SUPPLEMENTARY ORTHOGONAL CEPSTRAL FEATURES

Khaled T. Assaleh[1]

Motorola, GSTG
Scottsdale, AZ 85252, USA
email: assaleh@ziggy.geg.mot.com

ABSTRACT

A new set of LP-derived features is introduced. The concept of these features is motivated by the power sum formulation of the LP cepstrum. Due to the fact that the LP model implies that the resulting poles are either real or occur in complex conjugate pairs, the power sum of the poles is equivalent to the power sum of their real components. Therefore, the LP cepstrum is associated to the power sum of the real component of the LP poles. This fact is utilized in deriving a new set of features that is associated to the imaginary components of the LP poles. We refer to this new set of features as the sepstral coefficients. We have found that the sepstral coefficients and cepstral coefficients are relatively uncorrelated. Hence, they can be used jointly to improve the performance of pattern classification applications where cepstral features are usually used. In this paper we present some preliminary results on speaker identification experiments.

1. INTRODUCTION

Feature extraction is the process of deriving a compact set of parameters that are characteristic of a given signal. These parameters are desired to preserve all the information relevant to the application, and to have no redundancy in representing the signal.

In speech processing the majority of pattern classification systems use some type of short time spectral analysis followed by a certain transformation as a feature extraction step. The most effective and widely used spectral analysis techniques are LP analysis and filter bank analysis. This paper deals with LP-derived features.

The short-time transfer function of the LP model is given by:

$$H(z;m) = \frac{1}{A(z;m)} = \frac{1}{1+\sum_{i=1}^{P} a_i(m)z^{-i}} \qquad (1)$$

where $A(z;m)$ is the short-time LP polynomial, m is the frame index representing the temporal dimension, P is the order of the LP model, and $a_i(m)$ is the set of prediction coefficients of the m^{th} frame.

Several feature sets can be derived from $H(z;m)$ [1, 2, 3]. Generally, cepstral features are found to be the most effective.

The short-time LP cepstrum is defined as the inverse z transform of the natural logarithm of the short-time LP transfer function $H(z;m)$. It can be viewed as the impulse response of $\ln H(z;m)$ which is given by:

$$\ln H(z;m) = \sum_{n=1}^{\infty} c_n(m)z^{-n} \qquad (2)$$

where $c_n(m)$ is the n^{th} cepstral coefficient of the m^{th} frame.

A simple and unique recursive relationship between $c_n(m)$ and the prediction coefficients $a_n(m)$ can be obtained by differentiating both sides of equation (2) with respect to z^{-1} and equating the coefficients of equal powers of z^{-1} [1].

An alternative method of obtaining the short-time cepstral coefficients is by relating them to the poles of $H(z;m)$ and hence to the center frequencies and bandwidths of the resonances. The transfer function $H(z;m)$ can be expressed in terms of its short-time poles $z_i(m)$ as:

$$H(z;m) = \frac{1}{\prod_{i=1}^{P}(1-z_i(m)z^{-1})} \qquad (3)$$

By substituting equation (3) in equation (2) one gets

$$\sum_{i=1}^{P} \ln(1-z_i(m)z^{-1}) = -\sum_{n=1}^{\infty} c_n(m)z^{-n}. \qquad (4)$$

[1] At the time of submitting the summary of this paper the author was with the CAIP center at Rutgers University, New Jersey.

0-7803-2431-5/95 $4.00 © 1995 IEEE

The factor $\ln(1 - z_i(m)z^{-1})$ can be expanded [5] as

$$\ln(1 - z_i(m)z^{-1}) = -\sum_{n=1}^{\infty} \frac{1}{n} z_i(m)^n z^{-n}. \qquad (5)$$

By combining equation (4) and equation (5), $c_n(m)$ can be expressed in terms of the roots of the LP polynomial as follows.

$$c_n(m) = \frac{1}{n} \sum_{i=1}^{P} z_i(m)^n. \qquad (6)$$

Thus $c_n(m)$ can be interpreted as the power sum of the LP polynomial roots normalized by the cepstral index [6].

