Adaptive Channel Normalization Based on Infomax Algorithm for Robust Speech Recognition

Ho-Young Jung

This paper proposes a new data-driven method for high-pass approaches, which suppresses slow-varying noise components. Conventional high-pass approaches are based on the idea of decorrelating the feature vector sequence, and are trying for adaptability to various conditions. The proposed method is based on temporal local decorrelation using the information-maximization theory for each utterance. This is performed on an utterance-by-utterance basis, which provides an adaptive channel normalization filter for each condition. The performance of the proposed method is evaluated by isolated-word recognition experiments with channel distortion. Experimental results show that the proposed method yields outstanding improvement for channel-distorted speech recognition.

Keywords: Robust speech recognition, adaptive channel normalization, RASTA-like filtering, blind decorrelation, information-maximization method

Manuscript received Oct. 16, 2006; revised Mar. 28, 2007.

Ho-Young Jung (phone: +82 42 860 1328, email: hjung@etri.re.kr) is with Embedded S/W Research Division, ETRI, Daejeon, Korea

I. Introduction

The performance of speech recognition systems has been improved dramatically in recent years. However, it is degraded severely in real-world applications, which results in a mismatch between training and testing conditions. The control of different acoustic environments is very difficult, and even identical training and testing condition cannot guarantee a similar performance to the training and testing condition of clean speech when the speech signal is distorted by an unknown channel. In order to solve this problem, speech recognition methods in adverse conditions have been widely studied, and they can be classified into the following three categories. First, inherently robust feature parameters of the speech signal are used, such as auditory models and high-pass approaches. Second, a data compensation method is used to recover clean speech from corrupted speech in the feature domain. Finally, model compensation techniques modify the model parameters of the recognizer using noise estimations.

In the above methods, one notable technique is the high-pass approach, which reduces temporal slow-varying noise components in the feature domain. Hermansky and Morgan proposed relative spectral (RASTA) processing to cope with convolution noise [1], and Hirsch and others used a first-order high-pass filter to reduce the effect of various channel conditions [2]. The high-pass approaches do not require a prior knowledge of testing environments and they suppress slowly varying components of the feature sequence corrupted by noise; therefore, they are more attractive in practical systems than other compensation methods, which have some difficulties in estimating channel characteristics in real environments. They intrinsically cause local decorrelation of the feature sequence [3] and provide an alternative which

models some temporal properties of human auditory processing [1]. Although high-pass approaches are attractive for corrupted speech recognition, conventional methods have some problems. The RASTA and Hirsch filters are specific to a given task, and do not provide an adaptive filter for the condition of each utterance.

In this paper, we propose a new data-driven method to design RASTA-like filters. By performing a blind decorrelation process of the feature sequence, the proposed method can remove slow-varying noise components without long-term statistics and yield a suitable filter for the distorted state of each utterance. This decorrelation is expected to satisfy to some degree the independence assumption of the popular recognizer based on the hidden Markov model (HMM). Our decorrelation filter has been developed as a finite impulse response (FIR) filter with the information maximization algorithm used in blind signal separation [4], and it performs some envelope differentiation according to the tap size of the filter. Therefore, the proposed filter results in a temporally relative spectrum like the RASTA filter and can be a good solution for the adaptability of the RASTA filter. Speaker-independent isolated-word recognition experiments were performed to evaluate the performance of the proposed method. Experimental results showed that the proposed method outperforms conventional methods under severe channel-distorted conditions.

II. Information-Maximization (Infomax) Algorithm

Most channel distortions show up as slow-varying perturbations which introduce temporal dependencies in the feature domain. Thus, by deriving independence from the corrupted feature sequence, one can obtain a feature representation from which the channel distortion is effectively removed. This is our basic principle in presenting an alternative to high-pass approaches. It is realized as a decorrelation filter to remove statistical dependencies using the information-maximization algorithm, which maximizes the joint entropy of the feature 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 [5]. In addition, for super-Gaussian signals such as speech signals, the entropy maximization can always minimize the mutual information [4].

