IEEK Transactions on Smart Processing and Computing

On-Line Blind Channel Normalization for Noise-Robust Speech Recognition

Ho-Young Jung

Spoken Language Processing Team, Automatic Speech Translation and Artificial Intelligence Research Center, Electronics and Telecommunications Research Institute, Korea hjung@etri.re.kr

* Corresponding Author: Ho-Young Jung

Received November 14, 2012; Revised December 15, 2012; Accepted December 20, 2012; Published December 31, 2012

Abstract: A new data-driven method for the design of a blind modulation frequency filter that suppresses the slow-varying noise components is proposed. The proposed method is based on the temporal local decorrelation of the feature vector sequence, and is done on an utterance-by-utterance basis. Although the conventional modulation frequency filtering approaches the same form regardless of the task and environment conditions, the proposed method can provide an adaptive modulation frequency filter that outperforms conventional methods for each utterance. In addition, the method ultimately performs channel normalization in a feature domain with applications to log-spectral parameters. The performance was evaluated by speaker-independent isolated-word recognition experiments under additive noise environments. The proposed method achieved outstanding improvement for speech recognition in environments with significant noise and was also effective in a range of feature representations.

Keywords: Robust speech recognition, Blind channel normalization, Modulation frequency filtering

1. Introduction

The performance of speech recognition systems has improved dramatically in recent years. On the other hand, they degrade severely in real-world applications, resulting in a mismatch between the training and testing conditions. The major causes of this mismatch are environmental conditions, such as additive background noise and convolutional channel distortion. The control of different acoustic environments is difficult. Moreover, even identical training and testing environments cannot guarantee high performance when the signal-to-noise ratio (SNR) is less than 10 dB. In addition, the background noise can augment the effect of the speaker variability, which is another reason for the performance degradation. To solve this problem, speech recognition methods under adverse conditions have been studied widely, and can be classified into the following three categories. First, the inherently robust feature parameters of the speech signal were used, such as auditory models and modulation frequency filtering approaches. Next, a data compensation method is used to recover the clean speech from the corrupted speech in the feature domain. Unstructured approaches, such as probabilistic optimal filtering (POF) [1], stereo-based piecewise linear compensation for environments (SPLICE) [2, 3] and cepstral vector normalization [4], and structured methods, such as codebook-dependent cepstral normalization (CDCN) [5], vector Taylor series (VTS) [6], and switching linear dynamic model (SLDM) [7, 8], belong to this category. Finally, model compensation techniques adapt model parameters of recognition under consideration of the noise effect [9, 10].

In the above methods, one notable technique is the modulation frequency filtering approach, which reduces the slow-varying noise components in a feature parameter domain. Modulation frequency filtering approaches do not provide spectral parameters that augment the synergy with delta ones. On the other hand, in small vocabulary isolated-word recognition, they are comparable to unfiltered features with a delta parameter for clean speech, and outperform those of noisy speech [11]. Mokbel *et al.* proposed a cepstral mean subtraction (CMS) method to reduce the effects of variations in the telephone line

This work was supported by the industrial strategic technology development program, 10035252, Development of dialog-based spontaneous speech interface technology on mobile platform funded by the Ministry of Knowledge Economy (MKE, Korea).

conditions [12], and Hermansky and Morgan used relative spectral (RASTA) processing to simultaneously cope with additive and convolutional noises [13, 14]. Hanson also tried to apply the modulation frequency filtering approach for recognition of Lombard speech [15]. The modulation frequency filtering approaches do not require prior knowledge of the testing environments, and are more attractive than compensation methods in real system implementation. Using modulation frequency filtering methods, one can suppress the slowly varying components of the frame sequence corrupted by noise. This effect produces a local decorrelation of the frame sequence [11, 16] and provides alternative modeling of some temporal properties of human auditory processing [13]. The principle of the RASTA method originates from the human auditory perception that represents the insensitivity of human hearing to slowly varying auditory stimuli. In addition, the particularly interesting point of modulation frequency filtering approaches is that they emphasize noise-immune parts of the speech signal as in human hearing perception. Although modulation frequency filtering approaches are attractive for noisy speech recognition, the conventional methods have some problems. CMS requires a long-term average calculated from the entire cepstrum vector sequence. RASTA-like filters might be specific to a given task, and use the same form regardless of the noise condition [13, 17]. The data-driven design of a RASTA filter was introduced to optimize to a new task but it needs to be re-designed for a new environment [18-20].

This paper proposes a new data-driven method to design modulation frequency filters for noise-robust feature extraction. Based on a decorrelation criterion, the developed method produces an adaptive high-pass filter in the modulation frequency for each utterance under a range of environments. This is a prime cause of the better performance of the proposed method over existing methods. The proposed method was implemented not as a transformation but as a finite impulse response (FIR) filter to preserve the temporal homogeneity among the frames of a given utterance. Avendano *et al*. also presented linear discriminant analysis (LDA) to feature the time trajectories using FIR filtering for consistency with the ad hoc designed RASTA filter [18]. The proposed method was performed on an utterance-by-utterance basis, and is based on the information-maximization approach used in a blind signal separation [21, 22]. Speaker-independent isolated-word recognition experiments were performed under a range of additive noise conditions to evaluate the performance of the proposed method. The simulation showed that the proposed method outperforms other methods under severe noisy conditions.

