Harmonics-based Spectral Subtraction and Feature Vector Normalization for Robust Speech Recognition*

Jounghoon Beh · Heungkyu Lee*** · Ohil Kwon**** · Hanseok Ko****

ABSTRACT

In this paper, we propose a two-step noise compensation algorithm in feature extraction for achieving robust speech recognition. The proposed method frees us from requiring a priori information on noisy environments and is simple to implement. First, in frequency domain, the Harmonics-based Spectral Subtraction (HSS) is applied so that it reduces the additive background noise and makes the shape of harmonics in speech spectrum more pronounced. We then apply a judiciously weighted variance Feature Vector Normalization (FVN) to compensate for both the channel distortion and additive noise. The weighted variance FVN compensates for the variance mismatch in both the speech and the non-speech regions respectively. Representative performance evaluation using Aurora 2 database shows that the proposed method yields 27.18% relative improvement in accuracy under a multi-noise training task and 57.94% relative improvement under a clean training task.

Keywords : Spectral subtraction, Harmonics, Robust speech recognition, Feature vector normalization

1. Introduction

The range of automatic speech recognition (ASR) system used is progressively widening from the laboratory environment to the real environment; for example, for automobile, telematics, PDAs, mobile phones and so on. However, the performance of ASR system is drastically deteriorated by unexpectedly changed noisy environments and channel distortion. Performance deterioration can be attributed to the mismatch problem between training condition and testing condition. To solve this problem, many noise compensation techniques have been introduced in the last three decades. Most of the techniques

* This work was supported by grant No. A17-11-02 from Korea Institute of Industral Technology Evaluation & Planning Foundation.

** Dept. of Electronics and Computer Engineering, Korea University

*** MediaZen Co. Ltd.

**** Hyundai Autonet Co. Ltd.

Coresponding author: Hanseok Ko (hsko@korea.ac.kr)

introduced involve some type of noise compensation in feature extraction because of flexibility rather than the HMM model based strategies. The techniques to solve the mismatch problem can be categorized into two principal approaches. First is the spectral subtractive-type of algorithm performing noise suppression using short-time spectral amplitude, such as spectral subtraction, nonlinear spectral subtraction and the Weiner filter. The other approach is the feature compensation algorithm such as cepstral mean normalization or vector Tayler series. This paper presents a new spectral subtraction scheme based on the observation that even though speech is heavily corrupted by noise, the shape of spectral harmonics of speech can be as well preserved as when speech is not corrupted. In addition, we achieve further improvement in the performance by applying the weighted variance FVN (Feature Vector Normalization), which essentially multiplies an appropriate weighting factor to the variances of the speech and the non-speech regions respectively. This procedure compensates for the unreliable variance estimation due to small sampling.

2. Spectral Subtractive-type Algorithm

2.1 Power spectral subtraction

When speech signal x(n) is corrupted by background additive noise b(n), the corrupted speech can be expressed as follows:

$$y(n) = x(n) + b(n) \tag{1}$$

If speech signal $x(n)$ and noise signal $b(n)$ are assumed to be uncorrelated, in frequency domain the estimate of clean speech signal can be represented as follows:

$$|\hat{X}(k)|^2 = |Y(k)|^2 - |\hat{B}(k)|^2 \tag{2}$$

However, the estimated noise signal should be weighted in order to reduce the musical noise (Berouti et al., 1979). The clean speech is estimated as follows.

$$|\hat{X}(k)|^2 = \begin{cases} |Y(k)|^2 \left(1 - \alpha \dfrac{|\hat{B}(k)|^2}{|Y(k)|^2}\right) & \text{if } |Y(k)|^2 - \alpha|\hat{B}(k)|^2 > \beta|Y(k)|^2 \\ \beta|Y(k)|^2 & \textit{otherwise} \end{cases} \tag{3}$$

