Title: **Aurora-5 Experimental Framework for the Performance Evaluation of Speech Recognition in Case of a Hands-free Speech Input in Noisy Environments**

Author: **Guenter Hirsch, Niederrhein University of Applied Sciences**

Date: **20th November 2007**

Version: **2.1**

1 Summary

This document has been written as short reference to the most important details of the Aurora-5 experiment. Aurora-5 has been mainly developed to investigate the influence on the performance of automatic speech recognition for a hands-free speech input in noisy room environments. Furthermore two test conditions are included to study the influence of transmitting the speech in a mobile communication system. This is only of interest for the situations where no distributed speech recognition (DSR) is applied.

The earlier three Aurora experiments had a focus on additive noise and the influence of some telephone frequency characteristics. Aurora-5 tries to cover all effects as they occur in realistic application scenarios. The focus was put on two scenarios. The first one is the hands-free speech input in the noisy car environment with the intention of controlling either devices in the car itself or retrieving information from a remote speech server over the telephone. The second one covers the hands-free speech input in a type of office or in a type of living room to control e.g. a telephone device or some audio/video equipment.

As for the Aurora-2 experiment the TIDigits database is taken as basis for these experiments. The creation of the noisy speech data for the mentioned scenarios will be described in the next section. Several training modes have been defined for the training with clean data only or for a training with a mixture of clean and noisy data. Test conditions are defined to study the influence of combining different input conditions as they might occur in real applications.
Shell scripts are provided to run recognition experiments with HTK. Exemplary results are presented for applying the advanced front-end as standardized by ETSI.

Besides the artificially distorted data we provide a subset of speech data from the meeting recorder project [1]. During this project speech signals have been recorded in hands-free mode in a meeting room at the International Computer Science Institute in Berkeley. Besides the recording of conversational speech during the meetings the speech data contain also sequences of digits from different speakers. Experiments have been defined to recognize these digits by applying the HMMs that have been trained on clean TIDigits.

2 Noisy TIDigits speech data

A simulation tool [2] has been applied for the artificial creation of the noisy TIDigits versions. A down-sampled version of the TIDigits is used as “clean” speech data at a sampling frequency of 8 kHz. A few details about the simulation of certain recording and transmission conditions will be mentioned below.

To simulate the hands-free speech in a room the clean speech signals are convolved with the impulse responses of different acoustic scenarios. Three different hands-free input

conditions are considered. The first one is defined by the measured impulse response in a car. The two other impulse responses have been created by applying an image source model and adding further late reflections to create quite natural sounding reverberation. The first response considers source and receiver positions in an office room that aim at the simulation of a person sitting at a desk and controlling a telephone device on the desk. The second condition is based on a response where a person is sitting in a living room and controlling an audio or video device at a distance of about 3 to 4 meters. The simulation tool allows the variation of the reverberation time in a certain range.

A noise signal can be added to the speech signal at a desired SNR after the optional simulation of the hands-free input. Speech and noise are added together in the same way as it was done within the Aurora-2 experiment. The speech level is estimated according to ITU recommendation P.56 where speech and noise signals are filtered with a G.712 characteristic for estimating their levels. The corresponding modules from the ITU software library have been applied.

Finally the influence can be simulated for transmitting the noisy speech over a cellular telephone network. The encoding and decoding of speech is applied as defined by the AMR coding scheme, and a simulation is applied for the distortions on the GSM radio channel.
A Web site is available at http://dnt.kr.hsnr.de/sireac.html where you can experience the simulation tool with your own speech data.

2.1 Test sets

In general the 8700 utterances of all adult speakers (as designated for testing with the TIDigits) are used for each test condition.

For the experiments with respect to the car environment all data is filtered with a G.712 characteristic because the use of telephone-like equipment is quite likely in this application scenario. Different combinations of distortion effects are considered. For the case of using a close talking microphone and controlling devices in the car itself only the influence of additive noise is observed. When allowing a hands-free speech input the combination of hands-free input and additive car noise is taken into account. For the case of sending the noisy speech furthermore to a remote server the transmission over a GSM network is simulated. This condition is only of interest in case the approach of DSR (Distributed Speech Recognition) is not applied. Signal-to-noise ratios (SNRs) in the range from 15 down to 0 dB are considered as quite representative for this application scenario.

