

Title: Description and Baseline Results for the Subset of the Speechdat-Car German Database used for ETSI STQ Aurora WI008 Advanced DSR Front-end Evaluation

Source: Lorin Netsch, Texas Instruments

Date: 12-January-2001

Version: 1.1

1. Background

This paper describes baseline experiments using a subset of the German SpeechDat-Car (SDC) corpus provided to the Aurora consortium. These experiments provide baseline results for use in the evaluation of a noise robust front-ends for feature extraction, WI008. This report explains the composition of the corpus used for baseline evaluation and presents baseline digit recognition performance. It is similar in nature to the reports on the Italian SDC [1], and Finnish SDC [2].

2. German SDC corpus subset

Texas Instruments received a subset of the German SDC corpus which was compiled by Stefan Euler of Bosch. In addition, TI received supporting documentation such as corpus documentation, transcriptions in MLF format, and a spreadsheet defining speaker and session characteristics. The corpus subset consists of speech from 112 speakers, 61 female and 51 male. Each speaker contributed one or two sessions of speech, with data collected simultaneously on a close-talking microphone (channel 0) and a hands-free microphone (channel 1). Each session was collected in one of four car driving conditions:

- High speed good road
- Low speed rough road
- Stopped with motor running
- Town traffic

In addition, the car was placed in various configurations:

- Climate control on/off
- Left front window open/closed
- Right front window open/closed
- Rear window open/closed
- Sunroof open/closed
- Windshield wipers on/off

The following were utterances included in the recordings:

1. Four isolated digits (I1-I4)
2. A sequence of 10 isolated digits (B1)
3. A sheet number (C1)
4. A spontaneous telephone number (C2)
5. A credit card number (C3)
6. A PIN code (C4)
7. Three telephone numbers (C5-C7)

Each session consists of up to a maximum of the 12 items listed above. Sometimes the recorded utterances contained extraneous words. For example, the spontaneous telephone numbers sometimes included *VORWAHL* (area code) or similar additions to the basic digit sequence. Additionally, in many cases natural numbers were used, such as NULL SIEBEN ELF (zero seven eleven). The Aurora subset of the German SDC corpus excluded utterances such as these, restricting the vocabulary to the German digits (null, eins, zwei, zwo, drei, fuenf, sechs, sieben, acht, and neun.) This yielded 3118 total files, 1559 from each microphone type.

The data received by Texas Instruments had already been processed by removing utterances not meeting the above vocabulary restrictions, down-sampling the sampled data to 8kHz, and extracting the individual utterances into separate files. Upon listening to the data, we noted several anomalies, such as:

- A hardware failure resulting in loss of all data in the hands-free channel of one session
- Some remaining files containing natural numbers
- A few files containing significant vocal interjections by the speaker
- Files whose transcription in the MLF file were incorrect
- Files in which the utterance was significantly truncated at the beginning or end
- Files in which the close-talking microphone recording contained significant electrical interference noise
- A significant number of files in which the A/D saturated
- A few sessions where some small portion of the original prompting tone remained

These anomalies were handled as follows. We discarded files in which the recording hardware failed completely such that there was no speech signal. We discarded files containing natural numbers. We discarded files containing vocal interjections by the speaker. We corrected the MLF transcription for those files that had a transcription error. A few of these files were deemed as unusable due to other conditions, and so the transcription was not modified for those. We discarded files in which there was significant truncation if it was easily audibly detectable and confirmed by plotting the waveform. Files containing a portion of the initial beep tone were used as is.

The electrical noise problem existed in a significant number of files and it was difficult to determine how to handle the noise, since the amplitude of the noise varied from file to file and within a file. The decision was made to discard those files in which the noise was deemed dominant portion of the signal as determined by a listener. Hence, the noise is present to some degree in many of the remaining close-talking microphone files.

Files with A/D saturation included files where large portions of the signal were saturated, and those where only an occasional sample was limited by the A/D. The definition of saturation was any file having a sample of magnitude greater than or equal to 32767. There were 380 such files. Only the files that contained significant saturation were discarded. To determine which files to discard, files were sorted by number of saturated samples, and also by number of saturated samples divided by total number of samples in the file ("saturation density"). These two sorts indicated that there was a "knee" in the plots of saturated samples and saturation density. The top 50 of each file list were discarded. Since there was a significant amount of overlap, this resulted in discarding 70 files.

The result of this post-processing of the corpus data was removal of 189 files. The distribution of files removed is shown in Table 1. The distribution of the remaining 2929 files used for training and testing is shown in Table 2.

