Silence Feature Normalization for Robust Speech Recognition in Additive Noise Environments

Chieh-cheng Wang, Chi-an Pan and Jeih-weih Hung

Department of Electrical Engineering, National Chi Nan University, Taiwan
s95323553@ncnu.edu.tw, s95323544@ncnu.edu.tw, jwhung@ncnu.edu.tw

Abstract

In this paper, we propose a simple yet very effective feature compensation scheme for two energy-related features, the logarithmic energy (logE) and the zeroth cepstral coefficient (c0), in order to improve their noise robustness. This compensation scheme, named silence feature normalization (SFN), uses high-pass filtered features as the indicator for speech/non-speech classification, and then the features of non-speech frames are set to be small while those of speech frames are almost kept unchanged. In experiments conducted on the Aurora-2 database, SFN achieves a relative error reduction rate of nearly 50% from the baseline processing.

Index Terms: speech recognition, feature normalization, robust speech features

1. Introduction

When a clean speech signal is corrupted by additive noise, the magnitude spectra of the noise-corrupted speech signal will deviate from those of the clean signal, and the deviation becomes more significant if the signal-to-noise ratio gets worse. Among the commonly used speech features, the logarithmic energy (logE) can be approximated by the logarithm for the sum of the squared magnitude spectra, while the zeroth cepstral coefficient (c0) is the logarithm of the product of the squared magnitude spectra. As the result, the magnitude spectrum deviation caused by noise directly results in the distortion of logE and c0.

Recently, some approaches have been proposed to enhance the logE feature. For example, in the method of log-energy dynamic range normalization (LEDRN) [1], the dynamic ranges of logE sequences for both training and testing utterances are normalized to a target one. In the method of log-energy rescaling normalization (LERN) [2], a weight is applied to the logE feature. In our previous work, silence energy normalization (SEN) [3], the logE features for the non-speech frame are set to zero. All the three methods tend to adjust the lower logE while keep the higher logE nearly unchanged, with the main reason that the lower-energy frames are more affected by noise than the higher-energy ones.

In this paper, we first provide more rigorous mathematical analysis for the effects of the additive noise on logE and c0. Second, we present a feature compensation scheme, called silence feature normalization (SFN), for logE and c0 in order to improve their noise robustness. Briefly speaking, SFN first uses the high-pass filtered features as the indicator for speech/non-speech classification, and then compensates the features of the non-speech (silence) frames. For the Aurora-2 recognition task, it is shown that SFN significantly improves the recognition accuracy under a wide range of noise-corrupted environments. Furthermore, it outperforms the well-known techniques including mean and variance normalization (MVN) [4], MVN together with ARMA filtering (MVA) [5] and histogram equalization (HEQ) [6].

The remainder of this paper is organized into 5 sections. In Section 2, we investigate the noise effect on the logE and c0 features. Then the proposed SFN technique for compensating logE and c0 is described in Section 3. Section 4 describes the experimental environment. In Section 5, we present the experimental results and compare the proposed approaches with some other techniques. Finally, a brief concluding remark is given in Section 6.

2. The Effect of Additive Noise on the Energy-Related Features

We start the discussion for the effect of additive noise on the logE and c0 features with an illustrating example. Fig. 1 shows the waveform of a clean speech utterance, and the logE and c0 contours for the clean utterance and its two noise-corrupted counterparts. From this figure, some differences in the contours between speech and non-speech portions can be found. For example, the high-logE speech portions are relatively less influenced by noise and reveal a fluctuating characteristic. On the other hand, the low-logE non-speech portions are more vulnerable to noise, and are relatively "flat" in the contours. Similar phenomena can be also found in the c0 contours. In the following two sub-sections, we attempt to explain these phenomena with a mathematical analysis.

