Unsupervised Equalization of Lombard Effect for Speech Recognition in Noisy Adverse Environments

Hynek Bořil, *Member, IEEE*, and John H. L. Hansen, *Fellow, IEEE*

***Abstract*—In the presence of environmental noise, speakers tend to adjust their speech production in an effort to preserve intelligible communication. The noise-induced speech adjustments, called Lombard effect (LE), are known to severely impact the accuracy of automatic speech recognition (ASR) systems. The reduced performance results from the mismatch between the ASR acoustic models trained typically on noise-clean neutral (modal) speech and the actual parameters of noisy LE speech. In this study, novel unsupervised frequency domain and cepstral domain equalizations that increase ASR resistance to LE are proposed and incorporated in a recognition scheme employing a codebook of noisy acoustic models. In the frequency domain, short-time speech spectra are transformed towards neutral ASR acoustic models in a maximum-likelihood fashion. Simultaneously, dynamics of cepstral samples are determined from the quantile estimates and normalized to a constant range. A codebook decoding strategy is applied to determine the noisy models best matching the actual mixture of speech and noisy background. The proposed algorithms are evaluated side by side with conventional compensation schemes on connected Czech digits presented in various levels of background car noise. The resulting system provides an absolute word error rate (WER) reduction on 10-dB signal-to-noise ratio data of 8.7% and 37.7% for female neutral and LE speech, respectively, and of 8.7% and 32.8% for male neutral and LE speech, respectively, when compared to the baseline recognizer employing perceptual linear prediction (PLP) coefficients and cepstral mean and variance normalization.**

***Index Terms*—Cepstral compensation, codebook of noisy models, frequency warping, Lombard effect, speech recognition.**

I. Introduction

LOMBARD EFFECT (LE), named after the French oto-rhino-laryngologist Etienne Lombard, who first studied the impact of environmental noise on speech production [2], is known to affect a number of speech production parameters such as vocal effort, pitch, shape and spectral slope of glottal waveforms, formant locations and bandwidths, spectral center of gravity, energy ratios in voiced/unvoiced phones, and others [3]–[7]. Numerous investigations of speech communication in noise confirm that speakers adjust their speech production [8], [9], and report that the rate of the adjustments differs when speakers simply repeat prompts/speak spontaneously or communicate with others [10]; however, it is still not completely clear to what extent the adjustments represent an automatic reflex [11] and to what extent they are conscious [12].

In the task of automatic speech recognition (ASR) in noisy adverse environments, noise and LE considerably impact the ASR performance [13]. Even if noise in the acoustic speech signal is suppressed, LE causes severe ASR degradation due to the mismatch between the parameters of LE speech and ASR acoustic models trained on noise-clean neutral (modal) speech [14], [15]. To better understand the causes of the mismatch, the following paragraphs summarize known speech production variations under LE that directly impact speech coding used in ASR. Subsequently, an overview of the past efforts in approaching LE-resistant ASR is presented.

In noise, speakers adjust their *vocal effort* [2]. For a wide range of noise levels, the dependency between voice sound pressure level (SPL) and noise SPL is almost linear, with different slope when just reading text [8] or when communicating with others [9]. The increase of vocal effort is nonuniform across phones, where vowels are often more emphasized that consonants [3], [4]. The adjustment in vocal effort is accompanied by increases in *pitch* [2], as pitch rises with an increase in both sub-glottal pressure and tension in the laryngeal musculature [16]. Pitch changes almost linearly with vocal intensity when expressed in semitones and SPL, respectively, [17].

LE introduces considerable changes in time-domain *glottal waveform profiles* [18]. In the spectral domain, the energy of LE waveforms migrates to higher frequencies, resulting in an upward shift of *spectral center of gravity* [4], [7], and in flattening of the *spectral slope* of short-time speech spectra [3], [19], [20]. The first *formant center frequency* F_1 varies inversely to the vertical position of the tongue and the second formant frequency F_2 increases with tongue advancement [21]. In LE, the increased vocal effort is accompanied by a wider mouth opening, which is realized by lowering the jaw and the tongue. As a result, F_1 shifts up in frequency [16], [22], the trend being independent on the phone context [6], [23]. F_2 rises in some phones [6] while decreasing in others [4], [24]. In [19], the F_1 increases are accompanied by consistent F_2 decreases, while in [3] and [25] the locations of both $F_{1,2}$ shift up in frequency for most phones. Average *bandwidths* of the first four formants reduce in LE for most phones [3], [4], [6], [25].

Syllable *duration* tends to be prolonged in LE [8]. Vowel duration is generally longer while the duration of consonants either increases or can decrease depending on context. The rate of the

Manuscript received April 28, 2009; revised September 16, 2009. First published October 20, 2009; current version published July 14, 2010. This work was supported by the Air Force Research Laboratory under a subcontract to RADC Inc. under FA8750-05-C-0029. A short version of the suggested compensation scheme was presented at IEEE ICASSP'09, Taipei, Taiwan [1]. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Gaël Richard.

The authors are with the Center for Robust Speech Systems (CRSS), The University of Texas at Dallas, Richardson, TX 75083-0688 USA (e-mail: hynek@utdallas.edu; john.hansen@utdallas.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2009.2034770

duration reduction in consonants is usually smaller than the duration extension in vowels, resulting in an increase of average word durations [10], [23]. Word duration changes may be either significant [3], [5], [6], or insignificant [24], depending on the conditions.

A majority of recent ASR engines employ cepstral-based encoding of the speech signal, such as mel frequency cepstral coefficients (MFCCs) [26] or perceptual linear prediction (PLP) cepstral coefficients [27]. Speech variations in LE directly impact the cepstra of short-time speech segments. Changes in vocal effort are displayed in the energy of the speech signal and in the zeroth cepstral coefficient c_0. Spectral slope of glottal waveforms affects the first and second cepstral coefficients c_1, c_2 [28]. Higher cepstral coefficients reflect the formant configuration (center frequencies and bandwidths), followed by coefficients capturing the fine structure of the spectral envelope, governed by pitch [29]. The nonuniform increase of energy in vowels and consonants will change the contour of the long-term c_0 distribution, and distributions of all cepstral coefficients will be affected by the duration changes in vowels and consonants.

In contrast to the numerous studies on noise suppression and speech enhancement, relatively limited attention has been paid to the impact and suppression of LE for ASR. Efforts to increase ASR resistance to LE can be categorized into the domains of feature extraction, LE equalization, acoustic model adjustments and adaptation, and training methods. In the *feature extraction/LE equalization* domain, speech coding employing LE-optimized filterbanks [14], [15], spectral modeling based on minimum variance distortionless response (MVDR) [30], exploiting higher order temporal derivatives of speech feature vectors [28], [31], spectral subtraction of noise and speech enhancement [32], cepstral mean subtraction and spectral tilt compensation [14], fixed formant shifting [3], [33], vocal tract length normalization [34], whole-word cepstral compensation, and source generator based cepstral compensation [5] have been proposed and shown to be effective.

In the domain of *acoustic model adjustments and adaptation*, alternative duration models [35], N-channel hidden Markov models (HMMs) [36], and codebooks of talking style dedicated models directed by talking style classifiers [37], [38] have been presented, as well as an adaptation of neutral acoustic models to speaker dependent and independent LE [6].

In the domain of *training methods*, training speaker dependent acoustic models on speech samples comprising various talking styles including LE (multi-style training) [28] has been found to be partially effective. Unfortunately, applying similar concept in speaker-independent multi-style training results in low performance [39]. If consistent LE/stress speech styles are present, ASR performance can be improved by training perturbation stress style dedicated acoustic models [40].

While these algorithms provide various degrees of success in suppressing LE, the resulting ASR performance in LE is still below that of neutral. A majority of past studies assume that there is a sufficient amount of labeled LE data available in advance 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. In real world conditions, the level of environmental noise may vary, inducing a varying level of LE [41]. In addition, LE is strongly speaker dependent [4], [20] and varies with the actual communication scenario (e.g., with the number of subjects engaged in the communication [9]). Hence, the assumption of available labeled samples matching any possible test conditions may be quite unrealistic.

This study presents novel *frequency* and *cepstral domain* transformations that equalize LE speech samples towards neutral speech distributions captured in ASR models. In contrast to many previous LE-suppression methods, 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 matching the actual conditions.

In the *frequency domain*, short time spectra are normalized in a procedure derived from a variation of the previously developed maximum-likelihood vocal tract length normalization (VTLN) [42]. Scalar frequency warping used in VTLN to compensate for inter-speaker vocal tract differences is replaced by frequency transformations to better address the formant shifts introduced by LE. In the *cepstral domain*, the dynamics of cepstral coefficients are normalized to a constant range using two quantile estimates for each cepstral dimension. Recently, advanced techniques normalizing the fine contours of cepstral histograms have been developed, utilizing either a rather extensive adaptation data sets matching the test conditions [43], or quantile-based online normalization applying two-pass search and continuity criteria [44]. In contrast to these complex methods, the goal of the cepstral compensation proposed here is to exclusively address the dynamic range mismatch in cepstral samples introduced by background noise, channel, and LE, extending the concepts of the popular and computationally inexpensive normalizations of cepstral mean (CMN) [45] and variance (CVN) [46], and recently introduced cepstral gain normalization (CGN) [47]. The novel frequency and cepstral normalizations are incorporated in a recognition scheme employing a codebook of acoustic models trained on clean data mixed with car noise at various signal-to-noise ratio (SNRs) (noisy models). The codebook-based recognition procedure selects the models best matching the actual mixture of speech and noisy background, and employs them for utterance decoding.

The remainder of this paper is organized as follows. Section II introduces frequency domain transformations that compensate for formant shifts in LE. Section III discusses variability of cepstral distributions in adverse environments, with special focus on the impact of additive environmental noise, and presents a cepstral compensation technique exploiting the distribution properties neglected by common mean and variance normalizations. Section IV describes a codebook-based strategy for noisy speech decoding. In Section V, the proposed algorithms are evaluated and compared to traditional normalizations on a database comprising neutral and LE speech samples presented at various levels of background noise. Section VI presents discussion and conclusions.

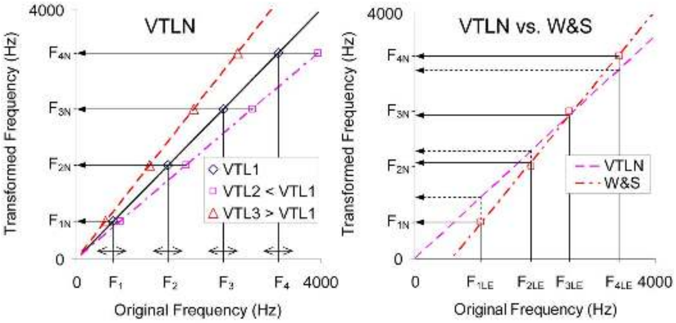

Fig. 1. Left—concept of scalar warping in VTLN. Right—motivation for increasing degrees of freedom: mapping of low and high LE formant shifts of different ratios.

II. Warp&Shift Frequency Transform

Formant locations are approximately inversely proportional to the vocal tract length (VTL) [48]. Vocal tract length normalization (VTLN) is a popular unsupervised method used to compensate for the formant shifts due to the inter-speaker VTL differences. VTLN [42] performs warping of the frequency axis F by a scalar factor α

$$F_{\mathrm{VTLN}} = \frac{F}{\alpha}. \tag{1}$$

The concept of scalar warping is demonstrated in the left part of Fig. 1, where original formant frequencies F_x displayed on the x-axis are mapped by a linear function passing through the coordinate origin to the normalized frequencies F_{yN}. In the ideal case, formants of the identical phones are mapped to the same F_{yN}'s independently of the actual VTL. In reality, both instantaneous VTL and formant configuration vary with articulation of distinct speech sounds [48]. Higher formants starting from F_3 tend to be distributed more evenly and reflect the actual VTL, while the first two formants $F_{1,2}$ are more sensitive to cross-sectional area and volume of posterior-anterior cavities varied by articulators in order to produce various phones, and less sensitive to the VTL changes [21], [49].

