UNSUPERVISED EQUALIZATION OF LOMBARD EFFECT FOR SPEECH RECOGNITION IN NOISY ADVERSE ENVIRONMENT

Hynek Bořil, John H. L. Hansen*

Center for Robust Speech Systems, Erik Jonsson School of Engineering & Computer Science
The University of Texas at Dallas, USA

ABSTRACT

When exposed to environmental noise, speakers adjust their speech production to maintain intelligible communication. This phenomenon, called Lombard effect (LE), is known to considerably impact the performance of automatic speech recognition (ASR) systems. In this study, novel frequency and cepstral domain equalizations that reduce the impact of LE on ASR are proposed. Short-time spectra of LE speech are transformed towards neutral ASR models in a maximum likelihood fashion. Dynamics of cepstral coefficients are normalized to a constant range using quantile estimations. The algorithms are incorporated in a recognizer employing a codebook of noisy acoustic models. In a recognition task on connected Czech digits presented in various levels of background car noise, the resulting system provides an absolute reduction in word error rate (WER) on 10 dB SNR data of 8.7 % and 37.7 % for female neutral and LE speech, and of 8.7 % and 32.8 % for male neutral and LE speech when compared to the baseline system employing perceptual linear prediction (PLP) coefficients and cepstral mean and variance normalization.

Index Terms— Lombard effect, speech recognition, frequency warping, cepstral compensation, codebook of noisy models

1. INTRODUCTION

Lombard effect (LE) is known to affect a number of speech parameters, such as voice intensity, spectral slope of glottal waveforms, formant locations and bandwidths, energy ratios in voiced/unvoiced phones, and others [1]–[3]. Since current ASR features are mostly in the form of cepstral coefficients extracted from short-time spectra, especially formant and spectral slope variations will directly impact ASR performance, introducing a degrading mismatch between the processed speech and neutral trained acoustic models. Efforts to increase ASR resistance to 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], [3] for overviews. A majority of past studies assume that there is a sufficient amount of labeled LE data available for estimating fixed signal equalization/model adaptation parameters and that the level of LE (a ratio of speech production variations introduced by the environmental noise) will not change over time. This may be unrealistic in real world applications where the characteristics of noise and the level of LE may vary continuously.

This study presents novel frequency and cepstral domain transformations for equalizing LE speech samples towards neutral ASR models in order to improve recognition performance. In contrast

to methods previously developed, the transformation parameters are estimated on-the-fly from the incoming speech signal and require neither *a priori* knowledge about the level of LE, nor availability of labeled training/adaptation LE samples.

The frequency domain transformation is conducted in a similar manner to maximum likelihood (ML) vocal tract length normalization (VTLN) [4], [5], with the difference being the frequency mapping function, which is chosen to better address the formant shifts introduced by LE. Subsequently, the dynamics of cepstral coefficients are normalized using two quantiles for each cepstral dimension estimated from the sorted cepstral samples. Recently, advanced techniques normalizing the fine structure of cepstral histograms have been developed, utilizing either a rather extensive adaptation data set matching the test conditions [6], or quantile-based online normalization using two-pass search and continuity criteria [7]. In contrast to these complex methods, the goal of the proposed cepstral compensation is to address the dynamics of the cepstral coefficients rather than the fine histogram structure, extending the concepts of the popular and computationally inexpensive normalizations of cepstral mean (CMN) [8] and variance (CVN) [9], and recently introduced cepstral gain normalization (CGN) [10]. The proposed frequency and cepstral transformations are incorporated in a recognizer employing a codebook of acoustic models trained on clean data mixed with car noise at various SNR's (noisy models).

The paper is organized as follows. First, novel frequency and cepstral transformations are introduced and compared to common algorithms in a recognition task. Second, a small-footprint ASR engine employing a codebook of noisy models is presented.

