THE JOURNAL OF THE ACOUSTICAL SOCIETY OF KOREA Vol.22, No.1E 2003. 3. pp. 43~48

Acoustic Channel Compensation at Mel-frequency Spectrum Domain

So-Young Jeong*, Sang-Hoon Oh**, Soo-Young Lee***
*Brain Science Research Center (BSRC) and also Division Of Electrical Engineering, Department of Electrical Engineering and Computer Science, Korea Advanced Institute of Science and Technology
**Department of Information Communication Engineering, Mokwon University
***BSRC and also Department of BioSystems, Korea Advanced Institute of Science and Technology
(Received January 17 2003; accepted March 4 2003)

Abstract

The effects of linear acoustic channels have been analyzed and compensated at mel-frequency feature domain. Unlike popular RASTA filtering our approach incorporates separate filters for each mel-frequency band, which results in better recognition performance for heavy-reverberated speeches.

Keywords: *Acoustic channel compensation, Channel deconvolution, Mel-frequency log-spectrum, Feature transformation, Robust speech recognition*

I. Introduction

Acoustic channel mismatches between training and testing environments result in performance degradation in automatic speech recognition. Although time-domain deconvolution filters may be developed, they require extensive computation, especially for acoustic channels with long time delays. Therefore, many researchers had come up with filtering approaches at feature domain. Effects of microphones and telecommunication channels can be modeled with impulse responses with short time delays and add bias terms to clean speech features in the log-spectrum domain, which may be compensated by log-spectral mean subtraction[1,2].

However, room acoustics usually come with longer time delays, which introduce interactions among several time frames. Highpass or bandpass filters at modulation-frequency domain had been developed by heuristics or based on information theory[1,3].

In this study, we analyzed the effects of acoustic channels with longer time delays on speech features, and came up with a new filtering method at mel-frequency spectral domain. Unlike other feature space methods our method incorporates separate filters for each mel-frequency band, which are optimized based on given training sets of clean and distorted speech data. The performance of the proposed compensation method was tested for several acoustic channels.

Corresponding author: So-Young Jeong (syjeong@extell.com)
Extell Technology Corp., 5F Soam Bldg., 44-10 Samsung-dong, Kangnam-gu, Seoul, Korea

II. Analysis of Distorted Features

To analyze channel distortion effects in the feature domain, linear time-invariant channel is assumed as $x(t) = \sum_{\tau} s(t-\tau)\, h(\tau)$. Here, $s(t)$, $h(t)$, and $x(t)$ are

clean signal, channel impulse response and distorted signal, respectively.

The short-time Fourier transform of distorted speeches at a time frame is given by

$$X(t,f)=\sum_m w(n_t-m)x(m)e^{-j2\pi fm}=\sum_r\sum_m w(n_t-m-r)s(m)e^{-j2\pi fm}h(r)e^{-j2\pi fr} \quad (1)$$

Here, $w(t)$ is Hamming window function, and $n_t(=lt)$ is the sampled time index corresponding to the t_{th} time frame with a sample length l between time frames. By decomposing the global sample index r into time frame index l and local sample index k at a time frame, *i.e.*, $r=ll+k$, eqn. 1 can be rewritten as

$$\begin{aligned}X(t,f)&=\sum_l\sum_k\sum_m w(n_t-m-ll-k)s(m)e^{-j2\pi fm}\cdot h(ll+k)e^{-j2\pi f(ll+k)}\\&=\sum_l\sum_k\underbrace{\left[\sum_m w(n_{t-l}-m-k)s(m)e^{-j2\pi fm}\right]}_{S_k(t-l,f)}\cdot\underbrace{\left[h(ll+k)e^{-j2\pi f(ll+k)}\right]}_{g_k(l,f)}\\&=\sum_l\sum_k S_k(t-l,f)g_k(l,f)\end{aligned} \quad (2)$$

where $S_k(t,f)$ is the Fourier transform of clean speech with a shifted window by k time samples as shown in Figure 1. By introducing $S_k(t,f)$, it is possible to model intermediate feature vectors between adjacent frames.

