Robust Log-Energy Estimation and its Dynamic Change Enhancement for In-car Speech Recognition

Weifeng Li, Longbiao Wang, Yicong Zhou, Herv Bourlard, and Qingmin Liao

***Abstract*—The log-energy parameter, typically derived from a full-band spectrum, is a critical feature commonly used in automatic speech recognition (ASR) systems. However, log-energy is difficult to estimate reliably in the presence of background noise. In this paper, we theoretically show that background noise affects the trajectories of not only the "conventional" log-energy, but also its delta parameters. This results in a poor estimation of the actual log-energy and its delta parameters, which no longer describe the speech signal. We thus propose a new method to estimate log-energy from a sub-band spectrum, followed by dynamic change enhancement and mean smoothing. We demonstrate the effectiveness of the proposed log-energy estimation and its post-processing steps through speech recognition experiments conducted on the in-car CENSREC-2 database. The proposed log-energy (together with its corresponding delta parameters) yields an average improvement of 32.8% compared with the baseline front-ends. Moreover, it is also shown that further improvement can be achieved by incorporating the new Mel-Frequency Cepstral Coefficients (MFCCs) obtained by non-linear spectral contrast stretching.**

***Index Terms*—Dynamic change enhancement, in-car speech recognition, log-energy, mel-filterbank (MFB), mel-frequency cepstral coefficients (MFCCs).**

I. Introduction

It is generally accepted that the human auditory system is sensitive to changes in speech inputs over time [2], and a certain degree of spectral contrast is necessary for robust speech recognition [3]. Multiple findings from auditory perception experiments also provide evidence that processes of successive spectral contrast can disambiguate co-articulated speech [4]. In [5], it was shown that eliminating the natural time-varying spectral changes over the duration of a vowel resulted in much lower recognition accuracy for American English vowels.

Under adverse conditions, background noise generally leads to a reduction in dynamic changes in speech signals. For even normal hearing listeners, serious reductions in dynamic changes lead to unreliable segmentation, making the task of parsing the speech signal more difficult [6]. However, it has been suggested that under adverse conditions the auditory system makes some adaptations serving to emphasize newly arriving components of the signal and enhance the regions of the signal undergoing spectro-temporal changes [7]. On the other hand, there is strong evidence that explicitly enhancing the dynamic change helps in the recognition of a speech signal [8]. In [3], it was shown that significantly higher scores were obtained with vowels enhanced to 6 dB of spectral contrast.

For Automatic Speech Recognition (ASR), widely used front-ends, like Mel-Frequency Cepstral Coefficients (MFCCs) [9] and Perceptual Linear Prediction (PLP) [10]), are extracted from short-time spectral energies in a compressed domain. For example, standard MFCCs are extracted from log scaled mel-filterbank (MFB) outputs. However, in the presence of background noise, the dynamic changes in spectral energies are generally reduced. Fig. 1 (the second row) shows the first-channel log MFB trajectory (or contour) of speech captured by a close-talking (headset) microphone and a distant microphone (attached to the ceiling above the driver's seat [1]) in a car-driving environment. Compared to close-talking speech, the floor level of the log MFB trajectory for distant speech is elevated and the valleys are buried by noise energy. While spectral changes in close-talking speech over time are rather apparent, they become obscure for distant speech owing to the noise effects. Besides MFCCs, the short-time log-energy and its temporal derivatives are often adopted as standard features as well. According to discriminant analysis of the features used for ASR [11], the frame log-energy and its temporal derivatives appear to be the most critical features in terms of recognition accuracy. It has been shown that ASR performance in clean conditions improves when the short-time log-energy and its temporal derivatives are used [12]. However, in low signal-to-noise ratio (SNR) conditions, the trajectory of the short-time log-energy, which is derived from a full-band spectrum, can be distorted and fails to describe the speech signal dynamics, as demonstrated in Fig. 1 (the lower part). Therefore, in the presence of the background noise, conventional MFCCs and log-energy usually introduce undesirable mismatches between relatively clean speech (used for training) and noisy speech (used for testing), resulting in a serious ASR performance drop. Motivated by the adaptation capabilities of

Manuscript received August 02, 2012; revised December 21, 2012; accepted April 14, 2013. Date of publication April 25, 2013; date of current version May 08, 2013. This work was supported in part by Shenzhen Basic Research Grant JCYJ20120831165730913, in part by the Macau Science and Technology Development Fund under Grant 017/2012/A1, and in part by the Research Committee at University of Macau under Grants SRG007-FST12-ZYC, MYRG113(Y1-L3)-FST12-ZYC, and MRG001/ZYC/2013/FST. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Murat Saraclar.

W. Li and Q. Liao are with the Shenzhen Key Laboratory of Information Science and Technology, Department of Electronic Engineering/Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China (e-mail: li.weifeng@sz.tsinghua.edu.cn; liaoqm@sz.tsinghua.edu.cn).

L. Wang is with the Nagaoka University of Technology, Nagaoka 940-2188, Japan (e-mail: wang@vos.nagaokaut.ac.jp).

Y. Zhou is with the Department of Computer and Information Science, University of Macau, Macau, China (e-mail: yicongzhou@umac.mo).

H. Bourlard is with the Idiap Research Institute, and Ecole Polytechnique Fédérale, Lausanne (EPFL), 1015 Lausanne, Switzerland (e-mail: herve.bourlard@idiap.ch).

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.2013.2260151

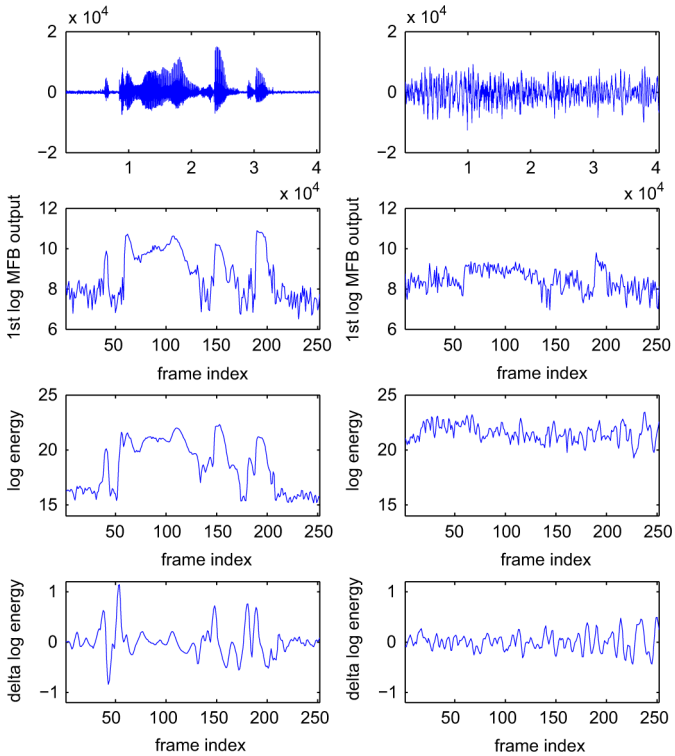


Fig. 1. Effects of car noise on log mel-filter bank (MFB) and log energy trajectories. The left subfigures (up to down): waveform, the first log MFB output, log energy, and the delta log energy of close-talking speech; The right subfigures (up to down): waveform, the first log MFB output, log energy, and the delta log energy of distant speech; The speech is "12439" in Japanese [1].

the auditory system described above, in [13] we proposed a new MFCC front-end based on the spectral contrast stretching of the log MFB outputs.