As another quite typical speech input scenario for the access to a remote speech server we look at the use of a mobile phone on the street or at public places like e.g. train stations or airports. This condition is simulated as combination of additive noise and the transmission over a GSM network. All test conditions mentioned so far are summarized in table 1.

As car noise a segment is randomly selected out of 8 recordings that have been made in two different cars under different conditions like e.g. windows closed or open. Each recording has a length of several minutes. This is different in comparison to the Aurora-2 experiment where only a single recording was used for a specific condition. But this should reflect better the noise variability of real applications. In general the car noise signals are fairly stationary.

As noise at public places a segment is randomly selected out of 4 recordings at the following places:

- at an airport
- at a train station
- inside a train
- on the street

These noise signals contain more non-stationary segments than the car noise as e.g. people chatting in the background.

	Car noise	Hands-free & Car noise	Hands-free & Car noise & GSM	Noise at public places & GSM
SNR	clean	clean	clean	clean
	15	15	15	15
	10	10	10	10
	5	5	5	5
	0	0	0	0

Table 1: 20 test conditions for the recognition inside a car or at public places

For the simulation of the GSM transmission one of the 8 following conditions is randomly selected with an equal occurrence of all conditions:

Data rate (kBit/s) of AMR mode	4,75	5,15	5,9	6,7	7,4	7,95	10,2	12,2
C/I (dB)	1	4	4	7	7	10	10	13

This should take into account that the AMR modes with lower data rates are selected at bad channel conditions with a lower C/I (carrier-to-interference) ratio.

Besides the experiments that are related to the car environment and/or the use of telephone equipment, a second set of test conditions is considered that contains the hands-free speech input in rooms in combination with the corresponding noise in such situations. No G.712 filtering is applied for these conditions. The test conditions without hands-free speech input can be compared to the cases with hands-free input in 2 different rooms. The conditions are summarized in table 2.

	Interior noise	Hands-free in office & Interior noise	Hands-free in living room & Interior noise
SNR	clean	clean	clean
	15	15	15
	10	10	10
	5	5	5
	0	0	0

Table 2: 15 test conditions for the recognition inside rooms

As interior noise a segment is randomly selected out of one of the following 5 signals that have been recorded at:

- a shopping mall
- a restaurant
- an exhibition hall
- an office
- a hotel lobby.

All of these noise signals partly contain non-stationary segments.

The reverberation time for the office room is randomly varied in the range from 0.3 to 0.4 s. The reverberation time for the living room is randomly varied in the range from 0.4 to 0.5 s. Thus, the impulse responses for creating successive utterances are not identical.

2.2 Training sets

Four training modes have been defined with two modes for the car/street noise conditions and two further modes for the interior noise conditions.
All TIDigits training data (8623 utterances) of the adult speakers is filtered with a G.712 characteristic for the application to the car/street noise conditions. This corresponds to the clean training of the Aurora-2 experiment. A mixed-mode training on clean and noisy data has been set up where the noisy data is created for all of the test conditions with car/street noise. The total number of training utterances is again 8623 with the utterances equally and randomly distributed across all conditions.
For the case of interior noise another training mode has been defined for the training on clean data only without G.712 filtering. Furthermore, a mixed-mode training is available with clean and noisy data from all conditions with interior noise.

3 Meeting Recorder Digits

Besides the artificially distorted TIDigits data a small set of real recordings in a meeting room is available. These data have been recorded in hands-free mode at the International Computer Science Institute in Berkeley as part of a larger data collection [1]. Only these recordings are considered here that contain sequences of digits. There are about 2400 utterances from 24 speakers available that contain about 7800 digits in total. The speech was recorded with several microphones that were placed in the middle of the table in the meeting room. Thus, 4 different versions of all utterances exist recorded with 4 different microphones. The recordings contain only a small amount of background noise but the effects of the hands-free recording in the reverberant room. Furthermore, the recording level is quite low for almost all utterances.
The scripts and a label file for running recognition experiments on these data are available on the DVDs. Recognition experiments have been defined where these HMMs are applied that have been trained on clean TIDigits data only without any additional filtering as described as one training mode in the previous section. Thus, a big mismatch between training and test data is taken into account. The recordings of each microphone are separately processed so that you can obtain 4 word accuracies as output of this experiment.