The organization of the paper is as follows. Section 2 introduces the basic principle of the information-maximization approach. Section 3 reports the design of a blind decorrelation filter using this principle. Section 4 contains the experimental results for noisy speech recognition, and Section 5 presents the results applying the proposed method to various feature representations. Section 6 concludes the paper.

2. Information-Maximization Approach

Most channel distortions or additive noises can manifest as a slow-varying perturbation introducing temporal dependencies in the feature vector domain. Therefore, by decorrelating the feature vector sequence, one can remove effectively the noise components from feature representation. This is a basic principle of modulation frequency filtering approaches. The blind decorrelation to remove the statistical dependencies is performed by maximizing the joint entropy of the feature vector sequence. Although the correct measure of statistical dependency is the mutual information, maximizing the joint entropy is computationally more efficient than minimizing the mutual information [23]. In addition, for super-Gaussian signals, such as speech signals, the entropy maximization can always minimize the mutual information [21].

When an input sequence U is passed through an invertible monotonic function, $g()$, the probability density function (PDF) $f(Z)$ of an output sequence Z is represented as [24]

$$f(Z) = \frac{f(U)}{\left|\frac{\partial Z}{\partial U}\right|} = \frac{f(U)}{|g'|}, \tag{1}$$

and the joint entropy $H(Z)$ is defined as

$$H(Z) = -\int f(Z) \ln f(Z)\, dZ. \tag{2}$$

From (2), $H(Z)$ is maximized when $f(Z)$ has a uniform distribution, i.e. $g'()$ and $f(U)$ are matched. This is the principle of entropy maximization, and the blind decorrelation using this principle might be implemented, as shown in Fig. 1. In Fig. 1, $g()$ is given as a basic form of the cumulative density function of the input feature sequence, and the linear transform W is learnt to match the PDF of sequence U to $g'()$. This process can be considered unsupervised learning. The non-linear function $g()$ can provide all the higher-order moments besides the second-order moment of simple decorrelation filtering [11], and its invertible property enables the maximization of $H(U)$ from the maximization of $H(Z)$. Therefore, the decorrelated frame sequence U^* can be obtained from a linear transform, W^*, which maximizes $H(Z)$.

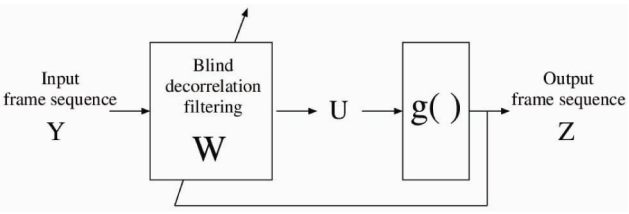

Fig. 1. Blind decorrelation procedure based on the information-maximization approach.

3. Blind Decorrelation Processing

3.1 Linearization of Environmental Model

The environmental model for noisy speech can be represented as

$$Y(\omega) = X(\omega)|H(\omega)|^2 + N(\omega), \tag{3}$$

where $X(\omega)$, $|H(\omega)|^2$, $N(\omega)$ and $Y(\omega)$ denote the power spectrum of clean speech, channel distortion, additive noise and noisy speech, respectively. For a common feature representation, taking the logarithm operator on (3) results in

$$\begin{aligned}\log Y(\omega) &= \log[X(\omega)|H(\omega)|^2 + N(\omega)] \\ &= \log X(\omega) + \log|H(\omega)|^2 + \log\left[1 + \frac{N(\omega)}{X(\omega)|H(\omega)|^2}\right],\end{aligned} \tag{4}$$

and defining x_l, q_l, n_l and y_l instead of $\log X(\omega)$, $\log H(\omega)$, $\log N(\omega)$ and $\log Y(\omega)$ derives the equation,

$$y_l = x_l + q_l + \log(1 + \exp(n_l - x_l - q_l)), \tag{5}$$

where subscript l denotes the log-domain. From (5), the noisy feature vector $\mathbf{y}$ is related to the clean feature vector $\mathbf{x}$, the additive noise, $\mathbf{n}$, and the channel distortion, $\mathbf{q}$, by

$$\mathbf{y} = \mathbf{x} + f(\mathbf{x}, \mathbf{n}, \mathbf{q}), \tag{6}$$

and the function $f(\mathbf{x}, \mathbf{n}, \mathbf{q})$ might be approximated linearly in a piecewise manner, as in a truncated vector Taylor series expansion [6]. Therefore, the channel distortion and additive noise can be represented as an additive slow-varying term in the log-scaled feature domain and be removed fairly well by linear high-pass filtering, such as modulation frequency filtering methods.

3.2 Design of Decorrelation filter

In this Section a procedure was derived to design the filter performing the blind decorrelation in the feature domain. The filter is a type of unsupervised adaptive high-pass filter in the modulation frequency, and is implemented as an FIR form by the information-maximization approach for each utterance. Although an FIR filter requires more coefficients than an infinite impulse response (IIR) filter, the derivation procedure for the coefficients of the FIR filter is much simpler.

