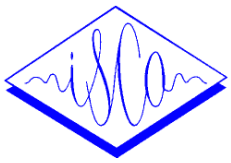

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

LINE SPECTRUM PAIR FREQUENCY - BASED DISTANCE MEASURES FOR SPEECH RECOGNITION

Fikret S. Gurgen, S. Sagayama*, Sadaoki Furui

NTT Human Interface Laboratories
Speech and Acoustics Lab.
9-11, Midori-Cho 3-Chome
Musashino-Shi, Tokyo, 180 Japan

ABSTRACT

In the present study, the performance of the line spectrum pair (LSP) frequencies representation for speech recognition is investigated. Various distance measures such as Euclidean, inverse variance weighted Euclidean, and Mel-scale-like weighted distance measures based on the LSP frequencies are used for speaker-independent isolated word recognition experiments with a Dynamic Time Warping (DTW) system. Transitional LSP frequency parameters defined by regression coefficients of LSP frequencies are also introduced. The transitional and the instantaneous parameters and distances are linearly combined for better recognition performance. The cepstral distance measures, and transitional and instantaneous cepstral parameters and distances are used for the comparison of the performances.

The linear combination of the instantaneous and the transitional parameters for LSP representation is found to be the best among the all distances used in the experiments.

INTRODUCTION

Line Spectrum Pair (LSP) frequencies representation of speech signal has recently been introduced by Itakura [1] for the purpose of maintaining voice quality at smaller bit rates. This new representation functions in the frequency domain and is an alternative linear predictive coding (LPC) representation. Various researchers have made use of this representation in various speech applications [10,11]. In the speech coding area, this representation is found to be better than LPC parametric representation such as Linear predictive coefficients, and PARCOR coefficients because LSP parameters produce smaller amount of distortion when they roughly quantized and linearly interpolated [2]. LSP parameters have both well-behaved dynamic range and filter stability preservation property and, therefore, they can be used to encode LPC spectral information more efficiently than any other LPC parameters. The experimental studies have been reported that LSP parameters can give the same spectral distortion by roughly 80% of the quantization bit rate compared with PARCOR coefficients. In speech enhancement area, LSP parametric representation has been used as an efficient and direct procedure for the imposition of constraints upon the sequence of speech spectra in sequential maximum a posteriori estimation.

Linear predictive coding (LPC) parametric representation has been extensively used in the speech recognition in the recent years. The most common LPC distance measures are the maximum likelihood ratio measure (Itakura distance), LPC cepstrum distance (CEP), weighted likelihood ratio (WLR) [4], weighted cepstrum distance, Cosh measure. Later, a transitional cepstrum measure was also proposed [6,7], and the transitional and the instantaneous distances are then linearly combined for a better recognition performance. A number of papers and reports showing the performance of the LPC based distance measures have been published by various researchers [4,5,6,7].

In the present paper, the performance of the line spectrum pair (LSP) frequency representation for speech recognition is reported. In the literatures, speaker-dependent and independent recognition experiments by using some LSP based distance measures have been reported [8,9]. In this study, we introduce the distance measures based on the linear combination of the transitional and the instantaneous LSP frequencies and compare the performances with those of cepstrum coefficients. As a general word recognition system, dynamic programming (DTW) based word recognition system is used. The database for this system consists of 216 phonetically balanced Japanese words. Therefore, variety of sounds may give better evaluation result for the performance of the distance measures, and the resulting performance will not be affected by the certain group of sounds.

LINE SPECTRUM PAIR REPRESENTATION

The line spectrum pair (LSP) parameters function in the frequency domain. The LSP method represents the speech spectral envelope through a distribution density of discrete frequencies. The cumulation of two or more frequencies on a certain region shows the occurrence of the resonances.

