A STUDY OF LSF REPRESENTATION FOR SPEAKER-DEPENDENT AND SPEAKER-INDEPENDENT HMM-BASED SPEECH RECOGNITION SYSTEMS

K.K. Paliwal
Computer Systems and Communications Group
Tata Institute of Fundamental Research
Homi Bhabha Road, Bombay-400005, India
(Current address: Acoustics Research Dept.,
AT&T Bell Labs., Murray Hill, NJ 07974, USA)

ABSTRACT — In this paper, the line spectral-pair frequency (LSF) representation is used as the parametric representation for speech recognition. Its performance is compared with that of the cepstral coefficient (CC) representation for the speaker-dependent and speaker-independent hidden Markov model (HMM) based isolated word recognition systems. It is shown that the CC and the LSF representations result in comparable recognition performances for the full covariance matrix case. But, for the diagonal covariance matrix case, the LSF representation provides significantly better recognition performance than the CC representation.

1. INTRODUCTION

The line spectral-pair frequency (LSF) representation has been proposed by Itakura [1] as an alternative linear prediction (LP) parametric representation. In the context of speech coding, it has been shown [2,3] that this representation has better quantization properties than the other LP parametric representations (such as log area ratios and reflection coefficients). The LSF representation is capable of reducing the bit-rate by 25-30% for transmitting the LP information without degrading the quality of synthesized speech [3]. Our interest in LSF representation has been to see whether we can get similar advantage from this representation for speech recognition. For this, we studied this representation in our earlier paper for the recognition of steady-state vowel frames in the speaker-dependent mode using the minimum distance classifier [4]. Though the LSF representation resulted in good performance [4], the scope of these results was very limited.

The aim of the present paper is to extend the use of the LSF representation for more general speech recognition systems and to widen the scope of its results. For this, we study here this representation in both the speaker-dependent and the speaker-independent modes for the hidden Markov model (HMM) based isolated word recognition systems. Since the HMM-based speech recognizers use the maximum likelihood decision rule for recognition, we also report here the results for the speaker-dependent and the speaker-independent vowel recognition experiments using the maximum likelihood classifier. In the present paper, we compare the performance of the LSF representation with that of the cepstral coefficient (CC) representation. The CC representation is chosen here for comparison because this is currently the most popular representation for the HMM-based speech recognizers [5].

The paper is organized as follows. In Section 2, the LSF representation is defined and its properties are briefly described. Different recognition experiments comparing the performance of the LSF representation with that of the CC representation are described in Section 3. Results obtained from these recognition experiments are discussed in Section 4 and conclusions are reported in Section 5.

2. THE LSF REPRESENTATION

In this section, we define the LSFs and describe some of their properties. For more details, see [2].

In the LP analysis of speech, a short segment of speech is assumed to be generated as the output of an all-pole filter $H(z) = 1/A(z)$, where $A(z)$ is the inverse filter given by

$$A(z) = 1 + a_1 + \cdots + a_M z^{-M}.$$

Here M is the order of LP analysis and $\{a_i\}$ are the LP coefficients.

In order to define the LSFs, the inverse filter polynomial is decomposed into two polynomials

$$P(z) = A(z) + z^{-(M+1)}A(z^{-1})$$

and

$$Q(z) = A(z) - z^{-(M+1)}A(z^{-1}).$$

The roots of the polynomials $P(z)$ and $Q(z)$ are called the LSFs. The polynomials $P(z)$ and $Q(z)$ have the following two properties: 1) All zeros of $P(z)$ and $Q(z)$ lie on the unit circle, and 2) Zeros of $P(z)$ and $Q(z)$ are interlaced with each other. These properties help in efficient numerical computation of the LSFs from $P(z)$ and $Q(z)$.

