Robust speech recognition by selecting mel-filter banks

Yun-Peng Wu[a], Jia-Min Mao[b] and Wei-Feng Li[c]
Department of Electrical Engineering/Graduate School at Shenzhen,
Tsinghua University, China
[a]wyp15@mails.tsinghua.edu.cn; [b]maj15@mails.tsinghua.edu.cn;
[c]weiflee@126.com

Mel-filterbank energies is a key feature that is widely employed in automatic speech recognition(ASR) system. It arises from a sub-band spectrum typically. But when the noise exists in the background, Mel-filterbank energies can not be easy to estimated accurately. In this paper, the fact that the trajectories of not only "traditional" log Mel-filterbank energies, but also its delta parameters can be influenced by noise will be theoretically analyzed. As a result, log Mel-filterbank energies and their delta parameters can not be calculated correctly. In this paper, we propose to remove those severely contaminated Mel-filterbank features and only keep those variations which perform better in the speech remained. We demonstrate the effectiveness of this novel operation through speech recognition experiments conducted on the Aurora-2 database.

Keywords: Speech recognition; Mel-filterbank (MFB); Melfilterbank energies; Mel-Frequency Cepstral Coefficients (MFCCs).

1. Introduction

It is widely known that changes in speech inputs over time is sensitive for human auditory system [1], and a certain extent of spectral conflict is essential for a robust speech recognition system [2]. Many discoveries from auditory aesthesia experiments also prove that procedures of continuous spectral conflict is able to help disambiguate co-articulated speech [3]. [4] presented that removing the natural time-varying spectral varieties over the duration of a vowel leaded to much lower American English vowels recognition accuracy.

In contrast situations, background noise always results in a decline in changing varieties in speech signals. For even ordinary hearing listeners, severe decline in dynamic varieties results in uncertain segmentation, and this will add difficulties in parsing the speech signal [5]. However, it has been shown that in multiple situations the auditory system has flexibility serving not only to emphasize current reaching element of the signal but also to strengthen the sphere of the signal undergoing spectro-temporal varieties [6]. Besides, there is a valid proof that clearly strengthening the dynamic change will be helpful to recognize speech signal [7].

Mel-Frequency Cepstral Coefficients (MFCCs) [8], derived from short-time spectral energies in a condensed sphere, are widely used in many Automatic Speech Recognition (ASR) systems. MFCCs are derived from log scaled mel-filterbank (MFB) energies. However, when background noise exists, the dynamic varieties in spectral energies will be widely weaken. In figure 1 and figure 2, the first row and second row show the clean speech and noisy speech waveforms, contaminated by babble and car noise (5 dB signal-to-noise ratio, SNR) respectively. The third row and fourth row show the first-channel log MFB trajectory (or contour) of clean speech and noisy speech respectively. The fifth and sixth rows show the 21st log MFB trajectory (or contour) of clean speech and noisy speech respectively. In contrast of the short-distance speech, the floor level of the log MFB trajectory for remote speech is advanced and the valleys are hidden by noise energy. While spectral varieties in close-distance talk over time are much more clear, they get into fuzzy for noisy speech because of the noise influence. Therefore, traditional log MFB energies and MFCCs usually import unwanted mispair between comparatively clean speech (used for training) and noisy speech (used for testing), owing to the background noise. This leads to a serious decline of ASR performance. On the other hand, the trajectories of different log Mel-filterbank energies show different dynamic changes, as shown in the first and 21st Mel-filterbanks. In this paper, we use log MFB energy features as the front-ends of speech recognition systems, and we raise the problem of the mismatch of log MFB energy features when background noise exists. More explicitly, this paper analyses systematically the influence of the noise on the trajectories of traditional log MFB energies and its delta parameters, resulting in failure of describing speech changes, and thus reduce the speech recognition accuracy in low SNR situations. We then plan to get rid of the log Mel-filterbank features which could no longer describe the changes in the speech (with low dynamic varieties). By dislodging the log MFB energies with low dynamic varieties, only the log MFB features which show better exhibitions of the variations in the speech are kept.