Instead of power spectrum, its magnitude spectrum can be made available. With

over-subtraction factor α and floor factor β, this algorithm regularizes the trade-off between noise reduction and residual noise. Note that the enhanced short-time power spectral amplitude $|\hat{X}(k)|^2$ depends on a posteriori SNR (Virag, 1999):

$$SNR_{post} = |Y(k)|^2 / |\hat{B}(k)|^2 \tag{4}$$

2.2 Nonlinear spectral subtraction (NSS)

NSS algorithm is essentially a SNR dependent scheme that in subtraction, a minimal over-subtraction factor is imposed on the high SNR region of frequency and also a maximal over-subtraction factor is imposed on the low SNR region of frequency. NSS algorithm reduces the distortion derived by subtracting the noise spectrum excessively from speech the spectrum as described by the following expression (Lockwood and Boudy, 1992):

$$|\hat{X}(k)|^2 = \begin{cases} |Y(k)|^2 \left(1 - \alpha \dfrac{\Phi(k)}{|Y(k)|^2}\right) & \text{if } |Y(k)|^2 - \alpha\Phi(k) > \beta |Y(k)|^2 \\ \beta |Y(k)|^2 & \textit{otherwise} \end{cases} \tag{5}$$

where

$$\Phi(\rho_{t(k)}, a_t(k), |\hat{B}_t(k)|^2) = a_t(k) F(\rho_t(k), a_t(k), |\hat{B}_t(k)|^2), \tag{6}$$

with

$$a_t(k) = \max_{i-40 \le r \le i} (|\hat{B}_r(k)|^2), \tag{7}$$

$$\rho_t(k) = |\overline{Y_t}(k)|^2 / |\hat{B}_t(k)|^2, \tag{8}$$

$$|\overline{Y_t}(k)|^2 = \lambda_Y |\overline{Y_{t-1}}(k)|^2 + (1 - \lambda_Y)|Y_t(k)|^2, \quad 0.1 \le \lambda_Y \le 0.5 \tag{9}$$

where t is the frame index.

Nonlinear function $\Phi(k)$ is calculated for each frame and can be chosen arbitrarily to implement the notion that relatively greater subtraction is applied to the low SNR region of spectrum and less subtraction to the high SNR region. $\Phi(k)$ is composed of two terms, that is, the over-estimate of noise $a_t(k)$ and the nonlinear function $F(\rho_t(k), a_t(k), |\hat{B}_t(k)|^2)$. Note that $\rho_t(k)$ refers to a biased estimate of the SNR at frame t. $F(\rho_t(k), a_t(k), |\hat{B}_t(k)|^2)$ refers to a nonlinear scheme and the core of its role resides in determining how to change the subtraction factors with respect to $\rho_t(k)$, noise estimate $|\hat{B}_t(k)|^2$ and over-estimate $a_t(k)$.

2.3 Problem formulation

As mentioned before, we can say that the subtraction factor α provides a compensation for this imperfect estimation of noise and the floor factor β guarantees a non-zero value of each subtracted spectrum derived by over-subtraction. In general, the spectral amplitude of speech components (e.g. spectral harmonics) is greater than that of noise or the sidelobes among harmonics so that the existing algorithms are rather suitable for reasonably high SNR (about 15~20 dB) noisy speech cases. However, it appears that in the case of low SNR environments (especially 0~10 dB), such that the noise level is close in amplitude to that of speech, there occurs an unnecessarily subtracted speech region or less subtracted the noisy region in noisy speech spectrum. Consequently, instead of the mundane use of SNR as a criterion for subtraction, we need a new and better measure for activating the subtraction procedure. In particular, the subtraction over spectrum method requires a more accurate measure than mere SNR in order to apply the subtraction rule, which is selective to speech-dominant region of spectrum vs. noise-dominant region of spectrum.

