ISCA Archive
http://www.isca-speech.org/archive

ASR2000 - Automatic Speech Recognition:
Challenges for the new Millenium
Paris, France
September 18-20, 2000

THE AURORA EXPERIMENTAL FRAMEWORK FOR THE PERFORMANCE EVALUATION OF SPEECH RECOGNITION SYSTEMS UNDER NOISY CONDITIONS

Hans-Günter Hirsch[1)], *David Pearce*[2)]

1) Ericsson Eurolab Deutschland GmbH, Hans-Guenter.Hirsch@eed.ericsson.se
2) Motorola Labs, UK, bdp003@email.mot.com

ABSTRACT

This paper describes a database designed to evaluate the performance of speech recognition algorithms in noisy conditions. The database may either be used for the evaluation of front-end feature extraction algorithms using a defined HMM recognition back-end or complete recognition systems. The source speech for this database is the TIdigits, consisting of connected digits task spoken by American English talkers (downsampled to 8kHz). A selection of 8 different real-world noises have been added to the speech over a range of signal to noise ratios and special care has been taken to control the filtering of both the speech and noise.

The framework was prepared as a contribution to the ETSI STQ-AURORA DSR Working Group [1]. Aurora is developing standards for Distributed Speech Recognition (DSR) where the speech analysis is done in the telecommunication terminal and the recognition at a central location in the telecom network. The framework is currently being used to evaluate alternative proposals for front-end feature extraction. The database has been made publicly available through ELRA so that other speech researchers can evaluate and compare the performance of noise robust algorithms.

Recognition results are presented for the first standard DSR feature extraction scheme that is based on a cepstral analysis.

1. INTRODUCTION

The robustness of a recognition system is heavily influenced by the ability

- to handle the presence of background noise and
- to cope with the distortion by the frequency characteristic of the transmission channel (often described also as convolutional “noise” – although the term convolutional distortion is preferred).

The importance of these issues is reflected by an increasing number of investigations and publications on these topics during the last years. This is again driven by the dependency on robustness in real-life scenarios for the successful introduction of recognition systems. Robustness can be achieved by an appropriate extraction of robust features in the front-end and/or by the adaptation of the references to the noise situation.

To compare the performance of different algorithms the definition and creation of training and test scenarios is needed. A first attempt was the Noisex-92 database [2]. This consists of recordings from one male and one female speaker that have been distorted by artificially adding background noise at different signal-to-noise ratios (SNRs) and in different noise conditions. The vocabulary contains the English digits. The Noisex-92 data can be mainly used to obtain comparable recognition results on the task of speaker dependent isolated word recognition in the presence of additive noise.

A database as well as a recognition experiment is presented in this paper to obtain comparable recognition results for the speaker-independent recognition of connected words in the presence of additive background noise and for the combination of additive and convolutional distortion. The distortions are artificially added to the clean TIDigits database [3].

The noisy database together with the definition of training and test sets can be taken to determine the performance of a complete recognition system. In combination with a predefined set-up of a HTK (Hidden Markov Model Tool Kit) based recognizer [4] it can be taken to evaluate the performance of a feature extraction scheme only.

The comparison of several feature extraction schemes has been the initial reason for the creation of the noisy database and for the definition of a HMM based recognizer. This evaluation is a task of the Aurora working group that belongs to the technical body STQ (**S**peech processing, **T**ransmission and **Q**uality aspects) as ETSI standardization activity. A DSR (**D**istributed **S**peech **R**ecognition) system consists of a front-end in any type of telecommunication terminal and a recognizer as back-end at a central location in the telecom network. Previous work has standardised the DSR front-end and compression based on the Mel-Cepstrum. The current activity is to develop an advanced DSR front-end that will be more robust in noise.

Besides using the artificially distorted TIDigits data the Aurora evaluation will also be based on recognition experiments with recordings in the noisy car environment. Subsets of the SpeechDat-car data collection [5] are going to be taken for this further evaluation. Thus the influence will be studied of looking at different languages and comparing the results of recordings under real noise conditions with the ones achieved on artificially distorted data.