Figure 1. (a) the waveform of a clean utterance (b) the logE contours of the clean utterance and its two noise-corrupted counterparts (c) the c0 contours of the clean utterance and its two noise-corrupted counterparts

2.1. The effect of additive noise on logE

Assume the noise-corrupted signal for the n-th frame of an utterance is presented by
\[ x_m[n] = s_m[n] + d_m[n], \quad (1) \]
where \( s_m[n] \) and \( d_m[n] \) are the clean signal and noise, respectively. Then the logarithmic energy (logE) of \( x_m[n] \) is
\[
E^{\omega}[n] = \log(\sum_{n} x_m^2[n]) \approx \log(\sum_{n} s_m^2[n] + \sum_{n} d_m^2[n])
\]
\[
= \log(e^{E^{\omega}(s)} + e^{E^{\omega}(d)}), \quad (2)
\]
where \( E^{\omega}[n] \), \( E^{\alpha}[n] \) and \( E^{\delta}[n] \) are the logE of \( x_m[n] \), \( s_m[n] \) and \( d_m[n] \), respectively. Thus the difference between \( E^{\omega}[n] \) and \( E^{\alpha}[n] \) caused by noise is,
\[
\Delta E[n] = E^{\omega}[n] - E^{\alpha}[n] \approx \log \left( 1 + e^{E^{\omega}(d)-E^{\omega}(s)} \right). \quad (3)
\]
From eq. (3), it is obvious that under the same noise level, \( E^{\alpha}[n] \), \( \Delta E[n] \) decreases as the \( E^{\alpha}[n] \) increases. This implies that for a noise-corrupted utterance, the logE of the speech frame is often less influenced by noise than that of the noise-only frame.

Next, let us consider the noise effect on the modulation spectrum of the logE sequence for an utterance. Taking Taylor series approximation of eq. (2) with respect to \( E^{\alpha}[n], E^{\omega}[n] = (0, 0) \) up to order 2, we obtain
\[
E^{\omega}[n] \approx 2 \log 2 + \frac{1}{2} \left( E^{\alpha}[n] + E^{\omega}[n] \right)
\]
\[
+ \frac{1}{8} \left( \left( E^{\omega}[n] \right)^2 + \left( E^{\alpha}[n] \right)^2 - E^{\alpha}[n] E^{\omega}[n] \right). \quad (4)
\]
Thus the modulation spectrum of the logE sequence \( \{E^{\omega}[n]\} \) can be approximated as
\[
X(j\omega) \approx (2\pi \log 2) \delta(\omega) + \frac{1}{2} \left( S(j\omega) + D(j\omega) \right)
\]
\[
+ \frac{1}{16\pi} \left( S(j\omega) \star S(j\omega) + D(j\omega) \star D(j\omega) - S(j\omega) \star D(j\omega) - D(j\omega) \star S(j\omega) \right). \quad (5)
\]
where \( X(j\omega) \), \( S(j\omega) \) and \( D(j\omega) \) are the modulation spectra of \( \{E^{\omega}[n]\} \), \( \{E^{\alpha}[n]\} \) and \( \{E^{\delta}[n]\} \), respectively, and \( * \) denotes the convolution operation. If the two sequences, \( \{E^{\alpha}[n]\} \) and \( \{E^{\delta}[n]\} \), are both low-pass, and their bandwidths are \( B_\alpha \) and \( B_\delta \), respectively, then the terms \( D(j\omega) \star D(j\omega) \) and \( S(j\omega) \star D(j\omega) \) in eq. (5) have bandwidths of \( 2B_\delta \) and \( B_\alpha + B_\delta \), respectively. This implies that \( \{E^{\alpha}[n]\} \) has a wider bandwidth than \( \{E^{\delta}[n]\} \). In other words, the logE of the noise-corrupted speech segment possesses more higher modulation frequency components than that of the noise-only segment. This is the possible reason why the speech regions look more fluctuating than the noise-only regions in the noise-corrupted logE contour.

### 2.2. The effect of additive noise on \( c_0 \)

The zeroth cepstral coefficients \( c_0 \) for the \( n^\text{th} \) frame of a noise-corrupted signal and its embedded clean signal and noise can be presented by
\[
c^{\omega}_0[n] = \Sigma_0 \log(M^{\omega}[k,n]) \approx \Sigma_0 \log(M^{\alpha}[k,n] + M^{\delta}[k,n]) \quad (6)
\]
\[
c^{\alpha}_0[n] = \Sigma_0 \log(M^{\alpha}[k,n]), \quad (7)
\]
\[
c^{\delta}_0[n] = \Sigma_0 \log(M^{\delta}[k,n]), \quad (8)
\]
respectively, where \( M^{\alpha}[k,n] \), \( M^{\omega}[k,n] \) and \( M^{\delta}[k,n] \) are the \( k^\text{th} \) mel-filter outputs of the noise corrupted signal \( x_m[n] \), the clean signal \( s_m[n] \) and noise \( d_m[n] \), respectively. Thus the difference between \( c^{\omega}_0[n] \) and \( c^{\delta}_0[n] \) caused by noise is,
\[
\Delta c_0[n] = c^{\omega}_0[n] - c^{\delta}_0[n] \approx \Sigma_0 \log(1 + M^{\delta}[k,n] M^{\omega}[k,n])
\]
\[
= \Sigma_0 \log(1 + 1/SNR[k,n]), \quad (9)
\]
where \( SNR[k,n] \) is the signal-to-noise ratio (SNR) for the \( k^\text{th} \) mel-frequency band of the \( n^\text{th} \) frame. Even though a high global SNR does not always ensure a high SNR for each band, it is still expected that the deviation \( \Delta c_0[n] \) for a noise-corrupted speech frame is very likely to be smaller than that for a noise-only frame.

