

Online Noise and Lombard Effect Compensation for In-Vehicle Automatic Speech Recognition

Hynek Bořil, Nitish Krishnamurthy, John H.L. Hansen

Center for Robust Speech Systems, Erik Jonsson School of Engineering & Computer Science, University of Texas at Dallas, U.S.A.

Abstract Presence of background noise in speech impacts the performance of automatic speech recognition (ASR). Adverse noisy environments are also known to induce so-called Lombard effect (LE), where speakers adjust their speech production in order to preserve intelligible communication. LE leads to further ASR degradation, often stronger than the one due to noise. Recently, a set of techniques reducing the impact of noise and LE have been introduced. In this paper, these algorithms are incorporated in a novel ASR setup and evaluated on neutral and Lombard speech corrupted by noise samples from a newly acquired car noise database. It is shown that the proposed scheme provides considerable performance improvement over the baseline and state-of-the-art approaches for all considered car environments and noise levels, reaching 4–17% absolute word error rate (WER) reduction.

Keywords Automatic speech recognition, Lombard effect compensation, UTD-CAR-NOISE corpus.

1 Introduction

Environmental noise and Lombard effect (LE) represent dominant factors degrading automatic speech recognition (ASR) when operated in real-world adverse conditions [1–3]. In the last three decades, prevalent attention has been paid to the development of noise suppression/speech enhancement techniques, while only several studies focused on addressing the impact of LE on ASR (see [1, 3, 4] for overviews). LE affects a wide variety of speech production parameters including vocal intensity, pitch, shape and spectral slope of glottal waveforms, formant locations and bandwidths, vowel-to-consonant energy ratios, and phone and word durations [2–4]. Such variations cause a mismatch between the ASR acoustic models trained typically on neutral speech and the processed LE speech, resulting in ASR degradation.

The efforts to alleviate the LE influence on ASR can be generally divided into three domains: design of feature extraction front-ends less sensitive to LE, transformation of LE speech towards neutral, and acoustic model adjustments and adaptation [1, 3, 4]. While the existing methods reach various degrees of LE suppression, none of them are

capable of complete elimination of the factors degrading ASR. Moreover, most of the methods assume that there is a sufficient amount of labeled LE samples available for compensation training, and that the level of LE (the rate of speech production parameter variations) in the recognized speech is constant. These assumptions may not be met in real-world scenarios, where the level and type of noise and LE can change in time.

Recently, novel online frequency and cepstral domain compensations have been proposed and shown to outperform conventional approaches in LE/noisy speech recognition tasks [5, 6]. The compensations do not require any training samples and make no *a priori* assumptions about the level of noise and LE in the speech signal. The compensation parameters are estimated on-the-fly for each utterance. In this paper, the compensations are incorporated in a novel recognition scheme and evaluated on neutral and Lombard speech mixed with noise from the newly acquired car noise database.

The remainder of the paper is organized as follows. First, the experimental setup including the description of the car noise/Lombard speech corpora and baseline ASR engine used in the experiments is discussed. Second, the novel ASR scheme incorporating LE compensations is introduced and evaluated together with baseline systems.

2 Experimental Setup

In this section, the newly acquired UTD-CAR-NOISE database is presented, followed by the description of the test neutral/LE speech corpus and the speech recognition setup.

2.1 UTD-CAR-NOISE Corpus

A new car noise corpus called UTD-CAR-NOISE has been designed to cover a range of vehicle types and driving situations. The corpus consists of recordings from 30 vehicles (20 sedans, 5 SUV's, and 5 pickup trucks) under the following conditions:

- NAWC: no air-conditioning (A/C), windows closed,

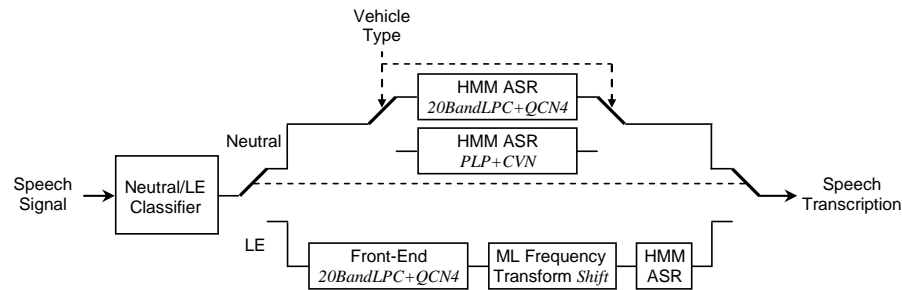


Fig. 1. Proposed compensation scheme.

- ACWC: A/C engaged, windows closed,
- NAWO: A/C off, windows open,
- HNK: windows closed, car horn,
- TRN: turn signal engaged,
- IDL: engine idling,
- REV: engine revving,
- LDR/RDR: left/right door opening and closing.