2. PROPOSED METHODS

2.1. Maximum Likelihood Frequency Transformations

The location of vocal tract resonances (formants) in neutral speech is approximately inversely proportional to the vocal tract length (VTL). To compensate for the inter-speaker VTL differences, VTL normalization (VTLN) can be used. VTLN warps short-time spectra of speech by a factor α: $F_{VTLN} = \frac{F}{\alpha}$. In the maximum likelihood (ML) VTLN [4], α is searched to maximize the likelihood of the utterance's forced alignment, given the utterance transcription and the ASR hidden Markov model (HMM). When applying VTLN during recognition, the unknown utterance transcription is first estimated by decoding unwarped data, followed by alignments (*two-pass* VTLN decoding). In [5], a so called *fully optimized* VTLN decoding was proposed. Here, the recognized utterance is consecutively warped by the whole search grid of α's and decoded, yielding a set of transcription estimates. The warping that yields a decoding path through the HMM with the highest likelihood is determined and the corresponding transcription is taken.

Formant shifts in LE do not simply correspond to those due to

VTL variations. Analyses of LE corpora show consistent increases of the first formant frequency F_1, and a frequent increase of the second formant frequency F_2 in many phones, while higher formants remain steady or shift in either direction in frequency, depending on the phonetic content and conditions [1]–[3]. This corresponds well with the concept that F_1 varies inversely to the vertical position of the tongue and F_2 is related to tongue advancement [11], as speakers tend to lower their jaw as well as lower and advance their tongue when increasing vocal intensity in noise. To allow for more accurate formant normalization in LE speech, we propose a linear frequency transform in the form: $F_{W\&S} = \frac{F}{\alpha} + \beta$. As shown in Fig. 1, $F_{W\&S}$ is capable to better address different shifts in low/high formants than VTLN.
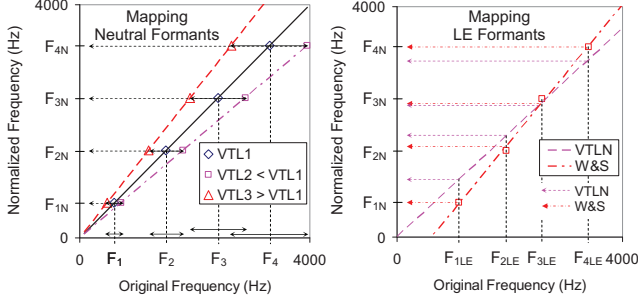

Fig. 1: VTLN/Warp&Shift frequency transforms – principle.

Implementation of *Warp&Shift*: The $F_{W\&S}$ transform is implemented in a similar way to [4], yielding a normalization called *Warp&Shift*. When applying *Warp&Shift* during HMM training, each training utterance is transformed consecutively by a set of $F_{W\&S}$ candidates and aligned by forced alignment, given the known utterance transcription. The transformation maximizing the likelihood of the forced alignment is selected to normalize the utterance, which is then employed in retraining the HMM. In the recognition stage, *Warp&Shift* is applied similarly to *fully optimized* VTLN. The search grid for $F_{W\&S}$ parameters is defined by two sets $s_1 = \{0, 50, ..., 200\}$ and $s_2 = \{3000, 3100, ..., 3400\}$. Only linear functions $F_{W\&S}$ passing through the points $A[s_1, 0]$, $B[s_2, 3200]$ are considered in the ML search for the optimal transformation. In a manner similar to [4], instead of transforming the frequency axis of the amplitude spectra, $F_{W\&S}$ is applied to the filter bank cutoff frequencies in the feature extraction front-end (e.g., $F_{W\&S}$ defined by points $A[0, 0]$, $B[3400, 3200]$ will expand the filter bank and compress the spectrum, corresponding to VTLN warping, while points $A[100, 0]$, $B[3300, 3200]$ will shift the filter bank cutoffs by +100 Hz, shifting the spectrum by 100 Hz downwards). The sets s_1, s_2 are chosen to allow for low formant translation by up to -200 Hz, since it has been observed that the increase in F_1, F_2 usually does not exceed this rate [3], and to allow for high formant translation in either direction. The search grid comprises 5 x 5 = 25 possible combinations.

In addition, we propose a frequency transform $F_{Shift} = F + \beta$, which only shifts the spectra in frequency. The search grid is chosen $\beta = \{0, 50, ..., 300\}$, yielding 7 choices. F_{Shift} is implemented in a similar way to *Warp&Shift*, but is applied only in the recognition stage, utilizing non-normalized HMM's. Motivation for *Shift* normalization is discussed in Sec. 3.

