Front-End Compensation Methods for LVCSR Under Lombard Effect

Hynek Bořil[1], *František Grézl*[2], *John H.L. Hansen*[*1]

[1]Center for Robust Speech Systems (CRSS), Eric Jonsson School of Engineering,
University of Texas at Dallas, Richardson, Texas, U.S.A.
[2]Speech@FIT, Brno University of Technology, Czech Republic

Abstract

This study analyzes the impact of noisy background variations and Lombard effect (LE) on large vocabulary continuous speech recognition (LVCSR). Robustness of several front-end feature extraction strategies combined with state-of-the-art feature distribution normalizations is tested on neutral and Lombard speech from the UT-Scope database presented in two types of background noise at various levels of SNR. An extension of a bottleneck (BN) front-end utilizing normalization of both critical band energies (CRBE) and BN outputs is proposed and shown to provide a competitive performance compared to the best MFCC-based system. A novel MFCC-based BN front-end is introduced and shown to outperform all other systems in all conditions considered (average 4.1% absolute WER reduction over the second best system). Additionally, two phenomena are observed: (i) combination of cepstral mean subtraction and recently established RASTA$_{LP}$ filtering significantly reduces transient effects of RASTA band-pass filtering and increases ASR robustness to noise and LE; (ii) histogram equalization may benefit from utilizing reference distributions derived from pre-normalized rather than raw training features, and also from adopting distributions from different front-ends.

Index Terms: speech recognition, Lombard effect, UT-Scope database, bottleneck features, quantile-based cepstral distribution normalization, histogram equalization.

10.21437/Interspeech.2011-100

1. Introduction

Acoustic signal variations due to the presence of environmental noise as well as speech production adjustments introduced by speakers to communicate effectively in noise, called *Lombard effect* (LE), are known to degrade the performance of automatic speech recognition (ASR) [1, 2]. While an extensive effort to increase ASR robustness to noise has been carried out by the speech community, the impact of speech changes due to LE has received far less attention. A number of speech production parameters are affected by LE [1–5] and their drift from neutral (modal) speech values introduce mismatch with the ASR acoustic models trained typically on neutral speech. So far, the efforts to reduce the impact of LE on ASR have been mostly limited to small vocabulary ASR tasks (see [6] for overview). Many techniques increasing ASR robustness to adverse environments and speaker variability are available, operating in the acoustic features domain (e.g., noise suppression, modified feature extraction filter-banks, cepstral distribution normalizations, temporal filtering, etc.), acoustic model domain (e.g., acoustic model adaptation, parallel model combination, model codebooks), and 'in-between' (e.g., vocal tract length normalization, VTLN, transforming features in frequency domain based on the likelihoods from acoustic models). Some of them were shown to be partially successful also in reducing the impact of LE, e.g., modified filter-banks [7,8], cepstral normalizations and VTLN [6], and temporal filtering [9].

While additional pre- and post-processing stages, such as noise suppression or VTLN, are expected to further boost the system performance, it is believed that careful front-end design conducted in the initial stage will provide a good asset for building a fully-loaded LE-robust LVCSR system. In our recent study [9], cepstral compensations and newly proposed modification of the popular RASTA [10] (relative spectra) temporal filtering were shown to improve robustness of a MFCC-based recognizer exposed to noise and LE. The objective of this paper is to compare performance of several existing feature extraction strategies combined with state-of-the-art cepstral compensations and temporal filtering when being exposed to varying speech modality (neutral vs. LE speech) and various levels of environmental noise in LVCSR tasks. In particular, the focus is on mel frequency cepstra (MFCC) [11], perceptual linear predictive (PLP) cepstra [12], perceptually motivated minimum variance distortionless response (PMVDR) cepstra [13], and bottleneck (BN) features [14]. It is noted that to our best knowledge, this represents the first effort to evaluate the BN performance in varying noise conditions and Lombard effect in LVCSR context. BN features are extracted from a hidden layer of a neural network (NN) that makes use of a longer temporal context of critical band energies (CRBE) or cepstral features. Inspired by the encouraging performance improvements observed for 'non-mainstream' combinations of cepstral features and normalizations in our initial experiments, we explore the use of similar normalization strategies on CRBE within the BN framework. In addition, we propose to extend the BN front-end for additional normalization of BN outputs, which is shown to provide further performance gains. Finally, a front-end that utilizes the best performing normalized cepstra as inputs to the BN framework is proposed and shown to provide superior performance to all other front-ends across all evaluation conditions.

