UT-SCOPE: TOWARDS LVCSR UNDER LOMBARD EFFECT INDUCED BY VARYING TYPES AND LEVELS OF NOISY BACKGROUND

Hynek Bořil, John H. L. Hansen[*]

Center for Robust Speech Systems (CRSS), The University of Texas at Dallas, Richardson, TX-75080, USA

ABSTRACT

Adverse environments impact the performance of automatic speech recognition systems in two ways – directly by introducing acoustic mismatch between the speech signal and acoustic models, and indirectly by affecting the way speakers communicate to maintain intelligible communication over noise (Lombard effect). Currently, an increasing number of studies have analyzed Lombard effect with respect to speech production and perception, yet limited attention has been paid to its impact on speech systems, especially within a larger vocabulary context. This study presents a large vocabulary speech material captured in the recently acquired portion of UT-Scope database, produced in several types and levels of simulated background noise (highway, crowd, pink). The impact of noisy background variations on speech parameters is studied together with the effects on automatic speech recognition. Front-end cepstral normalization utilizing a modified RASTA filter is proposed and shown to improve recognition performance in a side-by-side evaluation with several common and state-of-the-art normalization algorithms.

Index Terms— speech recognition, Lombard effect, cepstral compensations, RASTA, UT-Scope database

1. INTRODUCTION

Lombard effect (LE) is known to impact a number of speech production parameters [1]–[5]. Depending on the extent of the production variations, which are often proportional to the level of noise [6], Lombard effect may cause a severe degradation in automatic speech recognition (ASR) performance. This is due to the fact that current speech descriptors (features) used in ASR are highly sensitive to not only linguistic content, but also talking style, emotions [7], and other information captured in the speech signal [2], [8].

Efforts to improve ASR under LE span areas of robust front-end design, equalization of LE speech features towards neutral, improved training methods, and acoustic model adjustments and adaptation; see [2], [5] for overviews. Unfortunately, most approaches assume that sufficient amount of Lombard speech data is available for training the compensations or acoustic models. This may not be realistic in the case of real world applications where the level of background noise and Lombard effect can vary continuously. So far, a majority of studies considering the impact of Lombard effect on speech systems were focused on small vocabulary corpora [2], often collected in a fixed type and level of background noise [1], [5]. The present study focuses on recently acquired large vocabulary speech material – a Lombard portion of the UT-Scope database [9] – collected in several types and levels of simulated background noise. The goal is to analyze speech production variations as a function of type and level of noise, and their impact on automatic speech recognition, with a particular focus on the evaluation of efficiency of currently available and newly proposed cepstral compensation strategies in large vocabulary continuous speech recognition (LVCSR). All compensations considered here make no assumptions about the level and type of background noise or talking style, which makes them candidates for a broad range of applications.

The paper is organized as follows. First, the UT-Scope speech corpus is introduced. Second, results of speech production analysis are presented and discussed. Third, performance of an LVCSR system on speech produced in varying types and levels of noisy background is evaluated and the efficiency of several common and state-of-the-art cepstral normalization strategies are compared together with a newly proposed normalization utilizing a modification of a filter used in the popular RASTA (RelAtive SpecTrA) speech processing procedure [10].

2. LOMBARD PORTION OF UT-SCOPE CORPUS

The UT-Scope corpus [9] consists of speech produced under cognitive and physical stress, emotions, and Lombard effect. The current Lombard portion comprises recordings from 58 subjects, of which 31 are native speakers of US English. All subjects participated in noisy condition recording, where they were exposed to background noise samples through open-air headphones, and also in 'clean' condition recordings with no noise exposure in an ASHA certified sound booth. Three types of noisy backgrounds were used: (i) noise recorded in the car traveling at 65 mph on a highway with windows half open, (ii) large crowd noise, and (iii) pink noise. Highway and crowd noises were produced at the levels of 70, 80, and 90 dB SPL; pink noise was produced at 65, 75, and 85 dB SPL. Each speaker was subjected to a pure-tone hearing test in the range of 100 Hz–8 kHz according to ASHA standards to rule out any subjects with hearing loss. Speech was recorded using three microphones – throat microphone, close-talk Shure Beta-54 microphone, and far-field microphone Shure MX391BP/S with a preamplifier MX1BP, with all recordings sampled at 44,100 Hz.

