한국음향학회지 제31권 제2호 pp. 87~92 (2012)
The Journal of the Acoustical Society of Korea Vol.31, No.2 (2012)

ARMA Filtering of Speech Features Using Energy Based Weights

Sung Min Ban and Hyung Soon Kim

School of Electrical Engineering, Pusan National University
(Received January 3, 2012; accepted February 6, 2012)

ABSTRACT: In this paper, a robust feature compensation method to deal with the environmental mismatch is proposed. The proposed method applies energy based weights according to the degree of speech presence to the Mean subtraction, Variance normalization, and ARMA filtering (MVA) processing. The weights are further smoothed by the moving average and maximum filters. The proposed feature compensation algorithm is evaluated on AURORA 2 task and distant talking experiment using the robot platform, and we obtain error rate reduction of 14.4 % and 44.9 % by using the proposed algorithm comparing with MVA processing on AURORA 2 task and distant talking experiment, respectively.
Key words: Feature compensation; Temporal modulation filter; ARMA filter
ASK subject classification: Speech Signal Processing (2)

I. Introduction

The performance of speech recognition system is degraded by the mismatch between the acoustic conditions of the training and the testing environments. Sources of this mismatch include additive noise, channel distortion, speaker characteristics and speaking style. In particular, as the distance between source and microphone increases, the environmental mismatch due to additive noise and channel distortion also increases. In this paper, to alleviate this problem, a feature compensation algorithm is proposed, which is robust to the environmental mismatch caused by additive noise and channel distortion. Among many feature compensation algorithms, we employ the temporal modulation filter technique, which has the advantage of being simple to implement while providing robustness to both additive noise and reverberant distortion.

***Corresponding author:** Hyung Soon Kim (kimhs@pusan.ac.kr)
School of Electrical Engineering, Pusan National University, Busan 609-735, Korea
(Tel: +82-51-510-2452; Fax: +82-51-515-5190)

The existing temporal modulation filter techniques include cepstral mean normalization (CMN), cepstral mean and variance normalization (CMVN) and relative spectral (RASTA) filter [1]. Recently the temporal modulation normalization (TMN) and MVA processing were proposed [2,3]. Both filters showed a good performance, but MVA has less computational complexity than TMN.

Our work is based on MVA, and complements the weakness of MVA. MVA processes the CMVN results through auto-regressive moving average (ARMA) filter, in which the features in the adjacent background noise region tends to distort those in the speech region. To reduce this distortion, we applied energy based weights according to the degree of speech presence to ARMA filter coefficients. In our previous study of applying these weights, we observed the performance improvement on the isolated word recognition task [4]. In this paper, to relieve additional distortions caused by the modified ARMA filter, we obtain new weights from the smoothed log energy

contour by the maximum filter (MF) after moving average (MA) filtering of zeroth-order cepstral coefficient, C0. The proposed algorithm is evaluated on AURORA 2 task and distant-talking experiment using the robot platform.

This paper is organized as follows: In Section 2 we introduce the weighted ARMA filter which uses the information of the degree of speech presence. In Section 3 how to obtain the MA/MF weights is explained. Finally, the performance of the proposed algorithm is shown in Section 4 and we conclude this paper.

II. Weighted ARMA filtering

The conventional MVA is expressed as follows [3]:

$$C_{ARMA}^{(t,k)} = \frac{\sum_{i=1}^{m} C_{ARMA}^{(t-i,k)} + \sum_{i=0}^{m} \hat{C}^{(t+i,k)}}{2m+1} \tag{1}$$

where $C_{ARMA}^{(t,k)}$ is the result of ARMA filtering with order index k and time index t and m is an order of ARMA filter. $\hat{C}^{(t,k)}$ is the CMVN result of the cepstral coefficient with order index k and time index t. The CMVN processing is as follows:

$$\hat{C}^{(t,k)} = \frac{C^{(t,k)} - \overline{C}^{(k)}}{\sigma^{(k)}}, \quad k = 0,1,\cdots,12 \tag{2}$$

where $\overline{C}^{(k)}$ and $\sigma^{(k)}$ are the cepstral mean and the cepstral standard deviation respectively. $C^{(t,k)}$ is the cepstral coefficient with order index k and time index t. CMVN compensates for the mismatch between the cepstral distributions of the training and test data by equalizing their mean and variance values. ARMA filter with the characteristics of low pass filtering removes the high frequency components in the cepstral time sequences, while preserving the intelligibility information under 16 Hz modulation frequency of speech. Thus it compensates for the residual mismatch between the training and test data after CMVN.