2. NOISY SPEECH DATABASE

The TIDigits database is taken as basis. This part is considered that contains the recordings of male and female US-American adults speaking isolated digits and sequences of up to 7 digits. The original 20kHz data have been downsampled to 8 kHz with an "ideal" low-pass filter extracting the spectrum between 0 and 4kHz. These data are considered as "clean" data. Distortions are artificially added.

Filtering

An additional filtering is applied to consider the realistic frequency characteristics of terminals and equipment in the telecommunication area. Two "standard" frequency characteristics are used which have been defined by the ITU [6]. The abbreviations G.712 and MIRS have been introduced as reference to these filters. Their frequency responses are shown in Figure 1.

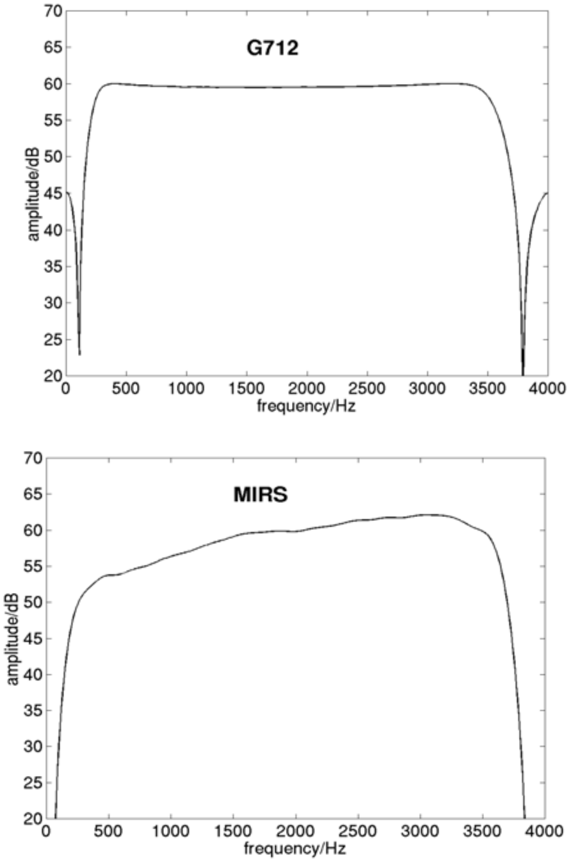

Figure 1: Frequency responses of G.712 and MIRS filter

The major difference is a flat curve of the G.712 characteristic in the range between 300 and 3400 Hz where the MIRS shows a rising characteristic with an attenuation of lower frequencies. MIRS can be seen as a frequency characteristic that simulates the behavior of a telecommunication terminal, which meets the official requirements for the terminal input frequency response as specified e.g. in the technical specification GSM 03.50 [7]. Both types of filtering are realized with the corresponding modules of the ITU STL96 software package.

Noise Adding

Noise is artificially added to the filtered TIDigits. To add noises at a desired SNR (signal-to-noise ratio) the term SNR has to be defined first because it is dependent on the selected frequency range. We define it as the ratio of signal to noise energy after filtering both signals with the G.712 characteristic. This assumes the recording of speech and noise signals with good and similar equipment that does not influence the spectrum of the original signals.

To determine the speech energy we apply the ITU recommendation P.56 [8] by using the corresponding ITU software. 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. This happens only for the worst SNR of –5dB and in less than 10 cases in total for all noises.

Noise signals are selected to represent the most probable application scenarios for telecommunication terminals. Noises have been recorded at different places:

- Suburban train
- Crowd of people (babble)
- Car
- Exhibition hall
- Restaurant
- Street
- Airport
- Train station

The long-term spectra of all noises are shown in figure 2. A dynamic range of 40 dB is shown in all plots even though the absolute levels are different. These spectra do not tell anything about the stationarity of the corresponding signals. Some noises are fairly stationary like e.g. the car noise and the recording in the exhibition hall. Others contain non-stationary segments like e.g. the recordings on the street and at the airport.