Each speaker session comprises 100 phonetically balanced read sentences from the TIMIT database [11] produced in clean conditions, and 20 TIMIT sentences produced in each of the nine noise type/level conditions. In addition, each condition contains 5 repetitions of 10-digit strings and approximately one minute of spontaneous speech where subjects had to describe the content of a picture presented on a computer screen.

In this study, only sessions from the 31 US-born subjects (25 females, 6 males) are utilized to eliminate the impact of foreign accent on analysis and recognition tasks. All experiments are conducted on TIMIT-type sentences. While spontaneous segments are undoubtedly valuable since they are expected to represent more natural speech, the authors have observed that subjects tended to be at times unsure of themselves when asked to 'be creative' during the spontaneous recording. To reduce the impact of these arbitrary effects, the spontaneous speech segments are not considered in the present study. The close-talk microphone channel providing high SNR speech recordings is used in all presented experiments. Finally, since Lombard effect was produced via noise exposure through open-air headphones, all speech recordings represent a 'clean' speech signal.

3. SPEECH PRODUCTION UNDER LE

Initial analyses of speech parameters in the Lombard portion of UT-Scope were presented in [9]. In particular, the following parameters were analyzed: sentence duration, duration of silence in speech, durations of broad phoneme classes, low/mid/high energy frame distributions in three noise levels, and spectral tilt. In this section, complemen-

[*] This project was supported in part by USAF under a subcontract to RADC, Inc. under FA8750-09-C-0067. Approved for public release; distribution unlimited.

tary speech production parameters that have direct impact on encoding in speech systems are analyzed: signal-to-noise-ratio (SNR), mean fundamental frequency in utterances, and vowel formant frequencies and durations. SNR, which is estimated using an arithmetical segmental algorithm [12], is expected to reflect vocal intensity changes independently of possible microphone pre-amp gain adjustments, as such adjustments would affect equally the level of speech and background noise in the recordings and preserved their ratio constant. WaveSurfer is used to extract both fundamental frequency (RAPT algorithm) and formant center frequencies (combination of linear predictive modeling of spectral envelope and dynamic programming). Phone boundaries used in the extraction of vowel durations and locations in the formant space are estimated by forced alignment [13] via speech recognizer from Sec. 4.

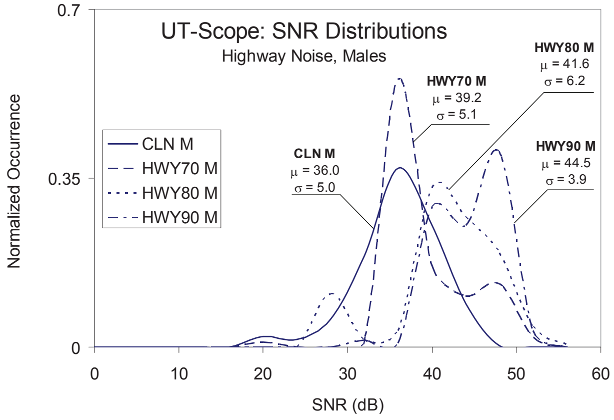

Fig. 1: Speech SNR distributions for various levels of simulated background noise; close-talk microphone.

Analysis of *SNR* distributions (see example for highway noise 'HWY' in Fig. 1) confirms the observations of past studies that vocal intensity in subjects increases with the level of background noise. A relationship between noise level and speech intensity, called *Lombard function* [14], was estimated by fitting a straight line into the SNR distribution means in the SPL (noise sound pressure level in dB)–speech signal SNR plane by means of linear regression. Similar slopes of Lombard function were observed in females (F) and males (M) for the highway noise ($a_F = 0.2$, $a_M = 0.3$) and large crowd noise ($a_{F,M} = 0.1$); in pink noise, the slope was flat in females while increasing in males ($a_F = 0.0$, $a_m = 0.1$). Similar values of Lombard function, ranging from 0.1 to 0.5 were observed in [14].