Two main approaches in search of the optimal α have been used in VTLN, formant-driven (FD), and maximum-likelihood (ML) search. Both approaches estimate the warping factor from long-term speech segments, such as utterances. FD search estimates α by interpolating the mean or median locations of typically higher formants by a line starting at the coordinate origin [48], [49]. In the ML VTLN [42], α is searched to maximize the likelihood of the utterance's forced alignment, given the transcription $\mathbf{W}$ and the ASR hidden Markov model (HMM) $\boldsymbol{\Theta}$:

$$\hat{\alpha} = \arg\max_{\alpha} \left[Pr(\mathbf{O}^{\alpha} | \mathbf{W}, \boldsymbol{\Theta}) \right] \tag{2}$$

where $\mathbf{O}^{\alpha}$ is a sequence of acoustic observations extracted from the utterance and warped by α. During speech recognition, the unknown $\mathbf{W}$ is first estimated by decoding unwarped data, followed by the warp selection from (2). Details of the VTLN procedure are discussed in Section V. Compared to the formant-driven approach, ML VTLN takes into account the actual characteristics captured in the ASR models, does not require reliable formant tracking (which is not available for noisy speech signals), and is more efficient in reducing ASR errors [50].

As discussed in the introduction, LE introduces considerable shifts in formant structure. F_1 consistently migrates to higher frequencies and F_2 shifts in either direction depending on the phonetic content. Higher formants are also affected by LE and shift either up or down in frequency, but their variations are not as significant [4], [25], [33]. ML VTLN employing (1) may be able to partially compensate for the low formant shifts due to LE by warping the overall spectral envelope towards neutral models, but the linear mapping function passing through the coordinate origin is unable to simultaneously address the different ratios of low and high formant shifts, especially when the ratio is higher for low formants. In the last two decades, a number of alternative VTLN transformations such as piece-wise linear, quadratic, and bilinear frequency mapping [51] have been proposed. These transformations allow for, to a certain extent, modeling different shift rates in low and high formants. Being prevalently single-parameter functions tied by the requirement of invertibility (i.e., identity mapping of 0 Hz and Nyquist frequency), these transformations display a tradeoff between the quality of low versus high formant mapping. In addition, none of these transformations are capable of effectively addressing low versus high formant shifts in the opposite direction from their neutral locations as seen in LE in [33]. In this paper, we propose a generalized linear frequency mapping function

$$F_{W\&S} = \frac{F}{\alpha} + \beta \tag{3}$$

where α represents warping as in VTLN, and β is a shift factor. As shown in the right part of Fig. 1—the dot-dashed line, extending the degrees of freedom allows for more accurate frequency mapping of different shift ratios as well as different direction of the shifts in low and high formants. Note that the accuracy of the mapping in (3) may be reduced if F_1 and F_2 shift in the opposite directions; however, the transformation will be arguably more accurate than (1). Equation (3) extends the degrees of freedom of (1) to two; hence, the parameter search grid becomes two-dimensional and more computationally challenging. However, as will be shown in Section V, the computational efforts can be considerably reduced with almost no performance cost [see the *Shift* transform introduced in Section V-D, (30)]. The ML frequency normalization employing (3) will be called $Warp\&Shift$ and abbreviated $W\&S$. Details of the Warp&Shift implementation will be presented together with performance evaluation in Section V.

III. Quantile-Based Cepstral Dynamics Normalization

Convolutional distortion introduced by the variability of the transfer channel, together with the presence of additive noise directly impacts cepstral coefficients extracted from speech, and may cause a severe mismatch between cepstral distributions of the processed speech signal and those captured in ASR acoustic models. The channel impulse response is represented in the frequency domain by a window weighting the speech spectrum, and in the cepstral domain by additive components shifting the means of cepstral coefficient distributions. If the channel characteristics vary slowly compared to speech, they can be effectively suppressed by cepstral mean subtraction (CMN) [45]. 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

$$c_{n,i}^{\mathrm{CMN}} = c_{n,i} - \bar{c}_n = c_{n,i} - \frac{1}{L}\sum_{k=1}^{L} c_{n,k} \tag{4}$$

where n is the cepstral dimension, L is the window length, and i is the index of the cepstral sample. Additive noise also contributes to the cepstral mean shifts, and moreover, affects the variance of the cepstral distributions [52]. The latter effect can be reduced by applying cepstral variance normalization (CVN) [46], which estimates the variance of each cepstral dimension $\hat{\sigma}_{Cn}$ in a long time window, and normalizes it to unity

$$c_{n,i}^{\mathrm{CVN}} = \frac{c_{n,i}^{\mathrm{CMN}}}{\hat{\sigma}_{Cn}} = c_{n,i}^{\mathrm{CMN}} \Big/ \sqrt{\frac{1}{L}\sum_{k=1}^{L}(c_{n,k} - \bar{c}_n)^2}. \tag{5}$$

Recently proposed cepstral gain normalization (CGN) has been shown to outperform CVN in suppressing the impact of additive noise [47]. Instead of variance, CGN estimates a so called cepstral gain in each dimension from the maximum and minimum sample values, $c_{n\,\max}$ and $c_{n\,\min}$, and normalizes it to unity

$$c_{n,i}^{CGN} = c_{n,i}^{\mathrm{CMN}} / (c_{n\,\max} - c_{n\,\min}). \tag{6}$$

As already noted in the introduction, speech production variations in LE directly impact cepstral distributions. Deviations of parameters such as vocal effort, spectral slope, or formant locations cause shifts in the corresponding cepstral coefficient means, while the time variability of the LE-induced changes is displayed in the distribution variances [33]. If the cepstral deviations are consistent in long term segments, CMN, CVN, and CGN may be able to transform the distributions towards neutral.

The presence of additive noise affects not only means and deviations of cepstral distributions, but also their shapes [52]. Time-variability in LE speech production, such as extended vowels and reduced unvoiced consonants, will affect the number of observations representing the particular phone classes, and their contribution in shaping the long term cepstral distributions. The above mentioned cepstral normalizations assume that the distributions to be normalized are symmetric about the mean, or at least that the asymmetry is in some sense similar in the data used to train the acoustic models and test data. If the distributions due to LE and additive noise drift from this assumption, the normalization efficiency will reduce. The following section analyzes the impact of additive noise on cepstral distributions based on a simple statistical model of MFCC speech coding.

A. *Speech Coding in Noise*

The presence of additive noise in the speech signal causes shifts of cepstral distribution means and reduction of the distribution variances [52], [53]. Moreover, in [52], higher noise levels cause a drift of distribution shapes from normal, and at low SNRs introduces bimodality into otherwise unimodal distributions. In contrast to this, in [53], no occurrence of additional distribution modes in low SNRs were observed. While the observations presented by the two studies are based on experimental experience, neither proposes an explanation of the underlying mechanisms causing the observed phenomena. In this section, the impact of additive noise is studied using a model of MFCC coding. The goal is to explain the causes of the variance and shape transformations introduced by additive noise. First, the phenomenon of cepstral variance reduction with increasing noise levels is discussed. Second, the fine contours of noisy speech distributions are analyzed. MFCC extraction comprises the following stages:

- pre-emphasis of time domain speech signal;
- windowing → short time speech segment;
- fast Fourier transform (FFT) → short time spectrum;
- squared amplitude spectrum → power spectrum;
- decimation by triangular Mel filterbank;
- log of filterbank outputs → log spectrum;
- inverse discrete cosine transform (IDCT) → MFCC.

Let the clean speech signal be mixed with additive noise at a certain SNR. By applying the first five MFCC extraction steps, the decimated power spectrum is obtained. Let $|X_{SN,k}|^2$ denote a bin in the decimated power spectrum, where SN stands for a mixture of speech and noise, and k is the bin index (index of the filterbank filter). $|X_{SN,k}|^2$ can be rewritten as

$$\begin{aligned} |X_{SN,k}|^2 &= |X_{S,k} + \sqrt{\xi}X_{N,k}|^2 \\ &= |X_{S,k}|^2 + \xi|X_{N,k}|^2 \\ &\quad + 2\sqrt{\xi}|X_{S,k}||X_{N,k}|\cos\theta \end{aligned} \tag{7}$$

where $X_{S,k}$ and $\sqrt{\xi}X_{N,k}$ are the decimated spectra of clean speech and pure noise, respectively, the non-negative real constant ξ is inversely proportional to the SNR of the speech/noise mixture, and θ is the phase difference between $X_{S,k}$ and $X_{N,k}$. Assuming that the speech and noise signals are statistically independent, the expected value of $S_{SN,k} = |X_{SN,k}|^2$ is

$$E(S_{SN,k}) \equiv E(S_{S,k}) + \xi E(S_{N,k}) \tag{8}$$

where $S_{S,k}$ and $S_{N,k}$ denote the clean speech and noise power spectrum bins. Note that the expected value of the cosine term in (7) is zero. Equation (8) forms a basis for numerous spectral subtraction techniques [54]. The variance of $S_{SN,k}$ can be evaluated as

$$Var(S_{SN,k}) = E\left(S_{SN,k}^2\right) - \left[E(S_{SN,k})\right]^2. \tag{9}$$

Substituting (7) for $S_{SN,k}$ in (9) yields

$$Var(S_{SN,k}) = Var(S_{S,k}) + \xi Var(S_{N,k}). \tag{10}$$

Similarly as in the case of (8), the expected value of all terms containing $\cos\theta$ is zero and hence, they have no impact on the variance of $S_{SN,k}$. It can be seen that adding noise to the clean speech signal causes an increase in the mean and variance of the filterbank outputs. In the following step, the impact of the mean and variance increase on the variance of log power spectrum will be analyzed. For simplicity, assume that $S_{S,k}$ and $S_{N,k}$ have *unimodal* normal distributions (in the case of $S_{N,k}$, the decimated filterbank output can be seen as a sum of random variables—spectral bin energies—and due to the central limit theorem it is expected to approach a normal distribution with an increasing number of bins per filter)

$$S_{S,k} \sim \mathcal{N}\left(\mu_{S,k}, \sigma^2_{S,k}\right), \quad \xi S_{N,k} \sim \mathcal{N}\left(\xi\mu_{N,k}, \xi^2\sigma^2_{N,k}\right). \tag{11}$$

In such a case, the distribution of $S_{SN,k}$ will also be normal

$$\begin{aligned} S_{SN,k} &= S_{N,k} + \xi S_{N,k} \\ &\sim \mathcal{N}\left(\mu_{SN,k}, \sigma^2_{SN,k}\right) \\ &= \mathcal{N}\left(\mu_{S,k} + \xi\mu_{N,k}, \sigma^2_{S,k} + \xi^2\sigma^2_{N,k}\right). \end{aligned} \tag{12}$$

Let $F_{S_{SN,k}}(x)$ be a cumulative distribution function (cdf) of $S_{SN,k}$: $F_{S_{SN,k}}(x) = P(S_{SN,k} \leq x)$, and $\widehat{S_{SN,k}}$ be a bin of the log power spectrum: $\widehat{S_{SN,k}} \equiv \log(S_{SN,k})$. CDF of $\widehat{S_{SN,k}}$ can then be expressed as

$$F_{\widehat{S_{SN,k}}}(y) = P(\widehat{S_{SN,k}} \leqslant y) = P(S_{SN,k} \leqslant e^y). \tag{13}$$

Next, the probability density function (pdf) of $\widehat{S_{SN,k}}$ can be obtained by differentiating (13) with respect to y

$$\begin{aligned} f_{\widehat{S_{SN,k}}}(y) &= \frac{d}{dy} F_{S_{SN,k}}(e^y) \\ &= e^y f_{S_{SN,k}}(e^y) = \frac{e^y}{\sigma\sqrt{2\pi}} e^{-\frac{(e^y - \mu_{SN,k})^2}{2\sigma^2_{SN,k}}}. \end{aligned} \tag{14}$$