The transformation from LP coefficients to LSFs is reversible; i.e., it is possible to compute exactly the LP coefficients from the LSFs. Also, since the $P(z)$ polynomial is even and the $Q(z)$ polynomial is odd, it is possible to decompose the power spectrum $|A(\omega)|^2$ as follows:

$$|A(\omega)|^2 = [|P(\omega)|^2 + |Q(\omega)|^2]/4.$$

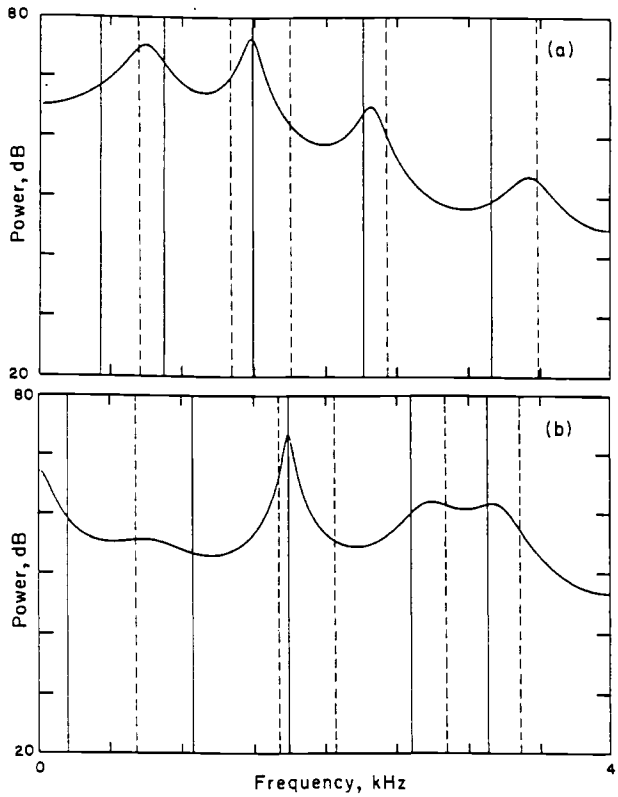

Fig. 1. LP power spectrum and the associated LSFs for (a) vowel /a/ and (b) fricative /s/.

CH2847-2/90/0000-0801 $1.00 © 1990 IEEE

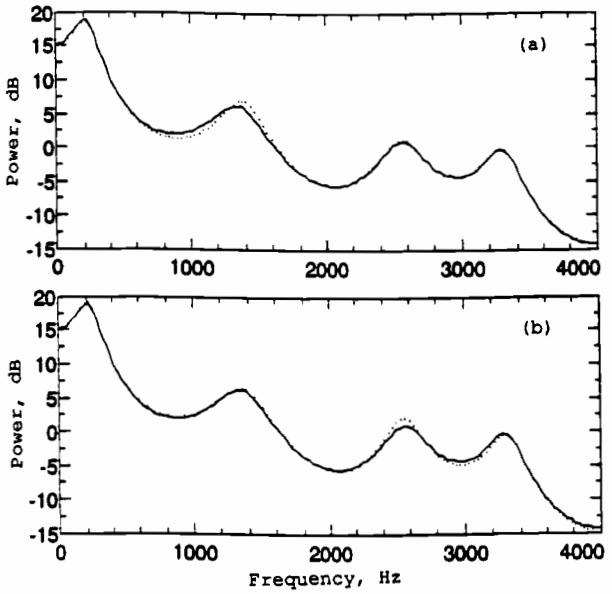

Fig. 2. Effect of changing LSF on LP power spectrum. The original spectrum is shown by solid line and the changed spectrum by dotted line. The original spectrum has LSFs at 212, 391, 930, 1285, 1505, 2003, 2484, 2719, 3177 and 3376 Hz. (a) Change of 4-th LSF from 1285 Hz to 1310 Hz, and (b) Change of 8-th LSF from 2719 Hz to 2691 Hz.

From this, it is easy to see that the roots of $A(z)$ (or, formants) are related to the roots of $P(z)$ and $Q(z)$. In order to illustrate this relationship between the formants and the LSFs more clearly, we show here the LP power spectrum and the associated LSFs in Fig. 1a for vowel /a/ and in Fig. 1b for fricative /s/. It can be seen here that a cluster of (2 to 3) LSFs characterizes a formant frequency and the bandwidth of a given formant depends on the closeness of the corresponding LSFs. In addition, the spectral sensitivities of LSFs are localized; i.e., a change in a given LSF produces a change in the LP power spectrum only in its neighborhood. This can be seen from Fig.2. Here, in Fig. 2a, a change in the fourth LSF from 1285 Hz to 1310 Hz affects the LP power spectrum near 1300 Hz. Similarly, in Fig. 2b, a change in eighth LSF produces a localized effect in its neighborhood in the LP power spectrum.

