SPEECH RECOGNITION IN NOISY ENVIRONMENT USING WEIGHTED PROJECTION-BASED LIKELIHOOD MEASURE

Won-Ho Shin, Weon-Goo Kim†, Chungyong Lee, and Il-Whan Cha

A.S.S.P. Lab., Dept. of Electronic Eng., Yonsei Univ.
134 Shinchon-dong Sudaemoon-ku, Seoul 120-749, Korea
Phone: +82-2-361-2863, Fax: +82-2-312-4584, E-MAIL : swh@caas.yonsei.ac.kr
†Dept. of Electrical Eng., Kunsan Univ., Kunsan, Korea

ABSTRACT

This paper investigates a projection-based likelihood meaure that improves speech recognition performance in noisy environment. The projection-based likelihood measure is modified to give the weighting and projection effect and to reduce computational complexity. It is evaluated in sub-model based word recognition using semi-continuous hidden Markov model with speaker independent mode. Experimental results using proposed measure are reported for several performance factors: additive noise and noisy channel environment, various noise signals, and combination with other compensation method. In various noisy environments, performance improvements were achieved compared to the previously existing methods.

1. INTRODUCTION

The performance of a speech recognition system is usually degraded in open environments since there is always a mismatch between the training and testing environments. Therefore, the problem of speech recognition in a noisy environments has greatly attracted researcher's attention. A number of alternative methods to minimize these effects have been proposed in the literature and used with varying degrees of success. Some methods, such as spectral subtraction and filtering approach, try to estimate the underlying "clean" speech, whereas others, such as composite hidden Markov model(HMM)'s, attempt to incorporate noise statistics into the reference models themselves. On the other hand, a more programmatic approach such as distance measures or probability calculation focuses on the characteristics of the speech signal that are the least sensitive to degradation due to noise. For such an approach, the projection operations have been applied to mitigate this condition mismatch. The basic concept of these measures relies on the observations by Mansour and Juang[1] that norms of truncated cepstral sequences derived from LPC analysis are perturbed by more additive white noise than other descriptors of the feature space. They have presented a family of distortion measures based on the projection between LPC cepstral vectors. Carlson and Clements[2] have shown how the projection measure can be exploited as a weighted measure in a recognition system using continuous density HMM. Tung et.al.[3] have proposed projection-based group delay approach that combines the advantages of the projection measure and those of the group delay spectrum. However, the above researchers have mostly focused on projection operation. Therefore, previous likelihood measure did not give weighting effect by lifter, although Tung. et.al[3] have used weighted cepstrum using group delay spectrum. In this paper, previous projection-based likelihood measure is modified to give the weighting and projection effect. The proposed method is used with semi-continuous HMM.

The organization of this paper is as follows. Weighted projecion-based likelihood measure is described in Section 2. Section 3 explains recognition system and feature analysis as well as its experimental results. Finally, conclusions are given in Section 4.

2. WEIGHTED PROJECTION-BASED LIKELIHOOD MEASURE

2.1. Projection-based Distance Measure

Empirical observations have revealed the followings[1]:

1) at a given global SNR, the norm reduction on cepstral vectors with larger norm is generally less than on vectors with smaller norm.

2) lower order coefficients are more affected than higher coefficients.

3) the direction of the cepstral vector is less susceptible to noise contamination than the norm of the vector.

Based on these observations, they have proposed a family of distance measures based on the projection be-

This work was supported in part by Korea Telecommunication

tween the two cepstral vectors being compared. In their approach the shrinkage of the norm is compensated using the following distortion measure[1].

$$d_1(\lambda) = (c_t - \lambda c_r) * (c_t - \lambda c_r), \tag{1}$$

where c_t and c_r are test and reference observation, respectively, and projection value λ is determined from the orthogonality principle as follows

$$\lambda = \frac{c_t * c_r}{c_r * c_r}. \tag{2}$$

The robustness property of the angle between the clean cepstral vector and its noisy version is calculated using the following distortion measure[1].