Exploiting the following property of the moment generating function (mgf) $M(t)$ [55], $E(\widehat{S_{SN,k}}^{\,r}) = M^{(r)}(0)$, the variance of $\widehat{S_{SN,k}}$ can be found as

$$\begin{aligned} Var(\widehat{S_{SN,k}}) &= E\left(\widehat{S_{SN,k}}^{\,2}\right) - \left[E(\widehat{S_{SN,k}})\right]^2 \\ &= \left.\frac{d^2}{dt^2} M(t) - \left[\frac{d}{dt} M(t)\right]^2\right|_{t=0} \end{aligned} \tag{15}$$

where

$$M(t) = \int_{-\infty}^{\infty} e^{ty} f_{\hat{S}_{SN,k}}(y)dy = \int_{-\infty}^{\infty} \frac{e^{y(t+1)}}{\sigma\sqrt{2\pi}} e^{-\frac{(e^y - \mu_{SN,k})^2}{2\sigma^2_{SN,k}}} dy. \tag{16}$$

Unfortunately, neither the integral in (16), nor its first and second derivatives with respect to t have a closed form solution. For this reason, instead of comparing clean speech and noisy speech log spectrum pdf's in analytic form, we analyze the variances indirectly, based on the properties of the log transformation. Fig. 2 shows an example of mapping the clean and noisy speech power spectrum pdf's (*power pdf's*) to the corresponding log power spectrum pdf's (*log pdf's*). Three phenomena can be observed in Fig. 2:

- adding noise to speech increases mean and variance of power spectrum bins; compare with (12);
- due to flattening of the log function slope with increasing x values, increasing the mean of power pdf's results in compressed variance of log pdf's;
- skewness of the log pdf's varies with the location of power pdf's on the x-axis. The closer the pdf mean is to the origin of the coordinates, the more pronounced the skewness in the log pdf.

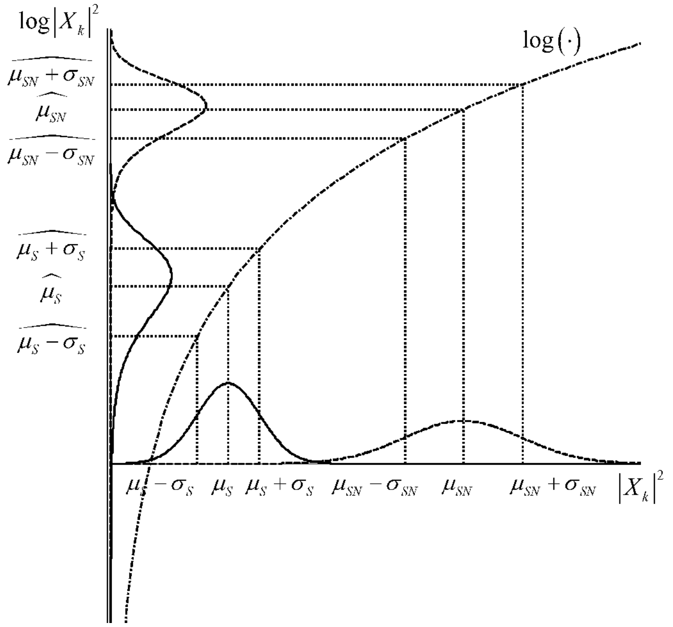

Fig. 2. Impact of noise on distributions of speech power spectrum bins; unimodal speech distributions. Note that locations of maxima in power pdf's are not mapped exactly to maxima of log pdf's as normal distributions are weighted by factor e^y while being projected to logarithmic scale [Equation (14)].

The first and second phenomenon have an opposite impact on the variance of the log pdf's. Increasing the level of additive noise in the speech signal [i.e., increasing ξ in (12)], will increase the variance of the power pdf. At the same time, the associated increase of the power pdf mean will shift the distribution up on the x-axis, inducing a stronger compression of the log pdf variance due to the flattening of the log function. Let us examine which of these two factors will have a prevalent impact on the resulting log pdf variance. Given the assumption of normality of the $S_{S,k}$ and $S_{N,k}$ pdf's, the x-axis intervals $i_S = [\mu_S - \sigma_S, \mu_S + \sigma_S]$ and $i_{SN} = [\mu_{SN} - \sigma_{SN}, \mu_{SN} + \sigma_{SN}]$ can be expected to capture approximately 68.2% of the clean speech and noisy speech samples, respectively. Since log is a monotonic function, the y-axis intervals $\widehat{i_S} = [\widehat{\mu_S - \sigma_S}, \widehat{\mu_S + \sigma_S}]$ and $\widehat{i_{SN}} = [\widehat{\mu_{SN} - \sigma_{SN}}, \widehat{\mu_{SN} + \sigma_{SN}}]$ contain the same number of samples like their x-axis counterparts. While the width of $\widehat{i_{SN}}$ will change with the mean and variance of the corresponding power pdf, the number of samples captured in the interval will remain constant. Hence, the trend in compressing or expanding $\widehat{i_{SN}}$ is proportional to the reduction or increase of the log pdf variance. The next step will analyze, under which conditions the following properties hold:

$$\begin{aligned} \widehat{\mu_S} - \widehat{\mu_S - \sigma_S} &> \widehat{\mu_{SN}} - \widehat{\mu_{SN} - \sigma_{SN}} \\ \widehat{\mu_S + \sigma_S} - \widehat{\mu_S} &> \widehat{\mu_{SN} + \sigma_{SN}} - \widehat{\mu_{SN}} \end{aligned} \tag{17}$$

or in other words, under which conditions adding noise to the clean speech signal will cause the interval $\widehat{i_{SN}}$ to become narrower than $\widehat{i_S}$. After a slight manipulation and substituting from (12), the first inequality (17) can be rewritten

$$\frac{\sigma_S}{\mu_S} > \frac{\sigma_{SN}}{\mu_{SN}} = \frac{\sqrt{\sigma_S^2 + \xi^2\sigma_N^2}}{\mu_S + \xi\mu_N} = \frac{\sigma_S\sqrt{1+\lambda_2^2}}{\mu_S(1+\lambda_1)} \tag{18}$$

where λ_1 and λ_2 are non-negative constants representing the relation between the clean speech and noise power spectral means and variances: $\lambda_1 = \xi\mu_N/\mu_s$, $\lambda_2 = \xi\sigma_N/\sigma_s$. The inequality is true for all λ_1 and λ_2 such that

$$\frac{\sqrt{1+\lambda_2^2}}{1+\lambda_1} < 1. \tag{19}$$

The second inequality in (17) results in an identical expression. Clearly, all non-negative $\lambda_1 > \lambda_2$ satisfy inequality (19), which means that if the ratio between the means of noise and clean speech power spectrum bins, λ_1, is bigger than the ratio between their variances, λ_2, the variance of the noisy speech log power spectrum bin will always decrease, with a rate inversely proportional to SNR. Note that the left and right sides of the inequality (18) are proportional to the variance of clean speech and noisy speech log spectra, respectively, and that for $\lambda_1, \lambda_2 \to 0$ the variance of noisy speech will approach that of the clean speech, and for $\lambda_1 \gg 1, \lambda_2^2 \gg 1$ will approach that of pure noise. When compared to clean speech, many environmental noises can be considered quasi-stationary; hence, their power spectrum variances are significantly smaller $(\lambda_1 \gg \lambda_2)$. Here, increasing the level of noise in speech will cause a decrease of the spectral bin variances, starting at values equal to the clean speech variances for high SNRs, and reaching the values of noise variances at low SNRs.

Cepstral coefficients are obtained by applying the DCT transform to the filterbank log-energies

$$c_n(i) = \sum_{k=1}^{K} \widehat{S_{SN,k}(i)} \cos\left[n\left(k-\frac{1}{2}\right)\frac{\pi}{K}\right] \tag{20}$$

where n is the cepstral dimension, i is the frame index over time, and K is the number of filterbank filters. Since variances of the log-energies of noisy speech satisfying inequality (19) decrease, the variance of their sum will also decrease.

While a unimodal Gaussian is a reasonable approximation for the spectral distribution of many environmental noises, it may not be as effective in representing long term distributions of clean speech spectra. For example, segmental energy in the zeroth power spectrum bin possesses a multimodal distribution where low energy components correspond to nonspeech and unvoiced speech segments, and high energy components represent voiced speech [53]. Another dominant source of multimodality is spectral slope, which is in general steep in voiced segments and more flat in unvoiced and nonspeech segments [56], affecting distributions of all spectral bins. Next, we analyze the impact of noise on multimodal spectral distributions of speech modeled by mixtures of Gaussians.

To simplify the notation, let $X \equiv S_{S,k}$, $Y \equiv \xi S_{N,k}$, and $Z \equiv S_{SN,k} = X + Y$. Let the multimodal pdf of X be a mixture of weighted normal distributions

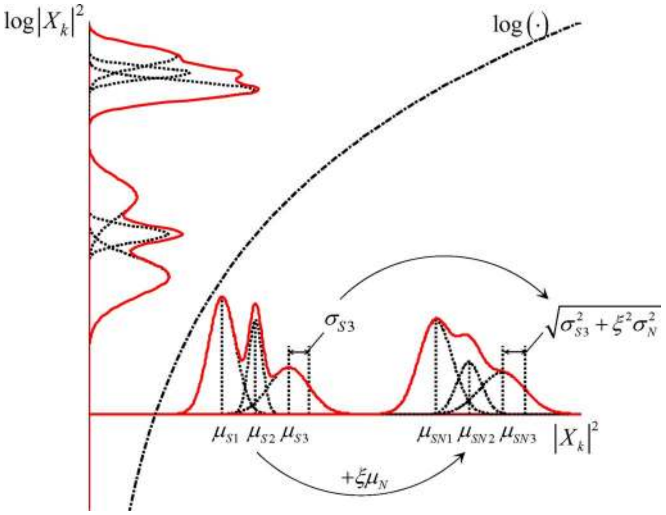
$\log|X_k|^2$
$\log(\cdot)$
σ_{S3}
$\sqrt{\sigma_{S3}^2+\xi^2\sigma_N^2}$
μ_{S1} μ_{S2} μ_{S3}
μ_{SN1} μ_{SN2} μ_{SN3}
$|X_k|^2$
$+\xi\mu_N$

Fig. 3. Impact of noise on distributions of speech power spectrum bins; multimodal speech distributions.

$$f_X(x) = \sum_{m=1}^{M} a_m \mathcal{N}\left(x; \mu_{X,m}, \sigma_{X,m}^2\right) \tag{21}$$

where M is the number of mixture components, a_m is the weight of the mth component, and $\sum_{m=1}^{M} a_m = 1$. The pdf of Z can be obtained by convolving the pdf's of X and Y:

$$f_Z(z) = \sum_{m=1}^{M} a_m \int_{-\infty}^{\infty} f_{X,m}(x) f_Y(z-x)dx \tag{22}$$

where $f_{X,m}$ denotes the mth unweighted component of the multimodal speech distribution. If f_Y has a normal distribution $\mathcal{N}(\mu_Y, \sigma_Y^2)$, (22) can be rewritten

$$f_Z(z) = \sum_{m=1}^{M} a_m \mathcal{N}\left(z; \mu_{X,m} + \mu_Y, \sigma_{X,m}^2 + \sigma_Y^2\right). \tag{23}$$

The last equation shows that within each power spectrum bin, components of the clean speech multimodal distribution are shifted by a constant rate μ_Y, and their variances increase by a constant σ_Y^2. Since the spacing between the component means in the noisy speech distribution remains preserved as in clean speech, increasing the variances causes a blending of the mixture components and a gradual vanishing of local peaks and valleys in the resulting noisy speech distribution (see Fig. 3). At low SNRs, the original multimodal power spectrum distributions will become unimodal. This trend will also be transferred to the log power spectrum, due to the monotonicity of the *log* function. As shown in (20), cepstral coefficients are obtained as weighted sums of random variables—log power spectral bins. Hence, the distribution of a cepstral coefficient equals the convolution of the distributions of weighted log power spectral bins. Arguably, convolving log power spectrum pdf's that approach unimodal contours will yield cepstral pdf's also approaching unimodal contours.

