

Silence Energy Normalization for Robust Speech Recognition in Additive Noise Environments

Chung-fu Tai and Jieh-weih Hung
 Dept of Electrical Engineering, National Chi Nan University
 Taiwan, Republic of China

e-mail : s3323542@ncnu.edu.tw, jwhung@ncnu.edu.tw

Abstract

The energy parameter has been widely used as an extension to the basic features of mel-frequency cepstral coefficients (MFCCs) to improve the recognition accuracy in speech recognition. In this paper, a simple and effective approach for energy normalization for silence (non-speech) portions in an utterance is proposed. This approach, named as silence energy normalization (SEN), uses the high-pass filtered log-energy as the feature for speech/non-speech classification, and then the log-energy of non-speech frames is set to be a small constant while that of speech frames is kept unchanged. In the experiments conducted on AURORA2 database, we showed that SEN provides an averaged word error rate reduction of 34.9% and 44.6% for Test Sets A and B, respectively, when compared with the baseline processing. It was also shown that SEN outperforms similar approaches like energy subtraction (ES) and feature vector selection (FVS). Finally, we showed that SEN can be integrated with cepstral mean and variance normalization (CMVN), to achieve further improved recognition performance.

Index Terms: silence energy normalization, frame vector selection, energy subtraction

1. Introduction

The performance of a speech recognition system is often severely degraded in the presence of noise. A variety of approaches have been proposed to mitigate this degradation, and they can be roughly divided into three classes: utilization of a noise robust representation of speech signals, enhancement of the speech features before they are fed to the recognizer, and adaptation of the speech models in the recognizer to make them better match the noise conditions. The main difference between the first two classes of approaches is that, for the first class, the noise robust speech features are used for both model training and testing, and for the second, enhancement procedures are often performed only on the noise corrupted speech features for testing, while the speech features for training are kept unchanged. In this paper, our proposed approach belongs to the first class. A new feature normalization scheme called silence energy normalization (SEN) is introduced.

As we know, the energy of speech signal contains important information regarding the phonetic content of speech, and therefore we have used it directly or the variation of it (for example, log-energy, delta energy, etc.) to be one of the speech features for recognition. However, these energy features are often vulnerable to noise and thus their discriminating capability is limited. Recently, some approaches have been proposed to enhance these energy features [1-5]. For example, in [1] the speech energy contour is extracted from the high-pass filtered signal so as to reduce the distortion in the delta energy,

and in [2] the dynamic range of log-energy sequences for both training and testing utterances is normalized to a target one in order to reduce the environmental mismatch, where the normalization function indicates lower-energy frames are more affected by noise than higher-energy ones. On the other hand, [6] introduces the method of frame vector selection (FVS) based on variable frame rate processing or voice activity detection (VAD), where the frame-to-frame variation (for example, the Euclidean norm of the delta feature) or the log-energy of each frame is used as an indicator for frame selection. If its value is below a predefined threshold, this frame is classified as silence or noise-only and is then discarded. Partly motivated by the concepts in [2] and [6], in this paper we propose the approach of silence energy normalization (SEN). In SEN, every frame vector of an utterance is first classified as speech or non-speech (silence). The classifier is based on the output of a high-pass filter with the log-energy being the input. Then for each of the silence frames the log-energy is normalized to a small constant, while the log-energy of the speech frames remains unchanged. Note that in SEN, the classification and normalization procedures need to be performed on the utterances in both clean training and noisy testing databases, and unlike FVS, the frames labelled as non-speech are not discarded. We will show that by SEN the normalized log-energy sequence of a noise corrupted utterance is quite close to that of the corresponding clean version. Also, the threshold used in the classifier for SEN is determined by the input utterance and needs not to be tuned heuristically, which is a particular benefit of SEN. Furthermore, SEN can be easily integrated with cepstral compensation techniques, for example, cepstral mean and variance normalization (CMVN) [7], to obtain further improved performance. The remainder of the paper is organized as follows. In section 2, the proposed approach of SEN is described. The experimental environment setup is described in section 3, and the recognition results of SEN and some other approaches are given and discussed in section 4. In addition, section 4 also contains the recognition results of the combination of SEN and CMVN. Finally, section 5 briefly contains some concluding remarks.