$$d_2 = \left| \frac{c_t}{|c_t|} - \frac{c_r}{|c_r|} \right|^2 = 2(1 - \cos\beta), \tag{3}$$

where $\cos\beta$ is defined by

$$\cos\beta = \frac{c_t * c_r}{|c_t||c_r|}. \tag{4}$$

Also the distortion measure is modified using the property that cepstral vectors with larger norms are more robust to additive white noise than those with smaller norms[1]. That is

$$d_3 = \left| c_t - \frac{|c_t|}{|c_r|} c_r \right|^2. \tag{5}$$

2.2. Projection-based Likelihood Measure

Similar to the derivation of the projection-based distortion measure, a scale factor can be incorporated into the HMM state distribution or, equivalently, into the Gaussian likelihood score to compensate for the reduction in the vector norm. The parameters of each state in the model are represented by a continuous Gaussian probability density function of the form.

$$\begin{aligned} f_1(c_t) = \; & (2\pi)^{-\frac{N}{2}} |C_i|^{-\frac{1}{2}} \\ & \cdot \; \exp\left(-\frac{1}{2} (c_t - \lambda_{i,t}\mu_i)^T {C_i}^{-1} (c_t - \lambda_{i,t}\mu_i) \right) \end{aligned} \tag{6}$$

or its log-likelihood result:

$$\begin{aligned} \log f_1(c_t) = \; & (c_t - \lambda_{i,t}\mu_i)^T {C_i}^{-1} (c_t - \lambda_{i,t}\mu_i) \\ & - \frac{1}{2} \log |C_i| - \frac{1}{2} N \log \; 2\pi, \end{aligned} \tag{7}$$

where C_i, and μ_i are covariance matrix and mean code vector of i-th state, respectively. Similar to (2), the optimal $\lambda_{i,t}$ values is determined as follows

$$\lambda_{i,t} = \frac{{c_t}^T {C_i}^{-1} \mu_i}{{\mu_t}^T {C_i}^{-1} \mu_i}. \tag{8}$$

With this value of λ and without considering the last two terms, the log-likelihood becomes the following.

$$\begin{aligned} \log \tilde{f}_1(c_t) = \; & \left(c_t - \frac{{c_t}^T {C_i}^{-1} \mu_i}{{\mu_i}^T {C_i}^{-1} \mu_i} \mu_i \right)^T \cdot {C_i}^{-1} \\ & \cdot \left(c_t - \frac{{c_t}^T {C_i}^{-1} \mu_i}{{\mu_i}^T {C_i}^{-1} \mu_i} \mu_i \right). \end{aligned} \tag{9}$$

2.3. Weighted Projection-based Likelihood Measure Using Weighted Cepstrum

The weighted cepstrum distance measures have been suggested to improve the recognition rate for distorted speech[4, 5]. Therefore, to utilize the robust property of weighted cepstrum, the projection-based likelihood measure is weighted by the liftering function w. However, the weighted function w is canceled out because of covariance matrix, and returned to the original equation. That is

$$\begin{aligned} l_1(wc_t) = \; & \left(wc_t - \frac{(wc_t)^T w^{-1} {C_i}^{-1} w^{-1} (wu_i)}{(wu_i)^T w^{-1} {C_i}^{-1} w^{-}1 (wu_i)} wu_i \right)^T \\ & \cdot \; w^{-1} {C_i}^{-1} w^{-1} \\ & \cdot \left(wc_t - \frac{(wc_t)^T w^{-1} {C_i}^{-1} w^{-1} (wu_i)}{(wu_i)^T w^{-1} {C_i}^{-1} w^{-1} (wu_i)} wu_i \right) \\ = \; & \left(c_t - \frac{{c_t}^T {C_i}^{-1} u_i}{{u_i}^T {C_i}^{-1} u_i} u_i \right)^T \cdot \; {C_i}^{-1} \\ & \cdot \left(c_t - \frac{{c_t}^T {C_i}^{-1} u_i}{{u_i}^T {C_i}^{-1} u_i} u_i \right). \end{aligned} \tag{10}$$

Therefore, in our method covariance matrix is not considered and the likelihood measure is modified using the two previous distortion measures given in (3) and (5).

