Methodology of Lombard Speech Database Acquisition: Experiences with CLSD

Hynek Bořil, Tomáš Bořil, Petr Pollák

Faculty of Electrical Engineering
Czech Technical University in Prague, Czech Republic
borilh@gmail.com, borilt@gmail.com, pollak@fel.cvut.cz

Abstract

In this paper, process of the Czech Lombard Speech Database (CLSD'05) acquisition is presented. Feature analyses have proven a strong appearance of Lombard effect in the database. In the small vocabulary recognition task, significant performance degradation was observed for the Lombard speech recorded in the database. Aim of this paper is to describe the hardware platform, scenarios and recording tool used for the acquisition of CLSD'05. During the database recording and processing, several difficulties were encountered. The most important question was how to adjust the level of speech feedback for the speaker. A method for minimization of the speech attenuation introduced to the speaker by headphones is proposed in this paper. Finally, contents and corpus of the database are presented to outline it's suitability for analysis and modeling of Lombard effect. The whole CLSD'05 database with a detailed documentation is now released for public use.

1. Introduction

A great effort is being made to increase robustness of automatic speech recognition systems in order to allow for building of voice-controlled devices operating reliably in adverse environments. In noisy conditions, recognition is not only degraded by presence of the disturbing background but also by Lombard effect (LE) representing speech production changes introduced by speaker in an effort to increase communication intelligibility.

Speech databases acquired in real conditions provide valuable material for recognition systems, but in case of louder backgrounds (crowded places, moving car, airplane cockpits) it may be problematic to analyze impact of speech feature variations caused by LE separately from the impact of the noise present in the recordings. Also assuring similar recording conditions and appropriate speaker reactions to the actual noise may be an issue in the real conditions. During the recording, speakers may tend to concentrate just on the correct pronunciation of the text without adequate reaction to the actual conditions.

Databases focused on LE usually introduce simulated noisy background to the speaker through headphones, hence high SNR of the recorded speech is preserved and the recording conditions can be easily controlled (Hansen, 1996; Chi & Oh, 1996; Wakao et al., 1996).

In (Bořil & Pollák, 2005, 1), basic properties of the CLSD'05 database were introduced. In (Bořil & Pollák, 2005, 2), overall Lombard speech features of CLSD'05 were analyzed and compared to two large Czech databases containing recordings from the moving car. In CLSD'05, appearance of LE has been found significantly stronger than in case of the other two databases.

In this paper, recording platform and contents of CLSD'05 are presented and extensions of the setup proposed.

2. CLSD'05 recording platform

To enable observations of LE influence on speech features on the speaker level, each speaker was recorded both in neutral and simulated noisy scenario.

Our experiences from the recordings in natural environments show that speakers often tend to ignore actual environmental changes and concentrate just on the correct reading of the prompts. This approach does not follow a real communication and thus from the viewpoint of the speaker, there is no need to preserve intelligibility of the speech much. To avoid this, it seems reasonable to introduce a communication element into the recording process.

2.1. Recording setup

In the simulated noisy conditions, noise samples mixed with the speech feedback are reproduced to speaker by closed headphones. An operator qualifies utterance intelligibility while hearing the same noise mixed with speaker's voice of intensity decreased according to the selected virtual distance, see Fig. 1. If the speech cannot be understood well, the operator asks for repeating. It was observed that after several requests for repeating of an item speakers started to react to the actual noise appropriately. In the neutral scenario, speaker does not wear headphones while reading the prompts.

In both scenarios, the speech is sensed by two microphones placed in the different distances. Recording set consists of 2 closed headphones AKG K44, close talk microphone Sennheiser ME-104 and hands-free microphone Nokia NB2. These microphones were chosen to fit Czech SPEECON recording conditions (SPEECON, 2001).