2. Silence Energy Normalization

2.1 Basic idea

When observing the log-energy contours of a clean utterance and its noise-corrupted counterparts as in Figure 1(a), some differences between the speech and non-speech portions may be found. For example, the high-energy speech portions are relatively less influenced by noise and sometimes keep the ripple characteristics. On the other hand, the low-energy non-speech portions



of the clean utterance are relatively “flat” in the contour, and they are much more vulnerable to noise since their log-energy levels are significantly elevated. Furthermore, the “flatness” of the non-speech portions is kept and sometimes further enhanced by the effect of noise. These observations lead us to develop the algorithm of silence energy normalization (SEN), which mainly deals with non-speech portions while keeping speech portions unaltered.

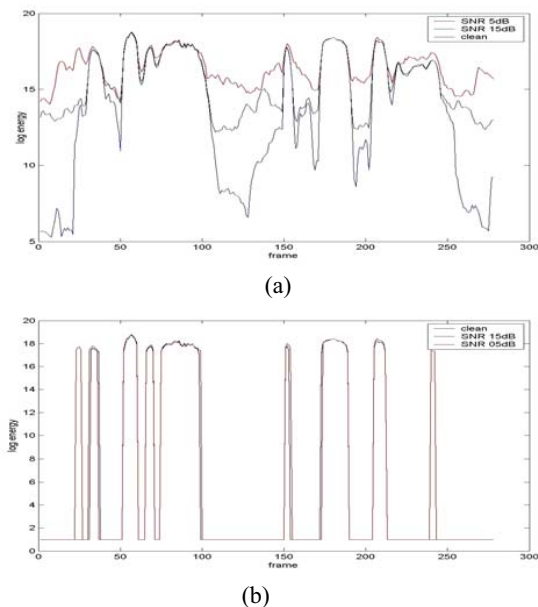


Figure 1 (a) the original log-energy contours of a clean utterance and its noise-corrupted counterparts with 15dB and 5dB of SNR, respectively (b) the SEN-processed log-energy contours of three utterances the same as those in Figure 1(a)

2.2 The procedures of SEN

The algorithm of SEN mainly consists of two steps. The first step is to classify each frame as speech or non-speech (silence), which is analogous to the voice-activity detection (VAD), and the second step is to normalize the log-energy of each silence frame to be a small constant. According to [6], the variable frame rate (VFR) processing in the frame vector selection (FVS) uses the delta features to indicate which are speech frames or silence frames. As we know, the delta operation often possesses band-pass characteristics, which removes the near-DC low-frequency components. Here, we use a simple IIR high-pass filter different from the delta filter in [6], and its input-output relationship is

$$y[n] = \frac{1}{2}(e[n + 1] - y[n - 1]), \tag{1}$$

where $e[n]$ is the log-energy of the n -th frame and $y[n]$ is the corresponding filter output. Figure 2 shows the frequency responses of this high-pass filter and the delta filter used in [6]. From this figure, it is shown that the used high-pass filter does not particularly de-emphasize the lower frequency components, and according our experimental results, it performs better than the delta filter.

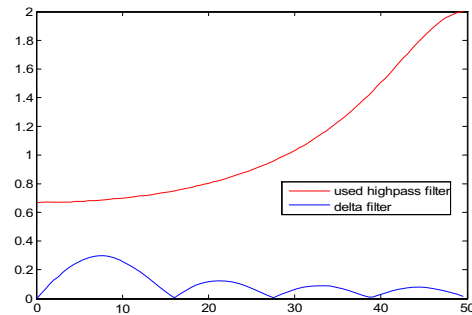


Figure 2. The amplitude responses of the used high-pass filter and the delta filter used in [6]

Next, according to the filter output $y[n]$, the normalized log-energy $\tilde{e}[n]$ for the n -th frame can be obtained by the following equation.