The LSP analysis is also based on the all-pole model. As it is known, in the LPC analysis, a short time frame of speech signal is assumed to be produced from the output of a time-invariant linear all-pole filter $H(z)=1/A_m(z)$ ($A_m(z)$ is the inverse filter). $A_m(z)$ is described by,

$$A_m(z)= 1+a_1 z^{-1} +.....+a_m z^{-m}$$

where $\{a_m\}$ are the LP coefficients and m is the order of the analysis. The inverse filter polynomial can be decomposed into two polynomials:

$$A_m(z)=1/2\ [P(z)+Q(z)]$$

which

* now with ATR laboratory.

$$P(z) = A_m(z) + z^{-(m+1)} A_m(z^{-1})$$
$$Q(z) = A_m(z) - z^{-(m+1)} A_m(z^{-1}).$$

P(z) and Q(z) are symmetric and asymmetric polynomials and have the following important properties:
(1) All zeros of P(z) and Q(z) lie on the unit circle,
(2) Zeros of P(z) and Q(z) are interlaced with each other, and using (1) and (2), it is shown that,
(3) Minimum phase property of $A_m(z)$ is easily preserved after quantization of the zeros of P(z) and Q(z) [10].

The zeros of P(z) and Q(z) which are on the unit circle can be shown as $e^{-j\omega}$ and the ω values are known as the LSP frequencies.

SOME OF LSP DISTANCE MEASURES

In the present paper, the concept of pattern recognition by minimum distance classifier is used. The motivation for using minimum distances as a recognition tool follows naturally from the fact that the most obvious way of establishing a measure of similarity between pattern vectors is by determining their proximity. For this purpose, the usage of various distance measures are considered.

One of the extensively used distance measures for speech recognition is Euclidean distance measure. The Euclidean distance measure between the LSP vectors is

$$d_{LSP} = \sum_{i=1}^{m} (x_i - y_i)^2$$

where x_i and y_i are LSP frequencies of input and reference vectors, respectively. m shows the dimension of the vectors. This distance is one of the simplest and most commonly used distance in the pattern recognition.

Weighted Euclidean distance measure is defined by using the inverse variance weight values which is computed statistically from the speech data [8,9].

$$d_{WLSP} = \sum_{i=1}^{m} w_i (x_i - y_i)^2$$

where $\{w_i\}$ show the inverse variance weight values.

Dynamic spectral features (spectral transition) as well as instantaneous spectral features are shown to be important in the speech recognition by Furui [6,7]. In these studies, the correspondence between the Fourier transformation of the first derivative for the cepstrum coefficients and the first derivative for the logarithmic spectral envelope has been shown and the idea of the emphasizing spectral dynamics for the speech recognition has been proposed. The regression coefficient for each cepstrum coefficient, which gives a reliable estimation of the derivative of parameter [3], has been computed.

In this study, we make use of the LSP frequencies to show the effectiveness of transitional parameters on the speech recognition. The spectral envelope transitions are modeled with the derivatives of the LSP frequencies. The derivatives of the LSP frequencies is also defined with the regression parameters. When the transitional and the instantaneous distances are linearly combined, the following distance measure is obtained:

$$d_{LSP} + d_{\Delta LSP} = \sum_{i=1}^{m} (x_i - y_i)^2 + \sum_{i=1}^{m} w_1 (\Delta x_i - \Delta y_i)^2$$

When the instantaneous and transitional parameters are linearly combined, a new feature vector can be defined [7]. Using this new vector the following distance is obtained.

$$d_{LSP+\Delta LSP} = \sum_{i=1}^{m} \{ (x_i + w_1 \Delta x_i) - (y_i + w_1 \Delta y_i) \}^2$$
$$= \sum_{i=1}^{m} \{ (x_i - y_i) + w_1 (\Delta x_i - \Delta y_i) \}^2$$
$$= d_{LSP} + d_{\Delta LSP} + 2 \sum_{i=1}^{m} w_1 (x_i - y_i)(\Delta x_i - \Delta y_i)$$

A bilinear transformed LSP frequency based distance measure is also proposed for the improvement of the recognition performance [5]. The bilinear transform is a method to transform the linear frequency axis into a warped one by the use of an all-pass filter. The all-pass filter equation is

$$\omega_{warp} = \omega + 2 \tan^{-1}[(a \sin\omega)/(1 - a \cos\omega)]$$

where ω shows linear frequency values, ω_{warp} shows warped frequencies, and a is a parameter for warping. The bilinear transform converts the linear frequency axis into a Mel-scale-like frequency axis by lengthening the low frequency axis.