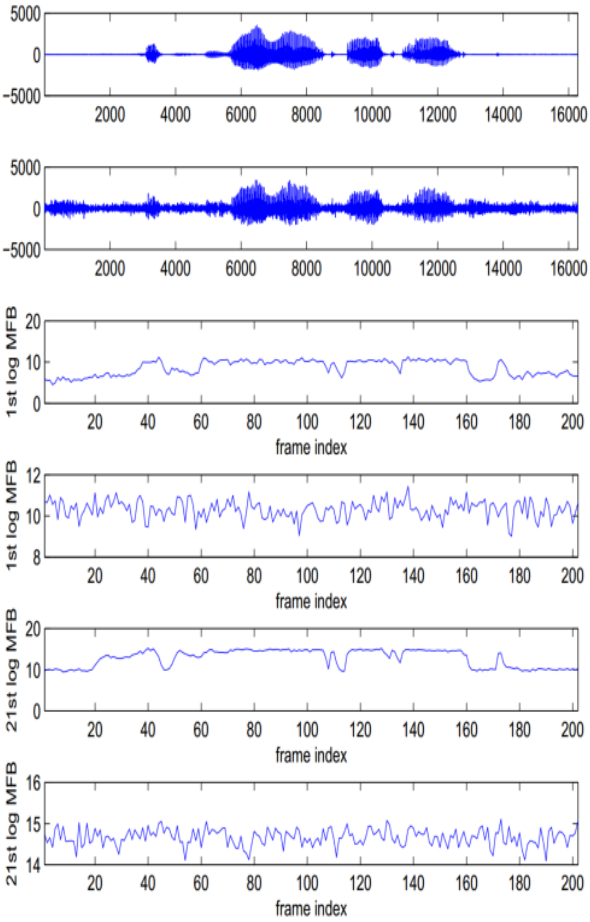

Figure 1: Effect of babble noise on log mel-filter bank (MFB) energy trajectories. Upper part (first and second rows): waveform of clean speech and noisy speech; Middle left (third and fourth rows): log MFB outputs of the first channel for clean speech and noisy speech; Lower left (fifth and sixth rows): log MFB outputs of the 21st channel for clean speech and noisy speech.

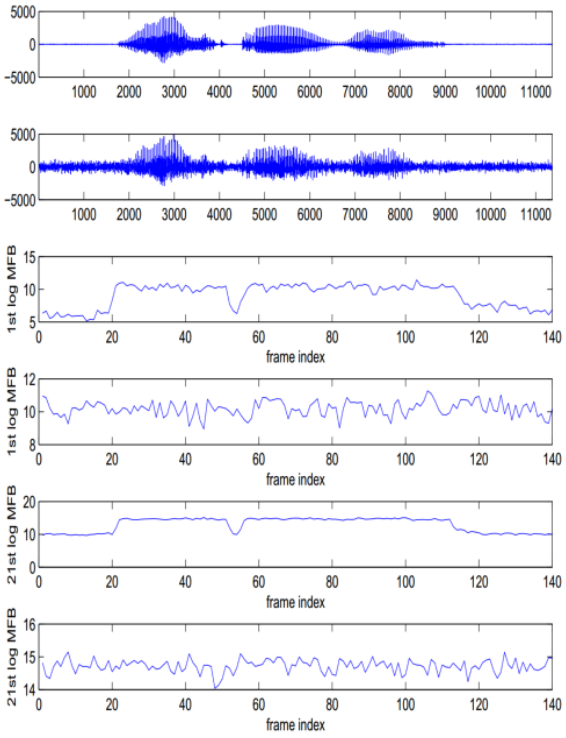

Figure 2: Effect of car noise on log mel-filter bank (MFB) energy trajectories. Upper part (first and second rows): waveform of clean speech and noisy speech; Middle left (third and fourth rows): log MFB outputs of the first channel for clean speech and noisy speech; Lower left (fifth and sixth rows): log MFB outputs of the 21st channel for clean speech and noisy speech.

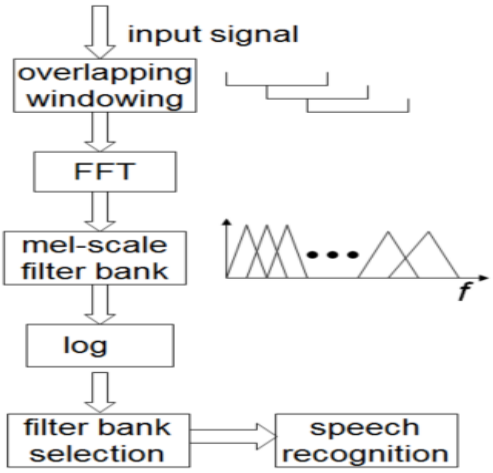

Figure 3: Effect of car noise on log mel-filter bank (MFB) and log energy trajectories.