In the next step, the mean and variance of multimodal speech power spectrum distributions mixed with noise are derived using the mgf of $f_Z(z)$

$$M(t) = \int_{-\infty}^{\infty} e^{tz} f_Z(z) dz = \sum_{m=1}^{M} a_m \int_{-\infty}^{\infty} e^{tz} f_{Z,m}(z) dz \quad (24)$$

where $f_{Z,m}$ is the unweighted mth mixture component of f_Z [see (23)]. The mean of the noisy power spectrum bin distribution is then

$$E(Z) = M_Z'(0) = \sum_{m=1}^{M} a_m \mu_{X,m} + \mu_Y \quad (25)$$

where the sum term on the right-hand side of (25) represents the mean of the clean speech multimodal distribution. The variance can be obtained as

$$\begin{aligned} Var(Z) &= M''(0) - [M'(0)]^2 \\ &= \sum_{m=1}^{M} a_m \left(\mu_{X,m}^2 + \sigma_{X,m}^2\right) \\ &\quad - \left(\sum_{n=1}^{M} a_n \mu_{X,n}\right)^2 + \sigma_Y^2 \end{aligned} \quad (26)$$

where the sum terms in the last two rows represent the variance of the clean speech multimodal distribution. Equations (25) and (26) show that the impact of noise on the overall mean and variance of the multimodal distributions of the speech power spectrum is identical as that in the case of unimodal speech spectra, see (12) (i.e., the noise mean and variance are added to the multimodal distribution mean and variance, respectively). Hence, the cepstral variances will reduce in the presence of noise if the inequality (19) is fulfilled.

This section derives the conditions under which adding noise to speech will cause a reduction of the cepstral distribution variances, and analyzes the impact of noise with normal power spectrum bin distributions on the noisy speech distribution shapes. In the latter case, it was shown that increasing the level of noise in speech will cause a blending of modes of the multimodal power spectral distributions [(23)], rather than introducing new modes into otherwise unimodal distributions [(12)]. This conclusion corresponds well with the experimental observations made in [53]. Alternatively, the bimodality observed in [52] when adding white Gaussian noise to speech at low SNRs is somewhat surprising. An example of the impact of additive car noise on the c_0 distributions in female LE speech is shown in Fig. 4 (the data set and the feature extraction front-end are described in Section V).

B. Novel Cepstral Compensation

The cepstral normalizations mentioned in the beginning of Section III were designed while making certain assumptions about the signal to be normalized. When the signal properties drift from these assumptions, the normalization efficiency can be expected to decrease.

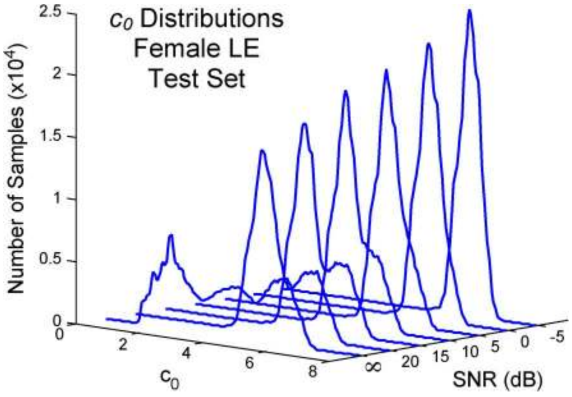

Fig. 4. Impact of additive car noise on c_0 distributions extracted using *20 Bands-LPC* front-end. Increasing level of noise in speech signal results in reduction of distribution variance, and transformation of multimodal clean speech distribution towards unimodal.

CMN assumes that the distributions to be normalized are symmetric to their means and have similar variances in training and test sets. In such a case, subtracting sample mean from the incoming test samples will assure that their *dynamic range* (an interval capturing certain amount of samples, centered to leave equal number of samples below and above the interval boundaries) will match the one in the data used to train the ASR acoustic models (training set). CMN can still provide good dynamic range matching even for asymmetric distributions, if the "asymmetricity" is similar in training and test data.

CVN extends CMN by estimating the distribution variance and normalizing it to unity. While variance represents well the sample dynamic range in the case of normal distribution, its accuracy reduces as the distribution skewness and shape deviate from normal. CVN may still be efficient if the deviation is similar in training and test data.

CGN estimates the sample dynamic range directly from the minimum and maximum sample values, and hence, does not require the distributions to be normal. However, CGN incorporates CMN and thus requires the distributions to be symmetric to their means.

It has been shown in Section III-A that the presence of additive noise in speech signal impacts not only means and variances of spectral and cepstral distributions, but also their skewness. Changes in vocal effort, spectral slope, phone durations, and other parameter variations introduced by LE will also affect means [33] and contours of cepstral distributions; see an example of c_0 histograms extracted from an extensive set of neutral and LE clean female utterances in the right part of Fig. 5—the dashed line represents generalized extreme value (GEV) interpolation of the histogram samples, μ is the sample mean, and q_5 and q_{95} are 5% and 95% quantiles bounding 90% of the histogram samples.

If the skewness of the training and test cepstral distributions is different, the efficiency of CMN, CVN, and CGN will reduce, even if the distribution variance is normalized accurately. The left part of Fig. 5 shows an example of two distributions with equal variance and opposite skewness (Distribution 1 and Distribution 2). It can be seen that although the means of the two

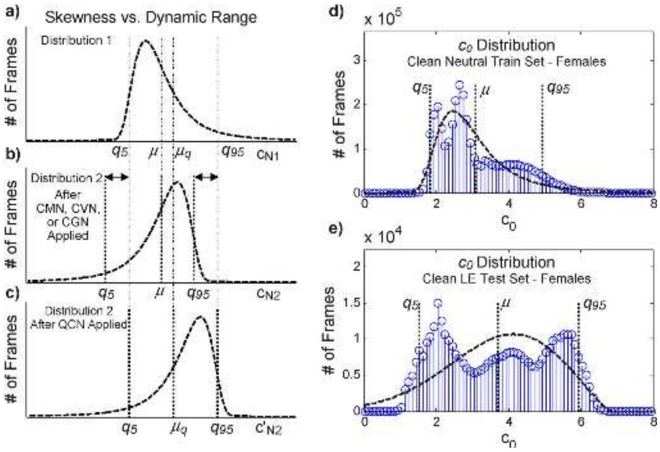

Fig. 5. Impact of skewness differences on the accuracy of dynamic range normalization. (a), (b) Distributions with opposite skewness (distribution 1 and 2), applying CMN/CVN/CGN, or QCN aligns the distribution means and variances, but *does not* address the dynamic range mismatch (intervals bounded by quantiles q_5 and q_{95}). (c) Applying QCN guarantees dynamic range alignment. (d), (e) Mismatch in skewness due to LE, large set c_o histograms.

distributions are aligned correctly (the first two upper distributions), the skewness difference results in the mismatch of the sample dynamic ranges (denoted by horizontal arrows).

To address the dynamic range mismatch due to the difference in distributions' skewness, we propose so called quantile-based cepstral dynamics normalization (QCN). To reduce the sensitivity to outliers seen in CGN, which estimates the dynamic range from only two extreme samples, the cepstral dynamic range of the nth cepstral dimension 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 $round(j \times L/100)$ and $round[(100-j) \times L/100]$, where L is the number of samples. Instead of subtracting the distribution mean as conducted in CMN, CVN, and CGN, which would introduce a requirement of the distribution symmetricity, an average of q_j^{Cn} and q_{100-j}^{Cn}, denoted μ_q in Fig. 5, is subtracted. QCN is defined

$$c_{n,i}^{QCNj} = \frac{c_{n,i} - \left(q_j^{Cn} + q_{100-j}^{Cn}\right)/2}{q_{100-j}^{Cn} - q_j^{Cn}} \tag{27}$$

where i is the index of a cepstral sample in the long time window. As shown in the left part of Fig. 5, QCN provides more accurate dynamic range normalization of distributions with different skewness than CMN, CVN, and CGN; compare the first and the third distribution.

IV. Recognizer With Codebook of Noisy Models

Ideally, ASR acoustic models should match the background noise characteristics from the input speech [57]. This requirement is difficult to meet in real world scenarios where noise level and type may vary continuously, introducing mismatch between the acoustic models typically trained with either clean or partial noise data. Numerous techniques to reduce mismatch have been proposed in two domains: 1) transformation of noisy speech towards clean, such as noise suppression [32], [54], [58]; 2) noise modeling in the ASR back-end, such as updating acoustic models for noise characteristics [58], using two-dimensional HMMs to decompose speech and noise components [59], and parallel model combination (PMC) [60]. In [32], front-end constrained iterative enhancement was shown to improve ASR in noise. However, in [58], back-end noise modeling with sufficient noise data was shown both analytically and experimentally to be preferable for ASR compared to noise suppression techniques.

In this section, a simple speech decoding scheme employing a codebook of noisy acoustic models is presented in order to better match input noisy speech and ASR acoustic models in changing background noise. The codebook consists of HMMs trained on data with different SNRs, where clean speech is mixed with car noises at SNRs of $-5, 0, \ldots, 20, \infty$ dB, yielding "noisy" models denoted $\lambda_1, \ldots, \lambda_7$. During recognition, the observation sequence $\mathbf{O}$ is decoded as

$$\hat{\mathbf{W}}_n = \arg\max_{\mathbf{W} \in \mathfrak{L}} \left[Pr(\mathbf{O}|\mathbf{W}, \boldsymbol{\lambda}_n) Pr(\mathbf{W}|\boldsymbol{\Theta}_l)\right] \tag{28}$$

where $\hat{\mathbf{W}}$ is the sequence of words for language $\mathfrak{L}$, and $\boldsymbol{\Theta}_l$ denotes the language model. Applying (28) consecutively for all noisy models $\lambda_1, \ldots, \lambda_7$ yields a set of transcription estimates $\hat{\mathbf{W}}_1, \ldots, \hat{\mathbf{W}}_7$. Subsequently, the λ_n that provides the highest likelihood decoding path (best match) is found

$$BM = \arg\max_n P(\mathbf{O}|\hat{\mathbf{W}}_n, \lambda_n) \tag{29}$$

with the corresponding transcription $\hat{\mathbf{W}}_{BM}$. The HMM with the closest noise structure is expected to give the highest score. The complete decoding scheme incorporating QCN and $Warp\&Shift$ is shown in Fig. 6.

In the present setup, only noise from the cabin of a moving car [61] is used for noisy model training, however, the codebook can be extended to multiple noise types depending on the intended application. Compared with standard HMM ASR using single-pass Viterbi decoding, codebook recognition increases computational complexity with decoding passes using multiple noisy models.

V. Experiments

Next, the proposed algorithms are evaluated side-by-side with traditional ASR techniques. First, the experimental framework is described, followed by the comparison of three feature extraction front-ends on a baseline ASR task. The following subsections evaluate the proposed frequency and cepstral based normalizations. Finally, a combination of the proposed algorithms is evaluated using the codebook recognition task.

