International Conference on Spoken Language Processing (*ICSLP-98*), 3: 1123-1126 (We1P46), Nov. 30, 1998, Sydney

THE DISTANCE MEASURE FOR LINE SPECTRUM PAIRS APPLIED TO SPEECH RECOGNITION

Fang Zheng, Zhanjiang Song, Ling Li, Wenjian Yu, Fengzhou Zheng, and Wenhu Wu

Speech Laboratory, Department of Computer Science and Technology,
Tsinghua University, Beijing, 100084, P. R. China
fzheng@sp.cs.tsinghua.edu.cn

ABSTRACT

The Line Spectrum Pair (LSP) based on the principle of linear predictive coding (LPC) plays a very important role in the speech synthesis; it has many interesting properties. Several famous speech compression / decompression algorithms, including the famous code excited linear predictive coding (CELP), are based on the LSP analysis, where the information loss or predicting errors are often very small due to the LSP's characteristics. Unfortunately till now there is not a satisfying kind of distance measure available for LSP so that this kind of features can be used for speech recognition applications. In this paper, the principle of LSP analysis is studied at first, and then several distance measures for LSP are proposed which can describe very well the difference between two groups of different LSP parameters. Experimental results are also given to show the efficiency of the proposed distance measures.

1. INTRODUCTION

Line Spectrum Pair (LSP) was first introduced by Itakura [4][8] as an alternative kind of LPC spectral representation. It was found that this new representation has such interesting properties as (1) all zeros of LSP polynomials are on the unit circle, (2) the corresponding zeros of the symmetric and anti-symmetric LSP polynomials are interlaced, and (3) the reconstructed LPC all-pole filter preserves its minimum phase property if (1) and (2) are kept intact through a quantization procedure. Soong proved all these properties via a "phase function" in his paper[7].

After introduced, LSP parameters accompanying with the vector quantization (VQ) technique play a very important role in speech coding/decoding and speech synthesis[7]. The famous code excited linear predictive (CELP) coding [1][2] is a good example in using the principles of linear predictive coding (LPC) and LSP to producing high quality speech in very low bit rate.

But till now the LSP parameters are seldom applied to speech recognition that is based on the statistical modeling and the VQ technique. Though there is a certain kind of distance measure adopted in the VQ technique for speech coding, but the measure is far from being suitable for speech recognition.

In speech recognition based on statistical modeling (such as hidden Markov models (HMM)) and the VQ technique, the distance measure for features is of a critical importance. The most recently used features nowadays are the cepstra. Because the Itakura distance measure has been proved suitable for LPC coefficients both theoretically and practically[3], therefore the Euclidean distance measure is good for the LPC derived cepstral vectors[3][10]. The solution of cepstrum distance measure makes the cepstra and their derived parameters play a very important role in speech recognition and become a kind of dominating feature.

However, the cepstra, either based on LPC principle or based on the Fourier transformation, have a great disadvantage, which often cannot distinguish two quite different phones, for example, the 'er' and 'ba' in Chinese. Now that LSP is a kind of successful feature in speech coding/decoding and its basic principle and behavior are somewhat different from those of LPC parameters, we have strong reason to do more research in LSP distance measure. Our motivation is to make it another kind of feature for speech recognition besides the cepstrum.

The paper is organized as follows. At first, we introduce the basic principle for LSPs and its relationship with LPC parameters and the transfer function of its all-pole model. Secondly, we proposed some LSP distance measures based on above principle. Thirdly, we give the experimental result to support our proposal. At last we come to the final conclusions.

2. PRINCIPLE OF LINE SPECTRUM PAIR (LSP)

2.1. Line Spectrum Pairs (Line Spectrum Frequencies)

Given a specific order P for the vocal track model of the speech to be analyzed, LPC analysis results in an all-zero inverse filter

$$A(z) \overset{def}{=} A_P(z) = 1 + \sum_{p=1}^{P} a_p z^{-p} \ , \qquad (1)$$

which minimizes the residual energy [5]. In speech compression and quantization based speech recognition, the LPC coefficients $\{a_1, a_2, ..., a_P\}$ are known to be inappropriate for quantization because of their relatively large dynamic range and possible filter instability problems. Different set of parameters representing the same spectral information, such as reflection coefficients and log area ratios, etc., were thus proposed for quantization in order to alleviate the above-mentioned problems. LSP is one such kind of representation of spectral information. LSP parameters have both well-behaved dynamic range and filter stability preservation property, and can be used to encode LPC spectral information even more efficiently than any other parameters.