In Fig. 1, after FIR filtering with the noisy feature vector sequence, Y, $U(t)$ and $Z(t)$ can be expressed as

$$U(t) = \sum_{k=0}^{K} \omega_k Y(t-k), \tag{7}$$

$$Z(t) = g(U(t)), \tag{8}$$

where t denotes a frame index, and ω_k and K denote the coefficient and order of the filter W, respectively. To apply information-maximization theory, the PDF $f(Z)$ is expressed as (1), and the joint entropy $H(Z)$ is given by

$$\begin{aligned}H(Z) &= -\mathrm{E}[\ln f(Z(t))] \\ &= \mathrm{E}\left[\ln\left|\frac{\partial Z(t)}{\partial Y(t)}\right|\right] - E[\ln f(Y(t))],\end{aligned} \tag{9}$$

where E[·] denotes the expectation operation. Because the second term is not affected by a change in ω_k, the first term was only considered to maximize $H(Z)$ with respect to ω_k.

By taking the gradient of the first term, the gradient descent rule for ω_k can be derived as

$$\begin{aligned}\Delta\omega_k &= \mathrm{E}\left[\frac{\partial \ln|Z'(t)|}{\partial \omega_k}\right] \\ &= \mathrm{E}\left[\frac{1}{Z'(t)}\frac{\partial Z'(t)}{\partial \omega_k}\right], k = 0, \cdots, K\end{aligned} \tag{10}$$

where $Z'(t)$ represents the partial derivative of $Z(t)$ with respect to $Y(t)$ calculated as

$$\begin{aligned}Z'(t) &= \frac{\partial Z(t)}{\partial Y(t)} = \frac{\partial Z(t)\partial U(t)}{\partial U(t)\partial Y(t)} \\ &= g'(U(t))\omega_0,\end{aligned} \tag{11}$$

and its gradient with respect to ω_k is obtained as follows:

$$\frac{\partial Z'(t)}{\partial \omega_0} = g'^{(U(t))} + \omega_0 \frac{\partial g'(U(t))}{\partial \omega_0}, \tag{12}$$

$$\frac{\partial Z'(t)}{\partial \omega_k} = \omega_0 \frac{\partial g'(U(t))}{\partial \omega_k}, \quad k = 1, \cdots, K. \tag{13}$$

Following the steepest descent update rule, ω_k is updated iteratively by

$$\omega_k^{j+1} = \omega_k^j + \eta\overline{\Delta\omega_k}, \quad k = 0, \cdots, K, \tag{14}$$

where j denotes an iteration index and η is a learning rate, respectively. The average $\overline{\Delta\omega_k}$ was used to apply the same filter to all the dimensions of feature vector despite slight different aspects.

Using these derivations for obtaining a blind decorrelation, the adaptive modulation frequency filter and the consequential robust feature extraction were achieved as follows:

- step 1: Obtain an initial estimate for ω_k and initialize sequences $U(t)$ and $Z(t)$ using (7) and (8).
- step 2: With (10), compute the gradient descent rule for ω_k.

- step 3: Update the filter coefficients by (14), and compute the sequences $U(t)$ and $Z(t)$ using them.
- step 4: If the convergence of filter coefficients is not satisfied, go to step 2.
- step 5: Extract the noise-removed feature sequence $U(t)$ based on (7).

Function $g'(U(t))$ was defined to obtain the desired filter. According to the principle of entropy maximization, $g'(U(t))$ should be matched with the PDF of the feature sequence, $f(U)$. Assume that in a log-spectral domain, the clean speech and noise signals follow a Gaussian distribution and that the channel distortion is known. By (4), the resulting PDF $f(Y)$ is clearly non-Gaussian, and linearly transformed $f(U)$ produces the same results. The PDF sometimes has a bimodal form, and its variance is decreased [25]. On the other hand, a Gaussian assumption can still capture a significant part of noisy speech statistics, and some results show how this assumption is effective [25]. This paper considers the activation function $g'(U(t))$ of the following two cases.

3.2.1 Gaussian distribution

The function $g'(U(t))$ was assumed to be

$$g'(U(t)) = Ce^{-U^2(t)}, \tag{15}$$

$$\frac{\partial g'(U(t))}{\partial \omega_k} = g'(U(t))\left[-2U(t)\frac{\partial U(t)}{\partial \omega_k}\right] = -2g'(U(t))U(t)Y(t-k). \tag{16}$$

Therefore, from (10), the learning rules for ω_k are given by

$$\Delta\omega_0 = \mathrm{E}\left[\frac{1}{\omega_0} - 2U(t)Y(t)\right], \tag{17}$$

$$\Delta\omega_k = \mathrm{E}[-2U(t)Y(t-k)] \quad k = 1, \cdots, K. \tag{18}$$

The filter coefficients were obtained by (14), and the noise-removed feature vector sequence U was obtained from (7) using the converged ω_k.