One weakness in conventional MVA processing is that, in ARMA filtering, the features in the speech region can be distorted by the features in the adjacent background noise region, because ARMA filtering uses the adjacent features in equation (1) to sum up them. This weakness results in the degradation of MVA performance. In our previous study, weighted ARMA filter was proposed to complement this weakness in MVA [4]. In the weighted ARMA filter, the weights according to the degree of speech presence are multiplied to the coefficients of ARMA filter for each frame, so the distortion due to the features in the adjacent background noise region may be reduced. The weighted ARMA filter is given by

$$C_w^{(t,k)} = \frac{\sum_{i=1}^{m} w(t-i)\, C_w^{(t-i,k)} + \sum_{i=0}^{m} w(t+i)\, \hat{C}_w^{(t+i,k)}}{2m+1} \tag{3}$$

where

$$w(t) = \frac{1}{1+e^{-\alpha x(t)}} \tag{4}$$

and

$$x(t) = C^{(t,0)} - \beta \overline{C}^{(0)} \tag{5}$$

$C_w^{(t,k)}$ is the result of the weighted ARMA filtering of the cepstral coefficient with order index k and time index t. $w(t)$ is the weight according to $x(t)$, the degree of speech presence at frame t. $C^{(t,0)}$, the zeroth-order cepstral coefficient, represents the log energy of frame t and $\overline{C}^{(0)}$ is the the mean of the zeroth-order cepstral coefficient over the total frames. Thus $x(t)$ is related to the degree of speech presence. We use a sigmoid function to normalize $w(t)$ into the range [0,1]. And α and β are the positive constants, m is the order of weighted ARMA filter.

The effects of the weighted ARMA filter are shown in Figure 1 where the time sequences of the first-

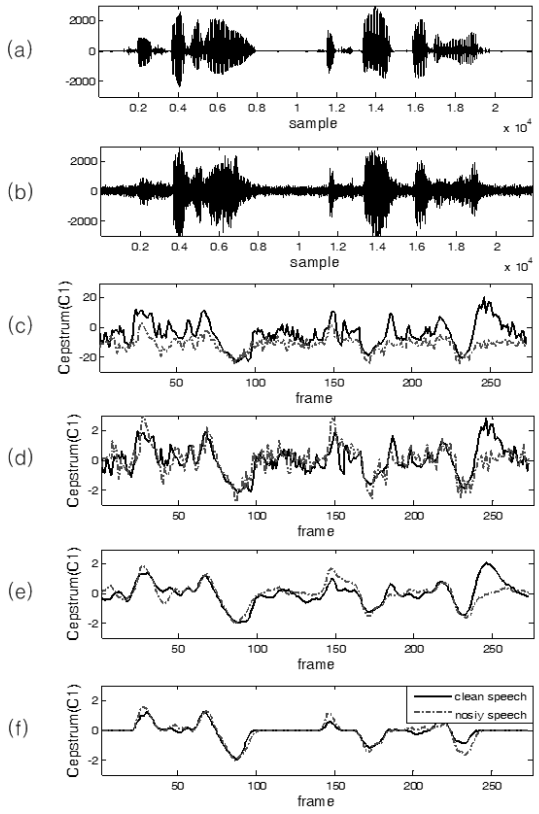

Fig. 1. Time sequences of the first-order cepstral coefficient in the distant-talking speech and the close-talking speech (a man speaks "27O6571") (a) close-talking speech waveform (b) distant-talking speech waveform (5dB SNR) (c) no processing (d) CMVN (e) ARMA filter (f) weighted ARMA filter.

order cepstral coefficient in distant-talking speech and close-talking speech are compared. The distant-talking speech is recorded at real robot platform, while playing the close-talking speech from the loudspeaker located 1 m away from the robot. The distant-talking speech contains robot noise and reverberant effect. In (c) it is observed that there are large differences between both the time sequences without any post-processing. In (d) these differences are greatly reduced by CMVN, but still there are some differences, which can lead to the other distortions in the first and the second derivatives of the cepstral coefficients. In (e) by applying the ARMA filter to the CMVN results, the high frequency components of the time sequences are removed, and less differences in

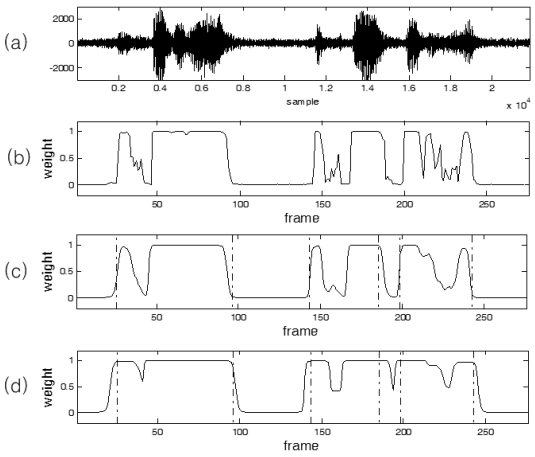

Fig. 2. Time sequence of the ARMA filter weight in distant-talking speech (a) distant-talking speech waveform (b) weight of conventional weighted ARMA filter (c) weight using MA filter (d) weight using MA/MF.