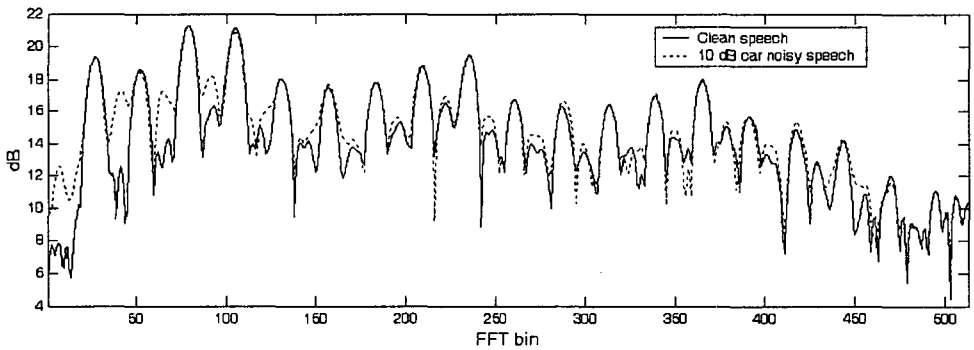

Fig. 1. Example spectrum of a speech frame (pronunciation /oh/ by female)

3. Harmonics-based Spectral Subtraction (HSS)

3.1 Speech dominant region vs. noise dominant region

In speech, it is observed that the voiced speech segment has peaks positioned periodically in spectrum due to the vibration of vocal cords. These points are very critical in speech sound perception. Fig. 1 illustrates this phenomenon with a sample spectrum of speech frame capturing the pronunciation /oh/ in one utterance contained in Aurora2 corpus. Note that at the peaks, their amplitudes are far greater than the amplitudes at the

points between adjacent small peaks, or sidelobes. Also, it is observed that the degree of corruption by the noise in the peaks is not as much as the degree of corruption at the points over the sidelobes. From this observation, it can be deduced that in the speech spectrum, the speech-dominant regions exist over or near the peaks and the noise-dominant regions exist over or near the sidelobes. However, the illustration as displayed should not be mislead. It does not show, for example, that the speech is more corruptable in some frequency regions. Fig. 1 illustrates that the degree of corruption in sidelobes can be greater than that of speech harmonics.

3.2 Peak points detection and segmentation

It is known that the spectral harmonics are located at the points of multiples of fundamental frequency (Hess, 1983). First, an using autocorrelation function, the indices of local maxima in autocorrelation values are found. Among those indices, we find the fundamental frequency f_0 by taking the reciprocal of the index that represents the maximum value. Second, by applying the nonlinear smoothing method (Hess, 1983), f_0 is modified to better reflect the true fundamental frequency. This algorithm works well even in high degree noisy environment. Autocorrelation function is expressed as follows (Rabinar and Schafer, 1978):

$$\phi(\tau) = \frac{1}{N} \sum_{n=0}^{N-1} x(n)x(n+\tau) \tag{10}$$

We then establish the index as k_0 for the frequency bin that corresponds to f_0. Using k_0, the peak points (harmonics) h_l are determined as $h_1 = k_0$, $h_2 = 2k_0$, $h_3 = 3k_0$, ..., $h_L = Lk_0$ where L is the index of the last component of harmonics. Using this procedure, we determine all harmonics in the spectrum of the input frame, which is assumed to be a speech-dominant region. We illustrate the validity of the proposed method with spectrograms and the result of peak point detections in Fig. 2. For the purpose of implementing the proposed scheme (Section 3.), the frequency axis is divided into several non-overlapping bands in the following form.

$$[1,\ h_1], \left[h_1,\ h_1 + \frac{k_0}{2}\right], \left[h_1 + \frac{k_0}{2},\ h_2\right], \ldots, \left[h_L,\ \frac{FFTorder}{2} + 1\right] \tag{11}$$

3.3 Voice activity detection (VAD) and noise estimation

Note that in each input frame, whether it is the speech frame or not, the fundamental

frequency is calculated. As a result, the value of autocorrelation is made available at each frame. To begin with, we calculate the ratio of this value over the frame energy, since this ratio appears to be a more effective measure, to distinguish the speech region from the non-speech region, than the mundane energy measure. Then, two separate procedures are employed for robust detection of speech vs. non-speech region. First, a smoothing procedure is carried out. If this value is greater than the threshold $1.4th_{t-1}$ defined by the following expressions, the relevant frame is assigned to be a 'speech frame'.