2. Feature Extraction Front-Ends

2.1. Feature Normalizations

Feature distribution normalizations, typically applied to log spectral energies or cepstral coefficients, are popular means to reduce the impact of speaker and channel variations and environmental noise on speech systems.

The following feature distribution normalizations are utilized in our study: cepstral mean normalization (CMN), cepstral mean–variance normalization (CVN), cepstral gain normalization (CGN) [15], recently proposed quantile-based cepstral dynamics normalization (QCN) [6], feature warping (Gaussianization, FW) [16], histogram equalization [17], relative spectral (RASTA) filtering [10], and recently proposed RASTA$_{LP}$ filtering [9]. Due to the deconvolution properties of log spectra/cepstra, convolutional signal distortions caused by changes in environmental acoustics, microphone/channel path, as well as speech production changes (speech intensity, spectral slope, etc.) can be modeled as shifts in feature distribution means and variances. Furthermore, the presence of additive noise has also direct impact on distribution means and contours [6]. CMN and CVN are popular means used to reduce the impact of distribution mean and variance changes. CGN is a modification of CVN where rather than variance, an amplitude interval bound by the global maximum and minimum in the speech segment is normalized. CGN was found to provide superior performance to CVN in ASR tasks on noisy neutral [15] and LE speech [9]. QCN extends the concept of CVN and CGN; the dynamic range of cepstral sample occurrence is estimated from histogram quantiles and subsequently, the samples are normalized to a zero inter-quantile mean and unit inter-quantile distance. This method is motivated by the observation that mean and variance normalization

*This project was funded by AFRL through a subcontract to RADC Inc. under FA8750-09-C-0067 (Approved for public release, distribution unlimited), and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. Hansen. F. Grézl was supported by Grant Agency of Czech Republic projects No. GP102/09/P635 and 102/08/0707, and by BUT FIT grant No. FIT-11-S-2.

Copyright © 2011 ISCA

may not be as efficient in aligning training and test distributions of different skewness (caused by signal distortions), while normalizing distributions by their selected low and high quantiles can assure good 'dynamic range' alignment for two distributions of any shape. Feature warping (FW) and histogram equalization (HEQ) normalize all distribution moments – towards Gaussian or reference distribution, respectively. RASTA band-pass filtering aims at suppressing speech signal components that vary either too slow or too fast to be attributed to speech. The slow varying component suppression can be viewed also as another realization of CMN (CMN suppresses a DC component). In our recent study [9], a modified RASTA filter that approximates only the low-pass portion of the original RASTA by a smoothing function was introduced (denoted here as RASTA$_{LP}$). The goal is to remove the 'CMN function' from RASTA and allow for combining other types of normalization that do not necessarily center distributions to their means (such as QCN) with the filtering of fast varying components due to noise. Compared to RASTA, RASTA$_{LP}$ is realized by a filter of significantly lower order, which helps reduce transient effects typical for RASTA filtering (see Fig. 1).

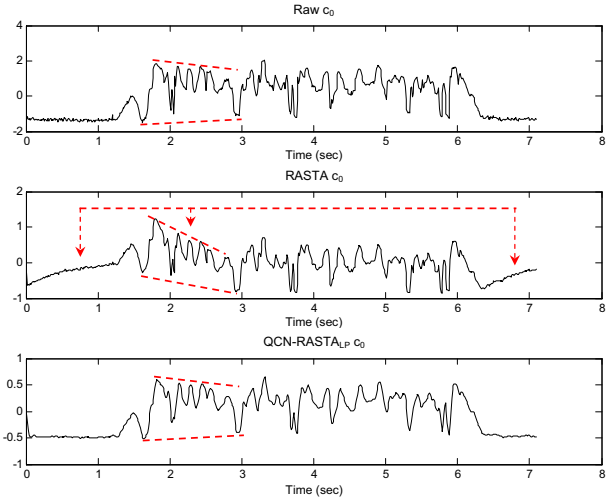


Figure 1: Transient effects due to RASTA filtering – raw, RASTA-filtered, and RASTA$_{LP}$-filtered c_0 tracks. Note increased transients due to RASTA filtering (middle plot) compared to RASTA$_{LP}$.

