ROBUST SPEECH RECOGNITION TECHNIQUES EVALUATION FOR TELEPHONY SERVER BASED IN-CAR APPLICATIONS

Lionel Delphin-Poulat

France Télécom R&D DIH/IPS, 2, Avenue Pierre Marzin
22307 Lannion, Cedex France
lionel.delphinpoulat@rd.francetelecom.com

ABSTRACT

In this paper, the feasibility of designing a speech-recognition based telephony server for in-car applications with an acceptable recognition rate is investigated. The whole acoustic channel (sound pickup, sound transmission over the cellular network, feature extraction) is evaluated: the loss or the gain in performance due to each element is quantified. More precisely, two sound pickup systems (a hypercardioid microphone and a microphone array) were tested. A standard MFCC and the Aurora Advanced front-ends were evaluated. Recognition performances were measured before and after transmission over a cellular (GSM) network. The gain of using either a robust sound recording device or noise robust front-end is demonstrated.

1. INTRODUCTION

Speech recognition is a natural way to interact with car devices while keeping hands on the wheel. Among different applications, we can distinguish command and control applications (for car radio, air conditioning) and database access applications (such as car navigation, large directory access). Numerous studies have been carried out with on-board systems using speech recognition. However, in the case of database access applications, server-based speech recognition might be useful. Indeed, database may change along time. For example, for car navigation systems, the cartographic database must be regularly updated by sending CDs to customers. If the database is on the server, speech recognition may also be performed on server to easily keep the consistency between the database and the speech recognition models. Moreover, in this case, the cost of a powerful speech recognition engine in the car is avoided. But the voice access to the recognition engine is done through a cell phone link. Therefore, speech recognition has to be effective in a very adverse environment. There are the unavoidable difficulties in a car, such as car noises (engine, fan…) and reverberant acoustics. The sound pickup is done with a hand-free device; acoustic echo cancellation may also be done if barge-in is required. Furthermore the cell phone link involves speech coding and transmission. Finally, if the car is moving, the radio channel is changing rapidly; cell switching may occur.

This study had different goals. We investigated the influence of the microphone location and we compare the advantage of using a microphone array over a simple microphone. In this study, the problem of acoustic echo cancellation in conjunction with noise reduction (for a survey of this problem, see for instance [1] and references therein) was not addressed. We evaluated two feature extractions (with and without additive noise reduction). One of them was designed to be robust against additive noise since it is known to be an important source of mismatch in speech recognition. Finally, the loss in performance due to the transmission channel was quantified.

2. DATABASES

Two databases were used to perform the evaluations.

A small database was recorded at France Télécom. This database was recorded in a Peugeot 406. 3 different sound pickup systems were used: a close talk microphone, a far talk microphone (AKG Q400-II) and a microphone array (from Andrea Electronics). They were placed on the car ceiling in front of the driver. The sound recorded through those 3 systems was transmitted over the GSM network. Furthermore, the car was equipped with a Nokia car-kit (with a hands-free system using an AKG microphone). This database contains only digit sequences of length 4. 10 speakers (5 male and 5 female) were recorded in this database. Each speaker uttered 10 sequences in 3 different car conditions (stopped car, car running on a highway and car running in city traffic). The on-board platform synchronously recorded the local channels. A synchronization procedure was designed and applied to carefully synchronize the GSM channels with the local ones. The on-board signal was recorded on 16 bits at 16kHz. The telephony signal was recorded in A-law.

We also used the French Vodis database [2]. This database was recorded simultaneously over a close-talk microphone, and an AKG microphone in three different locations, in three different cars (VW Passat, Peugeot 406, Laguna Renault). This database was essentially used to choose the correct microphone location, to adapt phoneme

models and to quantify the effect of a noise-robust front-end.

3. EXPERIMENTAL SET-UP

The experiments were performed with a HMM-based speech recognition engine, using context-dependent phoneme models. Densities are mixtures of Gaussians (with at most 8 Gaussians per density).

Two front-ends were used. We used 12 Mel Frequency Ceptral Coefficients (MFCC) and energy, and their first and second order derivatives (the frame rate is 16 ms). Those MFCC include a cepstral mean normalization technique and are referred as MFCC, in this paper. Internal results showed that this front-end performs slightly better than the Aurora front-end [3].

