ISCA Archive

http://www.isca-speech.org/archive
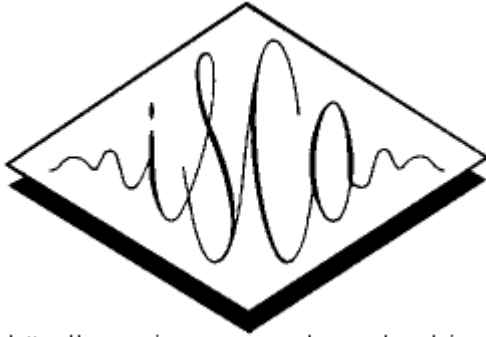
**Interspeech'2005
- Eurospeech**

**Lisbon, Portugal
September 4-8, 2005**

Design and Collection of Czech Lombard Speech Database

Hynek Boril, Petr Pollak

Czech Technical University in Prague, Czech Republic

In this paper, design, collection and parameters of newly proposed Czech Lombard Speech Database (CLSD) are presented. The database focuses on analysis and modeling of Lombard effect to achieve robust speech recognition improvement. The CLSD consists of neutral speech and speech produced in various types of simulated noisy background. In comparison to available databases dealing with Lombard effect, an extensive set of utterances containing phonetically rich words and sentences was chosen to cover the whole phoneme vocabulary of the language. For the purposes of Lombard speech recording, usual 'noisy headphones configuration' was improved by addition of an operator qualifying utterance intelligibility while hearing the same noise mixed with speaker's voice of intensity lowered according to the selected virtual distance. This scenario motivated speakers to react more to the noise background. The CLSD currently consists of 26 speakers.

Design and Collection of Czech Lombard Speech Database

Hynek Bořil & Petr Pollák

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

Abstract

In this paper, design, collection and parameters of newly proposed *Czech Lombard Speech Database (CLSD)* are presented. The database focuses on analysis and modeling of Lombard effect to achieve robust speech recognition improvement. The CLSD consists of neutral speech and speech produced in various types of simulated noisy background. In comparison to available databases dealing with Lombard effect, an extensive set of utterances containing phonetically rich words and sentences was chosen to cover the whole phoneme vocabulary of the language. For the purposes of Lombard speech recording, usual 'noisy headphones configuration' was improved by addition of an operator qualifying utterance intelligibility while hearing the same noise mixed with speaker's voice of intensity lowered according to the selected virtual distance. This scenario motivated speakers to react more to the noise background. The CLSD currently consists of 26 speakers.

1. Introduction

Efficiency of automatic speech recognizers decreases significantly with the presence of ambient noise. Performance is affected negatively both by speech signal corruption by noise and by *Lombard Effect (LE)*. While a lot of attention has been paid to noise suppression in speech signals recorded in adverse conditions, LE classification and elimination is promising further improvements in natural environment speech recognition accuracy.

The LE relates to speaker modifications of speech characteristics in an effort to increase communication intelligibility in noisy environment [1]. Considering speech feature domain, LE introduces nonlinear distortion depending on the speaker and the level and type of ambient noise. Changes of overall vocal intensity, fundamental frequency f_0 contours, variance and distributions as well as variations of formant and antiformant locations, formant bandwidth, spectral tilt and frequency band energy distribution have been observed for LE [2]. Such speech feature changes influence negatively performance of neutral speech trained recognizer.

Basic approaches to Lombard speech recognition can be divided into 3 groups – robust features, LE equalization and model adjustment [1]. First two methods consider use of a neutral speech recognizer with front-end performing speech normalization, third one assumes recognizer training to Lombard speech, which is problematic due to usual lack of sufficient amount of training data and large range of speech feature changes depending on speaker and type of noise.

Goal of the LE analysis is proposal of a degradation model representing relations between Lombard speech and clean speech [1, 3]. If such a relation is found, features or feature equalization more robust to LE can be found.

Recently, numerous multilingual speech databases recorded partly or fully in actual noisy environments are available, e.g. SPEECON (public places and car scenarios) [4]. Strong noise background present in the recordings makes it difficult to evaluate impacts of LE on speech recognition separately. Moreover, in case of Czech SPEECON LE can be observed very rarely, as speakers did not react much to the ambient noise and just read the text [5].