The noise effect on the modulation spectrum of the \( c_0 \) sequence for an utterance is discussed as follows. For simplicity, we re-write the expressions for \( c_0 \) in eqs. (6) and (8) in terms of the logarithmic mel-filter outputs as
\[
c^{\omega}_0[n] = \Sigma_0 \tilde{M}^{\omega}[k,n] \approx \Sigma_0 \log(e^{\tilde{E}^{\omega}[k,n]} + e^{\tilde{E}^{\alpha}[k,n]}), \quad (10)
\]
and
\[
c^{\delta}_0[n] = \Sigma_0 \tilde{M}^{\delta}[k,n], \quad (11)
\]
where
\[
\tilde{M}^{\omega}[k,n] = \log(M^{\omega}[k,n]), \quad \tilde{M}^{\alpha}[k,n] = \log(M^{\alpha}[k,n]), \quad \text{and} \quad \tilde{M}^{\delta}[k,n] = \log(M^{\delta}[k,n]).
\]
Following the analysis procedures for the modulation spectrum of the logE sequence in subsection 2.1, it is easy to show that for each mel-filter output, the bandwidth of \( \\{\tilde{M}^{\omega}[k,n]\} \) is larger than that of \( \{\tilde{M}^{\alpha}[k,n]\} \). Thus \( c^{\omega}_0[n] \) has a wider bandwidth than \( c^{\delta}_0[n] \), and again it shows that the \( c_0 \) of the noise-corrupted speech segment possesses more higher modulation frequency components than that of the noise-only segment. From the above discussions, we conclude that similar to logE, the \( c_0 \) of the speech portion is less influenced by noise, and has a more fluctuating characteristic.

### 3. Silence Feature Normalization

In this section, a compensation scheme, called silence feature normalization (SFN), is introduced for the logE and \( c_0 \) features in order to enhance their noise robustness. This compensation scheme tends to normalize the features in the silence portion, and to keep the features in the speech portion nearly unchanged since they are relatively less influenced by noise. Let \( \{x[n]\} \) represent the logE or \( c_0 \) feature stream of an utterance. We first process \( \{x[n]\} \) with a high-pass IR filter which input-output relationship is
\[
y[n] = -\alpha y[n-1] + x[n], \quad (12)
\]
with the initial condition \( y[0] = 0 \), where \( y[n] \) is the filter output for the \( n^\text{th} \) frame. This filter does not completely remove near-DC components while it emphasizes the higher frequency parts of the input \( \{x[n]\} \). According to the discussions in section 2, the logE and \( c_0 \) features within the speech segments contain higher modulation frequency components compared with those within the noise-only segments. As a result, the resulting \( \{y[n]\} \) is expected to be more helpful in speech/non-speech discrimination than \( \{x[n]\} \). Next, according to the filter output \( \{y[n]\} \), we proceed SFN for \( \{x[n]\} \) in two modes. In the first mode (SFN-I), the new feature \( \tilde{x}[n] \) is obtained as follows,
\[
\text{SFN-I: } \tilde{x}[n] = \begin{cases} x[n] & \text{if } y[n] > \theta \\ \log(e) + \delta & \text{if } y[n] \leq \theta \end{cases}. \quad (13)
\]
where $\theta$ is the threshold, $\varepsilon$ is a small positive constant and $\delta$ is a zero-mean random number with a small variance. That is, if $y[n]$ is greater than the threshold $\theta$, then the $n^{th}$ frame is classified as speech, and $x[n]$ remains unchanged. Otherwise, it is classified as silence (noise-only) and $x[n]$ is normalized to a small random variable. This mode of SFN is very close to SEN [3] in our previous work, except that a small random variable is used here rather than a constant. This avoids the possible problem of SEN that the variance of the Gaussian distribution in the silence model becomes zero.

