EFFECT OF NOISE ON LINE SPECTRUM FREQUENCY AND A ROBUST SPEECH RECOGNITION METHOD FOR THE LOW BIT-RATE ENCODED SPEECH

An-Tzyh Yu and Hsiao-Chuan Wang
National Tsing Hua University, Hsinchu, Taiwan

ABSTRACT

Line Spectrum Frequencies (LSFs) are parameters for the low bit-rate (LBR) speech coding. It is desirable to use LSFs for speech recognition without decoding the encoded speech into waveform. However, the performance of using LSFs in speech recognition degrades when the speech signal is contaminated by additive noise. This paper introduces a graphic representation of the effect of additive noise to LSFs. It shows that LSFs of noisy speech shift smoothly to LSFs of noise as SNR increases. Then an interpolation algorithm is developed to adapt the speech models in LSF domain. The concept of so-called LSF spectrum is applied to estimate the interpolation factors. Experiment has demonstrated the effectiveness of this proposed method.

1. INTRODUCTION

Speech Recognition based on the low bit-rate (LBR) coded speech becomes more popular. In light of the rapid development of Internet and mobile phones, the transmission of LBR coded speech increases quickly. The speech recognition based on the coded speech becomes more attractive [1][2]. Instead of extracting features from decoded speech waveform, it is more interested to apply encoded parameters directly for speech recognition [3].

Line spectrum frequencies (LSFs) are the parameters employed popularly for LBR speech coding. LSF was first introduced by Itakura as an alternative representation of linear prediction (LP) coefficients [4]. LSF has many interesting properties. It can be quantized and interpolated without losing too much accuracy in representing the vocal tract model. The quantization error on a LSF makes only the local effect in frequency domain. These properties make LSF be extensively applied in the area of speech coding and synthesis. Nowadays, most of the LBR coding algorithms are embedded in LSF.

A few researches have been reported in dealing with LSF and using LSF in the speech recognition. Zheng et al. proposed a distance measure for LSF and reported that satisfactory performance has been achieved [5]. In our previous work [6], speech recognition based on encoded LSFs can achieve very good performance. However, the speech represented by encoded LSFs is not robust to noise. The performance of speech recognition degrades drastically when the speech signal is contaminated by additive noise. In order to investigate the behavior of the effect due to additive noise, we introduce a graphic representation of the effect of additive noise to LSFs. It shows that LSFs of noisy speech signal shift to LSFs of noise signal as SNR decreases. The pattern of LSF loci depends on the type of noise.

To improve the performance of noisy speech recognition, the model adaptation method is a proper approach. The fact of LSF shift has motivated develop a model adaptation method. We adapt the speech models in LSF domain based on the interpolation between the clean speech model and the noise model. The experiments were conducted on the cases of speech signal contaminated by white noise, factory noise, and cabin noise of F16 fighter. The speech was encoded in FS1016, a low bit-rate coding method based on code-excited linear prediction (CELP). The experimental results have shown the effectiveness and robustness of our proposed model adaptation method.

2. PRINCIPLES OF LINE SPECTRUM FREQUENCY

2.1 Representation of LSF

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

$$A(z) = 1 + a_1 z^{-1} + \ldots + a_M z^{-M}, \quad (1)$$

where M is the order of $A(z)$ and $\{a_i\}$ are the LP coefficients.

Let us define two functions, $P(z)$ and $Q(z)$, that are;

$$P(z) = A(z) - z^{-(M+1)} A(z^{-1})$$
$$Q(z) = A(z) + z^{-(M+1)} A(z^{-1}) \quad (2)$$

$P(z)$ and $Q(z)$ have the following interesting properties:

1) all the zeros of $P(z)$ and $Q(z)$ are on a unit circle;
2) zeros of $P(z)$ and $Q(z)$ are interlaced; and
3) the minimum phase property of $A(z)$ can be easily preserved if the first two properties are retained after quantization.

Zeros of $P(z)$ and $Q(z)$ can be expressed as $e^{j\varpi_i}$, and ϖ_i's are then called line spectrum frequencies (LSFs). In other words, a short term of speech signal can be represented by a set of LSFs.

2.2 LSF Spectrum

Spectrum, which represents the energy distribution of a signal in frequency domain, provides useful information for many applications. By manipulating the LSFs, we can find a pattern in LSF domain that is well matched to the spectrum of the speech signal. We call this pattern the LSF spectrum.