3. RECOGNITION EXPERIMENTS AND RESULTS

In this section, the LSF representation is compared as to its recognition performance with the CC representation. For this, four different types of recognition experiments are conducted. These experiments and their results are described below.

3.1. Speaker-Dependent Vowel Recognition Experiment

In this experiment, the recognition task is to classify steady-state vowel segments into 10 vowel classes in the speaker-dependent mode. The speech data base used for this purpose is derived from 900 utterances which consist of 30 repetitions of 10 different /b/-vowel-/b/ syllables spoken by three speakers (two male and one female). These utterances are lowpass filtered at 4 kHz and digitized at 10 kHz sampling rate. The steady-state part of the vowel segment is manually located for each of the 900 utterances and a 20 ms segment is excised from its center. A 10-th order LP analysis is performed for each 20 ms segment using the autocorrelation method (with 20 ms Hamming window and without preemphasis).

Since the HMM-based speech recognizers use the maximum likelihood decision rule for recognition, we use here the maximum likelihood (ML) classifier in order to be consistent with the HMM-based speech recognition experiments (described later in this section). The ML classifier classifies the input vector $\mathbf{x}$ (having 10 LSFs or CCs as its components) into vowel class i if $p(\mathbf{x}|i) > p(\mathbf{x}|j)$ for all $j \neq i$, where

Table 1: Recognition performance of the speaker-dependent vowel recognizer with the CC and the LSF representations.

Parameters	Recognition accuracy (in %) for	
	diag. cova. matrix	full cova. matrix
CCs	92.3	94.2
LSFs	96.7	95.3

Table 2: Recognition performance of the speaker-independent vowel recognizer with the CC and the LSF representations.

Parameters	Recognition accuracy (in %) for	
	diag. cova. matrix	full cova. matrix
CCs	80.2	92.7
LSFs	89.6	90.2

$p(\mathbf{x}|i)$ is the probability density function (or, likelihood function) for class i. In the present study, the class-conditional likelihood functions are assumed to be multivariate Gaussian; i.e.,

$$p(\mathbf{x}|i) = (2\pi|\mathbf{C}_i|)^{-M/2}\, exp[-(\mathbf{x}-\mathbf{m}_i)^t \mathbf{C}_i^{-1}(\mathbf{x}-\mathbf{m}_i)/2],$$

where $\mathbf{m}_i$ and $\mathbf{C}_i$ are the mean vector and the covariance matrix of class i.

The parameters of the likelihood function (namely, mean vector and covariance matrix) are estimated from the data in training set. At times, the amount of data available for training is rather limited and, therefore, it becomes difficult to get reliable estimates of all the components of the covariance matrix. In such cases, it is necessary to confine to diagonal covariance matrix; i.e., assume the off-diagonal elements to be zero. In the present paper, we study the use of both diagonal and full covariance matrices in this experiment as well as in all the other recognition experiments described later in this paper.

In order to compute the speaker-dependent vowel recognition performance for each speaker, the following procedure is used. For each vowel class, twenty-nine repetitions are used as the training set and the thirtieth repetition is used as test set. Each of the 30 repetitions is used in turn as the test set. All the 300 vectors of a given speaker are thus classified into 10 vowel classes.

Speaker-dependent vowel recognition results are averaged here over the three speakers. These are shown in Table 1 for the diagonal and the full covariance matrices. It can be seen from this table that vowel recognition performance with the LSF representation is marginally better than that with the CC representation for the full covariance matrix case. But, for the diagonal covariance matrix case, the LSF representation performs significantly better than the CC representation.

It might be noted that the full covariance matrix contains more statistical information about the training data than the diagonal covariance matrix and, hence, it is expected to perform better on test data than the diagonal covariance matrix. But, in this experiment, it is not so for the LSF representation. This happens because of small amount of training data which causes unreliable estimates of the off-diagonal elements in the full covariance matrix. This phenomenon is explained in more detail in Subsection 3.3 where the effect of training data size on recognition performance is studied more explicitly.