3.2.2 Exponential power distribution

For a more correct approximation, the activation function $g'(U(t))$ is defined as follows:

$$g'(U(t)) = Ce^{-|U|^{\alpha}(t)}, \tag{19}$$

where α is a positive constant. According to the value of α, this function can represent Gaussian and different non-Gaussian distributions. The learning rule is calculated as follows:

First, the gradients of $g'(U(t))$ with respect to ω_k and α are given by

$$\frac{\partial g'(U(t))}{\partial \omega_k} = -\alpha g'(U(t))|U(t)|^{\alpha-1}Y(t-k), \tag{20}$$

$$\frac{\partial g'(U(t))}{\partial \alpha} = -g'(U(t))|U(t)|^{\alpha}\ln(U(t)), \tag{21}$$

and the gradient descent rules for ω_k and α are derived as

$$\Delta\omega_0 = \mathrm{E}\left[\frac{1}{\omega_0} - \alpha|U(t)|^{\alpha-1}Y(t)\right], \tag{22}$$

$$\Delta\omega_k = \mathrm{E}[-\alpha|U(t)|^{\alpha-1}Y(t-k)], k = 1, \cdots, K, \tag{23}$$

$$\Delta\alpha = \mathrm{E}[-|U(t)|^{\alpha}\ln(U(t))]. \tag{24}$$

Following the steepest descent update rule, until the filter coefficients and α converge from the initial values, they are updated by (14) and

$$\alpha^{j+1} = \alpha^{j} + \eta_{\alpha}\overline{\Delta\alpha}, \tag{25}$$

where j denotes the iteration index and η_{α} denotes the learning rate for α, respectively.

3.3 Frequency Response of Implemented Filter

Fig. 2 shows the frequency responses of the conventional modulation frequency filtering approaches and the proposed FIR filter on the modulation frequency,

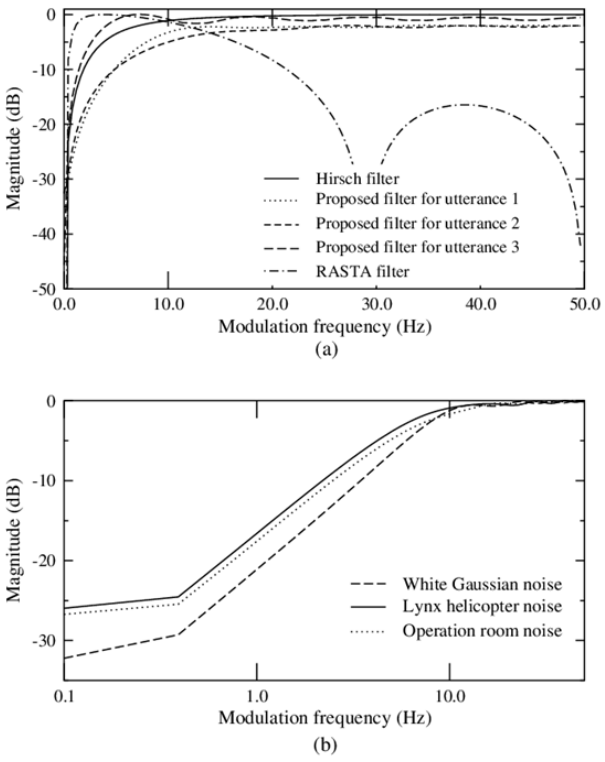

Fig. 2. Frequency response of (a) conventional modulation frequency filtering and proposed FIR filters with Gaussian activation function for three arbitrary utterances, and (b) filters obtained using the proposed method with Gaussian activation function under a range of noise environments.

which describes the temporal variations of the frame sequences. The filters obtained using the proposed method applied the learning rules (17) and (18) to three arbitrary isolated words. Their common attribute is the suppression of low modulation frequencies containing heavy noise components. Although the conventional methods use an identical filter for all cases, the proposed method results in an adaptive high-pass modulation frequency filter for an arbitrary utterance on a particular condition. Fig. 2(b) represents the filters obtained using the proposed method under a range of noise environments with 0 dB SNR. Each filter has a different cut-off frequency, which appears to indicate the different effects of each noise on the modulation frequency.

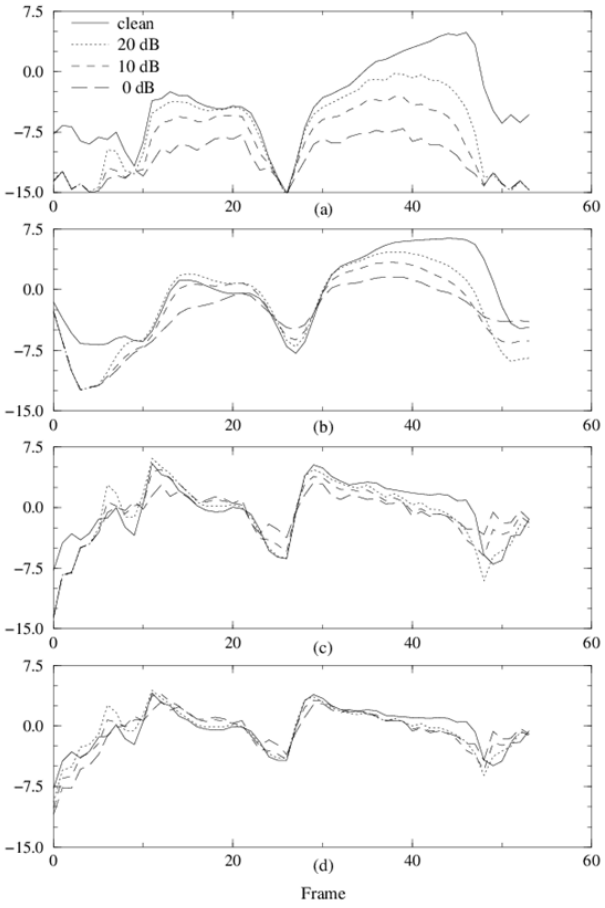

Fig. 3. Temporal trajectories of MFCC C_1 (a) before filtering, (b) after RASTA filter, (c) after the Hirsch filter, and (d) after the proposed method with Gaussian activation function.