The major part of the signals' energy concentrates in the low frequency region. From the spectral viewpoint some noise signals seem to be quite similar even though they have been recorded in totally different environments.

The noise signals are added to the TIDigits at SNRs of 20dB, 15dB, 10dB, 5dB, 0dB and –5dB.

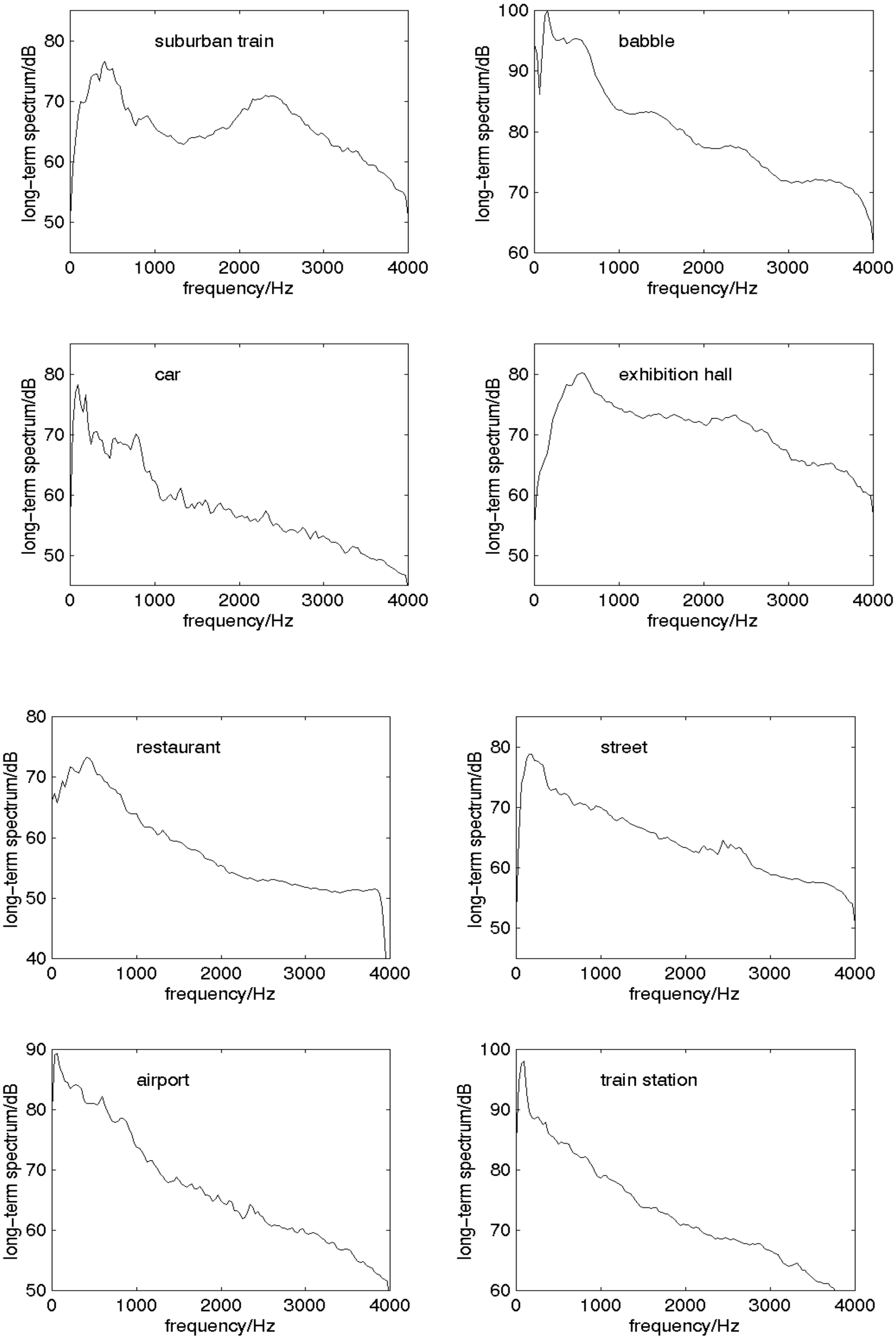

Figure 2: Long-term spectra of all noise signals

In the case of MIRS filtering in combination with additive noise both the speech and noise are filtered with the G.712 characteristic first to determine the weighting factor for the noise to achieve the desired SNR. Then speech and noise are filtered with the MIRS characteristic before adding them using this weight..