2.2. Quantile-Based Cepstral Dynamics Normalization

Cepstral mean and variance normalization [8], [9] are popular methods compensating for slow-rate convolutional and additive variations

occurring in the speech production/microphone channel chain. Since the variances in speech production due to LE can be viewed also as convolutional distortions of the speech signal, CMN/CVN may be able to suppress their impact on ASR. Cepstral mean normalization (CMN) estimates cepstral means $\bar{c}_n$ from a long time window and subtracts them from each cepstral sample $c_{n,i}$ in the window: $c_{n,i}^{CMN} = c_{n,i} - \bar{c}_n$, where n is the cepstral dimension and i is the index of the cepstral sample in the window. Cepstral variance normalization (CVN) estimates variance of each cepstral dimension, $\hat{\sigma}_{Cn}$, and normalizes it to unity: $c_{n,i}^{CVN} = c_{n,i}^{CMN}/\hat{\sigma}_{Cn}$. When the cepstral distributions drift from Gaussian, variance in CVN may not well represent the actual dynamic range of cepstra. This is addressed by the recently introduced cepstral gain normalization (CGN) [10], which estimates the dynamic range in each dimension from the maximum and minimum sample values, $c_{n\,\max}, c_{n\,\min}$ and normalizes it to unity: $c_{n,i}^{CGN} = c_{n,i}^{CMN}/(c_{n\,\max} - c_{n\,\min})$.

The accuracy of CMN, CVN, and CGN will reduce if the skewness of cepstral distributions from the set used for HMM training and those occurring in recognition data is different. Even if the distribution variances are normalized, subtracting the distribution mean will cause a mismatch in the cepstral dynamic range (e.g., see the two upper left distributions in Fig. 2, where μ denotes the distribution mean and q_5, q_{95} are 5% and 95% quantiles, bounding 90 % of the samples). When analyzing cepstral distributions in neutral ASR-training data and in LE test data used in our experiments (see Sec. 3), considerable differences in distribution skewness were observed (see non-normalized c_0 distributions in the right of Fig. 2).
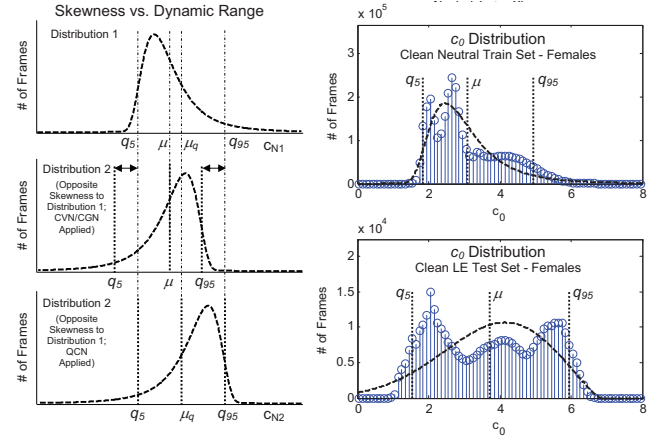

Fig. 2: Left – CVN, CGN, QCN principle. Right – c_0 distribution in neutral train set and LE test set; dashed line – generalized extreme value interpolation of histogram samples.

To address this, we propose quantile-based cepstral dynamics normalization (QCN). To reduce the sensitivity of the normalization to outliers (CGN utilizes two samples of extreme values), the cepstral dynamic range is determined from the low and high quantile estimates q_j^{Cn} and q_{100-j}^{Cn}, where j is in percent. The quantile estimates are found in each cepstral dimension by sorting cepstral samples from the lowest to the highest, and picking samples with indexes $round\,(j.L/100)$ and $round\,[(100 - j).L/100]$ where L is the number of samples. Instead of subtracting the distribution mean, an average of q_j^{Cn} and q_{100-j}^{Cn} (in Fig. 2 denoted μ_q) is subtracted, yielding more accurate dynamic range normalization for distributions of different skewness (compare the first and third distribution in the left part of Fig. 2). QCN is defined:

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

3. EXPERIMENTAL SETUP AND RESULTS

Database: Recognition experiments are conducted on the Czech Lombard Speech Database (CLSD'05) [3], comprising recordings of neutral speech and speech uttered in simulated noisy conditions (90 dB SPL of car noise produced to speakers through headphones). Speech was collected by a close-talk microphone, yielding high SNR signals (mean SNR of 28 dB). The recorded subjects communicated utterances over noise to a human operator to ensure proper reaction to the noisy background. The recordings were downsampled to 8 kHz and filtered by a G.712 telephone filter. For present experiments, clean recordings were mixed with 20 noise samples at SNR's of -5, 0,..., 20, ∞ dB, where ∞ dB represents clean data with no noise added. The noise samples were recorded in the cabin of a moving car [3].

