SUBBAND ANALYSIS FOR ROBUST SPEECH RECOGNITION IN THE PRESENCE OF CAR NOISE

Engin Erzin
Bilkent University,
Ankara, TURKEY.

A. Enis Çetin
Koç University,
İstanbul, TURKEY.

Yasemin Yardımcı
Boğaziçi University,
İstanbul, TURKEY.

ABSTRACT

In this paper, a new set of speech feature representations for robust speech recognition in the presence of car noise are proposed. These parameters are based on subband analysis of the speech signal. Line Spectral Frequency (LSF) representation of the Linear Prediction (LP) analysis in subbands and cepstral coefficients derived from subband analysis (SUBCEP) are introduced, and the performances of the new feature representations are compared to *mel* scale cepstral coefficients (MELCEP) in the presence of car noise. Subband analysis based parameters are observed to be more robust than the commonly employed MELCEP representations.

1. INTRODUCTION

Extraction of feature parameters from the speech signal is the first step in speech recognition. It is desired to have perceptually meaningful parameterization and yet robust to variations in environmental noise. The *mel* scale is accepted as a transformation of the frequency scale in a perceptually meaningful scale, and it is widely used in feature extraction [9]. However the environmental noise may effect the performance of the *mel* scale derived features. In this paper, the performance of the subband analysis based methods are investigated for robust speech recognition in the presence of car noise.

Of the two techniques based on subband analysis that are presented here, the first is the Line Spectral Frequency (LSF) representation of the Linear Prediction (LP) analysis in subbands, and the second is the extraction of cepstral coefficients derived in subband analysis of speech signal. These representations are described in Sections 2 and 3, respectively.

The performance evaluation is done with a speaker independent continuous density Hidden Markov Model (HMM) based isolated word recognition system. The vocabulary consists of ten Turkish digits (0:sıfır, 1:bir, 2:iki, 3:üç, 4:dört, 5:beş, 6:altı, 7:yedi, 8:sekiz, 9:dokuz). The simulation examples are described in Section 4.

2. SUBBAND ANALYSIS DERIVED LSF REPRESENTATION

Linear Predictive modeling techniques are widely used in various speech coding, synthesis and recognition applications. Line Spectral Frequency (LSF) representation of the Linear Prediction (LP) filter is introduced by Itakura [1]. LSFs have some desirable properties which make them attractive to represent the Linear Predictive Coding (LPC) filter. The quantization properties of the LSF representation is recently investigated [2, 3, 4].

It is well known that LSF representation and cepstral coefficient representation of speech signals have comparable performances for a general speech recognition system [5]. Car noise environments, however, have low-pass characteristics which may degrade the performance of general full-band LSF or *mel* scaled cepstral coefficient (MELCEP) representations [6]. In this section, LSF based representation of speech signals in subbands is introduced.

Let the m-th order inverse filter $A_m(z)$,

$$A_m(z) = 1 + a_1 z^{-1} + \cdots + a_m z^{-m} \tag{1}$$

is obtained by the LP analysis of speech. The LSF polynomials of order $(m+1)$, $P_{m+1}(z)$ and $Q_{m+1}(z)$, can be constructed by setting the $(m+1)$-st reflection coefficient to 1 or -1. In other words, the polynomials, $P_{m+1}(z)$ and $Q_{m+1}(z)$, are defined as,

$$P_{m+1}(z) = A_m(z) + z^{-(m+1)} A_m(z^{-1}), \tag{2}$$

and

$$Q_{m+1}(z) = A_m(z) - z^{-(m+1)} A_m(z^{-1}). \tag{3}$$

The zeros of $P_{m+1}(z)$ and $Q_{m+1}(z)$ are called the Line Spectral Frequencies (LSFs), and they uniquely characterize the LPC inverse filter $A_m(z)$.

