RECOGNITION EXPERIMENTS WITH THE SPEECHDAT-CAR AURORA SPANISH DATABASE USING 8KHZ- AND 16KHZ-SAMPLED SIGNALS

Climent Nadeu and Marta Tolos

TALP Research Center
Universitat Politècnica de Catalunya, Barcelona, Spain
Email: climent@talp.upc.es

ABSTRACT

Like the other SpeechDat-Car databases, the Spanish one has been collected using a 16kHz sampling frequency, and several microphone positions and environmental noises. In this work, we aim at clarify whether there is any advantage in terms of recognition performance by processing the 16kHz-sampled signals instead of the usual 8kHz-sampled ones. Recognition tests have been carried out within the Aurora experimental framework, which includes signals from both a close-talking microphone and a distant microphone. Our preliminary results indicate that it is possible to get a performance improvement from the increased bandwidth in the noisy car environment.

1. INTRODUCTION

Recently, SpeechDat-Car (SDC) databases have been collected for several European languages, using various microphone positions and environmental noises [1]. A subset of the SDC Spanish database [2] is being used, along with a few others, in the Aurora framework [3] for standardizing an advanced front-end for distributed speech recognition (DSR) that must be robust to the environmental conditions (ETSI STQ WI008). The recognition tests are specified in such a way that speech signals from both a close-talking microphone and a distant microphone are involved [2].

Although the speech signals were collected using a 16kHz sampling frequency, the signals used in the Aurora work are downsampled to the usual 8kHz rate. Nevertheless, the standard may eventually support terminals operating at sampling frequencies of 11kHz and 16kHz, as the already developed DSR standard for clean speech ETSI STQ W007 [4]. Actually, the possibility of using a higher frequency bandwidth is an advantage offered by DSR.

However, it is not clear that the speech information carried by the additional band that is made available by the 16kHz sampling frequency is worth using. The doubt comes not only from the fact that speech intelligibility is not much affected by speech content above ca. 4 kHz, but from the presence of non-speech sounds within that additional band, e.g. music. Actually, the initial test reported in [2] was rather discouraging.

In this work, we use the Aurora subset of the SDC Spanish database, which consists of digit utterances, to train and test a recognition system whose back-end is the one specified in the Aurora framework [6]. In our tests, two alternative front-ends are employed, which differ only in the way the filter-bank energies (FBE) are processed at each frame: the conventional mel-frequency cepstral coefficients (MFCC) [7], and the spectral derivative parameters obtained by frequency filtering (FF) [8]. The effects of compression and transmission of the parameters are not considered in the paper.

2. THE AURORA SUBSET OF THE SDC SPANISH DATABASE[1]

The SpeechDat-Car Spanish database comprises recordings of 600 different sessions from 300 speakers. A session consist of 119 read and spontaneous items recorded by five microphones installed in a car, and for various driving conditions. Signals from four distant microphones were recorded and stored in a mobile platform installed in the car. The signal from the close-talking microphone was transmitted simultaneously by GSM to a fixed platform connected to an ISDN telephone interface.

In this work, we train and test the recognition system with the subset of the SDC Spanish database that is being used by the Aurora consortium, along with similar SDC subset databases of other European languages, for the evaluation of robust DSR front-ends.

The following utterances from the original SDC database were included in that Spanish SDC-Aurora subset:

1) 1 sequence of 10 isolated digits
2) 1 sheet number, 4 digits
3) 1 credit card number, 16 digits
4) 1 PIN code, 6 digits
5) 4 utterances of isolated digits, 1 digit per utterance

[1] The SDC Spanish database has been developed at TALP-UPC within the scope of the SpeechDat-Car project (LE4-8334), sponsored by the European Commission and the Spanish Government. The digit subset of that database which is used in this work has been provided by TALP-UPC to the Aurora consortium for the evaluation of the alternative proposals for standardizing a robust DSR front-end, and it is publicly available through ELRA.

The Spanish SDC-Aurora database contains 4914 recordings (files) and more than 160 speakers. Recordings from the close-talking microphone and from one of the distant microphones are included in it. As in the whole SDC database, the files are categorized into three noisy conditions – quiet, low noisy and high noisy – depending on the driving conditions. Table 1 shows the number of files in each of these three conditions [2]. Note that an utterance recorded by both the close-talking and a distant microphone is stored in two files.

Noise conditions	Number of files	Percentage
quiet	792	16.12%
low noise	2422	49.28%
high noise	1700	34.60%
Total	4914	100%

Table 1 Number of files for each noise condition.