Driving Condition	Microphone Type			
	Close Talking		Hands Free	
	Female	Male	Female	Male
Stop Motor Running	1	4	-	-
Town Traffic	15	19	9	8
Low Speed Rough Road	36	16	31	8
High Speed Good Road	20	4	11	7

Table 1: Distribution of discarded files

Driving Condition	Microphone Type			
	Close Talking		Hands Free	
	Female	Male	Female	Male
Stop Motor Running	92	99	93	103
Town Traffic	283	203	289	214
Low Speed Rough Road	305	207	310	215
High Speed Good Road	127	128	124	137

Table 2: Distribution of corpus files

Examination of the files indicated that while some sessions were marked as “High Speed Good Road” the recordings included background conditions where the car was stopped and appeared to be waiting to make a turn. Similar situations were noted for the “Town Traffic” and “Low Speed Rough Road” driving conditions. This prompted measurements of the SNR for the different driving and microphone conditions of the corpus, which are shown in Figure 1.

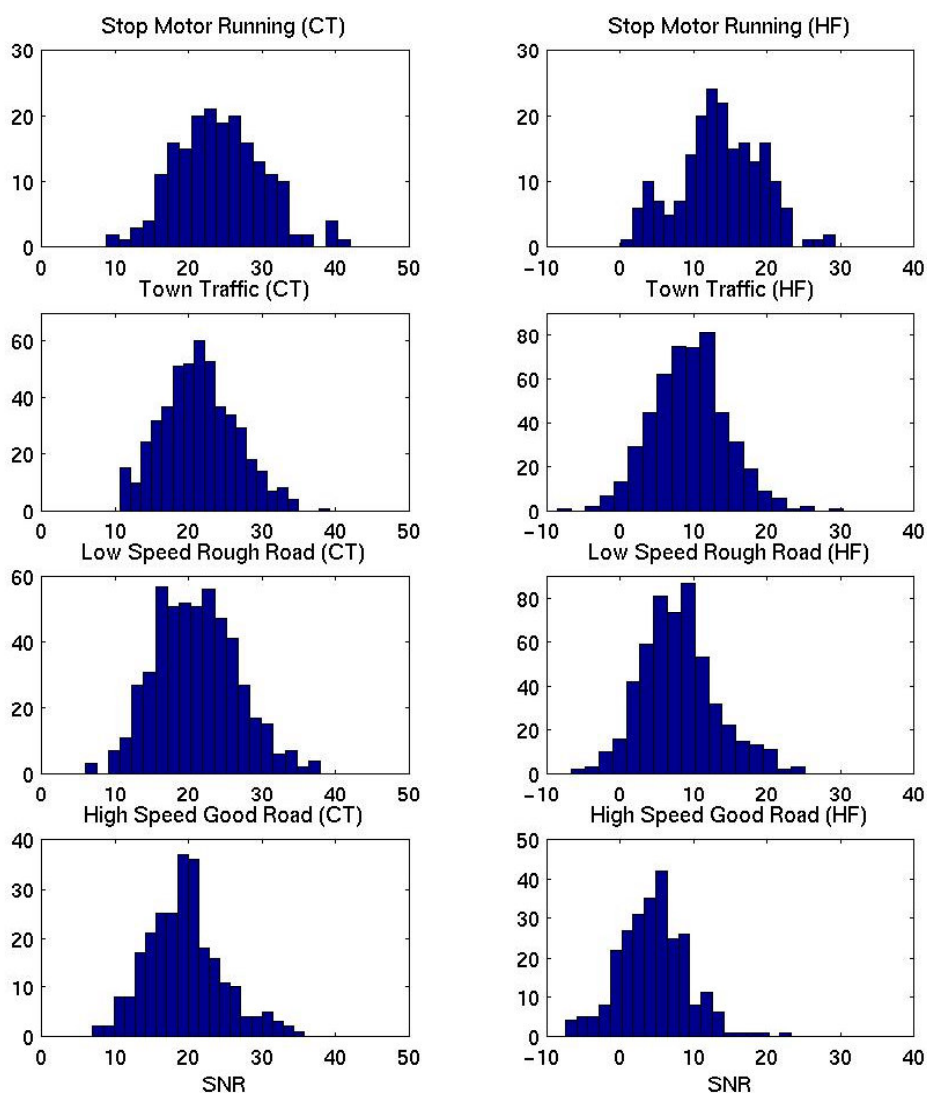


Figure 1: SNR of corpus files by driving condition and microphone

3. Experimental set-up

The experiments performed in this report used the HTK tools to train and test models. We used modified versions of the PERL scripts that were implemented for the Italian baseline SDC results described in [1] to create the front-end feature files, train the HMM models, and perform testing. The experimental results reported in this document use only a single set of training and recognition parameters. We used HMM's with 18 HTK states, and the 3,7,7,7 training iteration configuration in the experiments.

