

Reduced Complexity Equalization of Lombard Effect for Speech Recognition in Noisy Adverse Environments

Hynek Bořil, John H.L. Hansen*

Center for Robust Speech Systems (CRSS)

Erik Jonsson School of Engineering & Computer Science, University of Texas at Dallas, U.S.A

hynek@utdallas.edu, john.hansen@utdallas.edu

Abstract

In real-world adverse environments, speech signal corruption by background noise, microphone channel variations, and speech production adjustments introduced by speakers in an effort to communicate efficiently over noise (Lombard effect) severely impact automatic speech recognition (ASR) performance. Recently, a set of unsupervised techniques reducing ASR sensitivity to these sources of distortion have been presented, with the main focus on equalization of Lombard effect (LE). The algorithms performing maximum-likelihood spectral transformation, cepstral dynamics normalization, and decoding with a codebook of noisy speech models have been shown to outperform conventional methods, however, at a cost of considerable increase in computational complexity due to required numerous decoding passes through the ASR models. In this study, a scheme utilizing a set of speech-in-noise Gaussian mixture models and a neutral/LE classifier is shown to substantially decrease the computational load (from 14 to 2–4 ASR decoding passes) while preserving overall system performance. In addition, an extended codebook capturing multiple environmental noises is introduced and shown to improve ASR in changing environments (8.2–49.2% absolute WER improvement). The evaluation is performed on the Czech Lombard Speech Database (CLSD’05). The task is to recognize neutral/LE connected digit strings presented in different levels of background car noise and Aurora 2 noises.

Index Terms: Lombard effect, speech recognition, codebook decoding, frequency transformation, cepstral normalization

1. Introduction

In adverse environments, speech corruption by noise and Lombard effect (LE) represent dominant sources impacting performance of ASR systems. Even after noise in the speech signal is suppressed, LE causes severe ASR degradation [1]. LE impacts the shape and spectral slope of glottal waveforms, as well as locations and bandwidths of formants, and hence, has a direct impact on speech coding used in ASR [2, 3, 4], causing a mismatch between the parameters of LE speech and ASR acoustic models trained on noise-clean neutral (modal) speech. Past studies focusing on the suppression of LE in ASR have searched for speech coding less sensitive to LE, transformations of LE speech towards neutral, enhanced training methods, and acoustic model adjustments and adaptation (see [3, 4] for overviews). Many of the proposed LE-suppression methods perform fixed signal transformations estimated from a limited amount of training LE speech samples, assuming that the level and quality of LE will not change over time, which may be not true in real environments with varying background noise [5].

Recently, novel unsupervised frequency and cepstral domain equalizations that reduce the impact of LE on ASR have been proposed and shown to outperform common compensations in ASR on neutral/LE noisy speech in various levels of car noise [6]. The drawback of the algorithms is the required extensive

number of decoding passes conducted per utterance in order to estimate maximum likelihood parameters of the frequency transformation, and to select the noisy acoustic models that best match the actual noisy conditions.

This study presents an alternative approach to the selection of optimal spectral transformation and noisy acoustic models, which significantly reduces the number of necessary ASR decoding passes. In addition, an extended system employing an environment detector and a codebook of acoustic models capturing multi-environmental noises is presented and shown to improve ASR in changing noise environments.

The paper is organized as follows. First, a system utilizing the compensations from [6] is introduced (baseline). Second, Gaussian mixture model (GMM) speech-in-noise model selection, neutral/LE speech classification, and noise type detection is presented. Third, the new system is evaluated and compared to the baseline.

2. Baseline system

LE introduces variations of the formant structure, where especially low formants tend to shift to higher frequencies [2, 3]. To compensate for the formant migration, [6] introduces a frequency transformation *Shift*: $F_{Shift} = F + \beta$. Taking advantage of the fact that translation of a feature extraction filterbank in the frequency has the same effect as shifting the amplitude spectrum (in the opposite direction), *Shift* is applied directly to the cut-offs of the front-end filterbank. The translation parameter β (Hz) is searched to maximize the likelihood of the transformed observations \mathbf{O}_{utt}^β given the ASR acoustic model λ_n . The search procedure is identical with the *fully optimized search* proposed in [7] in the context of vocal tract length normalization (VTLN). The search procedure is described in Table 1, where Ψ denotes the search grid ($\Psi = \{0, 50, \dots, 300\}$ (Hz)), selected based on the observed maximum range of formant variations in LE, \mathbf{W} is the sequence of words for language \mathcal{L} , Θ_l denotes the language model, and $\hat{\mathbf{W}}$ is the resulting transcription estimate. Given the search grid, *Shift* performs 7 decoding passes to determine the optimal β .

