STOCHASTIC FEATURES FOR NOISE ROBUST SPEECH RECOGNITION

N.Iwahashi, H.Pao, H.Honda, K.Minamino, M.Omote

Sony Corp. D21 Laboratory
6-7-35, Kitashinagawa, Shinagawa-ku, Tokyo 141 Japan
E-mail: {naoto,pao,honda,nanno,omote}@pdp.crl.sony.co.jp

ABSTRACT

This paper describes a novel technique for noise robust speech recognition, which can incorporate the characteristics of noise distribution directly in features. The feature itself of each analysis frame has a stochastic form, which can represent the probability density function of the estimated speech component in the noisy speech. Using the sequence of the probability density functions of the estimated speech components and Hidden Markov Modelling of clean speech, the observation probability of the noisy speech is calculated. In the whole process of the technique, the explicit information on SNR is not used. The technique is evaluated by large vocabulary isolated word recognition under car noise environment, and is found to have clearly outperformed nonlinear spectral subtraction (between 13% and 44% reduction in recognition errors).

1. INTRODUCTION

Noise robustness in speech recognition has attracted a great deal of interest [1]. Interfering noise degrades the performance in existing recognition systems, particularly where there is a mismatch in the training and testing environment. In order to compensate the mismatch, several methods have been studied in the two main approaches: 1) compensation in the feature extraction stage; 2) compensation of speech model for noisy environment. Spectral subtraction [2, 3, 4, 5], is a technique of the former kind. In this technique, although the variance of noise power spectrum can implicitly be considered in the process of calculating overestimation factor [5], the noise variance is not compensated directly. Other compensation techniques in the feature extraction stage are almost the same as spectral subtraction technique in terms of the compensation of the variance of noise power spectrum. As the second category, the model combination based approach [6, 7, 8, 9] can compensate the variance of noise explicitly by incorporating characteristics of the noise variance into the noisy speech model. It is applicable to a wide rage of noise environments. In the model combination approach, however, knowledge of the SNR is required in order to combine noise and speech models precisely.

These previously investigated approaches are based on the standard framework of recognition, where the feature extraction process produces a point in the feature space, which is input to a classifier. On the other hand, the speech component in the noisy speech signal desirably should be estimated while keeping ambiguity, because noise has a random characteristics which lead to a variance of power spectrum.

The technique presented in this paper is a process in the feature extraction stage, and can compensate the variance of noise power spectrun by incorporating the noise distribution into the feature. In the technique, the feature has a stochastic form, which can represent the probability density function of the estimated speech component in noisy speech. Using the sequence of the probability density functions of the estimated speech components and Hidden Markov Modelling of clean speech, the observation probability of the noisy speech is calculated. In the whole process of the technique, the explicit information on SNR is not used. This technique will be referred to as Stochastic Feature Extraction (SFE).

SFE applied to MFCC [10] is evaluated for speech in the presence of car noise. Comparison in performance is made with the non-linear spectral subtraction[4]. It was reported in [7] that parallel model combination was comparable to nonlinear spectral subtraction technique under car noise environment.

2. STOCHASTIC FEATURE

The observed noisy speech signal consists of noise and clean speech components. The clean speech component, $\mathbf{s}$, in spectral domain is represented as

$$\mathbf{s} \quad = \quad \mathbf{y} - \mathbf{n}. \tag{1}$$

where $\mathbf{y}$ and $\mathbf{n}$ denote the observed noisy speech signal and noise components in spectral domain respectively. Because $\mathbf{n}$ is a random variable, $\mathbf{s}$ is taken as a random variable. Thus, the probability density function (pdf) of speech component, f, is represented using the pdf of noise, g, as follows:

$$f(\mathbf{s}) = \begin{cases} J \cdot g(\mathbf{y} - \mathbf{s}) & \mathbf{s} \in \mathcal{S} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $\mathcal{S}$ represents the observation space such that $(\mathbf{y} - \mathbf{s})$ falls into the range of possible noise signal. J is a normalization factor to satisfy $\int_{\mathcal{S}} f(\mathbf{s}) d\mathbf{s} = 1$. In order to evaluate the observation probability of the noisy speech using a Hidden Markov Model (HMM) trained by clean speech, the ith state output probability, $P_i(\mathbf{y}_t, g)$, for observation $\mathbf{y}_t$ in the

tth frame and noise pdf g, is represented as

$$P_i(\mathbf{y}_t, g) = \int_S b_i(\mathbf{s}) g(\mathbf{y}_t - \mathbf{s}) d\mathbf{s} \quad (3)$$

$$= \frac{1}{J_t} \cdot \int b_i(\mathbf{s}) f_t(\mathbf{s}) d\mathbf{s} \quad (4)$$

