STQ AURORA DSR WORKING GROUP

Title: **Experimental Framework for the Performance Evaluation of Speech Recognition Front-ends on a Large Vocabulary Task**

Author: **Günter Hirsch, Ericsson**

Date: **31th October 2002**

Version: **2.1**

1 Summary

This report defines and describes the experiments for the recognition of a large vocabulary with the purpose of evaluating and comparing the robustness of feature extraction algorithms or complete recognition systems.
The WSJ0 (**W**all **S**treet **J**ournal) database (available from LDC under the name CSR-I (WSJ0) complete) is chosen as basis for the experiments. The recognition of a 5000 word vocabulary is selected as task as it has been also used for the ARPA evaluations on continuous speech recognition. Besides the original data sampled at 16 kHz a second version of the data is created by downsampling the data from 16 kHz to 8 kHz. The reason for this is the required ability of the front-end to process speech at both sampling rates. Most of today's telecommunication terminals operate in the frequency range up to 4 kHz. But future speech services will aim at a higher speech intelligibility and a higher subjective speech quality by analysing speech at a higher bandwidth up to 8 kHz.

The WSJ data have been recorded with a Sennheiser microphone and with a second microphone in parallel. The recordings with the second microphone are used for enabling recognition experiments with different frequency characteristics in the transmission channel. To evaluate the robustness in the presence of background noise 6 different noises are artificially added to the original data recorded with the different microphones. These noises represent realistic scenarios of application environments for mobile telephones. Noises have to be also available at 8 and 16 kHz sampling rate. The methods are described for estimating the energy of speech and noise signals and adding both signals at a desired SNR.

Details are presented about all training and test sets for the recognition experiments. As already used in earlier experiments we define two training modes. One takes clean data only to train the recognizer. In the second mode clean as well as noisy data are applied to perform a multi-condition training that takes into account also the influence of recording the speech data with the different microphones.

The predefined ARPA test set is selected here to perform the recognition on a 5000 word vocabulary. 6 further noisy versions of this test set are created by adding individual noises at a certain range of SNRs. Noise adding is also applied to the test data that have been recorded with the second microphone. In total 14 versions of the test set are available. Recognition results can be separately determined for each condition.

Recognition results are presented for the standardized ETSI mel-cepstrum front-end (ES 201 108) when applying the recognizer of Mississippi State University at the backend.

2 Noisy speech database

The WSJ0 database is taken as basis for all experiments. Besides the original data sampled at 16 kHz with a precision of 16 Bit a second version is created by down sampling the original data to 8 kHz.

2.1 Filtering

An additional filtering is applied to consider the realistic frequency characteristics of terminals and equipment in the telecommunication a rea. Two “standard” frequency characteristics are used which have been defined by the ITU. The abbreviations G.712 and P.341 have been introduced as reference to these filters. Their frequency responses are shown in figure 1. The G.712 characteristic is defined for the frequency range of the usual telephone bandwidth up to 4 kHz and has a flat characteristic in the range between 300 and 3400 Hz. P.341 is defined for the frequency range up to 8kHz and represents a band pass filter with a very low cut off fre quency at the lower end and a cut off frequency at about 7 kHz at the higher end of the bandpass.