In this way the discriminative powers of log MFB features are expected to be strengthened. We conducted our experiments on Aurora-2 database [9] in different training and test situations. The results show the effectiveness of our method. The remainder of this paper is as follows: Section 2 analyzes the resulting mismatch between clean and noisy conditions owing to traditional log MFB energy estimation theoretically. Section 3 presents a new method to delete some log MFB energy feature with a low spectral, and describes the HMM adaption. Section 4 presents our experiments on Aurora-2 databases in details and analyzes the results. Conclusions are given in Section 5.

2. Log-energy Estimation Mismatch between Clean and Noisy Conditions

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

$$x(i) = s(i) + n(i). \quad (1)$$

After the short-time Fourier Transformation (STFT), in frequency domain it can be

$$X(k,l) = S(k,l) + N(k,l). \quad (2)$$

Here k is the frequency bin index and l is the frame index. The log mel-filterbank energy of noisy speech at the l-th frame is computed by

$$E(m,l) = \log(\sum_{k=M_s}^{M_t} W_m |X(k,l)|^2) \quad (3)$$

$$= \log(\sum_{k=M_s}^{M_t} W_m (|S(k,l)|^2 + |N(k,l)|^2), \quad (4)$$

where W_m is the weight of the m-th filterbank, as shown in Figure 3. M_s and M_t are the beginning and end of frequency bins of the m-th filterbank. $|S(k,l)|^2$ is the energy of k-th frequency bin at frame l. Here we assume that the clean speech and noise are statistically independent.

The dynamic changes of log mel-filterbank energy can be computed as the difference between the log mel-filterbank energy of noisy speech at frame l and the subsequent one (e.g., atframe l+p, $p > 0$).

$$\begin{gathered} C_{E(m,l)} = E(m,l+p) - E(m,l) \\ = \log(\sum_{k=M_s}^{M_t} W_m (|S(k,l+p)|^2 + |N(k,l+p)|^2)) \\ -\log(\sum_{k=M_s}^{M_t} W_m (|S(k,l)|^2 + |N(k,l)|^2)) \end{gathered}$$

$$= \log \frac{\sum_{k=M_s}^{M_t} W_m(|S(k,l+p)|^2+|N(k,l+p)|^2)}{\sum_{k=M_s}^{M_t} W_m(|S(k,l)|^2+|N(k,l)|^2)} \quad (5)$$

$$\simeq \log(1+\frac{\sum_{k=M_s}^{M_t} W_m(|S(k,l+p)|^2-|S(k,l)|^2)}{\sum_{k=M_s}^{M_t} W_m(|S(k,l)|^2+|N(k,l)|^2)}) \quad (6)$$

Where the assumption of this approximation is that the noise energy can keep nearly stable over time, i.e.,

$$\sum_{k=M_s}^{M_t} W_m|N(k,l+p)|^2 \simeq \sum_{k=M_s}^{M_t} W_m|N(k,l)|^2. \quad (7)$$

Formula (6) suggests that the existing of noise leads to the reduce of the dynamic change in log-energy, and it will become lower when the noise energy increases. When $|S(k,l)|^2 = 0$ (i.e., nonspeech segments) and $|S(k,l+p)|^2 > 0$ (i.e., speech segments), according to (6) the dynamic change in log-energy from non-speech segments to speech segments can be decreased to

$$\log\left(1+\frac{\sum_{k=M_s}^{M_t} W_m|S(k,l+p)|^2}{\sum_{k=M_s}^{M_t} W_m|N(k,l)|^2}\right)$$

$$= \log(1+\frac{E_s(m,l+p)}{E_n(m,l)})$$
$$\simeq log(1+SNR(m,l+p)), \quad (8)$$