A specific driving route was fixed for the acquisition to assure comparable conditions, such as external environmental noises, across the recorded vehicles. The route was selected to consist of a combination of 6-lane city roads with higher traffic density and 2-lane concrete community roads with lower traffic density, yielding a total of 4 miles per session. The average speed of the cars during the recording ranged around 40 mph. The car noise acquisition was conducted outside rush hours to minimize non-stationary external traffic noise. The data was recorded using a Shure MX 391S far-field microphone fixed on the driver side sun-visor. The major portion of the recording time was dedicated to the scenarios with windows closed or windows open and A/C off, since these are believed to represent the most typical driving conditions.

2.2 Speech Material and Recognition Framework

In the speech recognition experiments, the Czech Lombard Speech Database (CLSD'05) [4] is used. The database comprises recordings of neutral speech and LE speech, i.e., speech produced in simulated noisy conditions (90 dB SPL of car noise produced to speakers through headphones). Speech was sensed by a close-talk microphone, resulting in high signal-to-noise ratios of the recordings (mean SNR of 28 dB in neutral sessions, 41 dB in LE sessions). An operator was listening to the speaker while being exposed to the same noisy background. If an utterance was not intelligible, the operator requested a repetition to ensure that the message is conveyed over the noise.

For the purpose of the present experiments, the recordings were downsampled to 8 kHz and filtered by a G.712 telephone filter. Clean speech recordings were mixed with selected UTD-CAR-NOISE noise samples at SNR's of -5, 0, ..., 20, ∞ dB,

where ∞ dB represents clean data with no noise added. In particular, noise samples from three vehicles, a 2000 Ford Mustang sedan, a 1998 Chevrolet Blazer SUV, and a 2005 Chevrolet Silverado pickup truck, were used in the experiments. The selected driving scenario was 'car in motion, ACWC'.

The recognizer used in experiments employs HMM acoustic models that comprise 43 context-independent monophone models and two silence models (3 emitting states, 32 Gaussian mixtures). The speech signal is divided into 25 ms segments with 10 ms overlap and encoded into 13 static cepstral coefficients c_0 - c_{12} and their first and second order time derivatives. Gender-dependent phone models are trained in 46 iterations on large vocabulary material from 37 female/30 male Czech SPEECON database sessions. The task is to recognize 10 Czech digits (16 pronunciation variants) in connected digits utterances. The female neutral/LE test sets contain 4930/5360 words, respectively, from 12 speakers, and the male neutral/LE test sets contain 1423/6303 words from 14 speakers (per each SNR level).

In the baseline systems, mel frequency cepstral coefficients (MFCC), perceptual linear prediction (PLP) cepstral coefficients, and Expolog [1] cepstral coefficients are used. While MFCC represents a common ASR front-end and PLP its most frequent alternative, Expolog-based cepstral coefficients is a state-of-the-art coding scheme that has been shown to provide superior performance on stressed and LE speech compared to MFCC and PLP [1, 4]. In addition, recently introduced front-end 20BandLPC is employed in the experiments. 20BandLPC is derived from PLP by replacing the trapezoid filterbank with a bank of 20 non-overlapping rectangular filters distributed uniformly in the frequency band of 0-3200 Hz. In [5, 6], 20BandLPC was shown to outperform MFCC and PLP front-ends on LE speech. In all recognition setups, the front-ends are accompanied by cepstral normalization. In particular, MFCC and PLP are followed by cepstral variance normalization (CVN) [7] and 20BandLPC by novel quantile-based cepstral dynamics normalization (QCN), which is described in the next section (see also [5]).

3 Proposed Approach

In [5], the quantile-based cepstral dynamics normalization (QCN) has been introduced. QCN is defined:

$$c_{n,i}^{QCN} = \frac{c_{n,i} - (q_4^{Cn} + q_{96}^{Cn})/2}{q_{96}^{Cn} - q_4^{Cn}}, \quad (1)$$

where n is the cepstral dimension, i is the index of the cepstral sample in the time window, and q_4^{Cn} and q_{96}^{Cn} are the 4% and 96% quantiles of the n -th dimension cepstral sample distribution. QCN determines the dynamic range of the cepstral samples occurrence from the sample distribution quantiles, normalizes it to unity, and centers it to the quantile mean. QCN was shown to outperform cepstral mean (CMN) [8], cepstral variance [7], and cepstral gain (CGN) [9] normalizations in ASR tasks on noisy neutral and LE speech.