In the second mode (SFN-II), the new feature $\hat{x}[n]$ is obtained by weighting $x[n]$,

$$\hat{x}[n] = w[n]x[n],$$

where the weighting function $w[n]$ is defined by

$$w[n] = \begin{cases} \frac{1}{N} & \text{if } y[n] > \theta \\ 1 & \text{if } y[n] \leq \theta \end{cases},$$

(14)

where $\theta$ is the threshold, $\sigma_1$ and $\sigma_2$ are the standard deviations of $\{y[n] | y[n] > \theta\}$ and $\{y[n] | y[n] \leq \theta\}$, respectively, and $\beta$ is a constant.

Fig. 2 depicts the weighting function $w[n]$ by assuming $\theta = 0$, $\sigma_1 = 1$, $\sigma_2 = 3$ and $\beta = 0.1$. It is shown that $w[n]$ is a modified sigmoid curve whose left and right sides correspond to different logistic functions. Similar to SFN-I, in SFN-II the new feature $\hat{x}[n]$ for a more "silence-like" frame ($y[n] < \theta$) approaches zero, while it remains almost the same as $x[n]$ for a more "speech-like" frame ($y[n] \geq \theta$).

However, SFN-II performs a soft-decision speech/non-speech classification, which may have a smaller effect of misclassification than SFN-I. For both SFN-I and SFN-II, the threshold $\theta$ is set simply to the mean of $\{y[n]\}$,

$$\theta = (1/N)\sum_{n=0}^{N-1}y[n],$$

(16)

where $N$ is the number of frames in the utterance.

Fig. 2. The weighting function $w[n]$ in eq. (16)

4. Experimental Setup

We perform recognition experiments on the AURORA-2 database [7]. For the recognition environment, two test sets (Set A and Set B) of utterances corrupted by different types of additive noise (subway, babble, etc.) and different SNR levels (from 20dB to -5dB) are prepared. In order to analyze the performance of the proposed approaches on the two different features, log$E$ and $c_0$, individually, each frame of the utterances in the training and testing sets is converted to two different types of 39-dimensional feature vector. The first consists of 12 cepstral coefficients ($c_1$-$c_{12}$) and the logarithmic energy (log$E$) plus their first and second derivatives, while the second consists of 13 cepstral coefficients ($c_0$, $c_1$-$c_{12}$) plus their first and second derivatives. With the feature vectors in the training set, the hidden Markov models (HMMs) for each digit and silence are trained following the Microsoft complex back-end training scripts [8]. Each digit HMM has 16 states and 20 Gaussian mixtures per state. The parameters $\alpha$ in eq. (12), $\varepsilon$ in eq. (13) and $\beta$ in eq. (15) are set to 0.5, $10^{-3}$ and 0.1, respectively, and the zero-mean random number $\delta$ in eq. (13) is Gaussian distributed with a variance of $10^{-8}$.

5. Experimental Results

5.1. The results of the log$E$-processing approaches

In this subsection, we compare the recognition performance of several log$E$-processing approaches including the proposed SFN-I and SFN-II, log-energy dynamic range normalization (LEDRN) [1], log-energy rescaling normalization (LERN) [2], mean and variance normalization (MVN) [4], MVN together with ARMA filtering (MVA) [5], and histogram equalization (HEQ) [6]. Here the used 39-dim feature vector consists of $c_1$-$c_{12}$ and log$E$ plus their first and second derivatives. In LEDRN, the normalization process can be linear or nonlinear, and thus we use LEDRN-L and LEDRN-N to represent the linear and nonlinear LEDRN, respectively. In HEQ, the feature stream for each utterance is normalized to be nearly Gaussian distributed with zero mean and unity variance. Note that these approaches are used only to update the log$E$ feature and its first and second derivatives, while the other 36 features are kept unchanged.