The Aurora 2 front-end [4] (also called Advanced Front-End for distributed speech recognition) was also evaluated. It includes a noise reduction technique. Energy and the first 12 coefficients and their first and second order derivatives were used. The dimension of the feature space is thus the same. The features were down-sampled at 16ms (the standardized frame rate is 10ms). The same HMM topology and the same derivation filters could thus be used for both front-ends.

Both analyses were applied on a 8kHz-sampled signal (on board-signals were down-sampled). The baseline models were trained with 192 hours of telephony speech.

4. ADAPTATION PROCEDURE

Acoustic models adaptation was performed using Vodis data. This database was split into adaptation sessions and testing sessions. Each session contains data from different corpora. Some corpora were used only for adaptation and others only for evaluation. Adaptation was thus done only with data of the adaptation sessions belonging to the adaptation corpora while testing was done only with data of the testing sessions belonging to the testing corpora.

We used the R (spontaneous phrases, 5 utterances per session) and S (phonetic sentences, 5 utterances per session) corpora as adaptation corpora. Evaluations were performed on two corpora: T (telephone numbers, 5 utterances per session) and W (control and command words, 70 words per session). Sessions 018 to 068, which were recorded in the Peugeot and sessions 086 to 183, which were recorded in the Renault were chosen as adaptation sessions. Remaining sessions (17 recorded in the VW Passat, 17 in the Peugeot and 17 in the Renault) were used for testing. In the following results, the three far-talk channels were used to adapt the model with the incremental adaptation technique [5] (the *a priori* Gaussian weight was limited to 50).

5. SOUND PICKUP

5.1. Microphone Location

In the Vodis database, speech was recorded through four microphones: a close-talk microphone and 3 far-talk (AKG Q400-II) microphones that were placed on the ceiling (one on the left, one in front of the driver and one on the right). The results reported in this paragraph were obtained on the W task. They were confirmed by results on the T task.

Results with baseline models are presented on Figure 1. The microphone location in front of the driver (middle) is thus shown to be the best. Those results confirmed the conclusion of a previous internal study in which the place of the microphone was carefully studied to obtain a good sound pickup for transmission. Other studies also confirmed those results [6]. It is to be noticed that baseline models are not well suited for local speech recognition since they were trained on telephony data. Both front-ends gave similar results except for the right microphone.

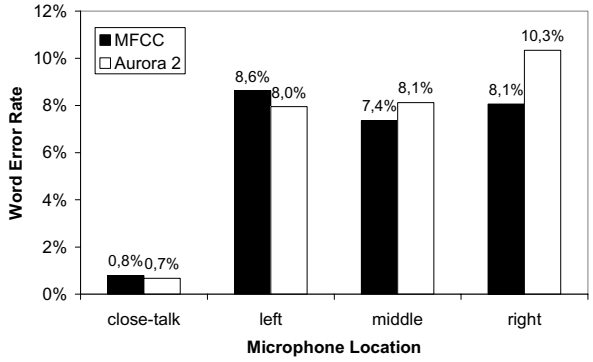

Figure 1: Microphone Locations Comparison before Adaptation

The same measurements were done with adapted models (Figure 2).

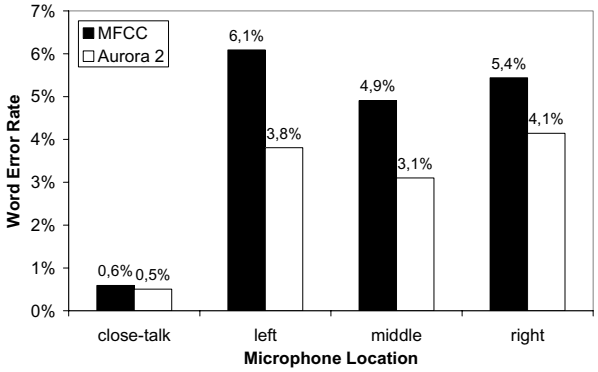

Figure 2: Microphone Locations Comparison with Adapted Models

With adapted models, the best microphone location is still the same (*i.e.* in front of the driver). After adaptation, the Aurora 2 front-end gave better performance than the MFCC front-end for all microphone locations.