In this paper, we address the problem of estimating robust log-energy features in the presence of (car) background noise. More specifically, we analyze theoretically how the noise affects the trajectories of the conventional log-energy and its delta parameters, causing them no longer to describe the variations in the speech, or even to reduce speech performance in low SNR conditions. We then propose an estimation of the log-energy from a sub-band spectrum for better representation of the variations in the speech and enhancing its discriminative power for speech recognition. Noise reduction and dynamic change enhancement (DCE) are then applied to suppress the stationary noise components and boost the non-stationary speech segments, respectively. Finally a mean filter based smoothing is performed to eliminate the spike noise and processing artifacts. Our experiments, conducted on realistic in-car data under different training and test conditions, demonstrate that, with the subsequent post-processing, the proposed log-energy and its temporal derivatives are capable of significantly reducing the mismatch between the training and test conditions, yielding much higher ASR performance.

The organization of this paper is as follows: In Section II we theoretically analyze the resulting mismatch between clean and noisy conditions in the conventional log-energy estimation. We then propose a new log-energy estimation scheme in Section III, where the subsequent post-processing is also described. Section IV describes the MFCC front-end proposed in our previous work [13]. Section V gives a detailed presentation of our experimental evaluations on realistic in-car data under different training and test conditions. Finally, in Section VI we draw our conclusions.

II. Log-energy Estimation of Mismatch Between Clean and Noisy Conditions

Let $s(i)$, $n(i)$ and $x(i)$, respectively, denote the clean speech, additive noise, and observed noisy speech signals. The distortion of noisy speech can be expressed as

$$x(i) = s(i) + n(i). \tag{1}$$

The energy of noisy speech at the l-th frame is computed by

$$e_x(l) = \sum_{i=1}^{I} x_w^2(i) \simeq \sum_{i=1}^{I} s_w^2(i) + \sum_{i=1}^{I} n_w^2(i), \tag{2}$$

where $\{x_w(i), i = 1, \ldots, I\}$ are the Hamming windowed noisy speech signal samples and I is the size of the window (likewise for the clean speech and noise signals). Here we make the assumption of statistical independence between the clean speech and noise.

The log-energy of the noisy speech x can be formulated as

$$E_x(l) = \log e_x(l) = \log\left(e_s(l) + e_n(l)\right), \tag{3}$$

where

$$e_s(l) = \sum_{i=1}^{I} s_w^2(i), \tag{4}$$

and

$$e_n(l) = \sum_{i=1}^{I} n_w^2(i). \tag{5}$$

The dynamic changes in log-energy C_{E_x} can be computed as the difference between the log-energies of noisy speech at frame l and the subsequent one (e.g., at frame $l + k, k > 0$).

$$\begin{aligned} C_{E_x} &= E_x(l+k) - E_x(l) \\ &= \log[e_s(l+k) + e_n(l+k)] \\ &\quad - \log[e_s(l) + e_n(l)] \\ &= \log \frac{e_s(l+k) + e_n(l+k)}{e_s(l) + e_n(l)} \qquad (6) \\ &\simeq \log\left(1 + \frac{e_s(l+k) - e_s(l)}{e_s(l) + e_n(l)}\right), \qquad (7) \end{aligned}$$

where the approximation is based on the (reasonable) assumption that the noise energy does not vary too much over time (i.e., $e_n(l+k) \simeq e_n(l)$).

From (7) we can see that when there is noise, the dynamic change in log-energy decreases, and it becomes even smaller as

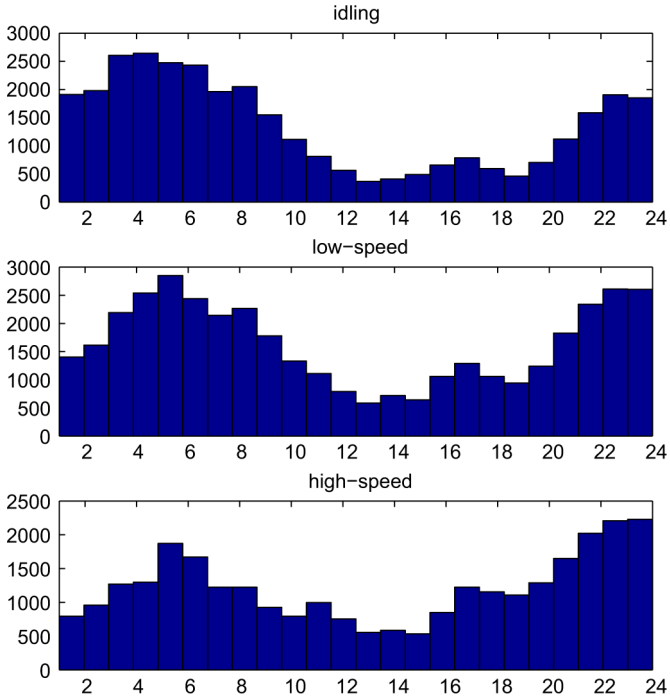


Fig. 2. Histograms of the selected 10 mel-filter banks for different driving conditions. The number of bins used for computing the histogram is 24, the same as total number of the mel-filter banks. The horizontal axis represents the filter-bank index and the vertical axis denotes the frequencies for each filter-bank. Idling: the vehicle is stopped; low-speed: driving on a city street; high-speed: driving on an expressway.

the noise energy increases. When $e_s(l) = 0$ (i.e., non-speech segments) and $e_s(l+k) > 0$ (i.e., speech segments), according to (7) the dynamic change in log-energy from non-speech segments to speech segments reduces to

$$\log\left(1+\frac{e_s(l+k)}{e_n(l)}\right) \simeq \log\left(1+SNR(l+k)\right), \qquad (8)$$

where

$$SNR(l+k) = \frac{e_s(l+k)}{e_n(l+k)} \qquad (9)$$

indicates the SNR at frame $l+k$. In the case of transition from speech (at frame l) to non-speech (at frame $l+k$) segments, (6) reduces to

$$\log\left(\frac{e_n(l+k)}{e_s(l)+e_n(l)}\right) \simeq -\log\left(1+SNR(l)\right). \qquad (10)$$

These equations illustrate that the presence of noise reduces the dynamic changes as a function of the SNR. The effects of the decrease in dynamic change are clearly demonstrated in Figs. 3-1, 3-5 and 3-7. When the noise is dominant (i.e., $e_n \gg e_s$), (6) reduces to $\log(E_n(l+k)/E_n(l))$. In this case, dynamic changes in the noisy speech signals over time reveal dynamic changes in the noise rather than those in the speech. Fig. 3-1 illustrates this phenomenon, especially for the first and last 50 frames.

In summary, in the presence of background noise the conventional static log-energy and its dynamic features (i.e., the delta and acceleration log-energy features) no longer reflect the variations in the speech signal very well. If input into an ASR system, they will produce a mismatch between relatively clean speech and noisy speech, which will inevitably degrade the ASR performance.