A consistent increase of *mean utterance fundamental frequency* (F_0) with the level of noisy background was observed for all three types of noise (see Table 1; mean values followed by standard deviations in brackets). Results of correlation analysis in Table 2 suggest a strong correlation between noise presentation level and mean F_0 (a – slope, R^2 – correlation coefficient, MSE – mean square error).

		HWY (dB)			CRD (dB)			PNK (dB)		
Gend	CLN	70	80	90	70	80	90	65	75	85
F	199.8 (52.7)	207.4 (53.2)	216.3 (53.4)	226.2 (55.8)	207.9 (52.0)	215.3 (52.2)	224.0 (54.2)	205.8 (52.4)	213.0 (51.6)	217.7 (51.8)
M	118.2 (26.2)	122.6 (26.3)	134.1 (27.5)	146.5 (29.5)	122.6 (23.9)	133.5 (25.0)	144.0 (27.6)	118.8 (25.7)	124.0 (24.8)	134.6 (27.5)

Table 1: F_0 distributions in varying noise types and levels.

Vowel locations in the first and second formant space F_1–F_2 were estimated by combining formant tracks and phone boundaries obtained from forced alignment. Systematic shifts of vowels in the F_1–F_2 space were observed (e.g., Fig. 2 shows female vowels extracted from speech produced in highway noise). The bars in Fig. 2 represent intervals of one standard deviation for clean 'cln00' and 90 dB SPL highway noise recordings. Formant shifts due to LE were previously observed in low vocabulary corpora [1], [2], [5].

	HWY (dB)	CRD (dB)	PNK (dB)
Gend	70 80 90	70 80 90	65 75 85
F	a=0.938, R^2=0.999 MSE=0.068	a=0.808, R^2=0.998 MSE=0.083	a=0.596, R^2=0.984 MSE=0.380
M	a=1.195, R^2=1.000 MSE=0.039	a=1.073, R^2=1.000 MSE=0.011	a=0.786, R^2=0.962 MSE=1.634

Table 2: Correlation analysis: sound-pressure-level (SPL) vs. F_0.

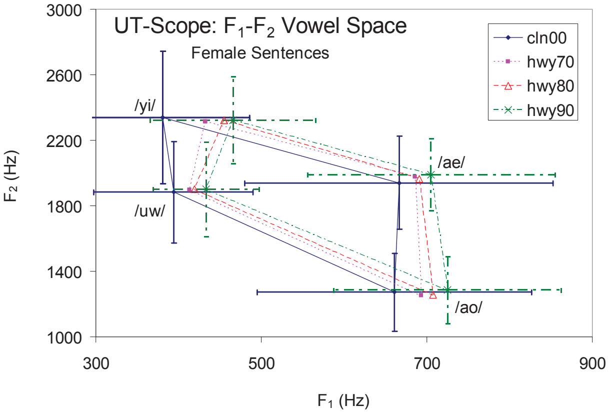

Fig. 2: Vowel shifts in F_1–F_2 space induced by perceived background noise. Mean vowel locations are interconnected by lines to visualize vowel-space transformation due to articulation in noise.

Finally, *vowel durations* in varying noise type and level were analyzed. A consistent increase in duration with the level of noise was observed for some vowels (especially /ae/ and /ao/), however, given the number of subjects, the effects of specific noise types and levels could not be proven to be statistically significant.

