

Comparison of Three Czech Speech Databases from the Standpoint of Lombard Effect Appearance

Hynek Bořil & Petr Pollák

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

Abstract

This paper focuses on three Czech speech databases recorded in actual and simulated noisy conditions and explores their suitability for LE analysis and modeling. Parameters of Czech SPEECON, CZKCC car database and newly established Czech Lombard Speech Database (CLSD) are compared. All three databases comprise speech recorded in neutral conditions and speech uttered in noise of the moving car. SNR distribution of the recorded channels, speech fundamental frequency, formant positions and bandwidths, phoneme and word length variations and their overall impact on small vocabulary recognizer's performance are analyzed. It is shown that all three databases display speech feature changes across the recording conditions. In SPEECON database these variations do not affect simple recognition task performance much, in CZKCC and CLSD significant recognition degradation has been observed. Due to results of the feature analyses, CZKCC recognition seems to be corrupted rather by background noise than by LE, while in CLSD only LE affects the recognition as the overall SNR is high.

1. Introduction

In recent days, high demands are put on robustness of automatic speech recognition systems in order to allow for building of voice-controlled interfaces operating reliably in adverse environments. Since recognizer's training can be performed successfully only on a sufficient amount of relevant data, availability of speech databases acquired in the real conditions is fundamental. Performance is affected negatively both by speech signal corruption by noise and by *Lombard Effect (LE)*, caused by speaker modifications of speech characteristics in an effort to increase communication intelligibility in noisy environment [1, 2]. Although recording in natural scenarios assures required presence of environmental noise in the speech signal, authenticity of speaker's responses to actual noisy conditions becomes an issue [3]. During the recording, speakers may tend to concentrate just on correct pronunciation of the text without adequate reaction to actual conditions. In such case, appearance of Lombard speech in the database can be negligible.

In this paper, three Czech speed databases are analyzed on Lombard speech appearance to qualify their suitability for LE analysis and modeling.

2. Databases

Czech SPEECON database, CZKCC car database and newly established Czech Lombard Speech Database (CLSD) are

analyzed and compared. All three databases comprise data recorded both in neutral conditions and actual or simulated noise of the moving car. Same type of close-talk microphone was used across these databases, transfer functions of the recording systems may be considered to be very close.

2.1. Czech SPEECON database

Czech SPEECON database [4] was recorded in public, office, car and entertainment scenarios. Office scenario representing recordings in calm environment (neutral speech) and car scenario introducing recordings in noisy conditions were chosen for the tests. In both environments, speech was recorded by 4 microphones placed in different distances from the speaker. In this paper, close-talk channels providing highest SNR were chosen for the tests.

2.2. CZKCC car database

CZKCC car database [5] comprises utterances recorded in car environment. Recordings in the standing car with engine off stand for neutral conditions and recordings in the moving car with engine on for noisy conditions. Speech was recorded by 2 microphones. In both cases close-talk channel was analyzed.

2.3. CLSD database

CLSD [6] consists of neutral speech and speech uttered in various types of simulated noisy background (CAR2E car noises [7], artificial band-noises). During the Lombard speech recording, usual 'noisy headphones configuration' was extended by presence 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. An average of 90 dB SPL and 3 meters of virtual distance were chosen as a default for Lombard speech recording scenarios. The CLSD currently consists of 26 speakers, each participating both in neutral and Lombard scenario recordings, with the exception of speakers who also participated in SPEECON Office recordings, as these fit CLSD neutral scenario conditions. CLSD was recorded by 2 microphones, in both conditions analyses were performed on the close-talk channel.

3. Analyses

Microphone channel SNRs, fundamental frequency (f_0) distributions, first four formant positions and bandwidths, phoneme and word durations and accuracy in digit recognition task in neutral and noisy conditions were analyzed.