4. Performance Evaluation

4.1 Speaker Independent Speech Recognition

The vocabulary consisted of 75 phonetically balanced Korean words that are mutually confusable, and the database consisted of 6750 words spoken by 90 male speakers in a quiet room. The utterances for the 68 speakers were used to form training data, and those from the other 22 speakers were used for the evaluation. The distorted speeches for a channel normalization evaluation were generated by applying the filter used by Hermansky to the clean speech data [13]. In addition, noisy speeches were simulated by adding the noise sources taken from the NOISEX-92 database to the clean speech data.

The feature vectors were extracted on 20 ms speech segments every 10 ms, and each frame consisted of 23 mel-scaled filterbank energies. The components of each frame were normalized by their frame energy and scaled logarithmically. The proposed blind modulation frequency filtering and other approaches were then applied, and 12 mel-frequency cepstral coefficients (MFCC) were extracted using a discrete cosine transform (DCT). The triphone was chosen as the basic unit of recognition, and all the corresponding 271 triphones in the vocabulary were used. Each triphone was modeled using a three-state left-to-right continuous-density hidden Markov model (CDHMM), and one Gaussian mixture with a diagonal covariance matrix was used for each state. All models were trained according to the maximum likelihood criterion, and the segmental k-means procedure was iterated ten times from an initial uniform segmentation for convergence.

Though speaker independent speech recognition, the performance of the proposed method was shown under both channel mismatch and additive noise conditions. Filtering in the logarithmic domain is a solution for convolutional noise, such as channel distortion, speaker variation, etc.. According to the environmental linearization in Section 3,1, however, the proposed method was also implemented to cope with additive noise in the logarithmic domain.

4.2 Experimental Results

Table 1 lists recognition results for the channel distortion condition. The learning rate η was 0.0003, and the average number of iterations for converging was 32.[1] The learning rate was dependent on the range of changes, $\Delta\omega$, and was not a dominant factor on the recognition results. The convergence condition was satisfied when for all filter coefficients, the variation ($= \left|\omega_k^{new} - \omega_k^{old}\right|$) was below the threshold, i.e. 0.0001 in this paper. The order K of the FIR filter performing blind decorrelation was chosen as 9. The time-span related to the temporal correlation among the successive feature vectors is between 30 and 90 ms [26], and $K = 9$ corresponds to a 90 ms time-span. The filter coefficients were initialized as zero except $\omega_0 = 1$, which were updated by learning rules using the Gaussian activation function. In Table 1, the proposed method outperforms the other filtering methods for a channel mismatch condition. Table 2 presents the recognition results for the speech corrupted by white

[1] The computation complexity to obtain the filter coefficients is ((3K+2)N+2)T for each iteration, where K denotes the order of decorrelation filter, N is a feature dimension, and T is the number of frames, respectively. Practically, the computation time is approximately 45ms on a 2.6GHz Xeon processor.

Gaussian noise (WGN). RASTA and Hirsch filters were also applied to the logarithmic domain for a fair comparison with the proposed filter. The proposed method was comparable to other methods above 0 dB SNR and yielded significant improvement at 0 dB SNR. Table 2 shows that the modulation frequency filtering on the logarithmic domain is appropriate for additive noise.

Table 1. Word accuracies (%) of modulation frequency filtering methods for a channel mismatch condition

No channel distortion	Channel distortion			
Baseline	Baseline	RASTA method	Hirsch's method	Proposed method
95.8	69.4	95.9	96.4	98.1

Table 3 shows the effect of modulation frequency filtering including a noise reduction technique. A noise reduction technique was used to compensate for the effects of additive noise, and modulation frequency filtering was then conducted on the logarithmic spectral domain. As a noise reduction technique, the spectral subtraction (SS) method was applied to the linear spectral domain. Modulation frequency filtering with a noise reduction module was valid, and the proposed method outperformed the existing methods under low SNR conditions.

Table 2. Word accuracies (%) of modulation frequency filtering methods with WGN

SNR (dB)	Baseline	CMS	RASTA method	Hirsch's method	Proposed method
clean	95.8	95.8	98.0	97.8	98.2
20	88.1	93.2	95.8	95.8	96.1
10	56.3	77.8	81.4	85.1	85.3
0	7.5	28.5	27.7	33.1	52.6

Table 3. Word accuracies (%) of the modulation frequency filtering methods including a spectral subtraction under WGN conditions

SNR (dB)	SS + CMS	SS + RASTA method	SS + Hirsch's method	SS + Proposed method
20	93.3	96.1	96.0	96.3
10	83.1	85.6	88.7	89.8
0	44.8	45.3	51.5	66.2

Table 4 presents the recognition results for two activation functions. In the case of the exponential power distribution, the learning rate, η_α, was 0.002, and α was initialized as 2. The results show that the exponential power distribution is more effective for noisy speech, and might require a more correct representation for the noisy feature distribution. On the other hand, a slight improvement shows that the Gaussian approximation can capture a part of the noise effect. Fig. 3 compares the temporal trajectories of the cepstral coefficient C_1 before and after modulation frequency filtering for a particular utterance. This shows that modulation frequency filtering methods effectively recover corrupted parts, and the proposed method is the most robust for noise damage. Note that while the dynamic range of feature values is decreased in the stationary regions of trajectories, the transition regions are enhanced. This indicates that a context-dependent recognition model is needed for modulation frequency filtering approaches.