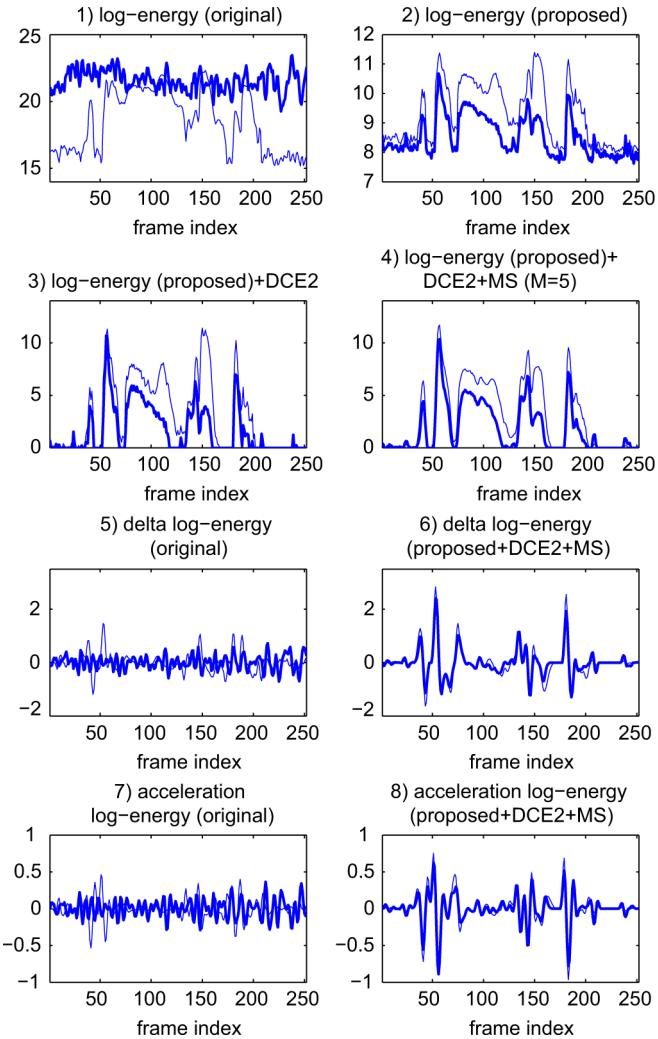


Fig. 3. The original and proposed log energy, delta log energy, acceleration log energy trajectories of the close-talking speech and distant speech (The speech is the same as Fig. 1). Inside each sub-figure, the thin line is for the close-talking speech and the bold line is for the distant speech. DCE2: dynamic change enhancement using (15). $\mathrm{MS(M=5)}$: the five-order mean smoothing.

III. Proposed Log-Energy and Subsequent Post-Processing Steps

A. The Proposed Log-Energy

From Parseval's theorem [14] we have

$$e_x(l) = \sum_{i=1}^{I} x_w^2(i) = \frac{1}{I}\sum_{k=1}^{I} |X_w(k)|^2, \qquad (11)$$

where $X_w(k)$ is the discrete Fourier transform of $x_w(i)$, both of which have a length of I.

Equation (11) implies that the conventional log-energy is derived from the full-band spectrum. To alleviate this problem and make the log-energy better suited to reflect variations in the speech over time, we propose estimating log-energy from a sub-band perspective. More specifically, we estimate it from the log MFB outputs with the following two considerations: (1) log MFB outputs are sub-band based, and can capture dynamic

TABLE I
TRAINING AND TEST CONFIGURATIONS FOR EACH OF THE FOUR EVALUATION CONDITIONS. HF: HANDS-FREE MICROPHONE; CT: CLOSE-TALKING MICROPHONE. IDLING: THE VEHICLE IS STOPPED; LOW-SPEED: DRIVING ON A CITY STREET; HIGH-SPEED: DRIVING ON AN EXPRESSWAY

conditions	Cond.1		Cond.2		Cond.3		Cond.4	
train/test	train	test	train	test	train	test	train	test
microphone	HF	HF	HF	HF	CT	HF	CT	HF
idling	○	○	○		○		○	
low-speed	○	○		○	○	○		○
high-speed	○	○		○	○	○		○

TABLE II
RECOGNITION ACCURACIES (IN PERCENTAGES) FOR DIFFERENT METHODS. THE UPPER PART PRESENTS THE RECOGNITION PERFORMANCE OF USING (OR WITHOUT) THE ORIGINAL LOG-ENERGY, USING THE ZERO-ORDER MFCC, AND USING THE PROPOSED LOG-ENERGY PARAMETER; THE MIDDLE PART PRESENTS THE RECOGNITION PERFORMANCE OF USING THE PROPOSED LOG-ENERGY WITH THE SUBSEQUENT DCE; THE LOWER PART PRESENTS THE RECOGNITION PERFORMANCE OF ADOPTING MEAN SMOOTHING AFTER " ProposedE + DCE2 ." AVE.: AVERAGED RECOGNITION ACCURACIES OVER THE FOUR CONDITIONS

	Cond.1	Cond.2	Cond.3	Cond.4	Ave.
baseline	81.23	66.85	57.94	43.85	62.46
NE	80.79	68.04	60.47	44.12	63.35
MFCC0	82.34	66.99	59.53	48.87	64.43
proposedE	83.06	68.12	62.98	50.39	**66.16**
Eorg+NORM	82.37	70.16	59.34	47.15	64.76
Eorg+MVN	83.05	71.72	62.63	51.71	67.28
proposedE+DCE1	84.79	81.13	70.91	56.26	73.27
proposedE+DCE2	84.89	81.38	71.50	57.77	**73.88**
RASTA	85.04	80.78	71.72	58.48	74.01
$M = 3$	84.61	80.84	71.74	59.20	74.09
$M = 5$	84.26	81.52	71.80	61.46	**74.76**
$M = 7$	84.30	81.23	72.62	60.33	74.62
$M = 9$	83.66	81.13	70.41	59.00	73.55

variations in the speech signals over time within a particular sub-band; (2) log MFB outputs with wider change ranges across time can better reflect dynamic variations in the speech signals than those with smaller ones. Therefore, we propose calculating the log-energy by *averaging the J log MFB outputs with the largest relative changes*[1]. Here the relative[2] change in their log MFB values for the j-th filter bank is defined by

$$R(j) = \frac{X_{\max}^{(L)}(j) - X_N^{(L)}(j)}{X_N^{(L)}(j)}, \tag{12}$$

where $X_{\max}^{(L)}(j)$ and $X_N^{(L)}(j)$ are the maximum values of the j-th log MFB outputs along the frames of the utterance and the estimated noise log MFB value, respectively. $X_N^{(L)}(j)$ can be obtained by averaging the j-th log MFB outputs over the first several non-speech frames[3].

While the conventional log-energy is derived from the full-band spectrum, the proposed log-energy is sub-band-based. This technique can be viewed as a kind of "missing feature theory" [15], which rejects unreliable log MFB outputs. In our case the missing feature masks (or confidence measures) are based on (12). Fig. 2 shows histograms of the selected ten MFBs for different driving conditions (idling, low-speed, and high-speed) when using the distant microphone [1]. It can be observed that when the car-speed increases, the higher-order MFBs are more likely to be selected (i.e., more reliable than the lower order ones). From Fig. 3-2 we can see that compared with the conventional log-energy the proposed one better reflects variations in the speech over time, while the mismatch between the close-talking speech and distant speech is significantly reduced. Since the proposed log-energy, denoted by $E(l)$, is still noisy, it is further post-processed by noise subtraction, dynamic change enhancement (DCE), and mean smoothing, as described below.