Recognition setup: An HMM-based recognizer is used in experiments, employing 43 context-independent monophone models and two silence models (3 emitting states, 32 Gaussian mixtures). Cepstral coefficients c_0–c_{12} and their first and second order time derivatives form the feature vectors. Gender-dependent phoneme models are trained with 46 iterations on large vocabulary material from 37 female/30 male speaker sessions from the Czech SPEECON database [12]. The task is to recognize 10 Czech digits (16 pronunciation variants) in connected digits utterances. The female neutral/LE test sets comprise a total of 4930/5360 words, respectively, uttered by 12 speakers, while the male neutral/LE test sets comprise 1423/6303 words uttered by 14 speakers. Performance is evaluated by means of word error rate (WER).

Feature extraction front-ends: In a previous study [3], 20Bands–LPC cepstral coefficients displayed superior performance on clean LE speech and comparable performance on neutral speech to Mel frequency cepstral coefficients (MFCC) and perceptual linear prediction (PLP) cepstral coefficients. 20Bands–LPC is derived from PLP by replacing the trapezoid filter bank with a bank of 20 non-overlapping rectangular filters uniformly spaced on a linear scale over 0–4 kHz. In the initial recognition experiment on the present noisy test sets, 20Bands–LPC outperformed both MFCC and PLP on all clean/noisy LE sets as well as on noisy neutral sets. On clean neutral data, MFCC, PLP, and 20Bands–LPC establish comparable WER's (%) for females, 2.82, 2.92, 2.94 respectively; on clean neutral male set, the WER's are 2.32, 1.48, 2.60 respectively, with PLP performing the best (all front-ends employed CVN). Due to consistent superior performance on most sets, 20Bands–LPC is chosen for experiments.

Frequency normalizations: Two-pass VTLN [4] (*VTLN Lee-Rose*), VTLN utilizing optimized decoding [5] (*VTLN Optimized*), and *Warp&Shift* are applied during both HMM training and recognition. In recognition setups employing these frequency normalizations, non-normalized HMM's are trained in 36 iterations. Subsequently, a given normalization is applied, providing a normalized train set which is used for retraining HMM's in 3 iterations. The normalization is then repeated, followed by 7 retraining iterations, yielding fully trained HMM's. In VTLN, a warping factor α (search grid set to 0.8–1.2, step 0.05) is applied to the filter bank (FB) cutoff frequencies. To avoid exceeding Nyquist frequency of 4 kHz during FB warping, the initial FB is limited to span of 0–3200 Hz. An identical initial FB is used in all frequency transformation setups. *Shift* normalization is implemented similarly, using fully optimized decoding. *Shift* is applied only in the recognition stage, utilizing non-normalized HMM's obtained in 46 retraining iterations. Parameters of the frequency and cepstral transformations are estimated online for each individual utterance (average length ~4 sec).

Recognizer employing a codebook of noisy models: It is well known that in noise, ASR performance can be improved by adding actual noise characteristics to the acoustic models of a recognizer [13]. We use a simple recognition scheme employing a codebook of HMM's trained on data with different SNR's. In particular, separate models are trained on the clean training set mixed with car noises (5 samples disjunct from those mixed with the test sets) at SNR's of ∞, 20, 15, ..., -5 dB, yielding a codebook of 'noisy' models. In the recognition stage, each utterance is consequently decoded by each of the codebook HMM's, yielding a set of transcription estimates. The HMM reaching the highest likelihood of the decoding path is selected and the corresponding transcription estimate is used. It is assumed that HMM's capturing noisy background of the closest characteristics to those present in the recognized speech signal will reach the highest decoding likelihood.

3.1. Experiments & Discussion

Frequency normalizations were evaluated on clean neutral/LE data without applying cepstral compensation to observe separately their contribution to the ASR performance; see Table 1. The column *20Bands–LPC* represents a baseline setup employing 20Bands–LPC (FB 0–4 kHz) and non-normalized models. It can be seen that VTLN reduces WER of the baseline system on LE speech, meaning that scalar warping is partially effective in normalizing LE spectra.

