixtures with diagonal covariance matrix were used to estimate the acoustic emission of each state of the HMM's.

The models were trained with a third part of the utterances; the remained utterances were used to evaluate the system performance.

Two experiences were carried out: In the first one, the test conditions were matched exactly to the training conditions, which is the ideal way of developing a robust ASR system, although it is difficult to realize in a real-life application [4].

The performance achieved with this set up, in terms of correct word recognition rate, was **99,88%.**

The second experience was carried out at just the opposite situation: the utterances used for training were the best ones (high SNR) and those used for test were the worst ones (low SNR). Although running at these adverse conditions, **99.16%** of the total test utterances were correctly recognized.

CONCLUSIONS

We conclude that these good results were achieved because we extracted robust features from the speech signal (mel-cepstral coefficients with mean cepstral subtraction). We observed also that the same performance was achieved by the two microphones. This fact leads us to the conclusion that the distortion caused by the cheap microphone was neutralized by the robust feature extraction scheme.

The speech database will serve also for other research projects involving robust speech recognition and speech enhancement.

Finally, we want to remark that the students that worked in this undergraduate research project became enthusiasts of the speech recognition area and both have already started his master level studies in speech processing.

ACKNOWLEDGMENT

We sincerely thank for the help of Prof. Carlos A. Ynoguti and for the funding of FINATEL and FAPEMIG .

REFERENCES

[1] B. H, Juang; L. R., Rabiner, *Fundamentals of Speech Recognition*. Ed. Prentice-Hall, Signal Processing Series, New Jersey, 1993.

[2] J. C. Junqua; J. P. Haton, *Robustness in Automatic Speech Recognition – Fundamentals and Applications*, Kluwer Academic Publishers, Norwell, Massachusetts, 1996.

[3] C. E. Mokbel; G. F. A. Chollet, "Automatic Word Recognition in Cars", *IEEE Trans. on Speech and Audio Processing*, vol 3, n.º 5, pp. 346-356, Sep. 1995.

[4] R. M. Stern; A. Acero; F. H. Liu; Y. Ohshima, "Signal processing for robust speech recognition", *in Automatic Speech and Speaker Recognition: Advanced Topics*. Eds. Norwell, MA, 1997.