Table 1: Procedure of *Shift* frequency transform.

-
- 1) For each $\beta \in \Psi$, transform test utterances $\mathbf{O}_{utt} \rightarrow \mathbf{O}_{utt}^\beta$;
Decode transformed set using acoustic model λ_n :
$$\hat{\mathbf{W}}_{utt}^\beta = \arg \max_{\mathbf{W} \in \mathcal{L}} \left[Pr(\mathbf{O}_{utt}^\beta | \mathbf{W}, \lambda_n) Pr(\mathbf{W} | \Theta_l) \right];$$
 - 2) Find β_{max} maximizing decoding likelihood:
$$\beta_{max} = \arg \max_{\beta} \left[Pr(\mathbf{O}_{utt}^\beta | \hat{\mathbf{W}}_{utt}^\beta, \lambda_n) Pr(\hat{\mathbf{W}}_{utt}^\beta | \Theta_l) \right]$$

$$\rightarrow \hat{\mathbf{W}}_{utt} = \hat{\mathbf{W}}_{utt}^{\beta_{max}}.$$
-

Changes in the spectral slope of the glottal waveform and formant locations and bandwidths in LE [2, 3, 4], as well as

*This project was funded by AFRL under a subcontract to RADAC Inc. under FA8750-05-C-0029.

the presence of additive noise [8] and channel transfer function variability affect distributions of cepstral coefficients. In [6], quantile-based cepstral dynamics normalization, *QCN*, is introduced as an alternative to common cepstral mean normalization (CMN), cepstral variance normalization (CVN) [9], and recent cepstral gain normalization (CGN) [10]. *QCN* exploits the fact that cepstral distributions, especially of low order cepstral coefficients, deviate from Gaussian, and hence, the dynamic range of the cepstral samples can be better estimated from the sample histogram quantiles rather than from the mean and variance. *QCN* is defined:

$$c_{n,i}^{QCN4} = \frac{c_{n,i} - (q_4^{C_n} + q_{96}^{C_n})/2}{q_{96}^{C_n} - q_4^{C_n}}, \quad (1)$$

where i is the index of the cepstral sample in the time window of the normalization, and $q_4^{C_n}, q_{96}^{C_n}$ (n = cepstral dimension) are 4% and 96% quantiles which were found to provide superior performance in ASR task on a development set.

A simple speech decoding scheme employing a codebook of noisy acoustic models is used in order to increase match of the input noisy speech and ASR acoustic models in changing level of background noise. The codebook consists of HMM's trained on data with different SNR's, where clean speech is mixed with car noises at SNR's of -5, 0, ..., 20, ∞ dB, yielding 'noisy' models denoted $\lambda_1, \dots, \lambda_7$. During recognition, the observation sequence \mathbf{O} is decoded as:

$$\hat{\mathbf{W}}_n = \arg \max_{\mathbf{W} \in \mathcal{L}} \left[Pr(\mathbf{O} | \mathbf{W}, \lambda_n) Pr(\mathbf{W} | \Theta_t) \right]. \quad (2)$$

Applying Eq. (2) consecutively for all noisy models $\lambda_1, \dots, \lambda_7$ yields a set of transcription estimates $\hat{\mathbf{W}}_1, \dots, \hat{\mathbf{W}}_7$. Subsequently, the λ_n that provides the highest likelihood decoding path (best match) is found:

$$BM = \arg \max_n Pr(\mathbf{O} | \hat{\mathbf{W}}_n, \lambda_n) \quad (3)$$

with the corresponding output transcription $\hat{\mathbf{W}}_{BM}$. The HMM with the closest noise structure is expected to give the highest score. The *Shift* transform is conducted subsequently, utilizing the best matching noisy model set. In total, utterance decoding requires 14 decoding passes – 7 to find the best matching λ_n , 7 to search for the best β .