In case of special databases dedicated to LE, noisy background is usually reproduced to the speaker through headphones, hence high SNR of the recorded speech is preserved [3, 6, 1]. Recently, several small vocabulary speech databases fully or partly dedicated to LE are publicly available, e.g. Speech under Simulated and Actual Stress (SUSAS) [1].

In this paper, structure, recording platform and basic parameters of CLSD are presented. The database consists of neutral and Lombard speech recorded in various simulated noisy backgrounds (car noises, artificial band-noises). A total of 26 speakers have recently been recorded. Utterances contain phonetically rich words and sentences covering the whole Czech phoneme vocabulary to allow for overall analysis and modeling of LE. To evaluate properties of the database, analyses of selected LE sensitive speech features were carried out.

2. Database structure

Recently 26 speakers (12 female, 14 male) participated in the noisy background recordings, 12 of them (11 female, 1 male) were recorded in neutral conditions, neutral speech of the rest speakers is covered in the Czech SPEECON database. Each recording scenario typically comprises 108 utterances per speaker, which represents 10 – 12 minutes of continuous speech. The number of words uttered by speaker in one scenario slightly varies due to selected items forming the actual utterance list. In the average, 780 words per speaker and scenario were uttered.

2.1. Corpus and vocabulary

The content of the database is similar to the SPEECON database. Some very specific application utterances as spelled items, internet addresses, spontaneous speech, etc., were omitted. The following items were chosen to be recorded:

> *Phonetically rich material* – sentences and words.
> *Numerals* – isolated & connected digits, natural numbers.
> *Commands* – various application words.
> *Special items* – dates, times, etc.

In order to cover whole phoneme material sufficiently, 30 phonetically rich sentences (often complex) were included into each session. To allow statistically significant small

vocabulary speech recognition experiments, 470 repeated and isolated digits were added to each session. In case of SPEECON, the amount of 40 digits is available per session.

2.2. Label file specification

The label file contains mainly orthographic and phonetic transcription which is completed by the information about recording conditions, speaker information, etc. Our label file originates from the SPEECON one and is extended by items concerning LE conditions – Table 1.

NTY	Noise type	%s	Filenames – including noise description code
NLV	Noise level	%f	The noise level – set by measured level from soundcard output
DES	Speaker-Operator Distance	%f	Distance (m) ⇒ level of speech signal attenuation in operator recording monitor

Table 1: Label file – CLSD specific items

2.3. Noise backgrounds

Background noises were selected for observations of speech production changes both for natural noisy environment and for artificial band-noises interfering with typical locations of f_0 and first formants occurrence. 25 noises recorded in car environment from CAR2E database [7] and 4 band-pass noises (62-125, 75-300, 220-1120, 840-2500 (Hz)) were chosen. Each car noise sample was about 14 *sec* long, stationary band-noises were 5 *sec* long. The noise sample was looped in case the utterance was to exceed the sample length. All noises were RMS normalized to provide corresponding *sound pressure level (SPL)* during the reproduction.

3. Recording platform

The database was recorded digitally into hard disc. In case of the noisy conditions scenario, speaker heard his own voice mixed with noise in closed headphones. The level of the speech feedback was adjusted individually to make speaker feel comfortable. An operator qualified intelligibility of the utterances while listening to noise of the same level mixed with the utterance of intensity lowered in proportion to selected virtual speaker-listener distance.

3.1. Hardware configuration

Recording set, see Figure 1, consists of 2 closed headphones AKG K44 and 2 SPEECON microphones – close talk Sennheiser ME-104 and hands-free Nokia NB2, placed in different distances from the speaker's mouth.
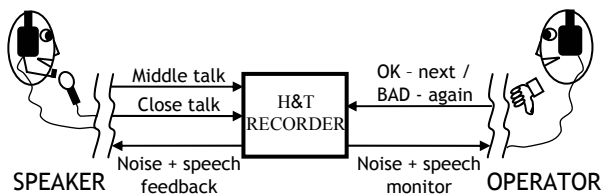