$P_{m+1}(z)$ and $Q_{m+1}(z)$ are symmetric and anti-symmetric polynomials, respectively. They have the following properties:

(i) All of the zeros of the LSF polynomials are on the unit circle,

(ii) the zeros of the symmetric and anti-symmetric LSF polynomials are interlaced,

(iii) the reconstructed LPC all-pole filter maintains its minimum phase property, if the properties (i) and (ii) are preserved during the quantization procedure, and

(iv) it has been shown that LSFs are related with the formant frequencies [5].

0-7803-2431-5/95 $4.00

In this scheme, the speech signal is filtered by a low-pass and a high-pass filter and the LP analysis is performed on the resulting two subsignals. Next the LSFs of the subsignals are computed and the feature vector is constructed from these LSFs.

It is experimentally observed that significant amount spectral power of car noise[1] is localized under 500 Hz. Due to this reason the LP analysis of speech signal is performed in two bands, a low-band (0-700 Hz) and a high-band (700-4000 Hz). In this case the high-band can be assumed to be noise-free.

This kind of frequency domain decomposition can be generalized to cases in which the noise is frequency localized.

3. SUBBAND ANALYSIS BASED CEPSTRAL COEFFICIENT REPRESENTATION

In this section, a new set of cepstral coefficients derived from subband analysis (SUBCEP) is introduced. The speech signal is divided into several subbands by using a perfect reconstruction filter bank [8] via a tree-structure. The selected filter bank corresponds to a biorthogonal wavelet transform [8]. The subbands are divided in a manner similar to the well-known *mel* scale decomposition [6].

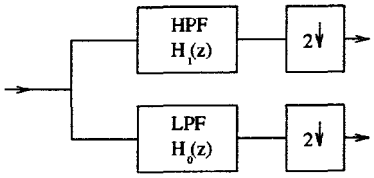

Figure 1: Basic block of subband decomposition.

The perfect reconstruction filter bank structure is shown in Figure 1. The low-pass filter, $H_0(z)$, and the high-pass filter, $H_1(z)$, are given by

$$H_0(z) = \frac{1}{2}[1 + zA(z^2)], \tag{4}$$

and

$$H_1(z) = -z^{-1} + \frac{1}{2}B(z^2)(1 + zA(z^2)) \tag{5}$$

where $A(z^2)$ and $B(z^2)$ are arbitrary polynomials of z^2. In this study we selected $H_0(z)$ as a *7-th* order Lagrange filter

$$H_0(z) = \frac{1}{2} + \frac{9}{32}(z^1 + z^{-1}) - \frac{1}{32}(z^3 + z^{-3}) \tag{6}$$

which is a half-band linear phase FIR filter. Note that (6) can be easily put into the form of (4) with

$$A(z^2) = \frac{9}{16}(1 + z^{-2}) - \frac{1}{16}(z^2 + z^{-4}) \ . \tag{7}$$

The second polynomial, $B(z^2)$ is chosen as

$$B(z^2) = \frac{1}{2}(1 + z^{-2}), \tag{8}$$

[1]This noise is recorded inside a Volvo 340 on a rainy asphalt road by *Institute for Perception-TNO, The Netherlands.*

This selection of $B(z^2)$ produces good low-pass and high-pass frequency responses for filters, $H_0(z)$ and $H_1(z)$, respectively [8]. This filterbank approximately divides the frequency domain into two half-bands, $[0, \pi/2]$ and $[\pi/2, \pi]$.

By applying the filterbank in a cascaded manner the frequency domain is divided into $L = 22$ subbands similar to the *mel* scale as shown in Figure 2 (This is equivalent to a wavelet packet bases decomposition of the input speech signal [8]).

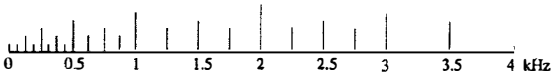

Figure 2: The subband decomposition of the speech signal.