3.1. SNR channel distribution

Arithmetical segmental SNR was evaluated as

$$SNR = 10 \log \sum_{j=1}^L \frac{\hat{\sigma}_{s,j}^2}{\hat{\sigma}_{n,j}^2}, \quad (1)$$

where j is the index of frames with speech activity since the average is evaluated only for short-time frames containing speech. $VAD_j = 1$ for each j . For each short-time frame it is possible to evaluate only the power of mixture $\sigma_{x,i}^2$ as the signal is supposed to contain noise all the time. Powers of speech and noise ($\hat{\sigma}_{s,i}^2$ and $\hat{\sigma}_{n,i}^2$) have to be estimated. Noise power is estimated in speech pause by standard exponential estimation

$$\hat{\sigma}_{n,i}^2 = p \cdot \hat{\sigma}_{n,i-1}^2 + (1-p) \cdot \sigma_{x,i}^2 \text{ for } VAD_i = 0, \quad (2)$$

$$\hat{\sigma}_{n,i}^2 = \hat{\sigma}_{n,i-1}^2 \text{ for } VAD_i = 1. \quad (3)$$

If the speech and noise signals can be considered to be uncorrelated, speech power can be estimated by subtraction of noise power from the mixture power

$$\hat{\sigma}_{s,i}^2 = \sigma_{x,i}^2 - \hat{\sigma}_{n,i}^2. \quad (4)$$

In principle, the algorithm estimates standard global SNR evaluated over speech activity regions only. The segmental approach and the averaging of linear power ratios give lower estimation error [8]. Precision of the estimation is very sensitive to correct VAD classification. Detector based on differential cepstral analysis was used. The details are described in [9].

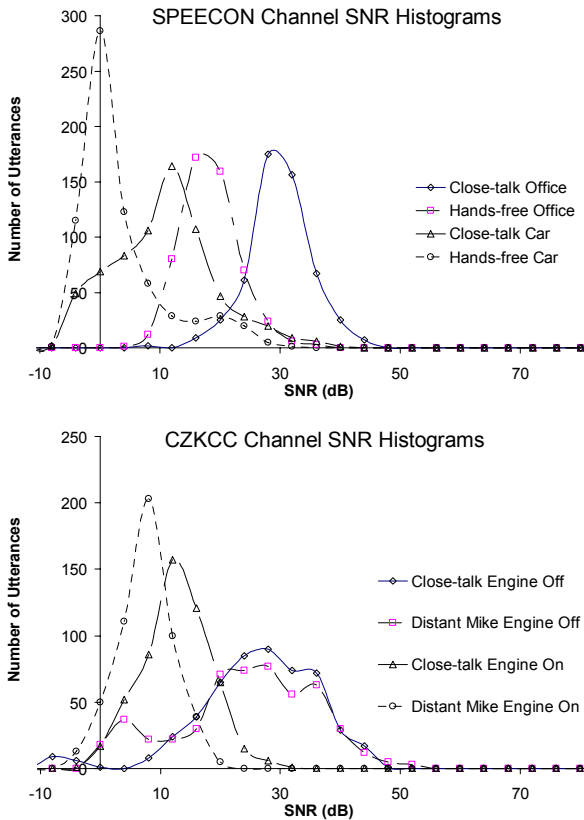


Figure 1: SPEECON and CZKCC channel SNRs.

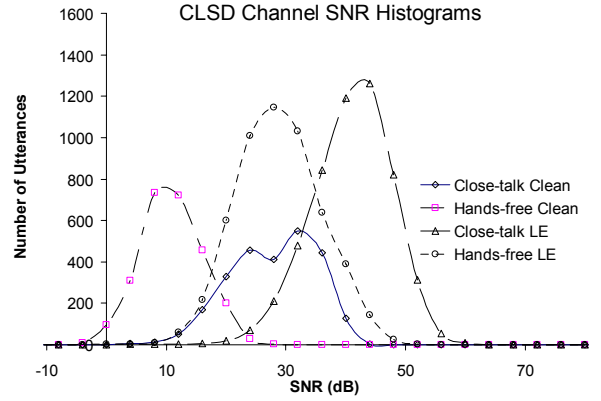


Figure 2: CLSD channel SNRs.