Figure 1: Recording setup

3.2. Noise level adjustment

To enable noise level adjustment, transfer function describing relation between sound card open circuit effective voltage V_{RMS_OL} and SPL in headphones was determined by measurement on a dummy head, see Figure 2. For chosen noise level, corresponding V_{RMS_OL} was set up at the beginning of each session recording,
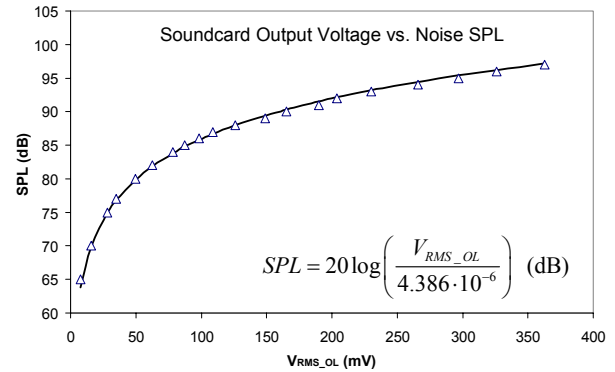

$$SPL = 20\log\left(\frac{V_{RMS\_OL}}{4.386\cdot10^{-6}}\right) \text{ (dB)}$$

Figure 2: V_{RMS_OL} – noise SPL dependency

An average of 90 dB SPL and 3 meters of virtual distance were chosen as default for Lombard speech recording scenarios. In some cases the settings had to be modified according to particular speaker's capabilities.

3.3. Recording studio

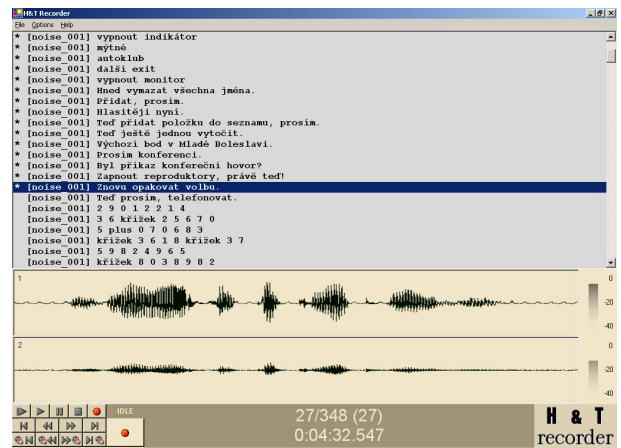H&T recorder developed for CLSD collection was implemented as a .NET application, see Figure 3.

Figure 3: H&T recorder window

H&T recorder supports two-channel recording and separate noise/speech monitoring for speaker and operator respecting virtual distance. To each utterance an item from the noise list is assigned during the recording.

Each recorded utterance was weighted by fading window derived from Blackman window [8]

$$M = \left\lfloor \frac{N-1}{2} \right\rfloor, \qquad (1)$$

$$w[n] = \begin{cases} 0.42 - 0.5\cos\left(\dfrac{2\pi n}{N-1}\right) + 0.08\cos\left(\dfrac{4\pi n}{N-1}\right), & 0 \le n < M, \\ 1, & M \le n < N_u - M, \\ 0.42 - 0.5\cos\left(\dfrac{2\pi n}{N-1}\right) + 0.08\cos\left(\dfrac{4\pi n}{N-1}\right), & N_u - M \le n < N_u \end{cases} \quad (2)$$

where M is length (in samples) of amplitude fade-in and fade-out, N corresponding length of the original Blackman window and N_u length of the whole utterance in samples. Weighting was performed to suppress clicking on the utterance boundaries. An example of harmonic signal amplitude weighting by fading window is shown in Figure 4.
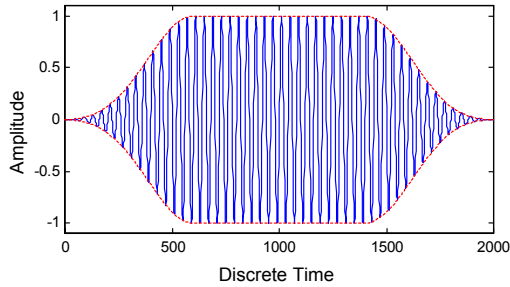

Figure 4: Modified Blackman weighting window

4. Database analyses

Variations of fundamental frequency distribution, first four formant positions and bandwidth and accuracy in digit recognition task were evaluated to measure amount and quality of LE captured in the database. Feature analyses were performed in the open source tool WaveSurfer [9] which provides ESPS algorithms for pitch extraction and formant tracking [10]. For speech recognition recognizer built upon HTK [11] was used.

