The MoveOn Motorcycle Speech Corpus

Thomas Winkler*, Theodoros Kostoulas†, Richard Adderley°, Christian Bonkowski*, Todor Ganchev†, Joachim Köhler*, Nikos Fakotakis†

*Fraunhofer IAIS
Schloss Birlinghoven, 53754 Sankt Augustin, Germany
thomas.winkler, christian.bonkowski, joachim.koehler@iais.fraunhofer.de

†Wire Communications Laboratory, University of Patras
Rion-Patras, 26500, Greece
tkost, tganchev, fakotaki@wcl.ee.upatras.gr

°A E Solutions (BI)
11 Shireland Lane, Redditch, Worcestershire, B97 6UB, UK
rickadderley@a-esolutions.com

Abstract

A speech and noise corpus dealing with the extreme conditions of the motorcycle environment is developed within the MoveOn project. Speech utterances in British English are recorded and processed approaching the issue of command and control and template driven dialog systems on the motorcycle. The major part of the corpus comprises noisy speech and environmental noise recorded on a motorcycle, but several clean speech recordings in a silent environment are also available. The corpus development focuses on distortion free recordings and accurate descriptions of both recorded speech and noise. Not only speech segments are annotated but also annotation of environmental noise is performed. The corpus is a small-sized speech corpus with about 12 hours of clean and noisy speech utterances and about 30 hours of segments with environmental noise without speech. This paper addresses the motivation and development of the speech corpus and finally presents some statistics and results of the database creation.

1. Introduction

Robust speech recognition for hands-free speech input in the vehicle environment is a challenging field of research (Cristoforetti et al., 2003; Kun et al., 2005; Li et al., 2007). Suitable speech data for training and testing such a system is vitally important. SpeechDat-Car (Moreno et al., 2000), for example, is a corpora with speech data recorded in a car environment. An even more difficult area of speech recognition is the motorcycle environment with an even higher level of noise. A German speech database – the SmartWeb Motorbike Corpus (SMC) (Kaiser et al., 2006) – has been recorded on a motorcycle facing the problems connected to the challenging conditions.

The development of the MoveOn Motorcycle Speech Corpus described in this paper has been initiated from a lack of suited speech corpora for the objectives of the EC-funded project MoveOn[1]. Command and control phrases and utterances in British English useful for template based dialogue systems are recorded on a motorcycle. Additionally, clean speech is recorded in a silent office environment using the same setup and utterances. The speech corpus is adapted but not limited to the police domain. Recordings with low distortion are in the focus of the corpus development. A good description of the environmental conditions and dominant noise is another aim during the corpus development.

2. MoveOn Scenario

MoveOn is an EC-funded project working on a multi-modal and multi-sensor zero-distraction interface for two wheel vehicles on the move. The generic aim of the project is practically focused on the use case of an assisting system for police motorcyclists. Speech input – which will be processed by a robust speech recognition system – is one of the main modalities. Voice commands will control the TETRA communication, basic routing etc., and simple dialogue phrases will be used to request data from a police database.

The extreme conditions on the motorcycle complicate a reliable speech recognition required by a zero-distractive system. The described MoveOn Motorcycle Speech Corpus is based on the requirements of West Midlands Police (WMP), UK, and enables an optimal adaptation of the speech recognition system. Furthermore, the corpus provides a profile of the difficult speech and noise environment on the motorcycle.

3. Database Design

The design and realization of the database follows the requirements of the project and the motorcycle environment. A balanced set of prompts is created and acoustic prompt sheets are used.

3.1. Database Content

The linguistic contents of the database covers terminology used in the police practice, characteristic MoveOn-specific command phrases and application words, as well as their most often used synonyms. For balancing the phonetic contents, a number of phonetically rich sentences, taken from the British English SpeechDat(II)-FDB4000 database (Heuvel et al., 2001), were added.