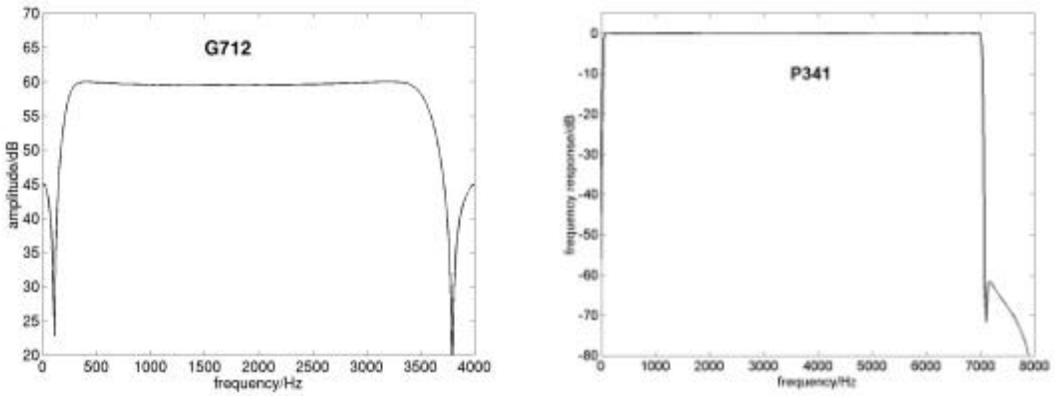

Figure 1: Frequency responses of G.712 and P341 filter

2.2 Noise Adding

Noise is artificially added to the clean WSJ0 data that have been recorded at a high SNR. To add noises at a desired SNR (signal-to-noise ratio) we define SNR as the ratio of signal to noise energy after filtering both signals with the G.712 filter characteristic. This can be applied to data sampled at 8 or 16 kHz. In the practical realization we use the available ITU tools for downsampling the data from 16 to 8 kHz and for applying the G712 filtering to data sampled at 8 kHz.

The speech energy is determined on basis of the ITU recommendation P.56 also available as part of the ITU software package. The noise energy is calculated as RMS value with the same software where a noise segment of same length than the speech signal is randomly cut out of the whole noise recording. We assume duration of the noise signal much longer than that of the speech signal.
The level of the speech signal is not changed as long as no overflow occurs in the Short-integer range. Based on the desired SNR the attenuation factor is calculated to multiply the noise samples before adding them to the speech samples. The speech level is only changed in case of an overflow but this situation did not occur for the WSJ database preparation.

Noise signals are selected to represent probable application scenarios for telecommunication terminals. Noises have been chosen for the following environments or conditions:

?? Car
?? Crowd of people (babble)
?? Restaurant

?? Street
?? Airport
?? Train station

Some noises are fairly stationary like e.g. the car noise. Others contain non-stationary segments like e.g. the recordings on the street and at the airport.
Recordings in these environments are available e.g. from the NTT noise database. The NTT noises have been sampled at 22.05 kHz so that versions can be easily created at 8 and 16 kHz sampling rate. The long-term spectra of the noise signals a re shown in figure 2 for the 6 noise conditions. These power density spectra have been estimated with the Welch method.

To add noise in case of 16 kHz sample rate speech and noise are downsampled to 8 kHz and filtered with the G.712 characteristic first to determine the weighting factor for the noise. Then speech and noise sampled at 16 kHz are filtered with the P.341 characteristic before adding them at the desired SNR.

