S5.6

STUDY OF LINE SPECTRUM PAIR FREQUENCIES FOR SPEAKER RECOGNITION

Chi-Shi Liu* Min-Tau Lin*
Wern-Jun Wang* Hsiao-Chuan Wang**

* Telecommunication Laboratories, Ministry of Communications, Taiwan, R.O.C.
** Dept. of Electrical Engineering of Tsing Hua University, Taiwan, R.O.C.

ABSTRACT

In this paper four varieties of Line Spectrum Pair(LSP) frequencies for speaker recognition applications and corresponding performance were studied , which were not used in any other papers about speaker recognition. The four varieties are (1) all LSP frequencies (2) odd or even line spectrum frequencies (3) the mean of adjacent LSP frequencies (4) the difference of adjacent LSP frequencies. A speaker-based vector quantization approach was employed for evaluating. The results show that each variety of LSP frequencies has very high accuracy rate.

I. INTRODUCTION

The efforts on speaker recognition has come of age. Different spectral features have been used for speaker recognition application, such as LPC, cepstrum, log area ratio, etc., but few of them talked about Line Spectrum Pair(LSP) frequencies representation. The Line Spectrum Pair was first introduced by Itakura[1] as an alternative representation of linear prediction (LP) coefficients. In past years, a number of studies had shown that LSP representation is more efficient than other LP parametric representations for encoding LPC spectral information[2] and better feature representation in speech recognition[3].

The speaker's voice is determined by structure of glottis, vocal tract, nasal, cavity, and other articulators. On the use of LP analysis for speaker recognition, variation of glottis and vocal tract are main information to distinguishing one speaker from another. In the LSP representation, it transfers these variation into frequency domain. Due to this artificial transformation, some varieties of LSP are investigated in this paper, including : (1) all LSP frequencies (2) odd line spectrum (OLS) frequencies and even line spectrum (ELS) frequencies, which are derived from setting the corresponding acoustic tube completely open or closed at the glottis (3) the mean of adjacent LSP frequencies, which are somewhat related to formant location (4) the difference of LSP frequencies, which are somewhat related to formant bandwidth. By these induced parametric representations from LSP, we want to study the corresponding performance in speaker recognition.

To evaluate these features used in speaker recognition, we used a speaker-based vector quatization(VQ) approach[4], where a VQ codebook can be easily generated and it needs no time-aligment, hence ideal for a text independent applications. A model(VQ codebook) is generated for each speaker and speaker's acoustic space is thus partitioned into non-overlapped convex regions.

Different LSP frequency representations were compared with LPC-derived cepstrum, which has been showed to be one of the best spectral information in speaker recognition [5].

II. REPRESENTATION OF LINE SPECTRUM PAIR FREQUENCIES AND OTHER INDUCED PARAMETERS

The LSP representation was first proposed by Itakura[1] as an alternative LP parametric representation. In the LP analysis of speech, a short stationary segment is assumed to be represented by a linear time invariant all pole filter $H(z)=1/A(z)$, where $A(z)$ is given by

$$A_m(z)=1+a_1z^{-1}+\ldots+a_mz^{-m} \qquad (1)$$

Where m is the order of $A_m(z)$ and $\{a_i\}$ are the LP coefficients

By deriving LSP spectrum frequencies, we extend the order (m) of the given filter (1) to (m+1) without introducing any new information by letting the (m+1)th reflection coefficient, k_{m+1}, be 1 or -1. This is

CH2847-2/90/0000-0277 $1.00 © 1990 IEEE

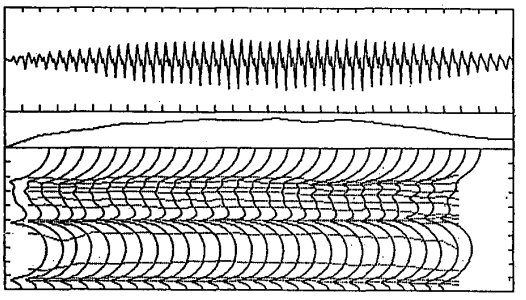

Fig. 1 Waveform, Energy contour, LPC spectrum and LSP spectrum of an isolated Mandarin syllable / i /

equivalent to setting the corresponding acoustic tube completely closed or open at the (m+1)th stage. We thus have P(z) and Q(z), for k_{m+1}=-1 or 1, respectively. That is,

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

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

All zeros of P(z) and Q(z) are on unit circle and zeros of P(z) and Q(z) are interlaced with each other. A(z) has minimun phase property after quantization of LSP frequencies[6]. Zeros of P(z) and Q(z) can be expressed as e^{jw_i} and w_i's are then called Line Spectrum Pair frequencies. LSP is the resonant frequencies at which the acoustic tube shows a line spectrum structure under a pair of artificially setting boundary conditions, i.e., complete opening or closure at glottis. If there is a resonant frequency, two or three LSP frequencies would form cluster around the corresponding resonance as shown in Fig.1.