Table 4. Word accuracies (%) of the proposed method when two activation functions are used

SNR (dB) / Activation function	clean	20	10	0
Gaussian	98.2	96.1	85.3	52.6
Exponential power	98.4	96.9	87.9	56.2

Table 5 compares the performance of the proposed method and other methods under Lynx helicopter noise (LYNX) and destroyer operation room noise (DOP). These noises are approximately stationary and contain significant low-frequency components. The proposed method was also effective for these noises and outperformed the other methods at low SNRs. Tables 2-5 show that the proposed method provides outstanding improvement under significant additive noise conditions. This suggests that additive noise acts as a bias component in the logarithmic domain and severe noises are reduced and normalized by adaptive filtering in the logarithmic domain.

Table 5. Comparison of the proposed method and others by word accuracies (%)

Methods / Noise		LYNX			DOP		
		SNR (dB)			SNR (dB)		
		20	10	0	20	10	0
Baseline		92.1	76.5	23.8	92.2	78.0	24.3
CMS		93.6	85.9	47.1	93.1	83.7	37.9
RASTA method		96.2	86.6	42.7	95.2	84.9	33.5
Hirsch's method		96.8	92.1	55.7	96.1	89.5	44.3
Proposed method	Gaussian	96.3	92.3	69.1	95.8	89.9	65.7
	Exponential power	96.7	93.1	69.8	96.0	91.3	65.8

5. Applying to Other Feature Representations

Modulation frequency filtering approaches can be also applied to other feature domains and representations as well as log-sub-band energies. They can be performed directly on the cepstral coefficients related linearly to the logarithmic spectrum. The original RASTA filter is for the perceptual linear predictive (PLP) representation. This Section evaluates the performance of the proposed method in the conventional MFCC and PLP features.

5.1 Cepstral Representation

Because the cepstrum coefficients are more compact and have less redundancy than the log-spectral coefficients, they are accepted in most recognition systems. The compact form reduces the computational load of modulation frequency filtering approaches, and is more effective for improving the noisy speech recognition scores. In Section 4, filtering on a MFCC, which is a final feature parameter of recognizer, might be more desirable.

Table 6 lists the results of modulation frequency filtering in the MFCC domain under the task of Section 4. Because of the linear relationship between cepstral domain and log-spectral domain, the recognition performance of conventional modulation frequency filtering methods was the same as that of Tables 2 and 5. The one difference is that they were processed not on an 18 dimension feature but on a 12 dimension feature. In contrast, the proposed method performs a new learning for the MFCC feature, and can adapt the filter coefficients to a new domain. The proposed method uses rules (17) and (18) for the learning of filter coefficients, and the learning rate η was 0.00006. This method showed a faster processing speed and provided better performance than processing in log-spectral domain.

Table 6. Word accuracies (%) of applying to MFCC domain

Noise	SNR (dB)	Baseline	RASTA method	Hirsch's method	Proposed method
WGN	clean	95.8	98.0	97.8	98.0
	20	88.1	95.8	95.8	96.3
	10	56.3	81.4	85.1	86.5
	0	7.5	27.7	33.1	53.2
LYNX	20	92.1	96.2	96.8	96.7
	10	76.5	86.6	92.1	93.3
	0	23.8	42.7	55.7	69.2
DOP	20	92.2	95.2	96.1	96.1
	10	78.0	84.9	89.5	92.1
	0	24.3	33.5	44.3	64.7

5.2 PLP Representation

To estimate the auditory spectrum, the PLP uses the critical-band spectral resolution, the equal-loudness curve, and the intensity-loudness power law, which are three concepts from the psychophysics of hearing. Fig. 4 presents a block diagram of the PLP method. Modulation frequency filtering is performed on log-subband energies between critical-band analysis and the equal-loudness curve, and intensity-loudness power law might compensate for the rapid spectral-amplitude variation of the transition regions caused by filtering.

PLP analysis was performed on 30 ms speech segments every 10ms, and the feature vector consisted of 12 cepstral coefficients. Modulation frequency filtering approaches were applied to 23 log-subband energies in the analysis procedure. The experimental results are listed in Table 7. The PLP-cepstrum provided better performance than the MFCC, and conventional modulation frequency filtering methods were more effective than in log-spectral and cepstral stages of the MFCC extraction process. The proposed method outperformed the conventional methods at all SNRs, and showed a superior recognition score to processing on other feature domains except for 0dB SNR. In the implementation of the proposed method, 23 log-subband energies of each frame were normalized to the frame total energy for learning stability, and the coefficients of the high-pass modulation frequency filter were obtained by learning rules using the Gaussian activation function. The learning rate was 0.00005.