3.1. Well-matched condition evaluation

The well-matched experiment utilized data from both microphone types and all driving conditions for both testing and training data. We selected a set of 70% of each of the female and male speakers such that the utterances of these speakers represented approximately 70% of the utterances for each condition. These files were used as training data, and the files for the remainder of the speakers comprised the test data. A breakdown of the distribution of files for the well-matched experiments is shown in Table 3, and the distribution of number of repetitions of each digit is shown in Figure 2. The results of performing evaluation recognition tests are shown in Table 4 and Table 5.

Driving Condition	Train				Test			
	Female(43)		Male(36)		Female(18)		Male(15)	
	CT	HF	CT	HF	CT	HF	CT	HF
Stop Motor Running	60	60	70	70	32	33	29	33
Town Traffic	201	205	141	148	82	84	62	66
Low Speed Rough Road	213	212	143	147	92	98	64	68
High Speed Good Road	91	86	88	97	36	38	40	40

Table 3: Distribution of files for well-matched condition

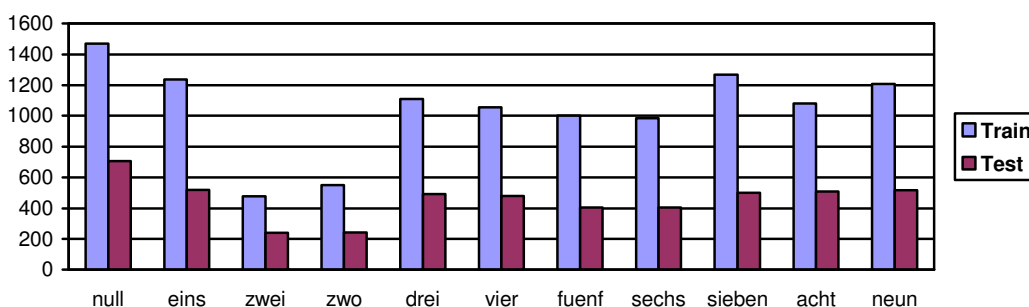


Figure 2: Number of repetitions for well-matched condition

Number Sentences/Words	Sub %	Del %	Ins %	Word Correct	Word Accuracy	Sentence Correct
897/5009	4.37	3.15	1.88	92.49	90.58	73.58

Table 4: Well-matched condition evaluation results

Driving Condition	Word Accuracy	
	CT	HF
Stop Motor Running	87.86	89.82
Town Traffic	95.76	92.49
Low Speed Rough Road	95.28	90.99
High Speed Good Road	95.36	65.82

Table 5: Well-matched condition evaluation results by driving condition

3.2. Medium-mismatch condition evaluation

The medium-matched evaluation utilized training data from the hands free microphone using all driving conditions except for the High Speed Good Road driving condition. Testing was performed using data from the hands free microphone and the High Speed Good Road driving condition only. We selected 30% of the speakers that provided the most hands free High Speed driving condition utterances as the test set. The hands free data from the remaining 70% of the speakers except for the High Speed driving condition data comprised the training data set. A breakdown of the distribution of files for the medium-mismatch experiments is shown in Table 6, and the distribution of number of repetitions of each digit is shown in Figure 3. The results of performing evaluation recognition tests are shown in Table 7.

Driving Condition	Train		Test	
	Female(43)	Male(36)	Female(18)	Male(15)
	HF	HF	HF	HF
Stop Motor Running	85	103	-	-
Town Traffic	223	155	-	-
Low Speed Rough Road	268	163	-	-
High Speed Good Road	-	-	124	117