The LSP representation is rather artificial. For a given P th order inverse filter as in (1), we can extend the order to (P+1) without introducing any new information by letting the (P+1) th Reflection coefficient be 1 or –1. This is equivalent to setting the corresponding acoustic tube model completely closed or completely open at the (P+1) th stage. We thus have

$$P(z) \overset{def}{=} A(z) + z^{-(P+1)} A(z^{-1}) \\ = 1 + \sum_{p=1}^{P} (a_p + a_{P+1-p}) z^{-p} + z^{-(P+1)} \qquad (2)$$

$$Q(z) \overset{def}{=} A(z) - z^{-(P+1)} A(z^{-1}) \\ = 1 + \sum_{p=1}^{P} (a_p - a_{P+1-p}) z^{-p} - z^{-(P+1)} \qquad (3)$$

It is obvious that $P(z)$ is a symmetric polynomial while $Q(z)$ is an anti-symmetric polynomial and

$$A(z) = \frac{P(z) + Q(z)}{2}. \qquad (4)$$

There are three important and interesting properties of $P(z)$ and $Q(z)$ listed as follows:

1. All zeros of $P(z)$ and $Q(z)$ are on the unit circle;
2. Zeros of $P(z)$ and $Q(z)$ are interlaced with each other; and
3. Minimum phase property of $A_P(z)$ is easily preserved after quantization of zeros of $P(z)$ and $Q(z)$.

The first two properties are useful for finding the zeros of $P(z)$ and $Q(z)$ and the third property ensures the stability of the synthesis filter. Since zeros of $P(z)$ and $Q(z)$ are on the unit circle, they can be expressed as $e^{j\omega}$ and hence ω's are called line spectrum frequencies (LSF).

If the order P is an even number that is greater than 2, we have the following special and additional properties for LSP:

4. –1 is a zero of $P(z)$ while 1 is a zero of $Q(z)$.
5. Besides ±1, $P(z)$ and $Q(z)$ have other P/2 pairs of conjugated zeros for each.

Therefore $P(z)$ and $Q(z)$ can be rewritten as

$$P(z) = (1 + z^{-1}) \prod_{i=1}^{P/2} (1 - z^{-1} e^{j\omega_i})(1 - z^{-1} e^{-j\omega_i}) \\ = (1 + z^{-1}) \prod_{i=1}^{P/2} (1 - 2\cos\omega_i z^{-1} + z^{-2}) \qquad (5\text{-}1)$$

$$Q(z) = (1 - z^{-1}) \prod_{i=1}^{P/2} (1 - z^{-1} e^{j\theta_i})(1 - z^{-1} e^{-j\theta_i}) \\ = (1 - z^{-1}) \prod_{i=1}^{P/2} (1 - 2\cos\theta_i z^{-1} + z^{-2}) \qquad (5\text{-}2)$$

Here $\omega_i (1 \le i \le P/2)$, the phases of conjugated zeros of $P(z)$, or the LSFs of the symmetric polynomial, and $\theta_i (1 \le i \le P/2)$, the phases of conjugated zeros of $Q(z)$, or the LSF of the anti-symmetric polynomial, are interlaced with each other in the interval $(0, \pi)$. That is to say

$$0 < \omega_1 < \theta_1 < \omega_2 < \theta_2 < \cdots < \omega_{P/2} < \theta_{P/2} < \pi \qquad (6)$$

Therefore we turn the coefficients of $A(z)$ equivalently into the phases of zeros of $P(z)$ and $Q(z)$ ω_i and θ_i. This is the procedure of LSP analysis for speech. Because of the interlacing property of these frequencies, they determine exclusively $P(z)$ and $Q(z)$, then $A(z)$. The set of frequencies of $P(z)$ and $Q(z)$ is named as a set of line spectrum pairs (LSP) or line spectrum frequencies (LSF).