3. Optimized system

As discussed in Sec. 2, a major portion of the computational load in the baseline system is due to the extensive search for best matching acoustic models and optimal frequency transformation of the incoming speech samples. In the following subsections, Gaussian mixture model (GMM) classifier-based efficient noisy model selection and neutral/LE utterance classification that directs the frequency transform are introduced.

3.1. Noisy model selection (SNR estimation)

The results presented in [6] suggest that assigning the noisy utterance to the matching noisy model set based on maximizing the likelihood of the decoding path across codebook models is successful, however, the computational cost of utterance decoding is high and grows linearly with codebook size, preventing possible extension to multiple or mixed noise types. To address this, a GMM-based noisy model selection is proposed. For each SNR captured in the noisy codebook, a unique GMM is trained (on the same data used for training the noisy ASR models), yielding a set of models $GMM_{-5dB}, GMM_{0dB}, \dots, GMM_{20dB}, GMM_{\infty dB}$. Instead of decoding the utterance consecutively by all codebook HMM's (see Eq. (2)), a computationally efficient scoring of the incoming utterance against the set of GMM's is conducted. The highest scoring GMM determines which λ_n should be used

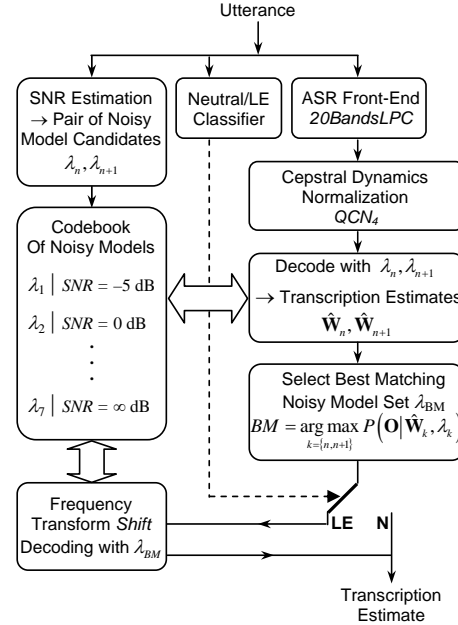


Figure 1: Optimized system – GMM-based noisy model selection and neutral/LE classifier driven *Shift* transform.

for utterance decoding. Note that the task here corresponds to SNR estimation. However, while a variety of available definitions of SNR are to a certain level sensitive to the proportion of speech/non-speech segments in the speech signal and characteristics of background noise, the GMM-based noise level classification proposed here *directly* captures the speech-in-noise characteristics as modeled by the ASR acoustic models. Hence, this method is promising to provide more reliable speech/noise level assessment with respect to the acoustic model assignment.

In the implementation of the noise level classification, multi-hypothesis testing (where the highest scoring GMM representing the actual SNR would be searched) is transformed into a series of pairwise GMM score comparisons incorporating a set of individual decision thresholds. This allows for fine tuning of the inter-class classification performance as known from two-hypothesis tasks. In particular, the score from GMM_{-5dB} is compared to GMM_{0dB} and the higher scoring model is selected as the winner. Subsequently, GMM_{0dB} is compared to GMM_{5dB} , etc., up to the pair $GMM_{20dB}, GMM_{\infty dB}$. The model that wins in most cases (i.e., twice), becomes the overall winner. Special cases: if GMM_{-5dB} wins zero times and all other models win once, $GMM_{\infty dB}$ is selected as a winner; if $GMM_{\infty dB}$ wins zero times and all others once, GMM_{-5dB} is the winner. For each pairwise comparison of scores s_1 and s_2 , an individual decision threshold $s_1/s_2 \ll Th_{12}$ is searched on the development set to maximize overall classification accuracy. This selection scheme exploits the fact that GMM's representing adjacent SNR's are cohorts and the distance between models increases with the difference in their SNR indexes (e.g., if a speech sample captures $SNR = 20$ dB, when scored against the pair GMM_{-5dB}, GMM_{0dB} , the latter model is acoustically closer and is more likely to be selected as a winner). Similarly, from the pair GMM_{0dB}, GMM_{5dB} the latter is likely to win. However, in both pairs GMM_{15dB}, GMM_{20dB} and $GMM_{20dB}, GMM_{\infty dB}$, the model GMM_{20dB} is most likely to win.