When input sequence Y is passes through invertible monotonic function g, the probability density function (PDF) of output sequence Z is represented in [6] as

$$f(Z) = \frac{f(Y)}{|\partial Z / \partial Y|} = \frac{f(Y)}{|g'|}, \quad (1)$$

where $g' = \partial g(Y) / \partial Y$, and the joint entropy $H(Z)$ is defined as

$$H(Z) = -\int f(Z) \ln f(Z) dZ. \quad (2)$$

As the definition of entropy, $H(Z)$ becomes maximized when $f(Z)$ is a uniform distribution, that is, when g' and $f(Y)$ match. Therefore, maximizing the entropy is transforming an input sequence so that g' and $f(Y)$ have identical distributions. This can be considered an unsupervised learning process and can be realized as blind filtering as shown in Fig. 1. In Fig. 1, g is given as a basis of the cumulative density function of the input feature sequence, and linear filter W is learned in order to match the PDF of sequence U to g'. Non-linear function g is mainly based on the sigmoid function and can lead to higher-order moments as well as second-order moments for the decorrelation object. Additionally, the invertible property of g enables the maximization of $H(U)$ through the maximization of $H(Z)$. Therefore, the decorrelated feature sequence U^* is obtained from linear filter W^*, which maximizes $H(Z)$.

III. Adaptive Blind Decorrelation Filtering

1. Environmental Model

Most acoustic features are based on the log-spectral domain. In this domain, an environmental model for distorted speech is represented as

$$Y(w) = X(w) + H(w), \quad (3)$$

where $X(w)$, $H(w)$, and $Y(w)$ denote the log spectra of clean speech, channel distortion, and distorted speech, respectively. This relation shows that the channel distortion is an additive term in a particular segment of short-time analysis and its temporal characteristic can be approximated to a bias when the channel is slowly varied. Subsequently, the proposed decorrelation filter subtracts the effect of channel distortion in the log-spectral domain and normalizes an unknown channel from the viewpoint of the linear-spectral domain.

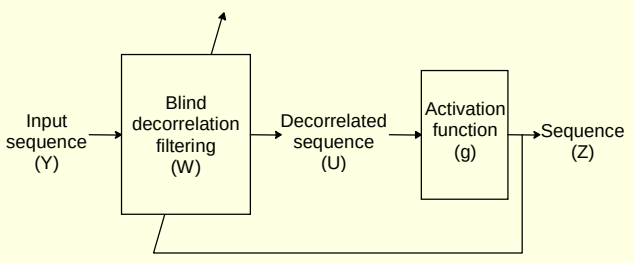

Fig. 1. Block diagram performing blind decorrelation based on an information-maximization approach.

2. Designing the Blind Decorrelation Filter

In this section, we present the procedure from which the design of the filter which performs a temporal blind decorrelation is derived. The filter is a kind of unsupervised adaptive high-pass filter using the information-maximization algorithm. It is realized as an FIR filter which adapts to each utterance. While an FIR filter requires more coefficients than an infinite impulse response (IIR) filter, the derivation procedure for coefficients of the FIR filter is much simpler.

In Fig. 1, after FIR filtering of the distorted input feature vector sequence, the distortion-removed sequence $U(t)$ is represented as

$$U(t) = \sum_{k=0}^{K} w_k Y(t-k), \tag{4}$$

where t is a frame index, and w_k and K denote the coefficient and order of the filter W. An object of information-maximization criterion $Z(t)$ is given by

$$Z(t) = g(U(t)). \tag{5}$$

The algorithm is to maximize the joint entropy $H(Z)$ represented as

$$\begin{aligned} H(Z) &= -E[\ln f(Z(t))] \\ &= E\left[\ln\left|\frac{\partial Z(t)}{\partial Y(t)}\right|\right] - E\left[\ln f(Y(t))\right], \end{aligned} \tag{6}$$

with respect to filter coefficient w_k. Only the first term of (6) is considered because the second term is not affected by the change of w_k.