A. Corpus

All algorithms are evaluated on the Czech Lombard Speech Database (CLSD'05) [62], comprising recordings of neutral speech and speech uttered in simulated noisy conditions (LE conditions), where 90-dB SPL of car noise samples from the CAR2E database [61] were heard by speakers through closed-ear headphones. Speech was collected using a close-talk microphone, yielding high SNR signals (mean SNR of 28 dB and 41 dB in neutral and LE recordings, respectively). The

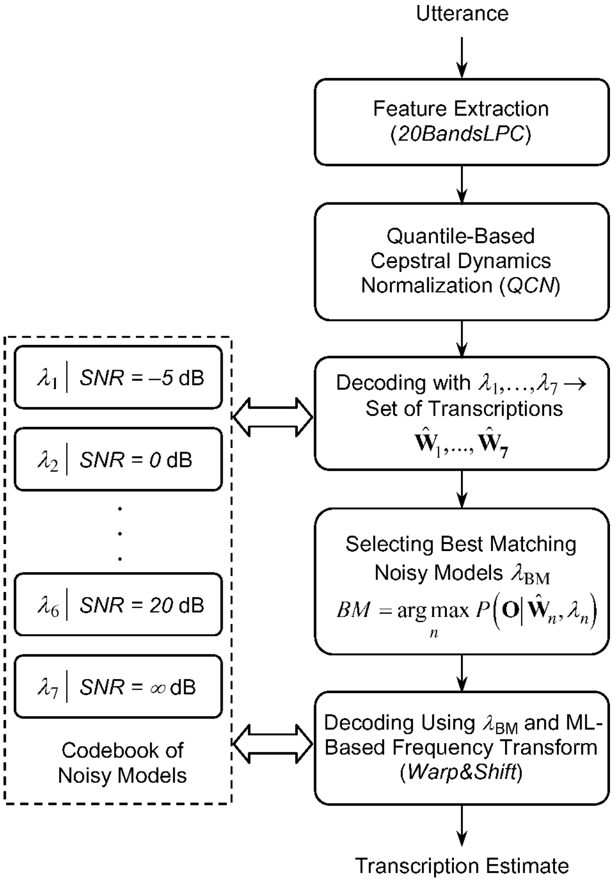
Utterance
Feature Extraction (20BandsLPC)
Quantile-Based Cepstral Dynamics Normalization (QCN)
Decoding with $\lambda_1,\ldots,\lambda_7 \rightarrow$ Set of Transcriptions $\hat{\mathbf{W}}_1,\ldots,\hat{\mathbf{W}}_7$
Selecting Best Matching Noisy Models λ_{BM}
$BM = \arg\max_n P\left(\mathbf{O}|\hat{\mathbf{W}}_n,\lambda_n\right)$
Decoding Using λ_{BM} and ML-Based Frequency Transform (Warp&Shift)
Transcription Estimate
λ_1 | SNR = –5 dB
λ_2 | SNR = 0 dB
λ_6 | SNR = 20 dB
λ_7 | SNR = ∞ dB
Codebook of Noisy Models

Fig. 6. ASR system employing codebook of noisy models and QCN and $Warp\&Shift$ normalizations.

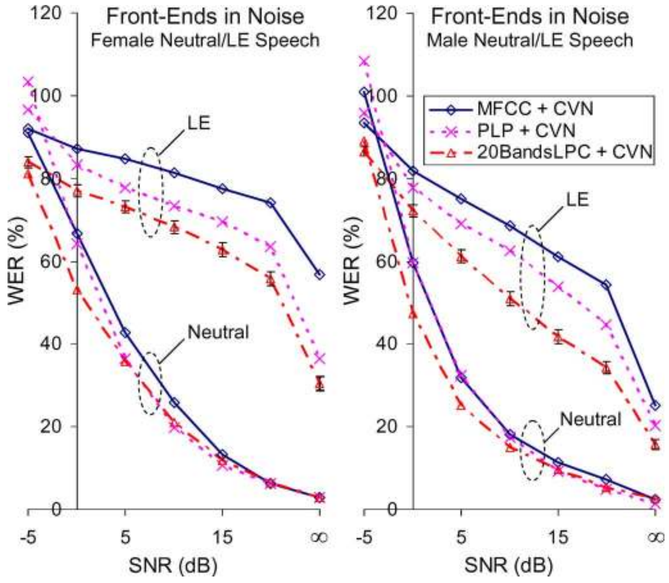

Fig. 7. Baseline front-ends comprising CVN: performance in noisy conditions.

TABLE I
BASELINE FRONT-ENDS COMPRISING CVN: PERFORMANCE ON CLEAN DATA; WER (%)

	Neutral Set, WER (%)			LE Set, WER (%)		
Gender	MFCC	PLP	20Bands LPC	MFCC	PLP	20Bands LPC
F	2.8	2.9	2.9	56.7	36.5	30.5
M	2.3	1.5	2.6	25.2	20.2	15.6

subjects were engaged in subject-to-human collection operator communication over noise, where the collection operator was instructed to ask the subject to repeat utterances if the message was not intelligible to ensure proper reaction to the noisy background. The recordings were downsampled to 8 kHz and filtered by a G.712 telephone filter and mixed with 20 noise samples at SNRs of $-5, 0, \ldots, 20, \infty$ dB, where ∞ dB represents clean data. The noise samples form a subset of the samples used in the CLSD'05 LE sessions' acquisition.

B. Recognition Setup

The baseline HMM-based recognizer comprises 43 context-independent monophone models and two silence models (mostly three emitting states, 32 Gaussian mixtures). Speech is parameterized with static cepstral coefficients $c_0 - c_{12}$ and their first- and second-order time derivatives. Gender-dependent phoneme models are trained with 46 iterations with large vocabulary material from 37 female/30 male speaker sessions from the Czech SPEECON database [63]. The sessions contain clean speech from the office environment.

The ASR task is to recognize ten Czech digits (16 pronunciation variants) in connected digit strings. The female neutral/LE test sets comprise a total of 4930/5360 words, respectively, uttered by 12 speakers, and male neutral/LE test sets comprise 1423/6303 words uttered by 14 speakers. Performance is assessed by means of word error rate (WER).

C. Baseline ASR Task: Front-End Comparison

The initial recognition experiment compares three feature extraction front-ends, MFCC, PLP, and 20 Bands-LPC, derived from PLP by replacing the trapezoid filterbank with a bank of 20 non-overlapping rectangular filters uniformly spaced on a linear scale over 0–4 kHz. In a previous study [38], 20 Bands-LPC cepstral coefficients displayed comparable performance to MFCC and PLP on clean neutral speech, and superior performance on clean LE speech. In the present experiment, performance of front-ends (incorporating CVN) is evaluated with various degrees of car noise; Fig. 7 summarizes WERs in all tested conditions and Table I details WERs on clean data, where F and M denote female and male sets. To demonstrate the statistical significance of the results, the 20 Bands-LPC LE plots in Fig. 7 are accompanied by 99% confidence intervals. It can be seen that in clean neutral conditions, PLP performs best on male data, and all front-ends perform comparably on female data. Since 20 Bands-LPC provides superior performance on LE speech for both genders in all noisy conditions, as well as on neutral speech at lower SNRs, it is selected as a front-end for the following experiments.

D. Frequency Normalizations

$Warp\&Shift$ frequency normalization is evaluated together with two-pass maximum likelihood (ML) vocal tract normalization (VTLN) [42], denoted "*Lee-Rose*," and *fully optimized*

TABLE II
HMM TRAINING EMPLOYING FREQUENCY NORMALIZATION

1) Train non-normalized HMM's $\boldsymbol{\Theta}$ in 36 iterations;

2) For each $\psi \in \Psi$, transform training utterances $\mathbf{O}_{utt} \rightarrow \mathbf{O}^{\psi}_{utt}$. Perform forced alignment for each $\mathbf{O}^{\psi}_{utt}$ and find ψ maximizing utterance alignment likelihood:

$$\psi_{utt} = \arg\max_{\psi}\left[Pr\left(\mathbf{O}^{\psi}_{utt}\middle|\mathbf{W}_{utt},\boldsymbol{\Theta}\right)\right];$$

2a) If speaker-dependent normalization, each ψ_{utt} is replaced by a median of all speaker's ψ_{utt}'s: $\psi_{spk} = \underset{utt \in Spk}{Med}\left(\psi_{utt}\right)$;

3) Transform training set by ψ_{utt}'s or ψ_{spk}'s; retrain $\boldsymbol{\Theta}$ on transformed set in n iterations, $n = 3$;

4) Repeat steps 2, 2a; repeat step 3, $n = 7$, which will yield normalized acoustic models $\boldsymbol{\Theta}_N$.

TABLE III
TWO-PASS DECODING PERFORMED BY ML VTLN "LEE-ROSE"

1) Decode test utterances using normalized models $\boldsymbol{\Theta}_N$:

$$\hat{\mathbf{W}}^{N}_{utt} = \arg\max_{\mathbf{W}\in\mathfrak{L}}\left[Pr\left(\mathbf{O}_{utt}\middle|\mathbf{W},\boldsymbol{\Theta}_N\right)Pr\left(\mathbf{W}\middle|\boldsymbol{\Theta}_l\right)\right];$$

2) For each $\psi \in \Psi$, transform test utterances $\mathbf{O}_{utt} \rightarrow \mathbf{O}^{\psi}_{utt}$;
Perform forced alignment for each $\mathbf{O}^{\psi}_{utt}$ and find ψ maximizing utterance alignment likelihood:

$$\psi_{utt} = \arg\max_{\psi}\left[Pr\left(\mathbf{O}^{\psi}_{utt}\middle|\mathbf{W}^{N}_{utt},\boldsymbol{\Theta}_N\right)\right];$$

3) Transform test set by ψ_{utt}'s; decode transformed set using normalized models $\boldsymbol{\Theta}_N$:

$$\hat{\mathbf{W}}_{utt} = \arg\max_{\mathbf{W}\in\mathfrak{L}}\left[Pr\left(\mathbf{O}^{\psi}_{utt}\middle|\mathbf{W},\boldsymbol{\Theta}_N\right)Pr\left(\mathbf{W}\middle|\boldsymbol{\Theta}_l\right)\right].$$

ML VTLN [64]. Training of the frequency normalized HMMs is implemented similarly for all three methods and summarized in Table II. First, frequency non-normalized acoustic models $\boldsymbol{\Theta}$ are trained. Second, optimal frequency transform parameters for each utterance (ψ_{utt}) are found in the parameter search grid Ψ. In the case of VTLN "*Lee-Rose*," speaker-dependent transform parameters ψ_{spk}'s are determined as a median of the speaker's ψ_{utt}'s. Third, the train set is normalized by ψ_{utt}'s or ψ_{spk}'s and used for retraining $\boldsymbol{\Theta}$ multiple times. Fourth, Step 3 is repeated, yielding frequency normalized acoustic models $\boldsymbol{\Theta}_{\mathbf{N}}$.

During recognition, the VTLN "*Lee-Rose*" method uses two-pass decoding; see Table III. In the first decoding pass on the non-normalized data, the unknown utterance transcription $\hat{\mathbf{W}}^{N}_{\mathrm{utt}}$ is estimated. In the second pass, optimal ψ_{utt} is searched given the transcription estimate $\hat{\mathbf{W}}^{N}_{\mathrm{utt}}$ and used for normalizing the utterance (yielding $\mathbf{O}^{\psi}_{\mathrm{utt}}$). Subsequently, the resulting transcription $\hat{\mathbf{W}}_{\mathrm{utt}}$ is estimated by decoding $\mathbf{O}^{\psi}_{\mathrm{utt}}$.

In the *fully optimized* VTLN and $Warp\&Shift$, the fully optimized decoding strategy [64] is applied; see Table IV. The utterance is consecutively transformed and decoded for each $\psi \in \Psi$, yielding transcription estimates $\hat{\mathbf{W}}^{\psi}_{\mathrm{utt}}$. Here, ψ_{max} associated with the highest likelihood decoding path is found and $\hat{\mathbf{W}}^{\psi\,\max}_{\mathrm{utt}}$ is taken.