The feature vector is constructed from the subsignals as follows: Let $x_l(n)$ be the subsignal at the l-th subband. For each subsignal the parameters, $e(l)$, is defined by

$$e(l) = \frac{1}{N_l}\sum_{n=1}^{N_l}|x_l(n)|, \quad l = 1, 2, ..., L \tag{9}$$

where N_l is the number of samples in the l-th band. The SUBCEP parameters, $SC(k)$, which form the feature vector are defined similar to MELCEP coefficients as

$$SC(k) = \sum_{l=1}^{L}\log(e(l))\cos(\frac{k(l-0.5)}{L}\pi), \ k = 1, 2, ..., 12. \tag{10}$$

The SUBCEP parameters are obtained in a computationally efficient manner because at every stage of the subband decomposition tree a downsampling by a factor of two is performed, and the filter bank structure of [8] can be implemented using integer arithmetic because all of the filters have rational coefficients.

Commonly used MELCEP parameters are obtained either in time domain with *critical band filter banks* or in frequency domain with *critical band windowing* of the speech spectrum. Since multirate signal processing techniques are not employed in the design of the so-called *critical band filter bank* [9] large filter orders are necessary for narrow subbands. This results in a computationally expensive and memory intensive implementation. Critical band windowing, on the other hand, requires complex arithmetic.

Apart from computational advantages, the SUBCEP approach also provides extra flexibility in dividing the frequency domain effectively. For instance, if the noise spectrum is localized in the frequency domain (e.g. car noise) then less emphasis can be given to the corrupted frequency regions by assigning larger subbands.

Other filter-bank structures and wavelet transforms can also be used to achieve a similar frequency decomposition and another set of SUBCEP parameters.

4. SIMULATION STUDIES

In simulation studies a continuous density Hidden Markov Model (HMM) based speech recognition system is used with

5 states and 3 mixture densities. The speech signal is sampled at 8 kHz and the so called car noise is down sampled to 8 kHz. The noisy speech is obtained with the car noise recording, assuming that the noise is additive. Simulation studies are performed on the vocabulary of Turkish digits from the utterances of 51 male and 51 female speakers. The isolated word recognition system is trained with 25 male and 25 female speakers, and the performance evaluation is done with the remaining 26 male and 26 female speakers.

4.1. Performance of LSF Representation in Subbands

A 12-th and 20-th order LP analysis are performed on every 10 ms with a window size of 30 ms (using a Hamming window) for low-band (noisy band) and high-band (noise free band) of the speech signal, respectively. First 5 LSFs of the low-band and the last 19 LSFs of the high-band are combined to form the sub-band derived LSF feature vector (SBLSF).

To compare the performance of LSF representation in subbands (SBLSFs) with full-band LSF, a 24-th order LP analysis is performed on full-band speech signal and recognition rate of full-band LSF feature vector is also recorded. The performance of LSFs with their time derivatives are also obtained using 12-th order LP analysis. Frequency domain cepstral analysis is performed to extract 12 *mel* scale cepstral coefficients. *Mel* scale cepstral feature vector (MELCEP) is obtained from these 12 cepstral coefficients and their time derivatives. The performances of the all four feature sets for various SNR values are plotted in Figure 3. In our simulation studies we observed that the performance of the subband derived LSF (SBLSF) representation is more robust in the presence of car noise.

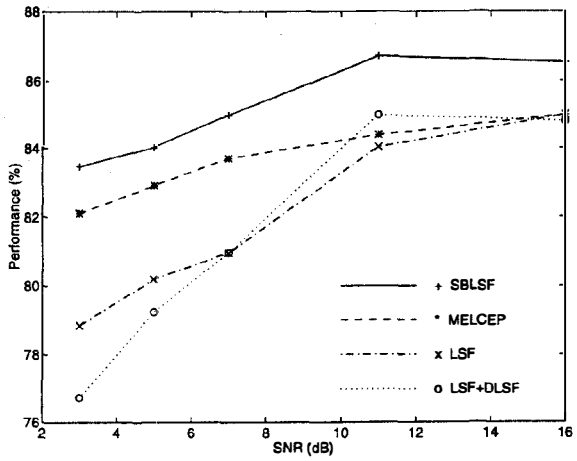

Figure 3: Performance evaluation of SBLSF, MELCEP and LSF representations.