Table 7. Word accuracies (%) of modulation frequency filtering on PLP representation

Noise	SNR (dB)	baseline	RASTA method	Hirsch's method	Proposed method
WGN	clean	97.9	99.0	98.2	99.1
	20	90.0	95.8	95.6	96.5
	10	61.1	82.9	87.9	88.9
	0	4.0	28.3	35.8	50.4
LYNX	20	94.5	96.5	97.3	98.2
	10	77.1	87.7	92.3	93.5
	0	22.1	43.5	54.8	64.9
DOP	20	94.3	95.5	96.9	97.0
	10	82.5	87.1	90.6	91.4
	0	36.1	42.1	54.1	66.3

From the above results, the proposed method reduces the effect of noise in various feature representations. Modulation frequency filtering approaches are used primarily for parameters related linearly to the log-spectral domain, but they can also be applied to the linear predictive coding (LPC)-cepstrum, in which the sub-band

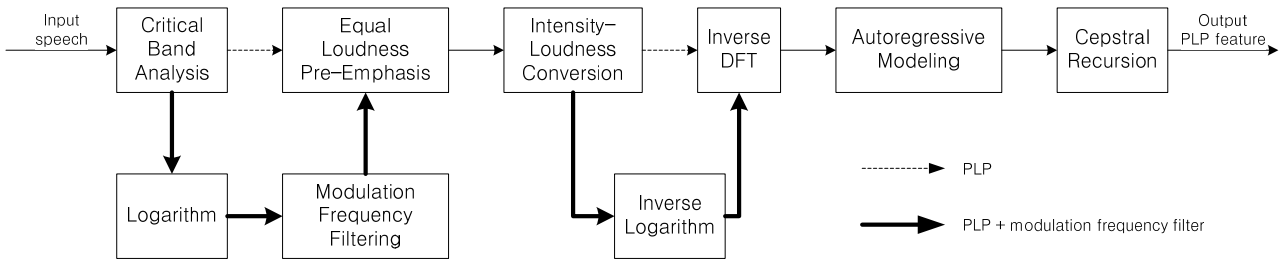

Fig. 4. Modulation frequency filtering in PLP feature analysis.

energy is not available [15]. In view of recognition rate, the proposed method is significant in the PLP representation. Although it is considered the best compromise between performance and computational load, one may require processing of the conventional MFCC domain.

6. Conclusion

This paper proposed a new design method of the modulation frequency filtering approach, which de-emphasizes the slow-varying noise perturbations in the spectral feature domain. This is a data-driven method providing an adaptive high-pass modulation frequency filter for each utterance. The proposed method was based on the local decorrelation of the frame sequence, and the information-maximization theory was used to perform such a decorrelation. This method may also describe some temporal properties of the human auditory system. The proposed method was implemented as an FIR filter form in the log-spectral domain of the MFCC extraction process, and was also applied to the MFCC and PLP representation. The recognition results for speaker-independent isolated-word show that the effect of the proposed method is outstanding under significant noise conditions, and suggests that the proposed method can be performed on various feature representations. In addition, plotting the feature sequences after modulation frequency filtering makes the procedure of channel normalization explicit. For a practical evaluation, the proposed method will need to be applied to real service environments, such as voice search systems, which will be performed in the near future.

References

[1] T. Hasan and Md. K. Hasan, "A Probabilistic Speech Enhancement Filter Utilizing the Constructive and Destructive Interference of Noise," *in Proc. of EUSIPCO*, pp.237-241, 2007. Article(CrossRefLink)

[2] Y. Shinohara, T. Masuko and M. Akamine, "Feature Enhancement by Speaker-Normalized SPLICE for Robust Speech Recognition," *in Proc. of ICASSP*, pp. 4881-4884, 2008. Article(CrossRefLink)

[3] L. Deng, A. Acero, L. Jiang, J, Droppo, and X. D. Huang, "High-Performance Robust Speech Recognition Using Stereo Training Data," *in Proc. of ICASSP*, 2001. Article(CrossRefLink)

[4] L. Buera, E. Lieida, A. Miguel, A. Ortega, and O. Saz, "Cepstral Vector Normalization Based on Stereo Data for Robust Speech Recognition," *IEEE Trans. On Audio, Speech and Language Processing*, Vol. 15, No. 3, pp. 1098-1113, 2007. Article(CrossRefLink)

[5] C. Kim, *Signal Processing for Robust Speech Recognition Motivated by Auditory Processing*, Ph.D Thesis, Carnegie Mellon University, 2010. Article(CrossRefLink)

[6] O. Kalinli, M. L. Seltzer, J. Droppo, and A. Acero, "Noise Adaptive Training for Robust Automatic Speech Recognition," IEEE Trans. on Audio, Speech, and Language Processing, Vol. 18, No. 8, pp. 1889-1901, 2010. Article(CrossRefLink)

[7] N. S. Kim, W. Lim, and R. M. Stern, "Feature Compensation Based on Switching Linear Dynamic Model," *IEEE Signal Processing Letters*, Vol. 12, No. 6, pp. 473-476, 2005. Article(CrossRefLink)