By taking the gradient of the first term, the gradient descent deviation for w_k is derived as

$$\Delta w_k = E\left[\frac{1}{Z'(t)}\frac{\partial Z'(t)}{\partial w_k}\right], \quad k = 0, \cdots, K, \tag{7}$$

where $Z'(t)$ indicates the partial derivative of $Z(t)$ with respect to $Y(t)$, and its gradient for w_k is obtained as follows:

$$\begin{aligned} \frac{\partial Z'(t)}{\partial w_0} &= g'(U(t)) + w_0 \frac{\partial g'(U(t))}{\partial w_0}, \\ \frac{\partial Z'(t)}{\partial w_k} &= w_0 \frac{\partial g'(U(t))}{\partial w_k}, \quad k = 1, \cdots, K. \end{aligned} \tag{8}$$

Then, according to the gradient descent rule, w_k is iteratively updated by

$$w_k^{j+1} = w_k^j + \eta(\Delta w_k \cdot I), \quad k = 0, \cdots, K, \tag{9}$$

where j is an iteration index, and η denotes a learning rate. We apply the same filter to all the dimensions of the feature vector despite different aspects, and I denotes the unit vector to obtain scalar coefficients.

Now, for a final update rule of filter coefficients, the function $g'(U(t))$ is defined. According to the principle of entropy maximization, it should have the form which can be matched with the PDF of the acoustic feature sequence. In this paper, the activation function $g'(U(t))$ is assumed to be a Gaussian function, and the final learning rules for w_k are given by

$$\begin{aligned} \Delta w_0 &= E\left[\frac{1}{w_0} - 2U(t)Y(t)\right], \\ \Delta w_k &= E[-2U(t)Y(t-k)], \; k = 1, \cdots, K. \end{aligned} \tag{10}$$

Since the PDF of the input utterance is changed by current distortion conditions, the filter coefficients are adapted to each utterance on the defined activation function. The final coefficients are obtained when the deviation of (10) falls below an arbitrary threshold, and the final feature sequence is derived from (4) using the final coefficients.

IV. Isolated Word Recognition Experiments

In our experiments, the vocabulary consisted of 75 phonetically balanced Korean words which are mutually confusable, and the database consisted of 6750 words spoken by 90 male speakers in a quiet room. The utterances of 68 speakers were used to form the training data, and those of the other 22 speakers were used for evaluation. Each utterance was low-pass filtered with a cut-off frequency of 7.2 kHz and was sampled at 16 kHz using a 16-bit A/D converter. The distorted speech data for evaluation was generated by applying the filter used in [1] to the clean speech.

Feature vectors were extracted on 20 ms speech segments every 10 ms, and each frame consisted of 23 mel-scaled filter-bank energies. Then, filter-bank energies were scaled logarithmically, and 12 mel-frequency cepstral coefficients (MFCCs) were extracted by taking a discrete cosine transform (DCT). The proposed filter was applied to two situations. In the first case, the proposed filter was applied to the log filter-bank energies, which are physically meaningful quantities for the environmental model of (3). In the other case, it was applied to the compact MFCCs obtained from log filter-bank energies in order to make the proposed method computationally efficient.

An acoustic model was trained with clean speech, and the triphone was chosen as the basic unit. All the corresponding 271 triphones in the vocabulary were used. Each triphone was modeled by a three-state left-to-right continuous-density hidden Markov model (CDHMM), and one Gaussian mixture with diagonal covariance matrix was used to represent the

distribution of each state.

In the proposed filter, the learning rate was 0.0003, and the average number of iterations and the threshold for converging were 32 and 0.0001, respectively. The order of the decorrelation filter was 9. The time-span related to the temporal correlation among successive feature vectors was between 30 and 90 ms [7], and the order of 9 corresponds to a 90 ms time-span. Note that, while the feature's dynamic range in the stationary regions is decreased after decorrelation filtering, its dynamic range in the transition regions is relatively enhanced. This is commonly shown in high-pass approaches and indicates explicitly that a context-dependent acoustic model is required for them.