B. Dynamic Change Enhancement

By averaging over the first several non-speech frames, the noise log-energy can be estimated. The noise-subtracted log-energy can then be obtained by:

$$u(l) = \begin{cases} E(l) - E_n, & \text{if } E(l) \geq E_n, \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

where E_n denotes the estimated log-energy of the noise.

By using this noise subtraction technique, the noise is reduced, while variations in the speech signal are preserved. Let $E_{\max}$ denote the maximum value of the proposed log-energies along the frames in an utterance. The DCE is implemented by

$$\breve{E}(l) = \frac{u(l)}{E_{\max} - E_n} \cdot E_{\max}, \tag{14}$$

or

$$\breve{E}(l) = \frac{u(l)}{E_{\max} - E_n} \cdot E(l). \tag{15}$$

Through this operation, the range of the dynamic change over the speech segments is stretched from $[0, E_{\max} - E_n]$ to $[0, E_{\max}]$ linearly ((14): DCE1) or non-linearly ((15): DCE2)[4], and the level of spectral variations in the speech is enhanced accordingly. As shown in Fig. 4, DCE2 emphasizes the speech variations with larger log MFB values more than those with smaller values, while DCE1 enhances them uniformly. Fig. 3-3 illustrates the DCE using (15).

C. Mean Smoothing

To reduce the high-frequency components mainly involving spike noise and the processing artifacts, $\breve{E}(l)$ is further processed by a mean filter, defined as

$$\hat{E}(l) = \frac{1}{M} \sum_{p=-(M-1)/2}^{(M-1)/2} \breve{E}(l+p). \tag{17}$$

[1]The zero-order MFCC can be viewed as an average of all the M log MFB outputs ($M = 24$ in our speech recognition experiments), while the proposed log-energy is the mean value of the J log MFB outputs only with prominent relative changes. (After exploring different values, we set $J = 10$ in our speech recognition experiments.) Considering that noise contaminates some log MFB outputs more than others, the proposed log-energy is expected to better reflect the variations in speech than the zero-order MFCC, which is confirmed in Table II.

[2]Here, selecting the largest "*relative*" log MFB outputs is based on the consideration that the log MFB outputs of a particular filter bank may have greater energy than those of other filter banks.

[3]In this paper the first 15 frames are used to estimate $X_N^{(L)}(j)$ and E_n in our experiments.

[4]If $E(l) > E_n$, (15) can be written as

$$\breve{E}(l) = \frac{(E(l))^2 - E_n \cdot E(l)}{E_{\max} - E_n}, \tag{16}$$

and thus $\breve{E}(l)$ has a quadratic form of the origin $E(l)$.

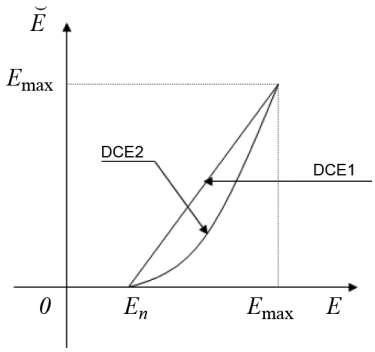


Fig. 4. The operations of dynamic change enhancement (DCE) DCE1—linear transformation using (14); DCE2—non-linear transformation using (15).

In summary, our proposed log-energy estimation and subsequent post-processing include three components: Section III-A estimates the log-energy from more reliable sub-bands, but the noise still remains and the dynamics of speech appear to be not so prominent, as shown in Fig. 3-2; Section III-B reduces the remaining noise and enhances the dynamic changes, as shown in Fig. 3-3; and Section III-C removes the high-frequency components consisting of the spike noise and processing artifacts created by the DCE operations.

Figs. 3-4, 3-6 and 3-8 show the resulting static, delta, and acceleration log-energies, respectively. Compared with the original features in Fig. 3-1, 3-5 and 3-7, it is clear that mismatches between the close-talking speech and distant speech have been reduced significantly.

Fig. 5 shows the estimated probability density functions (PDFs) of the resulting static, delta, and acceleration log-energies using realistic in-car data [1]. As shown in Fig. 5-1, in car-driving environments the histogram of the conventional log-energy for close-talking speech is inherently multi-modal, while the one for distant speech is more Gaussian owing to the noise effect. The differences between the two PDFs are significantly reduced by using the proposed log-energy and post-processing, as shown in Fig. 5-2. The PDFs of the delta and acceleration log-energy for close-talking speech have more peaks than those for distant speech (Figs. 5-3 and 5-5), while the use of the proposed log-energy and post-processing reduces the mismatches between them, causing them almost to overlap (Figs. 5-4 and 5-6).

IV. Proposed MFCC Front-End

In the presence of background noise, the floor level of the log MFB trajectories of distant speech may be elevated while the valleys are buried by the noise energy, as shown in Fig. 1. In this case, compared with clean speech the dynamic changes in the log MFB outputs are reduced[5]. If the derived MFCC front-ends are input into an ASR system, they will produce a mismatch between relatively clean speech (for training) and noisy speech (for testing). In [13], we proposed a new MFCC front-end by

[5]The mathematical analysis is similar to that in Section II.

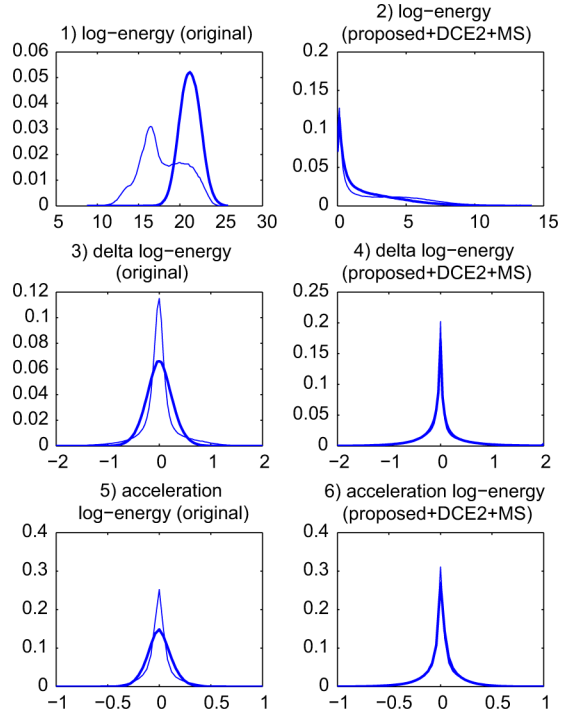


Fig. 5. Estimated probability density functions (PDFs) of the original and proposed (with subsequent post-processings) log energy, delta log energy, acceleration log energy. Inside each sub-figure, the thin line is for the close-talking speech and the bold line is for the distant speech. DCE2: dynamic change enhancement using (15). MS: mean smoothing with the order $M = 5$.

enhancing the dynamic changes in the log MFB outputs with a subsequent two-dimensional smoothing, similar to the process discussed in Sections III-B and III-C.