where

$$E_s(m,l) = \sum_{k=M_s}^{M_t} W_m|S(k,l)|^2 \quad (9)$$

$$E_n(m,l) = \sum_{k=M_s}^{M_t} W_m|N(k,l)|^2 \quad (10)$$

$$E_s(m,l+p) = \sum_{k=M_s}^{M_t} W_m|S(k,l+p)|^2 \quad (11)$$

$$E_n(m,l+p) = \sum_{k=M_s}^{M_t} W_m|S(k,l+p)|^2, \quad (12)$$

and

$$SNR(m,l+p) = \frac{E_s(m,l+p)}{E_{n(m,l)}} \quad (13)$$

indicates the SNR at frame l + k. The transition from speech (at frame l) to non-speech (at frame l+k) segments, Eq. (5) can be reduced to

$$\log(\frac{\sum_{k=M_s}^{M_t} W_m|N(k,l+p)|^2}{\sum_{k=M_s}^{M_t} W_m(|S(k,l)|^2+|N(k,l)|^2)}$$

$$\simeq -log(1 + SNR(m, l)). \quad (14)$$

It is stated in above equations that the dynamic change is shrank by the presence of noise. As shown in the figure 1 and 2, the influence of reducing of the dynamic change can be clearly observed. When the noise plays a leading role, i.e.,

$$E_n(m, l) \gg E_s(m, l), \quad (15)$$

Eq. (5) shrinks to log(En(m,l+p)=En(m,l)). In this case, dynamic changes in the noisy speech signals over time uncover dynamic changes in the noise rather than in the speech. In Figures 1 and 2, the fourth and sixth rows illustrate this problem, highlighted in the first and last 50 frames. In conclusion, with the existence of background noise the traditional static log Mel-filterbank 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.

There will be a mismatch between noisy speech and relatively clean speech if they were inputted into an ASR system, leading to a degrade to the performance of ASR.

3. Proposed Mel-filterbank Removal

2.1. Mel-filterbank removal

In consideration of two facts, we proposed to delete those severely contaminated Melfilterbank energies: (1) outputs of log MFB are sub-band based, thus, with in some particular sub-bands, it can capture dynamic variations in the speech signals over time; (2) log MFB can produce wider change ranges across time. Compared with smaller ones, log MFB can reflect dynamic variations in the speech signals better.

Table 1: Recognition accuracies (in percentages) of different methods on the multi-train set of aurora 2.0 [9]. 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	Bable	Car	Exhibition	Restaurant	Street	Airport	Station	SubWay	Street	
baseline	73.02	74.41	73.46	72.91	73.82	73.92	76.98	72.85	59.83	66.50	71.77
LSA	73.94	65.67	75.54	75.18	73.94	73.94	75.54	75.18	73.94	65.67	72.03
AFE	74.23	73.78	71.80	73.43	71.74	71.74	73.86	70.80	69.23	70.77	72.28
proposed (10 Melfilter banks)	73.75	70.15	72.28	69.14	68.80	68.80	72.93	70.97	66.83	63.23	69.83
proposed (20 Melfilter banks)	75.18	74.41	74.37	73.86	73.38	73.38	77.06	73.52	62.71	67.41	72.67

Therefore, based on the consideration the log MFB outputs of a particular filter bank may have greater energy than those of other filter banks, we select the log MFB outputs with larger dynamic changes and remove those log MFB with less dynamic changes, as is shown in Figure 3. The dynamic change in their log MFB values for the j-th filter bank is defined by

$$D(j) = \frac{E_{max}(j) - E_N(j)}{E_N(j)} \tag{16}$$

where Emax(j) and EN(j) are the maximum values of the jth log MFB outputs along the frames of the utterance and the estimated noise log MFB value, respectively. EN(j) is estimated by the mean of the j-th log MFB outputs over the first non-speech frames[1]. In this paper the first 15 frames are used to estimate EN (j) in our experiments.

Though traditional speech recognition systems usually employ the entire Mel-filterbanks energies, we plan to choose a number of robust ones with larger dynamic changes and delete some seriously contaminated ones by the background noise with less dynamic changes. This technique can be regarded as a type of "missing feature theory" [10], which accepts reliable log MFB outputs. In this case the missing feature masks (or confidence measures) are derived from Eq. (16). When the test speech utterance is decoded, the acoustic models (trained using Hidden Markov Models (HMM) in this paper) are changed accordingly. More specifically, the new HMMs are generated by deleting the corresponding log MFB features.