[8] B. Schuller, M. Wollmer, T. Moosmayr, and G. Rigoll, "Speech Recognition in Noisy Environments Using A Switching Linear Dynamic Model for Feature Enhancement," *in Proc. of Interspeech*, pp. 1789-1792, 2008. Article(CrossRefLink)

[9] J. Li, L. Deng, D. Yu, Y. Gong, and A. Acero, "High-Performance HMM Adaptation with Joint Compensation of Additive and Convolutive Distortions via Vector Taylor Series," *in Proc. of ASRU*, pp. 65-70, 2007. Article(CrossRefLink)

[10] M. J. F. Gales and S. J. Young, "Robust Continuous Speech Recognition Using Parallel Model Combination," *IEEE Trans. on Speech and Audio Processing*, Vol. 4, No. 5, pp. 352-359, 1996. Article(CrossRefLink)

[11] H.-Y. Jung and S. Y. Lee, "On the Temporal Decorrelation of Feature Parameters for Noise-Robust Speech Recognition," *IEEE Trans. on Speech and Audio Processing*, Vol. 8, No. 4, pp. 407-416, 2000. Article(CrossRefLink)

[12] Md. M. Rahman, S. K. Saha, Md. Z. Hossain, and Md. B. Islam, "Performance Evaluation of CMN for Mel-LPC based Speech Recognition in Different Noisy Environments," *International Journal of Computer Applications*, Vol. 58, No. 10, pp. 6-10, 2012. Article(CrossRefLink)

[13] H. Hermansky and N. Morgan, "RASTA Processing of Speech," IEEE Trans. on Speech and Audio Processing, Vol. 2, No. 4, pp. 578-589, 1994. Article(CrossRefLink)

[14] H. You and A. Alwan, "Temporal Modulation Processing of Speech Signals for Noise Robust ASR," in Proc. of Interspeech, pp. 36-39, 2009. Article(CrossRefLink)

[15] B. A. Hanson and T. H. Applebaum, "Subband or Cepstral Domain Filtering for Recognition of Lombard and Channel-Distorted Speech," *in Proc. of ICASSP*, Vol. 2, pp. 79-82, 1993. Article(CrossRefLink)

[16] C. Nadeu, D. Macho, and J. Hernando, "Time and Frequency Filtering of Filter-Bank Energies for Robust HMM Speech Recognition," *Speech Communication*, Vol. 34, No. 1-2, pp. 93-114, 2001. Article(CrossRefLink)

[17] H. G. Hirsch, P. Meyer, and H. W. Ruehl, "Improved Speech Recognition Using High-Pass Filtering of Subband Envelopes," *in Proc. of Eurospeech*, pp. 413-416, 1991. Article(CrossRefLink)

[18] Y.-H. B. Chiu, "Minimum Variance Modulation Filter for Robust Speech Recognition," *in Proc. of ICASSP*, pp. 3917-3920, 2009. Article(CrossRefLink)

[19] H. Hermansky, "TRAP-TANDEM: Data-Driven Extraction of Temporal Features from Speech," *in Proc. of ASRU*, pp. 255-260, 2003. Article(CrossRefLink)

[20] J.-W. Hung and L.-S. Lee, "Optimization of

Temporal Filters for Constructing Robust Features in Speech Recognition," *IEEE Trans. on Audio, Speech, and Language Processing*, Vol. 14, No. 3, pp. 808-832, 2006. Article(CrossRefLink)

[21] A. J. Bell and T. J. Sejnowski, "An Information-Maximization Approach to Blind Separation and Blind Deconvolution," *Neural Computation*, Vol. 7, pp. 1129-1159, 1995. Article(CrossRefLink)

[22] A. Hyvarinen and E. Oja, "Independent Component Analysis: Algorithms and Applications," *Neural Networks*, Vol. 13, pp. 411-430, 2000. Article(CrossRefLink)

[23] H. H. Yang and S. Amari, "Adaptive On-Line Learning Algorithms for Blind Separation – Maximum Entropy and Minimum Mutual Information," *Neural Computation*, Vol. 9, No. 7, pp. 1457-1482, 1997. Article(CrossRefLink)

[24] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, 1991. Article(CrossRefLink)

[25] H. Shen, G. Liu, and J. Guo, "Two-Stage Model-Based Feature Compensation for Robust Speech Recognition," *Computing*, Vol. 94, No. 1, pp. 1-20, 2012. Article(CrossRefLink)

[26] H. Bourlard, H. Hermansky, and N. Morgan, "Towards Increasing Speech Recognition Error Rates," *Speech Communication*, Vol. 18, No. 3, pp. 205-231, 1996. Article(CrossRefLink)

Ho-Young Jung is a principle researcher with the Electronics and Telecommunications Research Institute (ETRI), Daejeon, Korea. He received his B.S. degree in electronic engineering from Kyungpook National University, Daegu, Korea in 1993 and M.S. and PhD degrees in electrical and electronic engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 1995 and 1999, respectively. His PhD dissertation was on robust speech recognition. He joined the ETRI, in 1999 as a senior researcher and has been a principle researcher at the Automatic Speech Translation and Artificial Intelligence Research Center since 2010. His current research interests include speech recognition, noise-robust processing, blind signal separation and machine learning. He has published or presented approximately 40 papers in speech recognition. He is a member of the IEEK, KSSS, and IEEE.