J. of Computer Science and Technology, 12(2): 177-184, Mar., 1997 (Accepted in EI)

A Log-Index Weighted Cepstral Distance Measure for Speech Recognition

***Zheng Fang* (郑方), *Wu Wenhu* (吴文虎), *and Fang Ditang* (方棣棠)**
Department of Computer Science and Technology, Tsinghua Univ., Beijing, 100084
fzheng@sp.cs.tsinghua.edu.cn, (010)62784141

Abstract —— A log-index weighted cepstral distance measure is proposed and tested in speaker-independent and speaker-dependent isolated word recognition systems using statistic techniques. The weights for the cepstral coefficients of this measure equal to the logarithm of the corresponding indices. The experimental results show that this kind of measure works better than any other weighted Euclidean cepstral distance measures across three speech databases. The error rate obtained using this measure is about 1.8 percent for three databases on an average, which is a 25% reduction of that obtained using other measures, and a 40% reduction of that obtained using Log Likelihood Ratio (LLR) measure. The experimental results also show that this kind of distance measure works better in both speaker-dependent and speaker-independent speech recognition systems.
Keywords —— Log-Index Weighted Cepstral Distance Measure, Speech Recognition,

I. Introduction

The cepstral distance measure is one of the most important issue in speech recognition based on template matching, and many distance measures have been proposed. Among them, the LPC-based log likelihood ratio (LLR) distance measure proposed by Itakura [1] has been one of the most successful measures. Another important distance measure is the Euclidean distance measure, which is widely used with LPC-derived cepstral coefficients.

The Euclidean cepstral distance measure has a large number of variants, for it is an approximation to the distance between the two log spectra represented by the cepstral coefficients[2].

One widely used variant of the cepstral distance measure is a weighted cepstral distance measure. Furui [3] used such a weighted cepstral distance measure for automatic speaker verification, where the weight for the cepstral coefficients was the inverse of its intratalker variance. Paliwal [4] applied a weighted cepstral distance measure to vowel recognition and got a 1.3 percent recognition rate average improvement from 91.4 to 92.7, where the measure used in his experiments was the statistically weighted Euclidean distance measure with vowel class specific weights. Tohkura [5] studied the weighted cepstral distance measure on three isolated digit databases, reducing the error rate to one-fourth of that obtained using the simple Euclidean cepstral distance measure and about one-third of that using the log likelihood ratio (LLR) distance measure. Juang et al [6] used the liftering process to achieve an average error of 1 percent in a speaker-independent isolated digit test, the error rate was about one-half that obtained without the liftering process.

Based on the previous studies on weighted distance measure, it appears that weighting works well, but we have not clear explanation on the reasons why and how it works and how to choose an optimal set of weights yet. The purpose of this paper is to show that well-chosen weighted distance measure can lead to substantial performance improvement in speech recognition.

This paper is organized as follows. In Section II, we introduce the speech recognition model used in our experiments. In Section III, we discuss several weighted Euclidean distance measures for cepstral coefficient sets to be compared in our experiments. In Section

IV, three databases are described and in Section V the experimental results are listed. Finally, we draw conclusions in Section VI.

II. The Speech Recognition Scheme Used

The speech recognition model we used in the experiments is a kind of statistic model, named Center-distance Continuous Probabilistic Model (CDCPM) [9].

Say a random variable ξ with a normal distribution $N(\mu,\sigma)$. The probabilistic density function (p.d.f.) of ξ is:

$$f_{\mu-\sigma}(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/2\sigma^2} \tag{1}$$

Denote the distance between ξ and μ by another random variable η, then its p.d.f. can be written as:

$$f_{\sigma}(y)=\frac{2}{\sqrt{2\pi}\sigma}e^{-y^2/2\sigma^2},y\geq 0 \tag{2}$$

we name this kind of distribution the center-distance normal (CDN) distribution - CDN(σ). The mean of η can be calculated as $\mu_\eta=\frac{2\sigma}{\sqrt{2\pi}}$.

Our model is based on this distribution. Each utterance for the word in the vocabulary will be segmented into several segments corresponding to several states, using Non-Linear Segmentation (NLS) technique [7]. For the specified word, each state (segment) can be represented by several CDN distributions. This is something like the mixed Gussian Hidden Markov Model. In our experiments, we use 2 distributions to stand for the feature space of each state. Therefore, we name this kind of model as Center-Distance Continuos Probabilistic Model (CDCPM) [8,9].