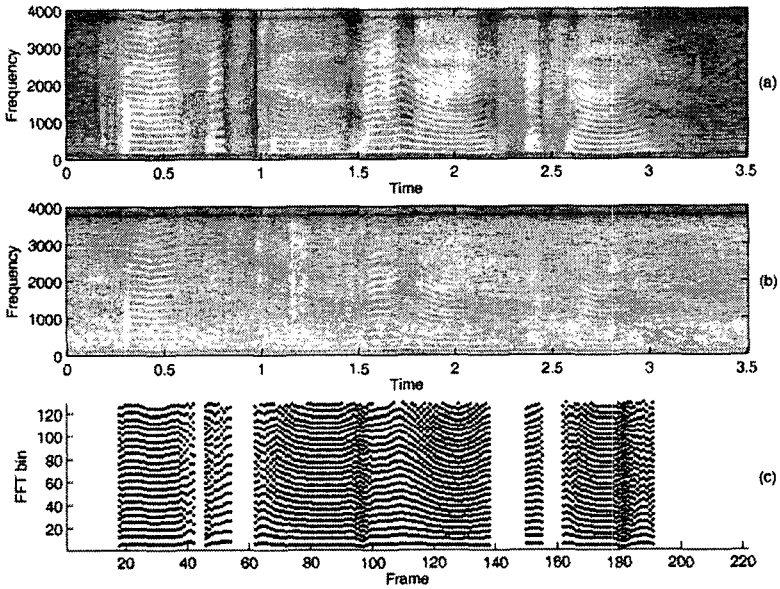

Fig. 2. (a) Spectrum of the pronunciation 'five, six, two, nine, nine, six, nine' by female, (b) Corrupted version by 5 dB car noise, (c) Result of the harmonics detection

$$\text{If } \frac{\phi(\tau')}{E_t} > 1.4th_{t-1} \text{ then 'speech frame' and}$$

$$th_t = 0.95th_{t-1} + 0.05\frac{\phi(\tau')}{E_t}, \tag{12}$$

else 'non-speech frame' and

$$th_t = th_{t-1} \tag{13}$$

Note that th_t indicates the threshold of the tth frame, E_t indicate tth frame energy and $\phi(\tau')$ is the autocorrelation value corresponding to the f_0 of relevant frame.

Secondly, for frames determined as 'speech frame', if the detected fundamental

frequency is not in 50~600 Hz, then it is concluded as 'non-speech frame' once again. For the noise estimation, we assume that any starting input speech is followed by a silence or a background noise segment corresponding to 10 frames (about 100 ms). Then, if the input frame is determined as 'non-speech frame', the estimated noise spectral magnitude is updated as follows:

$$|\widehat{B}_t(k)|^2 = \lambda_B |\widehat{B}_{t-1}(k)|^2 + (1-\lambda_B)|Y_t(k)|^2,\ 0.65 \le \lambda_B \le 0.998 \tag{14}$$

where t is the frame index and k is the index of FFT bins and we use $\lambda_B = 0.95$.

3.4 Spectral subtraction

3.4.1 Speech frame

In order to implement the proposed scheme, we design the following simple linear function.