TABLE IV
FULLY OPTIMIZED VTLN AND WARP&SHIFT DECODING

1) For each $\psi \in \Psi$, transform test utterances $\mathbf{O}_{utt} \rightarrow \mathbf{O}^{\psi}_{utt}$;
Decode transformed set using normalized models $\boldsymbol{\Theta}_N$:

$$\hat{\mathbf{W}}^{\psi}_{utt} = \arg\max_{\mathbf{W}\in\mathfrak{L}}\left[Pr\left(\mathbf{O}^{\psi}_{utt}\middle|\mathbf{W},\boldsymbol{\Theta}_N\right)Pr\left(\mathbf{W}\middle|\boldsymbol{\Theta}_l\right)\right];$$

2) Find ψ_{max} maximizing decoding likelihood across Ψ:

$$\psi_{\max} = \arg\max_{\psi}\left[Pr\left(\mathbf{O}^{\psi}_{utt}\middle|\hat{\mathbf{W}}^{\psi}_{utt},\boldsymbol{\Theta}_N\right)Pr\left(\hat{\mathbf{W}}^{\psi}_{utt}\middle|\boldsymbol{\Theta}_l\right)\right]$$
$$\rightarrow \hat{\mathbf{W}}_{utt} = \hat{\mathbf{W}}^{\psi\max}_{utt}.$$

In both VTLN algorithms, the search grid of $\alpha \equiv \psi$ [(1)] is chosen as $\Psi = \{0.8, 0.85, \ldots, 1.2\}$. Similar to [42], the frequency normalization is conducted by transforming the front-end filterbank (FB) cutoff frequencies. To avoid exceeding Nyquist frequency of 4 kHz during FB expansion, the initial FB is set to span over 0–3200 Hz in both VTLN setups.

For $Warp\&Shift$, the frequency transform [(3)] employs two parameter search grid. As in VTLN, the transformation is realized by manipulating the FB, which in the initial form spans 0–3200 Hz. The search grid for the FB low and high cutoff frequencies, F_L and F_H, is defined $F_L \in \{0, 50, \ldots, 200\}$ Hz and $F_H \in \{3000, 3100, \ldots, 3400\}$ Hz, providing 25 possible combinations. For example, if the first coordinate is $F_L = 0$ Hz, $Warp\&Shift$ performs a transformation identical to VTLN, where increasing F_H above 3200 Hz will expand the FB and compress the spectrum, and decreasing F_H will compress the FB and expand the spectrum. Incrementing F_L and F_H with identical step sizes will translate the spectrum down the frequency axis. The remaining possible combinations of F_L and F_H will realize a combined warping and translation of spectra in frequency. The search grid is chosen to allow for higher formant translation in either direction and low formant translation down to -200 Hz, as it has been observed that the increase in F_1 usually does not exceed this rate [6].

When evaluating the frequency normalizations, no cepstral compensation is applied in order to observe separately their contribution to ASR performance (see Table V). The column denoted "*none*" represents a baseline setup employing 20 Bands-LPC (FB 0–4 kHz) and non-normalized models. With the exception of clean neutral set, omitting CVN increases WER (compare column "*none*" with columns *20 Bands-LPC* in Table I). Both variants of VTLN improve baseline performance on LE sets, and both slightly improve or preserve WER for neutral speech, optimized VTLN being more effective. $Warp\&Shift$ (denoted $W\&S$) provides superior WER reduction on LE speech while preserving baseline performance on neutral sets. For illustration, Fig. 8 displays histograms of utterance-dependent α's as selected by the fully optimized VTLN during decoding of neutral and LE sets. It can be seen that for LE, FB cutoffs are generally warped by $\alpha < 1$, which represents FB expansion and frequency compression of the spectra. This corresponds well with intuition that LE speech spectra should be transformed down in frequency in order

TABLE V
FREQUENCY NORMALIZATIONS—CLEAN DATA; WER (%)

		Frequency Normalization					
	Set	None	VTLN Lee-Rose	VTLN Optim.	W&S	Shift	Shift, QCN4
F	Neutral	3.8	3.3	3.0	3.5	3.6	2.9
	LE	32.4	23.4	23.2	17.3	14.1	11.9
M	Neutral	1.9	2.2	2.2	2.0	2.0	1.7
	LE	19.0	18.4	17.5	12.9	13.3	11.5

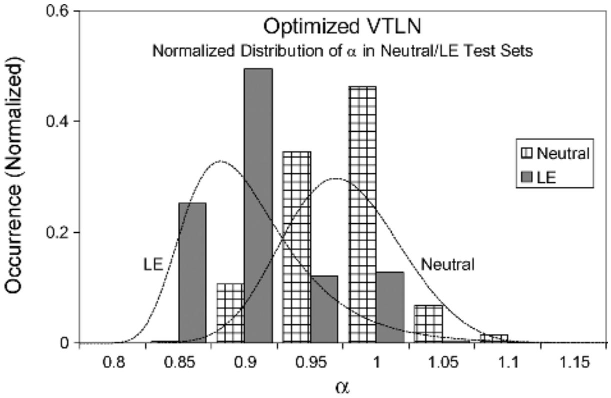

Fig. 8. Distribution of utterance-dependent α—females, neutral/LE clean test sets. Dashed line—GEV histogram interpolation.

TABLE VI
DISTRIBUTION OF W&S PARAMETERS; JOINT MALE AND FEMALE SETS

		F_H (Hz)									
Set	F_L (Hz)	3000		3100		3200		3300		3400	
W&S Train	0					2		57	30	6	33
	50							2	4		
W&S Test	0	3		1		4		5		3	
	50							9	3	1	9
	100						1		1		6
	150								6		

to compensate for upward shifts of low formants. Table VI presents the distribution of parameters in *Warp&Shift* normalization during HMM training and recognition. In the upper table part dedicated to the training stage, the brighter fields establish F_L, F_H coordinates seen during the first *Warp&Shift* retraining iteration, see the third step in Table II, and the darker fields those seen during the second iteration. The number in a shaded field represents the count of speakers for which a coordinate was prevalently observed across their sessions. It can be seen that in both *Warp&Shift* retraining iterations, a majority of data is normalized in a VTLN manner, since F_L was set to zero in most cases. This confirms that the main variability in neutral training data is due to inter-speaker VTL differences, and frequency warping as performed by VTLN represents a reasonable compensation to it.

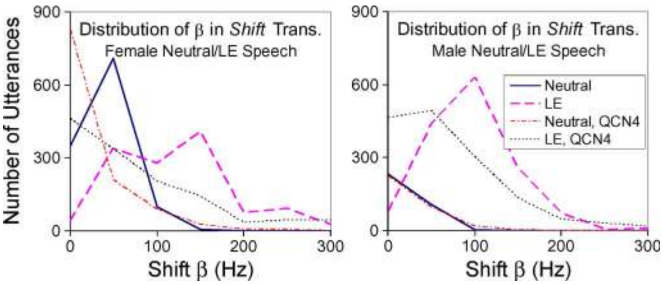

Fig. 9. Distribution of β in *Shift* decoding.

The bottom part of Table VI summarizes *Warp&Shift* decoding, with brighter fields denoting neutral sessions, and darker entries LE sessions. Again, neutral data are generally normalized as in VTLN, with an occasional slight shift of F_L. Alternatively, translation does play a more meaningful role in normalization of LE spectra.

In order to verify, whether translation of short time spectra alone is capable of suppressing formant shifts introduced by LE, the *Warp&Shift* transform [(3)] is reduced to a *Shift* transform

$$F_{\text{Shift}} = F + \beta. \tag{30}$$

The search grid is $\beta \in \{0, 50, \ldots, 300\}$ Hz, reducing the number of *Warp&Shift* choices from 25 to 7. F_{Shift} is implemented in a similar way to *Warp&Shift*. Since there is no need for spectral translation of neutral training data, *Shift* is applied only during recognition, utilizing non-normalized HMM's trained in 46 iterations. As shown in the penultimate column of Table V, *Shift* preserves baseline WERs on neutral speech and considerably improves performance for LE speech compared to the baseline and VTLN systems. *Shift* further reduces WER of *Warp&Shift* for female LE data at an affordable cost of a slight WER increase for male LE. Fig. 9 depicts the distribution of β when decoding neutral and LE sets (the trends with QCN_4 are discussed in the following section); on male neutral speech, β is exclusively set to zero and no translation is performed, while for male LE data β reaches maximum at 100 Hz. For female sets, there is a slight shift of 50 Hz dominating in the neutral set, and for LE speech, the majority of β's lie in the range 50–150 Hz. Both β distributions and performance in ASR tasks show the capability of *Shift* to compensate for formant shifts in LE while preserving performance for neutral speech. The frequent 50-Hz shift for female neutral speech is surprising; however, WER reduction compared to baseline confirms its relevance. In some level, similar observations were made in [42] and [6], where additional compression of spectra was demanded by VTLN even though models were trained on data similar to test. The shift here can be expected due to the effort of the *Shift* method to incorporate part of the spectrum that was originally out of the FB reach for some speakers.

Fig. 10 shows locations of the F_1, F_2 vowel space in neutral and LE samples, and for samples transformed by *Warp&Shift* and *Shift* processing, accompanied with example error ellipses covering 39.4% of the formant occurrences. The locations of neutral and LE formants were obtained by combining the output

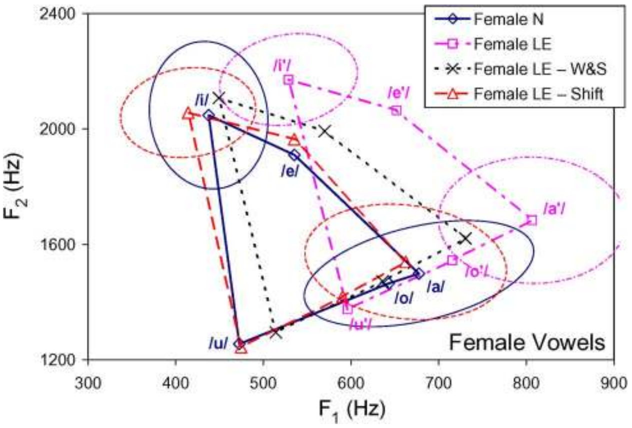

Fig. 10. F_1, F_2 normalization by $Warp\&Shift$ and $Shift$.

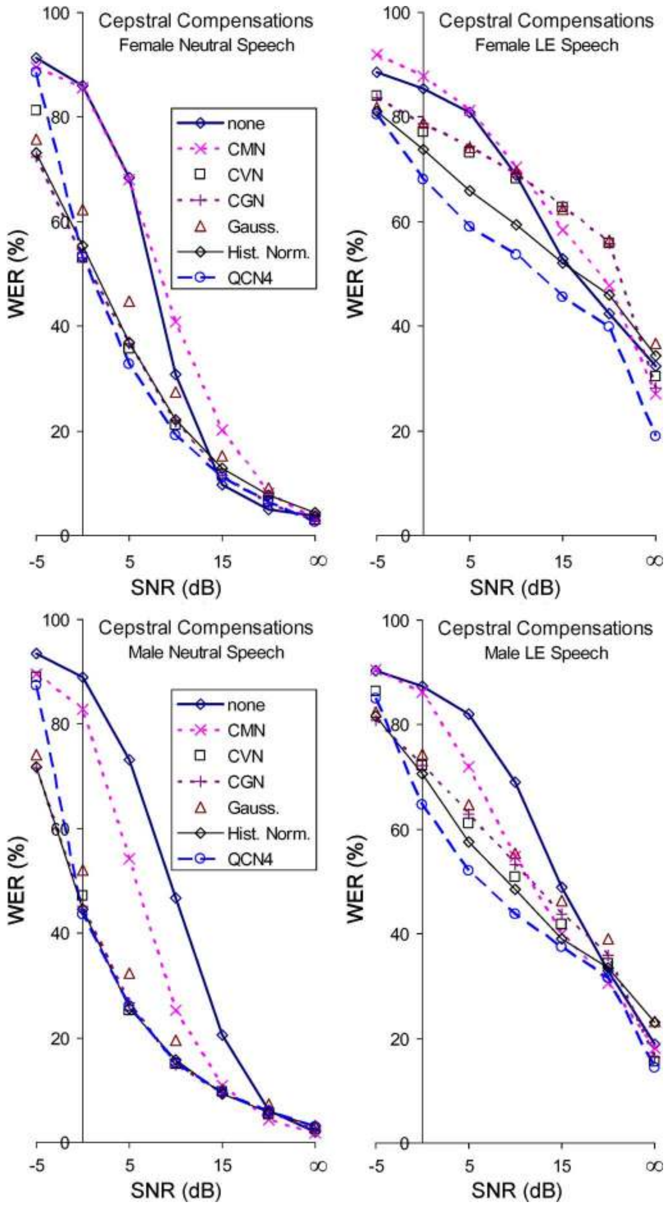

Fig. 11. Performance of cepstral compensations in noise.

of formant tracking and the phone boundaries estimates obtained from forced alignment. The formant locations of the normalized data were estimated by applying the transforms selected by $Warp\&Shift$ and $Shift$ to the formant frequencies of the corresponding LE utterances. It can be seen that both normalizations manage to transform formants towards the neutral locations, $Shift$ being more accurate. This confirms previous studies on direct formant location normalization for stress and LE [65].

