Acoust. Sci. & Tech. **32**, 3 (2011)

PAPER

Approach of features with confident weight for robust speech recognition

Ge Lingnan*, Katsuhiko Shirai† and Akira Kurematsu‡

Graduate School of Fundamental Science and Engineering, Waseda University, Ookubo 3–4–1, Shinjuku-ku, Tokyo, 169–8555 Japan

(*Received 17 February 2010, Accepted for publication 1 November 2010*)

Abstract: The enhancement of speech has become one of the focuses of automatic speech recognition (ASR) development. In recent studies, the missing feature approach (MFA) has been proved to be a suitable method. However the hard mask decision in the MFA is mostly a rough binary classifier on the basis of a certain threshold value that could cause a failed decision of reliability and result in a signal screening risk. As improvements of the hard mask the effectiveness of soft masks, including soft mask works with a Bayesian classifier, attempt to compensate the loss of real speech in the hard mask decision by discovering the probability density function (*p.d.f.*) of the unreliable feature component. Unfortunately, this is a very difficult task because of the overlap of at least two complex random processes. The sigmoid function suggested by some soft masks is not a reasonable *p.d.f.* In this paper, we provide an analysis of the confident degree of a feature component in a subband based on four criteria and then propose four types of confident weight (CWs). Based on CWs, we introduce four classes of approaches of feature with confident weight (AFCWs), which estimate the confidence degree of each feature vector simply and efficiently, describe the effect of noise in a rigorous manner, and eliminate the risk of selecting thresholds and the difficulty of finding a joint *p.d.f.* of reliable and unreliable components. Experimental results have shown that the proposed approaches improve the performances of ASR systems even in an adverse environment.

Keywords: Missing feature, Robust speech recognition, Confidence weighting, Cepstrum, Hidden Markov model

PACS number: 43.60.Mn [doi:10.1250/ast.32.92]

1. INTRODUCTION

Speech recognition systems perform poorly when acoustic speech is corrupted by noise. the robustness and adaptability of recognition systems are two huge barriers to the rapid development of automatic speech recognition (ASR) [1], therefore speech enhancement, acoustic model adaptation, and robust feature extraction have become new research focuses [2]. As to recent works, please read [3–7]. The missing feature approach (MFA) has also become a new hot topic. An earlier work on the MFA was by Cooke *et al.* in 1994 [8] and a recent survey can be found in [9].

The MFA has been proposed as an alternative solution for noisy ASR, especially under a band-limited noisy environment. On the basis of the assumption that noise corruption is different for different frequency subbands, the MFA implies that only reliable feature components corresponding to certain frequency subbands should be selected for the recognition stage and that the unreliable features must be ignored. Because the features are roughly classified as either reliable or unreliable depending on a certain threshold value, there is some real speech information in the ignored 'unreliable components' and, inversely, the effect of noise may remain in the saved 'reliable components'. Therefore the hard mask MFA involves a risky decision since it neither uses sufficient noise information nor effectively saves the real speech. As an improvement, marginalization was suggested in [10]. Furthermore, soft mask MFAs, for example, see [11,12], require the joint or conditional probability density functions (*p.d.f.*s) of reliable and unreliable feature components in the spectral domain and then calculate their integral mean to compensate for the loss of the neglected real speech information. Meanwhile, bounded marginalization must assume several special probability distributions such as a uniform distribution for the reliable features and a Gaussian distribution for an integrand in some cases [11]. Seltzer *et al.* [13] proposed a Bayesian classification

*e-mail: lingnan69@hotmail.com
†e-mail: shirai@waseda.jp
‡e-mail: akira.kurematsu@gmail.com

technique to avoid the predicament of the initial unknown masks. Li and Wang [14] gave a formal treatment on binary time-frequency masks and their conditions for becoming ideal and optimal. Works on ASR for multiple paths and multiple speakers, e.g. [15,16], are worth paying attention.

However, simple noise estimation techniques do not provide a probability measure of an element's reliability. In fact, finding such joint or conditional $p.d.f.$s required by soft mask methods is extremely difficult because of the overlap of at least two complex random processes. Simply assuming a uniform or Gaussian distribution is inappropriate. The sigmoid function proposed in some soft mask works is not an actual probability distribution function, thus appears to be a false substitute. Barker *et al.* also wrote that 'In practice the noise estimation error is only likely to be Gaussian if we have a good model of the noise. The missing data approach however attempts to avoid employing noise dependent models. In the current work we employ a simple stationary noise estimate for all noise types. For nonstationary noises the error in the estimate is likely to have a non-Gaussian distribution. Accepting this, we have not attempted to compute ideal fuzzy masks, but have instead generated a mask of values between 0 and 1 by compressing x with a simple sigmoid function' [11].

To describe the effect of noise appearing in all subbands and in all feature vectors under a simple and uniform framework, we propose four types of confident weights (CWs) to evaluate the credibility of a feature component in each subband. These CWs are based on energy, signal-to-noise ratio (SNR), statistics on SNR, and statistics on a new ratio between signal and noise (RSN). They simply and efficiently depend on log-spectral energies and can also be formulated. Each type of CW can be regarded as a discrete probability distribution. Moreover, on the basis of the CWs, we set up four approaches of features with confident weights (AFCWs) without any threshold and/or joint $p.d.f.$ of the reliable and unreliable components. These AFCWs extend and enhance our previous work [17].