LSF spectrum is expressed by a set of paired parameters, (f_j, p_j), $1 \le j \le M$. f_j *and* p_j are frequency and magnitude, respectively. The definition of these two parameters is given by the following equations.

$$f_j = \frac{l_{j-1} + l_j}{2}, \tag{3a}$$

$$p_j = \frac{1}{l_j - l_{j-1}}. \tag{3b}$$

where l_j is the *j*-th frequencies in this LSF representation with $l_0 = 0$ and $l_{M+1} = \pi$. It is clear that the magnitude p_j is defined on the center of two adjacent frequencies, l_{j-1} and l_j. The distance between two adjacent frequencies determines the magnitude of LSF spectrum. This representation is shown in Figure 1. It is interested that the LSF spectrum is well matched to the spectrum of the signal.

The LSF spectrum energy is defined as;

$$\hat{E} \equiv \sum_j p_j^2 (l_j - l_{j-1}), \tag{4a}$$

The normalized LSF spectrum energy is defined as

$$\hat{E}_n \equiv \sum_j p_j^2 \cdot \tag{4b}$$

Then, the relationship between speech frame energy E and LSF spectrum energy $\hat{E}$ is

$$E = k\hat{E} \tag{5}$$

, where k is a scaling factor.

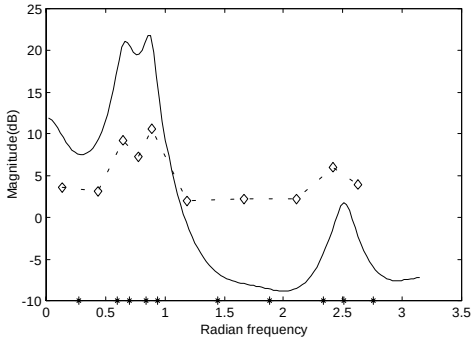
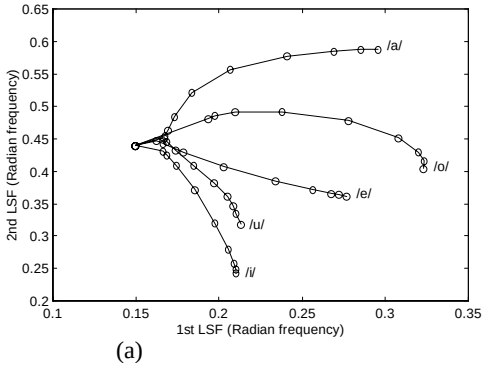


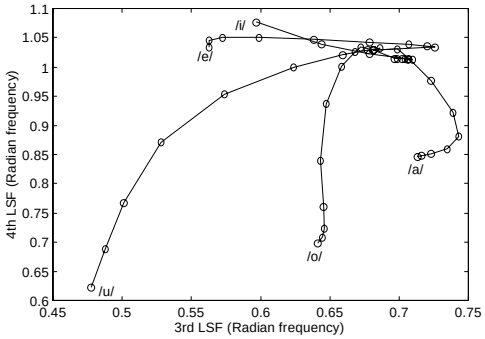
Figure 1. The spectrum (solid line), LSFs (stars), and LSF Spectrum (solid line with diamond) of a speech signal

3. THE EFFECT OF NOISE ON LSF

Some experiments on the vowels, /a/, /o/, /e/, /i/ and /u/, are conducted to show the effect of noise to the LSFs. The noise is artificially added to the speech signal to form a noisy speech signal. The white noise is generated by computer, while the factory noise and the cabin noise of F16 fighter are extracted from NOISEX-92 database. Figure 2 displays the loci of LSFs of five vowels in noisy environments. A locus starts from clean speech, then the noisy speech with SNR from 20 dB to -20 dB stepped by 5 dB. For simplicity, only the loci of l_2 versus l_1 and l_4 versus l_3 are plotted. These figures reveal that the loci of LSFs of noisy speech signal move to the LSFs of noise signal as the SNR decreases. Since LSFs of noise depends on noise type, the pattern of LSF loci is according to the noise type also.



(a)



(b)

Figure 2. The loci of LSFs. (a) l_2 versus l_1 (b) l_4 versus l_3

If we look at the spectra of noisy speech signal in different SNRs, we can find that the spectrum of speech signal is affected by noise in different levels. Figure 3 shows the spectra of a speech signal contaminated by cabin noise of F16 fighter. The lowest curve is spectrum of original speech signal. When noise level increases, the magnitude of spectrum increases. Finally the signal spectrum becomes close to noise spectrum. The SNR decreases from 20 dB to -10 dB with step of 5 dB.