4.1. Fundamental frequency distribution

Fundamental frequency was analyzed in voiced parts of all neutral and Lombard speech utterances. As shown in Figure 5, significant shift in f_0 distribution can be observed for LE speech. Solid line represents neutral speech and dash line Lombard speech f_0 distribution. Local maxima in both curves relate to major f_0 occurrences in male and female utterances respectively.
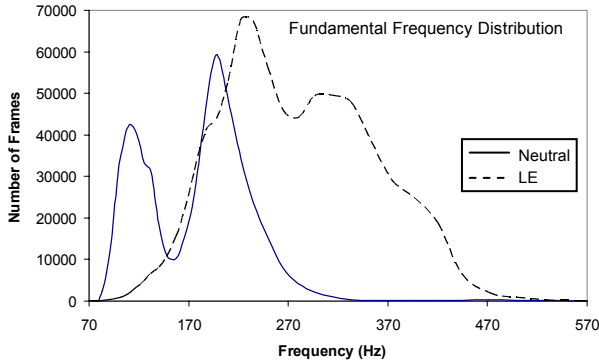

Figure 5: Neutral and Lombard speech f_0 distribution

4.2. Formant tracking

Monophone recognizer trained on 70 SPEECON office sessions was used for the CLSD forced alignment. Monophone models involved 32 mixtures and energy coefficient, 12 mel cepstral coefficients, delta and delta-delta coefficients were chosen as feature vectors. Forced alignment was performed on all CLSD utterances containing digits.

12th order LPC was chosen for formant tracking performed by the WaveSurfer. Information about first four formant frequencies and bandwidths were assigned to corresponding phonemes. As shown in Figure 6, average positions of first two formants vary significantly for selected Czech vowels /a/, /e/, /i/, /o/, /u/ in case of neutral and LE speech.
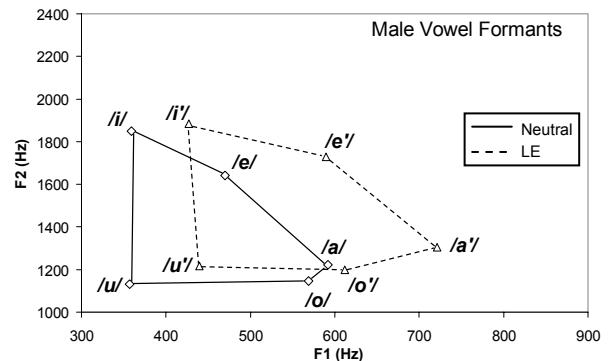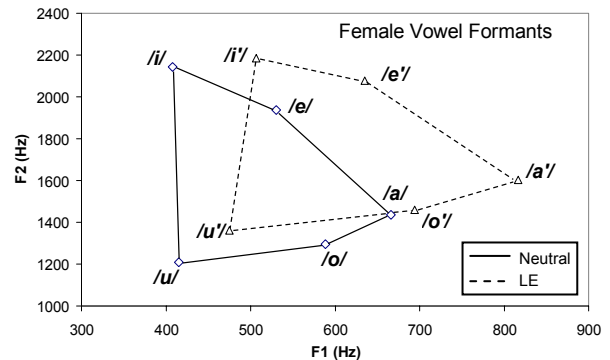

Figure 6: Female & male vowel formants under LE

4.3. Recognition performance under LE

Finally, impact of LE on recognition performance was evaluated. Recognizer mentioned in the previous subsection was used in the digit recognition task. Training set consisted of utterances containing isolated, repeated and connected digits. Neutral testing set was formed by 4930 and 1423 digits and Lombard set included 5360 and 6303 digits uttered by female and male speakers respectively. Recognition results are shown in Table 2, where F denotes female and M male speakers. Word recognition ratio has decreased by 12.5 % for male and by 35.5 % for female speakers.