[1] http://www.m0ve0n.net

Due to the specifics of the MoveOn application – involving hands-busy and eyes-busy motorcyclists – the prompt sheet provided was a sequence of recorded audio prompts. Table 1 presents the structure of a MoveOn prompt sheet. Here the items "AW001-AW065" are application-specific words and phrases, which are repeated twice during each prompt sheet. The prompts "SP001-SP010" invite the speaker to provide a spontaneous answer to a specific question. Four of these prompts are repeated 10 times inside a prompt sheet. Specifically, the motorcyclist is asked to provide information about the speed at that moment, present location, traffic conditions, and to spell a plate number of a car nearby. The "SR001-SR010" items correspond to the phonetically rich sentences, which are unique for each prompt sheet.

3.2. Prompt Sheets

In total, 23 prompt sheets were created. Each prompt sheet consists of a random combination of all items shown in Table 1. Each prompt is accompanied by a short introductory phrase instructing the motorcyclist if a word, a phrase or a question follows. Each prompt ends with a DTFM beep sound, which invites the speaker to speak. The length of the silence after each prompt is about twice the time necessary for the specific utterance. This gives some time to the speaker to speak when the driving conditions permit it. In total, each prompt sheet consists of 302 prompts – resulting from the duplication of the "AW001-AW065" items + 10 repetitions of 4 "SP" items + rest of the items. The overall length of a prompt sequence is approximately 85 minutes.

4. Hardware Setup

The selected hardware and the setup of the hardware influence the quality of the speech and noise database. Robust close-talk microphones with low distortions even at high acoustic pressure and a good frequency response in the required frequency range are used. The recording device must enable recordings with at least 16 kHz of sample rate and a sample size of 16 bit to provide a sufficient quality of the recorded data. The recording level of the device is adjusted carefully to a rather low level to avoid overdriven signals. The signals are recorded in PCM-WAV. A fixed cabling instead of a wireless connection is preferred to avoid additional sources of distortion.

Besides the recording hardware and setup, helmet and motorcycle have major influences on the recorded signal. The environmental conditions of each recording session - especially engine noise and air-wind noise - are noticeably dependent on helmet and motorcycle. The following equipment is used during the recordings:

4.1. Motorcycles and Helmets

Standard and police motorcycles from two major motorcycle manufacturers (BMW and Honda) and a variety of different helmets (from Arai, HJC, Shoei and Schuberth) are used. The assembly of the microphones is prepared using the Shoei XR1000, which is the most common helmet during the recording campaign. But the microphone setup and fixation is adaptable to the other used helmets as well.

4.2. Microphone Setup

A setup of three microphones is selected. Two AKG C417"' lavalier microphones with windshields are affixed to the helmet in a distance of about 45 mm left and right from the center of the mouth. The AKG C417"' provides the required good frequency response and robustness towards high acoustic pressure levels, which is very important considering the extreme conditions on the motorcycle. Additionally, speech is simultaneously recorded by an Alan throat microphone of the type AE 38. A throat microphone picks up vibrations directly from the larynx and, hence, is less prone to the environmental noise on the motorcycle. The integrated ear phone of the Alan throat microphone transmits the voice prompts to the speaker.

4.3. Recording Hardware

The captured data is recorded by the small-sized audio recorder ZOOM H 4. Two independent audio sources can be recorded at the same time. Suited devices with more than two recording channels were not available. Hence, two recording devices are used simultaneously to cover all three microphone channels. The fourth available recording channel – the second channel on the second device – is used to record the output of the first device to enable a synchronization of all channels. The speech data is recorded at a sample rate of 44.1 kHz at 16 bit. Up to two additional tracks can be used for playing back the voice prompts while recording.

5. Data Recordings

Speech data was recorded in two different environments. First, noisy speech data was captured in a realistic environment on the motorcycle. Second, additional speech data with the same hardware setup and the same voice prompts was recorded in a silent office environment.

5.1. Noisy Motorcycle Recordings

The main part of the database was recorded on the motorcycle to capture noisy speech and environmental noise.