Also in [5], the novel frequency transformation *Shift* was shown to address more accurately the formant shifts due to LE than vocal tract length normalization (VTLN) techniques [10]. *Shift* is derived from maximum likelihood VTLN, where the original scalar frequency warping is replaced by a spectral translation $F_{Shift} = F + \beta$. For each utterance, the translation parameter β is obtained in so called *fully optimized* search as introduced in [10]. The search consists of three steps. First, for each $\beta \in \Psi$, where Ψ is the search grid chosen $\Psi = \{0, 100, 200, 300\}$, the test utterance is transformed: $\mathbf{O}_{utt} \rightarrow \mathbf{O}_{utt}^\beta$. Second, the transformed utterance is decoded using acoustic model λ :

$$\hat{\mathbf{W}}_{utt}^\beta = \arg \max_{\mathbf{W} \in L} \left[Pr(\mathbf{O}_{utt}^\beta | \mathbf{W}, \lambda) Pr(\mathbf{W} | \Theta_l) \right], \quad (2)$$

where \mathbf{W} is the sequence of words for language L , and Θ_l is the language model. Third, the β_{max} that maximizes decoding likelihood is found:

$$\beta_{max} = \arg \max_{\beta} \left[Pr(\mathbf{O}_{utt}^\beta | \hat{\mathbf{W}}_{utt}^\beta, \lambda) Pr(\hat{\mathbf{W}}_{utt}^\beta | \Theta_l) \right]. \quad (3)$$

Finally, the resulting transcription $\hat{\mathbf{W}}_{utt}^{\beta_{max}}$ is taken.

The compensations QCN and *Shift* were incorporated in [5] in a codebook decoding scheme. Here a set of HMM's are trained on speech of various SNR's. In the decoding stage, models that best match the current noise level in speech are searched. The codebook decoding considerably improves recognition of noisy speech, but at the costs of a high number of decoding passes conducted per utterance. In order to reduce the search grid for best matching speech-in-noise models, GMM based noisy model selection was introduced in [6].

In this paper, a novel recognition scheme employing clean speech-trained acoustic models is presented (see Fig. 1). In the initial stage,

neutral/LE speech classification is conducted. The neutral/LE classifier utilizes two Gaussian mixture models (GMM's) comprising 32 mixtures each. The GMM's are trained on clean neutral and LE speech, which is parameterized by 20BandLPC into 13 cepstral coefficients and their first and second order time derivatives. Since the *Shift* transform compensates for the formant shifts due to LE, it is not conducted for the speech samples classified as neutral. Depending on the vehicle, either 20BandLPC and QCN are used for neutral speech coding (Mustang) or PLP and CVN (Blazer, Silverado). The different coding strategies were assigned to the vehicles based on a preliminary ASR experiment comparing front-end efficiency in varying types of noise. Speech classified as LE is always parameterized by 20BandLPC, followed by QCN and the *Shift* transform. Compared to [5], where fourteen decoding passes were required for each utterance, here only a single decoding pass through ASR acoustic models is conducted for speech classified as neutral and four decoding passes are conducted for LE speech, considerably reducing the computational costs.

4 Evaluation

The experimental setup is organized as described in Sec. 2.2. In the first step, the novel system depicted in Fig. 1 is compared to the baseline recognizer employing MFCC and CVN.

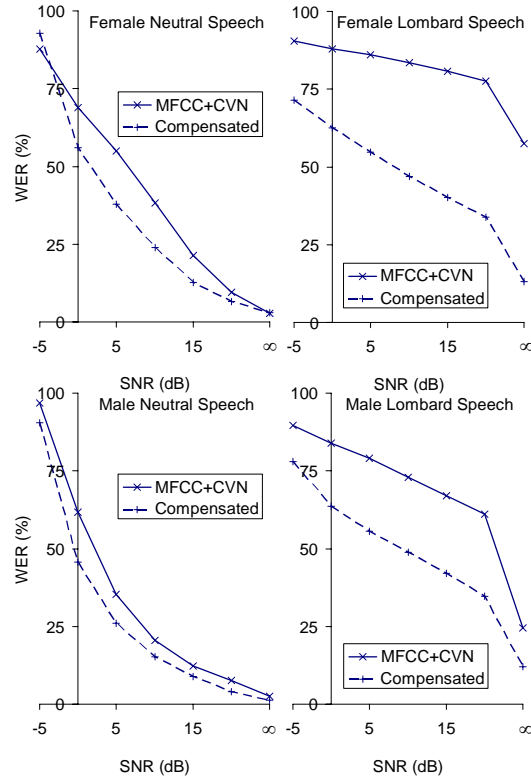


Fig. 2. Performance of baseline and proposed system; neutral and LE data; Mustang.

The evaluation is conducted on neutral and LE speech corrupted by various levels of noise from the Mustang environment. The performance is measured by the means of word error rate (WER; see Fig. 2; the novel system is denoted ‘Compensated’). It can be seen that for both genders, the proposed system significantly outperforms baseline on neutral and LE speech in a wide range of SNR’s.