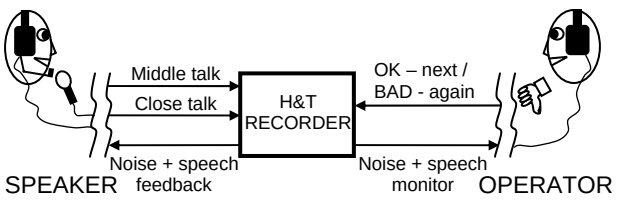

Figure 1: Recording setup

2.2. SPL adjustment

In the beginning of each Lombard session recording it was necessary to adjust level of the reproduced background noises. For this purpose, a transfer function between sound card open circuit effective voltage V_{RMS_OL} and SPL in headphones was determined by measurement on a dummy head, see Fig. 2. For the required noise level, corresponding V_{RMS_OL} was then set up. Constant 90 dB SPL and 1-3 meters of virtual distance were chosen for the

Lombard speech recording scenarios. In some cases the settings had to be modified according to the particular speaker's capabilities. The noise reproduction was interrupted between neighboring items and the recording session usually did not exceed 20-30 minutes with refreshment pauses included.

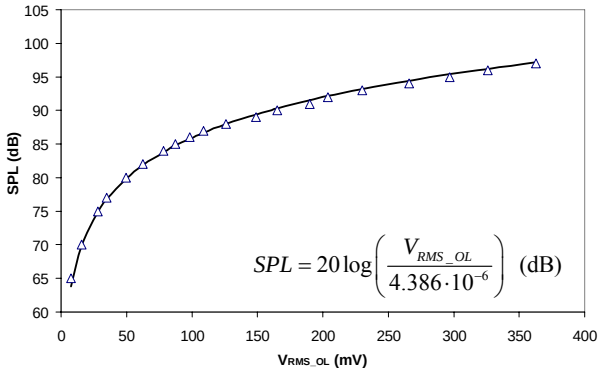

Figure 2: V_{RMS_OL} vs. SPL dependency

2.3. Noise samples

To enable observations of speech production changes between neutral and Lombard speech, natural environment noises and artificial band-noises interfering with typical locations of speech fundamental frequency and first formants occurrence were used. 25 quasi-stationary noises selected from CAR2E database recorded in the cabin of moving car (Pollák, P. et al., 1999) and 4 band-pass noises (62-125, 75-300, 220-1120, 840-2500 Hz) were used. Car noise samples were about 14 seconds long each, stationary band-noises 5 seconds long. The 29 noises were assigned to the session prompts periodically, one noise sample per prompt.

The noise sample was looped in case the utterance would exceed the sample length. All noises were RMS normalized to provide corresponding SPL during the reproduction to the speaker.

2.4. Extensions – speech feedback

In the noisy scenario recordings, speech feedback mixed with the noise was reproduced to the speaker to reduce the attenuation caused by the closed headphones. Level of the speech feedback was adjusted individually to make speaker feel comfortable.

The speech feedback affects significantly speaker's perception of intelligibility of his/her speech production in the noise, so the individual adjustments of its level caused the same speakers to react differently to noises of the same levels. Since the level of the monitored speech in the operator's headphones was derived from the level of the speaker's speech feedback, both operator and speaker were influenced by the feedback adjustment. To eliminate these undesirable variations, we propose a new method of precise speech feedback adjustment.

In general, sound waves propagate to human senses through air passing outer and middle ears and by skull bone vibrations. We presume that wearing closed headphones causes significant attenuation of the sound passing through the ears while the bone conduction transfer remains almost the same (although the mass of the head and head-phones systems differs).

If the attenuation caused by headphones is known, we can adjust the speech feedback to reach the same level of perceived sound like without wearing headphones.

A measurement on a dummy head in the anechoic room using system Pulse v.8 (Brüel & Kjær, 2004) was performed to analyze frequency response of the headphones attenuation and its directionality. The dummy head used for the measurements satisfies ITU recommendation (ITU – P.58, 1996) and models the auditory canals. Monaural directional frequency responses were measured for the dummy without and with headphones, the attenuation characteristic was determined as their subtraction, see Fig. 3.

The measurement was carried out for angles 0-180°, in case of 0° the head face and for 90° the measured ear were directed to the source of the measuring noise. The measurement was not performed for angles greater than 180° as in the anechoic room the sound would spread to the measured ear only by dummy head vibrations and the influence of the headphones mass was supposed to be insignificant.