E. Cepstral Normalizations

In this subsection, QCN is evaluated together with CMN, CVN, and CGN, and in addition compared to two variants of feature warping. Feature warping, also called histogram equalization, was independently proposed in [43] for ASR and in [66] for speaker verification. The goal of the method is to transform detailed structure of the cepstral distributions for the incoming data towards those captured in the acoustic model in order to decrease acoustic mismatch. Feature warping methods search for a function mapping the cdf's estimated from the incoming data towards cdf's of the target distributions. Our implementation of feature warping follows that in [66]. Here, each cepstral dimension is treated separately in the warping process. For each dimension, cepstral samples captured in the warping window of length N are sorted in descending order and ranked in a linear fashion. The maximum value is assigned a rank of 1 and the minimum value N. The ranking is used as an index in the lookup table with the target cdf to determine the corresponding warped value. In our implementation, the target cdf lookup table is resampled for each incoming utterance to match the utterance length, as well as the warp window length N. Hence, unlike in [66], where only the center sample in the sliding warp window is warped at a time, here, all window samples are transformed collectively. Two alternatives of feature warping are evaluated. In the setup denoted "*Gauss.*," the target distribution for all cepstral dimensions is chosen to be a normal distribution $\mathcal{N}(0,1)$. In the setup denoted "*Hist. Norm.*," the gender-dependent target distributions are chosen to be those seen in the training data (i.e., each cepstral dimension is represented by a unique, in general, non-Gaussian distribution). In both setups, feature warping is applied both to the training and test utterances.

In advance to the evaluation of the compensation schemes, an optimal choice of the dynamic range for QCN, represented by j [see (27)], is determined. The search range of $1, 2, \ldots, 15$ is chosen, where the upper limit defines the dynamic range covering 70% of samples (bound by quantiles q_{15} and q_{85}), which roughly corresponds to the interval of $[\mu - \sigma, \mu + \sigma]$ in a normal distribution. In applying QCN to a training set to obtain QCN-normalized HMMs, and a small subset of the overall test set (neutral/LE recordings from two male/two female speakers), it was observed that $j = 4$ (bounding 92% of cepstral samples) provided the most consistent WER reduction for both neutral/LE data in all noisy conditions. Therefore, this configuration was used in the following experiments (denoted $QCN4$).

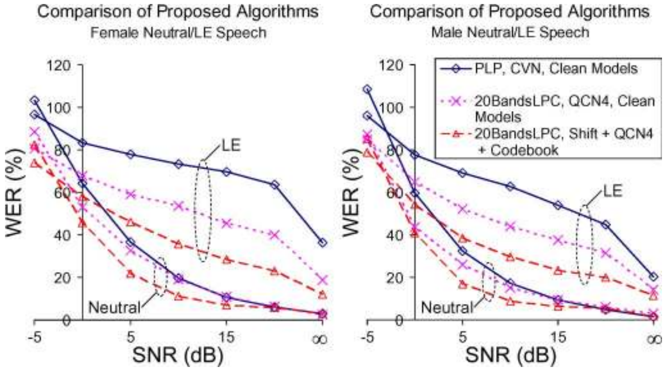

Fig. 12. Comparison of recognition performances—all noisy conditions.

TABLE VII
COMPARISON OF RECOGNITION PERFORMANCES; WER (%)

Set			System Setup			
			PLP, CVN	20BandsLPC, QCN4	20BandsLPC, QCN4, Codebook	20BandsLPC, Shift, QCN4, Codebook
Clean	F	Neutral	2.9	2.5	2.9	2.9
		LE	36.5	18.8	17.4	12.1
	M	Neutral	1.5	3.0	1.6	1.7
		LE	20.2	14.3	14.6	11.5
10dB SNR Car Noise	F	Neutral	19.8	19.1	10.8	11.1
		LE	73.4	53.6	41.5	35.7
	M	Neutral	17.3	15.3	7.7	8.6
		LE	62.7	43.8	31.3	29.8

The overall performance of the six cepstral compensation methods in noisy conditions is shown in Fig. 11; it is noted that here, no frequency compensations were employed. It can be seen that $QCN4$ reaches the best WERs down to 0-dB SNR for female neutral data (followed by *Hist. Norm.* and CGN), with comparable performance to *Hist. Norm.*, CVN, and CGN for male neutral data. $QCN4$ outperforms all other methods on female LE speech down to 0-dB SNR and on male LE speech from 15-dB to 0-dB SNR. In a broad range of SNRs, *Hist. Norm.* provides the second best performance for both female and male LE speech. Note that for high SNRs in neutral data, all compensations slightly increase WER compared to non-compensated features. While Gaussianization was shown to improve speaker verification [66], it is consistently outperformed by CVN in the present ASR experiments. The possible explanation is due to the essential difference in speaker verification versus ASR tasks, where the first focuses on the speaker-dependent vocal characteristics while the latter on phonetic content. Low cepstral coefficient distributions are typically multimodal, where each mode represents different phone class (or speech silence). Transforming these distributions into "perfect" Gaussians may cause phone classes to become less distinguishable. On the other hand, histogram normalization towards the training set distributions generally outperforms other conventional methods in the present experiments.

TABLE VIII
CODEBOOK MODEL ASSIGNMENT, 20 BANDS-LPC + SHIFT + QCN4; NUMBER OF UTTERANCES ASSIGNED TO MODEL SET

Test Set			Noisy Models						
Gender	Cond.	SNR (dB)	-5	0	5	10	15	20	∞ (dB)
F	Neutral	-5	433	703	24	0	0	0	0
		0	32	715	381	32	0	0	0
		5	0	88	607	412	51	2	0
		10	0	0	112	527	439	82	0
		15	0	0	3	87	580	485	5
		20	0	0	0	8	160	968	24
		∞	0	0	0	2	4	82	1072
F	LE	-5	117	806	318	19	0	0	0
		0	6	322	565	312	47	8	0
		5	1	74	296	508	296	85	0
		10	0	13	104	324	510	297	12
		15	0	3	32	147	429	616	33
		20	0	1	7	63	256	884	49
		∞	0	0	0	5	3	60	1192

The performance of joint $QCN4$ and $Shift$ on clean data is presented in the last column of Table V, showing that combining frequency and cepstral normalizations further improves WER. Applying both compensations together affects the β distributions in $Shift$, where the rate of frequency shifts is reduced as part of the compensation is already handled during the feature extraction stage by $QCN4$; see Fig. 9.

F. Codebook Recognizer

Finally, performance of an ASR system combining *20 Bands-LPC* and selected combinations of $Shift$, $QCN4$, and a codebook of noisy models is evaluated and compared to a traditional system utilizing a standard PLP front-end and CVN; see Fig. 12 and Table VII. It can be seen in Fig. 12 that for SNRs of 15 dB and lower on neutral data and for all LE data sets, the codebook recognizer provides superior performance to the prior PLP system and the *20 Bands-LPC* system with $QCN4$. Table VII details WERs on clean and 10-dB SNR data, also for a setup employing *20 Bands-LPC*, $QCN4$, and the noisy codebook. It can be seen that incorporating $Shift$ into the codebook system considerably improves performance on LE clean and noisy sets at a cost of a slight WER increase on neutral sets. The codebook scheme proves to be efficient in reducing the impact of noise on recognition performance and in some cases it also helps to reduce WER on clean speech (compare performances on clean male speech). This is due to the fact that clean environment recordings reach different SNRs and in some cases, noisy models may provide a better match than clean ones.

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 or close SNR (±5 dB) as appearing in the actual test set were selected for the decoding, proving the efficiency of the approach (see Table VIII).

VI. Conclusion

This paper has presented novel robust methods for suppressing the impact of Lombard effect and background noise in automatic speech recognition. Unsupervised frequency and cepstral domain normalizations were developed using a maximum-likelihood transformation of the short-time spectra and quantile-based cepstral dynamics normalization. The normalization parameters are estimated on-the-fly from the incoming speech signal, and all algorithms require no *a priori* knowledge about the level of noise and presence of Lombard effect. As a part of the design, a simple statistical model that reflects the impact of additive noise on cepstral distributions was introduced. The model provides an explanation of several effects of noise observed experimentally in previous studies.

The proposed algorithms were incorporated into an ASR engine employing a codebook of noisy acoustic models. The models were trained on clean neutral speech mixed with car noises at different levels. Evaluation tasks on noisy speech data sets showed that the proposed methods are efficient in compensating for the impact of both LE and noise, and outperform traditional normalizations.

References

[1] H. Bořil and J. H. L. Hansen, "Unsupervised equalization of Lombard effect for speech recognition in noisy adverse environment," in *Proc. ICASSP'09*, Taipei, Taiwan, Apr. 2009, pp. 3937–3940.

[2] E. Lombard, "Le signe de l'elevation de la voix," *Ann. Malad. Oreille, Larynx, Nez, Pharynx*, vol. 37, pp. 101–119, 1911.

[3] J. H. L. Hansen, "Analysis and compensation of stressed and noisy speech with application to robust automatic recognition," Ph.D. dissertation, Georgia Inst. of Technol., Atlanta, GA, 1988.

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

[5] 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.

[6] H. Bořil, "Robust speech recogniton: Analysis and equalization of Lombard effect in Czech corpora" Ph.D. dissertation, Czech Technical Univ., Prague, Czech Republic, 2008 [Online]. Available: http://www.utdallas.edu/~hxb076000

[7] Y. Lu and M. Cooke, "Speech production modifications produced by competing talkers, babble, and stationary noise," *J. Acoust. Soc. Amer.*, vol. 124, no. 5, pp. 3261–3275, Nov. 2008.

[8] J. J. Dreher and J. O'Neill, "Effects of ambient noise on speaker intelligibility for words and phrases," *J. Acoust. Soc. Amer.*, vol. 29, no. 12, pp. 1320–1323, 1957.

[9] J. C. Webster and R. G. Klumpp, "Effects of ambient noise and nearby talkers on a face-to-face communication task," *J. Acoust. Soc. Amer.*, vol. 34, no. 7, pp. 936–941, 1962.

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

[11] H. L. Pick, G. M. Siegel, P. W. Fox, S. R. Garber, and J. K. Kearney, "Inhibiting the Lombard effect," *J. Acoust. Soc. Amer.*, vol. 85, no. 2, pp. 894–900, 1989.

