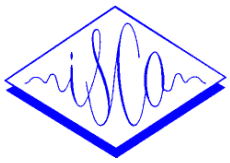

2001: A Speaker Odyssey
The Speaker Recognition Workshop
Crete, Greece
June 18-22, 2001

A VQ Speaker Identification System in Car Environment for Personalized Infotainment

Javier R. Saeta[1*] *Christian Koechling*[2] *Javier Hernando*[1]

(1) TALP Research Center
Universitat Politecnica de Catalunya
Barcelona, Spain
Email: jrodsae@gps.tsc.upc.es

(2) Robert Bosch, GmbH
Hildesheim, Germany

Abstract

Car applications demand more and more the use of speech technologies. Drivers must concentrate on controlling the car and the non-use of hands makes the voice a valuable tool. Here we analyze the possibility of identifying the user of a car through her/his voice in order to develop some useful applications, and establish preferences, some of them related to music. The identification will be done in parallel to speech commands which will be given to devices in the car in the future. Once the user is identified, the system loads a personal profile. It includes music preferences which can be downloaded from the Internet databases using e.g. MPEG-7 .

1. Introduction

The work has been developed in the FV/SLM 'Advanced Development Multimedia Systems' department at Robert Bosch GmbH, Hildesheim, Germany. The user identification part has been examined as contribution for the VHE (Virtual Home Environment) -Middleware project. This project investigates among other things future 'infotainment' systems for the home and vehicular environment. The word infotainment refers to sytems where the information and/or the entertainment play an important role. The user identification might be used in the future in combination with other recognition methods like fingerprint sensors, chipcards, login-passwords or face recognition systems. VHE-Middleware is a ITEA (Information Technology for European Advancement) project. ITEA belongs to the EUREKA program, and it is dedicated to strengthening software and software engineering competencies at European industrial level. ITEA has become deeply involved in focusing European innovation on the development of open middleware and embedded software, applicable to the various domains of the e-economy.

The potential for application of speaker recognition systems exists any time speakers are unknown and their identities are important. Applications [1] running on future navigation systems and computers inside the car can be arranged in a better way according to the identity of the person.

Our purpose here is to identify the speaker, but this identification is not a security matter –although it could be used in this way in a future. We intend to adjust services for a certain number of users of a car, but once these users are inside the car. The system that we will develop is classified as *closed-set text-independent* speaker identification. It is text-independent because it is not important what the speaker is saying. It is closed-set because the speaker -the car driver- should be in the database [2]. The system can also be enlarged to regular passengers. The users could be a family, a single person or a higher number of people in case we deal with a company car. We have therefore a limited number of users. This is the reason why we use speaker identification instead of speaker verification. Moreover, the system should be capable of recognizing the user as soon as possible, i.e. with a few number of words, and without any kind of questions. It will work in real time, so we need a fast recognition and, obviously, at the same time, a high accuracy.

2. Speaker identification system design

Speakers' voice samples were recorded in the office and some background noises were taken in a car. The recordings were sampled at 22 kHz, by a direct audio cable connector from a microphone to a soundblaster card, with 16 bits of resolution. A silence detector takes the non-speech signal out. It computes the energy and the zero-crossing rate of the signal. The detector gives the entire region where speech exits in an input signal. This speech could include voiced regions as well as unvoiced regions. We assume the energy level and the zero-crossing rate are constants in average.

The data is processed in 10 ms frames, Hamming windowed and preemphasized with a zero at z=0.97. The feature set consists of 12th-order mel-frequency cepstral coefficients (MFCC), and the normalized short-time log energy, with a mel-filterbank of 14 triangular-shaped filters. This set is augmented by the corresponding delta MFCC from 5 successive cepstral coefficients, delta-delta MFCC from 5 successive deltas and delta energy to form a 39-dimensional vector for each individual frame. We use here a 256-FFT. This number is suitable to facilitate the fast radix-2 FFT.

After the computation of MFCC, we will make use of pattern recognition techniques [3,4], particularly *Vector Quantization* (VQ). With VQ, the data is significantly compressed, yet still accurately represented. It has been selected for its computational cost, low complexity and easy

[*] This work has been developed in a seven-month stay at Robert Bosch GmbH

implementation [5,6]. The codebook will be composed by 16 codewords. We consider this number is high enough for establishing a division among the feature vectors in our application. As we can see in Figure 1, the complexity will increase with a higher number of codewords, but the accuracy remains almost constant.