3. DEFINITION OF TRAINING AND TEST SETS

Two training modes are defined as

- training on **clean** data only and as
- training on clean and noisy (**multi-condition**) data.

The advantage of training on clean data only is the modeling of speech without distortion by any type of noise. Such models should be suited best to represent all available speech information. The highest performance can be obtained with this type of training in case of testing on clean data only. But these models contain no information about possible distortions. This aspect can be considered as advantage of multi-condition training where distorted speech signals are taken as training data. This leads usually to the highest recognition performance when training and testing are done in the same noise condition. The question arises whether the performance gain can also be achieved for a different type of noise or a different SNR than seen during training.

For the first mode 8440 utterances are selected from the training part of the TIDigits containing the recordings of 55 male and 55 female adults. These signals are filtered with the G.712 characteristic without noise added.

The same 8440 utterances are taken for the second mode too. They are equally split into 20 subsets with 422 utterances in each subset. Each subset contains a few utterances of all training speakers. The 20 subsets represent 4 different noise scenarios at 5 different SNRs. The 4 noises are suburban train, babble, car and exhibition hall. The SNRs are 20dB, 15dB, 10dB, 5dB and the clean condition. Speech and noise are filtered with the G.712 characteristic before adding.

Three different test sets are defined. 4004 utterances from 52 male and 52 female speakers in the TIDigits test part are split into 4 subsets with 1001 utterances in each. Recordings of all speakers are present in each subset. One noise signals is added to each subset of 1001 utterances at SNRs of 20dB, 15dB, 10dB, 5dB, 0dB and –5dB. Furthermore the clean case without adding noise is taken as seventh condition. Again speech and noise are filtered with the G.712 characteristic before adding.

In the first test set, called **test set A**, the four noises suburban train, babble, car and exhibition hall are added to the 4 subsets. In total, this set consists of 4 times 7 times 1001 = 28028 utterances. It contains the same noises as used for the multi-condition training which lead to a high match of training and test data.

The second test set, called **test set B**, is created in exactly the same way, but using the four different noises, namely restaurant, street, airport and train station. In this case there exists a mismatch between training and test data also for the multi-condition training. This will show the influence on recognition when considering different noises than the ones used for training.

The third test set, called **test set C**, contains 2 of the 4 subsets with 1001 utterances in each. This time speech and noise are filtered with a MIRS characteristic before adding them at SNRs of 20dB, 15dB, 10dB, 5dB, 0dB and –5dB. Again the clean case without additive noise is considered as seventh condition. Suburban train and street are used as additive noise signals. This set is intended to show the influence on recognition performance when a different frequency characteristic is present at the input of the recognizer.

4. HTK REFERENCE RECOGNIZER

The reference recognizer is based on the HTK software package version 2.2 from Entropic. The training and recognition parameters are defined to compare the recognition results when applying different feature extraction schemes. Some parameters, e.g. the number of states per HMM model, have been chosen with respect to the commonly used frame rate of 100 Hz (frame shift = 10ms). The recognition of digit strings is considered as task without restricting the string length.

The digits are modeled as whole word HMMs with the following parameters:

- 16 states per word (according to 18 states in HTK notation with 2 dummy states at beginning and end)
- simple left-to-right models without skips over states
- mixture of 3 Gaussians per state
- only the variances of all acoustic coefficients (No full covariance matrix)

As an initial starting point a vector size of 39 is defined by using 12 cepstral coefficients (without the zeroth coefficient) and the logarithmic frame energy plus the corresponding delta and acceleration coefficients. The vector size may be changed when testing with an alternative front-end that generates a different number of features.