2.2. Feature Extraction Strategies

MFCC [11] and PLP [12] feature extraction front-ends predominate in current ASR speech systems. While there were many alternative feature extraction strategies proposed since the introduction of MFCC and PLP and shown to perform comparably or better in selected tasks, it seems to be quite difficult to overly improve on the 'baseline' features. One of the potential competitors may be found in so called bottleneck (BN) features [14]. BN feature extraction makes use of a neural network (NN). While NN-based speech features have been used for over a decade [18], they typically do not reach the performance/robustness of cepstral features. However, they were shown to provide complementary information to cepstral features and boost the performance when combined together [19]. A typical NN-based extractor uses NN or a set of NN's to estimate class (phone or sub-phone) posteriors from longer-term signal characteristics. Inputs to the NN (raw features) are usually formed by several consecutive frames of either critical band energies (CRBE) or cepstral features. The context can span from ± 4 up to ± 50 frames. Cepstral mean–variance normalization or RASTA filtering can be applied to the CRBE/cepstral tracks [20,21]. The class posterior probabilities are usually transformed by log non-linearity and decorrelated by PCA to better fit expectations of the following GMM–HMM acoustic model. In the case of BN, outputs of a hidden layer in a multi-layer NN are used rather than the class posteriors. This layer is typically smaller than other layers and is called a bottleneck.

Our BN framework is based on [14] (see Fig. 2). 22 critical band energies (CRBE) or 13 static cepstral coefficients form the raw features. Subsequently, selected normalizations and temporal filtering from the set described in Sec. 2.1 are applied to each feature dimension (track). Then, the context of $\pm$ 15 frames representing the temporal evolution in each feature track is weighted by a Hamming window and projected on 16 discrete cosine transform (DCT) bases. A 5-layer NN is trained by standard back-propagation algorithm to classify 136 sub-phone classes corresponding to the states of the HMM acoustic model. The training data are force-aligned by an MFCC baseline system. The NN has 250 000 weights and 30 neurons in the bottleneck layer. The two adjacent hidden layers have equal sizes. For the purpose of feature extraction, only the first half of the NN is used. As shown in [22], decorrelation of BN outputs further improves system performance. For our experiments, the Maximum Likelihood Linear Transform (MLLT) [23] with sub-phoneme state classes (same as NN targets) of the same covariance is utilized. In the following text, the term 'BN features' denotes BN outputs transformed by MLLT. In addition, we propose applying normalizations from Sec. 2.1 also on the top of the BN features, which may provide further robustness to feature variations due to the presence of noise or speech variations not seen by the BN neural network during the training. It is noted that in traditional BN systems, the (transformed) BN outputs are taken 'as they are' and fed directly to the acoustic back-end. Finally, in correspondence with finding in [22], delta parameters are computed.

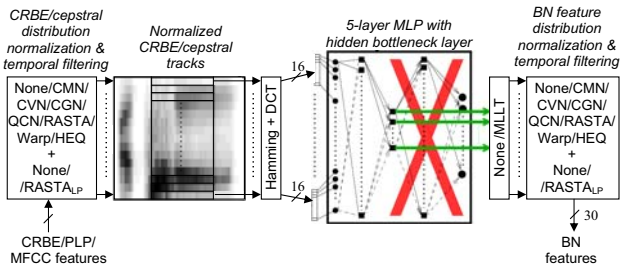


Figure 2: Proposed extended bottleneck front-end framework.

In addition, perceptually motivated minimum variance distortionless response (PMVDR) cepstral coefficients [13] are evaluated in our experiments. PMVDR was shown to provide a competitive performance to MFCC in small to middle vocabulary ASR tasks in car noise. PMVDR utilizes the minimum variance distortionless response estimator to represent the spectral envelope of the speech signal.

3. Experimental Results

3.1. Corpus Description

The test samples utilized in this study come from the Lombard portion of the UT-Scope corpus [24], where all subjects produced speech in a *neutral condition* (no noise exposure) and also in *simulated noisy conditions* (background noise samples produced through open-air headphones). Three noisy scenarios were introduced: (i) highway car noise, (ii) crowd noise, and (iii) pink noise. Highway and crowd noises were produced through headphones at 70, 80, and 90 dB SPL (sound pressure level); pink noise at 65, 75, and 85 dB SPL. Speech was recorded using throat, close-talk, and far-field microphones. In this study, the close-talk channel that provides high SNR for both *neutral* and *simulated noisy* condition recordings is used. For the ASR tasks, sessions from 31 native speakers of US English (25 females, 6 males) are used. Each session comprises 100 phonetically balanced read sentences from the TIMIT database [25] produced by each subject in the neutral condition, and 20 TIMIT sentences produced in each of the nine noise type/level conditions.