2.2. The Relationship Between LSP and the Transfer Function *H(z)* of the Vocal Track Model

The transfer function of a simplified vocal track model based on P th order LPC speech analysis is given by[6][9]

$$H(\mathrm{z}) = \frac{1}{A_P(z)} = \frac{1}{1 + \sum_{p=1}^{P} a_p z^{-p}}, \qquad (7)$$

which is an all-pole filter model of speech signal.

It is possible to get the speech impulses from the original speech data using the filter with transfer function $A_P(z)$, which is the basis of the linear predictive coding. Eq. (7) implies a kind of relationship between the excitations and the predictive errors of a speech signal. It can be proved that using the linear predictive coding method the LPC coefficients $a_p (1 \le p \le P)$ can be estimated most precisely under the condition of least mean square error[6][9].

Obviously if the predictive errors are not ignored the LPC coefficients contain most information of the analyzed speech, so the LPC coefficients or LSPs can be used to represent the speech signals for speech recognition. In order to discover the distance measure for LSPs, we first study the relation between the transfer function $H(z)$ and LSPs.

By Equations (2), (3) and (4), we have

$$H(z) = \frac{1}{A_P(z)} = \frac{2}{P(z) + Q(z)} \qquad (8)$$

It is equivalent to the frequency response function form as

$$H(e^{j\omega}) = \frac{1}{A_P(e^{j\omega})} = \frac{2}{P(e^{j\omega}) + Q(e^{j\omega})} \qquad (9)$$

LSFs are the phases of the zeros of $P(z)$ and $Q(z)$, i.e., LSFs are the zeros of $P(e^{j\omega})$ and $Q(e^{j\omega})$, so if a pair of LSFs are very close at ω_0, $\left|P(e^{j\omega_0}) + Q(e^{j\omega_0})\right|$ will be very close to zero while $\left|H(e^{j\omega_0})\right|$ very big, resulting in a peak around ω_0

in the amplitude frequency response curve. On the contrary, if a pair of LSFs are far from each other, the amplitude frequency response curve will be relatively plat around the two LSFs. The relation is illustrated in Figure 1.

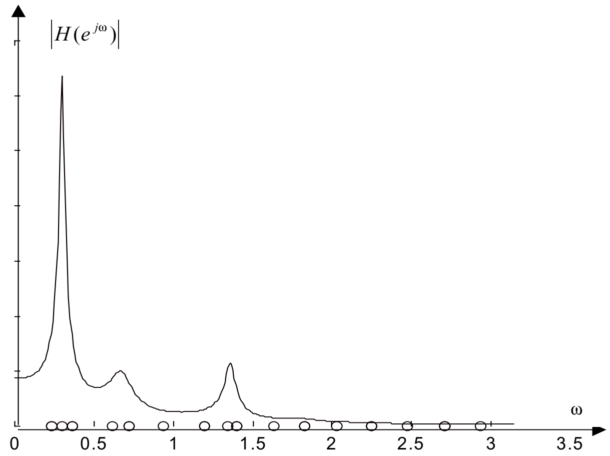

Figure 1: The relationship between LSPs and the transfer function *H(z)*. The small circles on the horizontal axis indicate the line spectrum frequencies.

3. DISTANCE MEASURE FOR LSPS

In this section, the LSP distance measures are proposed mathematically according to the above stated physical concept. For convenience, we regard a set of LSFs in ascending order as a feature vector in the feature space, denoted by $L=(l_1,l_2,\ldots,l_P)$, where *P* is the order of LSP analysis,

$$l_{2i-1}=\omega_i,\quad l_{2i}=\theta_i,\qquad 1\le i\le P/2 \tag{10-1}$$

and in the meanwhile we assume

$$l_0=0\,,\quad l_{P+1}=\pi\,. \tag{10-2}$$

3.1. LP Method: Considering Line Positions and Line Pair Distances Simultaneously

According to the relationship between LSPs and *H(z)*, the LSF line positions and line pair distances are the most important factors that affect *H(z)*, so the distance measure should reflect these two factors simultaneously. Based on this we define the distance between two LSP vectors in general form as

$$d(L^{\mathrm{R}},L^{T})=\alpha\cdot d_p(L^{\mathrm{R}},L^{T})+d_d(L^{\mathrm{R}},L^{T})\,, \tag{11}$$

where L^{R} is the reference LSP vector while L^T the test LSP vector, $d_p(\cdot,\cdot)$ is the partial distance related to the line positions while $d_d(\cdot,\cdot)$ related to the line pair distances, and α is the balance factor between the two partial distance.