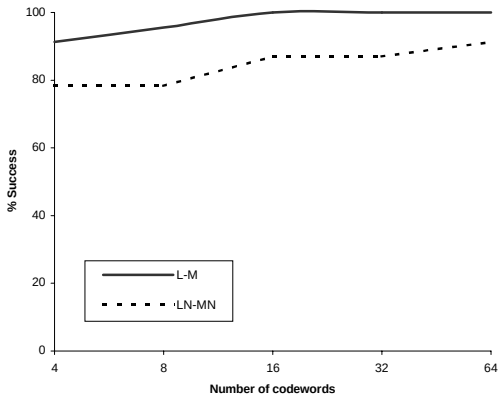

Figure 1: Decision about the number of codewords

The first step consists of creating an initial codebook from a sound file formed by chunks of speech that belong to future users of the car, using the k-means algorithm. After the process of creating a generic codebook, we create a model for every speaker by clustering her/his training acoustic vectors according to the previous codebook by means of the LBG algorithm [7]. In the testing stage, the user will be identified with the speaker whose codebook yields the minimal distortion with the given vectors. The distortion is calculated with the Euclidean distance.

The system has a 'doubt' mechanism. It establishes a certain threshold. If the distance between the two lowest distortions is under a certain experimental value, which varies depending on the number of speakers, then the system returns a null value. This means that it is difficult to distinguish among the two speakers with the lowest distortions and the system cannot ensure who is speaking. In this case, the user should talk again.

Once the user has been identified, a personal profile, previously defined by the user, is automatically loaded. This will let the user to control the favourite music inside the car. For instance, (s)he will be able to get information about, in addition to the song number or its duration, even the title, the lyrics, the credits, the video, the biography of the authors and some other aspects related to concerts, next tours, music awards or merchandise. Some music players and browsers let the user to obtain this kind of information through the Internet [1], when they play a CD from the computer. In the future, the system will also be able to learn about which are the user preferences based on her/his habits.

MPEG-7 [8,9] becomes an excellent tool after the identification process and makes possible to define a user profile. It specifies a way of describing various types of audiovisual information, including pictures, video, speech, audio..., irrespective of its representation format and storage support. This will allow the user to get access to lots of applications in the future.

The block diagram of the whole system is represented in Figure 2.

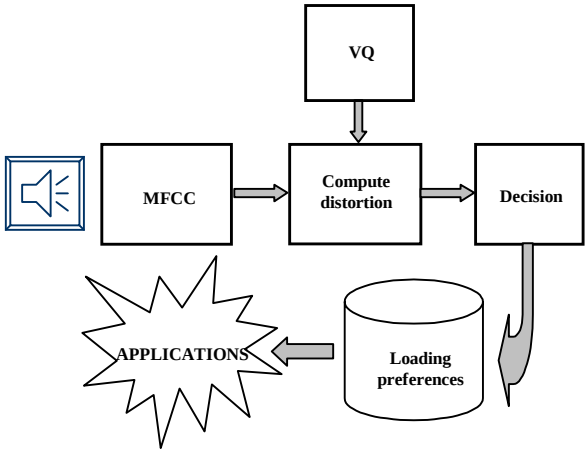

Figure 2: Block diagram of the speaker identification system

3. Database and scenarios

We have recorded utterances from different speakers in order to create a database. Two scenarios have been selected to test the application.

3.1. Database description

Our database has 23 speakers (21 males and 2 females). Every speaker has 7 training sentences in order to prove the performance of the designed system. The sentences are identical for every speaker. Hence we have:

- one long sentence of approximately 20 seconds,
- two medium length sentences of 6-7 seconds, and
- four short single words (3-4 seconds).

They all were mainly selected because of their wealth and variety of sounds.

3.2. Scenario I

We have recorded the voice samples in an office environment. The place is surrounded by lots of noises: people speaking, typing or clicking their mouses, the fans of the computers... We can consider this recording environment as a noisy one, with an average SNR of 20 dB. We have seen that the noise level here is not far from the noise level inside the car when the engine is on, the car is stopped, the windows are closed and the radio is off. Of course the type of noise is different.

3.3. Scenario II

The background noises in a car have been recorded and added to the samples taken from the Scenario I. This creates a new scenario. The SNR is now lower and decreases with the car speed. The average SNR at 50 km/h is around 8 dB and nearly 6 dB at 100 km/h. We should note that this scenario will be an aproximation to speech collected in a car. Factors such as the Lombard effect are not here properly modeled.