SPEECH RECOGNITION EXPERIMENTS

Speaker-independent, dynamic time warping based isolated word recognition system (Figure 1) is used as the workbench of the study. 216 words (NTT database) which were uttered by three male speakers are used. The sampling rate of input utterances is 12 kHz. Each speech frame is extracted every 10 ms with 30 ms Hamming window and converted into acoustic feature parameters. 10^{th} order and 16^{th} order instantaneous parameter sets (LSP frequencies and LPC cepstrum coefficients) are derived. For the transitional parameters, the regression coefficients are computed for the adjacent nine frames and combined with the instantaneous parameters. For example, in case of the 16^{th} order parameter set case, the derived acoustic feature parameters are 16 LSP frequencies and 16 ΔLSP's, and 16 LPC cepstrum coefficients and 16 ΔLPC cepstrum coefficients

The effectiveness of each distance measure is examined by the cross-talker word recognition experiments. Each of the 216 word set uttered by each of three male speakers is taken as the reference set and the recognition experiments are performed by using each of the other two utterance sets. As a result, the total number of the recognition experiments for each distance measure is six and the average performance of these six talker recognition experiments is taken.

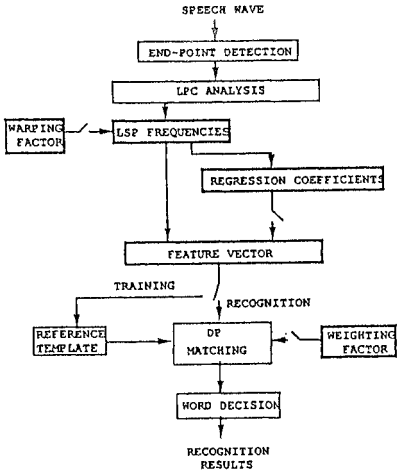

FIGURE 1 Block diagram indicating principal operations of word recognition system

EVALUATION OF DISTANCE MEASURES

The performances of various LSP frequency based distance measures were compared with conventional LPC parameter based distance measures such as the Euclidean cepstrum (CEP), the weighted (inverse variance) Euclidean cepstrum, and the linear combination of the instantaneous and the transitional cepstrum coefficients (regression coefficients).

The LSP based Euclidean distance measure was evaluated first and the recognition experiments were conducted with 10^{th} order LPC analysis (with 10 LSP frequencies) and 16^{th} order LPC analysis (with 8, 10, 12, and 16 LSP frequencies) (figure 2) cases. The 16^{th} order analysis has clearly given the better recognition results than those of 10^{th} order case. Thus, the LSP based recognition experiments were continued with the 16^{th} order coefficients set. When the reduced number of LSP frequencies in the 16^{th} order analysis were used, the other frequencies were assumed to be zero (in other words, the bandwidth of the speech signal was made narrower than before). It is observed that high frequency components reduce the recognition performance. From the results, the first 10 LSP frequencies were decided to be used for the rest of the recognition experiments because of the better performance with less number of the frequencies.