$$\tilde{e}[n] = \begin{cases} e[n] & \text{if } y[n] > T \\ \varepsilon & \text{if } y[n] \leq T \end{cases} \tag{2}$$

where T is the threshold and ε is a small constant. That is, if $y[n]$ is smaller than the threshold T , then the n -th frame is classified as silence and its log-energy is normalized to be ε . Otherwise, the n -th frame is classified as speech and its log-energy remains unchanged. Here the threshold T is set in an utterance-wise manner, and it equals to the average of $y[n]$ in an utterance. That is,

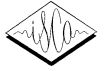
$$T = \frac{1}{N} \sum_{n=1}^N y[n], \tag{3}$$

where N is the number of total frames in an utterance. Obviously, the choice of the threshold T in eq. (3) involves the whole utterance and thus possibly introduces a long delay, which may be not feasible solution in some real-time applications. However, the advantage of such choice is that it is quite simple and performs very well under various signal-to-noise ratio (SNR) conditions, which will be shown in section 4.

Figure 1(b) shows the SEN-processed log-energy contours of the clean utterance and its two different noise-corrupted versions which are the same as those in Figure 1(a). From this figure, it can be shown that there is little difference among these normalized contours. It is shown that SEN preserves most speech portions while normalizes the log-energy of silence portions to be a small value ε (ε is set to be 1 here).

3. Experimental Setup

The proposed silence energy normalization algorithm has been tested with the AURORA2 database. For the recognition experiments, two sets (Test Sets A and B) of utterances artificially contaminated by different types of noise (subway, babble, car, etc.) and different SNR levels (ranging from -5dB to 20dB) were prepared. For both the clean training and noisy testing databases, each utterance was first converted into a stream of 12 mel-frequency



cepstral coefficients (MFCCs) plus log-energy. Then the log-energy sequence for each utterance was processed by the proposed SEN algorithm described in section 2 or some other approaches that will be described in section 4. The original 12-dimensional MFCCs (c1~c12) and the updated log-energy, plus their delta and delta-delta were the components in the finally used 39-dimensional feature vectors. Since the proposed algorithm only involves the front-end feature extraction, all the following procedures for training and recognition are identical to the reference experiments stated in the AURORA2 documentation [8].

4 Experimental Results

4.1 The results of the energy-processed approaches

In this subsection, we compare the recognition performance of several energy-processing approaches including the proposed SEN, Energy Subtraction (ES) [3] and two versions of Feature Vector Selection (FVS). In FVS here, the frames classified as silence are directly removed from the utterance without any normalization. The first version of FVS uses the speech/silence classifier in SEN, and is thus denoted as SEN-FVS, while the second makes use of the output of ES method, and is thus denoted as ES-FVS. The ES and ES-FVS are briefly introduced here.

The approach of energy subtraction (ES) is quite similar to the typical spectral subtraction (SS), and the algorithm is stated as follows,

$$\tilde{E}[n] = \begin{cases} E[n] - \alpha \bar{N} & \text{if } E[n] \geq T_{ES} \\ \beta E[n] & \text{if } E[n] < T_{ES} \end{cases}, \quad (4)$$

where $E[n]$ and $\tilde{E}[n]$ are the original and updated energy values of the n -th frame, respectively, T_{ES} is a threshold, \bar{N} is the noise energy estimate, α is the over-subtracting factor and β is the flooring factor. In the following experiments, both \bar{N} and T_{ES} are set to be the average of energy values for the first five frames of each utterance, and α and β are set to be 0.95 and 0.05, respectively.

For the approach of ES-FVS, a simple rule based on the updated energy values $\{\tilde{E}[n]\}$ from ES is used for frame vector selection. If the two consecutive energy values, $\tilde{E}[n]$ and $\tilde{E}[n-1]$, are both smaller than a threshold T_{ES-FVS} , then the n -th frame is classified as silence and then discarded. The threshold is determined by the following equation,

$$T_{ES-FVS} = wT_{ES} + (1-w) \frac{1}{N} \sum_{n=1}^N \tilde{E}[n], \quad (5)$$

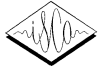
where N is the number of total frames in an utterance, T_{ES} is the threshold used in Eq. (4), and $0 < w < 1$. In our experiments, w is set to be 0.4.