Each AFCW estimates the confidence degree of each feature vector component, sums all the weighted output probabilities of the feature vector components in the log-spectral domain, and uses the summation as the likelihood score in the final recognition stages. This approach manages all the components under a uniform framework with different reliabilities and describes the effect of noise in a more simple and accessible manner.

Our experiments have shown that a speech recognition system with an AFCW exhibits high performance and a significant improvement in ASR accuracy under common noisy environments, including several stationary and/or nonstationary noise disturbances. In particular, even under low-SNR environments, the recognition accuracy of a system based on an AFCW is better than those of the general hard mask MFA, spectral subtraction (SS) [18], and cepstral mean normalization (CMN) [19].

The rest of this paper is structured as follows. In Sect. 2 we first give the four types of CWs based on energy, SNR, statistics on SNR, and statistics on the new RSN. We then set up four classes of AFCW with the framework of the hidden Markov model (HMM) [20]. In Sect. 3, we describe several comparative experiments and compare the experimental results not only between different CWs but also between AFCWs and other usual approaches for noisy ASR. Finally, in Sect. 4 we present our conclusions and discuss future studies.

2. APPROACHES OF FEATURES WITH CONFIDENT WEIGHT

2.1. Thresholds of Current Reliability Standards in Frequency Spectrum

Suppose that $S_i(\omega)$ denotes a short-time frequency spectrum of frequency ω belonging to subband i, $\delta_i = (\omega_i^{\mathrm{L}}, \omega_i^{\mathrm{H}}]$, $i = 1, \ldots, N$, where superscripts L and H denote the lowest and highest frequencies of the signal in subband i, respectively, and N is the number of subbands. Generally, to decide whether or not a feature component should be regarded as unreliable one, three current standards are employed. They are the energy standard, SNR standard, and statistical standard, which depend on the following three conditions, respectively:

$$|S_i(\omega) + N_i(\omega)|^2 - |\hat{N}_i(\omega)|^2 < \theta_1 \tag{1}$$

$$SNR_i(\omega) = 10\log_{10}(|\hat{S}_i(\omega)|^2/|\hat{N}_i(\omega)|^2) < \theta_2 \tag{2}$$

$$P(SNR_i > 0) < \theta_3, \tag{3}$$

where $|S_i(\omega) + N_i(\omega)|$ denotes the spectral magnitude of the pure speech with additional noise, $\hat{S}_i(\omega)$ and $\hat{N}_i(\omega)$ denote estimations of the speech signal and noise in subband i, respectively, and θ_j, $j = 1, 2, 3$, denote revelant threshold values. In general, we consider $\theta_1 = 0$. The value of $\hat{S}_i(\omega)$ may be obtained, for example, by SS method [18].

It is risky to choose the type of standard and the threshold. Sometimes it is difficult to avoid damaging real speech and to avoid retaining noise information. The CWs proposed in this paper will be first utilized instead of the thresholds in MFAs.

2.2. Confident Weight for Four Reliability Criteria

The reliability standards presented in this paper are different from the current threshold standards. We shall define a series of CWs $\{\hat{r}_i\}$ to describe the effect of noise in an accessible manner and to deal with the reliable and unreliable components under a uniform framework. The CW $\hat{r}_i$ is a standardization of $\{r_i\}$ as follows.

$$\hat{r}_i = r_i \Big/ \left[\sum_{j=1}^{N} r_j\right], \quad i = 1, 2, \ldots, N. \tag{4}$$

Eq. (4) has a uniform framework, and the nonnegative series $\{r_i\}$ is set depending on four standards: energy, SNR, and two statistical criteria. Therefore, we set four types of CWs. They are Energy CW, SNR CW, Stat1 CW based on SNR, and Stat2 CW based on RSN, which are defined by Eqs. (5), (6), (12), and (17), respectively. On the basis of the CW and HMM framework, we set up four promising classes of AFCWs. Consequently, the CW r_i can be regarded as a discrete probability distribution. These classes can reflect the reliability of the ith spectral feature vector corresponding to the four standards.

Energy CW. Suppose that the standardized energy $E_i = |X_i|^2/[\sum_{j=1}^{N} |X_j|^2]$, where $|X_i|^2 = |S_i + N_i|^2$, belongs to the ith subband $\delta_i = (\omega_i^{\mathrm{L}}, \omega_i^{\mathrm{H}}]$. We define

$$r_i = 1/[1 - q_{\mathrm{e}} \ln(E_i)], \tag{5}$$

where q_{e} is a constant to be determined. Note that r_i is a nondecreasing function of E_i and we define $\hat{r}_i$ as Energy CW.