Due to the line spectrum structure in LSP, there are some other interesting parameters induced from all LSP frequencies(ALSP), $W_a=\{w_{a1},w_{a2},\ldots,w_{am}\}$. These paramemters were described as follows :

(1) The first parameter set is called odd line spectrum(OLS) frequencies, W_o, and even line spectrum(ELS) frequencies, W_e, which are derived from zeros of P(z) and Q(z) respectively. W_o, $\{w_{o1},w_{o2},\ldots,w_{ok}\}$, is the set of line frequencies in P(z) and W_e, $\{w_{e1},w_{e2},\ldots,w_{el}\}$, is the set of line frequencies in Q(z). k and l are the order of P(z) and Q(z). Sum of them is equal to the order of A(z) and difference (k-l) is zero or one. The set of resonant line frequencies in P(z) or Q(z) is just the set of resonant line frequencies in acoustic tube under letting glottis open or close.

(2) the second parameter set called the mean of adjacent line spectrum(MALS) frequencies is expressed as W_m, $\{w_{m1},w_{m2},\ldots,w_{m(m-1)}\}$, and element, w_{mi}, is $(w_{a(i+1)}+w_{ai})/2$. Sihce formants are located in strong resonant positions, and in these locations some LSP frequencies are concentrated as shown in Fig.1, MALS correlate strongly with formant frequencies around clustered LSP frequencies.

(3) the third parameter set called the difference of adjacent line spectrum(DALS) frequencies is expressed as W_d, $\{w_{d1},w_{d2},\ldots,w_{d(m-1)}\}$, and element, w_{di}, is defined as $(w_{a(i+1)}-w_{ai})$, which was first proposed for efficient LPC spectral coding by Soong and Juang[6]. When formant bandwidth is smaller, the LSP frequencies located in formant position would be closer. This phenomenon can also be seen from Fig.1. From this observation, the DALS frequencies not only represent intra-normalization in short time spectrm but also are somewhat related to the formant bandwidths.

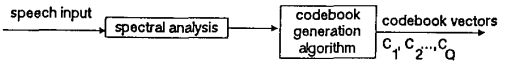

Fig. 2 Training proceduce for text independent speaker recognition

III. THE VQ-BASED SPEAKER RECOGNITION SYSTEM

The VQ-based speaker recognition system is first proposed by Soong, etc[4]. The basic principle of VQ-based speaker recognition is the compression of a large set of short-time spectral vector representing the spoken utternaces of a speaker into a small set of vectors. The compression set of vectors is referred to as a "codebook" and each speaker would have its own spectral codebooks. If there are N speakers to be recognized and M kinds of features to character a speaker, then the system would have N*M codebooks. The basic training block diagram of this system is shown in Fig.2. Given S training vectors $R_i=\{r_1,r_2,\ldots,r_S\}$ for each one speaker, the codebook generation algorithm[7] was used to produce codebook vectors $C_i=\{c_1,c_2,\ldots,c_Q\}$ as speaker characteristics. In recognition phase shown in Fig.3, a set of test vectors, $t_1,t_2,\ldots,t_M$, analyzed from the utterance of unknown speaker is encoded by the codebook, C_k, of specified speaker k to generate codeword sequence, $c_{t1},c_{t2},\ldots,c_{tM}$, c_{ti} is the codeword from the codebook C_i such that the distance between c_{ti} and t_i is minimun in the set C_i. The distance accumulated over the test vectors and normalized by the number of test vectors

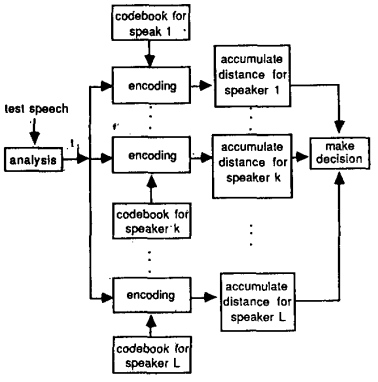

Fig.3 Text independent operation of the VQ-based speaker recognition

$$D=\frac{1}{M}\sum_{i=1}^{M} d\,(t_i, c_{ti})$$

is used to make a recognition decision. The Euclidean distance between two feature vectors X_i and X_j was used in this paper

$$d=[X_i-X_j]^t[X_i-X_j]$$

By this way, we had six different features(ALSP, OLS, ELS, MALS, DALS and cepstrum)used in the VQ-based speaker recognition system.