3.2. Directing the *Shift* transform

The *Shift* transform compensates for formant shifts due to LE by shifting the short time spectral envelope down in frequency. Unlike VTLN, the shift is constant across the whole frequency band

and does not aim at compensating for the inter-speaker vocal tract differences. When analyzing neutral speech samples, the maximum likelihood-assigned β 's in *Shift* were almost exclusively $\beta = 0$ Hz, or reached the smallest allowed non-zero value (50 Hz). In order to eliminate multiple decoding passes performed by *Shift* when processing neutral speech, a GMM-based neutral/LE classifier is incorporated into the system. If the utterance is classified as neutral, the classifier assigns $\beta = 0$ Hz and no further decoding is conducted.

The complete system is shown in Fig. 1. In the rightmost part of the scheme, front-end feature extraction is first conducted, followed by the cepstral dynamics normalization (see Eq. (1)). Subsequently, noisy model selection incorporating the information from the SNR estimator (Sec. 3.1) is conducted, yielding the best matching acoustic model λ_{BM} . Based on the output of the neutral/LE classifier, λ_{BM} is used either for direct decoding of the utterance (neutral speech) or in the search for the optimal β in *Shift* (Table 1) and subsequent decoding of the optimally warped utterance.

3.3. Extended codebook system

Due to the computational savings provided by the GMM-based noisy model selection in Sec. 3.1, the codebook can be easily extended for multiple noise types, where each environmental noise will be represented by a sub-codebook of models covering various SNR's. To achieve reasonable accuracy for model assignment, the GMM-based model selection is split into two stages: (i) environmental noise identification, (ii) SNR estimation.

4. Experiments

The Czech Lombard Speech Database (CLSD'05) [4] that captures neutral speech and speech uttered in simulated noisy conditions (90 dB SPL of car noise produced to speakers through headphones) is used in all experiments. A close-talk microphone was used in recordings, yielding high SNR signals (mean SNR of 28 dB). The subjects were provided listener feedback to ensure proper reaction to the noisy background. The recordings were downsampled to 8 kHz and filtered by a G.712 telephone filter. In the 'single environment' experiment, clean recordings were mixed with 20 car noise samples [4] at SNR's of -5, 0, ..., 20, ∞ dB, where ∞ dB represents clean data with no noise added. In the 'multiple environments' setup, clean recordings were mixed correspondingly also with airport, babble, restaurant, street, and train samples from the Aurora 2 database [11]. The long noise samples were cut into a number of 2 sec. samples which were then randomized and mixed with clean speech samples.

The HMM-based recognizer used in experiments comprises 43 context-independent monophone models and two silence models (3 emitting states, 32 Gaussian mixtures). The feature vector consists of 13 static cepstral coefficients c_0-c_{12} and their first and second order time derivatives. Gender-dependent phoneme models are obtained from 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 a total of 4930/5360 words, respectively, from 12 speakers, while the male neutral/LE test sets contain 1423/6303 words from 14 speakers (per each SNR level).

Based on the superior performance in [6], a front-end denoted 20Bands-LPC0-3200, derived from PLP by replacing the trapezoid filterbank with a bank of 20 non-overlapping rectangular filters, is used in the codebook system. The filters are uniformly spaced on a linear scale over 0-3200 Hz.

4.1. Evaluation & discussion

Noisy model selection: In the preliminary experiment on the development set (small subset of open test set data - 2 female and 2 male sessions), PLP provided superior classification performance to MFCC and 20Bands-LPC. In particular, PLP cep-

stral coefficients c_0-c_{14} and their first and second time derivatives provided superior performance. The overall accuracy of the noisy model assignment reached approximately 60%. It was observed, that most of model assignments were falling either on the diagonal of the confusion matrix (correct assignment) or were right adjacent to it. For this reason, the original SNR estimate was extended for an adjacent, complementary candidate (denoted '+1 Neighbor'), increasing the probability that the SNR output contains a correct value. For GMM_{-5dB} the neighbor was GMM_{0dB} , for the rest the lower model (difference -5 dB SNR) was picked. Together with the pairwise decision threshold, tuning on the development set yielded an accuracy of 94.96% and 94.62% on development neutral and LE sets, and 92.13% and 91.37% on neutral and LE open test sets. The selection scheme yields a pair of noisy model candidates, the optimal being selected based on Eq. (2, 3).