[12] J.-C. Junqua, S. Fincke, and K. Field, "Influence of the speaking style and the noise spectral tilt on the Lombard reflex and automatic speech recognition," in *Proc. ICSLP'98*, Sydney, Australia, 1998, pp. 467–470.

[13] J.-C. Junqua*, Sources of Variability and Distortion in the Communication Process. Robust Speech Recognition in Embedded Systems and PC Applications*. Dordrecht, The Netherlands: Springer, 2002, pp. 1–36.

[14] 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. Speech Audio Process.*, vol. 8, no. 4, pp. 429–442, Jul. 2000.

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

[16] R. Schulman, "Dynamic and perceptual constraints of loud speech," *J. Acoust. Soc. Amer.*, vol. 78, no. S1, pp. S37–S37, 1985.

[17] P. Gramming, S. Sundberg, S. Ternström, and W. Perkins, "Relationship between changes in voice pitch and loudness," *STL-QPSR*, vol. 28, no. 1, pp. 39–55, 1987.

[18] K. Cummings and M. Clements, "Analysis of glottal waveforms across stress styles," in *Proc. ICASSP'90*, Albuquerque, NM, 1990, vol. 1, pp. 369–372.

[19] D. Pisoni, R. Bernacki, H. Nusbaum, and M. Yuchtman, "Some acoustic-phonetic correlates of speech produced in noise," in *Proc. ICASSP'85*, Tampa, FL, 1985, vol. 10, pp. 1581–1584.

[20] W. V. Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow, and M. A. Stokes, "Effects of noise on speech production: Acoustic and perceptual analyses," *J. Acoust. Soc. Amer.*, vol. 84, no. 3, pp. 917–928, 1988.

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

[22] Z. Bond and T. Moore, "A note on loud and Lombard speech," in *Proc. ICSLP'90*, Kobe, Japan, 1990, pp. 969–972.

[23] J. Junqua and Y. Anglade, "Acoustic and perceptual studies of Lombard speech: Application to isolated-words automatic speech recognition," in *Proc. ICASSP'90*, Albuquerque, NM, 1990, vol. 2, pp. 841–844.

[24] Z. S. Bond, T. J. Moore, and B. Gable, "Acoustic—Phonetic characteristics of speech produced in noise and while wearing an oxygen mask," *J. Acoust. Soc. Amer.*, vol. 85, no. 2, pp. 907–912, 1989.

[25] J. Hansen and O. Bria, "Lombard effect compensation for robust automatic speech recognition in noise," in *Proc. ICSLP'90*, Kobe, Japan, 1990, pp. 1125–1128.

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

[27] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, no. 4, pp. 1738–1752, 1990.

[28] R. P. Lippmann, E. A. Martin, and D. B. Paul, "Multi-style training for robust isolated-word speech recognition," in *Proc. ICASSP'87*, Dallas, TX, 1987, pp. 705–708.

[29] R. W. Schafer and L. R. Rabiner, "System for automatic formant analysis of voiced speech," *J. Acoust. Soc. Amer.*, vol. 47, no. 2B, pp. 634–648, 1970.

[30] 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.

[31] B. A. Hanson and T. H. Applebaum, "Features for noise-robust speaker-independent word recognition," in *Proc. ICSLP'90*, Kobe, Japan, 1990, pp. 1117–1120.

[32] J. H. L. Hansen and M. A. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Trans. Signal Process.*, vol. 39, no. 4, pp. 795–805, Apr. 1991.

[33] Y. Takizawa and M. Hamada, "Lombard speech recognition by formant-frequency-shifted LPC cepstrum," in *Proc. ICSLP'90*, Kobe, Japan, 1990, pp. 293–296.

[34] T. Suzuki, K. Nakajima, and Y. Abe, "Isolated word recognition using models for acoustic phonetic variability by Lombard effect," in *Proc. ICSLP'94*, Yokohama, Japan, 1994, pp. 999–1002.

[35] D. B. Paul, "A speaker-stress resistant HMM isolated word recognizer," in *Proc. ICASSP'87*, Dallas, TX, 1987, pp. 713–716.

[36] B. D. Womack and J. Hansen, "N-channel hidden Markov models for combined stress speech classification and recognition," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 6, pp. 668–677, 1999.

[37] B. D. Womack and J. Hansen, "Classification of speech under stress using target driven features," *Speech Commun., Special Iss. Speech Under Stress*, vol. 20, no. 1–2, pp. 131–150, 1996.

[38] H. Bořil, P. Fousek, and H. Höge, "Two-stage system for robust neutral/Lombard speech recognition," in *Proc. Interspeech'07*, Antwerp, Belgium, 2007, pp. 1074–1077.

[39] B. Womack and J. Hansen, "Stress independent robust HMM speech recognition using neural network stress classification," in *Proc. Eurospeech'95*, Madrid, Spain, 1995, pp. 1999–2002.

[40] S. E. Bou-Ghazale and J. Hansen, "HMM-based stressed speech modeling with application to improved synthesis and recognition of isolated speech under stress," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 3, pp. 201–216, 1998.

[41] J. 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. Audio, Speech, Lang. Process.*, vol. 17, no. 2, pp. 366–378, Feb. 2009.

[42] L. Lee and R. Rose, "Speaker normalization using efficient frequency warping procedures," in *Proc. ICASSP'96*, Los Alamitos, CA, 1996, vol. 1, pp. 353–356, IEEE Comput. Soc..

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

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

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

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

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

[48] H. Wakita, "Normalization of vowels by vocal-tract length and its application to vowel identification," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 25, no. 2, pp. 183–192, Apr. 1977.

[49] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *Proc. ICASSP'96*, Los Alamitos, CA, 1996, vol. 1, pp. 346–348, IEEE Comput. Soc..

[50] P. Zhan and M. Westphal, "Speaker normalization based on frequency warping," in *Proc. ICASSP'97*, 1997, vol. 2, pp. 1039–1042.

[51] M. Pitz and H. Ney, "Vocal tract normalization equals linear transformation in cepstral space," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 13, no. 5, pp. 930–944, Sep. 2005.

[52] J. P. Openshaw and J. S. Mason, "Optimal noise-masking of cepstral features for robust speaker identification," in *Proc. ESCA Workshop Automatic Speaker Recog., Identification, and Verification—ASRIV-1994*, 1994, vol. 1, pp. 231–234.

[53] F. de Wet, J. de Veth, L. Boves, and B. Cranen, "Additive background noise as a source of non-linear mismatch in the cepstral and log-energy domain," *Comput. Speech Lang.*, vol. 19, no. 1, pp. 31–54, Jan. 2005.

[54] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, 1979.

[55] J. A. Rice, *Mathematical Statistics and Data Analysis*. Belmont, CA: Duxbury, 1995.

[56] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[57] K. Yao, K. K. Paliwal, and S. Nakamura, "Noise adaptive speech recognition based on sequential noise parameter estimation," *Speech Commun.*, vol. 42, no. 1, pp. 5–23, 2004.

[58] C. E. Mokbel and G. Chollet, "Automatic word recognition in cars," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 5, pp. 346–356, Sep. 1995.

[59] A. P. Varga and R. K. Moore, "Hidden Markov model decomposition of speech and noise," in *Proc. ICASSP'90*, Albuquerque, NM, 1990, vol. 21, pp. 845–848.

[60] M. Gales and S. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 5, pp. 352–359, Sep. 1996.

[61] P. Pollák, J. Vopička, and P. Sovka, "Czech language database of car speech and environmental noise," in *Proc. Eurospeech'99*, Budapest, Hungary, 1999, pp. 2263–2266.

[62] H. Bořil, T. Bořil, and P. Pollák, "Methodology of Lombard speech database acquisition: Experiences with CLSD," in *Proc. LREC 2006—5th Conf. Lang. Res. Eval.*, Genova, Italy, 2006, pp. 1644–1647.

[63] 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. LREC'2002*, Las Palmas, Spain, 2002.

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

[65] J. H. L. Hansen and M. Clements, "Stress compensation and noise reduction algorithms for robust speech recognition," in *Proc. ICASSP'89*, Glasgow, U.K., 1989, pp. 266–269.

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

Hynek Bořil (S'08–M'09) was born in Most, Czech Republic. He received the M.S. degree in electrical engineering and Ph.D. degree in electrical engineering and information technology from the Department of Electrical Engineering, Czech Technical University, Prague, in 2003 and 2008, respectively.

In August 2007, he joined the Center for Robust Speech Systems, Eric Jonsson School of Engineering and Computer Science, University of Texas at Dallas, Richardson, as a Research Associate. His research interests are focused in the field of speech processing, including analysis and normalization of Lombard effect, noise, and channel variations for robust automatic speech recognition, stress classification, prosody modeling, and speech production development in infants.

John H. L. Hansen (S'81–M'82–SM'93–F'07) received the B.S.E.E. degree from the College of Engineering, Rutgers University, New Brunswick, NJ, in 1982 and the M.S. and Ph.D. degrees in electrical engineering from the Georgia Institute of Technology, Atlanta, in 1988 and 1983, respectively.

He joined the Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas (UTD), Richardson, in the fall of 2005, where he is a Professor and Department Head of Electrical Engineering, and holds the Distinguished University Chair in Telecommunications Engineering. He also holds a joint appointment as a Professor in the School of Brain and Behavioral Sciences (Speech and Hearing). At UTD, he established the Center for Robust Speech Systems (CRSS) which is part of the Human Language Technology Research Institute. Previously, he served as Department Chairman and Professor in the Department of Speech, Language, and Hearing Sciences (SLHS), and Professor in the Department of Electrical and Computer Engineering, University of Colorado Boulder (1998–2005), where he cofounded the Center for Spoken Language Research. In 1988, he established the Robust Speech Processing Laboratory (RSPL) and continues to direct research activities in CRSS at UTD. His research interests span the areas of digital speech processing, analysis and modeling of speech and speaker traits, speech enhancement, feature estimation in noise, robust speech recognition with emphasis on spoken document retrieval, and in-vehicle interactive systems for hands-free human-computer interaction. He has supervised 51 (22 Ph.D., 29 M.S.) thesis candidates, is author/coauthor of 328 journal and conference papers in the field of speech processing and communications, coauthor of the textbook *Discrete-Time Processing of Speech Signals*, (IEEE Press, 2000), coeditor of *DSP for In-Vehicle and Mobile Systems* (Springer, 2004), *Advances for In-Vehicle and Mobile Systems: Challenges for International Standards* (Springer, 2006), *In-Vehicle Corpus and Signal Processing for Driver Behavior Modeling* (Springer, 2008), and lead author of the report "The impact of speech under 'stress' on military speech technology," (NATO RTO-TR-10, 2000).

Prof. Hansen was a recipient of the 2005 University of Colorado Teacher Recognition Award as voted by the student body. He also organized and served as General Chair for ICSLP/Interspeech-2002: International Conference on Spoken Language Processing, September 16–20, 2002, and is serving as Co-Organizer and Technical Program Chair for IEEE ICASSP-2010, Dallas, TX. In 2007, he was named IEEE Fellow for contributions in "Robust Speech Recognition in Stress and Noise," and is currently serving as Member of the IEEE Signal Processing Society Speech Technical Committee and Educational Technical Committee. Previously, he has served as Technical Advisor to U.S. Delegate for NATO (IST/TG-01), IEEE Signal Processing Society Distinguished Lecturer (2005–2006), Associate Editor for IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (1992–1999), Associate Editor for the IEEE SIGNAL PROCESSING LETTERS (1998–2000), and Editorial Board Member for the *IEEE Signal Processing Magazine* (2001–2003). He has also served as a Guest Editor of the October 1994 special issue on Robust Speech Recognition for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING. He has served on the Speech Communications Technical Committee for the Acoustical Society of America (2000–2003), and is serving as a member of the ISCA (International Speech Communications Association) Advisory Council.