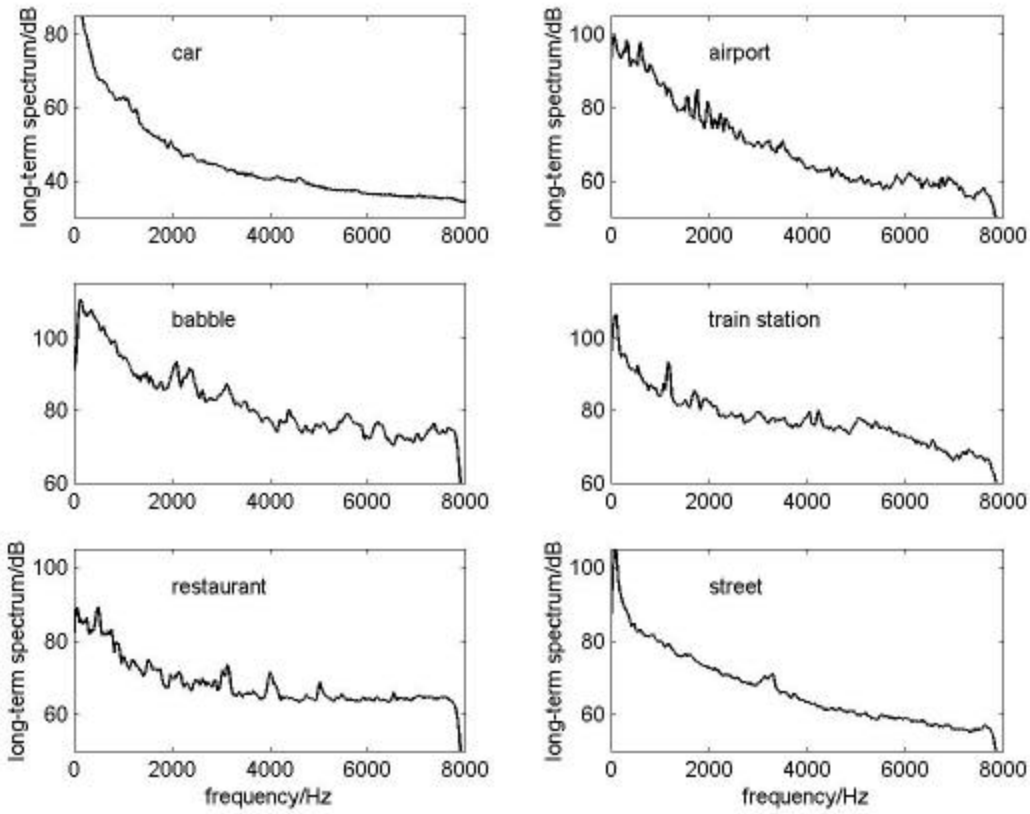

Figure 2: Estimated power density spectra for the 6 noise signals

3 Definition of training and test sets

The WSJ0 database consists of speech that has been recorded with two microphones in parallel. The first one is a close-talking microphone of type Sennheiser HMD414. No regulation exists about the choice of the second microphone which has been e.g. a desk mounted microphone. Most recordings consist of read texts from the Wall Street Journal.

In the ARPA evaluations a set of about 7200 utterances (~12 hours of speech) has been selected for the training. 7138 recordings are available on the CDs. These data are taken from the recordings with the Sennheiser microphone. We consider the same set for our training mode on **clean** data only. We refer to this training mode by naming it "training_clean_sennh". To define the multi-condition training we take the same set of 7138 utterances with half of them recorded with the Sennheiser microhone and the other half recorded with the second microphone. A variety of different microphones has been used as second microphone (e.g. Sony ECM-55, Sony ECM-50PS, Crown PZM-6FS, Crown PCC-160, RadioShack omni-electret, Nakamichi CM100, AT&T 5400 cordless phone, Panasonic KXT2365 speaker phone, …). No further noise is added to one fourth of each of the two subsets. To the rest of the data noise is artificially added. The type of noise is randomly chosen out of 6 noises in total and at a randomly chosen SNR between 10 and 20 dB. Goal is an equal distribution of noise types and of SNRs over the whole range. Thus the average SNR is 15 dB. We refer to this training mode by naming it "training_multicondition". Figure 3 gives an overview about the data sets for the two training modes.