Table 1 shows the recognition results for the baseline and various methods. The method "clean-log$E$" in the third row of this table is to replace each noise-corrupted log$E$ feature with its clean counterpart. The recognition accuracy of this method can be viewed as the upper bound of the accuracy achieved by the other approaches. From this table, some points can be found as follows,

1. The performance difference between the baseline and the method "clean-log$E$" reveals the fact that the distortion of log$E$ caused by additive noise results in a very significant degradation in recognition accuracy.

2. Nonlinear LEDRN behaves better than the linear one. LERN performs better than nonlinear LEDRN for Set B, while the situation is converse for Set A.

3. The improvement achieved by MVN is relatively less significant. In MVA, the low-pass ARMA filtering provides MVN with an extra improvement of nearly 1.3%. However, HEQ obviously outperforms MVN and is slightly better than LEDRN and LERN. Compared with MVN, HEQ additionally compensates the higher-order moments of the feature, and this extra compensation indeed results in an apparent improvement.

4. The proposed approaches, SFN-I and SFN-II, provide very significant improvement for both Set A and Set B, which shows that they effectively alleviate the deteriorating effect of additive noise on log$E$. Also, both of them outperform all the other approaches mentioned previously. Finally, SFN-II is slightly better than SFN-I, which partly supports our statement that a soft-decision mode for speech/non-speech classification in SFN-II has a minor misclassification effect.

The proposed SFN is easily integrated with cepstral processing techniques since they are performed on different features. Here, we combine SFN with any of the approaches, MVN, MVA and HEQ, and the corresponding recognition results are shown in Table 2. Comparing Table 2 with Table 1, we see that such combinations bring about further improvements over each individual approach, and SFN again shows its excellence in performance. For example, the method "SFN-II for log$E$ plus MVA for $c_1$-$c_{12}$" outperforms "MVA for log$E$ and $c_1$-$c_{12}$" by 3.18% in absolute accuracy improvement.
In the names of these methods, \( c_0 \) stands for the features \( c_1-c_{12} \). 

### 5.2. The results of the \( c_0 \)-processing approaches

In this sub-section, the recognition results of several \( c_0 \)-processing approaches are presented and discussed. These approaches include MVN, MVA, HEQ, and the proposed SFN-I and SFN-II. Here the used 39-dim feature vector consists of \( c_0-c_{12} \) plus their first and second derivatives. Analogous to the previous sub-section, these approaches are used only to update the \( c_0 \) feature and its first and second derivatives, while the other 36 features are kept unchanged. Table 3 shows the recognition results for the baseline and various approaches. The method "clean-\( c_0 \)" in this table is to replace each noise-corrupted \( c_0 \) feature with its clean counterpart. First, comparing the results of the baseline and the method "clean-\( c_0 \)\), we find that, similar to \( \log E \), the mismatch of \( c_0 \) caused by additive noise deteriorates the recognition accuracy seriously. Second, MVN, MVA and HEQ significantly improve the recognition accuracy, and similar to the \( \log E \) case in Table 1, here HEQ provides a better compensation for \( c_0 \) than MVN and MVA. Third, the accuracy improvements achieved by SFN-I and SFN-II are very outstanding, which shows that they significantly enhance the noise robustness of \( c_0 \) as they do for \( \log E \). In particular, SFN-II performs the best among all the methods and provides 49.50% in relative word error rate reduction. Analogous to the previous subsection, here MVN, MVA or HEQ are performed on the features \( c_1-c_{12} \) to see if it is additive to SFN processed on \( c_0 \). The corresponding recognition accuracy results shown in Table 4 indicate that very promising improvement can be achieved with such combinations. For example, the method "SFN-II for \( c_0 \) plus MVA for \( c_1-c_{12} \)" provides more than 65% in relative error rate reduction, and it is better than "MVA for \( c_0 \) and \( c_1-c_{12} \)\) (61.03%) and "SFN-II for \( c_0 \)" (49.50%).

### 6. Concluding Remarks

In this paper, we investigate the effect of additive noise on the logarithmic energy (\( \log E \)) and the zeroth cepstral coefficient (\( c_0 \)), and then provide the two features with a compensation scheme, called silence feature normalization (SFN), to enhance their noise robustness. SFN compensates the features in the silence portion of an utterance and keeps them nearly unchanged in the speech portion. It has been shown that SFN is very effective in promoting the recognition performance under various noise conditions for the AURORA-2 task, and it can be integrated with the well known MVN, MVA and HEQ to provide further improvement.

### References