Let $X^{(L)}(j,l)$ denote the log MFB output at the j-th filter bank channel and the l-th frame. Let $X^{(L)}_{\max}(j)$ and $X^{(L)}_{N}(j)$ denote the maximum value in the utterance and the estimated one for noise, respectively. We enhance the dynamic changes in $X^{(L)}(j,l)$ via

$$\breve{X}^{(L)}(j,l) = \frac{U(X^{(L)}(j,l) - X^{(L)}_{N}(j))}{X^{(L)}_{\max}(j) - X^{(L)}_{N}(j)} \cdot X^{(L)}(j,l), \quad (18)$$

where $U(\cdot)$ is the step function:

$$U(v) = \begin{cases} v, & \text{if } v \geq 0\,, \\ 0, & \text{otherwise.} \end{cases} \quad (19)$$

Then, a two-dimensional mean filter is applied to remove the undesirable high-frequency components and processing artifacts, i.e.,

$$\hat{X}^{(L)}(j,l) = \frac{1}{MN} \sum_{(m,n)\in R} \breve{X}^{(L)}(m,n), \quad (20)$$

where R denotes the $M \times N$ window and $\breve{X}^{(L)}(m,n)$ denotes the neighbors around (j,l).

In our experiments a 3×3-windowed two-dimensional mean filter, with weights of the form

$$w(k,l) = \frac{1}{9}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}, \quad (21)$$

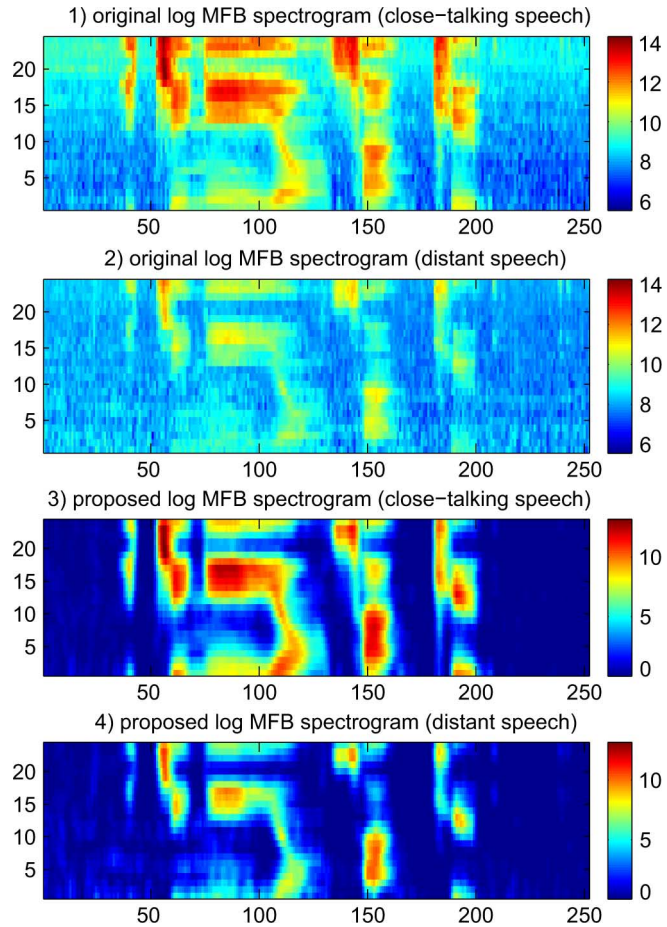


Fig. 6. The proposed log MFB spectrogram. The speech is the same as that in Fig. 1. The horizontal axis represents the frame number and the vertical axis denotes the log MFB index.

was adopted.

The proposed MFCCs are finally calculated from $\hat{X}^{(L)}(j,l)$ using the discrete cosine transform (DCT)

$$Y(i,l) = \sqrt{\frac{2}{J}} \sum_{j=1}^{J} \hat{X}^{(L)}(j,l) \cos\left(\frac{\pi i}{J}(j-0.5)\right) \qquad (22)$$

where i and J denote the MFCC index and the total number of filter bank channels, respectively.

Fig. 6 shows an example of the proposed log MFB spectrograms composed of $\hat{X}^{(L)}(j,l)$ in (20) (The speech is the same as that in Fig. 1). It is observed that the enhancement of changes in log MFB outputs is effective in reducing the mismatch between close-talking and distant speech. The first three derived MFCC trajectories for the original and compensated versions are plotted in Fig. 7. As shown in this figure, the original versions yield remarkable mismatches between the close-talking and distant speech in the first and third MFCC trajectories. By using the proposed MFCCs, however, the mismatches are reduced and the speech variations become more pronounced. As for the second MFCC trajectories, which almost match, the proposed method retains the matched property. Note that the dynamic ranges of the compensated MFCC trajectories become larger owing to the enhancement of dynamic changes in the log MFB domain.

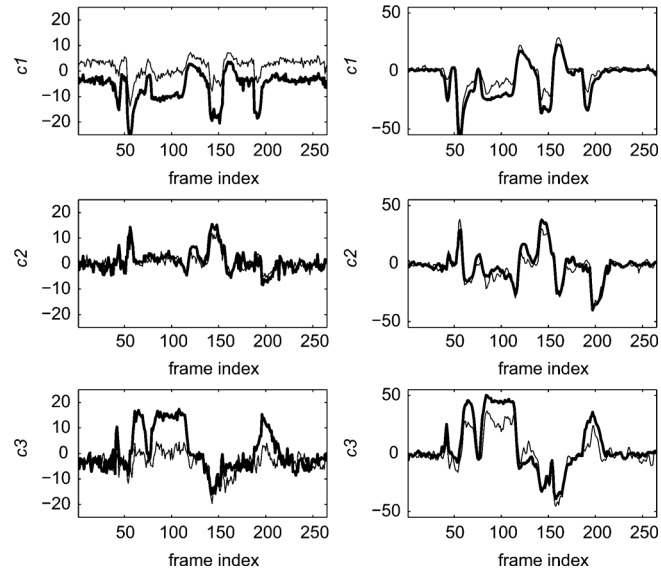


Fig. 7. The first three MFCC trajectories of the close-talking speech and distant speech without and with compensation (The speech is the same as that in Fig. 1). The left three sub-figures depict the original versions of standard MFCCs, and the right sub-figures correspond to the ones using the proposed MFCCs. Inside each small figure, the bold line is for clean speech and the thin line is for distant speech.

V. Speech Recognition Experiments

A. Experimental Setup

The proposed algorithms were evaluated on the in-car CENSREC-2 speech database [1]. This database comprises a task for continuous digit recognition in real car driving environments. In-car speech data was collected in a specially equipped vehicle under 11 environmental conditions created from combinations of three vehicle speeds (idling, low-speed driving on a city street, and high-speed driving on an expressway) and four different in-car environments (normal, with air-conditioner (fan) on, with audio CD player on, and with windows open). The speech data recorded with the close-talking (CT) microphone and hands-free (HF) microphone attached to the ceiling above the driver's seat were used for this corpus, while speech recorded with a HF microphone was used for evaluation. There were four evaluation environments (conditions), as shown in Table I, and the speech recognition performance depended on whether the recording environments and the microphones used for the training and testing data matched.

The speech signals were sampled at 16 kHz. In the baseline system, spectral components lower than 250 Hz were filtered out to compensate for the spectrum of engine noise, which is concentrated in the lower frequency region. The duration of the analysis window was 20 ms with a frame shift of 10 ms. A 24-channel MFB analysis was applied, and the logarithmic outputs of filter banks were computed. The estimated log MFB outputs were transformed into 12 MFCCs, from which the MFCC delta and acceleration coefficients were extracted. Finally, a vector of 39 parameters was used in the hidden Markov models (HMMs). The speech recognition was carried out using the whole-word HMMs.