When LSF representation is applied, we can plot the LSFs versus SNRs. Figure 4 demonstrate the shift of LSFs from speech signal at the right hand side to noise at the left hand side. The SNR decreases from 20 dB to -20 dB with step of 5 dB. Due to the property of LSF, the loci of LSFs in Figure 4 shift smoothly. The denser region of LSFs corresponds to bigger magnitude in the spectrum.

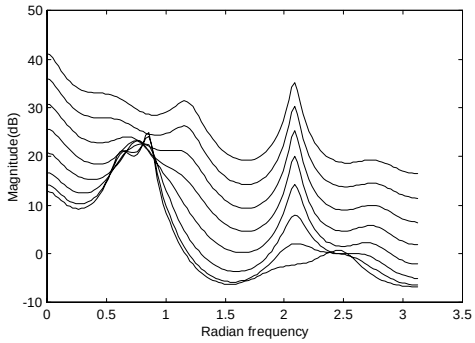


Figure 3. Spectra of speech signal contaminated by cabin noise of F16 fighter in different noise levels

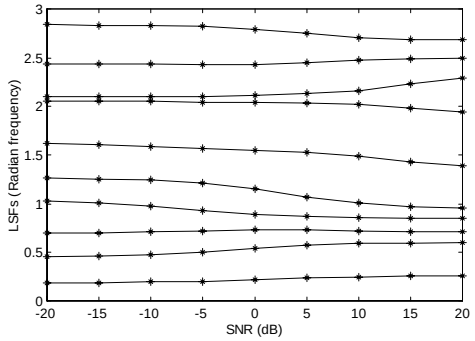


Figure 4. Shift of LSFs due to the effect of cabin noise of F16 fighter in different noise levels.

4. MODEL ADAPTATION

Based on the information revealed in Figure 4, we can develop a model adaptation algorithm to adjust the clean speech model to fit the environment of a noisy speech. Since the loci of LSFs in Figure 4 are relatively smooth, a linear interpolation is suitable for shifting the feature vector of clean speech model to the feature vector of noisy speech according to a known SNR. Let LSFs be a feature vector of framed speech signal. This model adaptation is expressed as

$$\bar{\mu}_S = \mathrm{A}\mu_S + (I - \mathrm{A})\mu_N , \tag{6}$$

where μ_S, μ_N and $\bar{\mu}_S$ are the mean vectors respectively for the clean speech model, the noise model and the adapted model. $\mathrm{A} = diag(\alpha_1, \alpha_2, \alpha_3, .., \alpha_M)$, $0 \le £\backslash_j \le 1$, $1 \le j \le M$, is a set of scaling factors which are generated by the following equation;

$$\alpha_j = \frac{1}{1 + e^{-a(r_j - 1)}}, \quad 1 \le j \le M , \tag{7}$$

where *a* is a smoothing factor and depends on the noise type and SNR. r_j is the energy ratio between the clean speech model and the noise model on *j*-th frequency in LSF spectrum. r_j is calculated as

$$r_j = \frac{E_{S\,j}}{E_{N\,j}} = \frac{k_S \hat{E}_{S\,j}}{k_N \hat{E}_{N\,j}} = \frac{k_S p_{S\,j}^2}{k_N p_{N\,j}^2} . \tag{8}$$

$E_{S\,j}$ and $E_{N\,j}$ are the energy at *j*-th frequency in LSF representation for clean speech model and noise model, respectively. $P_{S\,j}$ and $P_{N\,j}$ are the LSF spectrum magnitude at *j*-th frequency respectively for the clean speech model and the noise model. k_S and k_N are the scaling factors as defined in Equation (5). Let $\hat{r}_j$ and k_r be defined as

$$\hat{r}_j = \frac{p_{S\,j}^2}{p_{N\,j}^2} , \tag{9}$$

$$k_r = \frac{k_S}{k_N} . \tag{10}$$

r_j can be formulated by

$$r_j = k_r \hat{r}_j . \tag{11}$$

$\hat{r}_j$ is the energy ratio between the clean speech model and the noise model on *j*-th frequency of a LSF spectrum.

Since the SNR can be expressed as

$$SNR = 10 \log \frac{E_S}{E_N} = 10 \log \frac{k_S \hat{E}_S}{k_N \hat{E}_N} , \tag{12}$$

$k_r = \frac{k_S}{k_N}$ can be obtained by

$$k_r = \frac{k_S}{k_N} = \frac{\hat{E}_N}{\hat{E}_S} 10^{\frac{SNR}{10}} . \tag{13}$$

Instead of using Equation (9), we can define $\hat{r}_j$ as

$$\hat{r}_j = \frac{p_{S\,j}^2 + p_{S\,j-1}^2}{p_{N\,j}^2 + p_{N\,j-1}^2} . \tag{14}$$

Our experiment shows that the definition of Equation (14) can obtain better performance.