If $k \in [1,\ h_1]$ or $k \in \left[h_{l-1} + \frac{k_0}{2},\ h_l\right]$, then

$$\gamma(k) = \frac{\alpha_{\min} - \alpha_{MAX}}{h_l - \left(h_{l-1} + \frac{k_0}{2}\right)}(h_l - k) + \alpha_{MAX}, \tag{15}$$

$$\delta(k) = \frac{\beta_{MAX} - \beta_{\min}}{h_l - \left(h_{l-1} + \frac{k_0}{2}\right)}(h_l - k) + \beta_{\min}. \tag{16}$$

If $k \in \left[h_l,\ h_l + \frac{k_0}{2}\right]$ or $k \in \left[h_L,\ \frac{FFTorder}{2} + 1\right]$, then

$$\gamma(k) = \frac{\alpha_{MAX} - \alpha_{\min}}{\left(h_l + \frac{k_0}{2}\right) - h_l}(h_l - k) + \alpha_{\min}, \tag{17}$$

$$\delta(k) = \frac{\beta_{MAX} - \beta_{\min}}{\left(h_l + \frac{k_0}{2}\right) - h_l}(h_l - k) + \beta_{MAX}. \tag{18}$$

where $l = 1, 2, \ldots, L$ and L is the index of te last component of harmonics.

$\gamma(k)$ applies the maximum over-subtraction factor to the midpoint at each component. Then, the over-subtraction factors of points existing in the interval between those points are interpolated linearly. In contrast, $\delta(k)$ applies the maximum floor factor to each

harmonic component and the minimum floor factor to the midpoint at each component. Then, other floor factors between these points are generated by linear interpolation.

In this procedure, we set $\alpha_{MAX}=8$, $\alpha_{\min}=1$, $\beta_{MAX}=0.15$ and $\beta_{\min}=0.05$. In the case of operation with FVN, we use $\alpha_{MAX}=2$, $\alpha_{\min}=1$, $\beta_{MAX}=0.3$ and $\beta_{\min}=0.1$. Finally, the proposed subtraction rule can be represented as follows:

$$|\hat{X}(k)|^2=\begin{cases}|Y(k)|^2\left(1-\gamma(k)\dfrac{|\hat{B}(k)|^2}{|Y(k)|^2}\right) & \text{if } |Y(k)|^2-\gamma(k)>\delta(k)|Y(k)|^2\\ \delta(k)|Y(k)|^2 & \textit{otherwise}\end{cases} \tag{19}$$

3.4.2 Non-speech frame

We apply the subtraction rule in the same way as conventional methods using Eq. 3 with α_{MAX} and $\beta_{\min}$.

4. Feature vector normalization (FVN)

4.1 Conventional FVN

This method assumes that the feature vector goes through an affine transformation due to channel distortion and additive background noise. So, if we can get these transform matrix **A** and bias component **b**, we can obtain the feature vector uncontaminated, by inversely transforming the noisy feature vector.

$$\hat{\mathbf{x}} = \mathbf{A}^{-1}\mathbf{y}-\mathbf{b} \tag{20}$$

where **b** is noisy feature vector and $\hat{\mathbf{x}}$ is clean feature vector. Consequently, FVN (Viikki and Laurila, 1998) enables the distribution of feature vectors in both conditions to be similar to each other so that the mismatch problem can be solved. We illustrate the concept of FVN in 2 dimensions in Fig. 2. If we assume one more thing, in that each component of feature vectors is uncorrelated and has Gaussian distribution, it simply becomes the normalization procedure of single Gaussian random variable.

$$\hat{x}_C(i)=s_C^{-1}\cdot(y_C(i)-m_C(i)), \tag{21}$$

where C means cepstrum and $\hat{x}_C(i)$ is the ith component of normalized version of the input feature vector **y**. Note that $m_C(i)$ and $s_C(i)$ indicate mean and standard deviation

for each feature vector component i, respectively.

The normalization coefficients $m_C(i)$ and $s_C(i)$ are typically calculated as follows:

$$m_C(i) = \frac{1}{N} \sum_{t=1}^{N} y_{C,t}(i) \tag{22}$$

$$s_C(i) = \sqrt{\frac{1}{N} \sum_{t=1}^{N} (y_{C,t}(i) - m_C(i))^2} \tag{23}$$

where N is the value of frame length, which corresponds to an incoming utterance to be recognized in experiments. The mean value becomes reliable when it is calculated for a long period, say, over 2 seconds.