The vocabulary of CENSREC-2 consists of 11 digit models, where each digit HMM has 18 states with 16 output distribu-

tions. The acoustic models were properly tuned to balance the number of insertions and deletions. Further details of the corpus and the baseline speech recognition system can be found in [1].

B. Speech Recognition Results

The speech recognition results using different methods are summarized in Table II. The experiments can be divided into three groups. The upper part of Table II covers the experiments using the original or proposed log-energy parameter:

- baseline: the original MFCC and log-energy features, and their delta and delta-delta (acceleration) parameters (MFCCs + E + $\triangle$ + $\triangle\triangle$);
- NE: the original MFCC features without the log-energies ((MFCCs + $\triangle$ + $\triangle\triangle$);
- MFCC0: the original MFCC features + the zero-order MFCC, and their delta and delta-delta parameters;
- proposedE (39 dimensions): the original MFCC features + the proposed log-energy given in Section III-A, and their delta and delta-delta parameters.

The middle part consists of the experiments using the proposed log-energy with the subsequent DCE:

- proposedE + DCE1: using (14) for the DCE (linear);
- proposedE + DCE2: using (15) for the DCE (non-linear).

For comparison we also performed the experiments applying the energy normalization[6] and MVN to the original log-energy parameter only (denoted by "Eorg+NORM" and "Eorg+MVN," respectively).

Mean smoothing (filtering) is performed after the DCE. The mean filter used in (17) is essentially a low-pass filter, smoothing out any spikes in the time series. Although for clean speech such spikes may contain important information about the speech variations, for noisy speech these spikes are more likely to be caused by noise (e.g., the first and last spikes in Fig. 3−3). Therefore, there is an inherent trade-off in choosing the order M of the filter in (17). A small M will retain the short-term cepstral information but it is vulnerable to noise, while a large M will ensure that the processed features are less corrupted by noise, but the short-term speech information will be lost. The frequency responses of $M = 3, 5$ are plotted in Fig. 8. The lower part of Table II presents the recognition performance using different mean filter orders. The RASTA filtering [16] was performed for comparison as well.

Table II shows the recognition results obtained for the different methods. From this table, the following observations can be made:

- The "baseline" recognition accuracies depend on the evaluation environments. If the recording environments and the microphones used for the training and testing data do not match, the recognition accuracy can degrade to 43.85% (Condition 4).
- If the original log-energy and its $\triangle$ and $\triangle\triangle$ are not used, the performance increases for the last three unmatched conditions. This illustrates that when the training and testing conditions are not matched, the conventional log-energy and its $\triangle$ and $\triangle\triangle$ become harmful, and should be discarded.
- The use of the zero-order MFCC instead of the original log-energy helps in speech recognition, but its performance is worse than using the proposed log-energy. This can be explained by the fact that some log MFB outputs are so badly contaminated by noise that they cannot reflect variations in the speech. This demonstrates the advantages of the proposed log-energy, in which such seriously-contaminated log MFB outputs are discarded.
- Compared with the normalization of log-energy, applying the MVN to the original log-energy parameter only is effective in improving recognition performance. Compared with using the proposed log-energy only, the subsequent DCEs further improve recognition accuracy, especially for the last three unmatched conditions. DCE2 performs better than DCE1, which is reasonable when considering that the non-linear transformation highlights more prominent speech variations, as shown in Fig. 4. It is noticeable that the proposed log-energy and subsequent DCEs perform even better than MVN except for the last condition, which clearly demonstrates the effectiveness of our proposed algorithms.
- From the lower part of Table II, it is observed that $M = 5$ yields the best results and achieves an average relative improvement of 32.8% compared with the baseline front-ends. This filter order apparently strikes a good balance between the speech information preservation and noise robustness.

C. Incorporating the Proposed MFCCs

We investigated the recognition performance of incorporating the proposed MFCCs by comparing with some conventional methods. Their results are shown in Table III. The upper part consists of the experimental results using the features obtained from various conventional methods[7]:

- SS: the MFCCs and log-energy extracted from the spectral subtraction [19];
- LSA: the MFCCs and log-energy extracted from the speech enhanced by using the minimum mean square error on the log-spectral amplitude [20];
- AFE: the ETSI advanced front-end [21];
- MVN: the cepstral post-processing methods based on the mean and variance normalization (MVN) [22], [23];
- Gau: the cepstral post-processing methods based on the feature space Gaussianization [24];
- LSA + Gau: the first speech enhancement based on the LSA, and then the cepstral-domain Gaussianization.

It can be observed that the speech enhancement method, the LSA , which was proven to be better than the spectral subtraction, is effective under all four evaluation conditions for noise

[6]This normalization is implemented by subtracting the maximum log-energy level in dB of the utterance from the energy dB level of the frame and then limiting the dynamic range of the result to 30–40 dB.

[7]Note that the conventional methods used for comparison are mainly based on speech enhancement, robust feature extraction/normalization, and combinations of these. There are model-based methods as well, such as [17], [18], which perform better than the conventional methods.

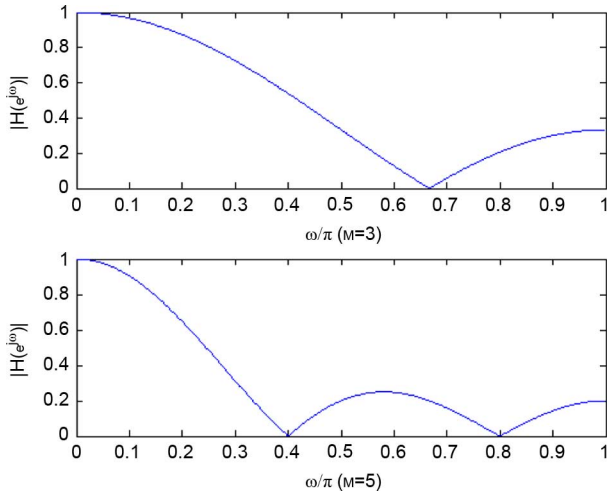


Fig. 8. Amplitude-frequency response of the mean filter with $M = 3$ and $M = 5$.

TABLE III
RECOGNITION ACCURACIES (IN PERCENTAGES) OF USING NEW MFCCS IN SECTION IV. THE UPPER PART PRESENTS THE RECOGNITION PERFORMANCE OF USING THE FEATURES OBTAINED FROM VARIOUS CONVENTIONAL METHODS. THE LOWER PART PRESENTS THE RECOGNITION PERFORMANCE OF ADOPTING THE PROPOSED MFCCS. AVE.: AVERAGED RECOGNITION ACCURACIES OVER THE FOUR CONDITIONS.

	Cond.1	Cond.2	Cond.3	Cond.4	Ave.
SS	84.05	81.80	63.09	57.73	71.67
LSA	81.71	80.08	67.77	60.24	72.45
AFE	84.80	70.42	62.23	56.30	68.44
MVN	83.95	80.87	70.54	64.11	74.86
Gau	84.10	79.93	72.24	65.83	75.52
LSA+Gau	84.04	80.71	74.33	68.16	**76.81**
NewC+Eorg	84.16	83.80	76.85	72.51	79.33
NewC+proposedE	85.04	84.12	77.16	72.71	79.76
NewC+proposedE2	85.66	84.30	82.34	78.82	**82.78**
NewC+proposedE2 +MVN	85.74	84.48	83.26	81.21	**83.67**

reduction. The ETSI advanced front-end (AFE) is not effective except for Condition 1 where both the recording environments and the microphones are matched. Using the normalization methods in the cepstral domain is helpful for improving the in-car speech recognition performance, although the MVN does not perform as well as the Gaussianization. Speech enhancement followed by a Gaussianization processing performs the best, especially for the last two unmatched conditions.