4. LVCSR UNDER LOMBARD EFFECT (LE)

We have illustrated the impact of LE on selected speech production parameters in UT-Scope. In this section, the sensitivity of automatic speech recognition (ASR) to such variations is studied. A triphone recognizer combining Hidden Markov Model Toolkit (HTK) based acoustic modeling and trigram language model (LM) implemented with the SRI Language Modeling Toolkit (SRILM) is trained on the TIMIT database (16 kHz) [11]. 13 static mel frequency cepstral coefficients (MFCC), including c_0, and their first and second order time derivatives form the feature vector. At the end of the training phase, 32-mixture triphone models are adapted towards UT-Scope channel/acoustics using combined maximum likelihood linear (MLLR) adaptation and maximum a posteriori (MAP) adaptation on a subset of neutral speech UT-Scope recordings (the database was downsampled from 44.1 kHz to 16 kHz). Speakers from the adaptation set are excluded from the open test set, which contains sessions from 3 male and 19 female subjects. Complete lexical overlap of the open test set and the LM training set eliminates the occurrence of out-of-vocabulary words [15]. Performance of the baseline MFCC system (incorporating cepstral mean/variance normalization – CVN) on the open test set produced in neutral conditions (no noise exposure for speakers) reaches 8.3 % word error rate (WER). A complementary system utilizing perceptual linear prediction (PLP) front-end yielded similar performance – 8.9 % WER.

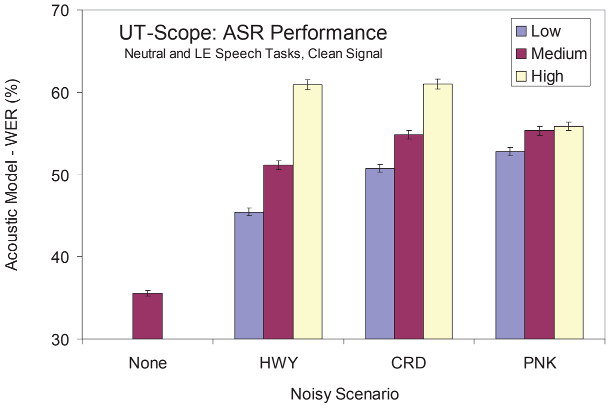

Fig. 3: Performance of baseline LVCSR (MFCC–CVN) as a function of type and level of highway background noise; clean speech signal; word error rates (WER) from acoustic model decoding, LM not engaged.

For the remainder of this paper, only the MFCC system is analyzed. Our preliminary experiments have confirmed a known fact that the recognition performance can be varied drastically based on the LM adjustments. Our goal is to analyze performance of the acoustic front-end rather than optimizing the LM, hence, all results in the rest of the study represent WER's of acoustic model decoding to eliminate the impact of LM. Performance of the MFCC–CVN system on neutral (scenario 'None') and Lombard (highway – 'HWY', large crowd – 'CRD', pink noise – 'PNK') sets is shown in Fig. 3, where 'Low, Medium, High' denote noise presentation levels (no noise exposure in the 'None' set). It can be seen that speech variations observed in previous section result in significant recognition degradation, and that the degradation is proportional to the noise presentation level (error bars denote 95 % confidence intervals). Note that these data sets represent clean speech recordings (see example SNR's in Fig. 1). Since some speech variations under LE can be partly viewed as convolutional distortions (spectral slope flattening, formant shifts, intensity changes), the following paragraphs investigate the efficiency of cepstral compensation methods that incorporate blind deconvolution.

Efficiency of the following cepstral compensations are compared on clean/noisy neutral/LE data sets: cepstral mean normalization (CMN), CVN, RASTA filtering (RASTA) applied in cepstral domain, cepstral gain normalization (CGN) [16], feature warping (Gaussianization) [17], histogram equalization [18], and recently proposed quantile-based cepstral dynamics normalization (QCN) [5]. CGN is similar to CVN where the variance normalization is replaced by dynamic range (estimated from minimum and maximum samples) normalization, and was reported to be effective on noisy signals. QCN builds on the concepts of CVN and CGN; dynamic range of cepstral sample occurrence is estimated from histogram quantiles and subsequently, the histograms are centered to the quantile mean and their variance is normalized to a unit inter-quantile interval:

$$c_{n,i}^{QCNj} = \frac{c_{n,i} - \left(q_j^{Cn} + q_{100-j}^{Cn}\right)/2}{q_{100-j}^{Cn} - q_j^{Cn}}, \qquad (1)$$

where q_j^{Cn} and q_{100-j}^{Cn} are jth and $(100 - j)$th quantile estimates in the nth cepstral dimension. The quantile estimates are obtained on the utterance level. QCN was shown to provide good performance in small vocabulary task on neutral and LE speech in car noise [5]. Feature warping (FW) and histogram equalization (HEQ) alter sample values to match Gaussian or a selected target distribution, respectively, and have the potential to compensate for cepstral distribution variations due to acoustics/channel and additive noise. While HEQ may be more popular in ASR than FW, the concept of Gaussianization preceding GMM-HMM ASR back-end has been previously proven successful in the context of bottleneck features [19]. In our implementation of FW and HEQ, the test utterance cepstral samples are sorted by amplitude in each dimension from smallest to highest, and their amplitudes are subsequently adjusted to match the target cumulative distribution function (CDF). For FW, the target Gaussian CDF is resampled for each test utterance to match the number of frames; for HEQ, the target distribution is provided in a pre-calculated look-up table at various sample lengths and the closest length matching the test utterance is selected (see [5] for details). The target cepstral distributions for HEQ used in this study represent the TIMIT train set distributions.

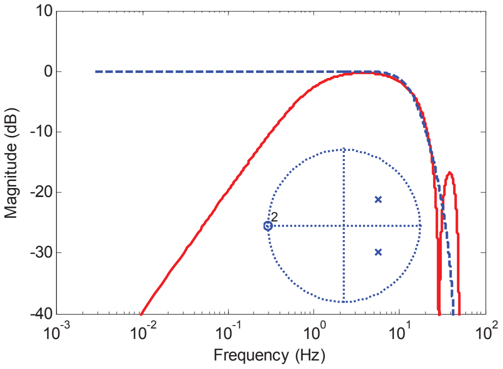

Fig. 4: Dash line – filter approximating low-pass function of original RASTA filter (solid line); unit circle – zeros and poles of proposed low-pass.

Band-pass filtering in RASTA eliminates slow-varying components (including DC) as well as components varying faster than typical for speech. In that sense, RASTA can be seen as CMN combined with low-pass filtering. Other compensations were shown to be superior to CMN in number of applications and hence, it would be convenient to have the option to replace CMN in RASTA by a normalization of choice. Note that engaging other cepstral compensation before RASTA filtering will keep CMN in effect (which will hurt normalizations that align cepstral distributions towards other distribution parameters than mean), and employing them after RASTA may corrupt the previous low-pass smoothing (e.g., adjusting sample amplitudes towards a target distribution may introduce further discontinuities in the coefficient time trajectories). For this reason, we propose a modified filter that approximates the low-pass functionality of RASTA while preserving the slow varying components (see Fig. 4). In that case, any cepstral compensation can be conveniently performed in advance to RASTA without CMN being engaged. The coefficients of the second-order infinite-impulse response (IIR) filter

$$H(z) = \sum_{m=0}^{M} b_m \Big/ \sum_{n=0}^{N} a_n \qquad (2)$$

were obtained from a Butterworth approximation and, assuming standard 10 ms window step, have the following values: $\mathbf{B} = [b_0, b_1, b_2] = [0.10408, 0.20816, 0.10408]$, $\mathbf{A} = [a_0, a_1, a_2] = [1, \text{-}\ 0.90342, 0.31973]$. This filter approximates the side lobe region of the original filter characteristics by a smoothing function. When informally inspecting cepstral tracks filtered by the two filters, the proposed low-pass filter exhibited significant reduction of transient effects compared to the original RASTA filter. The new low-pass filter is combined with QCN, yielding a compensation called QCN_RASTA.