The partial distance related to the line positions is defined as

$$d_p(L^{\mathrm{R}},L^{\mathrm{T}})=\sum_{\mathrm{i=1}}^{\mathrm{P}}(l_i^R-l_i^T)^2 \tag{12}$$

and the partial distance related to the line pair distances is defined in three different ways as

$$d_d(L^{\mathrm{R}},L^T)=\sum_{\mathrm{i=1}}^{\mathrm{P}}\left(1-\frac{l_i^R-l_{i-1}^R}{l_i^T-l_{i-1}^T}\right)^2\,, \tag{13-1}$$

or

$$d_d(L^{\mathrm{R}},L^T)=\sum_{\mathrm{i=1}}^{\mathrm{P}}\left|1-\left(\frac{l_i^R-l_{i-1}^R}{l_i^T-l_{i-1}^T}\right)^2\right|, \tag{13-2}$$

In practice, α ranges between 20 to 200, hereafter we take $\alpha=50$.

This kind of distance measure is simple and directly related to the relationship between LSPs and the transfer function of the vocal tract model.

3.2. RI Method: Approaching *H(z)* Function With Rectangle Impulses

In the above section, we define a distance measure for LSPs by the line positions and the line pair distances mathematically. It is known that the relation between the distribution of *p* line positions and the transfer function $H(e^{j\omega})$ (simply denoted by $H(\omega)$ hereafter) is very close. The denser where the spectral lines are, the bigger where the $|H(\omega)|$ values are, and a peak is formed. Because the vocal track transfer function directly reflects the characteristics of the speech signal, the distance measure for LSPs should also reflect this characteristics. If there exists a method by which the shape of *H* function can be reconstructed from LSP parameters, the LSP distance can be measured by the Euclidean distance between the shapes of *H* functions converted from the LSP parameters.

Considering the computational complexity, we do not want to construct *H* from LSFs by a very complicated algorithm. It will be suitable only if the constructed *H* can well approach the actual one. Based on the above discussion, a method is proposed to reconstruct the *H* function by splicing several rectangle impulses related to the line pairs.

Given a LSP vector $L=(l_1,l_2,\ldots,l_P)$, construct a rectangle impulse at each line position as

$$H_i^L(\omega)=\begin{cases}\dfrac{A}{l_{i+1}-l_{i-1}}, & \omega\in I_i\\ 0, & else\end{cases}\quad,\ 1\le i\le P\,, \tag{14}$$

where *A* is a constant and can be taken to be 1 normally, I_i is the interval of *i-th* impulse which can be defined in different ways. The impulses at each line position make up of the whole frequency response function of the vocal tract model as

$$H^L(\omega)=\sum_{i=1}^{P}H_i^L(\omega) \tag{15}$$

Sampling $H^L(\omega)$ into $H^L(k) \hat{=} H^L(k \cdot \Delta\omega)$ ($1 \le k \le K$ and $K \cdot \Delta\omega = \pi$), the distance between the reference vector L^{R} and the test vector L^T is defined as the distance between their corresponding reconstructed discrete transfer function as follows

$$d(L^{\mathrm{R}}, L^{\mathrm{T}}) = \left\| \mathrm{H}^{\mathrm{L^R}}, \mathrm{H}^{\mathrm{L^T}} \right\| = \sum_{\mathrm{k=1}}^{\mathrm{K}} (\mathrm{H}^{\mathrm{L^R}}(k) - \mathrm{H}^{\mathrm{L^T}}(k))^2 \qquad (16)$$

Once the sampling spacing $\Delta\omega$ is well pre-defined, the discrete transfer function can be directly and easily reconstructed using a simple algorithm.

There are several different forms for this kind of distance measure depending on the different definitions of the impulse interval I_i.