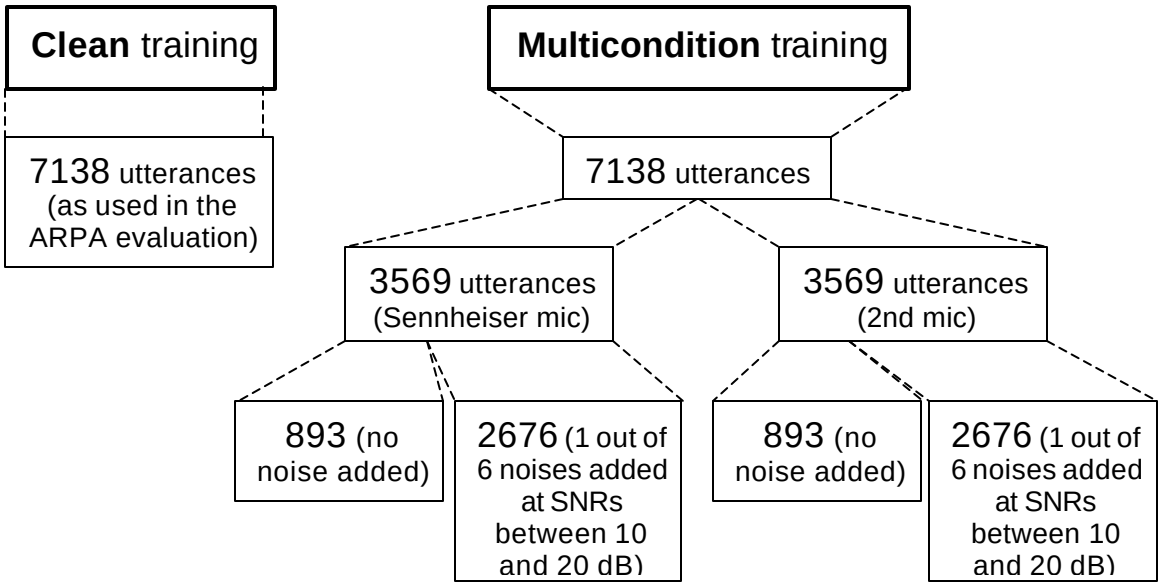

Figure 3: Data sets for the two training modes

A set of 330 utterances has been designated in the ARPA evaluation to perform a baseline recognition on the 5000 word vocabulary. This test includes the usage of a closed vocabulary bigram language model as supplied by Lincoln. The 330 utterances contain recordings from 8 speakers with about 40 utterances per speaker.

An alternative experiment has been proposed for the ARPA evaluation by using the same set of 330 utterances but recorded with the second microphone. Three different microphones have been used to record the 8 speakers in individual sessions. The same microphone has been used throughout each session. The frequency characteristics of these microphones have been analyzed in a differential way by comparing the parallel recordings with the Sennheiser microphone and the second microphone. A long term spectral analysis has been applied to the two versions of the same utterance. Comparing the two long-term spectra results in a rough estimate of the differential frequency response for each microphone in comparison to the Sennheiser microphone. The three differential frequency responses are plotted in figure 4.