The front-ends were evaluated on two tasks: *(i) clean recordings* – high SNR signals, neutral speech and noise-free Lombard speech produced in 90 dB SPL of highway and crowd noise, and 85 dB of pink noise; *(ii) noisy recordings* – neutral speech and speech produced in 90 dB SPL of highway noise, both mixed with the NOISEX'92 'Volvo' noise at 15 dB and 5 dB SNR. Ten 15 *sec* samples were cut from the original 4 *min* sample and cyclically mixed with the test files to assure variability of the noisy background. The neutral speech open test set comprise 1271 utterances (15489 words) and each presentation noise type/level open test set contains approximately 400 utterances (3500 words). Average WER's for the front-ends on these two tasks are shown in Table 3. It can be seen that all cepstral compensations provided performance improvement over the baseline front-end with no compensation ('none'). The best performance on task *i* was reached by QCN9 (QCN utilizing $9th$ and $91st$ quantiles), followed by QCN4_RASTA (QCN employing $4th$ and $96th$ quantiles). Note that 'plain' QCN4 was outperformed by several compensations and the newly proposed scheme employing low-pass RASTA filtering significantly improved the QCN4 performance. Original RASTA was outperformed by CMN. A close inspection of the cepstral coefficient time trajectories revealed transient effects caused by the band pass filter. The effects were most prominent in the lowest cepstral coefficients and were most probably the cause of performance degradation compared to plain cepstral mean normalization. A total of 26K words forming the complete evaluation set assures the statistical significance of the observed performance differences. Only the 5 best performing normalizations on clean neutral and LE speech were evaluated together with the baseline on the noisy recordings in task *ii*. Here, histogram equalization displayed the best performance, followed by CGN and QCN4 performance. The biggest gains of histogram equalization were observed at lowest SNR's (5 dB) while its superiority gradually reduced with increasing SNR. It can be seen that while especially CGN, histogram equalization, QCN, and newly proposed QCN_RASTA provide significant performance gains in channel/noise/talking style mismatched conditions (see the mismatch in c_0 and c_1 distributions extracted from full-size train/test sets in Fig. 5) over the baseline – in average 10 % absolute WER improvement on both high SNR and noisy recordings, there is no single winner across all conditions representing multiple sources of distortion.

Clean Recordings	
Cepstral Comp.	Across Cond.
none	62.0
RASTA	60.0
warp	55.7
CMN	54.3
QCN4	54.3
HistEq	**53.9**
CVN	**53.3**
CGN	**52.8**
QCN4_RASTA	**52.6**
QCN9	**51.1**

Noisy Recordings	
Cepstral Comp.	Across Cond.
none	77.8
QCN9	69.2
CVN	68.5
QCN4_RASTA	68.4
CGN	67.0
HistEq	64.4

Table 3: Performance of cepstral compensations on clean recordings (left) and noisy recordings (right); WER (%) averaged across conditions, LM not engaged.

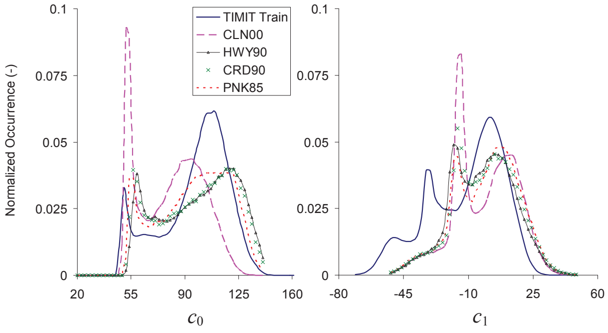

Fig. 5: Cepstral distributions in training and test sets.

5. CONCLUSIONS

This study analyzed the impact of varying types and levels of background noise on speech production parameters in the UT-Scope database and the corresponding consequences for large vocabulary automatic speech recognition. A number of speech parameters were found to vary with the type and level of background noise, which is an observation consistent with earlier studies. A surprisingly strong linear relationship between noise presentation level and mean pitch (in Hz) was observed for large crowd and highway noises. A new version of recently established quantile-based cepstral normalization (QCN) utilizing modified RASTA filtering was presented and shown to improve the original performance. A number of cepstral normalizations were compared in the task of talking style and noisy background mismatch (combined with artifacts of training/testing database mismatch). Especially CGN, histogram equalization, QCN, and newly proposed QCN_RASTA provided significant performance gains in channel/noise/talking style mismatched conditions, however, none of the normalizations managed to outperform others in all conditions considered.