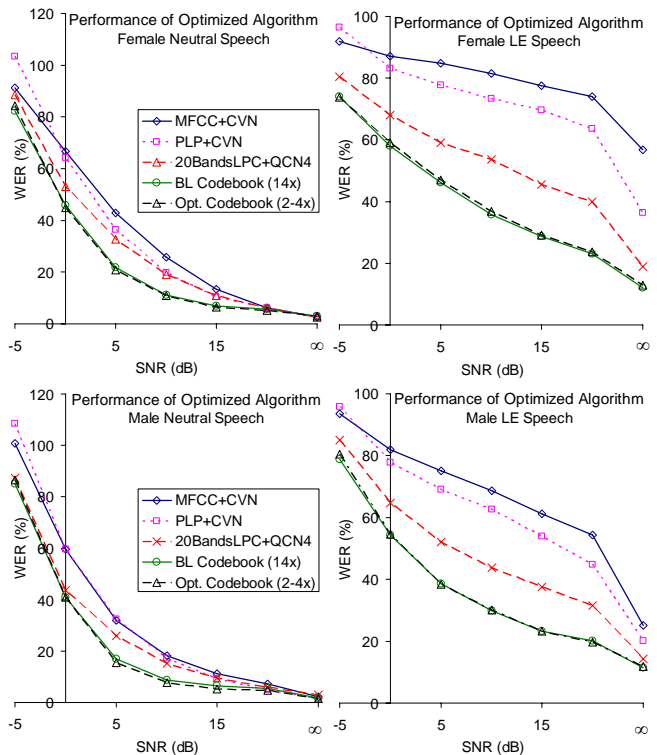


Figure 2: Performance of optimized system for neutral and LE speech versus noise SNR and gender.

Neutral/LE classification: MFCC, PLP, Expolog [1], and 20BandsLPC0-3200 were compared for the task of neutral/LE speech classification. All classifiers were trained and tested on the clean development set (closed test). The best performance was reached by MFCC (100% on clean, 88% on all SNR's set), followed by 20BandsLPC0-3200 (100% on clean, 85% on all SNR's set). To reduce computational demands during feature extraction, 20BandsLPC0-3200 (c_0-c_{12} and their first and second order derivatives) was selected for the neutral/LE classifier (as already used in the ASR front-end).

Optimized system (single environment): The optimized codebook system utilizing the '+1 Neighbor' model selection strategy requires 2 ASR decoding passes to select the matching noisy model set. The neutral/LE classifier reduces the number of additional *Shift* decoding passes for speech classified as neutral from 7 to 0. To further reduce the computations in the case of LE speech, the search grid for β is limited to $\beta \in \{0, 100, 200\}$, representing 2 additional decoding passes compared to neutral speech. The performance of the optimized system is shown in

Fig. 2, denoted ‘Opt. Codebook’, and compared to the baseline system ‘BL Codebook’ [6], and single-pass 20BandsLPC system with QC/N_4 , and MFCC and PLP HMM systems employing CVN (the last three trained on clean speech). It can be seen that both codebook systems considerably outperform the single-pass recognizers, and that the optimized codebook recognizer provides comparable performance to the ‘BL Codebook’.

Table 2: Environmental detector – performance, all SNR’s.

Cond.	Test Set	Assigned Environment			Acc (%)
		Car	Street	Restaurant	
N+LE	Car	20558	38	4856	80.77
	Street	2699	14112	8641	55.45
	Restaurant	81	142	25229	99.12

Optimized system (multiple environments): Here, three separate sub-codebook model sets are trained on the clean train set mixed with car, street, and restaurant noise samples, respectively. Each sub-codebook contains models from -5 dB SNR to 20 dB SNR. A set of clean-trained models is shared among the sub-codebooks. The GMM-based environmental noise classifier (setup similar to the SNR estimator) is trained on the training clean data mixed with the corresponding noises at -5 dB SNR. Performance on the open test set is shown in Table 2. It can be seen that utterances with car and restaurant noises are assigned with acceptable accuracy, while street samples are frequently confused either with car or restaurant noises. This can be explained by the high non-stationarity of the street noise samples, which contain both babble noise from pedestrians (observed also in restaurant) and traffic noises (also in car environment). It is noted that noise classification accuracy reduces with increasing SNR of the signals, as the noisy background is less pronounced. Table 3 shows performance of the neutral/LE classifier on open test sets comprising all SNR’s. The accuracy ranges consistently around 80 % for all environments.