SNR CW. Let

$$r_i = \beta_i/[q_{\mathrm{n}} + \beta_i], \quad \textit{and} \quad \beta_i = \exp\{b * SNR_i\}, \tag{6}$$

where b is a positive constant. In this paper, taking $b = \ln 10/10$, we obtain

$$r_i = \frac{1}{q_{\mathrm{n}}|N_i|^2/|S_i|^2 + 1}, \tag{7}$$

where q_{n} is a positive constant to be determined. It is easy to see that a larger value of r_i indicates greater reliability of the feature component of subband i. Thus, the SNR CW series r_i provides the CWs of all the subbands.

We estimate SNR_i by the on-line noise estimation method [21], which implies the relation

$$|\hat{N}_i|^2 \stackrel{\mathrm{def}}{=} \arg\max_{|X_i(\omega)|} (pdf(|X_i(\omega)|^2)) \tag{8}$$

for estimating the noise magnitude, where $pdf(|X_i(\omega)|^2)$ is the $p.d.f.$ of the ith subband spectrum. Note that in this equation the signal X_i belongs to the noise before the onset of speech according to [21]. Consequently, $|\hat{S}_i|$ can be obtained by SS [18] after the onset of speech, and X_i is real noisy speech. Furthermore,

$$SNR_i = 10\log_{10}(|\bar{S}_i|^2/|\hat{N}_i|^2), \quad |\bar{S}_i|^2 = \frac{1}{n_i}\sum_{\omega\in\Delta_i} |\hat{S}_i(\omega)|^2, \tag{9}$$

where $|\bar{S}_i|^2$ denotes the average spectrum energy of subband i, and n_i denotes the number of samples in the ith subband.

To estimate the CWs of the delta spectral feature vector components, we define the SNR of delta speech as follows:

$$SNR_i \stackrel{\mathrm{def}}{=} \frac{|\bar{S}_{\mathrm{c},i}|^2 + |\bar{S}_{\mathrm{p},i}|^2}{|\hat{N}_{\mathrm{c},i}|^2 + |\bar{N}_{\mathrm{p},i}|^2}, \tag{10}$$

where index c denotes the current speech frame and index p denotes the previous speech frame.

The confidence of the delta-delta spectral feature components is defined in a similar manner.

We define the standardized series $\hat{r}_i$ given by Eq. (4) as the SNR CW,

$$\hat{r}_i = (r_i - r_l)/(r_h - r_l), \quad r_l \leqslant r_i \leqslant r_h, \tag{11}$$

where r_l and r_h are selected as boundaries for r_i.

Stat1 CW, confident weight based on the statistical property of SNR. Suppose that the SNR obeys a normal distribution. We define

$$r_i = \int_{-\infty}^{y(i)/2} f_{SNR_i}(x)dx \tag{12}$$

$$\hat{r}_i = r_i \Big/ \left[\sum_{j=1}^{N} r_j\right], \quad i = 1, 2, \ldots, N, \tag{13}$$

where $y(i)$ denotes the amplitude of a noisy speech in subband i and $f_{SNR_i}(x)$ denotes the probability density function $p.d.f.$ of SNR_i. We consider the position series $\hat{r}_i$ to be the CW in the statistical standard based on SNR denoted as Stat1 CW.

Stat2 CW, Confident weight based on the statistical property of RSN. Suppose that the centralized amplitudes of noise N_i and speech S_i,

$$|N_i(\omega)| - |\bar{N}_i| \quad \textit{and} \quad |S_i(\omega)| - |\bar{S}_i|, \tag{14}$$

satisfy normal distributions with different parameters and that both random variables are independent in the ith frequency interval. The standardized variables of $|N_i|$ and $|S_i|$ are N_i^* and S_i^*, respectively. Let $|N_i^*|$ be the absolute value of

$$RSN_i = S_i^*/|N_i^*|. \tag{15}$$

Then, from the central limit theorem, the two standardized variables are approximately normal variables if the number of samples is sufficiently large. Thus, the $p.d.f.$ of $T_i = RSN_i$ is [22]

$$f_{T_i}(x) = \frac{1}{\pi(1 + x^2)}. \tag{16}$$

In practice, we only consider the case $x > 0$ see item (3) of 'Stat2 CW' in Sect. 3.2.

We define Stat2 CW $\{\hat{r}_i\}$ as follows:

$$r_i = \frac{1}{1 + t_i^2}, \quad \hat{r}_i = r_i \Big/ \left[\sum_{j=1}^{N} r_j\right], \quad i = 1, 2, \ldots, N, \tag{17}$$

where t_i is an observable value of T_i.

2.3. Approach of Features with Confident Weight

For the HMM [20] framework λ with a Gaussian mixture model (GMM), the likelihood of the input feature, being the N-dimensional vector $X = \{X_1, \ldots, X_N\}^{\mathrm{T}}$ under λ is

$$P(X|\lambda) = \sum_{i=1}^{M} k_i \exp\left\{-\frac{1}{2}(X - \vec{\mu}_i)^{\mathrm{T}} \Sigma_i^{-1} (X - \vec{\mu}_i)\right\}, \quad (18)$$

where $\vec{\mu}_i = (\vec{\mu}_{i1}, \ldots, \vec{\mu}_{iN})^{\mathrm{T}}$, $\Sigma_i = diag(\sigma_{i1}^2, \ldots, \sigma_{iN}^2)$, and k_i is a constant that depends on not only the mixing coefficients α_i of the ith Gaussian density in the GMM but also the Σ_i, k is exactly given by $k_i = \alpha_i / ((2\pi)^{N/2} |\Sigma_i|^{1/2})$.