When evaluating the distance between one feature vector and the center vector, we use weighted Euclidean distance measure.

III. The Weighted Euclidean Distance Measures

The famous Itakura distance measure Log Likelihood Ratio (LLR) is defined as follows:

$$d_{LPC}(\mathbf{a}_R,\mathbf{a}_T)=\log\left|\frac{\mathbf{a}_R R\,\mathbf{a}_R^T}{\mathbf{a}_T R\,\mathbf{a}_T^T}\right| \tag{3}$$

where $\mathbf{a}_T=(1,a_{T1},\cdots,a_{Tp})$ and $\mathbf{a}_R=(1,a_{R1},\cdots,a_{Rp})$ are feature row vectors composed of the linear predictive coefficients obtained from a test utterance and a reference one, respectively, and R is the autocorrelation matrix (obtained from the test sample) corresponding to $\mathbf{a}_T$. LLR measure is tested in our experiment only for comparison purpose.

In our experiments, we mainly use the weighted Euclidean distance to measure the distance between two feature vectors. The general form of weighted Euclidean distance measures can be written as:

$$d_{CEP}^2=\sum_{i=1}^{p}w_i(c_T(i)-c_R(i))^2 \tag{4}$$

where c_T and c_R are p-dimensional feature row vectors which are composed of the cepstral coefficients obtained from a test utterance and a reference one, respectively, and w_i is the weight of the *i*th component, different sets make different measures.

The most commonly used measure is the quefrency weighted cepstral distance measure, which is one form of weighted cepstral measure and has previously been applied to vowel recognition experiments [4].This kind of measure has the following form:

$$d^2_{QCEP} = \sum_{i=1}^{p} (i\,c_T(i) - i\,c_R(i))^2 = \sum_{i=1}^{p} i^2 (c_T(i) - c_R(i))^2 \quad (5)$$

where $w_i = i^2$. This kind of measure is also referred to as the index-weighted cepstral distance measure or triangular-weighted cepstral distance one.

Another form of weighted cepstral weighted measure is the widely used Mahalanobis distance, which is define as follows:

$$d^2_{MCEP} = (c_T - c_R) V^{-1} (c_T - c_R)^T \quad (6)$$

where $\mathbf{V} = (v_{ij})$ is the covariance matrix of the feature vectors. This measure can be used to clustering and recognition purposes.

There are some difficulties in calculating the inverse of the covariance matrix. Our solution is to use the diagonal part of the covariance matrix V. In this sense, the covariance weighted distance measure is described by the following equation:

$$d^2_{MCEP} = \sum_{i=1}^{p} v_{ii}^{-1} (c_T(i) - c_R(i))^2 \quad (7)$$

where $w_i = v_{ii}^{-1}$.

The third kind of measure is to use the raised sine lifter. The weight function can be defined as[10]:

$$w_i = \left[1 + \frac{p}{2} \sin(\frac{\mathrm{i}\pi}{p}) \right]^2, 1 \le i \le p \quad (8)$$

In our experiments, we test another form of weight function, which is defined as:

$$w_i = \left[\ln(ci + 1) \right]^2, 1 \le i \le p \quad (9)$$

where c is a constant number.

IV. DATABASES

4.1 Cepstral Analysis

We adopt the following steps for cepstral analysis in our experiments: (1) Speech is first filtered typically to a bandwidth of 3400Hz and then digitized typically at 8KHz sampling rate, or first filtered typically to a bandwidth of 6800Hz and then digitized typically at 16KHz sampling rate. (2) The digitized speech is then emphasized using a simple first-order digital filter with transfer function $H(z) = 1 - 0.95z^{-1}$. The preemphasized speech is then blocked into frames of 32 msec in length spaced every 16msec. (3) Each frame of speech is weighted by the Hamming Window. (4) The linear predictive coding (LPC)analysis is then performed on each frame using Levinson-Durbin recursive algorithm[11]. (5) LPC cepstral coefficients c[i] are computed from the *p*th-order Linear Predictor Coefficients a[i] by the following equations[12]:

$$c[1] = a[1]$$

$$c[n] = a[n] + \sum_{m=1}^{n-1} \frac{m}{n} a[m] c[n-m], \quad 2 \le n \le p \quad (10)$$

$$c[n] = \sum_{m=1}^{p} \frac{n-m}{n} a[m] c[n-m], \qquad n > p$$

4.2 Database Description

In order to evaluate the performance of the weighted cepstral distance measure, three word vocabularies are used. The first one is a small size vocabulary consisting of the ten Chinese digits (0-9), and the second is a medium size vocabulary consisting of 35 Chinese finals, while the third is another medium size vocabulary consisting of 128 Chinese phrases (3 to 4 Chinese characters per phrase). These databases have the following descriptions:

Database I (DB-I): 2800 isolated digit utterances spoken by 14 male speakers. Each speaker uttered 0-9 twice. The sampling frequency is 16KHz and the cutoff is 8KHz. The feature vectors calculated from these utterances were 16th-order cepstral coefficients derived from 12th-order LPC coefficients.