3.2. Experimental Setup

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) [25]. For cepstral front-ends (MFCC, PLP, PMVDR), 13 static cepstral coefficients, including c_0, and Δ's and $\Delta\Delta$'s form the feature vector. In the case of BN framework, 30 static and 30 Δ coefficients are used. The back-end acoustic model/LM setup is fixed for all front-ends. 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 *clean neutral speech* UT-Scope recordings. The adaptations are supervised and utilize labels obtained through forced alignment [26]. Speakers from the adaptation set are excluded from the open test set, which contains sessions from 3 male and 19 female subjects. ASR systems utilizing different front-ends are likely to have different optimal operating points (OP). To assure a fair comparison of various front-end systems, an 'optimal' OP is searched for each setup by selecting the number of HMM retraining iterations and word insertion penalty that minimize word error rate (WER) on small clean neutral development set from UT-Scope (excluded from subsequent evaluations).

The front-ends are evaluated on two tasks: *(i) clean recordings* – high SNR signals – clean neutral speech and clean Lombard speech produced in 70, 80, and 90 dB SPL of simulated highway and crowd noise, and 65, 75, and 85 dB of pink noise (noise was produced through headphones and does not appear in the LE recordings); *(ii) noisy recordings* – clean neutral speech and clean LE speech produced in 90 dB SPL of simulated highway noise, both mixed with the NOISEX'92 'Volvo' noise at 0, 5, ..., 20 dB SNR; clean neutral speech and clean LE speech produced in 90 dB SPL of simulated crowd noise, both mixed with the NOISEX'92 'Babble' noise at 0, 5, ..., 20 dB SNR. This yields a total of 30 evaluation sets. The initial ASR system utilizing MFCC–CVN front-end establishes performance on the clean neutral set at 8.3 % word error rate (WER) and PLP–CVN system 8.9 % WER. Since our focus is on optimizing the performance of acoustic front-ends rather than LM, the reminder of the paper reports WER's obtained from the acoustic model decoding with LM bypassed.

3.3. Results and Discussion

In the case of cepstral front-ends, normalizations are applied as shown in the left-most box in Fig. 2 (RASTA and RASTA$_{LP}$ are considered exclusive). For BN features, normalizations are performed both on features entering the BN neural network and on BN outputs. The experimental results are summarized in Tables 1 and 2 and Fig. 3. The tables show average WER's for open clean test sets, noisy sets (averaged across all SNR's and both types of noise), and an overall average WER across clean and noisy sets (denoted *Avg.*). Table 1 reports the performance of the baseline setups together with the best performing configurations found and Table 2 details observations that motivated the design of the winning system as well as some other interesting phenomena.

In the first step, performance of the baseline front-end setups where either CMN, CVN, or no normalization was applied to the raw features is studied (see the *Baseline* section in Table 1). CRBE–BN represents BN features without any normalization applied to CRBE. CRBE–BN$_{NoMLLT}$ is CRBE–BN where MLLT is also excluded. Major observations: (i) raw PLP outperforms raw MFCC in most conditions (Table 1, 1^{st} 2 lines with numbers), but mostly loses to MFCC once normalizations are applied; (ii) BN front-end is superior on clean neutral/LE sets but more sensitive to noise than MFCC.

Second, the broader set of normalizations introduced in Sec. 2.1 were applied. Major observations are summarized in individual notes in Table 2. **Note 1**: combination of CMN and RASTA$_{LP}$, denoted CMN$_{LP}$, outperforms RASTA. This is likely to be attributed to the significant reduction of the transient effects produced by the high order band-pass RASTA filter (see Fig. 1). **Note 2**: deriving HEQ reference distributions from a train set normalized by the combination of CGN–RASTA$_{LP}$ (CGN$_{LP}$) provided slight performance improvement compared to using raw distributions. It is noted that all normalizations here are applied on per-utterance basis while the HEQ distributions are derived from a large training set. **Note 3**: applying reference distributions derived from more discriminative front-end (MFCC) in HEQ on less discriminative front-end (denoted HEQ(MFCC)) can provide significant performance gains compared to using standard target front-end reference distributions. Note that the default PLP–HEQ performs poorer than PLP–CVN, but also represents the winning PLP setup when later combined with RASTA$_{LP}$. **Note 4**: applying CVN to CRBE in BN front-end as seen in past literature on bottleneck features may not be the best choice available. In our experiments, the