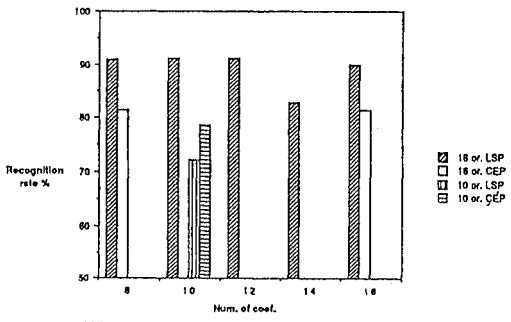

FIGURE 2 Recognition rates for LSP and ceps. coefficients

The Euclidean LPC cepstrum distance (CEP) was also tried with the 10^{th} (with the all 10 coefficients) and 16^{th} order (with the different number of the coefficients such as 8, 16) analysis and was found that the first 8 cepstrum coefficients in the 16^{th} order analysis give the best recognition result among them (figure 2). The rest of the recognition experiments with the cepstrum coefficients were conducted with these 8 coefficients. When the reduced number of the LPC cepstrum coefficients were used, the other coefficients were set to zero (in other words, the spectrum estimate was smoothed in a certain degree). As a result, the Euclidean distance measure defined on LSP frequencies was found to be clearly superior to CEP.

The first 10 LSP frequencies of the 16^{th} order analysis has become overwhelmingly superior to the 10 LSP frequencies of 10^{th} order analysis. The lower order frequencies are more effective than the higher order frequencies in the recognition.

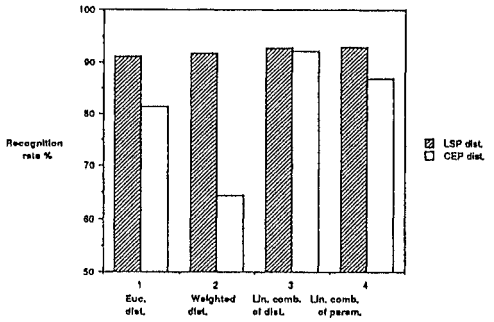

FIGURE 3 Recognition rates for various LSP distance measures

At the second part of the experiments, the weighted (inverse variance) LSP Euclidean distance was used for the recognition experiments with 10 LSP frequencies of the 16^{th} order analysis. It was observed that the weighted distance gives an improvement over the unweighted Euclidean distance (figure 3). On the other hand, the weighted CEP is also used for the comparison purpose and found to be inferior to both the weighted LSP Euclidean and the unweighted CEP distances. The inverse standard deviations of the LSP frequencies and the cepstrum coefficients are shown in figure 4. The monotonically increasing weighting of the inverse standard deviation of the cepstrum coefficients has clearly reduces the recognition rate. This is probably because the emphasis on the high quefrencies has a negative effect on the recognition performance.

The third part of the experiments cover the linear combination of the instantaneous and the transitional parameters and distances for the LSP frequencies and the cepstrum coefficients. For both of the representations, the optimum weight w_1 is searched by a small size (but gives distinguishable recognition performance) experiments. From a large region of the optimum weights, w_1=10. is used for both of the parameter sets. The performances of the instantaneous and the transitional parameter set of 10 LSP frequencies and 8 cepstrum coefficients are

compared. The combination of the LSP parameters is found to be advantageous to the cepstrum coefficients in terms of performance (figure 3). It was also found that the combination of LSP parameters gives slightly better performance than the combination of LSP distances. The former is also advantageous in terms of computation; First, the combination of parameters can be obtained at the analysis stage of speech signal and thus, it does not bring extra computation to the recognition process. Second, it also reduces the computation amount of the recognition process compared to the combination of distances.

At the last part of the experiments, bilinear transformed (Mel-scale-like warped) LSP distance measure was used. The speech recognition results for frequency axis warping parameter a=0.2, 0.5, and 0.8 were obtained (a=0.0 corresponds to the plain Euclidean LSP distance measure) (figure 5) A warping parameter a=0.4-0.8 causes a frequency warping of the LSP frequencies that is comparable to the logarithmic transformation of the Mel or Bark scales. A value of the warping parameter "a" that is greater than 0.8 causes the frequency axis to be warped more severely than that of Mel scaling. For the warping value of a=0.2, it was found that the recognition result for the warped LSP distance measure is slightly better than those of the Euclidean and the weighted Euclidean LSP distance measures, but the warping values a=0.5 and 0.8 do not give results as good as those distance measures.