IV. DATABASE AND EXPERIMENTS

To evaluate the performance of the different spectral informations used in speaker recognition system, the database used in this study is a 20 speaker(16 male and 4 feamle), isolated digit database. Over a period of six weeks, Mandarin isolated digits were collected. Each speaker spoke 20 isolated digits(2 utternances per digit) in each week. The speech signal of the database is sampled at 10kHz by 16 bit analog-to-digital converter. The sampled specch data were preemphasised by transfer function $H(z)=1-0.94z^{-1}$. A 20msec Hamming window shfited every 15 msec was used to compute LPC coefficients of 12th order. Cepstrum coefficients and autocorrelation coefficients of order 12 were derived from LPC coefficients. Other induced LSP coefficients defined as section II were also derived. In this study, we used six diffenent features including ALSP, OLS, ELS, MALS, DALS and Cepstrum in speaker recognition.

We used the first 100 digits of each speaker to generate five referenced codebooks with 32 entries. The rest 20 digits of each speaker was used as test token. In the testing experiment, 2,4,6,8,10, and 12 randomly selected digit lengths were used.

feature \ token length (digit) / accuracy	2	4	6	8	10	12
CEP	89	92	96.3	97.5	97.5	97.5
ALSP	92	96	97.5	97.5	100	100
OLS	85	92	95	95	95	95
ELS	85.5	95	95	97.5	100	100
MALS	90	96	96.3	98.8	100	100
DALS	93	98.8	98.8	98.8	100	100

Table 1 The accuracy of six different features for different test token lengths. Size 32 VQ codebooks were used.

Table 1 showed the speaker identification accuracy for six different features. From this table we found that any induced LSP frequencies except OLS or all LSP frequencies have higher accuracy rate than the cepstrum coefficients. This illustrated that the performance of LSP spectrum was not only good in speech coding but also in speaker recognition. The other interested point was that DALS had the highest accuracy among all six different features.

Another experiment was performed by looping back the database through a local exchange telephone line. For this new database, ALSP, MALS, DALS and normalized-cepstrum were evaluated and the results were shown in Fig.4. The performance of DALS was better than other for test tokens of 4 digits or more. The performances of ALSP, MLSP and normolized cepstrum were similar under test tokens of 10 digits or more, but the performance of ALSP and MALS had higher performance than normalized cepstrum for test tokens of between 4 and 8 digits. The high performance of DALS seems to indicate that the DALS features tend to equalize the all channel effect.

V. CONCLUSION

The effectiveness of LSP parameters for speaker recognition was studied in this paper. All LSP parameters, except OLS coefficients, are good spectral features in speaker recognition. The recognition accuracy remains good when only half of LSP cofficients, ELS, as feature parameters. The recognition accuracy under telephone line by these LSP parameters was also tested. DALS parameter was the best

parameter not only in normal recording condition but also in telephone line. LSP has been shown to be an efficient parameter set for speaker recognition.

VI. ACKNOWLEDGMENTS

The authors would like to thank Dr. S.C. Lu, the director of Telecomm. Labratories, and Dr. I.C. Jou for their invaluable advice and timely encouragement. We also thank Drs. F.K. Soong, B.H. Juang, and C.H. Lee, for their suggestions.

REFERENCES

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

[2] Sugamura, N. and Itakura F., "Speech Data Compression by LSP Speech Analysis-Synthesis Technique," Trans. IECE Vol. J64-A, No.8, 1981, pp.599-605, 1981.

[3] Paliwal, K.K., "A Study of Line Spectrum Pair Frequencies for Speech Recognition," Proc. ICASSP, pp.485-488,1988.

[4] Soong, F.K., Rosenberg A.E., Rabiner, L.R. and Juang, B.H.," A vector quantization Approach to Speaker Recognition," Proc. ICASSP, PP.387-390, 1985.

[5] Atal, B.,"Automatic Recognition of Speakers from Their Voices," Proc. IEEE, Vol. 64 pp.460-475, Apr. 1976.

[6] Soong, F.K. and Juang, B.H.," Line Spectrum Pair and Speech Data Compression," Proc. ICASSP, 1984.

[7] Linde, Y., Buzo, A. and Gray, R.M.," An Algorithm for Vector Quantization," IEEE Trans. on Communication, Vol. COM-28, No.1, pp84-95, Jan, 1980.

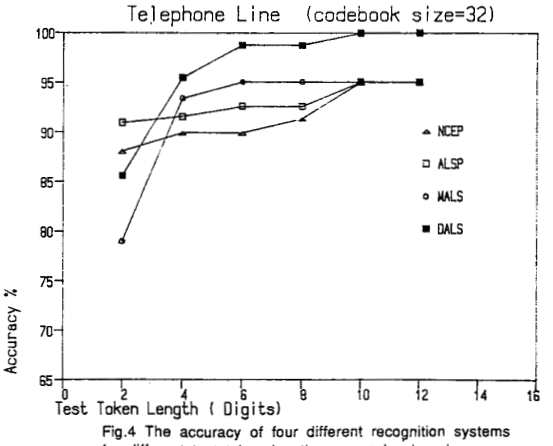

Fig.4 The accuracy of four different recognition systems for different test token lengths over a local exchange telephone line