Set	Method	20Bands LPC	VTLN Lee-Rose	VTLN Optimized	Warp&Shift	Shift	Shift + QCN4
F	Neutral	3.79	3.27	3.02	3.53	3.57	2.90
	LE	32.37	23.40	23.23	17.29	14.07	11.92
M	Neutral	1.90	2.18	2.18	1.97	2.04	1.69
	LE	18.98	18.37	17.45	12.85	13.31	11.45

Table 1: WER (%), clean data, no cepstral compensation.
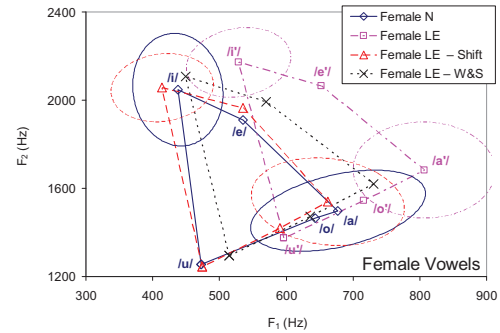

Fig. 3: Low formant equalization in *Warp&Shift* and *Shift*.

Warp&Shift further improves performance in LE, confirming that $F_{W\&S}$ can better address low and high LE formant shifts. When analyzing the distribution of $F_{W\&S}$ parameters assigned by the ML search, it was observed that $F_{W\&S}$ performs frequency mapping in a way that is mostly equivalent to VTLN ($A = [0, 0]$) during HMM training, and LE test data are mostly shifted in frequency (by β) as $\alpha \to 1$. The first observation confirms that VTLN is a good choice to address inter-speaker differences in neutral speech. The latter observation motivates the introduction of the *Shift* transform as described earlier. Fig. 3 shows locations of F_1, F_2 in vowels in neutral and LE samples, and in samples transformed by *Warp&Shift* and *Shift*, and error ellipses covering 39.4 % of the formant occurrences. It can be seen that both normalizations manage to transform formants towards the neutral locations, *Shift* being more accurate. Table 1 shows that *Shift* further reduces WER on female LE speech at an affordable cost of slight WER increase on other sets.

Cepstral normalizations: First, an optimal choice of j for QCN (see Eq. (1)) was searched in the range 1,2,...,15. When applying

QCN to the training set and a small subset of the test set (neutral/LE recordings from 2 male/2 female speakers), $j = 4$ provided the most consistent WER reduction on both neutral/LE data in all noisy conditions (denoted $QCN4$); $j = 4$ determines quantiles q_4, q_{96} bounding 92 % of the cepstral samples. The overall performance of cepstral compensations in the evaluation task is shown for female sets in Fig. 4; here, no frequency compensations were employed.
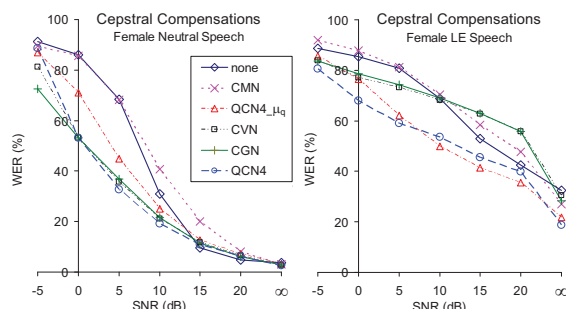

Fig. 4: Performance of cepstral compensations on noisy data.

$QCN4\_\mu_q$ denotes a setup where quantile mean μ_q was subtracted from the cepstra with no normalization of the dynamic range applied. It can be seen that $QCN4\_\mu_q$ considerably outperforms CMN on noisy neutral data and all standard compensations on LE data down to 0 dB SNR. $QCN4$ improves the performance of $QCN4\_\mu_q$ on neutral speech, reaching the best WER's down to 0 dB SNR. Similar trends were found for male data. Note that for high SNR's in neutral data, all compensations slightly increase WER compared to non-compensated features. Performance of joint $QCN4$ and *Shift* is presented for clean data in the last column of Table 1, showing that combining frequency and cepstral normalizations further reduces WER.