Database II (DB-II): 840 isolated Chinese finals uttered by 1 male speaker. The sampling frequency is 8KHz and the cutoff is 3.4KHz. The feature vectors calculated from these utterances were 16th-order cepstral coefficients derived from 12th-order LPC coefficients.

Database III (DB-III): 2560 Chinese phrase utterances spoken by 20 male speakers. Each speaker uttered the vocabulary once. The sampling frequency is 8KHz and the cutoff is 3.4KHz. The 20 speakers were from almost different provinces around China, and the utterances were spoken in Mandarin with different accents. The feature vectors calculated from these utterances were 10th-order cepstral coefficients derived from 10th-order LPC coefficients.

DB-I and DB-III were used for speaker-independent testing, and DB-II for speaker-dependent testing. The first half of each database is used as training set while the second half the testing set.

V. Experiments

5.1 Statistical Characteristic of the Cepstral Coefficients

The following three figures show the statistical variances as a function of the cepstral coefficient indices for the three databases. The variance values in the figures are relative value. From Fig.1 we can find that the variance trends to decrease with the cepstral coefficient index.

5.2 Test on Reverse-index Weights

In order to compare the importance for lower and higher cepstral coefficients, we did three experiments using the following three different weight functions, respectively:

(1) Index weights

$$w_i = i^2, 1 \le i \le p$$

(2) Equal weights

$$w_i = 1, 1 \le i \le p$$

(3) Reverse-index weights

$$w_i = (p + 1 - i)^2, 1 \le i \le p$$

The results were that, the error rates obtained through the three experiments identically increased in the following order for any databases:

Error (1) < Error (2) << Error (3)

Therefore an obvious conclusion can be drawn as 揾igher cepstral coefficients should be emphasized more strongly than lower ones to get higher recognition rate."

5.3 The Way to Emphasize Higher Cepstral Coefficients

It is true that the higher cepstral coefficients should be emphasized more than the lower ones, but how ? To find the truth, we designed another three sets of weights:

(1) Exp-weights:

$$w_i = (\exp(i) - 1)^2, 1 \le i \le p$$

(2) Index-weights:

$$w_i = i^2, 1 \le i \le p$$

(3) Log-weights:

$$w_i = (\ln(i+1))^2, 1 \le i \le p$$

The results were that the error rates decreased in the following order for any database:

Error (1) >> Error (2) > Error (3)

Obviously, the higher coefficients should be deweighted, the curve of weight function should just bend to the index-axis. (See appendix.)

To find the relationship between the error rate and the bending extent of the weight curve, we define:

$$w_{ci} = (\ln(ci+1))^2, 1 \le i \le p$$

where c is a bending constant. Experiments showed that the error rate increased when c varied from 1 upwards or downwards.

5.4 Performance Comparison for Several Weight Types

The Tab. 1 gives the experimental results for several weighted Euclidean cepstral distance measures as well as LLR-based distance measure for the three databases described above, the shown results are based on the testing sets. From Tab. 1, we find that the log-index weighted Euclidean cepstral distance measure as well as the inverse-variance weighted one is better than others.

VI. Summary

In this paper the log-index weighted cepstral distance measure with weighting coefficients set equal to the logarithm of the coefficient indices has been studied, with comparison to several other measures.

Through several experiments, we summarize our experimental results and findings as follows:

1. The log-index weighted cepstral distance measure works substantially better than both the Euclidean cepstral distance and any other measure across three different databases.
2. The most important feature of the weighting is that it weights the higher order cepstral coefficients more strongly than the lower order ones.
3. The weighting need not weight the higher order cepstral coefficients so strong as the index weighted distance measure does.
4. With respect to the relationship between the recognition rate and the bending extent of the weight function, the recognition performance is better when the bending constant c=2,1,or 1/2.