Two pause models are defined. The first one called “sil” consists of 3 states with a transition structure as shown in Figure 3. This HMM shall model the pauses before and after the utterance. A mixture of 6 Gaussians models each state.

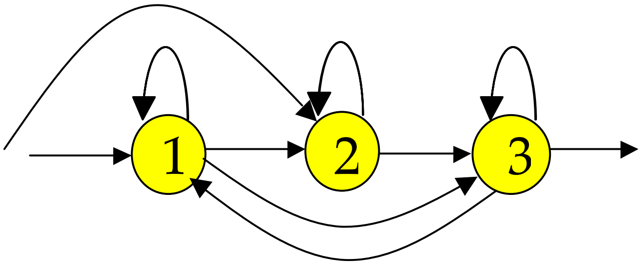

Figure 3: Possible transitions in the 3-state pause model “sil”

The second pause model called “sp” is used to model pauses between words. It consists of a single state which is tied with the middle state of the first pause model.

The training is done in several steps by applying the embedded Baum-Welch reestimation scheme (HTK tool HERest):

- Initialize all word models and the 3-state pause model with the global means and variances (determined with HcompV). Word and pause models contain only 1 Gaussian per state in this initialization stage.
- Three iterations of Baum-Welch reestimation with the pruning option –t of HERest set to 250.0 150.0 1000.0
- Introduce the interword pause models, increase the number of Gaussians to 2 for the 3-state pause model and apply three further iterations of Baum-Welch reestimation
- Increase the number of Gaussians to 2 for all states of the word models, increase the number of Gaussians to 3 for all states of the pause model and apply three further iterations of Baum-Welch reestimation
- Increase the number of Gaussians to 3 for all states of the word models, increase the number of Gaussians to 6 for all states of the pause models and apply seven further iterations of Baum-Welch reestimation

During recognition an utterance can be modeled by any sequence of digits with the possibility of a “sil” model at the beginning and at the end and a ”sp” model between two digits.

5. AURORA WI007 FRONT-END

As already mentioned in the introduction the definition of the whole experiment was initially caused by a demand of the Aurora DSR standardization activity. It will be used to select a robust front-end as component in telecommunication terminals for the realization of a distributed speech recognition. This selection process is work item WI008 of the Aurora group. The proposers of alternative candidates for the advanced DSR front-end are evaluating its performance on this database as part of the final submissions on 27th October 2000.

Because it was known in advance that the selection and standardization of a robust front-end would need a longer period, another front-end has been standardized first [9] as basis for the immediate realization of DSR applications. This was done as work item WI007.

The Aurora WI007 front-end is a cepstral analysis scheme where 13 Mel frequency cepstral coefficients (MFCCs), including the coefficient of order 0, are determined for a speech frame of 25ms length. The frame shift is 10 ms. Besides the cepstral coefficients the logarithmic frame energy is taken as further acoustic coefficient. Thus each feature vector consists of 14 components in total.

Further details of the cepstral analysis scheme are:

- Signal offset compensation with a notch filtering operation
- Preemphasis with a factor of 0.97
- Application of a Hamming window
- FFT based Mel filterbank with 23 frequency bands in the range from 64 Hz up to half of the sampling frequency

Besides the cepstral analysis a compression scheme is part of the front-end to transfer the acoustic parameters as a data stream with a rate of 4800 Bit/s. Therefore a quantisation scheme is used in the standard [9] to code the 14 acoustic coefficients of each frame with 44 Bits. The quantisation is based on a split vector codebook where the set of 14 vector components is split into 7 subsets with two coefficients in each. There exist 7 codebooks to map each subset of vector components to an entry of the corresponding codebook.

6. RECOGNITION PERFORMANCE

The recognition results are presented in this section when applying the WI007 front-end and the HTK recognition scheme as described above. The MFCC of order 0 is not part of the feature vector that consists of the remaining 13 components as well as of the corresponding delta and acceleration coefficients. Thus a vector contains 39 components in total. Based on those results a relative improvement can be stated for the proposals of the Aurora WI008 activity.