The lower part of Table III corresponds to the speech recognition experiments incorporating the new MFCC front-end discussed in Section IV, denoted by "NewC" in Table III.

- NewC+Eorg: the new MFCCs and the original log-energy features, together with their delta and delta-delta (acceleration) parameters;
- NewC + proprosedE: the new MFCCs + log-energy as proposed in Section III-A, and their delta and delta-delta parameters;
- NewC + proprosedE2: the new MFCCs + the proposed log-energy with a subsequent post-processing consisting of a non-linear DCE and a mean smoothing with $M = 5$, and their delta and delta-delta parameters;
- NewC + proprosedE2 + MVN: MVN is appended to "NewC+ProprosedE2."

From the lower part of Table III, we can see that employing the proposed MFCCs significantly improves recognition performance, which demonstrates their effectiveness in reducing the mismatch between clean and noisy speech as shown in Figs. 6 and 7. "NewC+ProprosedE" again performs better than adopting the original log-energy "NewC+Eorg," although the improvement is not as significant as shown in Table II. This could be explained by the hypothesis that the role of the three-dimensional "proposedE" in improving recognition performance is overshadowed by that of the 36-dimensional "NewC." Adding the proposed post-processing to our log-energy features yields further improvements and achieves an average improvement of 54.1% compared with the baseline front-ends. A subsequent MVN in the cepstral domain yields further ASR improvements, especially for the last two conditions.

D. Discussion

In this paper we focused on the in-car hand-free speech recognition. The noise in the realistic in-car data used in the above evaluations was mainly stationary. To investigate the effect of the proposed methods on other types of noise, we performed the experiments on Aurora 2.0 [25]. In the Aurora 2.0 database, two training sets and three test sets are defined. The multi-train set consists of both clean and noisy speech, while the clean-train set consists of clean speech only. Test set A is composed of speech with the same types of additive noise as those in the multi-train set. Test set B is composed of speech with the non-matched additive noise, while Test set C is composed of speech with the partially matched additive noise and non-matched convolutional noise. The "baseline" feature vector is composed of 39 parameters (12 MFCCs, and their delta and acceleration coefficients as well as the log-energy and its delta and acceleration).

Table IV gives the recognition performance of different methods on the multi-train set. The upper part contains the data for the conventional methods listed in Table II. The lower part corresponds to our proposed methods.

- proposedE: the proposed log-energy with its subsequent post-processing (including both DCE and mean smoothing described in Section III);
- NewC+proprosedE: the proposed MFCCs in Section IV and the proposed log-energy with its subsequent post-processing.

The recognition performance for each noise type was averaged over all SNR levels (including clean, 20 dB, 15 dB, 10 dB, 5 dB, 0 dB, and -5 dB). It can be seen from Table IV that on the average employing the speech enhancement method ("LSA") and advanced front-end ("AFE") is helpful for improving the recognition accuracy; however, their benefits are not as significant as the normalization methods ("MVN" and "Gau"). On average our methods, "proposedE" and "NewC+proposedE," perform better than the speech enhancement method ("LSA") and advanced front-end ("AFE"), but not as good as "MVN" and "Gau." It is noticeable that, compared with the "baseline," our methods, "proposedE" and "NewC+proposedE," are effective in dealing with the convolutional noise (in Set C) and stationary noise (Car noise), but they are not effective for the non-stationary noise (e.g., Babble, Restaurant, Street). Compared

TABLE IV
RECOGNITION ACCURACIES (IN PERCENTAGES) OF DIFFERENT METHODS ON THE MULTI-TRAIN SET OF AURORA 2.0 [25]. THE RECOGNITION PERFORMANCE FOR EACH NOISE TYPE IS AVERAGED OVER ALL SNR LEVELS. AVE.: AVERAGED RECOGNITION ACCURACIES OVER THE 10 NOISE TYPES.

	Set A				Set B				Set C		Ave.
	Subway	Babble	Car	Exhibition	Restaurant	Street	Airport	Station	Subway	Street	
baseline	81.06	80.43	78.12	79.50	77.45	79.71	80.14	76.97	75.53	77.69	78.66
LSA	80.94	72.67	82.54	82.18	80.94	72.67	82.54	82.18	80.94	72.67	79.03
AFE	81.23	80.78	78.80	80.43	78.74	80.19	80.86	77.80	76.23	77.77	79.28
MVN	85.56	80.15	84.39	83.87	78.69	83.51	83.31	81.89	84.68	82.80	82.88
Gau	84.96	78.96	84.40	83.00	76.95	83.37	82.61	81.47	84.03	82.57	82.23
proposedE	80.61	77.61	81.98	83.87	77.02	79.37	81.70	81.55	78.73	78.78	80.12
newC + proposedE	78.91	77.84	85.59	82.56	77.11	79.48	83.12	82.82	79.04	79.75	80.62

with "proposedE," "NewC+proposedE" does not show its advantages for the Subway and Exhibition noise in Set A. However, for other noise types, the benefits of "NewC" are obtained, and especially for the stationary Car noise the gain is maximized. In summary, the proposed log-energy and MFCC front-ends work well for the convolutional and relatively stationary noise, but not for the non-stationary noise, which can be explained by: 1) inaccuracy of estimating noise energies by simply averaging several frames; and 2) failure of enhancing the dynamic changes of the speech signals when the non-stationary noise is involved.

VI. CONCLUSIONS

The log-energy and its delta parameters are critical features for good performance of ASR systems. In the presence of background noise, however, these parameters may introduce serious distortions, reducing their discriminative potential, or even seriously reducing performance, especially for low SNR conditions. In this paper, we theoretically analyzed the impact of background noise on the trajectories of the conventional log-energy and its delta parameters. Based on this, we proposed a robust log-energy parameter estimation algorithm, which significantly reduces the mismatch between clean speech and noisy speech. The effectiveness of the proposed log-energy and its corresponding delta parameters was demonstrated on the CENSREC-2 continuous digit recognition task in real in-car environments. Although the current implementation is in the log MFB domain, the proposed schemes can be straightforwardly applied for J-RASTA [16] or in the root power domain [26].

In this paper we focus on in-car hand-free speech recognition. The noise in the realistic in-car data used in the above evaluations is mainly stationary. Dealing with non-stationary noise will definitely be our future direction. It is also noticeable that the proposed method can be applied only after an utterance ends, and therefore other future work lies in developing a real-time version.

REFERENCES

[1] S. Nakamura, M. Fujimoto, and K. Takeda, "Censrec2: corpus and evaluation environments for in car continuous digit speech recognition," in *Proc. Interspeech'06*, 2006, pp. 2330–2333.

[2] B. C. Moore, *An Introduction to the Psychology of Hearing*. New York, NY, USA: Academic, 1988, pp. 191–191.

[3] P. Loizou and O. Poroy, "Minimum spectral contrast needed for vowel identification by normal-hearing and cochlear implant listeners," *J. Acoust. Soc. Amer.*, vol. 110, no. 3, pp. 1619–1627, 2001.

[4] K. Kluender, J. Coady, and M. Kiefte, "Sensitivity to change in perception of speech," *Speech Commun.*, vol. 41, pp. 59–69, 2003.