Appendix

The CDCPM is sensitive to the ratio between each weight coefficient instead of the absolute value of each weight. We will prove this truth as follows:

Assume that two sets of weights, namely w1 and w2, satisfy w2[i] = k w1[i], 1<=i<=p, where k is a constant. The weighted Euclidean distances for any two sets of cepstral coefficient row vectors c1 and c2 will satisfy $d_{w2}^2(c1, c2) = k\, d_{w1}^2(c1, c2)$. So the model parameters for two sets of weights will satisfy $\sigma_{w2}^2 = k\, \sigma_{w1}^2$, In this case, the values of p.d.f. at point x will be:

$$f_{\sigma_{w2}}(d_{w2}(x,\mu)) = \frac{2}{\sqrt{2\pi}\,\sigma_{w2}} \exp(-d_{w2}^2(x,\mu)/2\sigma_{w2}^2)$$
$$= \frac{2}{\sqrt{2\pi}\sqrt{k}\,\sigma_{w1}} \exp(-d_{w1}^2(x,\mu)/2\sigma_{w1}^2)$$
$$= \frac{1}{\sqrt{k}} f_{\sigma_{w1}}(d_{w1}(x,\mu))$$

The constant $\frac{1}{\sqrt{k}}$ will not affect the recognition rate.

REFERENCES

[1] **F. Itakura**, Minimum prediction residual principle applied to speech recognition, *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp.67-72, Feb. 1975

[2] **N. Nocerino, F.K. Soong, L.R. Rabiner, and D.H. Klatt**, Comparative study of several distortion measures for speech recognition, in Proc. *ICASSP 1985*, vol. 1, Mar. 1985, pp.25-28

[3] **S. Furui**, Cepstral analysis technique for automatic speaker verification, *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp.254-272, Apr. 1981

[4] **K.K. Paliwal**, On the performance of the quefrency-weighted cepstral coefficients in vowel recognition, *Speech Commun*., vol. 1, pp.151-154, May 1982

[5] **Y. Tohkura**, A weighted cepstral distance measure for speech recognition*, IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, No.10, Oct. 1987, pp.1414-1422

[6] **B.H. Juang, L.R. Rabiner and J.G. Wilpon**, On the use of bandpass liftering in Speech Recognition, *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, No.7, July 1987, pp.947-953

[7] **JIANG Li, WU Wenhu, CAI Lianhong, and FANG Ditang**, A Real-time Speaker-independent Speech Recognition System Based on SPM for 208 Chinese Words, in Proc. *ICSP'90*, pp.473-476, 1990

[8] **ZHENG Fang, YANG Hongbo, WU Wenhu, and FANG Ditang**, A Continuous Distance Density Segmental Probabilistic Model, in Proc. *National Conference on Man-Machine Speech Communication (NCMMSC'94), Speech Recognition and Synthesis*, pp.238-241, Oct. 1994 (in Chinese)

[9] **F. Zheng, W.H. Wu, and D.T. Fang**, The CDCPM with applications to speech recognition, already accepted by *Chinese J. of Advanced Software Research*, 1996 (in Chinese)

[10] **B.H. Juang, L.R.Rabiner, and J.G.Wilpon**, On the use of bandpass liftering in speech recognition, *IEEE Trans. on ASSP*, vol. ASSP-35, pp.947-953, Oct. 1987

[11] **J. Makhoul**, Linear Prediction: A Tutorial Review, Proc. *IEEE*, vol. 63, pp.562-580, Apr. 1975

[12] **B. Gold and C.M. Rader**, Digital Processing of Signals, *New York: McGraw-Hill*, 1969, P246

