Feature Compensation Combining SNR-Dependent Feature Reconstruction and Class Histogram Equalization

Youngjoo Suh and Hoirin Kim

ABSTRACT—In this letter, we propose a new histogram equalization technique for feature compensation in speech recognition under noisy environments. The proposed approach combines a signal-to-noise-ratio–dependent feature reconstruction method and the class histogram equalization technique to effectively reduce the acoustic mismatch present in noisy speech features. Experimental results from the Aurora 2 task confirm the superiority of the proposed approach for acoustic feature compensation.

Keywords—Class histogram equalization, feature reconstruction, robust speech recognition.

I. Introduction

The performance of speech recognizers degrades severely when they are employed in acoustically mismatched environments compared to training environments [1]. An effective way to reduce acoustic mismatch is the feature compensation approach. An acoustic mismatch resulting from the corruption of additive noise and channel distortion causes a nonlinear transformation in feature spaces such as the cepstral coefficients [1]. An efficient nonlinear feature compensation approach is the histogram equalization (HEQ) technique, in which compensation is made by converting the probability density function (PDF) of test features into that of reference features [2]-[5]. However, HEQ has some fundamental limitations when it compensates a short noisy utterance, a common input unit in current speech recognition applications [5]. In this case, the class HEQ (CHEQ) approach is reported to provide better compensation effects [4], [5]. Nevertheless, our further analysis of the compensation features provided by CHEQ indicates that nonmonotonic transformation caused by the acoustic mismatch and inaccurate class information still act as the major hindrance in its compensation capability. Here, nonmonotonic transformation in CHEQ refers to the discrepancy between reference and test cumulative distribution function (CDF) estimates for a specific feature value. Therefore, the performance of CHEQ can be further improved by alleviating these two problems. Nonmonotonic transformation mainly results from the presence of random noise. Thus, it can be more monotonic if the noise components of noisy speech are reduced [6]. The estimation of class information can also be enhanced by utilizing noise-reduced speech features. Consequently, CHEQ can provide further performance improvement by utilizing noise-reduced speech features. Therefore, we propose a new feature compensation technique which combines a signal-to-noise-ratio (SNR)–dependent feature reconstruction (SFR) approach with CHEQ, called SFR-CHEQ.

Manuscript received May 14, 2008; revised June 12, 2008; accepted June 23, 2008.

This work was supported by the IT R&D program of MKE/IITA, Rep. of Korea [2008-S-001-01, Development of Large Vocabulary/Interactive Distributed/Embedded VUI for New Growth Engine Industries].

Youngjoo Suh (phone: + 82 42 866 6830, email: yjsuh@icu.ac.kr) and Hoirin Kim (email: hrkim@icu.ac.kr) are with the School of Engineering, Information and Communications University, Daejeon, Rep. of Korea.

II. CHEQ

Let us define noisy test feature vector Y_n consisting of K-dimensional components at time frame n in feature sequence S composed of N frames as

$$Y_n = [y_n^{(1)}\ y_n^{(2)}\ \ldots y_n^{(k)} \ldots y_n^{(K)}]^T, \quad (1)$$

where $y_n^{(k)}$ is the k-th test feature component, and T stands for the vector transpose. The class information at acoustic class ω_i

in CHEQ is given as the posterior probability of ω_i, given Y_n, as

$$P(\omega_i \mid Y_n) = \frac{\alpha_i \mathbb{N}(Y_n; \mu_i, \Sigma_i)}{\sum_{m=1}^{I} \alpha_m \mathbb{N}(Y_n; \mu_m, \Sigma_m)}, \quad (2)$$

where I denotes the total number of acoustic classes, α_i represents the mixture component weight, and $\mathbb{N}(Y_n; \mu_i, \Sigma_i)$ is a Gaussian distribution with mean vector μ_i and covariance matrix Σ_i at the i-th class. With class information in (2), an estimate of the reference feature component by CHEQ, given test feature component y_n, is defined by [5]

$$\hat{x}_{CHEQ,n} = \sum_{i=1}^{I} P(\omega_i \mid Y_n) C_{X(i)}^{-1}\left[\hat{C}_{Y(i)}(y_n)\right], \quad (3)$$

where $C_{X(i)}^{-1}$ is the inverse of the reference CDF, and $\hat{C}_{Y(i)}(y_n)$ denotes the test CDF estimate of y_n at the i-th class obtained by using the order statistics-based method [2], [5] as

$$\hat{C}_{Y(i)}(y_n) = \frac{\sum_{r=1}^{R(y_n)} P(\omega_i \mid Y_{T(r)})}{\sum_{r=1}^{N} P(\omega_i \mid Y_{T(r)})}, \quad (4)$$

where $R(y_n)$ is the rank of y_n $(1 \le n \le N)$ when the sequence of test feature components are sorted in ascending order, and $T(r)$ denotes the original frame index when its rank is given as r.