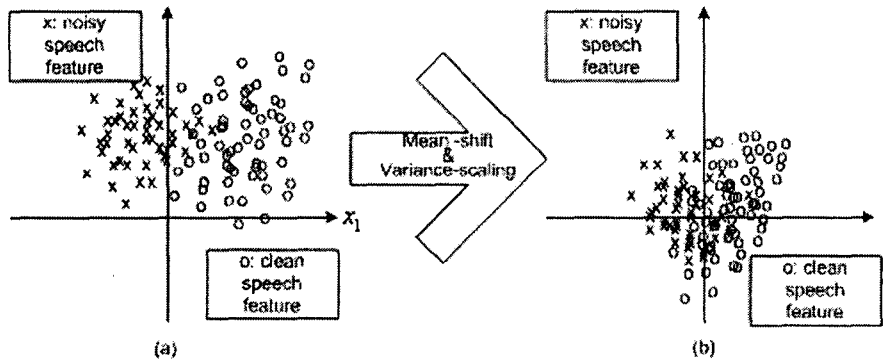

Fig. 2. Concept of FVN illustrated by feature vector $\mathbf{X} = [x_1\ x_2]^T$ (a) each distribution is separated (b) each distribution is located closely by FVN

4.2 Problem formulation

There is a mismatch problem when applying the inverse transform due to the difference in mean and variance between the speech region and the non-speech region, especially in noisy environments. To resolve this mismatch, an accurate mean and variance estimator during even very short periods should be provided for each region of the input signal. However, our preliminary experiments have shown that the estimated mean and variance using short periods are unreliable and using prior information of noisy speech data, is not helpful in resolving the mismatch problem. Furthermore, in the case of mean, the degree of difference is much more severe than in the case of variance. As a result, we modify the conventional FVN by applying a multiplying factor to correct the variance while the mean is fixed. The motivation for using this procedure is that variance is highly dependent on the mean and badly estimated mean value degrades the overall performance.

4.3 Weighted variance FVN

Assume that $s^2_{C,SP}(i)$ and $s^2_{C,NO}(i)$ denote the variance of speech frames and non-speech frames respectively. Then,

$$N \cdot s^2_C(i) = N_{SP} \cdot s^2_{C,SP}(i) + N_{NO} \cdot s^2_{C,NO}(i) \tag{24}$$

where N_{SP} and N_{NO} are the numbers of speech frames and non-speech frames respectively. We can rewrite Eq. 24 in the following simple form:

$$s^2_{C,SP}(i) = s^2_C(i) \cdot \theta_{SP} \tag{25}$$

$$\theta_{SP} = \frac{1}{N_{SP}} \left(\frac{N - s^2_{C,NO}(i)}{s^2_C(i)} \right) \tag{26}$$

Also, in the case of non-speech region,

$$s^2_{C,NO}(i) = s^2_C(i) \cdot \theta_{NO} \tag{27}$$

$$\theta_{NO} = \frac{1}{N_{NO}} \left(\frac{N - s^2_{C,SP}(i)}{s^2_C(i)} \right) \tag{28}$$

Using Eq. 21 and 25~28, the normalized components in each region are as follows:

$$\hat{x}_{C,SP}(i) = s^{-1}_{C,SP} \cdot (y_{C,SP}(i) - m_C(i)) \tag{29}$$

$$\hat{x}_{C,NO}(i) = s^{-1}_{C,NO} \cdot (y_{C,NO}(i) - m_C(i)) \tag{30}$$

Since each variance calculated with only a single utterance recognized is unreliable due to the small quantity of data, we multiply the standard deviation by a constant $\theta_{SP} = 1.96$ and $\theta_{SP} = 1.44$ determined empirically. As a result, the variance used in Eq. 21 has a weighted form as shown by Eq. 29 and 30. Note that we apply Eq. 29 and 30 for testing only. Speech/non-speech discrimination is applied using enhanced energy obtained by HSS and Eq. 12 and 13.