4 Recognition experiments with HTK

Shell scripts have been written for all 4 training modes to create whole word HMMs with HTK where version 3.3 of HTK has been used [4]. The gender independent HMMs are defined by the following parameters:

- 16 states per word
- simple left-to-right models without skips over states
- mixture of 4 Gaussians per feature and state

A single HMM is used to model the pauses. This consists of 3 states and has the same structure as in the Aurora-2 experiment. The only difference is the use of mixtures with 4 Gaussians per feature and state.

The training is done as for Aurora-2 without knowledge about the time labelling of the training utterances. After an initialisation of the HMMs with the HTK routine HCompV about 30 iterations of the embedded Baum-Welch reestimation are applied by using the HTK routine HERest and increasing the number of Gaussians in three steps.

5 Recognition results for the advanced ETSI front-end

The advanced front-end as standardized by ETSI in its version 1.1.3 [3] is applied to create exemplary recognition results for the Aurora-5 experiment.

5.1 HMMs trained with G.712 filtered data

When training the HMMs with the clean TIDigits data that have been filtered with a G.712 characteristic the following word accuracies in percent are achieved for the test conditions with car noise or noise at public places:

	CarNoise			**StreetNoise**
SNR	**-**	**& Handsfree**	**& Handsfree & GSM**	**& GSM**
-	99.44	97.91	91.53	97.94
15	98.70	94.66	85.83	94.26
10	97.58	91.26	81.00	89.83
5	94.13	83.32	71.14	79.70
0	83.96	65.49	51.26	58.56

5.2 HMMs trained with mixed G.712 filtered data (car/street cases)

Applying a training on a mixture of clean and noisy data the following word accuracies are achieved where training utterances from all noise conditions with car noise or noise at public places take part in the training:

	CarNoise			**StreetNoise**
SNR	**-**	**HandsfreeCar**	**HandsfreeCar_GSM**	**GSM**
-	99.10	98.61	96.24	98.14
15	98.83	97.60	94.72	96.48
10	98.37	96.69	93.28	94.17
05	97.15	94.05	89.22	88.95
00	92.96	85.35	77.31	74.99

5.3 HMMs trained with clean data

When training the HMMs with the clean TIDigits only the following word accuracies are achieved for the test conditions with interior noise and without and with hands-free speech input:

	InteriorNoise		
SNR	**-**	**HandsfreeOffice**	**HandsfreeLivingroom**
-	99.48	93.62	84.75
15	97.38	88.45	77.61
10	94.29	81.73	69.44
05	85.58	68.66	54.82
00	64.74	46.95	35.40

5.4 HMMs trained with mixed data (interior cases / non G712)

Applying a training on a mixture of clean and noisy data the following word accuracies are achieved where training utterances from all noise conditions with interior noise take part in the training:

	InteriorNoise		
SNR	**-**	**HandsfreeOffice**	**HandsfreeLivingroom**
-	98.99	97.77	96.10
15	96.96	95.95	94.14
10	94.39	92.90	89.96
05	88.15	85.21	80.24
00	73.91	67.81	60.75

5.5 Meeting recorder digits

Applying the HMMs trained on clean TIDigits only without any additional filtering we obtain the following word accuracies for the utterances from the 4 microphones. The abbreviations 6 7 E and F are used as notation for the microphones.

Microphone	Word accuracy / %
6	64.31
7	47.66
E	58.12
F	62.72

6 References

[1] A. Janin et al., "The ICSI meeting corpus", ICASSP 2003, Hongkong, 2003

[2] H.G. Hirsch, H. Finster, "The simulation of realistic acoustic input scenarios for speech recognition systems", *Interspeech conference 2005*, pp. 2697-2700, Lisbon, Portugal, 2005

[3] ETSI standard document, "Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Advanced Front-end feature extraction algorithm; Compression algorithm", *ETSI ES 202 050 v1.1.3 (2003-10)*, Oct. 2003

[4] S. Young et al., "The HTK book (version 3.3)", Cambridge University Engineering Department, http://htk.eng.cam.ac.uk, 2005