First, we used likelihood measure of normalized weighted cepstrum.

$$\begin{aligned} l_2(wc_t) = \; & \left(\frac{wc_t}{|wc_t|} - \frac{w\mu_i}{|w\mu_i|} \right)^T \cdot \; ({C_i}^N)^{-1} \\ & \cdot \left(\frac{wc_t}{|wc_t|} - \frac{w\mu_i}{|w\mu_i|} \right), \end{aligned} \tag{11}$$

where, ${C_i}^N$ is the covariance matrix of normailzed vector. Second, norm weighted likelihood measure similar to (5) was exploited, where covariance matrix is multiplied by weight function and so, the weight is canceled out. However, weight of projection value remains.

$$\begin{aligned} l_3(wc_t) = & \left(wc_t - \frac{|wc_t|}{|w\mu_i|}w\mu_i\right)^T \cdot w^{-1}C^{-1}w^{-1} \\ & \cdot \left(wc_t - \frac{|wc_t|}{|w\mu_i|}w\mu_i\right) \\ = & \left(c_t - \frac{|wc_t|}{|w\mu_i|}\mu_i\right)^T \cdot C^{-1} \\ & \cdot \left(c_t - \frac{|wc_t|}{|w\mu_i|}\mu_i\right) \end{aligned} \tag{12}$$

So, these likelihood measures can have weighting effect as well as projection effect. Also, they require less computation than the previous methods.

2.4. Implementation Issues

For a training procedure with semi-continuous HMM, several issue are involved in the practical use of the measure. The probability calculation of the observation is composed of two steps. First step is to choose some nearest code vectors using weighted projection distortion measure of section 2.1. Second step is to calculate weighted projection-based likelihood measure corresponding to the distance measure of the first step. However, these methods are used only with static observation, and for temporal observations, standard euclidean measure is used, because differential information dos not restore the directional information. To train the models, it is necessary to prepare for the initial codebook which are generated from the same measure by LBG algorithm. Especially for l_2 measure of (11), normalized observations are used to train the codebook.

3. EXPERIMENTAL RESULTS

3.1. Database and Recognition System

In this experiments Korean 14 digits(digit and command words for phone call) and 50 isolated words were used in speaker independent mode. For 14 digit database(DB) 100 noise free tokens of male speakers were used for training, whereas 40 different noise-contaminated tokens of each word were used for testing. Fifty words DB was composed of 150 noise free tokens and 42 different noise-degraded ones of each word for training and testing.

To generate features, speech was sampled with 8kHz for 14 digit DB and 10kHz for 50 word DB, and parameter analysis was performed on each 20ms frame of speech with Hamming window at every 10ms. For each speech frame, 18 or 20 channel filter bank spectra with mel-scaled frequency depending on the sample rate were obtained. Each speech spectral vector was then transformed to a cepstral vector. In addition a set of time differential features were generated and used as independent observations. Also, differential energy and differential-differential energy were used for energy feature vector.

To generate the noise contaminated speech of 14 digit DB, various signals including exhibition hall, computer room, and white noise were added to the speech waveform. Fifty word DB were recorded in telephone channel, and similarly various noise signals were mixed to the speech.

In all experiments, the words were modeled by an context-dependent phoneme using three states, multiple mixture, semi-continuous HMM with a diagonal covariance matrix. To model the background noise that is assumed to be present at start and end of an utterance, two noise models with two state were used. Each one was concatenated to front and end of speech model. The HMM structure was left-to-right with no skip states.

3.2. Results and Discussion

Table 1 shows the experimental results using 14 digit DB. To compare the weighting effect of liftering, we used euclidean(euc) and root power sum(rps) weight function. The "(euc)" and "(rps)" use the same standard likelihood measure. Only the difference between them is the chosen code vectors using weight function. Compared to the previous measure of "l_1" that has no weighting, the proposed normalized and norm weighted measure of "$l_2(rps)$" and "$l_3(rps)$" improved the performance. These two measures showed comparable performance, however "$l_3(rps)$" of norm weighted case gave a little better recognition rate in most cases. Therefore, from now on we'll consider only norm weighted "$l_3(rps)$" for experiment.

To see the usefulness of the proposed method, we observed the reognition results after model compensation. Table 2 shows the recognition results combined with parallel model combination(PMC) [6]. It significantly improved the performance compared to standard "(rps)" measure and showed similar or a little better reognition rate than "$PMC(rps)$". From the result we can see the proposed measure can be combined with other processing method. However, the performance improvement was not significant, because the performance has been already improved by the compensation method, and its processing affects the projection likelihood measure.

We also tested in the noisy channel environment. We used 50 word database recorded in telephone channel.

Table 1: Recognition results using weighted projection-based likelihood measures, where "(euc)" and "(rps)" means standard likelihood measures using code vector selection by weighting, "l_1" represents previous projection-based likelihood measure, and "$l_2(rps)$" and "$l_3(rps)$" describes the proposed measures using rps weight.