We have studied the six different environmental conditions resulting from combining the three noise conditions and the two positions of the microphones used to record the files. The most immediate conclusion was that the noise is globally low-pass and, therefore, the lower speech bands are strongly affected by the noise. Additionally, it was interesting to notice that, for both the high and low noise conditions, there is a spectral peak of noise near 7.5kHz in a number of files, which can be as high as -3dB with respect to the speech spectral maximum. It is also worth mentioning there is a relatively high number of files with background music noise.

3. AURORA EVALUATION SPECIFICATIONS FOR THE SDC-AURORA SPANISH DATABASE

Like for the other SDC-Aurora databases, three different experiments are defined in [2], depending on the degree of matching between training and testing. Baseline recognition results for the three experiments are also given.

3.1. The three experiments

In the well-matched (WM) experiment, 70% of all 4914 files (so including both the close-talking and the distant microphone versions) have been used for training and 30% for testing. There are about 1800 repetitions of each digit for training and about 800 for testing.

In the medium-mismatch (MM) experiment, all the distant microphone recordings from quiet and low noisy conditions are used for training and all the distant microphone recordings from the high noisy condition are used for testing. By using this configuration, 1607 files are employed for training and 850 files for testing. There are about 800 repetitions of each digit for training and about 400 for testing.

In the high-mismatch (HM) experiment, 70% of the close-talking microphone recordings from all conditions are used for training (1696 files), and 30% of the distant microphone recordings from low and high noisy conditions for testing (631 files). There are about 850 repetitions of each digit for training and about 350 for testing.

3.2. Baseline Results

To generate the baseline results with the Spanish SDC-Aurora database reported in [2], the ETSI STQ W007 standard front-end for clean speech, which is based on the mel-cepstrum parameterization [4], was used with the same back-end setup from [6].

Table 2 shows the digit recognition accuracy for each of the three experiments. Note that, for this database, the microphone position is a much more important unmatching factor than the noise condition, since the HM accuracy is much lower than the MM one.

	Accuracy
WM	86.85%
MM	73.74%
HM	42.23%

Table 2 Baseline recognition accuracy for the three experiments of the Spanish SDC-Aurora database.

4. SPEECH PARAMETERIZATION TECHNIQUES USED IN THE TESTS

In this paper, we will mainly report results with mel-cepstrum, which is the basis of the speech representation proposed in the standard ETSI STQ W007 for DSR [4], but a few results with the alternative frequency filtering (FF) technique [8] will also be presented. The FF parameters show a property that it is relevant to the use of different sampling frequencies. In fact, as the FF parameters lie in the frequency domain, they allow of a straightforward change from a sampling frequency to another in the own parameter domain.

4.1. Modified standard mel-cepstrum

In the mel-cepstrum parameterization, a set of filter-bank energies (FBE) is computed by a weighted integration of the energy within each of the Q mel-scaled sub-bands, and then those FBEs are transformed to the cepstral domain through a DCT, resulting in the mel-frequency cepstral coefficients (MFCC). In our experiments, the front-end has been implemented with HTK, so the distribution of the triangular half-overlapped weighting functions used to compute the mel-scaled FBEs is slightly different from the one of the MFCC standard front-end for clean speech ETSI STQ WI007.

Additionally, we have performed a few changes with respect to that standard since they seem to improve its performance for noisy speech [9]: no pre-emphasys; square magnitude of the DFT values instead of the magnitude itself; and a Hamming window 30 ms long instead of 25 ms (the frame rate is kept: 10 ms). To compute the delta and acceleration features, we employed the same time filters as in the standard MFCC front-end. No additional time filtering like cepstral mean subtraction was used.

4.2. Frequency filtering

Recently, spectral derivative-like parameters, obtained by filtering along the frequency variable the sequence of spectral energies, have been successfully used in both clean and noisy HMM speech recognition [8][9]. They are a set of almost uncorrelated parameters which actually represent an alternative to the cepstral coefficients, and have the additional advantage of lying in the frequency domain.

In the usual implementation, if the logarithmic FBE of the *k*-th sub-band is denoted by *S(k)*, the FF parameters are computed by

$$S_{FF}(k) = S(k+1) - S(k-1) \quad (1)$$

i.e. by filtering the sequence of FBEs with the filter $z\text{-}z^{-1}$.