5. Experiments

5.1 Experimental conditions

All experiments are conducted by Aurora 2 evaluation procedure and under continuous digits recognition tasks. Test sets are reproduced using TIDigits of which the entire speech data are downsampled to 8 hKz and various realistic noises are added artificially. Feature vector order is 39 and composed of 13 order static MFCC (c1~c12+log energy), its derivatives and accelerations. For comparison, a spectral subtraction algorithm (Berouti et al., 1979) and a nonlinear spectral subtraction algorithm (Lockwook and Boudy, 1992) is evaluated.

5.2 Experimental results

'Avg' denotes the mean value of word accuracy over 0~20 dB. From Table 2, the average word accuracy of the proposed HSS method shows that it is superior to other spectral subtraction approaches.

Table 1. Summary of overall word accuracies on Aurora2 mis-matched training/testing condition(%)

	Baseline	SS	NSS	HSS
Baseline	60.06	77.89	78.20	80.59
CMN	71.16	78.56	78.73	82.00
FVN	77.56	79.96	80.03	80.70
WVFVN	80.33	82.84	82.47	83.41

In Table 1, each abbreviations are defined as follows:

Table 2. Word error rate(%) of 'HSS+WVFVN'

Aurora2 Word Error Rate				
Training mode	Set A	Set B	Set C	Overall
CMN	9.07	9.80	10.25	9.60
FVN	16.53	15.81	18.25	16.59
WVFVN	12.08	12.08	14.25	13.09

Table 3. Relative improvements(%) of 'HSS+WVFVN'

Aurora2 Relative improvements				
Training mode	Set A	Set B	Set C	Overall
CMN	20.76	31.24	31.91	27.18
FVN	51.32	68.23	50.62	57.94
WVFVN	36.04	49.74	41.27	42.56

- 'SS': general power spectral subtraction.
- 'NSS': nonlinear spectral subtraction.
- 'HSS': harmonics-based spectral subtraction.
- 'CMN': cepstral mean normalization.
- 'FVN': feature vector normalization.
- 'WVFVN': weighted variance feature vector normalization.

In Table 1, rows represent the frequency domain approach and columns denote the feature domain approach and their combined approaches. Each value is calculated as the average of word accuracy over 0~20 dB. In Tables 2 and 3, we also report the improved results of a joint approach (HSS+WVFVN) involving the proposed algorithms using Aurora 2 database reporting methodology.

5.3 Discussion

A notable advantage of the proposed scheme is that it does not require an exact SNR estimate in various noise conditions. If a noise characteristic is stationary and each magnitude of harmonics is not masked by noise, SS technique itself may work alone and NSS would further improve the performance. However, real-life conditions are not all that cooperative. In the case of SS and NSS, due to the fact that the non-stationary characteristic of noise *a posteriori* and a biased SNR are estimated inaccurately, there may be a distortion in the shape of harmonics by subtracting noise magnitude more the in the harmonics region and less in the sidelobes region. As a result, this distortion will lead directly to poor performance in the ASR system. However, HSS also has a remedial capability that would correct the described deficiency. The key to improving the performance of HSS is how exactly we can find the harmonics corresponding to the spectrum.

Proposed WVFVN is compared to the widely used CMN and FVN. FVN technique has an advantage over other feature domain compensation techniques such as RATZ and VTS (Moreno, 1996) in that it softens the requirement of prior information about the distribution of the noisy speech feature. The reason is that it brings both the noisy speech feature and

clean speech feature to the same normal distribution. In short, it reduces the difference between training and testing condition. However, there is another mismatch problem due to the fact that the mean and the variance of noisy speech region are not the same as those of the non-speech region. Nevertheless, in FVN technique, the same normalization parameter is applied to both region. For solving this problem we simply apply weighting constants to the variances of both regions. Consequently, both trajectories become more similar to each other, compared to the ones derived from FVN.