Since CZKCC and CLSD were recorded by two microphones, SPEECON SNR distributions are also depicted only for the first two channels. In CZKCC a directional microphone was used in the distant position, which explains higher average SNR in the distant microphone ‘engine on’ channel than in SPEECON.

Sometimes it is necessary to modify gain of the microphone preamplifier during the recording session to avoid signal clipping when speaker changes voice intensity. In consequence, it becomes impossible to evaluate voice intensity changes directly from the amplitude of the recorded speech signal. In case the ambient noise can be considered stationary, relative voice intensity changes can be estimated from the SNR even with gain being changed during the session. Moreover, if the absolute level of the ambient noise was known, absolute level of vocal intensity could be estimated (but it was not our case).

In SPEECON and CZKCC environmental characteristics changed significantly when comparing office and car or standing car with engine off and moving car scenarios, but in case of CLSD ambient noise can be considered stationary and thus SNR histograms relate to overall vocal intensity changes in neutral and Lombard speech. It is obvious that voice intensity rises significantly for the Lombard speech, see Fig. 2.

3.2. Fundamental frequency

f_0 was tracked in the WaveSurfer [10]. Tracking was performed in voiced parts of all neutral and noisy speech utterances. In the graph descriptions, letters ‘F’ and ‘M’ represent female and male data respectively.

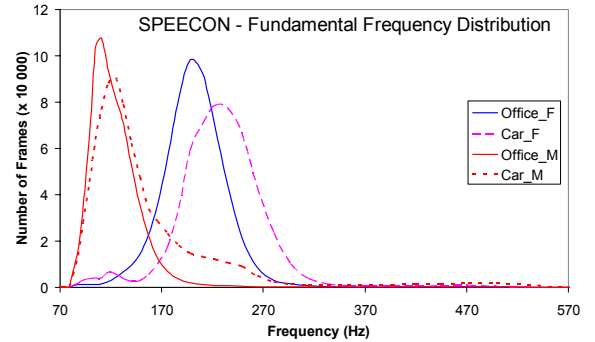


Figure 3: SPEECON f_0 distribution.

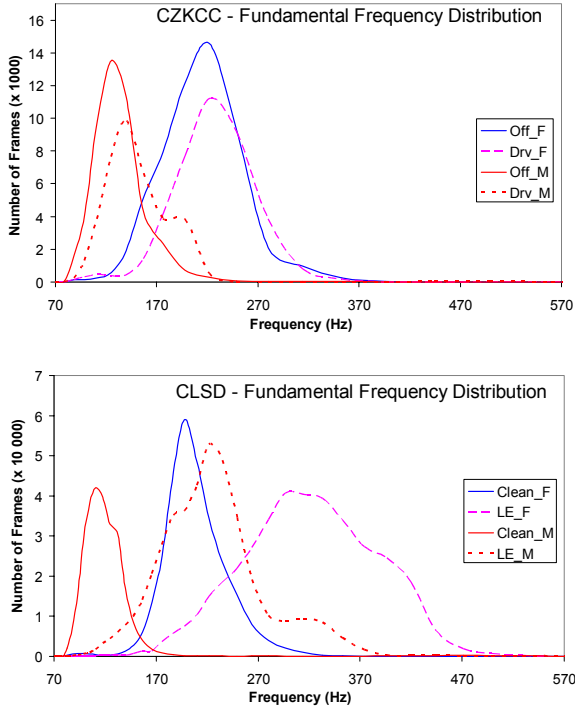


Figure 4: CZKCC and CLSD f_0 distribution.

In case of SPEECON, see Fig. 3, and CZKCC, Fig. 4, shifts in f_0 distribution are observable but not significant. In case of CLSD, Fig. 4, maximum of the LE male f_0 distribution appears at the higher frequency than maximum of neutral female distribution while female maximum moves to location of typical first formant appearance of certain phonemes in neutral speech. During the recognition, f_0 component may be wrongly interpreted as F_1 .

3.3. Formants

Formant analysis was performed on utterances containing digits. Monophone HTK [11] recognizer trained on 70 SPEECON office sessions was used for the forced alignment. 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. In the following figures, positions of first two female formants of the selected vowels appearing in Czech digits are presented.