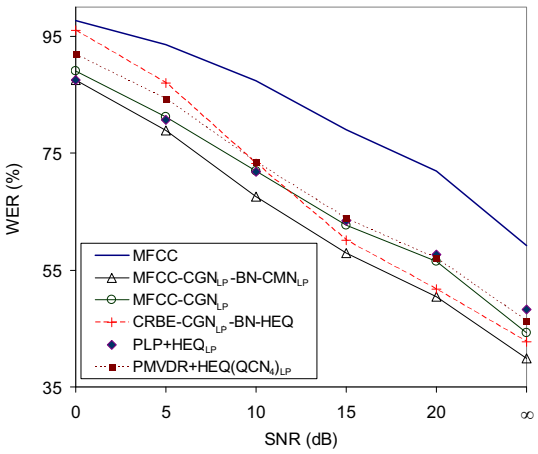


Figure 3: Performance of baseline MFCC system, best systems per each strategy (MFCC, PLP, PMVDR, CRBE BN), and proposed MFCC–BN system on the mixture of neutral and LE speech presented at various levels of highway and crowd noise; WER (%) averaged across noise types and neutral/LE speech sets.

CRBE–CVN–BN setup performed best when combined with CGN at the output, yet it is outperformed by CRBE–CGN$_{LP}$–BN–CGN. **Note 5**: applying normalizations to either CRBE or BN outputs benefits the performance, and combining both has a cumulative effect. **Note 6**: while PLP is prevalently used in cepstral-based BN front-ends [22], the superior performance of normalized MFCC cepstra in our baseline experiments suggests that normalized MFCC may be a stronger candidate also for cepstral-based BN features – which is confirmed here. **Note 7**: concatenating best performing CRBE–BN and MFCC features into a single feature supervector brought significant performance gain on the clean neutral and LE sets, but was not as successful on noisy sets.

The best performing front-end/normalization setups found per each feature extraction strategy are summarized in lower part of Table 1 (the index in X_{LP} denotes a use of RASTA$_{LP}$ jointly with normalization X; HEQ(QCN$_4$) denotes HEQ utilizing reference distributions from train data pre-normalized by QCN$_4$ which utilizes $4th$ and $96th$ percentiles [6]). CRBE-CGN$_{LP}$–BN–HEQ provides best performance on 3 out of 4 tasks – clean neutral/LE speech and noisy LE speech, MFCC–CGN$_{LP}$ is the second best setup. PLP and PMVDR provide similar average WER over the 4 tasks, PMVDR being more successful on clean neutral/LE speech and PLP on noisy sets. Finally, the newly proposed BN feature extraction scheme utilizing normalized MFCC as inputs to the BN neural network, MFCC-CGN$_{LP}$–BN, is presented at the bottom of Table 1. It can be seen that even without applying any normalization to the MLLT transformed BN outputs, the system considerably outperforms all the other feature extraction schemes considered. The overall best performance is reached in combination with CMN$_{LP}$ and the best noise robustness is provided with Warp$_{LP}$. It is noted that all the best systems found across extraction strategies as well as the proposed MFCC–BN system benefit from utilizing RASTA$_{LP}$ at their outputs and in the latter case also at the input to the BN neural network.

Fig. 3 compares the performance of a baseline MFCC system, best systems per strategy, and the proposed MFCC–BN system in various levels of background noise. For each SNR, WER's from noisy neutral and noisy LE test sets are averaged. It can be seen that the best CRBE–BN front-end outperforms all cepstral front-ends on the mixture of neutral/LE speech from 15 dB SNR up while the MFCC and PLP front-ends do better in lower SNR's. The newly proposed MFCC–BN front-end provides superior performance to all other front-ends in all SNR's (being closely approached by PLP at 0 dB SNR – a WER difference of 0.1 %).