Table 6: Distribution of files for medium-mismatch condition

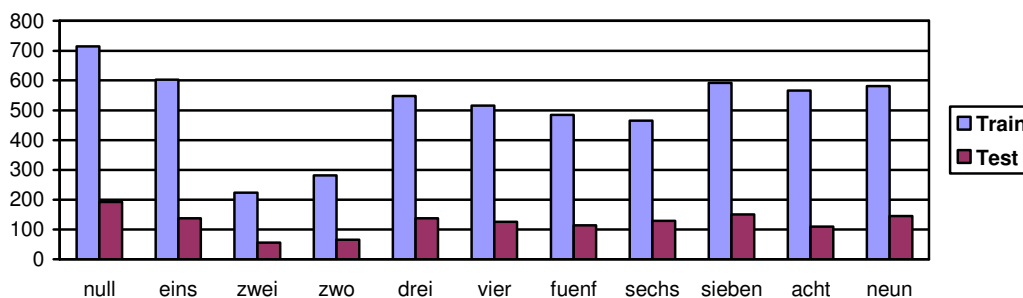


Figure 3: Number of repetitions for medium-mismatch condition

Number Sentences/Words	Sub %	Del %	Ins %	Word Correct	Word Accuracy	Sentence Correct
241/1366	11.05	7.69	2.20	81.26	79.06	53.94

Table 7: Medium-mismatch condition evaluation results

3.3. High-mismatch condition evaluation

The high-matched experiment utilized data from the close talking microphone and all driving conditions for the training data. We used the same set of 70% of each of the female and male speakers as was used in the well-matched condition to form the training set of speakers. The hands free data for all driving conditions except the Stopped Motor Running driving condition from the remaining 30% of the speakers comprised the test data. A breakdown of the distribution of files for the high-mismatch experiments is shown in Table 8, and the distribution of number of repetitions of each digit is shown in Figure 4. The results of performing evaluation recognition tests are shown in Table 9 and Table 10.

Driving Condition	Train		Test	
	Female(43)	Male(36)	Female(18)	Male(15)
	CT	CT	HF	HF
Stop Motor Running	60	70	-	-
Town Traffic	201	141	84	66
Low Speed Rough Road	213	143	98	68
High Speed Good Road	91	88	38	40

Table 8: Distribution of files for high-mismatch condition

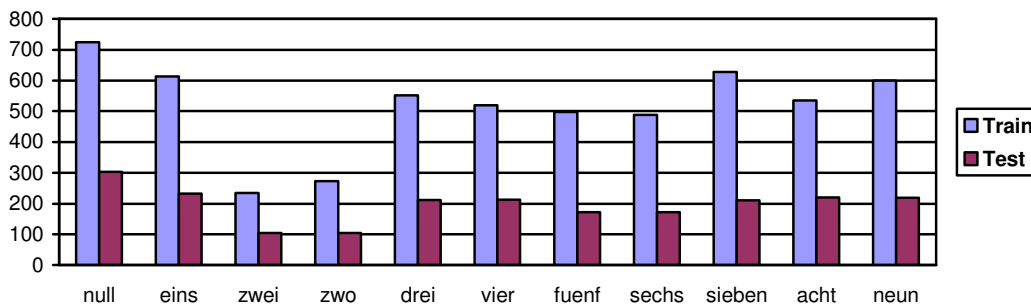


Figure 4: Number of repetitions for high-mismatch condition

Number Sentences/Words	Sub %	Del %	Ins %	Word Correct	Word Accuracy	Sentence Correct
394/2162	9.48	13.78	2.45	76.73	74.28	56.09

Table 9: High-mismatch condition evaluation results

Driving Condition	Word Accuracy
Town Traffic	81.68
Low Speed Rough Road	80.81
High Speed Good Road	46.65

Table 10: High-mismatch condition results by driving condition

4. Concluding remarks

The experimental results of this report indicate the recognition word accuracy when dividing the SDC German corpus by driving condition. A summary of the word accuracy by matching condition is shown in Table 11.

Matching Condition	Word Accuracy (%)
Well-matched	90.58
Medium-mismatch	79.06
High-mismatch	74.28

Table 11: Summary of word accuracy by testing condition

We will examine the possibility of further experiments to perform recognition in various mismatch conditions by dividing the corpus by SNR. These results will be reported in a later document.

Acknowledgements

Several people contributed during the process of running these experiments. Stephan Euler provided the SDC German corpus and supporting documentation. He rapidly responded to questions regarding the corpus configuration and collection conditions. Ram Ramalingam of Nokia and Hans-Guenter Hirsch of Ericsson provided information, valuable guidance, and made available supporting software for setting up the experiments.

References

- [1] *Description and Baseline Results for the Subset of the Speechdat-Car Italian Database used for ETSI STQ Aurora WI008 Advanced DSR Front-End Evaluation*, STQ Aurora DSR Working Group input document AU/237/00
- [2] *Baseline Results for subset of SpeechDat-Car Finnish Database for ETSI STQ WI008 Advanced Front End Evaluation*, STQ Aurora DSR Working Group input document AU/225/00