5.1.1. Recruitment

Professional police motorcyclists from West Midlands Police were recruited for the data collection. To ensure fullest cooperation all senior Officers from each of the 21 Operational Command Units (OCU) were presented with an overview of the project and the mechanics regarding the recording process to seek approval for their Officers to participate.

Recruiting experienced police motorcyclists is advantageous in several aspects. First, good skills in riding a motorcycle increase the safety of the speakers. Second, an enhanced quality of the captured speech is achieved by a better comprehension of the common police work flow and the domain specific words and contexts. Due to the fact that hardly any female police motorcyclists are available, only male speakers could be recruited for the motorcycle recordings.

Code	*Description*	*Number of items*
AW001-AW065	Application words-phrases	65
BD001-BD005	Sequence of 5 isolated digits in one utterance (written in digits)	5
PL001	Plate number	1
ID001-ID010	Single isolated digit	10
TP001	Time phrase	1
GW001-GW026	General words	26
LC001-LC014	Call signs	14
MW001-MW011	Special mandatory words	11
MS001-MS015	Special mandatory words-synonyms	15
OW001-OW022	Optional words-phrases	22
CP001-CP007	Confirmation phrases	7
SR001-SR010	Phonetically rich sentences	10
SP001-SP010	Spontaneous questions	10

Table 1: Structure of the MoveOn prompt sheet

5.1.2. Recording Procedure

A fixed route through Birmingham (UK) is defined for the campaign enabling a more accurate definition of the current environment and background noise. The selected route offers a variety of different acoustic environments like city traffic, tunnels, motorways and rural sections. The route was driven clockwise and counter-clockwise and shuffled prompts were used for the different sessions to avoid similar environmental conditions for the same words and speech sounds. A break of about five minutes (stopping the engine but continuing the recordings) is part of most sessions to acquire some data without engine noise. Speaker and session protocols for each speaker and session are available providing additional information about speaker characteristics, recording setup and specific conditions in respect to each particular data set. A video recorded while going the recording route by car is available to complement the speaker and session protocols.

5.2. Silent Office Recordings

Besides the noisy speech data several clean speech recordings were done.

5.2.1. Recruitment

Native British English speakers were contacted and recruited via the Bonn English Network in Germany. Each speaker was individually introduced to the project and the recording procedure. Both, male and female speakers, participated in the office recording sessions.

5.2.2. Recording Procedure

Two small office rooms which are almost identical were used for the recordings. Both rooms are in a silent environment enabling noise free speech data recordings. Speech was recorded while wearing a Shoei XR1000 motorcycle helmet. The hardware setup and the prompt sheets were identical to the motorcycle recordings. Due to the helmet, reverberation effects of the office room could be neglected. Most of the sessions include recordings with open and closed visor. Speaker and session protocols are available for all speakers and sessions containing all relevant information for the clean speech recordings.

6. Database Annotation

The annotation of the MoveOn recordings involves two independent steps: (i) annotation of speech data, and (ii) annotation of background noise and transient interferences. Both steps are realized using PRAAT (Boersma and Weenink, 2007). The result of the annotation procedure is saved in the TextGrid file format (Boersma and Weenink, 2007).

6.1. Speech Annotation

During the annotation of the speech data two relevant categories are considered: (i) linguistic contents, and (ii) emotional contents. These two categories are split in different tiers. In the linguistic tier, the boundaries of utterances pronounced by the speaker and the corresponding prompts were aligned automatically. Furthermore, the annotators are asked to define the exact boundaries of the uttered phrase and transcribe the uttered phrase. The transcription is orthographic, using in principle the words as they occur in an ordinary dictionary of the English language (Heuvel et al., 2001). Word truncations, mispronunciations, non-understandable speech and non-speech acoustic events are denoted following the SpeechDat conventions (Heuvel et al., 2001).