3.2. Speaker-Independent Vowel Recognition Experiment

In this experiment, the recognition task is to recognize steady-state vowel segments into 10 vowel classes in the speaker-independent mode. The speech data base used for this purpose is the same as described in Subsection 3.1. However, this data base is used here in a different fashion to compute the speaker-independent vowel recognition performance. Here, the first 15 repetitions from each of the three speakers are pooled together to get 45 repetitions for training. The remaining 45 repetitions from the three speakers are used for testing.

The speaker-independent vowel recognition results with the CC and the LSF representations are shown in Table 2 for the diagonal and the full covariance matrices. It can be seen from this table that the LSF representation does not perform as well as the CC representation for the full covariance matrix case, but its performance is significantly better than that of the CC representation case for the diagonal covariance

case.

3.3. Speaker-Dependent HMM-based Isolated Word Recognition Experiment

Here, the recognition task is to recognize isolated words from a limited vocabulary in the speaker-dependent mode. An HMM-based isolated word recognizer is used for this purpose [5]. The HMM has five states and is a left-to-right model where single skips between the states are allowed. Single mixture multivariate Gaussian functions are used to characterize the probability density functions of different states. The Viterbi algorithm is used for training as well as for testing the recognizer.

In order to study the speaker-dependent isolated word recognition performance of the CC and the LSF representations, the following four different vocabularies are used: 1) the vocabulary V1 containing 3 syllables (/be/, /de/ and /ge/) formed from voiced stop consonants. 2) the vocabulary V2 containing nine English e-set alphabets ('B', 'C', 'D', 'E', 'G', 'P', 'T', 'V' and 'Z'), 3) the vocabulary V3 containing 9 Norwegian e-set alphabets ('B', 'C', 'D', 'E', 'G', 'J', 'P', 'T' and 'V'), and 4) the vocabulary V4 containing 42 Norwegian alpha-digits (29 alphabets + 10 digits + 3 control words "start", "stopp" and "gjenta"). 120 repetitions of these vocabulary words are recorded over a period of 5 weeks. Three male speakers are used for recording. The utterances are lowpass filtered at 3.5 kHz and digitized at 8 kHz. A 10-th order LP analysis is performed every 15 ms with a frame width of 45 ms (using a preemphasis filter $H(z) = 1 - 0.95z^{-1}$ and a Hamming window). Endpoints are detected automatically using an energy criterion with some human supervision [6].

This speech data base is divided into two sets: 1) the training set containing the first 65 repetitions, and 2) the test set containing the remaining 55 repetitions. In order to study the recognition performance as a function of training data size, the recognizer is trained on varying number of repetitions from the training set and tested on the same 55 repetitions of test set. Results are shown in Tables 3, 4, 5 and 6 for the V1, V2, V3 and V4 vocabularies, respectively. The following four observations can be made from these tables: 1) When the amount of data available for training is large (65 repetitions), the full covariance matrix leads to better recognition performance than the diagonal covariance matrix. But, its performance is poorer with respect to the diagonal covariance matrix for less training data (20 repetitions), in spite of the fact that it characterizes statistically the training data better. 2) For the diagonal covariance matrix case, the LSF representation always performs significantly better than the CC representation. 3) Advantage in recognition performance (for the diagonal covariance matrix case) due to the LSF representation over the CC representation is more for smaller size of training set. 4) For the full covariance matrix case, the LSF and the CC representations are comparable in terms of their recognition performances (i.e.; differences in their performances are only marginal, these are some times in favor of the LSF representation and other times in favor of the CC representation).

3.4. Speaker-Independent HMM-based Isolated Word Recognition Experiment

In this experiment, the task is to recognize isolated words from a limited vocabulary in the speaker-independent mode. The HMM-based isolated word recognizer used here is the same as that used in Subsection 3.3. The vocabulary used in this experiment consists of 11 Norwegian digits. The speech data base is obtained by recording 223 repetitions of each of these digits. Forty-three speakers (both male and female) are used here for recording. The training set consists of 150 repetitions from 30 speakers and the test set 73 repetitions from 13 speakers. The speakers used in training and test sets are different. Processing of these utterances to derive LP parameters is done in the same fashion as described in Subsection 3.3.