time sequences of the cepstral coefficient are observed. As we mentioned earlier, if the weights according to the degree of speech presence are applied to the ARMA filter, it can prevent the features in the noise region from affecting the features in the adjacent speech region. Thus the differences between the two time sequences are reduced. To observe this effect, in (f) we filtered the CMVN results through the weighted ARMA filter using the ideal weights from clean speech. In the background noise region, the weights are close to zero and the two time sequences of the cepstral coefficients become almost the same. Eventually these similar time sequences can reduce the speech recognition errors.

III. Weights using smoothed energy

In conventional weighted ARMA filter, the weights can lead to two kinds of new distortions. First, the weights tend to have small values in a short silence interval between the speech segments, and this may distort the temporal modulation structure of speech. Second, incorrectly small weights at speech boundaries

can mask the cepstral coefficients in speech region. From these reasons, the recognition errors may increase. In this paper, we propose the weights based on the smoothed energy using the moving average and maximum filter (MA/MF) in order to complement the weakness of conventional weighted ARMA filter. We use MA processing in the equation (6) to relieve the first mentioned distortion, and filter the result of MA processing through the MF processing in the equation (7) to relieve the second mentioned distortion.

$$C_{MA}(t) = \frac{1}{2k+1} \sum_{i=-k}^{k} C^{(t+i,0)} \tag{6}$$

$$C_{MF}(t) = \max_{(-p \le i \le p)} C_{MA}(t+i) \tag{7}$$

Using the $C_{MF}(t)$ instead of the $C^{(t,0)}$ in the equation (5), a new degree of speech presence is determined. The effects of the MA/MF in the equation (6) and (7) are examined in Figure 2. Figure 2(a) indicates the waveform of distant-talking speech which is the same as in Figure 1. The time sequence of the weight in the conventional weighted ARMA filter is represented in (b), where we can observe that the weights have small values in a short interval between the speech segments. But after MA filtering these small weights are smoothed out in (c). Also, to protect the weights in the speech region from being estimated incorrectly, some margin around the speech boundary frame is set by using MF. The proposed feature compensation process is summarized in Figure 3.

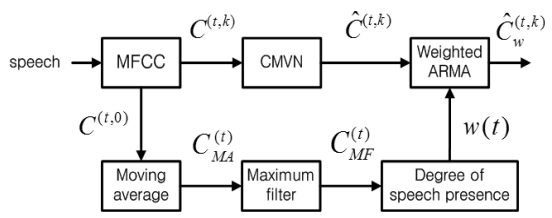

Fig. 3. Proposed feature compensation process.

IV. Experimental results

The performance of the proposed feature compensation algorithm is evaluated on AURORA 2 task [5] and a pilot test is also performed on the real robot platform in distant-talking condition.

We get 39-dimensional Mel-frequency cepstral coefficients (static:13, delta:13, delta-delta:13) with C0 using power spectrum. We set k as 4, p as 3 in MA/MF processing and α as 0.4 in weighted ARMA processing. In previous study, it was found that there was little difference in the performances of speech recognition for α value of more than 0.3 [4]. Acoustic models on AURORA 2 task are trained using the clean condition training DB. In this training, it is possible to estimate almost correct weights according to the degree of speech presence, so we use the weighted ARMA filter without applying MA/MF in training.

Table 1 shows the word recognition accuracies for AURORA 2 clean condition training using the proposed feature compensation algorithm. The accuracies are averaged over the five SNR levels from 0 to 20 dB in each test set. In this table, wARMA indicates the conventional weighted ARMA filter in which the MA/MF is not employed. Next, the weighted ARMA filters using MA filter, MF, MA/MF are expressed as wARMA (MA), wARMA (MF), wARMA (MA/MF), respectively. And wARMA (ideal) means the performance of wARMA (MA/MF) when using the weights

Table 1. Word recognition accuracies on AURORA 2 task for clean condition training (%).

Algorithm	Set A	Set B	Set C	Average
Baseline	56.07	52.76	62.22	55.98
MVA	81.21	81.66	82.29	81.66
wARMA	82.66	83.17	83.15	82.96
wARMA (MA)	83.20	83.89	83.66	83.57
wARMA (MF)	83.21	84.10	83.55	83.63
wARMA (MA/MF)	83.81	84.79	84.34	84.31
wARMA (ideal)	83.62	84.47	84.73	84.18

obtained from clean speech.