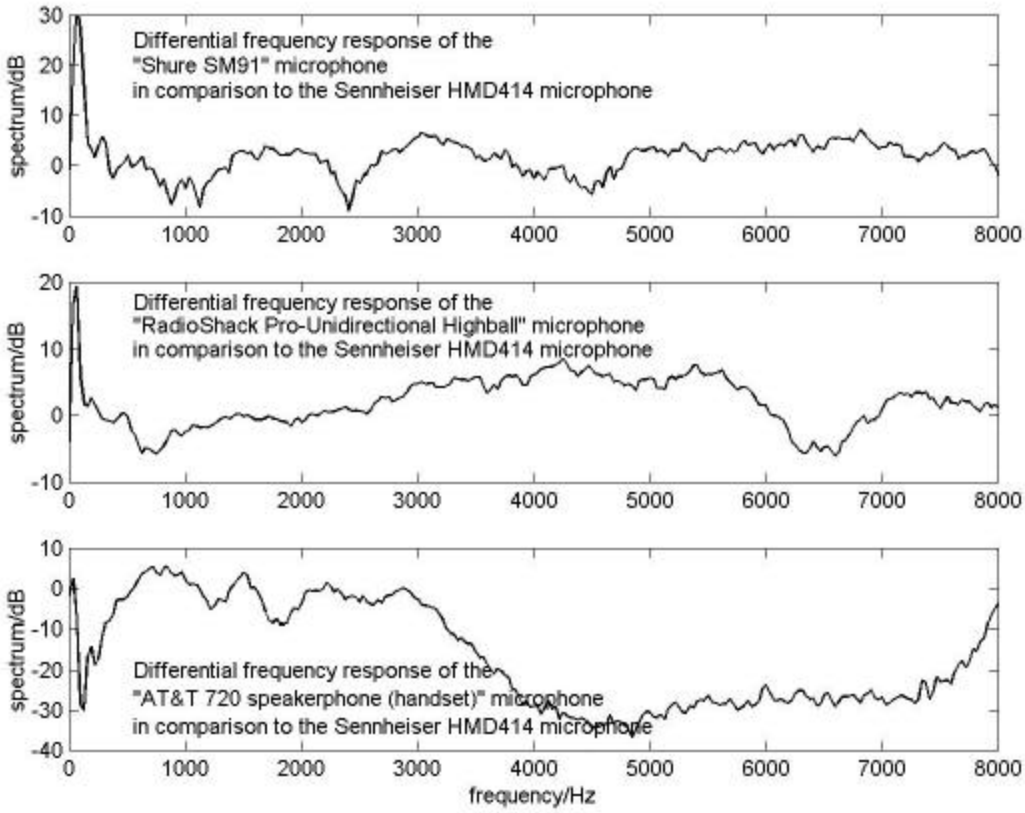

Figure 4: Differential frequency responses of different microphones

The peaks at very low frequencies should be ignored and are probably due to the highpass characteristics of both microphones. Having a big attenuation in this frequency region will cause an unreliable analysis result.
The highest influence on the spectrum has the AT&T handset microphone. It limits the speech signal to telephone bandwidth.

These two sets of speech data recorded either with the Sennheiser microphone or with one of the three other microphones are available on the CDs.
6 further versions are created for each of the two sets so that we have 14 sets in total for testing. The 6 additional versions are created by artificially adding one type of noise for each version at a randomly chosen SNR between 5 and 15 dB. The goal is again an equal distribution of SNRs over the whole range so that the average SNR is 10 dB.
Thus we get 6 additional test sets for the speech data recorded with the Sennheiser microphone and 6 additional sets for the recordings with the second microphone. Each of these additional sets is characterized by being contaminated by one of the 6 noise types at an average SNR of 10 dB.

Figure 5 gives an overview about all 14 sets of test data.

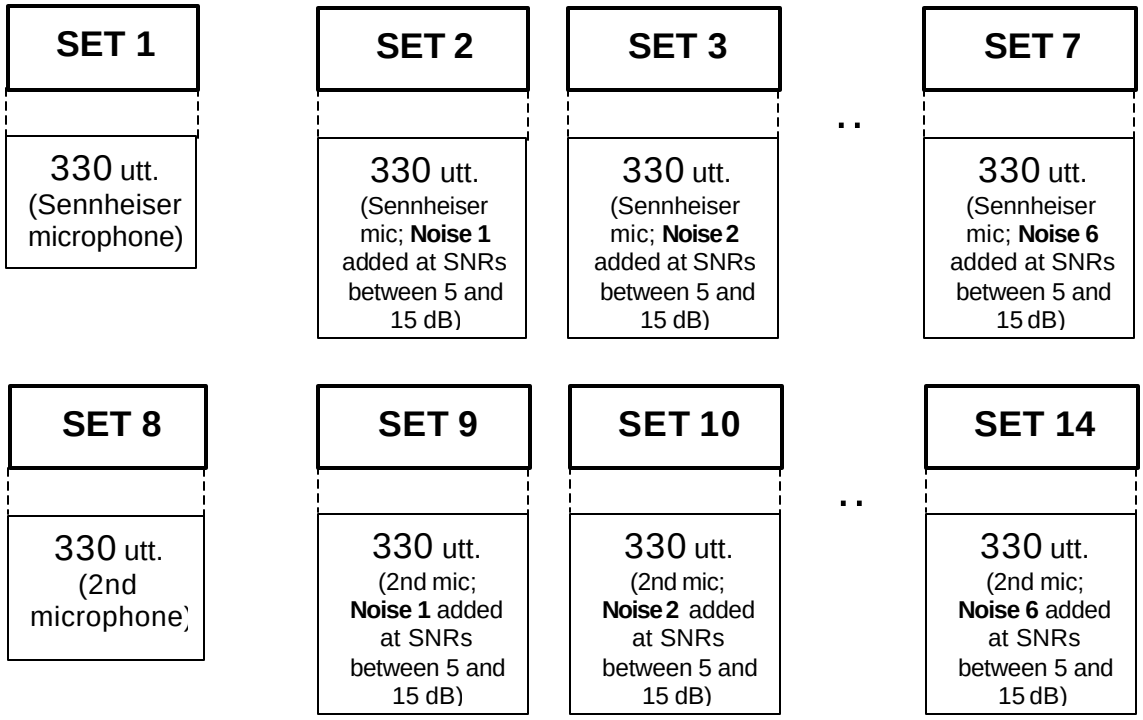

Figure 5: **Data sets for recognition**. Note that although there are 330 utterances in each test set provided in the database a second set of file lists is also provided with 166 of these in each test set.