As previously mentioned, the performance of CHEQ can be further improved by reducing deviations of rank information $T(r)$, which result in nonmonotonic transformation, and by enhancing the estimation accuracy of $P(\omega_i|Y_n)$.

III. SFR-CHEQ

SFR-CHEQ combines an SNR-dependent feature reconstruction algorithm with CHEQ. The former method consists of a vector quantization (VQ) type of minimum mean square error (MMSE) estimation and frame-SNR-dependent feature reconstruction, while the latter technique transforms reconstructed features into compensated features with refined class information and test CDFs.

In the first stage of the SNR-dependent feature reconstruction algorithm, an estimate of the clean test feature is obtained from the noisy test feature by using a VQ type of MMSE estimator. In the MMSE estimation, the noisy speech feature can be modeled by a mixture of J Gaussians [1]. In addition, clean and noisy speech features are assumed to be jointly Gaussian within the acoustic class involved with each mixture component. Then, an estimate of the clean test feature vector based on the VQ type of MMSE estimation is given by

$$\hat{X}_n = \sum_{j=1}^{J} P(\Omega_j \mid Y_n) \cdot \nu_j, \quad (5)$$

where $P(\Omega_j|Y_n)$ is the posterior probability of the j-th Gaussian, given test feature vector Y_n, ν_j is the mean vector of the j-th Gaussian, and J is the total number of Gaussians in VQ.

Because both nonmonotonic transformation and inaccurate class information result from noise corruption, their adverse effects are more dominant at lower SNRs. On the contrary, VQ-based feature estimation introduces quantization errors at high SNRs. For these reasons, an SNR-dependent reconstructed feature is obtained by a weighted average as

$$\tilde{Y}_n = (1-\beta_n) Y_n + \beta_n \hat{X}_n, \quad (6)$$

where SNR-dependent weight β_n is designed to weigh more heavily on the VQ-based estimated feature at lower SNRs by using a sigmoid function which is inversely proportional to the frame SNR as

$$\beta_n = \frac{\delta \exp(-\kappa(\xi_n - \theta))}{1 + \exp(-\kappa(\xi_n - \theta))}, \quad (7)$$

in which δ is a constant for the degree of weight, κ and θ are scale and bias factors, respectively, and ξ_n denotes the n-th frame SNR. Finally, compensated features are obtained from reconstructed features by using (3) with refined class information and test CDFs as

$$\hat{x}_{SFR,n} = \sum_{i=1}^{I} P(\omega_i \mid \tilde{Y}_n) C_{X(i)}^{-1}\left[\hat{C}_{SFR,Y(i)}(\tilde{y}_n)\right], \quad (8)$$

where $\tilde{y}_n$ represents the reconstructed version of y_n obtained by using (6), $P(\omega_i \mid \tilde{Y}_n)$ denotes the posterior probability of acoustic class ω_i, given $\tilde{Y}_n$, and $\hat{C}_{SFR,Y(i)}(\tilde{y}_n)$ is the test CDF estimate of $\tilde{y}_n$ at the i-th class in SFR-CHEQ, which is re-estimated by using reconstructed features and their re-ordered rank information as

$$\hat{C}_{SFR,Y(i)}(\tilde{y}_n) = \frac{\sum_{\tilde{r}=1}^{R(\tilde{y}_n)} P(\omega_i \mid \tilde{Y}_{T(\tilde{r})})}{\sum_{\tilde{r}=1}^{N} P(\omega_i \mid \tilde{Y}_{T(\tilde{r})})}. \quad (9)$$

IV. Experimental Results

In the experiments, the Aurora 2 database converted from the TI-DIGITS database was used. We used clean training data and three test sets of the Aurora 2 noisy speech database, where sets A and B were corrupted by different four kinds of additive noise and set C was contaminated by two kinds of additive noise and channel distortion. In feature extraction, speech signals were blocked into a sequence of frames, each 25 ms long with a 10 ms interval. For each frame, speech signals were pre-emphasized with a factor of 0.97 and a Hamming window was applied. From the frame sequence of 23 mel-scaled filter-bank log energies, 39-dimensional MFCC-based feature vectors, each consisting of 12 MFCCs, log energy, and their first and second derivatives, were extracted. The baseline speech recognizer employs 13 whole-word hidden Markov models

(HMMs), consisting of 11 digit models with 16 states, a silence model with 3 states, and a short-pause model with 1 state. Each state in digits has 3 Gaussian mixture components, while those in silence and short-pause have 6 Gaussians. Diagonal covariance matrices were used in the HMMs. The number of histogram bins in reference CDFs was empirically chosen as 64. Compensation was conducted on all of the 39-dimensional MFCCs independently for both training and test data on an utterance basis. Parameters I, J, δ, κ, and θ were empirically set to 7, 1024, 0.3, 0.5, and 20 dB, respectively. The frame SNR was estimated as the ratio of the frame energy to the averaged noise energy obtained from the initial silence region of each utterance.