Table 1 shows the performance of filtering in log filter-bank energies. As popular methods in the log-spectral domain, the RASTA [1] and Hirsch filters [2] were evaluated for comparison with the proposed method. The proposed filter significantly enhanced recognition performance and outperformed conventional methods.

Table 1. Word correction rate (%) for filtering in log filter-bank energies.

(%)

No filtering	Filtering		
MFCCs	RASTA filter	Hirsch filter	Proposed filter
69.4	95.9	96.4	98.1

Table 2. Word correction rate (%) for filtering in MFCCs.

(%)

No filtering	Filtering			
MFCCs	CMVN	RASTA filter	Hirsch filter	Proposed filter
69.4	92.3	95.6	96.5	98.0

Table 2 represents the recognition rate for filtering in MFCCs. The cepstral mean and variance normalization (CMVN) was also evaluated as a normalization technique in the cepstral domain [8]. Results show that the proposed method works well if the channel distortion is modeled as an additive term regardless of feature representation. For more robustness, the proposed method may be combined with the frequency filtering approach [9]. Moreover, in the cepstral domain, the proposed filter yielded better performance than other filters.

V. Conclusion

In this paper, we proposed a new channel normalization method based on the high-pass approach which de-emphasizes slow-varying noise perturbations in the feature sequence. The proposed method is based on the local decorrelation of the feature sequence. It was developed as an FIR filter for which the coefficients are learned using the information-maximization theory. The method provides a high-pass filter adapted to the noisy condition of each utterance, and presents a new method for the data-driven design of a RASTA-like filter. The experimental results demonstrate that the adaptability of the proposed method enables improved performance.

References

[1] H. Hermansky and N. Morgan, "RASTA Processing of Speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, 1994, pp. 578-589.

[2] H.G. Hirsch, P. Meyer, and H.W. Ruehl, "Improved Speech Recognition Using High-Pass Filtering of Subband Envelopes," *Proceeding of the European Conference on Speech Communication and Technology*, 1991, pp. 413-416.

[3] C. Nadeu, P. Paches-Leal, and B.-H. Juang, "Filtering the Time Sequence of Spectral Parameters for Speaker-Independent CDHMM Word Recognition," *Proceeding of the European Conference on Speech Communication and Technology*, 1995, pp. 923-926.

[4] A.J. Bell and T.J. Sejnowski, "An Information-Maximisation Approach to Blind Separation and Blind Deconvolution," *Neural Computation*, vol. 7, 1995, pp. 1129-1159.

[5] H.H. Yang and S. Amari, "Adaptive On-Line Learning Algorithms for Blind Separation-Maximum Entropy and Minimum Mutual Information," *Neural Computation*, vol. 9, 1997, pp. 1457-1482.

[6] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, 1991.

[7] H. Bourlard, H. Hermansky, and N. Morgan, "Towards Increasing Speech Recognition Error Rates," *Speech Communication*, vol. 18, 1996, pp. 205-231.

[8] C.-P. Chen, K. Filali, and J.A. Bilmes, "Frontend Post-Processing and Backend Model Enhancement on the AURORA 2.0/3.0 Databases," *Proceeding of ICSLP*, 2002, pp. 241-244.

[9] H.-Y. Jung, "Filtering of Filter-Bank Energies for Robust Speech Recognition," *ETRI Journal*, vol. 26, 2004, pp. 273-276.

Ho-Young Jung received the BS degree in electronics engineering from Kyungpook National University, Daegu, Korea, in 1993, and the MS and PhD degrees in electrical engineering from Korea Advanced Institute of Science and Technology, Daejeon, Korea, in 1995 and 1999, respectively. His PhD dissertation was on robust speech recognition. He joined Electronics and Telecommunications Research Institute, Daejeon, Korea, in 1999 as a senior researcher and has belonged to the Speech/Language Information Research Center from 2002. His current research interests include speech recognition, noise-robust processing, blind signal separation, and machine learning. He has published or presented about 30 papers in speech processing. He was the guest reviewer of the IEEE Trans. on Audio and Speech Processing from 2004 to 2005.