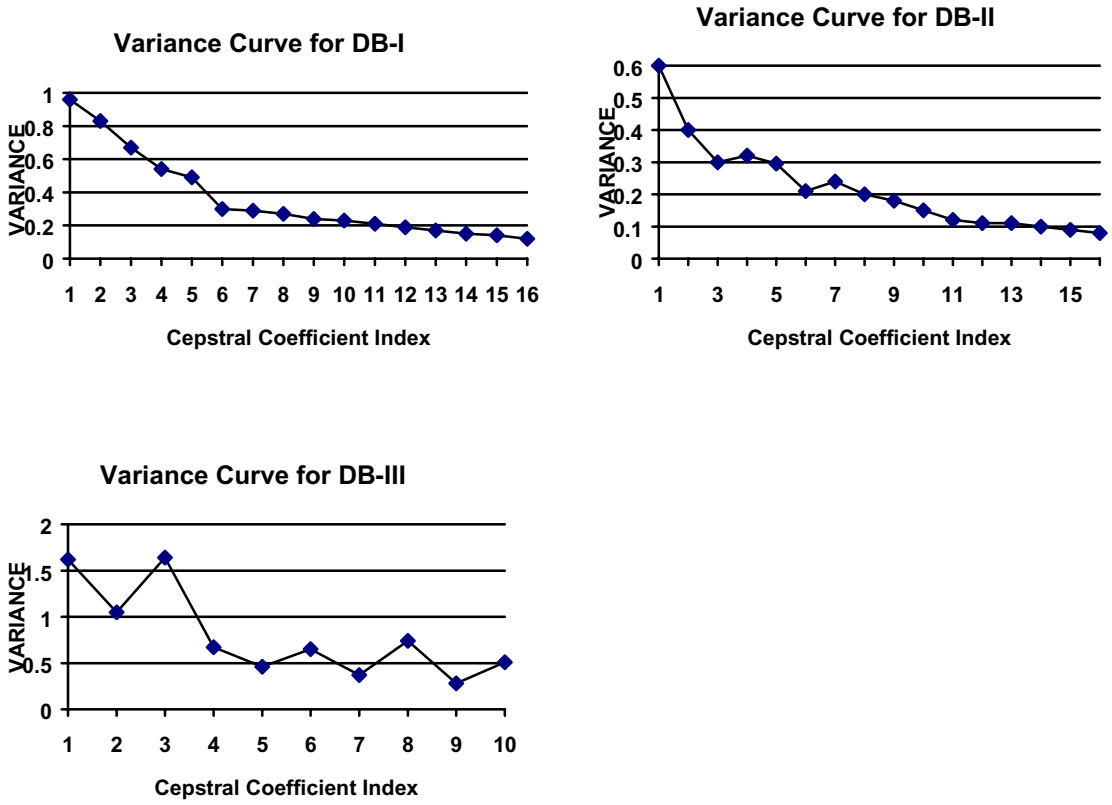

Fig. 1 The variance curve of the cepstral coefficients

Tab. 1 Error rates of 7 different weight types

weight type	LLR-based	index	equal	inv-var	raised sine	ln(2i+1)	ln(i+1)	ln(i/2+1)
DB-I	3.00	2.75	4.25	**1.75**	3.00	2.50	2.75	2.50
DB-II	1.55	0.95	1.19	1.43	0.83	0.83	**0.60**	**0.60**
DB-III	4.49	2.85	3.95	2.85	3.36	2.07	**1.95**	2.27

作者英文简介:

Zheng Fang was born in Jiangsu Province, P.R.China, in 1967. He received the B.S. degree and the M.S. degree from Tsinghua Univ., P.R. China, both in computer science and technology, in 1990 and 1992, respectively.
He is now both a lecturer and a Ph.D. student in Tsinghua. He is also the executive director of the Analog Devices Inc.-Tsinghua DSP Technology Research Center.
He has been working in Speech Recognition at Speech Lab., Dept. of Computer Science and Technology, Tsinghua, since 1988.

Prof. Wu Wenhu was born in Beijing, P.R.China, in 1936. He studied in the Department of Electrical Engineering, Tsinghua University, from 1955 to 1958, and then in the Department of Automation, Tsinghua University, from 1958 to 1961.
Since then, he has been teaching at Tsinghua University and now a Full Professor in the Department of Computer Science and Technology. He is the director of the Speech Lab now.
He is devoted in researching Chinese speech recognition and understanding, especially the speaker-independent Chinese speech recognition. As a result, he has been awarded several times.
He is also devoted in the computer spread education. He is the chairman of Computer Spread Education Commission of CCF (China Computer Federation). He has led the China Team to take part in the IOI'89 - IOI'95 (International Olympiad in Informatics) and won many golden medals.

Prof. Fang Ditang was born in Shanghai, P.R.China, in 1930. He received the B.S. degree from Jiaotong University and the M.S. degree from Tsinghua University, both in electrical engineering, in 1953 and 1956, respectively.
Since then, he has been teaching at Tsinghua University and now a Full Professor in the Department of Computer Science and Technology. In 1979, he founded the Laboratory for Human-Machine Speech Communications and has been its director from 1979 to 1990. The laboratory received the National Scientific Research and Technology Progress Award twice, in 1987 and 1989, respectively, the National Scientific Invention Award in 1990, and three other awards.
He is the Deputy Chief of the Artificial Intelligence and Pattern Recognition Committee of the Chinese Computer Science Society.