4.2. Performance of SUBCEP Representation

The filter bank structure of Figure 1 is applied to the speech signal in a cascaded form (up to 6 levels) to achieve the subband decomposition shown in Figure 2. This decomposition results in 22 subsignals. The window size is chosen as 48 *ms* (384 samples) with an overlap of 32 ms so that the subsignal with the smallest subband has 6 samples. The SUBCEP parameters are derived as in Equation (10) and the feature vector is constructed from these SUBCEP parameters and their time derivatives. The performance of the SUBCEP and MELCEP representations are compared in Figure 4. The SUBCEP representation exhibits robust performance in the isolated word recognition application and it outperforms the MELCEP representation.

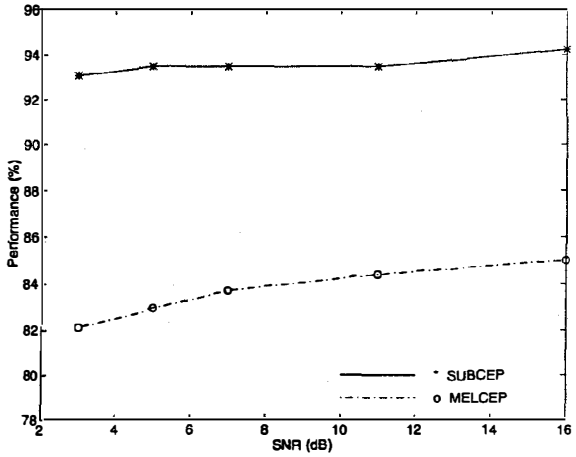

Figure 4: Performance evaluation of SUBCEP and MELCEP representations.

4.3. Conclusion

In this section, two new sets of speech feature parameters based on subband analysis, SBLSF's and SUBCEP's are introduced. It is experimentally observed that the SUBCEP representation provides the highest recognition rate for speaker independent isolated word recognition in the presence of car noise.

5. REFERENCES

[1] F. Itakura "Line spectrum representation of linear predictive coefficients of speech signals," *Journal of Acoust. Soc. Am.*, p. 535a, 1975.

[2] E. Erzin and A. E. Çetin "Interframe differential coding of Line Spectrum Frequencies," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 2, pp. 350-352, April 1994. Also presented in part at *Twenty-sixth Annual Conference on Information Sciences and Systems*, Princeton, NJ. March 1992.

[3] E. Erzin and A.E. Çetin "Interframe differential vector coding of Line Spectrum Frequencies," *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing 1993 (ICASSP '93)*, vol. II, pp. 25–28, April 1993.

[4] K.K. Paliwal, B.S. Atal "Efficient Vector Quantization of LPC Parameters at 24 bits/frame," *Proc. of the Int. Conf. on Acoustic, Speech and Signal Processing 1991 (ICASSP '91)*, pp. 661-664, May 1991.

[5] K.K. Paliwal "On the use of Line Spectral Frequency parameters for speech recognition," *Digital Signal Proc. A Review Jour.*, vol. 2, pp. 80–87, April 1992.

[6] J.R. Deller, J.G. Proakis, and J.H.L. Hansen. *Discrete-Time Processing of Speech Signals.* Macmillan, 1993.

[7] B. Tüzün, E. Erzin, M. Demirekler, T. Memişoğlu, S. Uğur, and A.E. Çetin "A speaker independent isolated word recognition system for Turkish," in *NATO-ASI, New Advances and Trends in Speech Recognition and Coding*, Bubion (Granada), June-July 1993.

[8] C. W. Kim, R. Ansari and A. E. Cetin, "A class of linear-phase regular biorthogonal wavelets," *Proc. of ICASSP'92*, pp. IV-673-677, 1992.

[9] E. Zwicker and E. Terhardt, "Analytical expressions for critical band rate and critical bandwidth as a function of frequency," J. Acoust. Soc. America, vol. 68, no. 5, pp. 1523-1525, Dec. 1980.