The word accuracy is listed in Table 1 for test set A when applying the multi-condition training. As well known the performance deteriorates for decreasing SNR. The degradation does not significantly differ for the different noises. A performance measure for the whole test set has been introduced as average over all noises and over SNRs between 0 and 20dB. This average performance between 0 and 20dB takes a value of **87.81%** for test set A.

The results for test set B are listed in Table 2 when applying the multi-condition training. The performance degradation is not much worse in comparison to the noises of test set A. The average performance of test set B is **86.27%** for the SNR range between 0 and 20dB. This value shows only a slightly worse performance for the case of noises not seen during training. The noises of the first test set seem to cover the spectral characteristics of the noises in the second test set to a high extent.

For test set C the detailed results are listed in table 3 in case of multi-condition training. The average word accuracy is **83.77%**. Degradation in performance can be seen due to the different frequency characteristic.

SNR/dB	Subway	Babble	Car	Exhibition	Average
clean	98.68	98.52	98.39	98.49	98.52
20	97.61	97.73	98.03	97.41	97.69
15	96.47	97.04	97.61	96.67	96.94
10	94.44	95.28	95.74	94.11	94.89
5	88.36	87.55	87.80	87.60	87.82
0	66.90	62.15	53.44	64.36	61.71
-5	26.13	27.18	20.58	24.34	24.55
Average between 0 and 20dB	88.75	87.95	86.52	88.03	**87.81**

Table 1: Word accuracy as percentage for test set A in multi-condition training

SNR/dB	Restaurant	Street	Airport	Train-station	Average
clean	98.68	98.52	98.39	98.49	98.52
20	96.87	97.58	97.44	97.01	97.22
15	95.30	96.31	96.12	95.53	95.81
10	91.96	94.35	93.29	92.87	93.11
5	83.54	85.61	86.25	83.52	84.73
0	59.29	61.34	65.11	56.12	60.46
-5	25.51	27.60	29.41	21.07	25.89
Average between 0 and 20dB	85.39	87.03	87.64	85.01	**86.27**

Table 2: Word accuracy as percentage for test set B in multi-condition training

SNR/dB	Subway(MIRS)	Street(MIRS)	Average
clean	98.50	98.58	98.54
20	97.30	96.55	96.92
15	96.35	95.53	95.94
10	93.34	92.50	92.92
5	82.41	82.53	82.47
0	46.82	54.44	50.63
-5	18.91	24.24	21.57
Average between 0 and 20dB	83.24	84.31	**83.77**

Table 3: Word accuracy as percentage for test set C in multi-condition training

The recognition results for the three test sets are listed in Tables 4, 5 and 6 when training the recognizer on clean data only.

SNR/dB	Subway	Babble	Car	Exhibition	Average
clean	98.93	99.00	98.96	99.20	99.02
20	97.05	90.15	97.41	96.39	95.25
15	93.49	73.76	90.04	92.04	87.33
10	78.72	49.43	67.01	75.66	67.70
5	52.16	26.81	34.09	44.83	39.47
0	26.01	9.28	14.46	18.05	16.95
-5	11.18	1.57	9.39	9.60	7.93
Average between 0 and 20dB	69.48	49.88	60.60	65.39	**61.34**

Table 4: Word accuracy as percentage for test set A in clean training

SNR/dB	Restaurant	Street	Airport	Train-station	Average
clean	98.93	99.00	98.96	99.20	99.02
20	89.99	95.74	90.64	94.72	92.77
15	76.24	88.45	77.01	83.65	81.33
10	54.77	67.11	53.86	60.29	59.00
5	31.01	38.45	30.33	27.92	31.92
0	10.96	17.84	14.41	11.57	13.69
-5	3.47	10.46	8.23	8.45	7.65
Average between 0 and 20dB	52.59	61.51	53.25	55.63	**55.74**