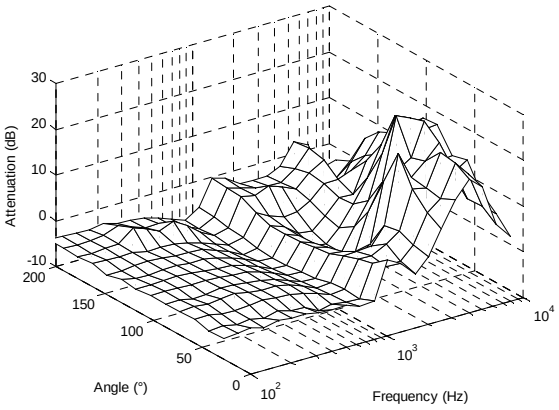

Figure 3: Attenuation by headphones – directional characteristic

The measurement has proven that frequency response of the attenuation depends significantly on the source direction. Attenuation directionality for selected frequencies is shown in detail in Fig. 4. Hence it is obvious that the attenuation depends on the actual room configuration – size, sound absorption coefficients and speaker position.

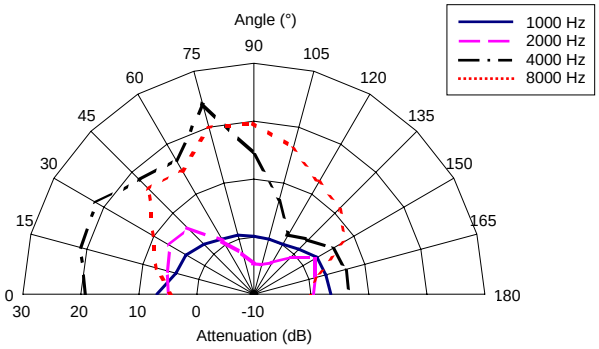

Figure 4: Directional characteristics – detail

As shown in Fig. 3, 4, for low frequencies the attenuation is less significant and also less directional, some harmonic

components are even slightly amplified by the presence of headphones.

In the second step, the frequency response of the attenuation was measured in the recording room for the position where the speakers used to sit during the CLSD'05 recording. In this case, a loudspeaker was placed in front of the dummy head's mouth to simulate voice propagation in the room. Third octave band noises were used for the room excitation in the interval of 80 – 8000 Hz. Monaural transfer functions were measured without and with headphones and attenuation characteristic determined as their subtraction. In Fig. 5, attenuation characteristic for the CLSD'05 recording room is shown and compared with the selected anechoic room ones (0°, 90°, 180°).

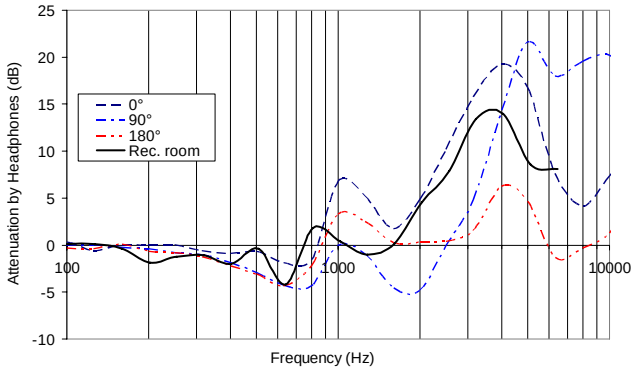

Figure 5: Anechoic and recording room - frequency responses of the attenuation

From the equal loudness curves (Fletcher, 1940) it is known that the maximum sensitivity area for human hearing is around 3-4 kHz and relates to the resonance of the auditory canal. From the presented measurements it can be presumed that the headphones change configuration of the resonator and move the resonant frequency higher, which creates a significant attenuation peak in the former area and significant drop in the area of new resonance.

Once the attenuation impulse response is measured for the recording room and speaker position, transfer function of the preemphasis filter for the speech feedback is determined as the inverse to the attenuation.

2.5. Extensions – signal level reconstruction

During the recording, it is necessary from time to time to modify gain of the microphone preamplifier to avoid signal clipping when speaker changes the voice intensity. In consequence, it is impossible to evaluate voice intensity changes directly from the amplitude of the recorded speech signal.

In case of CLSD'05 the ambient noise can be considered stationary and thus SNR distributions relate to overall vocal intensity changes in neutral and Lombard speech. In Fig. 6, histograms for CLSD'05 microphone channels in neutral (clean) and noisy (LE) scenarios are shown. It is obvious that voice intensity rises significantly for the Lombard speech.