[5] P. Assmann and W. Katz, "Time-varying spectral change in the vowels of children and adults," *J. Acoust. Soc. Amer.*, vol. 108, pp. 1856–1866, 2000.

[6] S. Greenberg, W. A. Ainsworth, A. N. Popper, and R. Fay, *Speech Processing in the Auditory System*. New York, NY, USA: Springer-Verlag, 2004.

[7] A. Q. Summerfield, A. Sidwell, and T. Nelson, "Auditory enhancement of changes in spectral amplitude," *J. Acoust. Soc. Amer.*, vol. 81, no. 3, pp. 700–708, 1987.

[8] J. Chen, T. Baer, and B. C. Moore, "Effects of enhancement of spectral changes on speech quality and subjective speech intelligibility," in *Proc. Interspeech*, 2010, pp. 1640–1643.

[9] S. 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.

[10] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, pp. 1738–1752, 1990.

[11] E. L. Bocchieri and J. G. Wilpon, "Discriminative analysis for feature reduction in automatic speech recognition," in *Proc. ICASSP'92*, 1992, pp. 501–504.

[12] S. Ikbal, H. Hermansky, and H. Bourlard, "Nonlinear spectral transformations for robust speech recognition," in *Proc. IEEE Autom. Speech Recogn. Understand. (ASRU) Workshop*, 2003, pp. 393–398.

[13] W. Li and H. Bourlard, "Non-linear spectral contrast stretching for in-car speech recognition," in *Proc. Interspeech'07*, 2007.

[14] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1989.

[15] B. Raj and R. Stern, "Missing-feature approaches in speech recognition," *IEEE Signal Process. Mag.*, vol. 22, no. 5, pp. 101–116, Sep. 2005.

[16] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 2783–2793, Oct. 1994.

[17] G. Saon and J. Chien, "Bayesian sensing hidden Markov models for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'11)*, 2011, pp. 5056–5059.

[18] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

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

[20] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 2, pp. 443–445, Apr. 1985.

[21] Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithm, ETSI ES 202 050 v1.1.1 2002.

[22] J. Openshaw and J. Masan, "On the limitations of cepstral features in noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1994, pp. II/49–II/52.

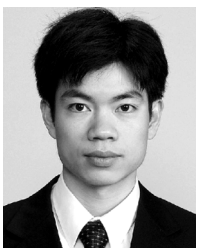
[23] P. Jain and H. Hermansky, "Improved mean and variance normalization for robust speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 2001.

[24] G. Saon, S. Dharanipragada, and D. Povey, "Feature space Gaussianization," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'04)*, 2004, vol. 1, pp. 329–332.

[25] H. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluations of speech recognition systems under noisy conditions," *ISCA ITRW ASR2000*, 2000.
[26] M. Fujimoto, K. Takeda, and S. Nakamura, "Root cepstral analysis: a unified view-application to speech processing in car noise," *Speech Commun.*, vol. 12, pp. 277–288, 1993.

Weifeng Li received the M.E. and Ph.D. degrees in Information Electronics at Nagoya University, Japan in 2003 and 2006, respectively.

Dr. Li joined the Idiap Research Institute, Switzerland in 2006, and in 2008 he moved to Swiss Federal Institute of Technology, Lausanne (EPFL), Switzerland as a research scientist. Since 2010 he has been an associate professor in the Department of Electronic Engineering/Graduate School at Shenzhen, Tsinghua University, China. His research interests lie in the areas of audio and visual signal processing, Biometrics, Human-Computer Interactions (HCI), and machine learning techniques. He is a member of the IEEE and IEICE.



Longbiao Wang received his B.E. degree from Fuzhou University, China, in 2000 and an M.E. and Dr. Eng. degree from Toyohashi University of Technology, Japan, in 2005 and 2008 respectively.

From July 2000 to August 2002, he worked at the China Construction Bank. He was an assistant professor in the faculty of Engineering at Shizuoka University, Japan from April 2008 to September 2012. Since October 2013 he has been an associate professor at Nagaoka University of Technology, Japan. His research interests include robust speech recognition, speaker recognition and sound source localization.

He received the "Chinese Government Award for Outstanding Self-financed Students Abroad" in 2008. He is a member of IEEE, the Institute of Electronics, Information and Communication Engineers (IEICE) and the Acoustical Society of Japan (ASJ).

Yicong Zhou (M'07) received his B.S. degree from Hunan University, Changsha, China, and his M.S. and Ph.D. degrees from Tufts University, Massachusetts, USA, all degrees in electrical engineering.

Dr. Zhou is currently an Assistant Professor in the Department of Computer and Information Science at University of Macau, Macau, China. His research interests focus on multimedia security, image/signal processing, pattern recognition and medical imaging. He is a member of the IEEE and SPIE (International Society for Photo-Optical Instrumentations Engineers).

Hervé Bourlard (M'89–SM'95–F'00) received the Electrical and Computer Science Engineering degree and the Ph.D. degree in applied sciences from the Faculté Polytechnique de Mons, Mons, Belgium.

After having been a member of the Scientific Staff at the Philips Research Laboratory of Brussels and an R&D Manager at L&H Speech Products, he is now Director of the Idiap Research Institute, Martigny, Switzerland, Full Professor at the Swiss Federal Institute of Technology at Lausanne (EPFL), and Director of a National Center of Competence in Research in "Interactive Multimodal Information Management." Having spent (since 1988) several long-term and short-term visits (initially as a Guest Scientist) at the International Computer Science Institute (ICSI), Berkeley, CA, he is now a member of the ICSI Board of Trustees. His main interests are in signal processing, statistical pattern classification, multichannel processing, artificial neural networks, and applied mathematics, with applications to speech and natural language modeling, speech and speaker recognition, computer vision, and multimodal processing. He is the author/co-author/editor of four books and over 250 reviewed papers (including one IEEE paper award) and book chapters.

Dr. Bourlard is an IEEE Fellow for "contributions in the fields of statistical speech recognition and neural networks." He is (or has been) a member of the program/scientific committees of numerous international conferences (e.g., General Chairman of IEEE Workshop on Neural Networks for Signal Processing 2002, Co-Technical Chairman of ICASSP 2002, General Chairman of Interspeech 2003), and on the editorial board of several journals (e.g., past co-Editor-in-Chief of *Speech Communication*). Over the last 20 years, he has initiated and coordinated numerous large international research projects, as well as multiple collaborative projects with industries. He is an appointed expert for the European Commission and, from 2002 to 2007, was also part of the European Information Society Technology Advisory Group (ISTAG).

Qingmin Liao received the B.S. degree in radio technology from the University of Electronic Science and Technology of China, Chengdu, China, in 1984, and the M.S. and Ph.D. degrees in signal processing and telecommunications from the University of Rennes 1, Rennes, France, in 1990 and 1994, respectively.

Since 1995, he has been with Tsinghua University, Beijing, China. In 2002, he became Professor in the Department of Electronic Engineering of Tsinghua University. Since 2010, he has been the director of the Division of Information Science and Technology in the Graduate School at Shenzhen, Tsinghua University. He is also affiliated with the Shenzhen Key Laboratory of Information Science and Technology (Director), China. Over the last 30 years, he has published over 100 peer-reviewed journal and conference papers. His research interests include image/video processing, transmission and analysis; biometrics; and their applications to teledetection, medicine, industry, and sports.