In general, $h(k)$ is a fast-varying function of k, and the dependency becomes more complicated for $g_k(l,f)$ with the complex exponential term. However, in order to allow small feature changes between adjacent frames, the number of time samples between frame shifts is usually set to a small number. Therefore, $S_k(t,f)$ varies much more slowly over k than $g_k(l,f)$, and may be approximated as a constant within a frame.

Moreover, if $G(l,f)$ represents short-time Fourier transform of channel impulse response corresponding to l_{th} time frame, *i.e.*, $G(l,f)=\sum_k g_k(l,f)$, then long reverberation channel distorts spectral features as follows

$$X(t,f)=\sum_l\sum_k S_k(t-l,f)g_k(l,f)=\sum_l S_0(t-l,f)G(l,f) \quad (3)$$

It can be noticed that acoustic channel distorts each spectral band separately, which are modeled as independent convolutive filters along time frames. Therefore, to compensate for channel distortions, one needs to define deconvolutive filters for each spectral band.

III. Compensation of Channel-distorted Features

It is assumed that there exist some measured data for clean speeches and corresponding distorted speeches with the acoustic environment of interests. Deconvolutive filters are adaptively trained to transform the distorted features

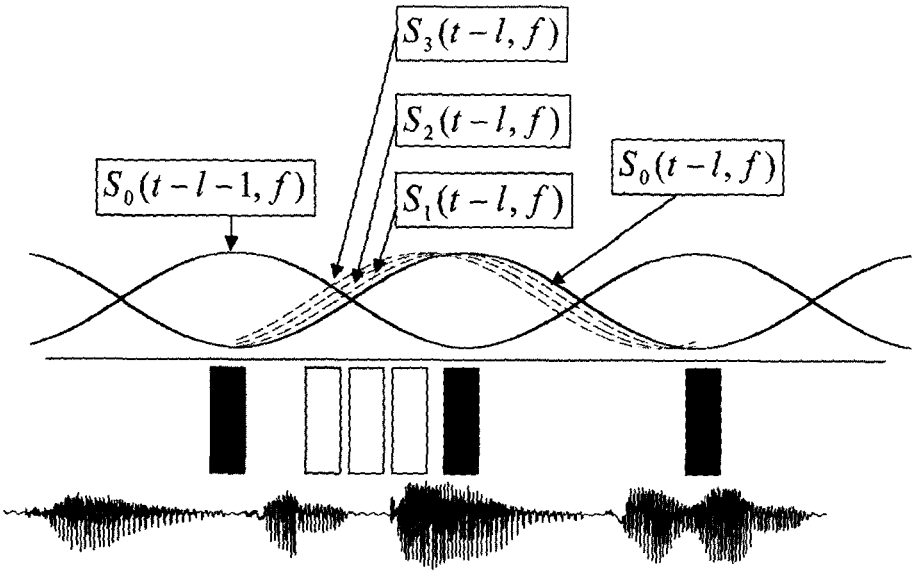

Figure 1. Frame analysis with time-sample-shifted Hamming window.

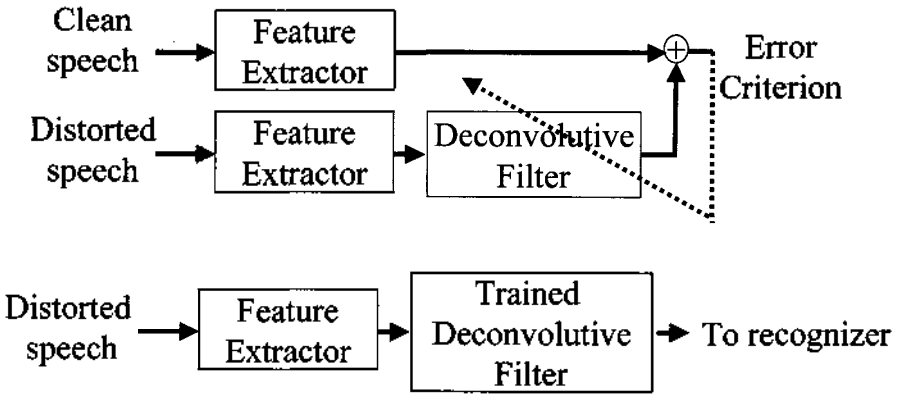

Figure 2. Basic concepts of acoustic channel compensation.

into clean speech features.