In the annotation of the emotional contents, i.e. the affect tier, the annotators were asked to place an appropriate affect marker for each transcribed utterance, based on their human intuition. Table 2 shows how the affect markers are related to emotional states (Cowie et al., 2000).

6.2. Noise Annotation

There are five levels in the annotation of background noise and transient interferences: air-wind noise, engine noise, other noise, sound event and visor. The reason behind selecting air-wind noise and engine noise as independent tiers is their crucial effect on the speech recognition performance. For these types of noises the annotators were asked to assess the perceived amplitude of noise in three levels: low, high, very high. The tier *other noise* is filled with the type of any other relatively stationary noise that might appear, such as: general traffic noise, rain, etc. In the tier *sound event*, the annotators were asked to identify a range

Affect	*Emotional State*	*Marker*
Positive-Active	Happy, Excited, Interested, Pleased, Delighted	posa
Positive-Passive	Relaxed	posp
Negative-Active	Angry, Afraid, Panic, Terrified, Furious, Disgust	nega
Negative-Passive	Despair, Depress, Sad, Bored	negp
Neutral	Neutral	neu

Table 2: Affect markers and emotion categories

of sound events that might occur (e.g. passing-by vehicles, horns, sirens). The effect of an opened or closed visor influences the acoustic environment. So the position of the visor is estimated during the noise annotation and indicated in the last tier *visor*. Table 3 summarizes the annotation tiers used for speech and noise annotation.

Annotation	*Tier*	*Examples*
Speech	Words	*'word transcription'*
	Affect	*Table 2*
Noise	Air Wind Noise	a+, a++, a+++
	Engine Noise	e+, e++, e+++
	Other Noise	traffic, rain, ...
	Sound Event	horn, passing car, ...
	Visor	open, closed, ...

Table 3: Speech and noise annotation tiers

7. Database Results

About 12 hours of clean and noisy speech and about 30 hours of environmental noise on the motorcycle (without speech) have been recorded.

7.1. Speech Data

Overall 49 sessions – each of about one hour of length – with 39 different speakers were recorded. The complete database includes about 12 hours of clean and noisy speech utterances. The number of recorded utterances is approximately 10.000 (Table 4).

7.1.1. Noisy Speech Data

During the motorcycle recording campaign 39 recording sessions with 29 different motorcyclists were performed in noisy environments. The recorded data contains approximately 8.500 utterances. Taking into account missing channels due to technical problems, 21 fully complete recordings from 17 different speakers are available. All other 18 recordings offer either one or both close-talk channel(s) or the throat microphone channel.

7.1.2. Clean Speech Data

10 recording sessions with 10 different speakers in a silent office environment are also part of the database. Six of the ten speakers are male and four are female. Nine of ten sessions include all three microphone channels. Technical problems caused a complete blackout of the throat microphone channel for one of the male speakers.

7.2. Recorded Noise

Most prominent noises in the recordings are engine and air-wind noise. Engine noise primarily depends on the motorcycle and the way of driving. It is significant for acceleration periods and medium to high speed levels. At higher speed levels the air-wind noise becomes the most prominent noise and masks most other noises.

Other frequent environmental noises – but generally less significant – are traffic, other engines nearby and passing vehicles. Dependent on the motorcycle changing gears can clearly be noticed. Breaks and other squeaking noises as well as miscellaneous indefinable noises can also be heard in the background. Rare sound events are, for example, horns and sirens.

7.3. Motorcycles and Helmets

Motorcycle and helmet influence the environmental conditions in respect to the most crucial sources of noise – engine and air-wind. Several types of motorcycles were used for the data recordings. Typical motorcycles during the recording campaign were:

- BMW RS1200
- Honda Pan European
- BMW K1100
- Honda Gold Wing GL1800

As the motorcyclists used their private or official helmet, many different manufacturers and helmets were present. The most frequently used helmets are:

- Shoei XR1000
- Schuberth C2
- Shoei Multitec

But several other helmets – full face and flip front – from Arai, Shoei, Schuberth and HJC were also worn during the recording sessions. Both – type of motorcycle and helmet – are registered in the session protocols.