In the second step, the compensated system is compared to baseline MFCC, PLP, and Expolog recognizers in three noisy environments – Mustang, Blazer, and Silverado, where the vehicle noises are mixed with clean speech samples at SNR’s from -5 dB to ∞ dB as described in Sec. 2.2. The performance improvement of the novel scheme over each baseline system is evaluated:

$$\Delta WER = \frac{\sum_{Style, SNR} (WER_{Style, SNR}^{Baseline} - WER_{Style, SNR}^{Compensated})}{N_{Styles} \times N_{SNR's}}, \quad (4)$$

where $Style \in \{Neutral, LE\}$ and N_{Styles} and $N_{SNR's}$ are the number of talking styles (2) and SNR conditions (7) in which the systems are tested. In other words, ΔWER represents mean WER reduction provided by the compensated system over the baseline across all noise and talking style conditions (including both genders).

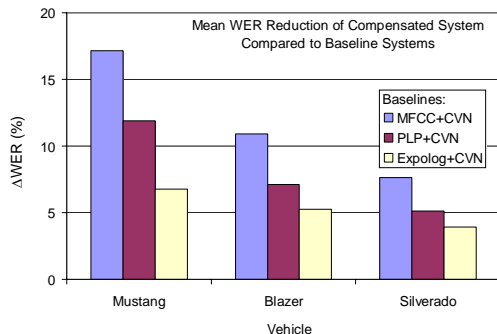


Fig. 3. Mean performance improvement of proposed system over baseline systems.

TABLE I. Mean performance improvement of proposed system over baseline systems; ΔWER (%).

Cond.	Baseline	Vehicle		
		Mustang	Blazer	Silverado
N+LE F+M	MFCC+CVN	17.2	10.9	7.6
	PLP+CVN	11.9	7.1	5.1
	Expolog+CVN	6.8	5.3	3.9

It can be seen that the novel compensated system outperforms the baselines in all three vehicle environments and that the WER improvement is consistently highest for the MFCC baseline and

gradually decreases for PLP and Expolog ones. The superior performance of the last system amongst the baseline setups could have been expected since the Expolog front-end was designed to deal with noisy stressed and LE speech.

5 Conclusions

This study has presented a novel noise and LE compensation scheme that incorporates GMM based neutral/LE classifier and frequency and cepstral domain normalizations. The proposed system is evaluated on neutral and LE speech corrupted by various levels of noise acquired in a sedan, SUV, and pickup truck environment. It is shown that the proposed scheme significantly outperforms all baseline systems, yielding mean WER improvements ranging from 3.9% to 17.2%.

References

- [1] S. E. Bou-Ghazale and J. H. L. Hansen, “A comparative study of traditional and newly proposed features for recognition of speech under stress,” *IEEE Transactions on Speech & Audio Processing*, vol. 8, no. 4, pp. 429–442, 2000.
- [2] J.-C. Junqua, “The Lombard reflex and its role on human listeners and automatic speech recognizers,” *The Journal of the Acoustical Society of America*, vol. 93, no. 1, pp. 510–524, 1993.
- [3] 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.
- [4] H. Bořil, “Robust speech recognition: Analysis and equalization of Lombard effect in Czech corpora,” Ph.D. dissertation, Czech Technical University in Prague, Czech Republic, <http://www.utdallas.edu/~hxb076000>, 2008.
- [5] H. Bořil and J. H. Hansen, “Unsupervised equalization of Lombard effect for speech recognition in noisy adverse environment,” in *Proc. of ICASSP’09*, Taipei, Taiwan, April 2009.
- [6] H. Bořil and J. H. Hansen, “Reduced complexity equalization of Lombard effect for speech recognition in noisy adverse environments,” accepted to *Interspeech’09*, Brighton, UK, September 2009.
- [7] O. Viikki and K. Laurila, “Noise robust HMM-based speech recognition using segmental cepstral feature vector normal.” in *ESCA-NATO Workshop on RSR*, vol. 1, 1997, pp. 107–110.
- [8] J. P. Openshaw and J. S. Mason, “Optimal noise-masking of cepstral features for robust speaker identification,” in *ASRIV-1994*, vol. 1, 1994, pp. 231–234.
- [9] S. Yoshizawa, N. Hayasaka, N. Wada, and Y. Miyanaga, “Cepstral gain normalization for noise robust speech recognition,” in *Proc. of ICASSP’04*, vol. 1, May 2004, pp. 1–209–12 vol.1.
- [10] L. Welling, H. Ney, and S. Kanthak, “Speaker adaptive modeling by vocal tract normalization,” *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 6, pp. 415–426, Sep 2002.