If the information about actual speech signal intensity is required to be preserved, approach shown in Fig. 7 can be used.

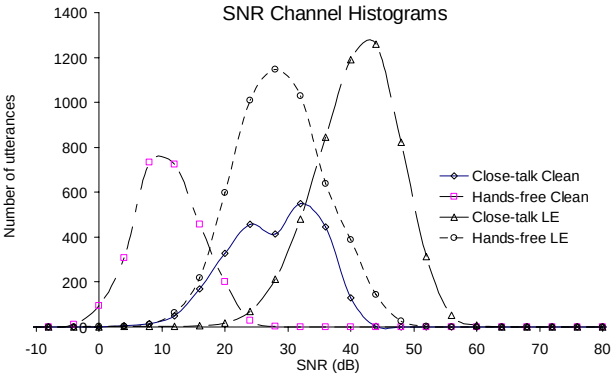

Figure 6: CLSD SNR channel histograms

For the known microphone preamplifier gain the sensitivity V_{ef}/SPL is measured. Then the information about actual preamplifier gain is sufficient for reconstruction of the acoustic signal SPL.

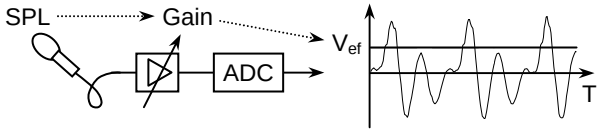

Figure 7: Recorded signal level vs. SPL

3. CLSD recording studio

H&T recorder used for the CLSD'05 acquisition supports two-channel recording and separate noise/speech monitoring for speaker and operator with respect to the virtual distance. To each utterance an item from the noise list is assigned during the recording. Informations about the actual speaker, scenario and conditions are stored in label files that are generated for each recorded utterance.

H&T recorder was implemented as a .NET application. For the purposes of synchronous noise reproduction, speech recording and dual monitoring, the DirectSound functions from the DirectX for Managed Languages library were used (Microsoft, 2005). In Fig. 9, process of speech recording and monitoring is shown.

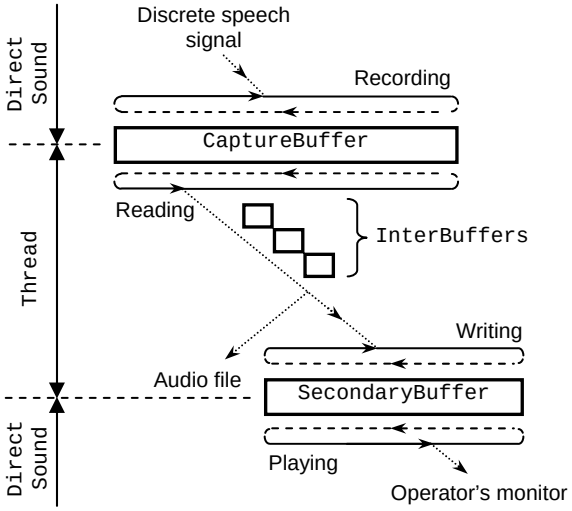

Figure 9: Recording/monitoring implementation

Data from the soundcard input are written cyclically to the `CaptureBuffer`. `Thread` calls automatically a function `ReadCaptureData` which checks actual position of the recording pointer. If there are already new data ready in the `CaptureBuffer`, they are copied to the `InterBuffers`. `Thread` stores data from `InterBuffers` into the `SecondaryBuffer` cyclically. `Thread` assures synchronization of writing and reading of the `SecondaryBuffer` as at the same time, data from `SecondaryBuffer` are read, mixed with the noise samples and played to the output.

4. CLSD'05 in detail

CLSD'05 consists of 26 speakers (12 female, 14 male) who participated both in neutral and noisy scenarios. Recording sessions comprise typically 205 utterances per speaker and scenario, which represents about 10-12 minutes of continuous speech. The number of words uttered by speaker per scenario slightly varies due to the actual items forming the utterance list. In the average, 780 words per speaker and scenario were uttered. The utterance files are stored in a raw file format 16 kHz/16b.