In the AFCW, the selection of the CW depends on the sum of the output probabilities of the weighted feature vector in the logarithmic domain. Let r_j denote the confidence of X_j; then, the likelihood of the input feature vector X under λ in the AFCW is

$$\sum_{i=1}^{M} c_i \exp\left\{-\frac{1}{2}\Sigma_j^N \left(\hat{r}_j \log(2\pi\sigma_{ij}^2) + \hat{r}_j \frac{(X - \mu_{ij})^2}{\sigma_{ij}^2}\right)\right\}, \quad (19)$$

where c_i depends on the mixing coefficients. In the case of $r_j = 1$, the probability of component j is the same as that for the standard HMM; in other words, component j is reliable in the standard MFA. When $r_j = 0$, the likelihood of component j is 1; hence, its logarithmic value is zero, and component j is ignored, which indicates that component j is unreliable in the standard MFA. Henceforth, the AFCW can be also considered as an extended and more precise MFA in some sense.

In Eq. (19), $r_j \log(2\pi\sigma_{ij}^2)$ represents the impact of noise on the variance σ_{ij}^2, and $r_j(X_j - \mu_{ij})^2/\sigma_{ij}^2$ represents the impact of noise on the difference $X_j - \mu_{ij}$; from the viewpoint of model self-adaptation, both the means and variances in the model are adjusted under noisy environments. Our experimental results showed that the recognition accuracy can degrade significantly if recognition is only weighted on the mean value and the Mahalanobis distance of the second part of Eq. (19), and the adjustment of the first part is ignored; this implies that $r_j \log(2\pi\sigma_{ij}^2)$ is necessary in Eq. (19).

From Eq. (19), the AFCW is a self-adaptive method that combines the confidence of the speech feature vector components and the speech recognition model. Moreover, it does not introduce greater computational complexity compared with the standard HMM.

3. EXPERIMENTAL RESULTS FOR DETERMINATION OF CW AND RECOGNITION

3.1. Databases and Recognition Platform

Database 1. All the experiments are performed using a Mandarin Chinese speech database that contains 100 isolated words and utterances from 13 speakers (seven males and six females). Each word has 21 utterances for model training and 18 utterances for testing. The front-end uses an 8 kHz sampling frequency and a 10 ms frame period. On the HTK 3.0 platform [23], each word is modelled on eight Gaussian mixtures for a 10-state HMM. The added silence model is based on four Gaussian mixtures for a 3-state-HMM. Each model is trained on clean speech data.

Database 2. Three types of noise data from the database NOISEX-92 [24] are used in the experiments; the used noises consist of white noise, high-frequency (HF) channel noise, and babble noise. The noisy speech data cover an overall SNR range from −5 to 20 dB; these data are derived by artificially adding noise to the clean speech data used for testing.

Features. The main experiments in the next two subsections were performed in a Mel-frequency spectrum vector consisting of 48 components (24 static and 24 delta components) that was similar to *FBAK.D* of HTK. However, in Sect. 3.4, the input features for baseline experiments are the standard Mel-frequency cepstral coefficients (MFCCs), which include 39 components (13 static, 13 delta, and 13 delta-delta components), since the MFCC is considered to provide more accurate features than the features in the spectral field. The feature MFCC is similar to MFCC_E_D_A of HTK.

3.2. Determination of Confident Weight

Energy CW. The main problem is to determine q_{e} in Eq. (4). Initially, it is natural to assume that $r_i = 1/2$ if the present subband spectral energy is equivalent to the average of all the subband energies. Thus, $1/(1 + q_{\mathrm{e}} \ln N) = 1/2$, and if $N = 24$ is considered, we have

$$q_{\mathrm{e}} = 1/(\ln N) = 1/(\ln 24) = 0.315. \quad (20)$$

Our experiments have shown that q_{e} is indeed 0.315.

SNR CW. q_{n} in Eq. (6) is ditermined as follows. r_i should be less than 0.5 if $SNR_i = 0$, which is because $r_i = 1/[q_{\mathrm{n}} + 1]$ and $q_{\mathrm{n}} > 1$; and we hope that r_i is not too small. Here, we set $r_i = 0.2$ as the desired minimum, which makes $q_{\mathrm{n}} = 4$. Hence, $1 < q_{\mathrm{n}} \leq 4$.

Over the interval $(1, 4]$, we obtained the optimum value $q_{\mathrm{n}} = 2$, from our experiments, see Table 1. The advanced recognition experiments in Sect. 3.3 will show the reasonableness of this value.