5. RECOGNITION RESULTS

We will begin with the experiments performed with SpeechDat-Car with the usual sampling frequency, i.e. 8 kHz, and then we will try if an improvement can be obtained using the 16 kHz signals. Analogously, we will start with the ETSI's standard MFCC front-end, changing it in order to optimize its performance in the SDC-Aurora framework, and then we will present results with the alternative parameterization.

5.1. Results for 8 Khz signals

The recognition tests have been carried out with the modified standard mel-cepstrum parameterization presented in Section 4.1. Like in the standard front-end, Q=23 sub-bands, M=12 cepstral coefficients, and an entire band ranging from 64 Hz to 4000 Hz were used for the 8kHz-sampled signals. Table 3a presents the recognition results. Notice they are better than those reported in Section 3 for the clean-speech standard front-end for both MM and HM experiments, but not for the WM one.

The results in Table 3a show a large number of insertions for any of the three experiments. We have observed that many insertions are due to the fact that non-speech segments of high energy are often mistaken as digits, mostly the *uno* (one) digit. Consequently, we decided to remove the static frame energy from the parameter set, though keeping its two dynamic versions. The corresponding results are shown in Table 3b. By comparing the HM results from both Tables 3a and 3b, a strong reduction of insertions can be noticed due to the removal of the static energy. The substitutions are also noticeably reduced, and, eventhough the deletions have increased, the effect in terms of word accuracy is a strong increment. The same kind of changes, although less accentuated, can be observed in the MM experiment, and no meaningful changes are observed in the HM experiment between both parts of Table 3.

If we do not constraint our experimental work to keep exactly the same back-end from Aurora but we play with the word insertion penalty parameter (p in HTK) in the Viterbi decoding, it is possible to get a further recognition performance improvement. However, as WM and HM show an opposite deletion-insertion balance, to get a noticeable improvement, a different p value should be allowed for each of the three experiments.

	Accuracy	Del	Subs	Ins
WM	86.08%	132	361	628
MM	75.72%	195	295	613
HM	47.76%	204	788	745

(a)

	Accuracy	Del	Subs	Ins
WM	86.51%	128	348	611
MM	79.71%	227	275	420
HM	74.23%	480	292	85

(b)

Table 3. Recognition results for the 8kHz-sampled signals using the modified standard MFCC: (a) with static frame energy, (b) without it.

5.2. Results for 16 Khz signals

In the following, we are going to report results with 16 kHz sampling rate. In this case, the entire band was chosen from 64 Hz to 7000 Hz, to discard the noise spectral peak around 7.5 kHz, and also because the band above 7 kHz is not meaningful for speech intelligibility. Other changes were made to adapt the MFCC front-end to the new sampling frequency: 1) the number of sub-bands Q was increased from 23 to 30 since, due to the non-uniformity of the mel scale, 7 c; and 2) the number of cepstral coefficients M was set to 15, in correspondence with the increased number of sub-bands. Again, only the first and second temporal derivatives of the frame energy were included, not the static energy itself, thus having a total number of 47 spectral parameters, 9 more than for 8kHz-sampled signals.

	Accuracy	Del	Subs	Ins
WM	87.56%	110	296	596
MM	79.90%	171	288	454
HM	68.57%	341	402	302

(a)

	Accuracy	Del	Subs	Ins
WM	91.17%	219	253	239
MM	83.82%	354	329	152
HM	70.38%	576	306	103

(b)

Table 4. Recognition results for the 16kHz-sampled signals using the modified standard MFCC without the frame energy : (a) p=0 (standard), (b) p=-50.

Results with this parameterization setup are shown in Table 4a. From them, we can not state that there is an improvement by using 16 kHz; actually, there is a significant loss in the HM experiment. However, we can notice by comparing both tables that they show a different insertion-deletion balance, especially for the experiment HM. Therefore, the word insertion penalty p can be tuned to increase the average accuracy of the whole set of three experiments. Using a -50 value for all the three experiments, a remarkable improvement in WM and MM, and a slight improvement in HM are obtained, as shown in Table 4b.

We have also performed tests with the FF parameterization, computing the FBEs exactly in the same way of the last MFCC test for 16 kHz (Table 4b), but using Q=18 bands, since the four additional bands cover the range from 4 to 7 kHz. Note that, unlike for MFCC, no truncation of the parameter vector is performed in the FF parameterization (for MFCC, Q was 30). Thus, instead of applying the DCT, those FBEs were frequency-filtered with the filter $z\text{-}z^{-1}$. The two endpoints, which actually are the 2nd and the 17th sub-band energies, were removed, since the former is highly corrupted by the car noise, and the latter corresponds to a band around 5-6 kHz, which does not carry much speech information. Thus, 16 static spectral parameters and their temporal derivatives (same filters than for MFCC) were used as observation vector, which includes 48 elements, one more than in the MFCC case.