Figure 2 illustrates the basic concepts of the adaptive training and feature compensation at test phase. At the training phase of the convolutive filters both the clean speeches and distorted speeches are fed to a same feature extractor, and the filter coefficients are adaptively adjusted to minimize the mean-square-error (MSE). At the test phase the convolutive filters transform the distorted features into clean speech features for better recognition performance.

Popular MFCC features are selected for speech recognition tasks. To obtain MFCC feature from the short-time Fourier transforms of speeches in eqn. 3, one need to calculate magnitude squares for spectral powers, sum over mel-frequency bands, apply logarithmic operations, and perform discrete cosine transforms. Although the linear filters are defined at short-time Fourier transform domain only and nonlinear transforms are required for the other cases, the linear convolutive filters may still be applicable as an approximation. The discrete cosine transforms introduce couplings among frequency bands, and the proposed convolutive filters can not be applied separately for each frequency band. Therefore, we had tested the convolutive filters at complex spectrum, spectral power, mel-frequency spectral power, and log-spectrum domains. At the log-spectrum domain the transformation equation is given as

$$\hat{x}_{ti}^{L} = \sum_{j} w_{ij} x_{(t-j)i}^{L} + \zeta_i = \sum_{j} w_{ij} \log\left[\sum_{k} v_{ik} x_{(t-j)k}^{P}\right] + \zeta_i$$
$$E_t = \frac{1}{2}\sum_{i}\left[y_{ti}^{L} - \hat{x}_{ti}^{L}\right]^2 \quad (4)$$

where x_{ti}^{L} , $\hat{x}_{ti}^{L}$, y_{ti}^{L} denote distorted log-spectrum vector, compensated log-spectrum vector, and clean log-spectrum vector at t_{th} frame and i_{th} band, respectively. v_{ik} is the weight between i_{th} mel-frequency band and k_{th} power-spectrum. w_{ij} is the j_{th} filter coefficient at the i_{th} mel-frequency band, and ζ_i denotes a bias term at the i_{th} mel-frequency band. With zero-mean normalization of feature vectors these bias terms become zero.

Steepest decent algorithm is able to find proper mapper parameter w_{ij} minimizing eqn. 4 as shown below

$$w_{ij}[n] = w_{ij}[n-1] - \eta\frac{\partial E_t}{\partial w_{ij}}$$
$$-\frac{\partial E_t}{\partial w_{ij}} = \sum_{t}\left[y_{ti}^{L} - \hat{x}_{ti}^{L}\right]x_{(t-j)i}^{L} \quad (5)$$

To speed up the parameter learning, we calculated optimal learning rate η_{opt} as follows. Here, we revisited error function defined at eqn. 4 as

$$E_t = \frac{1}{2}\sum_{i}\left[y_{ti}^{L} - \sum_{j}\left[w_{ij}[n-1] - \eta\frac{\partial E_t}{\partial w_{ij}}\right]x_{(t-j)i}^{L} - \zeta_i\right]^2$$
$$= \frac{1}{2}\sum_{i}\left[y_{ti}^{L} - \sum_{j} w_{ij}[n-1]x_{(t-j)i}^{L} - \zeta_i + \eta\left(\sum_{j}\frac{\partial E_t}{\partial w_{ij}}\right)x_{(t-j)i}^{L}\right]^2 \quad (6)$$

Hence, minimization of E_i with respect to η gives following optimal learning rate.

$$\eta_{opt} = \frac{\sum_t \left[\sum_j \left(-\frac{\partial E_i}{\partial w_{ij}} \right) x^L_{(t-j)i} \right] \cdot \left[y^L_{ti} - \left(\sum_j w_{ij}[n-1] x^L_{(t-j)i} + \zeta_i \right) \right]}{2 \sum_t \left[\sum_j \left(-\frac{\partial E_i}{\partial w_{ij}} \right) x^L_{(t-j)i} \right]^2} \quad (7)$$