4. Experimental results

Here we present the results obtained with the identification system. We have decided to use MFCC, although some more tests were made with the use of Linear Prediction Coefficients (LPC) and frequency formant analysis, but results were not as good as MFCC. The combination of LPC and MFCC became also inefficient.

4.1. Threshold adjustment

The first step consists of adjusting the threshold value. In Figure 3, the vertical axis represents the percentage of success, i.e. the number of correctly recognized speakers out of the total number of speakers. The line on top (dashed) corresponds to the recognition of a short sentence. The dash-dotted line belongs to a medium sentence. The training procedure was made in both cases with a long sequence. We choose 0.1625 as the suitable threshold because the recognition rate becomes almost constant from this value, and it is acceptable for this application.

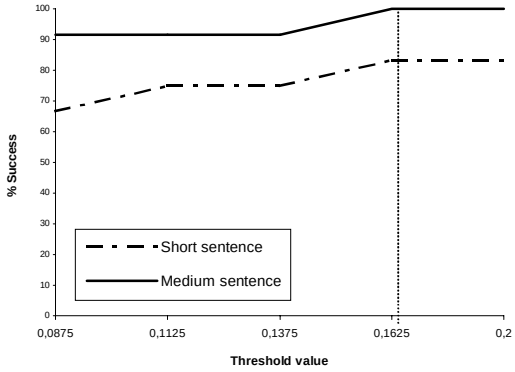

Figure 3: Adjustment of the threshold value

4.2. Recognition rates

Figures 4 and 5 indicate the percentage of success depending on the number of speakers in the database used for these experiments. At the legend in both figures, the first letter corresponds to the length of the utterance used in the training procedure, whereas the second one is the length of the utterance used for testing the system during the recognition process. The utterances can be –depending on their length- long (L), medium (M) or short (S).

In figure 4, L-M means that we train the system with a long sequence and we test it with a medium length sequence. A 100% of success is obtained in this case. The worst results belong to the training with M and the testing with only a word (S). It has been observed that the majority of errors belongs to those speakers whose SNR are low. Different SNR among speakers do not only depend on the noise or reverberation in the room, but also they are due to the distance from the speaker to the microphone.

Figure 5 shows the importance of matching the training and testing scenarios. The utterances are recorded at 50 km/h. We have here the medium and long utterances of the Scenario I, plus the background noise in the car (LN and MN). The recognition rate falls to around 85% in this scenario, with the presence of the noise in the car in both training and testing procedures (LN-MN). We can also observe an increase of the percentage of success when the number of speakers becomes higher. This happens because our number of experiments is small. The percentage of success would tend to decrease if we augmented the number of trials. On the other hand, when we do not match the scenarios (L-MN), we observe the recognition rate is considerably lower and decreases when we increase the number of users.

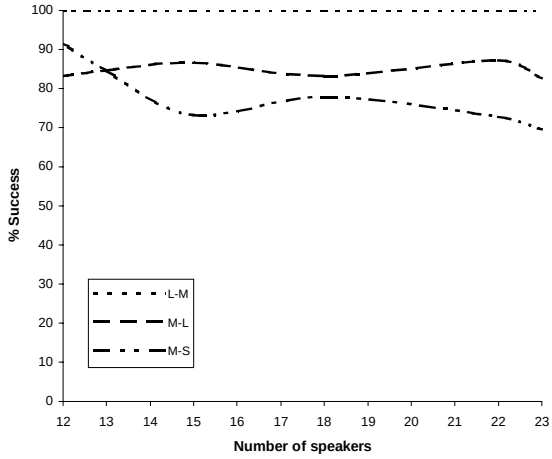

Figure 4: Recognition rates with Long (L), Medium (M) and Short(S) utterances

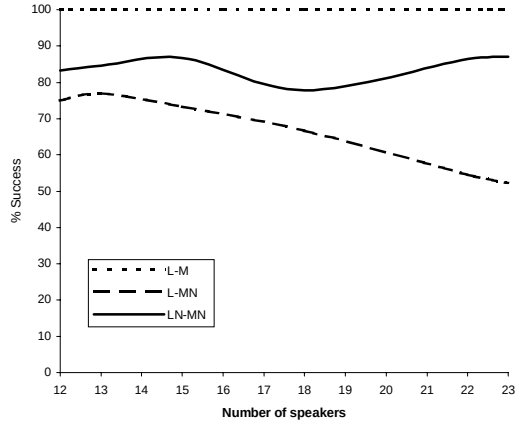

Figure 5: Recognition rates with Long (L), Medium (M), Long with car Noise (LN) and Medium with car Noise (MN) utterances

We have also tested the system at 100 km/h. In this case, the recognition rate fell dramatically when we trained in the Scenario I and recognized in the Scenario II. On the contrary, when training utterances corresponded to the Scenario II, the results improved. They were similar to those reached at 50 km/h.