4.1. Corpus

In order to represent whole phoneme material of Czech language, 30 phonetically rich sentences (often complex) appear in each session. To allow statistically significant small vocabulary speech recognition experiments, 470 repeated and isolated digits were included to each session. A complete content of each recording session is shown in Tab. 1.

Corpus contents	Corpus/item id.	Number
Phonetically rich sentences	S01 – 30	30
Phonetically rich words	W01 – 05	5
Isolated digits	CI1 – I4, 30 – 69	44
Isolated digit sequences (8 dig.)	CB1 – B2, 00 – 29	32
Connected dig. seq. (5 dig.)	CC1 – 4, C70 – 99	34
Natural numbers	CN1 – N3	3
Money amount	CM1	1
Time phrases; T1 : analogue, T2 : digital	CT1 – T2	2
Dates: D1 – analogue, D2 – relat. and gen. date, D3 – digital	CD1 – D3	3
Proper name	CP1	1
City or street names	CO1 – O2	2
Questions	CQ1 – Q2	2
Special keyboard characters	CK1 – K2	2
Core word synonyms	Y01 – 95	89
Basic IVR commands	101 – 85	
Directory navigation	201 – 40	
Editing	301 – 22	
Output control	401 – 57	
Messaging & Internet browsing	501 – 70	
Organizer functions	601 – 33	
Routing	701 – 39	
Automotive	801 – 12	
Audio & Video	901 – 95	

Table 1: CLSD'05 session content

4.2. Release notes

The CLSD'05 database as a whole is now available for public use, including phonetic transcriptions and detailed documentation (Download section). The H&T recorder is available for free use upon prior arrangement with the authors.

5. Conclusions

Hardware platform, scenarios and recording tool used for acquisition of the CLSD'05 database were described as well as the database corpus and contents.

During the database recording, difficulties with the appropriate speech feedback adjustment for the speaker were encountered. In this paper, characteristics of sound attenuation caused by wearing headphones were presented and a method of preemphasis speech feedback filter design was proposed.

6. Acknowledgements

The presented work was supported by GAČR 102/05/0278 "New Trends in Research and Application of Voice Technology", GAČR 102/03/H085 "Biological and Speech Signals Modeling", and research activity MSM 6840770014 "Research in the Area of the Prospective Information and Navigation Technologies".

7. References

Bořil, H., Pollák, P. (2005) (1). Design and Collection of Czech Lombard Speech Database. In *Proc. INTERSPEECH '05*, Lisboa, pp. 1577-1580.

Bořil, H., Pollák, P. (2005) (2). Comparison of Three Czech Speech Databases from the Standpoint of Lombard Effect Appearance. In *ASIDE 2005 - Applied Spoken Language Interaction in Distributed Environments*. Grenoble. International Speech Communication Association. Book of abstracts [CD-ROM].

Brüel & Kjær (2004). PULSE X Sound & Vibration Analyzer. http://www.bksv.com/pdf/bu0228.pdf.

Chi, S. M., Oh, Y. H. (1996). Lombard Effect Compensation and Noise Suppression for Noisy Lombard Speech Recognition. In *Proc. ICSLP '96*. Philadelphia. Vol. 4, pp.2013-2016.

Download section. http://noel.feld.cvut.cz/speechlab.

Fletcher, H. (1940). Auditory patterns, Review of Modern Physics, vol. 12, pp. 47-65.

Hansen, J. H. L. (1996). Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. *Speech Communications*, Special Issue on Speech under Stress, 20(2), pp. 151-170.

ITU – P.58 (1996). Recommendation – Head and torso simulator for telephonometry. http://www.itu.int/rec/T-REC-P.58-199608-I/E.

Microsoft DirectX SDK Documentation (2005). http://msdn.microsoft.com/directx/sdk/.

Pollák, P. et al. (1999). Czech Language Database of Car Speech and Environmental Noise. In *Proc. EUROSPEECH '99*. Budapest, Vol. 5, pp. 2263-6.

SPEECON (2001). http://www.speechdat.org/speecon.

Wakao, A. et al. (1996). Variability of Lombard effects under different noise conditions. In *Proc. ICSLP '96.* Philadelphia. Vol. 4, pp. 2009-2012.