IV. Experimental Results

To evaluate performance of the proposed convolutive filters, we conducted isolated word recognition (IWR) experiment using speech signal distorted by simulated channels. Korean 50-word database uttered three times by 16 people is used as baseline experiment[5]. Among the total of 2400 utterances, 1350 utterances are used for recognizer training and the other 1050 utterances are used for recognition test. We make four sets of training and test division by random selections in order to effectively utilize small database. Each speech frame is generated with a 30 msec Hamming window and 10 msec shifting. Twenty-three mel-frequency filter banks are used, and 13th-order MFCC features are calculated. Then, each word is normalized to 64 with a trace segmentation algorithm. A multilayer perceptron with 832-50-50 nodes is used for recognition and recognition results are summarized by averaging over 20 trials, *i.e.*, 5 trials with random initialization for each of 4 data sets.

For training of convolutive filters, we made use of 15 second-long speech signals extracted from the train database, which are not included in the four sets of recognizer test speeches. Each convolutive filter incorporates 9-frame delays.

Simulated acoustic channels with three different RT60 reverberation time, *i.e.*, 170 msec, 350 msec, and 700 msec, are generated by image method[4] as shown in Figure 3. These channels are convolved with clean speech database in the time domain, which results in channel-

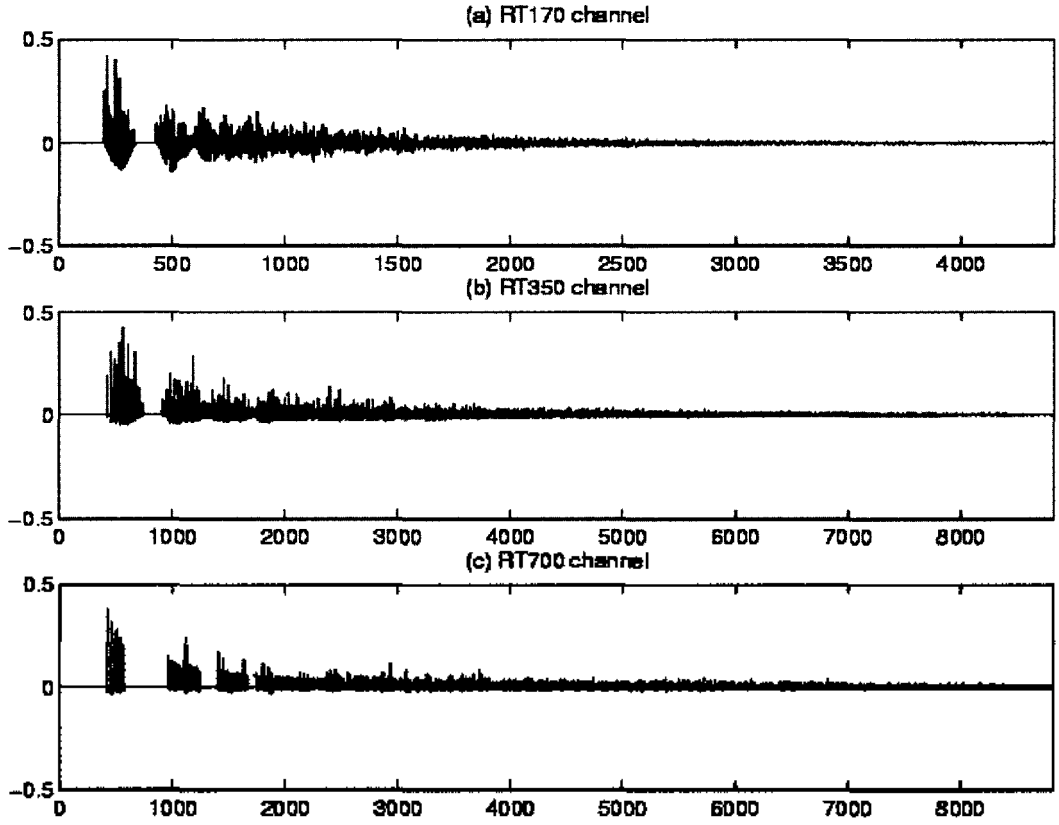

Figure 3. Impulse response for simulated acoustic channel with several reverberation time (RT60) (a) 170 msec (b) 350 msec (c) 700 msec.