We should add that the two female speakers did not elicit any mistake. Pitch makes female voices different from male ones. Due to this fact, the number of crossed-sex errors becomes smaller.

4.3. Computational cost

In an identification system, the speaker is compared with all the models in order to decide who (s)he is. The more users you have, the slower the system will be. Nowadays, due to the technological evolution of the microprocessors, it is expected that an identification system works well in real time in terms of computational cost when the number of speakers is small, as in our case. Anyway, we can observe little differences depending on the length of the sentences in Figure 6.

It has also been proved that the silence detector strongly influences the computational cost. We can reduce this cost by a factor of four if we remove the non-speech part of the speech file.

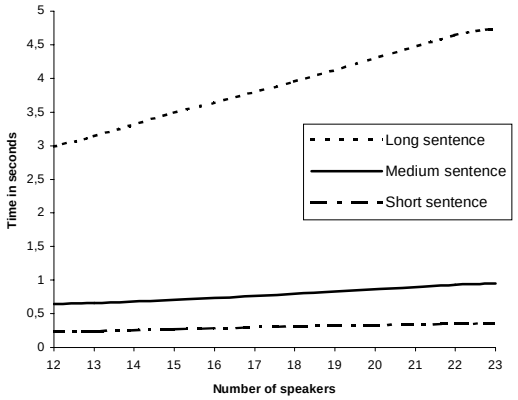

Figure 6: Computational cost during the recognition process

5. Conclusion

Tests made on the set of speakers of our database show no errors with long and medium, training and testing utterances in the Scenario I, with a SNR over 20 dB. This rate decreases by around 15% in the Scenario II, when we add the noise of the car. Some more tests have been carried out for similar conditions (parking, accelerating...) to those reached in the Scenario II concerning the noise level, and the results have been similar.

It has been decided to use MFCC because we have shown they work better than LPC or formant frequencies in our speaker recognition tests. Vector Quantization approach has become here a fast and accurate technique. The computational cost is acceptable and the results are satisfactory for this application, with the levels of SNR considered in both scenarios. In an application with a higher number of speakers, another pattern matching technique would probably become more appropriate.

6. References

[1] Gracenote, CD DataBase (CDDB) Home Page, *www.cddb.com*

[2] H. Gish and M. Schmidt, "Text-Independent Speaker Identification", *IEEE Signal Processing Magazine*, vol. 11, no. 4, pp. 18-32, October 1994.

[3] L. Rabiner and B.H. Juang, "Fundamentals of Speech Recognition", Prentice-Hall, Englewood Cliffs, New Jersey, USA, 1993.

[4] J. P. Campbell, Jr., "Speaker Recognition: A Tutorial", in *Proceedings of the* IEEE, vol. 85, no. 9, pp. 1437-1462, September 1997.

[5] E. Gabrilovich and A.D. Berstein, "Speaker Recognition: Using a Vector Quantization Approach for Robust Text-Independent Speaker Identification", *Technical Report DSP Group*, Inc., Santa Clara, California, USA, September 1995.

[6] N. C. Ward amd D. R. Dersch, "Text-Independent Speaker Identification and Verification using the TIMIT Database", in *Proc. ICSLP'*98, vol. 2, pp. 233-237, November 1998

[7] Y. Linde, A. Buzo, and R.M. Gray, "An Algorithm for Vector Quantizer Design," *IEEE Transactions on Communications*, vol. 28, no. 1, pp. 84-95, January 1980.

[8] MPEG Requirements Group, "MPEG-7 Multimedia Description Schemes XM (Version 4.0)", Doc. ISO/MPEG N3464, MPEG Beijing Meeting, July 2000.

[9] MPEG Requirements Group, "MPEG-7 Multimedia Description Schemes WD (Version 4.0)", Doc. ISO/MPEG N3465, MPEG Beijing Meeting, July 2000.