From the table, the weighted ARMA filter overall outperforms MVA. The performance improvements of the proposed algorithms over MVA are statistically significant, e.g., p-value<0.001 for wARMA (MA/MF) over MVA for all the test sets. When both the MA filter and the MF are used, the best performance is achieved. It is also found that the performances of wARMA (ideal) and wARMA (MA/MF) are similar levels, from which we can expect that even if more sophisticated voice activity detector is employed, additional performance improvement may not be achieved.

To confirm the performance of the proposed algorithm on distant-talking condition, we performed a pilot test on the robot named Engkey, which was developed by the Center for Intelligent Robotics at the Korea Institute of Science and Technology (KIST). In this experiment, speech data from four microphones are preprocessed by multi-channel Weiner filter (MWF) [6]. To evaluate the algorithm, clean speech and noise data are played by two loudspeakers, respectively. The loudspeaker for speech is placed just in front of the robot and 1 m away from it, and the loudspeaker for noise is placed 2 m away from the robot. The angle between the loudspeakers for speech and noise is 60 degrees. The robot has fan noise as internal noise, and the sounds of vacuum cleaner and TV are additionally used as external noises. Table 2 shows the isolated word recognition performance in the distant-talking condition. We used the Korean phonetically balanced words database provided by SiTEC (Speech Information Technology and Industry Promotion Center) as test data. And we used CleanSent01 database which consists of phonetically balanced sentences also provided by SiTEC as training data. For comparison, the ETSI advanced front-end (AFE) algorithm was also employed [7]. It can be seen from the table that our proposed method, wARMA(MA/MF), outperforms both MVA and AFE over all noise conditions.

Table 2. Isolated word recognition accuracies in distant-talking condition (%).

Noise	Baseline	MVA	AFE	Proposed
Close-talking	97.45	96.18	96.91	95.64
Robot	5.09	58.00	54.73	74.00
Vacuum cleaner	2.91	42.55	47.45	74.36
TV	4.17	42.86	38.18	67.64
Average	27.41	59.90	59.32	77.91

Fig. 4. Engkey robot with eight microphones (Four microphones indicated by arrows are used in this paper).

V. Conclusions

In this paper, a feature compensation algorithm, robust to the environmental mismatch, is proposed. The proposed algorithm applies the energy based weights according to the degree of speech presence to ARMA filter in the MVA processing and employs the moving average and the maximum filter to relieve the additional distortion. The proposed feature compensation algorithm shows better performance than the conventional MVA in various noise envi-

ronments. As future work, we are planning to apply the proposed technique to more delicate algorithm such as data-driven temporal filtering.

Acknowledgment

This research was performed for the Intelligent Robotics Development Program, one of the 21st Century Frontier R&D Programs funded by the Ministry of Knowledge Economy of Korea.

References

1. H. Hermansky, N. Morgan "RASTA processing of speech", *IEEE Trans. Speech and Audio Process*., vol. 2, no. 4, pp. 578-589, 1994.
2. X. Lu, S. Matsuda, M. Unoki, S. Nakamura, "Temporal contrast normalization and edge-preserved smoothing of temporal modulation structures of speech for robust speech recognition", *Speech Comm*., vol. 52, no. 1, pp. 1-11, 2010.
3. C. P. Chen, J. Bilmes, "MVA processing of speech features", *IEEE Trans. Audio Speech Language Process*., vol. 15, no. 1, pp. 257-270, 2007.
4. S. M. Ban, H. S. Kim, "Robust speech recognition using weighted auto-regressive moving average filter", *Journal of the Korean Society of Speech Sciences*, vol. 2, no. 4, pp. 145-151, 2010.
5. H. G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluations of speech recognition systems under noisy conditions," *ISCA ITRW ASR2000*, Sep. 2000.
6. K. B. Kim, N. I. Cho, "Frequency domain multi-channel noise reduction based on the spatial subspace decomposition and noise eigenvalue modification," *Speech Comm*., vol. 50, no. 5, pp. 382-391, 2008.
7. ETSI, "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," *ETSI ES 202 050 Recommendation*, 2002.

Profile

▸ Sung Min Ban

Sung Min Ban received the B.S. and the M.S. degrees in electrical engineering in 2008 and 2010, respectively, from Pusan National University (PNU), Busan, Korea. He is currently pursuing the Ph.D. degree in electrical engineering at PNU. His current research interests include speech recognition and signal processing.

▸ Hyung Soon Kim

The Journal of the Acoustical Society of Korea, Vol. 30, No. 6, 2011.