Stat1 CW

$$r_i = \int_{-\infty}^{y(i)/2} f_{SNR_i}(x)dx, \quad (21)$$

$$f_{SNR_i}(x) = \frac{1}{\sqrt{2\pi}(\hat{\sigma}_i)} \exp\left(-\frac{(x - \hat{\mu}_i)^2}{2\hat{\sigma}_i^2}\right), \quad (22)$$

where $\hat{\mu}_i$ and $\hat{\sigma}_i^2$ denote the estimated mean value and variance for the ith frequency segment, respectively,

Table 1 Rates of different q_n-values under white noise (%).

q_n-values	SNR (dB)					
	−5	**0**	**5**	**10**	**15**	***20***
1.0	6.0	16.3	47.7	62.3	89.3	*99.7*
1.5	9.3	26.7	56.3	69.0	90.3	*99.7*
2.0	13.7	34.7	63.0	75.0	92.3	*99.7*
2.5	14.3	34.7	60.3	73.7	89.0	*98.3*
3.0	14.3	33.0	57.7	72.7	87.7	*95.3*
3.5	13.0	27.0	50.3	65.3	85.3	*91.7*
4.0	*12.7*	*26.3*	*50.0*	*63.0*	*85.7*	*90.3*

and $y(i)$ denotes the amplitude of the noisy speech in subband i.

Stat2 CW

From the symmetry and characteristics of the t-distribution, we determine Stat2 CW

(1) Separate the positive real number field into K subintervals,

$$(x_{k-1}, x_k], \quad k = 1, 2, \ldots, K;$$

(2) Calculate the observed value t_i of T_i in the ith frequency segment.

(3) Select r_i such that

$$r_i = \frac{1}{1 + x_k^2}, \quad \text{if } x_{k-1} < |t_i| \leqslant x_k,\ i = 1, 2, \ldots, N.$$

(4) Determine Stat2 CW $\{\hat{r}_i\}$, where $\hat{r}_i = r_i / [\sum_{j=1}^{N} r_j]$, $i = 1, 2, \ldots, N$.

On the basis of our experience, we considered $K = 25$ subintervals with $(0, 12)$ separated into 24 equal subintervals and the interval $(12, \infty)$. We set $r_i = 0$ if $12 = x_{24} < |t_i|$. By forming tables of r_i-values and $\hat{r}_i$-values, Stat2 CW $\{\hat{r}_i\}$ based on RSN can be rapidly and easily determined according to an observed value $\{t_i\}$. Furthermore, our experiments showed that this improved the ability for speech enhancement, see Sect. 3.3.

According to the statistical theory, the mean of random samples drawn out from a normal distribution $N(\mu, \sigma^2)$ has a normal distribution $N(\mu, \sigma^2/n)$, where n denotes the sample size. Hence, in practical applications, we can directly estimate all parameters of the noise N_i and speech S_i and then standardize them. Consequently, the advantageousness of Stat2 CW $\{\hat{r}_i\}$ is highly significant.

3.3. Comparison between the Four Types of Confident Weights

Comparison between Energy CW and SNR CW. Energy CW is only dependent on the local energy of the speech signal, which simplifies its computation but does not reflect the reliability degree of the feature vector. The SNR CW considers the magnitude of the effect of the noise on speech signals. To select a better CW between Energy CW and SNR CW, some recognition comparison experiments have been performed under white noise and babble noise environments. These results are shown in detail in Figs. 1(a) and 1(b).

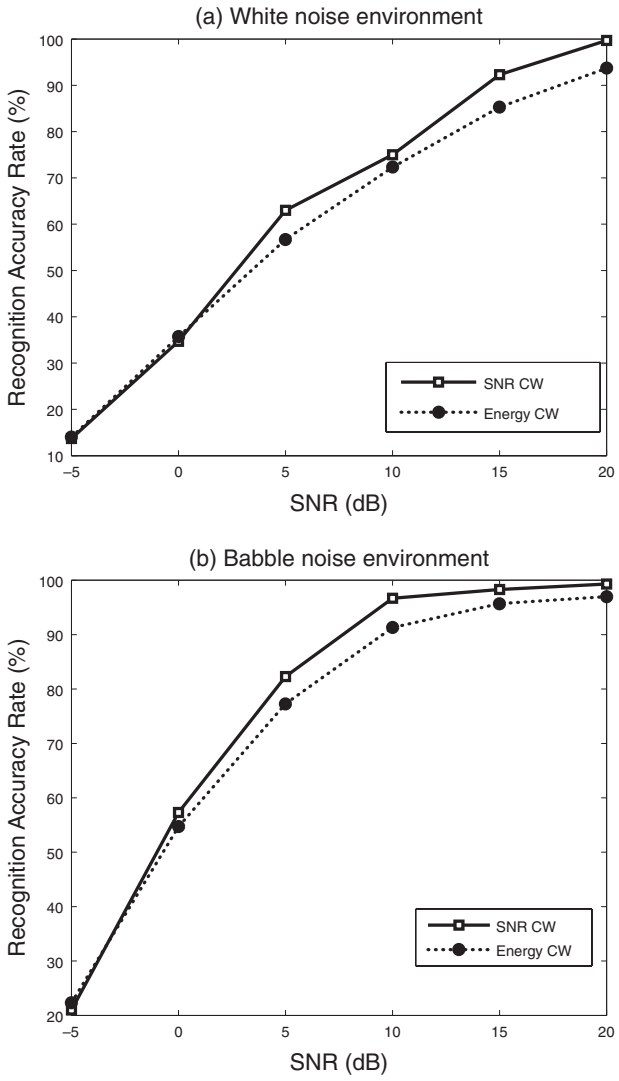

Fig. 1 Comparison between the recognition accuracy of SNR CW and Energy CW.