Speaker-independent isolated word recognition results with the CC and the LSF representations are shown in Table 7 for the diagonal and the full covariance matrices. It can be seen from this table that the CC and the LSF representations are comparable for the full covariance matrix case. But, for the diagonal covariance case, the LSF representation results in better recognition performance than the CC representation.

Table 3: Recognition performance of the speaker-dependent isolated word recognizer with the CC and the LSF representations for the V1 vocabulary as a function of training data size.

Training data size	Recognition accuracy (in %) for			
	diag. cova. matrix with		full cova. matrix with	
	CCs	LSFs	CCs	LSFs
20	83.6	89.1	80.0	83.6
35	86.7	90.3	85.5	86.7
50	89.7	94.6	93.9	96.4
65	92.7	95.2	97.0	97.0

Table 4: Recognition performance of the speaker-dependent isolated word recognizer with the CC and the LSF representations for the V2 vocabulary as a function of training data size.

Training data size	Recognition accuracy (in %) for			
	diag. cova. matrix with		full cova. matrix with	
	CCs	LSFs	CCs	LSFs
20	79.2	88.9	90.7	90.1
35	84.2	88.9	95.6	92.5
50	82.4	89.1	93.9	92.7
65	85.5	88.9	96.2	95.4

Table 5: Recognition performance of the speaker-dependent isolated word recognizer with the CC and the LSF representations for the V3 vocabulary as a function of training data size.

Training data size	Recognition accuracy (in %) for			
	diag. cova. matrix with		full cova. matrix with	
	CCs	LSFs	CCs	LSFs
20	66.9	75.2	64.7	66.3
35	73.6	81.2	76.6	73.7
50	77.6	83.6	86.5	85.7
65	80.2	84.4	88.3	90.3

Table 6: Recognition performance of the speaker-dependent isolated word recognizer with the CC and LSF representations for the V4 vocabulary as a function of training data size.

Training data size	Recognition accuracy (in %) for			
	diag. cova. matrix with		full cova. matrix with	
	CCs	LSFs	CCs	LSFs
20	87.4	89.4	87.2	87.0
35	89.7	92.6	91.7	91.3
50	90.8	94.0	94.0	93.8
65	93.3	94.0	95.9	96.1

Table 7: Recognition performance of the speaker-independent isolated word recognizer with the CC and the LSf representations for the 11 Norwegian digit vocabulary.

Parameters	Recognition accuracy (in %) for	
	diag. cova. matrix	full cova. matrix
CCs	95.3	96.3
LSFs	96.4	96.4

4. DISCUSSION OF RESULTS

We have seen in the preceding section that the LSF and the CC representations result in comparable recognition performance for the full covariance matrix case. But, for the diagonal covariance matrix case, the LSF representation provides significant improvement in recognition performance over the CC representation. This improvement is more for smaller sizes of the training set.

A natural question that arises here is — why these representations are comparable in terms of their recognition performances for the full covariance matrix case, but so different for the diagonal covariance matrix case. The answer to this question lies in the fact that the spectral sensitivities of LSFs are localized, while those of CCs are not. That is, a

Table 8: Recognition performance of the speaker-dependent isolated word recognizer for the V2 vocabulary using different LP parametric representations.

Parameters	Recognition accuracy (in %) for	
	diag. cova. matrix	full cova. matrix
LP coeffs.	83.8	94.1
CCs	85.5	96.2
RCs	88.5	92.5
Log area ratios	86.5	93.1
Inverse sine RCs	88.7	92.9
Area coeffs.	83.2	87.7
Impulse response	73.3	95.4
Auto. coeffs.	74.3	-
LSFs	88.9	95.4

Table 9: Recognition performance of the speaker-dependent isolated word recognizer for the V3 vocabulary using different LP parametric representations.

Parameters	Recognition accuracy (in %) for	
	diag. cova. matrix	full cova. matrix
LP coeffs.	77.4	89.3
CCs	80.2	88.3
RCs	77.2	89.1
Log area ratios	78.0	89.5
Inverse sine RCs	78.2	89.3
Area coeffs.	70.5	81.4
Impulse response	78.0	85.5
Auto. coeffs.	66.3	-
LSFs	84.4	90.3

Table 10: Recognition performance of the speaker-dependent isolated word recognizer for the V4 vocabulary using different LP parametric representations.