5. EXPERIMENTAL RESULTS

The speech data provided by 50 male and 50 female speakers was collected in a sound treated environment and sampled at 8 kHz. There are three sessions of data collection. This database is referred to as the clean speech database. A speaker utters a set of ten Mandarin digits in each session. Two sessions are used for

training and the other session is for testing. End points are roughly detected so that each utterance still contains short periods of pre-silence and post-silence. The noise is added to the clean speech with specific SNR values for generating artificial noisy speech. Three noises are tested. They are white noise, factory noise, and cabin noise of F16 fighter. White noise is artificially generated by computer, and the other noises are picked from NOISEX-92 database. Both the clean speech and the noisy speech are coded by FS1016 (software). LSF parameters are extracted from those coded speech as feature vectors.

For the Mandarin digit recognition, Each digit is modeled by an HMM with mixture of continuous Gaussian densities. The HMM contains seven states beginning with a pre-silence state and ending with a post-silence state. The silence states for all digits are tied together, i.e., share the same statistic parameters. Each state is modeled by a mixture of 4 Gaussian densities. All covariance matrices are diagonal. The parameters of model are trained by the segmental K-means algorithm.

Table 1 shows the performance of this Mandarin digit recognition. It is obvious that the proposed model adaptation algorithm has successfully enhanced the recognition performance for all noisy environments. The improvement is much significant for colored noises

6. CONCLUSION

This paper introduces a graphic representation of the effect of additive noise to LSFs. It shows that the LSFs of noisy speech signal shift to the LSFs of noise signal as the signal to noise ratio decreases. The smooth curves of LSF loci motivate the development of linear interpolation algorithm for model adaptation. The concept of so-called LSF spectrum is applied to find the interpolation factors for adapting a clean speech model to fit the noisy environment. It comes out a model adaptation method in LSF domain. The experiment of Mandarin digit recognition has demonstrated the effectiveness of the proposed speech recognition method.

		SNR				
		20 dB	15 dB	10 dB	5 dB	0 dB
White noise	Baseline	75.4	62.6	46.4	36.9	27.4
	Model adapt.	86.7	78.9	70.4	51.9	34.6
Cabin noise	Baseline	95.4	92.0	81.1	70.3	50.5
	Model adapt.	96.8	96.2	93.7	88.0	77.8
Factory noise	Baseline	95.8	95.5	91.6	81.3	66.7
	Model adapt.	97.0	96.4	95.1	91.3	85.9

Table 1. Recognition rates for speech contaminated by noises.

ACKNOWLEDGMENTS

This research has been sponsored by the National Science Council, Taiwan, under contract number NSC-87-2213-E-007-020.

REFERENCES

[1] B.T. Lilly and K.K. Paliwal 1996. Effect of Speech Coders on Speech Recognition Performance. *Proceedings ICSLP'96*, pp. 2344-2347.

[2] M. Kuitert and L.Boves 1997. Speaker Verification with GSM Coded Telephone Speech. *Proceedings EuroSpeech'97*, pp. 975-978.

[3] A. Gallardo-Antolin, F. Diaz-de-Maria and F. Valverde-Albacete 1998. Recognition from GSM Digital Speech. *Proceedings ICSLP'98*, pp. 1443-1446.

[4] F. Itakura 1975. Line Spectrum Representation of Linear Predictive coefficients of Speech Signals. *J. Acoust. Soc. Amer.*, vol.57, Suppl., no. 1, S35

[5] F. Zheng, Z. J. Song, L. Li, W. J. Yu, F. Z. Zheng, and W. H. Wu 1998. The Distance Measure For Line Spectrum Pairs Applied to speech recognition. *Proceedings ICSLP'98*, pp. 1123-1126.

[6] A. T. Yu and H. C. Wang 1998. A Study on the Recognition of Low Bit-Rate Encoded Speech. *Proceedings ICSLP'98*, pp. 1523-1526.

[7] C. S. Liu, M. T. Lin, W. J. Wang and H. C. Wang 1990. Study of Line Spectrum Pair Frequencies for Speaker Recognition. *Proceedings ICASSP'90*, pp. 277-280.

[8] F. K. Soong and B. H. Juang 1993. Optimal Quantization of LSP parameters. *IEEE Trans. Speech and Audio Processing*, Vol. 1, No. 1, pp. 15-24.