It can be observed that SNR CW is generally better than Energy CW, particularly when the overall SNR is high. However, when the SNR is low, this situation is reversed.

Comparison between Stat1 CW and Stat2 CW. The fitness of normality of the sample data of the centralized amplitudes of white noise N_i or speech S_i in Eq. (13) is acceptable, and the hypothesis of normality of $SNR_i = 10\log_{10}(|\bar{S}_i|^2/|\hat{N}_i|^2)$ cannot be accepted at a significance level of $\alpha = 0.10$ on the basis of a statistical test of skewness-kurtosis [22]. Under the condition of a normal population X and a sufficiently large sampling number n, the basic statistical conclusion is approximately

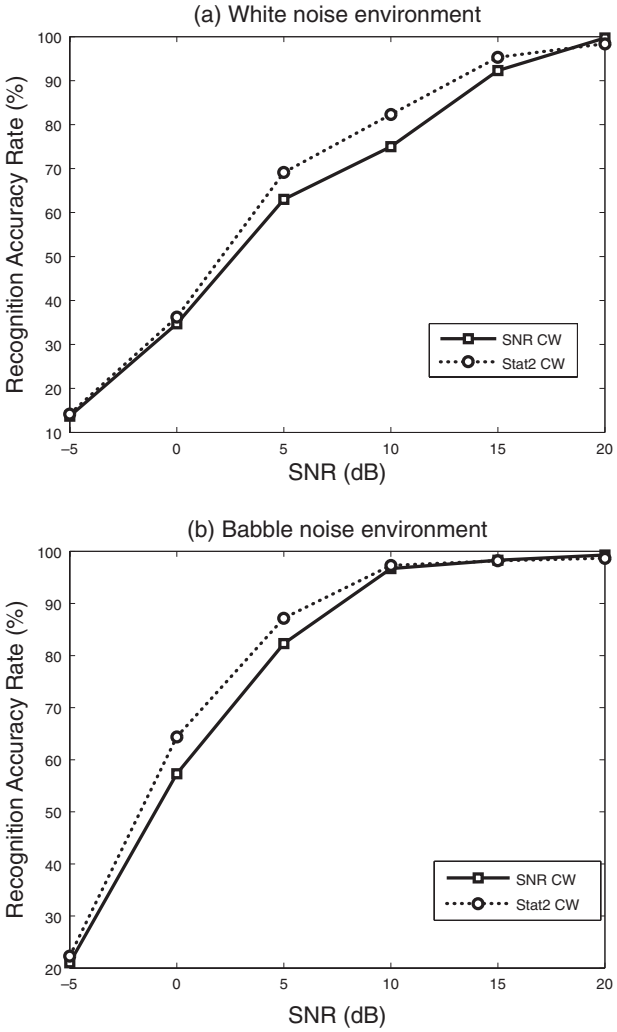

Fig. 2 Comparison between the recognition accuracy of SNR CW and Stat2 CW.

$$B_3 B_2^{3/2} \sim N\left(0, \frac{6(n-2)}{(n+1)(n+3)}\right)$$

$$B_4 B_2^2 \sim N\left(3 - \frac{6}{n+1}, \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}\right). \quad (23)$$

The two left terms are called the skewness statistics and kurtosis statistics respectively, where $B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k$. Under the condition mentioned above, our experiments demonstrated that the absolute values of the standardized variables of both statistics were always greater than 3.5, which showed that the normality of neither of them could be accepted at a significance level of 0.05. In fact, most of the values were over 6.

Comparison between Stat2 CW and SNR CW. From Fig. 2, it can be observed that Stat2 CW is generally slightly better than SNR CW. This can be attributed to the closes fit to a normal distribution. The advantage of Stat2 CW in babble noise environments is less significant than that in white noise environments. If the SNR is greater, the recognition results with Stat2 CW indicate that the assumption on the distributions of the centralized amplitudes given by Eq. (14) may be too simple.

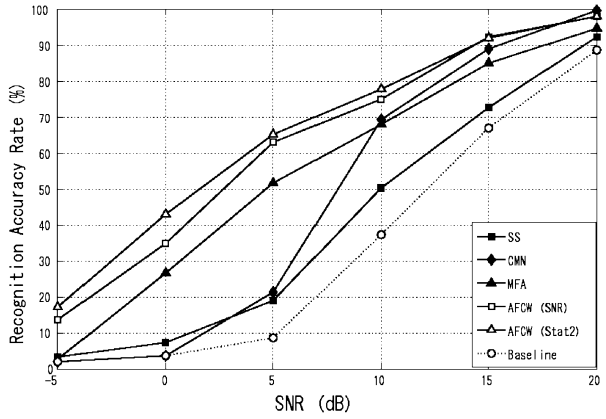

Fig. 3 Comparison of approaches for white noise.