5.2. Sound Pickup Systems

The subsequent results are obtained with the small France Télécom database. In this digit string task, we used context-dependent phoneme models and not specialized word models. The goal was not to achieve the best recognition results but to measure relative differences. Sound pickup is compared in different conditions: we used phoneme models that were trained using a large database and the same context-dependent phoneme models after adaptation on the Vodis database (*i.e.* the same adapted models as in the previous paragraph). Results on the close-talk channel serve as references.

The results with non-adapted models and before any transmission are reported on Figure 3.

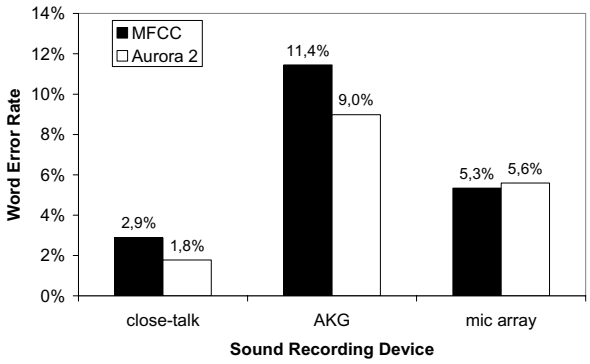

Figure 3: Sound Pickup before Transmission before Adaptation

In this case, the microphone array gives better performances than the microphone. We can see that the Aurora 2 front end leads to substantial improvements in the case of the simple microphone over a standard MFCC front-end. But for the microphone array, there is no difference between the two front-ends: with the microphone array, a noise spectral component, which is out of the beam defined for this frequency is cancelled. Thus the effects of noise reduction are very small.

Here we used the phoneme model adapted on the Vodis database. We recall that those models where adapted using the far-talk channels recorded through an AKG microphone. Results on data collected before transmission are reported on Figure 4.

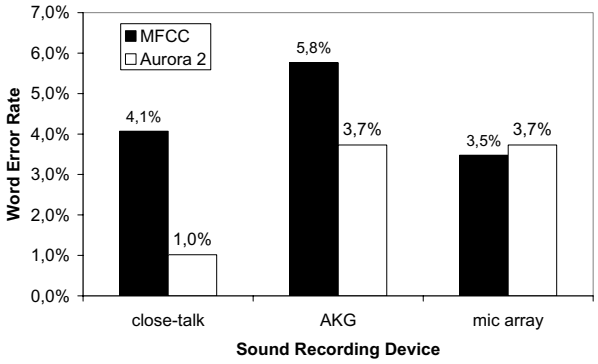

Figure 4: Sound Pickup Comparison before Transmission with Adapted Models

First, adaptation leaded to a dramatic error rate reduction. For the microphone array, we can see that both feature extractions gave the same recognition rate, whereas the noise reduction can enhance the speech signal recorded by the AKG microphone (the explanation is the same than for the previous results). With the standard MFCC, the microphone array gave better results than the AKG microphone. But with the Aurora 2 front-end, the error rates obtained for both studied sound pickup systems were the same. It should be kept in mind that adaptation was done with speech material collected in the same acoustic conditions (same car Peugeot 406 and same microphone) as far as the AKG microphone is concerned. For the microphone array there was still a sound pickup mismatch between adaptation and testing. To know if it is worth using a microphone array, a larger database recorded through a microphone array should be collected to obtain models that are specific to this sound pickup and to measure performances.

6. TRANSMISSION

For all the sound recording systems, the same experiments as in the previous paragraph were carried out after transmission over the GSM network. In addition, the error rate on data that was transmitted through the car-kit was also measured. Results with baseline models on transmitted data are reported on Figure 5.

After transmission, there are no more differences observed between the microphone-array and the microphone. For the microphone array, there is a dramatic loss in performance due to the channel. However, the GSM channel does not seem to affect the recognition rates on the AKG microphone (the sound quality of the AKG-recorded speech is already bad before transmission). The recognition results are much worse for the car kit. We did not have access to the signal processing features of the car kit, but it seems that some sequences are cut by a kind of Voice Activity Detector. The signal processing that is

going in the car kit is harmful for speech recognition. The speech enhancement techniques in car equipment should be carefully designed to obtain good performances for both speech transmission and speech recognition.