6. REFERENCES

[1] J.-C. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizers," *JASA*, vol. 93, no. 1, pp. 510–524, 1993.

[2] J. H. L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Comm.*, vol. 20, no. 1-2, pp. 151–173, 1996.

[3] H. Bořil, *Robust Speech Recognition: Analysis and Equalization of Lombard Effect in Czech Corpora*, Ph.D. thesis, Czech Technical University in Prague, Czech Republic, http://www.utdallas.edu/∼hynek, 2008.

[4] Y. Lu and M. Cooke, "Speech production modifications produced by competing talkers, babble and stationary noise," *JASA*, vol. 124, no. 5, pp. 3261–3275, 2008.

[5] H. Bořil and J. H. L. Hansen, "Unsupervised equalization of Lombard effect for speech recognition in noisy adverse environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1379–1393, August 2010.

[6] M. Garnier, *Communiquer en environnement bruyant: de l'adaptation jusqu'au forçage vocal [Communication in noisy environments: From adaptation to vocal straining]*, Ph.D. thesis, Univ. of Paris 6, France, 2007.

[7] Z. Callejas and R. López-Cózar, "Influence of contextual information in emotion annotation for spoken dialogue systems," *Speech Communication*, vol. 50, no. 5, pp. 416 – 433, 2008.

[8] Hynek Bořil, Omid Sadjadi, Tristan Kleinschmidt, and John H. L. Hansen, "Analysis and detection of cognitive load and frustration in drivers' speech," in *Interspeech'10*, Makuhari, Japan, Sept. 2010, pp. 502–505.

[9] J. H. L. Hansen and V. Varadarajan, "Analysis and compensation of Lombard speech across noise type and levels with application to in-set/out-of-set speaker recognition," *IEEE Trans. ASLP*, 17(2), pp. 366–378, Feb. 2009.

[10] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Transactions on SAP*, vol. 2, no. 4, pp. 578 –589, Oct. 1994.

[11] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Comm.*, vol. 9, no. 4, pp. 351 – 356, 1990.

[12] M. Vondrášek and P. Pollák, "Methods for speech SNR estimation: Evaluation tool and analysis of VAD dependency," *Radioengineering*, vol. 14, pp. 6–11, 2005.

[13] J. Volín, R. Skarnitzl, and P. Pollák, "Confronting HMM-based phone labelling with human evaluation of speech production," in *Proc. of INTERSPEECH'05*, Lisbon, Portugal, Sept. 2005, pp. 1541–1544.

[14] H. Lane and B. Tranel, "The Lombard sign and the role of hearing in speech," *J. of Speech and Hearing Research*, vol. 14, pp. 677–709, 1971.

[15] P. Motlíček, "Automatic out-of-language detection based on confidence measures derived from LVCSR word and phone lattices," in *Proc. INTERSPEECH'09*, Brighton, UK, Sept. 2009, pp. 1215–1218.

[16] S. Yoshizawa, N. Hayasaka, N. Wada, and Y. Miyanaga, "Cepstral gain normalization for noise robust speech recognition," in *Proc. of ICASSP'04*, May 2004, vol. 1, pp. 209–212.

[17] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *In ODYSSEY-2001*, Crete, Greece, 2001, pp. 213–218.

[18] S. Dharanipragada and M. Padmanabha, "A nonlinear unsupervised adaptation technique for speech recognition," in *ICSLP'00*, 2000, pp. 556–559.

[19] F. Grézl and P. Fousek, "Optimizing bottle-neck features for LVCSR," in *Proc. of ICASSP'08*, Las Vegas, NV, April 2008, pp. 4729–4732.