Table 1 shows the averaged recognition accuracy for the baseline processing and several approaches stated previously for Test Set A (four types of stationary noise), and Test Set B (four types of non-stationary noise) of

AURORA2 database. From this table, several phenomena can be observed:

1. The proposed approach SEN significantly improves the recognition accuracy for both stationary and non-stationary noise conditions in almost every SNR case. For example, compared with the baseline results, SEN gives 13.57% and 19.28% of absolute word accuracy improvements for Test Sets A and B, respectively. Furthermore, it performs particularly well for non-stationary noise conditions since it gives the best recognition performance among all approaches in almost all SNR cases in Test Set B. On the other hand, for Test Set A, SEN is especially well for the median and low SNR (10dB~0dB) cases.
2. The approach ES also performs quite well and very similarly to SEN for Test Set A, and it gives 11.98% of absolute word accuracy improvement when compared with the baseline result. However, it does not perform as well as SEN for Test Set B, although it still outperforms the baseline processing by 14.32% of accuracy rate. One of the possible reasons is that under non-stationary noise cases, the noise estimate in ES here (simply the average of the first several frames) is not very accurate. In addition, the two parameters, α and β in eq. (4), are set to be constants here for simplicity, which in fact should be updated according to different noise conditions.
3. When implementing frame-vector-selection (FVS) after the classification procedure of SEN (denoted as SEN-FVS in the table), the recognition performance is deteriorated when compared with that of SEN alone. Such results probably tell us that the "detected" silence portions do not always provide us with redundant or erroneous information for recognition. Modifying these portions (as in SEN) instead of discarding them (as in FVS) may be a better choice. Another possible reason is that the classifier used here does not perform very well, and thus some portions like the silence-to-speech or speech-to-silence transitions are misclassified as silence and are then discarded.
4. Finally, observing the results of ES-FVS, where the frame selection is based on the results of ES, it is found that ES-FVS is better than SEN-FVS for all cases, which possibly shows that the speech/silence classifier in ES-FVS is more reliable than that in SEN-FVS since it depends on the noise-reduced energy sequence $\{\tilde{E}[n]\}$ in eq. (4) rather than the filtered log-energy sequence $\{y[n]\}$. These results somewhat coincide those obtained in [6], where FVS is performed on the noise-reduced features to obtain a better recognition accuracy. Furthermore, we also find that ES-FVS outperforms ES only when the SNR is worse (0dB and 5dB). One possible reason is that under worse SNR conditions, ES is less capable of dealing with the silence frames, and thus dropping them directly as in FVS may be more beneficial.

From the above, we can roughly conclude that the proposed SEN performs excellently for almost all conditions. For example, under high and median SNR cases, it preserves the energy contour of speech portions



like ES, thus it performs almost as well as ES. When the SNR is getting worse, SEN successfully deals with the frames of silence portions and thus it is not necessary to discard them as in FVS.

Test Set A	Baseline	SEN	ES	SEN-FVS	ES-FVS
Clean	98.91	98.71	99.02	96.14	99.09
20dB	94.99	96.90	97.14	88.23	94.43
15dB	86.93	93.81	93.94	83.62	90.06
10dB	67.28	85.18	84.12	75.54	80.60
5dB	39.36	64.84	61.24	57.43	63.55
0dB	17.07	32.75	29.10	28.11	33.54
average	61.13	74.70	73.11	66.58	72.43

(a)

Test Set B	Baseline	SEN	ES	SEN-FVS	ES-FVS
Clean	98.83	98.77	99.02	96.14	99.09
20dB	92.35	96.95	96.73	88.73	94.20
15dB	80.79	94.29	92.21	85.51	89.62
10dB	58.06	86.52	80.29	78.26	79.29
5dB	32.04	65.93	55.68	60.93	59.68
0dB	14.63	33.29	24.55	30.96	32.90
average	55.57	75.39	69.89	68.88	71.14

(b)

Table 1. Recognition accuracy (%) for baseline and various approaches, silence energy normalization (SEN), energy subtraction (ES), SEN-based frame vector selection (SEN-FVS) and ES-based frame vector selection (ES-FVS) for (a) Test Set A and (b) Test Set B in Aurora 2 database.