It was observed that the recognition result of the LSP distance measure with the linear combination of the instantaneous and the transitional parameters is superior to the all of the LSP distance measures, and also the best of the distance measures used in this study.

CONCLUSIONS

The application of LSP frequencies representation for speech recognition is addressed. Various LSP based distance measures such as the Euclidean distance measure, the weighted (inverse variance) Euclidean distance, the linear combination of the transitional and the instantaneous parameters, and distances, and the bilinear transformed (Mel-scale-like frequency warped) distance are studied and the best performance result is obtained by the linear combination of the transitional and the instantaneous parameters.

Some of the common LPC based distance measures (CEP, weighted CEP, and the linear combination of the transitional and instantaneous cepstrum parameters and distances) are used for the comparison purpose. At the result, it was found that the linear combination of the instantaneous and the transitional LSP parameters gives the best recognition performance among the distance measures used for comparison. They are also found to be advantageous in computation.

It was also concluded that small values of frequency warping (such as a=0.2) cause the recognition performance to be improved and when the warping value increases (a=0.5 and a=0.8), the recognition performance starts decreasing.

REFERENCES

[1] Itakura, F., "Line Spectrum Representation of Linear Predictive Coefficients of Speech Signals," J. Acoust. Soc. Am., 57, 535(A), 1975.

[2] Furui, S., "Digital Speech Processing, Synthesis, and Recognition," Marcel Dekker, Inc., New York, 1989.

[3] Sagayama S. and Itakura F., "On individuality in a dynamic measure of speech," 3-2-7, Proc. ASJ annual meeting, pp. 589-590, 1979 (in Japanese).

[4] Sugiyama, M., "LPC Spectral Matching Measures for Speech Recognition," NTT Musashino ECL, 1984.

[5] Shikano, K., "Evaluation of LPC Spectral Matching Measures for Phonetic Unit Recognition," Carnegie Mellon University, Technical Report, CMU-CS-86-108, 1986.

[6] Furui, S., "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum," IEEE ASSP Trans., Vol. ASSP-34, No. 1, Feb 1986.

[7] Furui, S., "Speaker-Indepcndent Isolated Word Recognition Based on Dynamics-Emphasized Cepstrum," The Trans. of The IECE of Japan, Vol. E69, No. 12 December 1986.

[8] Paliwal, K. K., "A Study of Line Spectrum Pair Frequencies for Speech Recognition," ICASSP, 1988.

[9] Paliwal, K. K., "A Study of LSF Representation for Speaker-Dependent and Speaker-Independent HMM-Based Speech Recognition Systems," ICASSP, 1990.

[10] Soong F.K., Juang B.H., "Line Spectrum Pair (LSP) and Speech Data Compression," ICASSP, 1984.

[11] Hansen J.H.L., Clements M.A., "Constrained Iterative Speech Enhancement,"

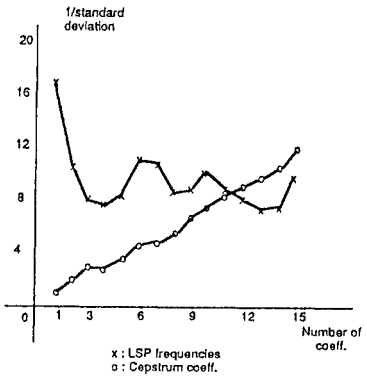

FIGURE 4 The relationship between the average inverse standard deviations of LSP frequencies and Ceps. coefficients

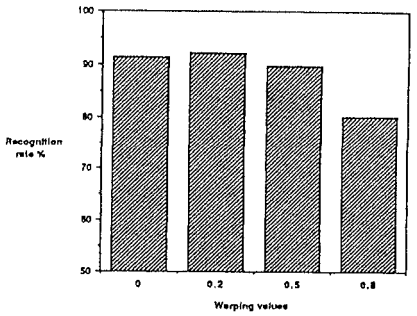

FIGURE 5 Recognition rates for frequency warped LSP distance measure