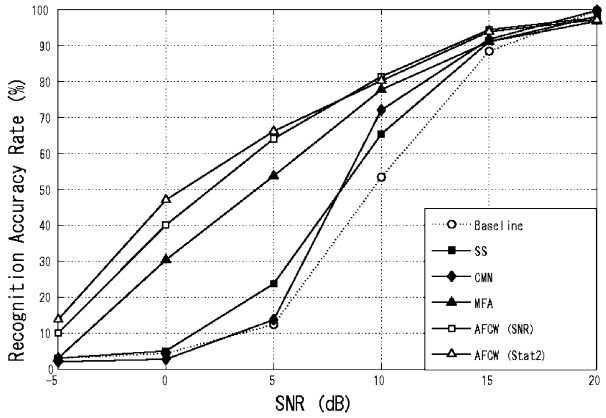

Fig. 4 Comparison of approaches under HF channel noise.

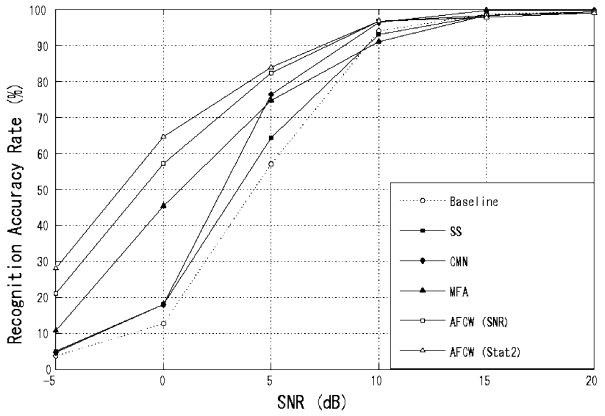

Fig. 5 Comparison of approaches under babble noise.

3.4. Comparison of Approaches

We completed our experiments by comparing our proposed AFCW with several other robust speech processing techniques such as SS [18], CMN [19], and the MFA.

The recognition results under three types of noise environments are shown in Figs. 3–5, which show that the

recognition systems based on introducing CWs have better performances. Noisy speech data cover an overall SNR range from −5 to 20 dB under the three types of noise environments, as shown in Figs. 3–5.

From the results shown in Figs. 3–5, we can conclude that by introducing CWs, the speech recognition system exhibits a better performance, the word accuracy rates for the three types of noise (including stationary and nonstationary noises) environments improve significantly, and the accuracy curves obtained by the AFCW show better results than those for other systems discussed in this paper. In particular, under low-SNR environments, the accuracy is improved to an acceptable level; moreover, the recognition performance for clean speech is maintained. In summary, the MFA is a useful robust speech recognition technique, and in general the AFCW is more effective than a hard mask in the MFA, SS, and CMN.

4. CONCLUSIONS AND FUTURE STUDIES

4.1. Conclusions

(1) By analyzing the confidence degree of feature components of noisy speech, we propose and discuss four types of CWs in this paper. As a result, the effect of noise and self-adaptation ability for noisy environments can now be simply and effectively described and handled under a uniform framework.

(2) In general, SNR CW and Stat2 CW are better than Energy CW and Stat1 CW. The assumption of the normality of the SNR is often not reasonable according to our experiments, while Stat2 CW, which is based on the central limit theorem is worth noticing.

(3) For every type of CW, the AFCW proposed here significantly enhances speech signals under an inverse environment including stationary and nonstationary noise. The AFCW does not require a threshold value or a joint distribution density of feature components. Moreover, it does not require a sigmoid function, which seems to be a false substitute, thus, the AFCW seems to be superior to a common MFA.

4.2. Future Studies

(1) A future study should be to improve SNR CW and Stat2 CW to increase the robustness of ASR, and Stat1 CW should be improved by using a mixed Gaussian distribution instead of a simple normal distribution. New more reasonable CWs and corresponding AFCWs should be explored.

(2) A method should be devised to strengthen our estimation by introducing more information from the speech model and noise model. We believe that the concept of CWs has great potential and that AFCWs based on CWs and the HMM framework are promising. In fact, other types of CW can be constructed if they appear to make sense.

(3) Our database of experimental results should be expanded to increase the value of comparative experiments.

ACKNOWLEDGEMENT

This work was supported by Project NSFC ANS06-10571103 (200601–200812).

REFERENCES

[1] D. Pearce and H.-G. Hirsch, "The Aurora experimental framework for performance evaluation of Speech Recognition system under noisy conditions," *Proc. ICSLP*, Vol. 4, pp. 29–32 (2000).

[2] L. Josifovski, "Robust automatic speech recognition with missing and unreliable data," *Ph. D. thesis, University of Sheffield*, UK (2002).

[3] M. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Am.*, **119**, 1562–1573 (2006).

[4] J. Laidler, M. Cooke and N. Lawrence, "Model-driven detection of clean speech patches in noise," *Proc. Interspeech* (2007).