4. Experimental Results

In order to estimate the proposed methods with different types of noises, we conducted experiments on Aurora 2.0 [9], which includes two training sets and three testing sets. The two training sets are clean speech and noisy speech respectively, while the clean set consists of clean speech only. For the testing set A, the noise of the speech is same as those in the training set. On the contrary, testing set B is composed of speech with unmatched additive noise. And testing set C consists of speech with partially matched additive noise and non-matched convolutional noise. The "baseline" feature vector is composed of 48 parameters (23 log MFB outputs and a log-energy and their delta parameters). The digits are modeled as whole word HMMs with 16 states per word (according to 18 states in HTK notation with 2 dummy states at beginning and end). Simple left-to-right models without skips over states (3 Gaussian mixtures per state) are used. Table 1 shows the recognition performance of different methods on the multi-trainset. The proposed methods have two following configurations.

- proposed (10 Mel-filterbanks): keep the log MFB features with the first 10 largest dynamic values and remove the others;
- proposed (20 Mel-filterbanks): keep the log MFB features with the first 20 largest dynamic values and remove the others;.

The experimental results for each noise type were 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 1 that employing the speech enhancement method ("LSA" [11]) and advanced front-end [9] ("AFE") is beneficial for improving the recognition accuracy on the average; however, their contributions are limited. The proposed method with 10 Mel-filterbanks remained performs not well mainly because too many features are neglected. Our method with 20 Mel-filterbanks achieves better recognition performance in average by employing the method (referred LSA and AFE above), especially for the noise in the car, airport and station. This obviously proved the robustness of our approach.

5. Conclusions

The log MFB energies and its delta parameters are discriminative features for good performance of ASR systems. With the background noise, these parameters may result in intense distortions, reducing their recognition ability, or even seriously reducing performance, especially in the low SNR condition. In this paper, we analyzed the influence of background noise on the trajectories of the conventional log MFB energies and its delta parameter in the theory. Based on this, we proposed to remove the log MFB energies with less dynamic changes, which could alleviate the mismatch between clean speech and noisy speech. The effectiveness of the proposed log-energy and its corresponding delta parameters

was demonstrated on the Aurora-2 continuous digit recognition task. Although the current implementation is in the log MFB domain, the proposed schemes can be further applied in the root power domain [12].

Reference

[1] B. C.J. Moore, An introduction to the psychology of hearing, Academic Press, New York, pp. 191, 1988.
[2] P.C. Loizou and O. Poroy, "Minimum spectral contrast needed for vowel identification by normal-hearing and cochlear implant listeners," Journal of the Acoustic Society of America, vol. 110(3), pp. 1619C1627, 2001.
[3] K.R. Kluender, J.A. Coady, and M. Kiefte, "Sensitivity to change in perception of speech," Speech Communication, vol. 41, pp. 59–69, 2003.
[4] P.F. Assmann and W.F. Katz, "Time-varying spectral change in the vowels of children and adults," Journal of the Acoustic Society of America, vol. 108, pp. 1856C1866, 2000.
[5] S. Greenberg, W. A. Ainsworth, A. N. Popper, and R. Fay, Speech Processing in the Auditory System, Springer-Verlag, New York, 2004.
[6] A. Q. Summerfield, A. Sidwell, and T. Nelson, "Auditory enhancement of changes in spectral amplitude," Journal of the Acoustical Society of America, vol. 81, no. 3, pp. 700–708, 1987.
[7] J. Chen, T. Baer, and B. C.J. Moore, "Effects of enhancement of spectral changes on speech quality and subjective speech intelligibility," in Proceedings Interspeech, 2010, pp. 1640–1643.
[8] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 28, no. 4, pp. 357 – 366, 1980.
[9] H.G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluations of speech recognition systems under noisy conditions," in ISCA ITRW ASR2000, 2000.
[10] B. Raj and R.M. Stern, "Missing-feature approaches in speech recognition," Signal Processing Magazine, IEEE, vol. 22, no. 5, pp. 101 – 116, 2005.
[11] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," IEEE Transactions Acoustics, Speech and Signal Processing, vol. 32, no. 2, pp. 443–445, 1985.
[12] M. Fujimoto, K. Takeda, and S. Nakamura, "Root cepstral analysis: a unified view-application to speech processing in car noise," Speech Communication, vol. 12, pp. 277–288, 1993.