4.2 The results of the integration of SEN and CMVN

The proposed SEN is easily integrated with cepstral processing approaches since they are performed on different features. Here, we combine SEN with the approach of cepstral mean and variance normalization (CMVN), and the corresponding recognition accuracy rates are shown in Table 2. For the purpose of comparison, Table 2 also contains the results of the baseline processing, CMVN alone, and SEN alone. From this table, we find that CMVN alone only leads to improvement for non-stationary noise cases (Test Set B). However, integrating CMVN with SEN brings significantly improved performance for both stationary (Test Set A) and non-stationary noise environments in almost all SNR cases, and it is better than SEN alone especially when the SNR is worse (0dB~10dB). These results apparently indicate that SEN and CMVN are additive.

5. Concluding Remarks

In this paper, we have proposed the approach of silence energy normalization (SEN) for the log-energy feature in speech recognition. One of the main benefits of SEN is that it is very simple to realize, and in SEN almost no parameters need to be tuned heuristically. Experimental results also show that it is very effective in promoting the

recognition accuracy under various noisy conditions. Furthermore, it can be integrated with the well known cepstral mean and variance normalization (CMVN) to obtain further improved performance.

Test Set A	Baseline	SEN	CMVN	SEN&CMVN
Clean	98.91	98.71	98.95	98.70
20dB	94.99	96.90	91.52	96.17
15dB	86.93	93.81	84.50	93.50
10dB	67.28	85.18	66.16	86.91
5dB	39.36	64.84	41.92	72.38
0dB	17.07	32.75	19.56	44.86
average	61.13	74.70	60.73	78.76

(a)

Test Set B	Baseline	SEN	CMVN	SEN&CMVN
Clean	98.83	98.77	98.95	98.70
20dB	92.35	96.95	93.02	96.54
15dB	80.79	94.29	81.84	94.27
10dB	58.06	86.52	64.73	87.49
5dB	32.04	65.93	43.01	72.42
0dB	14.63	33.29	22.25	44.89
average	55.57	75.39	60.97	79.12

(b)

Table 2. Recognition accuracy (%) for baseline processing, silence energy normalization (SEN), cepstral mean and variance normalization (CMVN), and the combination of SEN and CMVN (SEN&CMVN) for (a) Test Set A and (b) Test Set B in Aurora2 database.

References

- [1] Tai-Hwei Hwang, "Energy Contour Extraction for In-Car Speech Recognition", 9th European Conference on Speech Communication and Technology (Eurospeech 2003)
- [2] Weizhong Zhu and Douglas O'Shaughnessy "Log-energy Dynamic Range Normalization for Robust Speech Recognition", 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005).
- [3] Tai-Hwei Hwang and Sen-Chin Chang, "Energy Contour Enhancement for Noisy Speech Recognition", 2004 International Symposium on Chinese Spoken Language Processing (ISCSLP 2004)
- [4] M. Ahadi, H. Sheikhzadeh, R. Brennan and G. Freeman, "An Energy Normalization Scheme for Improved Robustness in Speech Recognition" 8th International Conference on Spoken Language Processing (ICSLP 2004)
- [5] R. Chengalvarayan, "Robust Energy Normalization Using Speech/nonspeech Discriminator for German Connected Digit Recognition", 6th European Conference on Speech Communication and Technology (Eurospeech 1999)
- [6] J. Veth et al, "Feature Vector Selection to Improve ASR Robustness in Noisy Conditions", 7th European Conference on Speech Communication and Technology (Eurospeech 2001)
- [7] O. Viikki, K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," Speech Communication, 1998
- [8] H.-G Hirsch, D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," ISCA ITRW ASR 2000, Paris, France, September 18-20, 2000