6. Conclusions

The proposed method has two strategies to solve a mismatch problem. First, HSS is applied in the frequency domain. HSS is an efficient spectral subtraction scheme focused to specifically low SNR noisy environments by distinguishing the speech-dominant segment from the noise-dominant segment in the speech spectrum. Consequently, we let the shape of the spectral harmonics be preserved more clearly in noisy environments, which are very critical in speech sound perception. We then improve the overall performance with FVN. FVN normalizes the distribution of each component of the feature to the normal distribution so that it compensates biased mean and reduces the variance of noisy speech. Also, in implementation, the prior noise information is made nonessential in a simple, but effective method. In addition to the FVN, we multiply the variance by a weighting factor to compensate for the mismatch of variance between the speech region and the non-speech region. Representative experiments confirm the superior performance of the proposed method to conventional methods. Each method can be used in tandem with conventional methods.

References

Beh, J., Ko, H., 2003a. A novel spectral subtraction scheme for robust speech recogition: spectral subtraction using spectral harmonics of speech. Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing, I648-I651.

Beh, J., Ko, H., 2003b. Spectral subtraction using spectral harmonics of speech for robust speech recognition in car environments. Lecture Notes in Computer Science, Vol. 2660, 1109-1116.

Berouti, M., Schwartz, R., Makhoul, J., 1979. Enhancement of speech corrupted by additive noise. Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing, 208-211.

Boll, S. F., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on Acoustics, Speech and Signal Processing 27 (2), 113-120.

Hess, W., 1983. Pitch Determination of Speech Signals, Springer-Verlag Berlin Heidelberg New York Tokyo.

Lockwood, P., Boudy, J., 1992. Experiments with a nonlinear spectral subtractor (NSS), hidden Markov models and the projection, for robust speech recognition in cars, Speech Communication 11, 215-228.

Moreno, P.J., 1996. Speech recognition in noisy environments, Ph.D. Dissertation.

Viikki, O., and Laurila, K., Cepstral domain segmental feature vector normalization for noise robust speech recognition, Speech Communication 25, 133-147.

Rabiner, L., Schafer, R., 1978. Digital Processing of Speech Signals, Prentice-Hall.

Virag, N., 1999. Single channel speech enhancement based on masking properties of the human auditory system, IEEE Transactions on Speech and Audio Processing 7 (2), 126-137.

Received: February 10, 2004
Accepted: March 10, 2004

▲ Jounghoon Beh
Department of Electronics and Computer Engineering, Korea University
5Ka-1, Anam-dong, Sungbuk-ku, Seoul, 136-701, Korea.
Tel: +82-2-927-6115, Fax: +82-2-3291-2450
E-mail: jhbeh@ispl.korea.ac.kr

▲ Heungkyu Lee
Department of Visual Information Processing, Korea University,
MediaZen Corporation
5Ka-1, Anam-dong, Sungbuk-ku, Seoul, 136-701, Korea
Tel: +82-2-953-8003(211), Fax: +82-2-923-8830
e-mail: hklee@ispl.korea.ac.kr, hklee@mediazen.co.kr

▲ Ohil Kwon
Hyundai Autonet, Core Technology Team.
San 136-1, Ami-ri, Bubal-eub, Ichon-si, Kyoungki-do, 467-701, Korea
Tel: +82-31-639-7817, Fax: +82-31-639-6695
E-mail: koi@haco.co.kr

▲ Hanseok Ko
Dept of Electronics and Computer Engineering, Korea University
5Ka-1, Anam-dong, Sungbuk-ku, Seoul, 136-701, Korea
Tel: +82-2-3290-3239, Fax: +82-2-3291-2450
E-mail: hsko@korea.ac.kr