Table 5: Word accuracy as percentage for test set B in clean training

SNR/dB	Subway(MIRS)	Street(MIRS)	Average
clean	99.14	98.97	99.05
20	93.46	95.13	94.29
15	86.77	88.91	87.84
10	73.90	74.43	74.16
5	51.27	49.21	50.24
0	25.42	22.91	24.16
-5	11.82	11.15	11.48
Average between 0 and 20dB	66.16	66.11	**66.14**

Table 6: Word accuracy as percentage for test set C in clean training

The performance is much worse in comparison to the multi-condition training. Besides the ability of training the noise characteristics as part of the word models a further advantage in multi-condition training is the possibility of training the noise characteristics as contents of the pause models.

The recognition accuracy is worse for those noises (babble, restaurant, airport, train) which contain non-stationary segments. The reason for seeing this effect only in the clean training may be the ability of partly training the non-stationary noise characteristics as contents of the pause model in multi-condition training.

An unexpected result is the improvement of the word accuracy when filtering with the MIRS characteristic instead of G.712 in case of street noise added. In clean training mode it seems to be of advantage to attenuate the components of low frequencies where a major part of the noise energy can be found for the street noise.

All average results for the Aurora WI007 Mel-Cepstrum front-end are summarized in Table 7.

training mode	test set A	test set B	test set C
multi-condition	**87.81**	**86.27**	**83.77**
clean	**61.34**	**55.74**	**66.14**

Table 7: Average word accuracy as percentage for the Aurora WI007 front-end

The average recognition results are listed in Table 8 for both training modes and all test sets when applying the Aurora WI007 front-end in combination with the compression scheme.

training mode	test set A	test set B	test set C
multi-condition	**87.77**	**85.77**	**82.65**
clean	**60.16**	**54.94**	**63.96**

Table 8: Average word accuracy as percentage for the Aurora front-end including compression

Only a small loss in recognition performance can be seen in case of high word accuracy. The loss is slightly higher in situations with a poor accuracy (clean training) and for the case of considering a different frequency characteristic (test set C).

ACKNOWLEDGEMENTS

This experimental framework has been prepared as a result of discussion and agreement among the Aurora project partners. We would like to thank all partners for their contribution to this framework but especially

- Nabil Benemeur (while at Matra-Nortel) who developed the original scripts for the noise addition.
- Erwin Marschall (Siemens) who performed the downsampling of the TIDigits from 20kHz to 8kHz.
- Gary Chong (Motorola & National University of Singapore) who developed the HTK scripts and file lists for training and test sets.

DATABASE DISTRIBUTION

This database as well as all scripts for the HTK recognizer is available from ELRA (European Language Resource Association) [10]. A mandatory requirement is the proof of purchasing the original TIDigits from LDC (Linguistic Data Consortium).

REFERENCES

[1] D Pearce, “Enabling New Speech Driven Services for Mobile Devices: An overview of the ETSI standards activities for Distributed Speech Recognition Front-ends” Applied Voice Input/Output Society Conference (AVIOS2000), San Jose, CA, May 2000

[2] A. Varga, H.J.M. Steeneken, “Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems”, Speech Communication, Vol.12, No.3, pp. 247-252, 1993

[3] R.G. Leonard, “A database for speaker independent digit recognition, ICASSP84, Vol.3, p.42.11, 1984

[4] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, P. Woodland, “The HTK book – version2.2”, Entropic, 1999

[5] http://www.speechdat.org/SP-CAR

[6] ITU recommendation G.712, “Transmission performance characteristics of pulse code modulation channels”, Nov. 1996

[7] ETSI-SMG technical specification, “European digital cellular telecommunication system (Phase 1); Transmission planning aspects for the speech service in GSM PLMN system”, GSM03.50, version3.4.0, July 1994

[8] ITU recommendation P.56, “Objective measurement of active speech level”, Mar. 1993

[9] ETSI standard document, “Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithm”, ETSI ES 201 108 v1.1.1 (2000-02), Feb. 2000

[10] http://www.icp.inpg.fr/ELRA/home.html