Figure 1 shows the log energy contours of the clean, noisy, and reconstructed speech compensated by CHEQ. The noisy speech utterance with a 5 dB SNR was chosen from set C. We observe that the compensated version of noisy speech features is notably different from that of clean speech features in silence and speech regions. On the contrary, the compensated version of reconstructed features is closer to that of clean features.

Table 1 shows recognition results for sets A, B, and C obtained by MFCC, SFR, HEQ, CHEQ, and SFR-CHEQ. The results are represented in averaged word error rates between 0 and 20 dB SNRs. The table shows that SFR-CHEQ produces outstanding improvement over MFCC with an error reduction of 64.89% and substantial improvement over CHEQ with a reduction of 9.83%, even though SFR only produces relatively lower improvement. Due to the relatively large amount of the Aurora 2 test data in the evaluation, we regard these improvements by SFR-CHEQ as statistically significant. The error rates of HEQ and CHEQ for set C are especially large compared to those for sets A and B. These results indicate that HEQ and CHEQ are more susceptible to the acoustic mismatch caused by additive noise and channel distortion together than additive noise only. In contrast, the performance of SFR-CHEQ for set C is very close to that for sets A and B, producing an error reduction of 12.61% over CHEQ. From these results, we conclude that SFR-CHEQ is an effective approach in feature compensation and its effectiveness seems to be almost independent of the types of acoustic mismatches.

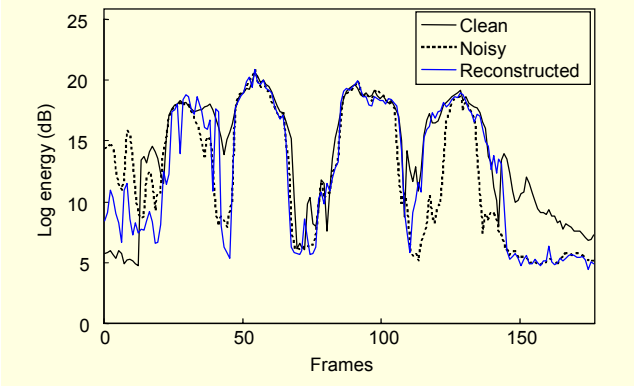

Fig. 1. Log energy contours of the clean, noisy, and reconstructed sp eech compensated by the CHEQ technique (1 frame =10ms).

Table 1. Word error rates and error reductions by the MFCC, SFR, HEQ, CHEQ, and SFR-CHEQ techniques on the Aurora 2 task (%).

Test sets	MFCC	SFR	HEQ	CHEQ	SFR-CHEQ
A	38.92	24.24	19.40	15.54	14.21
B	44.42	24.80	18.31	15.15	13.69
C	32.88	28.86	21.54	16.33	14.27
Average	39.91	25.39	19.39	15.54	14.01
Error reduction	-	36.40	51.41	61.06	64.89

V. Conclusion

We proposed a new histogram equalization technique called SFR-CHEQ which combines SNR-dependent feature reconstruction and CHEQ for robust feature compensation. SFR-CHEQ is focused on reducing acoustic mismatch-driven nonmonotonic transformation and improving the estimation accuracy of class information. Besides its remarkable performance, one crucial merit of SFR-CHEQ is that it is mainly based on speech models and does not require any detailed noise statistics. Therefore, even for other types of noise, SFR-CHEQ can be more robust than other techniques based on both speech and noise statistics.

References

[1] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing*: *A Guide to Theory*, *Algorithm*, *and System Development*, Upper Saddle River, NJ: Prentice Hall, 2001.

[2] J.C. Segura et al., "Cepstral Domain Segmental Nonlinear Feature Transformations for Robust Speech Recognition," *IEEE Signal Processing Letters*, vol. 11, no. 5, 2004, pp. 517-520.

[3] Á. de la Torre et al., "Histogram Equalization of Speech Representation for Robust Speech Recognition," *IEEE Trans. Speech, Audio Processing*, vol. 13, no. 3, 2005, pp. 355-366.

[4] Y. Suh and H. Kim, "Class-Based Histogram Equalization for Robust Speech Recognition," *ETRI J.*, vol. 28, no. 4, Aug. 2006, pp. 502-505.

[5] Y. Suh, M. Ji, and H. Kim, "Probabilistic Class Histogram Equalization for Robust Speech Recognition," *IEEE Signal Processing Letters*, vol. 14, no. 4, Apr. 2007, pp. 287-290.

[6] J.C. Segura et al., "Feature Extraction Combining Spectral Noise Reduction and Cepstral Histogram Equalization for Robust ASR," *Proc. ICSLP*, Sept. 2002, pp. 225-228.