7.4. Data Loss and Distortions

Some problems occurred during the recording campaign due to malfunctions and technical limits. During the motorcycle campaign the second recording device failed several times due to malfunctions and needed to be replaced. This problem led to missing throat microphone signals in several sessions. Even though the microphones are capable of dealing with high acoustic pressure, overdriven close-talk

		recordings	*speakers*	*hours (abs./spoken)*	*utterances*
noisy	**overall rec.**	**39**	**29**	**40/10**	**8500**
	fully complete	21	17	21/5	4500
	helmet mics.	33	26	33/8	7000
	throat mic.	26	20	26/7	5500
clean	**overall rec.**	**10**	**10**	**10/2**	**2000**
	fully complete	9	9	9/2	1800
	helmet mics.	10	10	10/2	2000
	throat mic.	9	9	9/2	1800
	complete database	**49**	**39**	**50/12**	**10500**

Table 4: Amount of recorded clean and noisy speech data (hours and utterances approximated).

microphone signals due to extreme air-wind noise could not be avoided entirely. The quality of the throat microphone signal is poor for some office recording sessions. Especially for women with very small necks a good adjustment with sufficient pressure for picking up vibrations from the larynx was not always possible.

8. Conclusion

A small but good database with about 12 hours of speech and several hours of additional noise recordings has been developed. The acquired data promises to be very useful for research and development in MoveOn and beyond. Distorted signals were generally avoided by using a fixed cabling, relatively low recording amplitudes and microphones with low distortion even for high acoustic pressure levels. Despite the careful preparation, some data loss and distortion occurred due to malfunction of the recording devices and technical limits of the equipment. Great importance has been attached to the extraction of as much information as possible about the environmental conditions. In addition to the annotated speech data, annotated noise samples are available for further analysis and testing.
It is planned to publish the corpus via ELDA/ELRA.

9. Acknowledgement

The MoveOn project is partially funded by the European Commission.
We would also like to thank the Bonn English Network, all Officers from West Midlands Police as well as all other native speakers who participated in the MoveOn data collection or were involved in the preparation in either way.

10. References

P. Boersma and D. Weenink. 2007. PRAAT: Doing phonetics by computer (version 4.6.34). http://www.praat.org/. Computer program.

R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder. 2000. 'feeltrace': An instrument for recording perceived emotion in real time. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 19–24. ISCA.

L. Cristoforetti, M. Matassoni, M. Omologo, and P. Svaizer. 2003. Use of parallel recognizers for robust in-car speech interaction. In *Proceedings of the 2003 IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP '03)*, volume 1, April.

V.D. Heuvel, L. Boves, A. Moreno, M. Omologo, G. Richard, and E. Sanders. 2001. Annotation in the SpeechDat projects. *International Journal of Speech Technology*, 4:127–143.

M. Kaiser, H. Mögele, and F. Schiel. 2006. Bikers accessing the web: The SmartWeb Motorbike Corpus. In *Proceedings of the 5th International Conference on Language Resources and Evaluation, (LREC 2006)*, pages 1628–1631, Genova, Italy, May. ELRA.

A.L. Kun, W.T. Miller, and W.H. Lenharth. 2005. Evaluating the user interfaces of an integrated system of in-car electronic devices. In *Proceedings of the 8th International IEEE Conference on Intelligent Transportation Systems*, Vienna, Austria, September.

W. Li, K. Takeda, and F. Itakura. 2007. Robust in-car speech recognition based on nonlinear multiple regressions. *EURASIP J. Appl. Signal Process.*, 2007(1):5–5.

A. Moreno, B. Lindberg, C. Draxler, G. Richard, K. Choukri, S. Euler, and J. Allen. 2000. SPEECHDAT-CAR. a large speech database for automotive environments. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation, (LREC 2000)*, Athens, Greece.