4. Conclusions

This study analyzed the impact of varying types and levels of background noise and Lombard effect on LVCSR. Robustness of MFCC, PLP, PMVDR, and bottleneck (BN) feature extraction strategies combined with state-of-the-art feature normalizations and temporal filter-

Strategy	Norm.	Clean		Noisy (0–20dB)		Avg.
		Neutral	LE	Neutral	LE	
Baseline						
MFCC	None	43.7	66.1	76.7	95.1	70.4
PLP		43.8	65.3	**74.0**	94.0	69.3
PMVDR		N/A	N/A	N/A	N/A	N/A
CRBE– BN_{No_MLLT}		38.5	59.0	79.0	91.2	66.9
CRBE–BN		**37.4**	**56.6**	76.7	**90.7**	**65.3**
MFCC	CMN	33.5	55.2	69.2	83.9	**60.4**
PLP		34.2	57.9	**69.0**	85.4	61.6
PMVDR		36.1	57.8	70.1	89.4	63.3
CRBE– BN_{No_MLLT}		32.6	60.7	72.4	**82.6**	62.1
CRBE–BN		**32.2**	**54.1**	71.5	84.3	60.5
MFCC	CVN	33.3	54.2	**66.2**	84.1	**59.4**
PLP		36.6	57.4	71.3	87.1	63.1
PMVDR		36.6	61.8	75.7	90.8	66.2
CRBE– BN_{No_MLLT}		32.4	**52.2**	71.3	**84.0**	60.0
CRBE–BN		**31.3**	53.7	71.8	87.1	61.0
Best Systems Per Strategy						
MFCC	CGN_{LP}	32.2	51.7	**62.0**	82.5	57.1
PLP	HEQ_{LP}	35.3	56.7	62.3	82.2	59.1
PMVDR	$HEQ(QCN_4)_{LP}$	33.7	54.7	65.6	82.6	59.1
CRBE–CGN_{LP}–BN	HEQ	**30.2**	**50.6**	65.7	**81.4**	**57.0**
Proposed MFCC–BN System						
MFCC–CGN_{LP}–BN	None	27.4	48.6	58.7	80.2	53.7
	$Warp_{LP}$	28.1	49.3	**55.7**	**79.7**	53.2
	CGN	27.5	**47.6**	57.0	80.3	53.1
	CMN_{LP}	**26.9**	47.7	56.8	80.1	**52.9**

Table 1: Performance of baseline and top ranking front-end setups across various evaluation scenarios; WER (%). Available PMVDR implementation contains 'hardwired' CMN, hence no-norm. configuration was not available for evaluation (however, CMN can be overridden by other subsequent normalizations).

Strategy	Norm.	Clean		Noisy (0–20dB)		Avg.
		Neutral	LE	Neutral	LE	
Note 1: *RASTA vs. CMN-RASTA$_{LP}$ (Denoted CMN$_{LP}$)*						
MFCC	RASTA	38.6	60.9	68.8	84.1	63.1
	CMN_{LP}	**33.4**	**54.9**	**68.3**	84.3	**60.2**
PLP	RASTA	40.2	66.3	71.0	87.4	66.2
	CMN_{LP}	34.6	56.0	**68.3**	**83.6**	60.6
Note 2: *HEQ Refer. Distribution Derived from Raw vs. Normalized Data*						
MFCC	HEQ	**35.0**	55.6	63.3	**81.1**	58.7
	$HEQ(CGN_{LP})$	35.2	**54.5**	**61.7**	81.7	**58.3**
Note 3: *HEQ Ref. Distributions Derived from Different Front-End*						
PLP	HEQ	41.2	67.2	70.7	91.1	67.5
	HEQ(MFCC)	**37.1**	**59.8**	**64.8**	**84.3**	**61.5**
Note 4: *Standard (CVN) vs. Proposed CRBE Normalization*						
CRBE–CVN–BN	CGN	28.9	53.0	69.4	87.0	59.6
CRBE–CGN_{LP}–BN		**28.7**	**52.4**	**65.4**	**84.5**	**57.8**
Note 5: *Effect of BN Input & Output Normalization*						
CRBE–BN	None	37.4	56.6	76.7	90.7	65.3
CRBE–BN	HEQ	32.0	55.8	72.5	88.6	62.2
CRBE–CGN_{LP}–BN	None	30.6	52.1	69.7	86.3	59.7
CRBE–CGN_{LP}–BN	HEQ	**30.2**	**50.6**	**65.7**	**81.4**	**57.0**
Note 6: *PLP vs. MFCC BN Features*						
PLP–CVN–BN	CMN_{LP}	29.2	52.5	62.6	85.1	57.3
MFCC–CVN–BN		**27.7**	**50.7**	**59.4**	**81.8**	**54.9**
Note 7: *Fusion of Best CRBE–BN and Best Cepstral Features*						
♠CRBE–CGN_{LP}–BN	HEQ	30.2	50.6	65.7	**81.4**	57.0
♠♠MFCC–CGN_{LP}	CGN_{LP}	32.2	51.7	**62.0**	82.5	57.1
♠ & ♠♠ (Fusion)		**26.9**	**48.0**	64.1	82.2	**55.3**