Table 1. Recognition rates of isolated words for 3 different acoustic channels.

Algorithms	Test			
	Org	Org-170	Org-350	Org-700
Without compensation	96.2	75.2	65.0	61.4
RASTA		80.0	75.3	69.7
Proposed filter		89.8	86.2	78.1

distorted speech.

Experimental results show that feature transformations at the log-spectrum domain provide best recognition performance. It may come from the fact that the log-spectrum values are most directly connected to the MFCC values. Therefore, results of feature transformation only at the log-spectrum domain are reported here.

Table 1 displays the recognition rates for speech distorted by three simulated channels when the recognizer is trained on clean speech. Baseline results show that mismatched channels degrade recognition rates about 20 percents in light reverberation to 35 percents in heavy reverberation. Although the RASTA algorithm with a fixed convolutive filter for all frequency bands provides enhanced recognition rates, the proposed convolutive filters for each mel-frequency band result in much better recognition rates.

It can be seen from Table 1 that recognition rates are improved by about 10 percents over RASTA algorithm. The improvements come from added complexity of the convolutive filters with available clean-to-distorted speech training data.

Figure 4 represents frequency magnitude responses for the twenty-three trained filters for the 3 acoustic channels. The frequency response of the RASTA filter is also shown at Figure 4(d) for comparison. The average frequency response of trained convolutive filters for the 23 mel-frequency bands is quite similar to that of the RASTA filter for acoustic channels with shorter time delays. However, as the time delay becomes longer, the trained filter results in smaller cutoff frequencies with more variations among frequency bands. It may come from the