where b_i denotes the output pdf for the ith state in the HMM. J_t denotes the normalization factor for the tth frame. The integration in Equation 4 is over the entire space. In the case that the boundaries of speech for recognition are given. $\frac{1}{J_t}$ can be ignored in the decoding stage. because it does not have the influence on recognition results. Therefore, in this case, it is enough to consider state output probability, $B_i(f_t)$, for estimated pdf f_t of the speech component, which is represented as

$$B_i(f_t) = \int b_i(\mathbf{s}) f_t(\mathbf{s}) d\mathbf{s} \quad (5)$$

The integration is over the entire space. If the parameters in the spectral domain, are mapped into a different domain, such as cepstral domain, noise is not necessarily additive. But still, the state output probability, $B_i^c(\cdot)$, in the new domain can be calculated in same way as follows:

$$B_i^c(f_t^c) = \int b_i^c(\mathbf{s}^c) f_t^c(\mathbf{s}^c) d\mathbf{s}^c \quad (6)$$

where b_i^c and f_t^c are the output pdf for ith state in a HMM and the pdf of the estimated speech component in the new domain respectively. The integration is over the entire space.

If f_t^c is Gaussian, which is represented by $\mathcal{N}(\xi, \mathbf{\Psi})$, a parameter set of $\{\xi, \mathbf{\Psi}\}$ is taken as a stochastic feature. In the case that b_i^c is also Gaussian, which is represented by $\mathcal{N}(\mu, \mathbf{\Sigma})$, the state output probability is calculated as

$$B_i^c(\mathcal{N}(\xi, \mathbf{\Psi})) = \frac{1}{(2\pi)^{\frac{D}{2}} |\mathbf{\Psi} + \mathbf{\Sigma}|^{\frac{1}{2}}} \times \exp\left\{-\frac{1}{2}(\xi - \mu)^T (\mathbf{\Psi} + \mathbf{\Sigma})^{-1} (\xi - \mu)\right\} \quad (7)$$

In the case that the state output pdf is represented by a mixture of Gaussians, the calculation is straight-forward. In the next section, the algorithm to obtain the stochastic feature, which has a form of Gaussian in cepstral domain, is described.

3. IMPLEMENTATION

If the state output pdfs in clean speech HMM is given in cepstral domain, the pdf of the estimated speech component in each noisy speech observation should be represented also in cepstral domain. Although there seem to be several ways of obtaining Gaussian representation for the pdf of the estimated speech component in cepstral domain, one straightforward way is to calculate mean vector and covariance matrix directly in cepstral domain using noise spectrum samples. The process of the stochastic feature extraction applied to MFCC domain is illustrated by Figure 1.

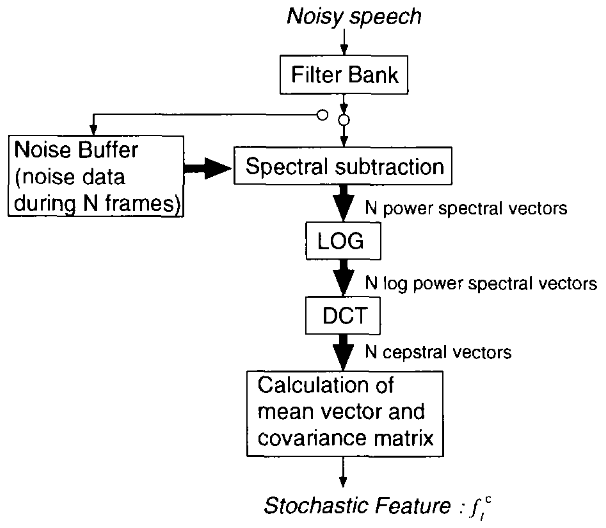
Noisy speech
Filter Bank
Noise Buffer (noise data during N frames)
Spectral subtraction
N power spectral vectors
LOG
N log power spectral vectors
DCT
N cepstral vectors
Calculation of mean vector and covariance matrix
Stochastic Feature : f_t^c

Figure 1: Block-diagram of the process of calculating Gaussian stochastic feature in the MFCC domain

Noise data samples in the figure are saved into the noise buffer during N frames before each input utterance. The processing stages are applied to the observed signal at each frame during input utterance using the saved noise data samples in the noise buffer. The steps in the process of calculating a stochastic feature for cepstral static coefficients is summarised as follows

1. Before each utterance, filter bank power spectral vectors of noise during N frames are saved into the noise buffer.

2. During the utterance, subtract each noise spectral vector in the noise buffer from the filter bank power spectral vector of input noisy speech, obtaining N power spectral vectors.