Table 2: Observed phenomena; WER (%).

ing was analyzed. It is noted that to our knowledge, this represents the first systematic evaluation of BN features under varying noisy conditions and Lombard effect. The newly proposed extension of a traditional bottleneck scheme for a larger set of normalizations at the input to the neural network as well as incorporating additional normalization of the MLLT-transformed bottleneck outputs provided significant performance gains. Inspired by the observed superior performance of normalized MFCC compared to other cepstral front-ends, a BN scheme that utilizes normalized MFCC features rather than normalized CRBE or PLP features as inputs was proposed and shown to outperform all other front-ends in all test conditions. In addition, recently proposed $RASTA_{LP}$ was shown to outperform RASTA filtering and provide cumulative performance gains when combined with distribution normalizations. $RASTA_{LP}$ was also found to be required in all best performing setups. Our experiments confirmed observations from older literature that combining CRBE-based BN features and cepstral features can benefit ASR on clean data (which in our case extends also to clean LE data), however, was not proven to be as successful on noisy data. In addition, it was observed that pre-normalizing training features before the extraction of HEQ reference distributions or even adopting reference distributions from more discriminative front-ends may improve HEQ efficiency.

5. 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. dissertation, CTU in Prague, Czech Rep., 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] M. Garnier, "Communication in noisy environments: From adaptation to vocal straining," Ph.D. dissertation, Univ. of Paris VI, France, 2007.

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

[7] 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 Trans. on SAP*, vol. 8, no. 4, pp. 429–442, 2000.

[8] H. Bořil, P. Fousek, and P. Pollák, "Data-driven design of front-end filter bank for Lombard speech recognition," in *Proc. ICSLP'06*, Pittsburgh, Pennsylvania, 2006, pp. 381 – 384.

[9] H. Bořil and J. H. L. Hansen, "UT-Scope: Towards LVCSR under Lombard effect induced by varying types and levels of noisy background," in *Proc. IEEE ICASSP'11*, Prague, Czech, 2011, pp. 4472–4475.

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

[11] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. ASSP*, vol. 28, no. 4, pp. 357–366, 1980.

[12] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *JASA*, vol. 87, no. 4, pp. 1738–1752, 1990.

[13] U. H. Yapanel and J. H. L. Hansen, "A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition," *Speech Commun.*, vol. 50, no. 2, pp. 142–152, 2008.

[14] F. Grézl *et al.*, "Probabilistic and bottle-neck features for LVCSR of meetings," in *Proc. IEEE ICASSP'07*, Apr 2007, pp. 757–760.

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

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

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

[18] H. Hermansky *et al.*, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. IEEE ICASSP'00*, 2000.

[19] B. Chen, Q. Zhu, and N. Morgan, "Learning long-term temporal features in LVCSR using neural networks," in *Proc. ICSLP'04*, 2004.

[20] C. Benitz *et al.*, "Robust ASR front-end using spectral-based and discriminant features: experiments on the AURORA task," in *Proc. Eurospeech'01*, Aalborg, Denmark, Sept. 2001.

[21] F. Grézl and J. Černocký, "Trap-based techniques for recognition of noisy speech," *Lecture Notes in Comp. Science*, vol. 2007, no. 9, pp. 270–277.

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

[23] R. A. Gopinath, "Maximum likelihood modeling with gaussian distributions for classification," in *Proc. IEEE ICASSP'98*, 1998, pp. 661–664.

[24] 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*, vol. 17, no. 2, pp. 366 –378, Feb. 2009.

[25] 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.

[26] 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.