		(euc)	(rps)	l_1	$l_2(rps)$	$l_3(rps)$
clean		99.82	98.93	99.46	99.11	99.46
Exhibition Hall(dB)	20	91.25	98.39	96.43	96.43	97.50
	10	63.04	91.07	85.18	88.75	90.00
	0	15.89	49.64	44.46	50.18	50.71
Computer Room(dB)	20	95.36	96.07	97.14	96.61	96.79
	10	80.89	86.96	90.00	88.75	90.36
	0	35.18	51.07	60.54	50.18	60.00
White Noise(dB)	20	67.50	92.32	87.86	95.00	96.43
	10	38.04	75.89	59.11	87.86	92.32
	0	4.46	34.29	16.61	48.21	53.75

Table 2: Recognition results after applying PMC, where "(rps)" indicates standard likelihood measure, "$PMC(rps)$" represents standard likelihood measure with model compensation, and "$PMC\_l_3(rps)$" describes combination with the proposed measure.

		(rps)	$PMC(rps)$	$PMC\_l_3(rps)$
Exhibition Hall(dB)	20	98.39	97.50	97.32
	10	91.07	88.39	92.50
	0	49.64	38.39	56.79
Computer Room(dB)	20	96.07	98.04	97.50
	10	86.96	92.68	97.50
	0	51.07	57.68	66.61
White Noise(dB)	20	92.32	97.50	96.68
	10	75.89	93.04	93.04
	0	34.29	69.11	69.29

To compensate for channel characteristic, cepstral mean subtraction(CMS) technique was used. Table 3 shows its recognition results. It shows that the proposed measure improved the performance compared to the standard and previous measure. The performance improvement was 21.34% at 10dB of exhibition hall noise, although a little degradation was observed compared to standard "rps" measure. In other noise cases, it showed consistent improvement.

4. CONCLUSION

This paper proposed and evaluated a weighted projection-based likelihood measure for semi-continuous HMM's. The proposed measure improved the performance of speaker independent, isolated word recognition in the presence of several noise types and channel environment compared to previous method. It doesn't require much computation and its usefullness was comfirmed through the combination of the other noise processing like PMC.

Table 3: Recognition results using mel-cepstrum with CMS for telephone channel, where "(euc)" and "(rps)" indicates standard likelihood measure, and "l_1" and "$l_3(rps)$" represents previous and the proposed measure, respectively.

		(euc)	(rps)	l_1	$l_3(rps)$
clean		97.95	97.90	97.81	97.29
car noise(dB)	20	91.24	96.33	94.62	96.95
	10	74.05	86.00	64.14	85.48
road-side noise(dB)	20	94.29	95.52	96.38	96.05
	10	63.00	70.33	71.86	72.05
white noise(dB)	20	96.76	96.71	94.43	97.52
	10	76.19	74.05	78.00	78.05

5. REFERENCES

[1] D. Mansour and B. H. Juang, "A Family of Distortion Measure Based upon Projection Operation for Robust Speech Recognition," *IEEE Trans. Acoust., Speech, Siganl Processing,* vol. 37, no. 11, pp. 1659-1671, Nov. 1989.

[2] B. A. Carlson and M.A. Clements, "A Projection-Based Likelihood Measure for Speech Recognition in Noise," *IEEE Trans. Corr. Speech and Audio Processing,* vol. 2, no. 1, pp. 97-102, Jan. 1994.

[3] S. L. Tung, I. S. Lei and Y. T. Juang, "Projection-Based Group Delay Scheme for Speech Recognition," *IEEE Trans. Corr. Speech Audio Processing,* vol. 4, no. 2, Mar. 1996.

[4] J. Junqua and H. Wakita, "A Comparative Study of Cepstral Lifters and Distance Measures for All Pole Models of Speech in Noise," *Proc. ICASSP,* pp. 476-479, May 1989.

[5] Y. Tohkura, "Weighted Cepstral Distance Measure for Speech Recognition," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. 35, no. 10, pp. 1414-1422, Oct. 1987.

[6] M. J .F. Gales, S. J. Young, "Cepstral Parameter Compensation for HMM Recognition in Noise," *Speech Communication,* vol. 12, no. 3, pp. 231-239, 1993.