Data set	*Neutral F*	*Neutral M*	*LE F*	*LE M*
Rec. ratio	92.70%	96.20%	57.18%	83.71%

Table 2: Recognition results – 'Digits' vocabulary

Vowel	Time (s)	F_1 (Hz)	σ_1 (Hz)	F_2 (Hz)	σ_2 (Hz)	B_1 (Hz)	σ_3 (Hz)	B_2 (Hz)	σ_4 (Hz)
/a/	251	651	153	1393	218	239	87	254	78
/e/	295	518	101	1874	267	169	78	198	74
/i/	254	394	55	2063	246	130	52	216	69
/o/	49	584	105	1254	329	244	91	328	106
/u/	61	396	76	1183	275	183	94	307	107

Table 3: Neutral speech average vowel formant positions and bandwidths

Vowel	Time (s)	F_1 (Hz)	σ_1 (Hz)	F_2 (Hz)	σ_2 (Hz)	B_1 (Hz)	σ_3 (Hz)	B_2 (Hz)	σ_4 (Hz)
/a/	615	755	97	1411	196	159	63	166	60
/e/	939	611	91	1894	234	114	49	175	61
/i/	503	457	69	1997	193	118	56	174	60
/o/	147	653	79	1331	215	157	72	210	80
/u/	102	452	82	1266	208	144	76	209	104

Table 4: Lombard speech average vowel formant positions and bandwidths

Such a significant degradation in female speech recognition may be attributed to the fact, that f_0 often shifts under LE into the location of typical neutral speech first formant and formant frequencies rise to locations where they never appeared during the neutral recognizer training.

In Tables 3 and 4 average first two formant positions, bandwidths and corresponding standard deviations are shown as detected for selected Czech vowels in the CLSD. For size reasons, male and female data are presented together in this case.

5. Conclusions

Structure, recording platform and basic parameters of newly proposed Czech Lombard Speech Database are presented in this paper. The database recently consists of neutral speech and Lombard speech produced in simulated noisy conditions by 26 speakers. Covering complete phoneme dictionary of the Czech language, the database focuses on LE analysis and modeling.

To evaluate amount and quality of LE captured in the database, variations of selected speech features sensitive to LE were analyzed. Both f_0 distribution and formants display significant changes in Lombard speech as already known from small vocabulary databases. Recognition ratio for Lombard speech decreased by 12.5 % for male and by 35.5 % for female speakers in digit recognition task, which also proves that CLSD contains challenging data for research in Lombard speech recognition. Sample of the CLSD is available at [12], complete database is available upon prior arrangement.

6. Acknowledgements

7. References

[1] Hansen, J. H. L., "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):151-170, November 1996.

[2] Womack, B. D., Hansen, J. H. L., "Classification of Speech under Stress Using Target Driven Features," *Speech Communications, Special Issue on Speech under Stress*, 20(1-2):131-150, November 1996.

[3] Chi, S. M., Oh, Y. H., "Lombard Effect Compensation and Noise Suppression for Noisy Lombard Speech Recognition", *Proc. ICSLP '96*, 4:2013-2016, Philadelphia, 1996.

[4] www.speecon.com

[5] Bořil, H., "Recognition of Speech under Lombard Effect", *Proc. of the 14th Czech-German Workshop on Speech Processing*, p. 110 – 113, Prague, Czech Republic, 2004.

[6] Wakao, A., Takeda, K., Itakura, F., "Variability of Lombard Effects under Different Noise Conditions", *Proc. ICSLP '96*, 4:2009-2012, Philadelphia 1996.

[7] Pollák, P., Vopička, J., Sovka, P., "Czech Language Database of Car Speech and Environmental Noise," *EUROSPEECH-99*, 5:2263-6, Budapest, Hungary 1999.

[8] Harris, F. J., "On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform", *Proc. IEEE*, 66:51-83, 1978.

[9] Sjölander, K., Beskow, J., "WaveSurfer - an Open Source Speech Tool", *Proc. of ICSLP 2000*, Bejing, China, 2000.

[10] ESPS (Entropic Signal Processing System 5.3.1), Entropic Research Laboratory, http://www.entropic.com

[11] Young, S. et al: The HTK Book ver. 2.2, Entropic Ltd 1999.

[12] http://noel.feld.cvut.cz/speechlab, download section.