Results for the FF parameterization are reported in Table 5 with p=-50 like in the MFCC results from Table 4b. By observing both tables, we can see that FF improves MFCC in both MM and HM experiments, and lies close to it in the WM experiment. The insertion-deletion balance is similar for both tables of results.

	Accuracy	Del	Subs	Ins
WM	90.14%	277	218	299
MM	84.39%	405	163	141
HM	73.92%	447	339	81

Table 5. Recognition results for the 16kHz-sampled signals using FF parameters and p=-50.

5.3. Summary

The above preliminary results are summarized in Figure 1, where the best reported MFCC results are depicted, i.e. MFCC for 8kHz corresponds to Table 3b, and MFCC for 16kHz corresponds to Table 4b. Note that, using 16kHz, the results are improved with respect to 8kHz, except for the HM experiment, where they are similar.

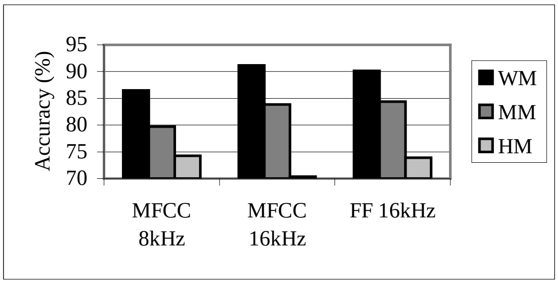

Figure 1. Summary of the results shown in Tables 3-5.

6. DISCUSSION AND FURTHER TESTING

A change in the front-end has an effect on the number of word insertions and its balancing with the number of word deletions. If, due to front-end comparison purposes, no change is allowed in the back-end, that effect may be misleading in deciding the most effective front-end. For example, in our tests, we would not have been noticed a clear improvement by using the 16kHz-sampled signals if the word insertion penalty had not been tuned.

Concerning the parameterization type, FF shows the potential of outperforming mel-cepstrum, as it has already been observed with other databases and recognition tasks [9]. A major advantage of the FF parameters is the fact that, as they lie in the frequency domain, they permit a straightforward change from a sampling frequency to another in the own parameter domain. Recognition tests with 8kHz, 16kHz and 11kHz will be performed with these spectral derivative parameterizations by using HMMs trained from signals of any of those sampling frequencies.

From these presented preliminary results, it appears that a larger bandwidth can yield a higher recognition accuracy, although an increase of the number of features has to be accepted. This advantage should be more apparent with phone-based large vocabulary speech recognition tests.

7. ACKNOWLEDGMENTS

The authors would like to thank to Dusan Macho for his assistance, especially at the beginning of the work. This work has been partially supported by CICYT under projects TIC2000-1005-C03-01 and TIC2000-1735-C02-01.

8. REFERENCES

[1] A. Moreno, et al. "SPEECHDAT-CAR. A Large Speech Database for Automotive Environments", Proc. II LREC, Athens, June 2000.

[2] D. Macho, “Spanish SDC-Aurora Database for ETSI STQ Aurora WI008 Advanced DSR Front-End”, UPC, Barcelona, Dec. 2000.

[3] D. Pearce, “Enabling New Speech Driven Services for Mobile Devices: An overview of the ETSI standards activities for DSR Front-ends", Applied Voice Input/Output Society Conf. (AVIOS2000), San José, CA, May 2000.

[4] "ETSI ES 201 108 V1.1.2 Distributed Speech Recognition; Front-end", April 2000.

[5] A. Moreno, “SpeechDat-Car Spanish Database. SpeechDat-Car Project LE4-8334”, UPC, Barcelona, May 2001.

[6] "Baseline Results for Subset of SDC Finnish Database for ETSI STQ WI008 Advanced Front-End Evaluation", STQ Aurora DSR Working Group, Document AU/225/00.

[7] S.B. Davis, P. Mermelstein, IEEE Trans. on ASSP, Vol. ASSP-28, No.4, pp. 357-366, August 1980.

[8] C. Nadeu, et al., Proc. Eurospeech, 1995, pp. 1381-84.

[9] C. Nadeu, et al., Speech Communication – Noise Robust ASR, Vol. 34, April 2001, pp. 93-114.