For each environmental noise in the codebook (car, street, restaurant), a separate GMM-based SNR estimator was trained (similarly to Sec. 3.1). Considering the imperfect performance of the environmental detector, and moreover, the possibility of the ASR system being exposed to out-of-codebook noises, it is required that the SNR estimator provide good accuracy also for non-matching noisy environments. When tested on development non-matching noisy data, the street-trained SNR estimator provided superior performance compared to the car- and restaurant-trained estimators. This is due to the fact that the street environment comprises broad range of noise components that can be observed also in other environments. Hence, a single, street-trained SNR estimator is incorporated in the final system (providing average accuracy of 84 % on the airport, babble, car, restaurant, street, and train open test sets).

Table 3: Neutral/LE classification – all environments.

Open Test N/LE Classification, All SNR’s, Acc (%)					
Airport	Babble	Car	Restaurant	Street	Train
83.19	80.8	81.14	83.52	78.08	79.05

The performance of the optimized extended codebook system is presented by means of overall word error rate (WER) reduction compared to the baseline single pass 20BandsLPC+ QC/N_4 HMM recognizer trained on clean neutral speech (see Table 4). The delta WER values for each environment are obtained by calculating average WER across genders and all SNR’s in the particular environment for both baseline and

codebook systems, yielding $WER_{Baseline}$ and $WER_{Codebook}$, and subtracting. Negative values signify WER reduction when using the codebook system, positive values have a WER increase. It can be seen that for five of the six environments, the codebook system provides considerable WER improvement. Note that the system improves performance also for environmental noises not present in the codebook (airport, babble, train), selecting the closest noise type/level match from the available codebook models. In the case of street, the slight WER increase may be explained by the ambiguity and relatively high non-stationarity of the environment, which makes the selection of the matching ASR models difficult.

Table 4: Extended codebook system – WER reduction.

Performance Improvement, $WER_{Codebook} - WER_{Baseline}$ (%)					
Airport	Babble	Car	Restaurant	Street	Train
-11.11	-8.23	-49.2	-14.48	1.61	-15.54

5. Conclusions

This study has presented a computationally efficient framework for compensating for the impact of Lombard effect and additive noise on ASR. The proposed system incorporates maximum likelihood spectral transformation directed by a neutral/LE classifier, cepstral dynamics normalization, and a codebook of noisy acoustic models utilizing GMM-based noisy model selection. Compared to a previously developed system, the number of ASR decoding passes is reduced from 14 to 2 (or 4) in the case of speech classified as neutral or LE, respectively, while overall WER performance is preserved. In addition, an extended scheme with a codebook capturing speech contaminated by multiple environmental noises at various SNR’s was presented and shown to measurably improve ASR performance (8.2–49.2 % WER improvement) for 5 out of 6 noisy environments, including environments not covered in the extended noise codebook.

6. 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 Trans. on SAP*, 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] 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.
- [6] H. Bořil and J. H. L. Hansen, “Unsupervised equalization of Lombard effect for speech recognition in noisy adverse environment,” in *Proc. of IEEE ICASSP’09*, Taipei, Taiwan, April 2009.
- [7] 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.
- [8] J. P. Openshaw and J. S. Mason, “Optimal noise-masking of cepstral features for robust speaker identification,” in *ASRIV-1994*.
- [9] 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.
- [10] S. Yoshizawa, N. Hayasaka, N. Wada, and Y. Miyanaga, “Cepstral gain normalization for noise robust speech recognition,” in *Proc. of IEEE ICASSP’04*, vol. 1, May 2004, pp. 1–209–12 vol.1.
- [11] H. G. Hirsch and D. Pearce, “The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions,” in *ISCA ITRW ASR2000*, Paris, France, 2000.