(1) The center of I_i is located at the *i-th* line, the width is half of the distance between the two lines to the left and right of the current line, i.e.,

$$I_i = \left\{ \omega \middle| |\omega - l_i| \le \frac{1}{2}(l_{i+1} - l_{i-1}) \right\} \qquad (17\text{-}1)$$

(2) I_i covers the right half of the interval between the *i-th* line and its left line and the left half of the interval between the *i-th* line and its right line, i.e.,

$$I_i = \left[\frac{l_{i-1} + l_i}{2}, \frac{l_i + l_{i+1}}{2} \right] \qquad (17\text{-}2)$$

(3) All impulse widths are same as $\dfrac{\pi}{P+1}$.

$$I_i = \left[\frac{(i-1/2)\cdot\pi}{P+1}, \frac{(i+1/2)\cdot\pi}{P+1} \right] \qquad (17\text{-}3)$$

4. EXPERIMENTAL RESULTS

We have done experiments on a database of 35 Chinese finals, where 3 speakers uttered twice through 16-bit sound blasters, the sampling rate is 11,025Hz. Features are extracted in a windowed frame of 256 samples (23 ms) every 128 samples. In this experiment, the training set consists of speech data by two of the three speakers, and the testing set consists of speech data by the left one of the three speakers. [3] The results are shown in Table 1. On an average, the accuracy based on the proposed LSP distance measures is about 4.5% improved compared to that based on the Itakura distance [3] for LPC coefficients.

Another experiment has been done over a giant 22-CD Chinese database sponsored by the State 863 HiTech Plan. The database consists of utterances uttered by about 140 speakers through 16-bit Sound Blasters at 16KHz sampling rate. 30 males' utterances are used for training and 8 males' utterances for testing. Acoustic modeling is based on the CDCPMs[11] based on Chinese syllables. The recognition rates for the Top 10 candidates using LP method are 89.49% for training set and 82.96% for testing set. No more comparison experiment is done.

5. CONCLUSIONS

In this paper, several kinds of LSP distance measures are proposed based on the theoretical analysis and experimental results. These distance measures are identical to the physical concept of the relationship between the transfer function *H(z)* and the LSP parameters of the vocal tract model to some extent.

Among these distance measure, the LP method should be the best choice. Not only the time complexity is very low, but also the performance is very good.

Table 1. The recognition accuracy for Chinese finals

Item			Rate
LP method	Eq. 13-1	Training Set	100%
		Testing Set	97.5%
RI method	Eq. 17-1	Training Set	100%
		Testing Set	85.0%
	Eq. 17-2	Training Set	100%
		Testing Set	85.0%

6. REFERENCES

[1] Campbell J.P., Tremain T.E., Welch V.C., "The DOD 4.8kb/s standard (proposed federal standard 1016)," in *Advances in Speech Coding*, (Atal B., Cuperman V., Gersho A., Eds.). Kluwer Academic Publishers, Boston, 1991, 121-133

[2] Fenichel R., Bodson D., "Details to assist in implementation of federal standard 1016 CELP," *NCS Technical Information Bulletin*, 92-1

[3] Itakura F., "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. ASSP* 23(1): 67-72, Feb. 1975

[4] Itakura F., "Line spectrum representation of linear predictive coefficients of speech signals," *J. Acoust. Soc. Am.*, 57, 537(A), 1975

[5] Markel J. And Gray A., *Linear Prediction of Speech*, Springer-Verlag, 1976

[6] Rabiner, L.R., Schafer, R.W., *Digital Processing of Speech Signals*, Prentice-Hall, Inc., 1978.

[7] Soong F.K., Juang B.H., "Linear spectrum pair (LSP) and speech data compression," *ICASSP-84*, 1(10): 1-4

[8] Sugamura N. And Itakura F., "Speech data compression by LSP speech analysis-synthesis technique," Trans. IECE'81/8, J64A(8): 599-606

[9] Yang X.J., Chi H.S., *et al. Speech Signal Digital Processing*, Publishing House of Electronics Industry, Beijing, Aug. 1995

[10] Zheng F., Wu W.-H., Fang D.-T., "A log-index weighted cepstral distance measure for speech recognition," *J. of Computer Science and Technology*, 12 (2): 177-184, Mar. 1997

[11] Zheng F., Chai H.-X., Shi Z.-J., *et al*, "A Real-World Speech Recognition System Based on CDCPMs," *J. of Computer Processing of Oriental Languages*, 11(3): 221-231, March. 1998