3. Transform the N power spectral vectors into N cepstral vectors.

4. Calculate mean vector and covariance matrix of the cepstrums among the N cepstrums, which results in pdf f_t^c for each frame.

The spectral subtraction in the process is carried out by using the following rule:

$$\hat{s}_i = \begin{cases} y_i - n_i & y_i - n_i > \beta y_i \\ \beta y_i & \text{otherwise} \end{cases} \quad (8)$$

where y_i and n_i are the ith components of power spectral vectors for the observed noisy speech and the noise, respectively. $\hat{s}_i$ is the ith component of the enhanced power spectral vector. β is a flooring factor, which prevents $\hat{s}_i$ from being negative.

When delta coefficients $\mathbf{c}_t'$ and acceleration coefficients $\mathbf{c}_t''$ for tth frame are calculated from static coefficients $\mathbf{c}_t$ in the stage of training HMM, by the following equations:

$$\mathbf{c}'_t = \sum_{i=-W_d}^{W_d} d_i \mathbf{c}_{t+i} \quad , \qquad \mathbf{c}''_t = \sum_{i=-W_a}^{W_a} a_i \mathbf{c}_{t+i}, \qquad (9)$$

the mean vectors and covariance matrices in the stochastic feature for delta coefficients, ξ'_t, Ψ'_t, and acceleration coefficients, ξ''_t, Ψ''_t are calculated using the stochastic feature for static coefficients, ξ_t, Ψ_t, by the following equations:

$$\xi'_t = \sum_{i=-W_d}^{W_d} |d_i| \xi_{t+i} \quad , \qquad \mathbf{\Psi}'_t = \sum_{i=-W_d}^{W_d} (d_i)^2 \mathbf{\Psi}_{t+i}, \qquad (10)$$

$$\xi''_t = \sum_{i=-W_a}^{W_a} |a_i| \xi_{t+i} \quad , \qquad \mathbf{\Psi}''_t = \sum_{i=-W_a}^{W_a} (a_i)^2 \mathbf{\Psi}_{t+i}, \qquad (11)$$

where d_i and a_i are constant coefficients. W_d and W_a are window length of the calculation of delta and acceleration parameters. The above calculation for delta and acceleration parameters is based on the assumption that the noise signals in different frames are not correlated. Using the obtained stochastic features, the output probability of each state is calculated by Equation 7. Compared to spectral or cepstral mean substraction the introduction of $\mathbf{\Psi}$ in Equation 7 is the most important difference. The effect is that if the noise variance of a certain feature in the feature vector is very large, then this feature does not contribute much to the output probabilities of the HMMs, so the feature is more or less ignored.

4. EXPERIMENT AND RESULTS

The stochastic feature extraction technique described in the previous sections has been evaluated in speaker-independent isolated word recognition experiments under car noise conditions. Left-to-right tied-state triphone HMMs were trained using a clean speech database consisiting of sixty four speakers. The total number of states in the HMM was 2,547. All output distributions were mixtures of two diagonal covariance Gaussians.

The stochastic features were represented by a diagonal covariance Gaussian. For each frame, a set of means and variances for 13 MFCC coefficients, its delta and acceleration coefficients were computed as a stochastic feature. The mean and variance for the zeroth cepstral coefficient were not used in the recognition stage. The total number of parameters in the feature vector for each frame is seventy six. In the process of the calculation of the stochastic feature, several values of β of the spectral subtraction were tried.

The vocabulary size in the isolated word recognition task was 5,075. Six kinds of car noise were used. Table 1 shows the environments under which noise data was recorded. The clean speech data for test is composed of eight speakers' speech data with 303 isolated word utterances per speaker. The speech and noise data was recorded with the same microphone. The different sample in the noise data was added artificially to each clean utterance data by appropriate SNR. Recognition used Viterbi decoder with beam search. Speech boundaries were given manually so as to include 50 msec of non-speech portions at both of beginning and end.

Table 2 shows the results using the clean speech HMMs without compensation, and with non-linear spectral subtraction [4]. Table 3 shows the results using stochastic feature with varying the value of β, and noise observation duration of 200 frames (2,000 msec). In the results, the best performance was obtained by setting 10^{-1} as the value of β. This results show that the use of the stochastic features outperformed nonlinear spectral subtraction. The recognition error reduces between 13% and 44%. Particularly in the cases that background music existed (Environment 5 and 6), the performance improvement was fairly large. The effect of varying the value of β was consistent over all kinds of noise in the experiments. In additional experiments using spectral over-estimation scheme [3] in the spectral subtraction in the process of stochastic feature extraction, no improvement was found.