Parameters	Recognition accuracy (in %) for	
	diag. cova. matrix	full cova. matrix
LP coeffs.	91.3	95.2
CCs	93.3	95.9
RCs	91.8	95.9
Log area ratios	91.2	95.8
Inverse sine RCs	92.0	95.9
Area coeffs.	82.9	90.7
Impulse response	92.5	94.9
Auto. coeffs.	88.9	-
LSFs	94.0	96.1

Table 11: Recognition performance of the speaker-independent isolated word recognizer for the 11 Norwegian digit vocabulary using different LP parametric representations.

Parameters	Recognition accuracy (in %) for	
	diag. cova. matrix	full cova. matrix
LP coeffs.	86.2	94.4
CCs	95.3	96.3
RCs	93.8	94.3
Log area ratios	94.3	94.8
Inverse sine RCs	94.3	95.3
Area coeffs.	65.5	77.8
Impulse response	89.4	93.2
Auto. coeffs.	85.9	-
LSFs	96.4	96.4

change in a given LSF produces a change in LP power spectrum only in the neighborhood of that LSF (as discussed in Section 2 and shown in Fig. 2). But, a change in a CC affects the entire LP spectrum. Because of this localized spectral sensitivity property, the LSF representation performs better than the CC representation for the diagonal covariance matrix case. In the case of full covariance matrix, these spectral interactions between different LP parameters are explicitly taken care of and, hence, the CC and the LSF representations result in comparable performances.

In order to see whether this explanation holds for other LP parametric representations as well, we study here the following nine LP parametric representations: 1) LP coefficients, 2) CCs, 3) reflection coefficients (RCs), 4) log area ratios, 5) inverse sine of RCs, 6) area coefficients, 7) impulse response of the LP synthesis filter, 8) autocorrelation coefficients, and 9) LSFs. Although each of these representations provide equivalent information about the LP power spectrum, only the LSF representation has the localized spectral sensitivity property. These LP parametric representations are studied here for both the diagonal and the full covariance matrix cases. Speaker-dependent results are shown in Tables 8, 9 and 10 for the V2, V3 and V4 vocabularies, respectively. Speaker-independent results are shown in Table 11 for the 11 Norwegian digit vocabulary. It can be seen from these tables that due to its localized spectral sensitivity property the LSF representation results in the best performance for the diagonal covariance matrix case. For the full covariance case, most of these representations (including CC and LSF) are comparable in performance.

5. CONCLUSIONS

In this paper, the LSF representation is used as the parametric representation for speech recognition. Its performance is compared with that of the CC representation for the HMM-based isolated word recognition systems. It is shown that the CC and the LSF representations result in comparable recognition performances for the full covariance matrix case. But, when the amount of training data is small (which happens quite often in practice), it is not possible to compute reliably the components of the full covariance matrix. In such cases, it is advantageous to use diagonal covariance matrix (as shown in Section 3). For the diagonal covariance matrix case, it is shown that the LSF representation provides significantly better recognition performance than the CC representation.

REFERENCES

[1] F. Itakura, "Line spectrum representation of linear predictive coefficients of speech signals", J. Acoust. Soc. Am., Vol. 57, p. S35, 1975.

[2] F.K. Soong and B.H. Juang, "Line spectrum pair (LSP) and speech data compression", Proc. ICASSP, pp. 1.10.1-1.10.4, 1984.

[3] G.S. Kang and L.J. Fransen, "Application of line-spectrum pairs to low bit-rate speech encoder", Proc. ICASSP, pp. 244-247, 1985.

[4] K.K. Paliwal, "A study of line spectrum pair frequencies for vowel recognition", Speech Communication, Vol. 8, pp. 27-33, 1989.

[5] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition", Proc. IEEE, Vol. 77, pp. 257-286, 1989.

[6] T. Svendsen, K.K. Paliwal, E. Harborg and P.O. Husoy, "An improved sub-word based speech recognizer", Proc. ICASSP, pp. 108-111, 1989.