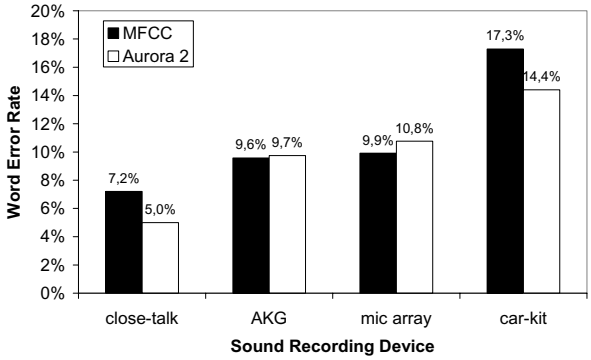

Figure 5: Sound Pickup Comparison after Transmission without Adaptation

Finally we evaluated the Vodis adapted models on data collected after transmission.

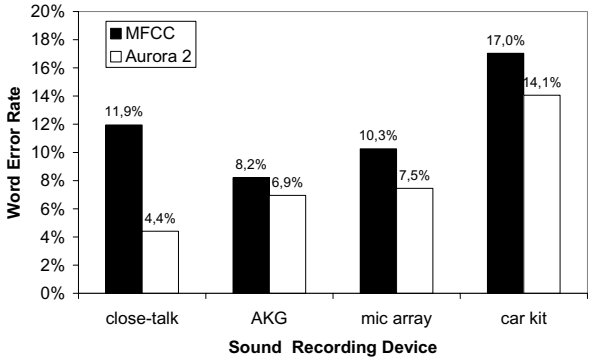

Figure 6: Sound Pickup Comparison after Transmission with Vodis Adapted Models

Though the Vodis database was recorded in the car, we see that adaptation was also useful for data collected over the GSM network for the far-talk microphone and the microphone array (for the Aurora 2 front-end). Of course the relative improvements were smaller than for data collected in the car. For the car-kit channel, adaptation did not bring any improvement. Adaptation was more efficient for the Aurora 2 front-end.

The channel had a negative impact for all sound pickup systems. Of course, the loss was even more important when the sound pickup was efficient. One solution to decrease this loss is to use distributed speech recognition *i.e.* to compute feature coefficients in the car, transmit them in data mode at a much lower bit rate than coded speech and to perform speech recognition on the server. The obtained results should be between the on-board and the server results. Hopefully, they should be closer to the on-board results.

7. CONCLUSIONS AND FUTURE WORK

In this work, we studied the whole channel starting from sound pickup and ending in feature extraction to design a good speech recognizer for server-based in-car applications. It was shown that before any adaptation, a microphone array could lead to better recognition rates. However this advantage disappears after adaptation. Interaction between sound pickup and speech recognition deserve more attention. Additional data is being collected through the microphone and the microphone array in a Clio.

With baseline models, there is no advantage the Aurora 2 front-end over the standard MFCC for all sound pickup systems. With adapted models, the Aurora 2 front-end performs significantly better than MFCC front-end for the AKG microphone. For the microphone array, the results are equivalent for both front-ends; the microphone array already provides some speech enhancement processing.

Finally, we have shown that a cellular transmission (including speech coding and radio transmission) significantly degrades speech recognition. Further experiments using the SpeechDat Car database are currently carried out to confirm those results and to obtain further model optimizations. A promising solution to the transmission problem is to use a distributed recognition system: a prototype is under development to collect realistic data.

8. REFERENCES

[1] R. Le Bouquin Jeannès, P. Scalart, G. Faucon and C. Beagant "Combined Noise and Echo Reduction in Hands-Free Systems: A Survey", *IEEE Trans. On Speech and Audio Processing*, Vol. 9, n°8, pp. 808-820, November 2001.

[2] http://www.loria.fr/equipes/parole/Html/anglais_outils.html

[3] ETSI, "ETSI ES 201 108 V1.1.2 Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Front End Feature Extraction Algorithms; Compression", April 2000.

[4] ETSI, "ETSI ES 202 050 V1.1.1 Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Advanced Front End Feature Extraction Algorithms; Compression", October 2002.

[5] C.Mokbel and O. Collin "Incremental Enrollment of Speech Recognizers", *ICASSP'1999*, Phoenix, USA, pp. 453-456, March 15-19, 1999.

[6] M. Matassoni, M. Omologo, A. Santarelli and P. Svaizer "On the Joint Use of Noise Reduction and MLLR Adaptation for In-Car Hands-Free Speech Recognition", *ICASSP'2002*, Orlando, Florida, USA, pp. 289-292, May 17-13, 2002.