Next, the influence of changing the noise observation duration on the recognition accuracy was investigated. Table 4 shows results when the noise observation duration was varied from $N = 10$ frames (i.e. 100msec) to $N = 200$ frames (i.e. 2,000msec). It was found that the performance was almost the same in the range from 500msec to 2,000msec, although the performance degraded when observation periods of environmental noise were very short (less than 200msec). The performance degradation for short periods of noise observation can be considered to be caused by the small-sample-size effect in the process of the estimation of speech pdf. To compensate this effect, the variances of static parameters were heuristically magnified by 1.5 in the process of calculating stochastic features with noise observation duration of 200msec. Table 5 shows that magnifying variances improved the performance.

Environment	1	2	3	4	5	6
Car	A	A	B	C	D	D
Speed(km/h)	60	100	100	100	idle	100
Music	no	no	no	no	yes	yes
SNR(dB)	4.3	2.5	3.1	1.3	16.1	0.4

Table 1: Car noise used in the experiments

	Environment					
	1	2	3	4	5	6
NC	70.8	27.8	12.5	14.9	73.9	9.6
NSS	86.6	74.2	61.1	67.6	73.9	52.9

Table 2: Recognition rates (%) by using clean HMMs without any compensation method (NC), and with non-linear spectral subtraction (NSS)

β	Environment					
	1	2	3	4	5	6
10^{-1}	81.6	58.2	43.2	46.0	79.3	39.0
10^{-2}	86.6	69.8	57.4	60.9	84.9	54.4
10^{-3}	89.9	76.9	66.8	69.0	86.1	62.3
10^{-4}	90.0	78.9	69.1	71.8	85.3	65.5
10^{-5}	88.9	76.5	68.2	70.4	83.7	65.1
10^{-6}	87.2	73.8	65.6	68.1	81.2	63.9

Table 3: Recognition rates (%) using different value of β in the process of calculating stochastic features, and noise observation duration of 2,000 msec

time (msec)	Environment					
	1	2	3	4	5	6
2,000	90.0	78.9	69.1	71.8	85.3	65.5
1,500	90.0	78.5	69.4	71.5	85.0	65.9
1,000	89.9	77.9	70.1	72.2	84.8	66.0
500	89.9	77.1	69.2	71.0	83.6	65.2
200	88.9	74.9	67.1	69.2	80.2	63.3
100	85.9	71.7	63.2	66.0	74.0	57.7

Table 4: Recognition rates (%) using different observation duration for noise, and β of 10^{-4}

Environment					
1	2	3	4	5	6
90.3	76.6	68.2	70.3	82.1	64.1

Table 5: Recognition rates (%) with magnifying (×1.5) the value of the variance of static parameter in the process of calculating stochastic features, using noise observation duration of 200msec

5. CONCLUSIONS

In this paper, we have examined a new framework for speech recognition under noisy environment, where each feature has a stochastic form and represents the probability density function of the estimated speech component in the observed noisy speech. The technique has been shown to produce higher recognition rates under car noise environment than nonlinear spectral subtraction.

The stochastic feature approach has a strong theoretical relationship with the model combination approach. Both approaches calculate the observation probability of input noisy speech using the information on the noise distribution, but it is carried out in different ways. The comparison with the model combination approach should be investigated.

Presently, only preliminary experiments have been conducted. The more detailed investigation is necessary to know the potential of the proposed technique.

6. REFERENCES

[1] Gong Y., "Speech recognition in noisy environments: A survey", Speech Communication, Vol.16, pp.261-292, 1995.

[2] Boll S.F., "Suppression of acoustic noise in speech using spectral subtraction", IEEE Trans. ASSP, Vol.27, pp.113-120, 1979.

[3] Beronti V.L., Schwartz R. and Makhoul J., "Enhancement of speech corrupted by additive noise", ICASSP79, pp.208-211, 1979.

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

[5] Xie F. and Compernolle D.V., "Speech enhancement by spectral magnitude estimation - A unifying approach", Speech Communication, Vol.19, pp.89-101, 1996.

[6] Varga A.P. and Moore R.K. "Hidden Markov model decomposition of speech and noise", ICASSP90, pp.845-848, 1990.

[7] Gales M.J.F. and Young S. "An improved approach to the hidden markov model decomposition of speech and noise", ICASSP92, Vol.I, pp.233-236, 1992.

[8] Gales M.J.F. and Young S. "Cepstral parameter compensation for HMM recognition in noise", Speech Communication, Vol.12, pp.231-239, 1993.

[9] Nolazco Flores J.A. and Young S.J., " Continuous speech recognition in noise using spectral subtraction and HMM adaptation", ICASSP94, Vol.I, pp.409-402, 1994.

[10] Davis S.B. and Mermelstein P. "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences", IEEE Trans. ASSP, Vol.28, pp.357-366, 1980.