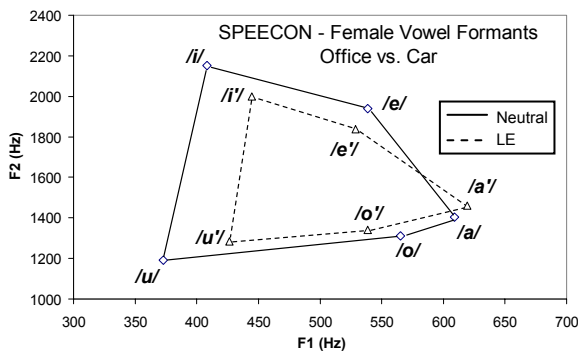


Figure 5: Positions of female F_1, F_2 – SPEECON.

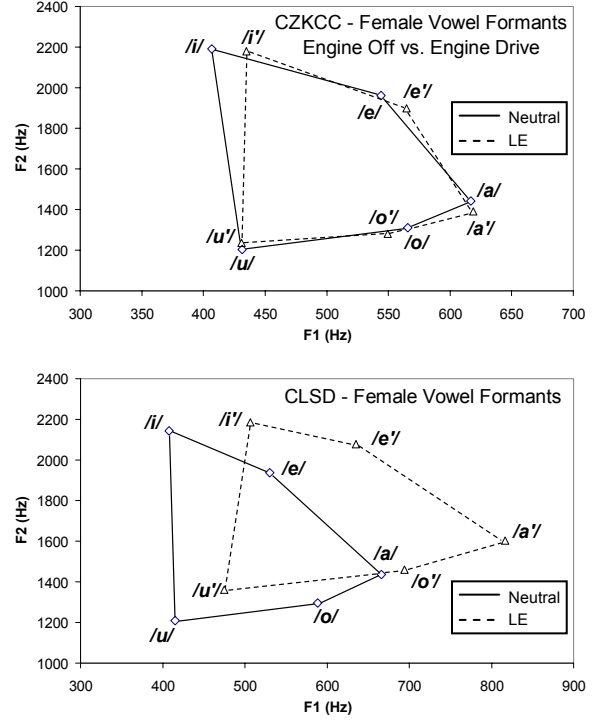


Figure 6: Female F_1, F_2 – CZKCC and CLSD.

Changes in first two formant F_1, F_2 locations can be observed for SPEECON, Fig. 5, and CZKCC, Fig. 6. Formant bandwidths did not display any systematical changes in different scenarios. Significant formant shifts can be observed in the CLSD, Fig. 6. Also significant narrowing of first two formant bandwidths has been observed.

3.4. Phoneme and word durations

Average phoneme durations were evaluated for utterances containing digits. Difference in phoneme duration in the same word uttered in two different scenarios was evaluated as shown in Eq. 5.

$$\Delta = \frac{T_{C_2} - T_{C_1}}{T_{C_1}} \cdot 100 \quad (\%), \quad (5)$$

T_{C_x} represents average phoneme duration in scenario x .

In SPEECON, phoneme duration differences did not exceed 38 %. In case of CZKCC, greatest duration changes were observed in the word 'štiri' (phoneme /t/ – 79 %) and in the word 'sedm' (phoneme /e/ – 73 %). Most significant phoneme duration differences were observed in the CLSD database, e.g. in word 'jedna' (/e/ – 161 %), 'pjet' (/e/ – 174 %), 'devjet' (2nd /e/ – 177 %). No systematical changes in word duration were observed in SPEECON.

Word	# N	T_N (s)	σ_{T_n} (%)	# LE	T_{LE} (s)	$\sigma_{T_{le}}$ (%)	Δ (%)
Nula	349	0,475	11,68	326	0,560	34,48	17,82
Jedna	269	0,559	13,62	251	0,607	26,27	8,58
Dvje	245	0,426	10,56	255	0,483	32,51	13,57

Figure 7: CZKCC - word duration changes.