Set		Method	PLP + CVN		20BandsLPC + QCN4		20BandsLPC + Shift + QCN4 + Codebook	
Clean/ 10 dB SNR	F	Neutral	2.92	19.80	2.52	19.13	2.94	11.08
		LE	36.46	73.40	18.84	53.62	12.09	35.73
	M	Neutral	1.48	17.29	2.95	15.25	1.69	8.64
		LE	20.21	62.65	14.31	43.80	11.45	29.83

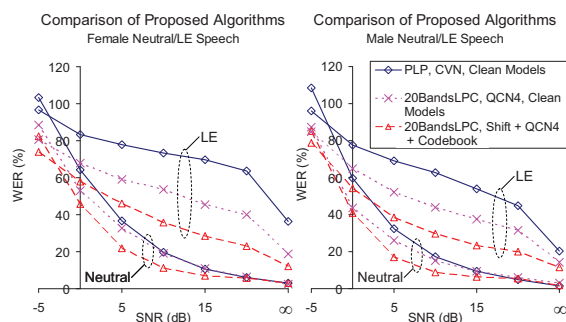**Table 2**: Resulting performance – WER (%); clean /10 dB SNR data.

Fig. 5: Performance of resulting system.

Resulting ASR system: Performance of an ASR system combining frequency normalization *Shift*, cepstral compensation *QCN4*, and a codebook of noisy models was compared to a system utilizing standard PLP front-end and CVN, and to 20Bands–LPC with *QCN4* (see Fig. 5). Table 2 shows WER's on clean and 10 dB SNR data (for each front-end, left/right columns represent clean/10 dB SNR

WER's). It can be seen that for SNR's of 15 dB and lower on neutral data and for all LE data sets the codebook recognizer provides superior performance. Analysis of the likelihood-based model assignments in the codebook system shows that in a majority of cases, noisy models trained on data of the same SNR or close SNR (± 5 dB) as appearing in the actual test utterance were selected for decoding, showing efficiency of the approach.

4. CONCLUSIONS

This study has presented novel unsupervised frequency and cepstral normalizations for noisy Lombard speech recognition, performing a maximum likelihood transformation of short-time spectra and quantile-based cepstral dynamics normalization. The normalization parameters are estimated on-the-fly from the incoming speech signal. The algorithms are incorporated in an ASR system utilizing a codebook of acoustic models trained on neutral speech mixed with noise at different levels of SNR. Evaluation tasks on speech presented in various levels of background car noise show that the proposed methods are efficient in compensating for the impact of both LE and noise and improve both neutral and Lombard speech recognition compared to common normalizations.

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. thesis, Czech Technical University in Prague, Czech Republic, http://www.utdallas.edu/~hxb076000, 2008.

[4] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures," in *ICASSP '96*, pp. 353–356.

[5] L. Welling, H. Ney, and S. Kanthak, "Speaker adaptive modeling by vocal tract normalization," *IEEE SAP*, vol. 10, no. 6, pp. 415–426, Sep 2002.

[6] S. Dharanipragada and M. Padmanabha, "A nonlinear unsupervised adaptation technique for speech recognition," in *ICSLP '2000*, vol. 4, pp. 556–559.

[7] F. Hilger and H. Ney, "Quantile based histogram equalization for noise robust large vocabulary speech recognition," *IEEE ASLP*, vol. 14, no. 3, pp. 845–854, May 2006.

[8] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *JASA*, vol. 55, no. 6, pp. 1304–1312, 1974.

[9] O. Viikki and K. Laurila, "Noise robust HMM-based speech recognition using segmental cepstral feature vector normalization," in *ESCA-NATO – RSR*, 1997, vol. 1, pp. 107–110.

[10] 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. I–209–12 vol.1.

[11] R. D. Kent and C. Read, *The Acoustic Analysis of Speech*, Whurr Publishers, San Diego, 1992.

[12] D. Iskra, B. Grosskopf, K. Marasek, H. van den Huevel, F. Diehl, and A. Kiessling, "SPEECON – Speech databases for consumer devices: Database specification and validation," in *Proc. of LREC '2002*, Las Palmas, Spain, 2002.

[13] M. J. F. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE SAP*, vol. 4, no. 5, pp. 352–359, 1996.