[5] R. Fernandez Astudillo, D. Kolossa and R. Orglmeister, "Uncertainty propagation for speech recognition using RASTA features in highly nonstationary noisy environments," In: 8. ITG-Fachtagung Sprachkommunikation (ITG08) (2008).

[6] R. Martin and C. Breithaupt, "speech enhancement in the DFT domain using laplacian speech priors," *Proc. IWAENC* (2003).

[7] T. Takiguchi, M. Nishimura and Y. Ariki, "Acoustic model adaptation using first-order linear prediction for reverberant speech," *IEICE Trans. Inf. Syst.*, **E89-D**, 908–914 (2006).

[8] M. Cooke, P. D. Green and M. D. Crawford, "Handling missing data in speech recognition," *Proc. 3rd Int. Conf. Spoken Language Processing*, pp. 1555–1558 (1994).

[9] C. Cerisara, S. Demange and J.-P. Haton, "On noise masking for automatic missing data speech recognition: A survey and discussion," *Comput. Speech Lang.*, **21**, 443–457 (2007).

[10] B. Raj, M. L. Seltzer and R. M. Stern, "Reconstruction of missing feature for robust speech recognition," *Speech Commun.*, **43**, 275–296 (2004).

[11] J. P. Barker, L. Josifovski, M. P. Cooke and P. Green, "Soft decisions in missing data techniques for robust automatic speech recognition," *Proc. ICSLP*, Vol. 1, pp. 373–376 (2000).

[12] M. Cooke, P. Green, L. Josifovski and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Commun.*, **34**, 267–285 (2001).

[13] M. L. Seltzer, B. Raj and R. M. Stern, "A Bayesian framework for spectrographic mask estimation for missing feature speech recognition," *Speech Commun.*, **43**, 379–393 (2004).

[14] Y. Li and D. L. Wang, "On the optimality of ideal binary time-frequency masks," *Speech Commun.*, **51**, 230–239 (2009).

[15] A. Hämäläinen, L. Bosch and L. Boves, "Modeling pronunciation variation with single-path and multi-path syllable models," *Speech Commun.*, **51**, 130–150 (2009).

[16] R. Takeda, S. Yamamoto, K. Komatani, T. Ogata and H. G. Okuno, *IEA/AIE 2007, LNAI 4570* (Springer-Verlag, Berlin/Heidelberg, 2007), pp. 384–394.

[17] Y. Ge, J. Song, L. Ge and K. Shirai, "Approach of feature with confident weight for robust speech recognition," *IEEE 6th*

Workshop on Multimedia Signal Processing (2004).
[18] R. Martin, "Spectral subtraction based on minimum statistics," *Proc. Eur. Signal Process. Conf.*, pp. 1182–1185 (1994).
[19] S. Furui, "On the role of spectral transition for speech perception," *J. Acoust. Soc. Am.*, **80**, 1016–1025 (1986).
[20] L. R. Rabiner and B. H. Juang, "A introduction of hidden Markov models," *IEEE Trans. Acoust. Speech Signal Process.*, **3**, 4–16 (1986).
[21] H. G. Hirsch and C. Ehrlicher, "Noise estimation technique for robust speech recognition," *Proc. ICASSP*, pp. 153–156 (1995).
[22] Y. Ge, *Probability Theory and Mathematical Statistics* (Tsinghua Univ. Press, Beijing, 2005).
[23] S. Young and D. Kershaw, *The HTK Book* (Copyright Microsoft Corporation, 1995–1999).
[24] A. Varga and H. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, **12**, 247–251 (1993).

Lingnan Ge received her B. Eng. degree in Mechanical Engineering and M. Eng. degree in Computer Science from Tsinghua University, Beijing, China, in 2000 and 2003, respectively. She is currently a Ph.D. candidate in Information and Computer Science, Waseda University, where she has been working since 2003. Since 1998, she has been focusing on acoustic signal processing and spoken language recognition research. Recently, her main interest has been the robustness of speech signals.

Katsuhiko Shirai received his B. Eng., M. Eng., and Dr. Eng. degrees from Waseda University, Tokyo, Japan, in 1963, 1965, and 1973, respectively (all in Electrical Engineering). Since 1991, he has been a professor at the Department of Information and Computer Science, Waseda University. His research interests include spoken language processing and the high-level synthesis of ASIC and CAI systems. He is a member of the Information Processing Society of Japan, Japanese Society of Artificial Intelligence, and the Institute of Electrical and Electronics Engineers in addition to other organizations.

Akira Kurematsu received his B.E. degree in Electrical Communication from Waseda University in 1961. He received a Ph.D. degree from Waseda University in 1971. In 1961, he joined the Research and Development Laboratories of KDD, where he was engaged in research on pattern recognition, speech signal processing and human communication systems. In 1983, he was appointed Deputy Director of KDD RD Labs. From 1986 to 1993, he was the President of ATR Interpreting Telephony Research Laboratories. He was a professor at the Department of Electronic Engineering, University of Electro-Communications, from 1993 to 2004. Since 2004, he has been a visiting professor at Waseda University.