Data set	SPEECON				CZKCC				CLSD			
	Offc F	Offc M	Car F	Car M	Std F	Std M	Drv F	Drv M	Ntrl F	Ntrl M	LE F	LE M
# Spkrs	22	31	28	42	30	30	18	21	12	14	12	14
# Digits	880	1219	1101	1657	1480	1323	1439	1450	4930	1423	5360	6303
WRR	94.55%	95.73%	95.37%	89.50%	97.03%	97.73%	86.52%	89.59%	92.70%	96.20%	57.18%	83.71%

Figure 8: Digit recognition performance in SPEECON, CZKCC and CLSD.

In CZKCC and CLSD changes in word durations were observed, but did not reach the ratios of phoneme changes, see examples in Fig. 7, 9.

Word	# N	T _N (s)	σ_{T_N} (%)	# LE	T _{LE} (s)	$\sigma_{T_{LE}}$ (%)	Δ (%)
Nula	497	0,397	10,94	802	0,476	15,67	19,87
Jedna	583	0,441	12,78	939	0,527	16,52	19,56
Dvje	586	0,365	11,39	976	0,423	13,82	15,87

Figure 9: CLSD - word duration changes.

It is caused by the fact that some phonemes extend and another shorten their durations while uttered in noisy conditions.

3.5. Digit recognition task

Recognition performance in all scenarios was evaluated. Recognizer mentioned in subsection 3.3 operated in the digit recognition task. Testing set consisted of utterances containing isolated, repeated and connected digits. Column 'Data set' denotes type of scenario (Offc – office, Std – standing car, Drv – moving car, Ntrl – neutral conditions, LE – simulated Lombard conditions), WRR signifies word recognition rate.

In SPEECON, recognition rate decreased by 6 % for male utterances, female recognition rate did not change significantly. In CZKCC, recognition performance decreased by 8 % for male and by 11 % for female utterances. In case of CLSD, recognition rate decreased by 12 % for male and by 36 % for female utterances.

4. Conclusions

In this paper, microphone channel SNRs and speech parameters sensitive to Lombard effect were analyzed and compared for Czech SPEECON, CZKCC and CLSD databases. Speech feature variations for neutral and noisy conditions were observed in all three databases. In case of SPEECON, these variations did not affect remarkably recognizer's performance.

In CZKCC, degradation in recognition efficiency was observed both for male (8 %) and female (11 %) utterances. Since the features of speech uttered in standing car with engine off and moving car did not display significant changes while SNR of the analyzed channel decreased, it can be presumed that the recognition performance was corrupted rather by increase of the adverse noise appearing in the speech signal than by LE.

In case of CLSD, vocal intensity, f_0 distribution, first two formant frequencies and bandwidths, duration of

several phonemes and words displayed significant changes under LE. In digit recognition task performance decreased by 13 % for male and by 36 % for female speakers. It should be emphasized that CLSD utterances are of high SNR and all speakers participated in both neutral and Lombard recording scenarios, thus the degradation can not be assigned to adverse noise in the speech signal neither to speaker variations. Analyses proved that CLSD contains valuable data for research in Lombard speech recognition.

5. 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".

6. 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] 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.
- [4] www.speecon.com.
- [5] www.temic-sds.com.
- [6] Bořil, H., Pollák, P. "Design and Collection of Czech Lombard Speech Database," 1577-1580, *INTERSPEECH-05*, Lisboa, Portugal, 2005.
- [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] Pollák, P., "Efficient and Reliable Measurement and Evaluation of Noisy Speech Background," In proc. of *11th European Signal Processing Conference – EUSIPCO*, Toulouse, 2002.
- [9] Vondrášek, M., Pollák, P., "Methods for Speech SNR Estimation: Evaluation Tool and Analysis of VAD Dependency," *Radioengineering*, 14(1):6-11, 2005.
- [10] Sjölander, K., Beskow, J., "WaveSurfer - an Open Source Speech Tool", *Proc. of ICSLP 2000*, Beijing, China, 2000.
- [11] Young, S. et al: The HTK Book ver. 2.2.