The recognition is done with the recognizer from Mississippi State University [1],[2] to investigate the influence and enable the comparison of different frontends. During the setup of these recognition experiments it turned out that a high computational performance is needed to create the recognition results for all utterances of all 14 conditions. It was decided to use only half of the 330 utterances in each test condition. While all 330 utterances for all conditions are available on the CDs special list files exist that define the reduced set of utterances used for the experiments in the Aurora group. File lists for these test sets with 166 utterances in each set are also defined on the CDs for the Aurora-4 database. It has been verified that the reduced number of utterances has no major influence on the achieved recognition performance in comparison to using the full set of 330 utterances.

As output 14 recognition results are created in total. The results for the two sets (sets 1 and 8 in figure 5) without noise added can also be taken to compare the performance against existing results of ARPA evaluations. All other results create a performance measure for individual noise scenarios as e.g. the car environment.

Considering the sampling at 8 and 16 kHz we have 4 sets of training data and 28 sets of test data. We determine 28 recognition results at each sampling rate by recognizing the 14 test sets in both training modes.

4 Recognition results

Recognition results are presented for the ETSI DSR mel-cepstrum standard together with its associated compression algorithm. Results are based on the official Aurora selection test sets with 166 utterances in each test set.

The word error rates are listed in tables 1 to 4 when applying the first standardized ETSI frontend that is based on Mel-cepstral analysis [3]. Each acoustic vector consists of 39 components containing 12 cepstral coefficients and the logarithmic frame energy together with the corresponding Delta and Delta-Delta coefficients.

Training mode	Test condition						
	Clean	Car	Babble	Restaurant	Street	Airport	Train station
Clean	15,40	49,40	60,60	59,00	57,40	61,90	62,00
Multi-condition	20,70	26,40	38,60	41,60	43,80	41,10	43,40

Table 1: Word error rates (%) for the Sennheiser microphone at a sampling with 8 kHz

Training mode	Test condition						
	Clean	Car	Babble	Restaurant	Street	Airport	Train station
Clean	36,60	59,90	71,60	67,80	72,50	70,20	69,50
Multi-condition	30,90	38,70	47,10	50,10	53,60	47,30	50,70

Table 2: Word error rates (%) for the second microphone at a sampling with 8 kHz

Training mode	Test condition						
	Clean	Car	Babble	Restaurant	Street	Airport	Train station
Clean	14,50	58,40	58,80	53,80	62,50	56,90	65,50
Multi-condition	19,10	23,40	31,70	35,50	35,30	33,10	36,40

Table 3: Word error rates (%) for the Sennheiser microphone at a sampling with 16 kHz

Training mode	Test condition						
	Clean	Car	Babble	Restaurant	Street	Airport	Train station
Clean	53,30	75,10	76,30	68,50	77,80	73,50	75,90
Multi-condition	40,90	47,40	50,30	48,90	54,70	49,30	51,80

Table 4: Word error rates (%) for the second microphone at a sampling with 16 kHz

5 References

[1] N. Deshmukh, A. Ganapathiraju, J. Hamaker, J. Picone and M. Ordowski: "A Public Domain Speech-to-Text System", 6th European Conference on Speech Communication and Technology, Vol. 5, pp. 2127-2130, Budapest, Hungary, September 1999

[2] N. Parihar and J. Picone: "DSR Front End LVCSR Evaluation - Baseline Recognition System Description," Aurora Working Group, European Telecommunications Standards Institute, November 2002

[3] ETSI standard document: “Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithm”, ETSI ES 201 108 v1.1.2 (2000-04), Apr. 2000. http://pda.etsi.org/pda/home.asp?wki_id=9948