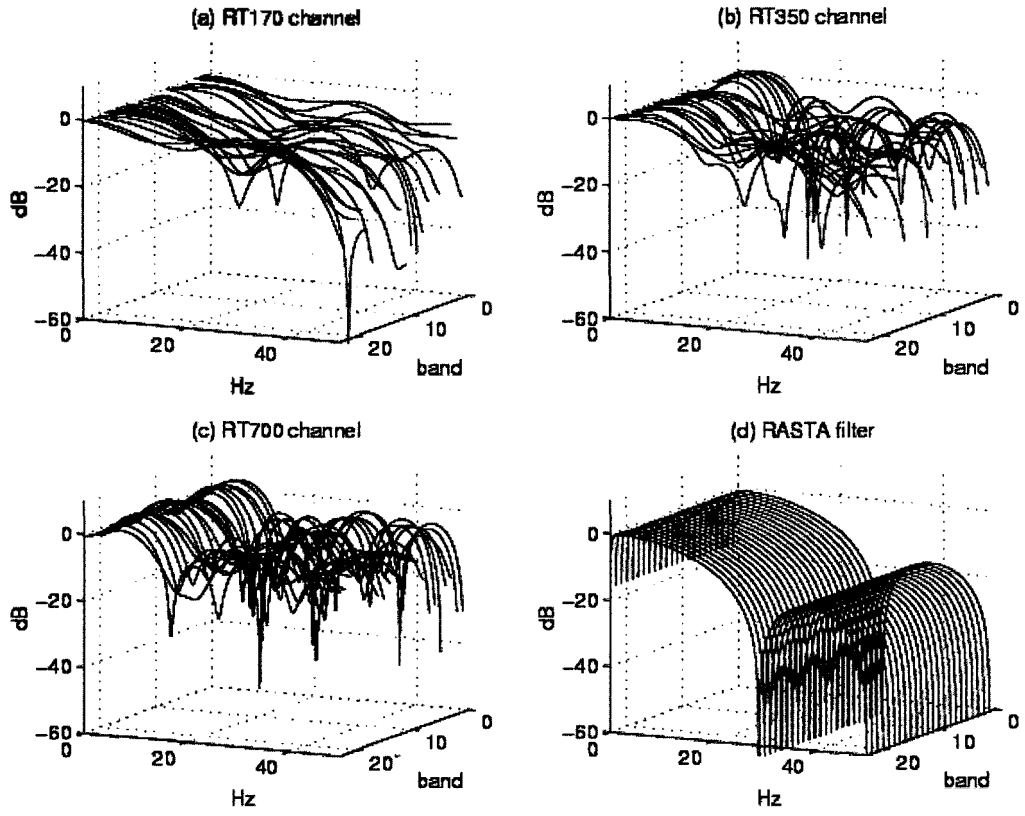

Figure 4. Frequency responses of trained filters and RATSA-filter.

narrower frequency bandwidth of the impulse response function with shorter time delays.

V. Conclusion

In this paper we demonstrated that acoustic channels with long time delays can be compensated for robust speech recognition. By training separate convolutive filters at each mel-frequency band, the developed algorithm successfully compensated acoustic channels up to 700 msec time delays.

Acknowledgment

This research was supported as a Brain Neuroinformatics Research Program by Korean Ministry of Science and Technology.

References

1. H. Hermansky, "Should recognizers have ears?," *Speech Communication*, 25, 3-27, 1998.
2. X. Huang, A. Acero and H.-W. Hon, *Spoken language processing*, *Prentice Hall PTR*, New Jersey, 2001.
3. H. Y. Jung and S. Y. Lee, "On the temporal decorrelation of feature parameters for noise-robust speech recognition," *IEEE Trans. Speech and Audio Processing*, 8 (4), 407-416, 2000.
4. J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of Acoustic Society of America*, 65 (4), 943-950, 1979.
5. D.-S. Kim and S.-Y. Lee and R. M. Kil, "Auditory processing of speech signals for robust speech recognition in real-world noisy environments," *IEEE Trans. Speech and Audio Processing*, 7 (1), 55-69, 1999.

[Profile]

• So-Young Jeong

So-Young Jeong received his B. S., M. S. and Ph. D. degrees in Electrical Engineering from Korea Advanced Institute of Science and Technolgy in 1996, 1998 and 2003, respectively. Since 2003, he has been with Extell Technology Corporation. His research interests are robust speech recognition, acoustic channel modeling and adaptive learning algorithms.

• Sang-Hoon Oh

Sang-Hoon Oh received his B.S. and M.S. degrees in Electrical Engineering from Pusan National University in 1986 and 1988, respectively. He received his Ph.D. degree in Electrical Engineering from Korea Advanced Institute of Science and Technology in 1999. From 1990 to 1998, he was a senior researcher in Electronics and Telecommunications Research Institute(ETRI), Daejeon, Korea. From 1999 to 2000, he was with Brain Science Research Center, KAIST. From 2000 to 2001, he was an R&D manager of Extell Technology Corporation. Since 2002, he has been with the Department of Information Communication Engineering, Mokwon University, Daejeon, Korea. His research interests are supervised/unsupervised learning for intelligent information processing, speech processing and pattern recognition.

• Soo-Young Lee

Soo-Young Lee received his B. S., M. S., and Ph. D. degrees from Seoul National University in 1975, Korea Advanced Institute of Science in 1977, and Polytechnic Institute of New York in 1984, respectively. From 1977 to 1980 he worked for the Taihan Engineering Co., Seoul, Korea. From 1982 to 1985 he also worked for General Physics Corporation at Columbia, MD, USA. In early 1986 he joined the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology, as an Assistant Professor and now is a Full Professor. In 1997 he established Brain Science Research Center, which is the main research organization for the Korean Brain Neuroinformatics Research Program. He was President of Asia-Pacific Neural Network Assembly, and is on Editorial Board for 2 international journals, i.e., Neural Processing Letters and Neurocomputing. His research interests have resided in artificial auditory systems based on biological information processing mechanism in our brain.