Since $z_i(m)$ is associated with time varying center frequencies $\omega_i(m)$ and bandwidths $B_i(m)$ by the relationship

$$z_i(m) = e^{-B_i(m)+j\omega_i(m)}, \qquad (7)$$

$c_n(m)$ can be expressed as:

$$c_n(m) = \frac{1}{n} \sum_{i=1}^{P} e^{n(-B_i(m)+j\omega_i(m))}$$

$$= \frac{1}{n} \sum_{i=1}^{P} e^{-n(B_i(m)} \left(cos(n\omega_i(m)) + jsin(n\omega_i(m))\right). \qquad (8)$$

The fact that $\{z_i(m)\}$ can either be real or occur in complex conjugate pairs results in the cancellation of the imaginary component of equation (8). Hence, $c_n(m)$ can be expressed as

$$c_n(m) = \frac{1}{n} \sum_{i=1}^{P} e^{-nB_i(m)} cos(n\omega_i(m)). \qquad (9)$$

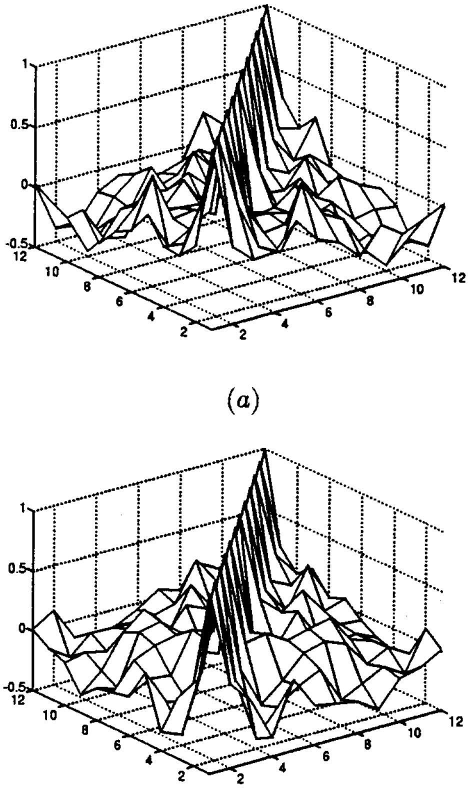
Thus the cepstral coefficients are associated with the real components of $z_i(m)$, and can be interpreted as a nonlinear transformation of the center frequencies and bandwidths.

2. SEPSTRAL COEFFICIENTS

The representation given in equation (9) suggests various alternate representations based on the formant center frequencies and bandwidths. For example, the orthogonal complement to c_n can be used, i.e.,

$$s_n(m) = \frac{1}{n} \sum_{i=1}^{P} e^{-nB_i(m)} sin(|n\omega_i(m)|), \qquad (10)$$

which we call the sepstral coefficients. The sepstral coefficients differ from the cepstral coefficients in that the sine function alters the contribution of the poles in the computed coefficients. For example, it filters out the real poles' contribution, as opposed to the cosine function which weights those poles more heavily. The absolute value of the frequencies in equation (10) is used to avoid cancellations.



(a)

(b)

Figure 1: The covariance matrix of (a) the cepstral coefficients, and (b) the sepstral coefficients, both normalized by their standard deviations.

Similar to the cepstral coefficients, the sepstral coefficients possess the desirable property of having a nearly diagonal covariance matrix (i.e., have little redundancy in representing the signal). Figure (1) shows mesh plots of the covariance matrices of $c_n(m)$ and $s_n(m)$ normalized by their standard deviations. They are extracted from a 10 sec utterance partitioned into overlapping frames of 30 msec length and 15 msec overlap. The sepstral coefficients are found to be nearly orthogonal to the cepstral coefficients. Hence,they can be used as a reinforcing set of features with the cepstral coefficients for a better recognition performance. To demonstrate the orthogonality of $s_n(m)$ with respect to

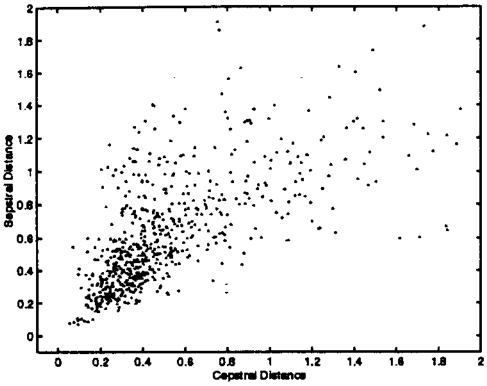


Figure 2: Scatter plots of cepstral vs. sepstral distances

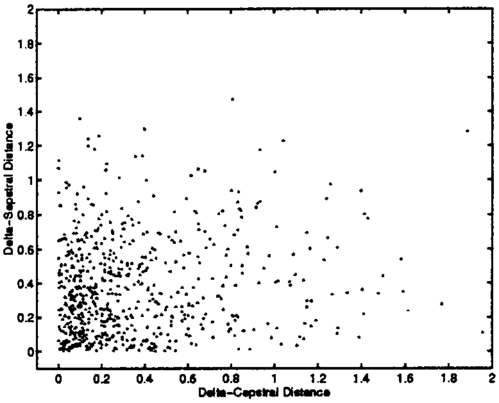


Figure 3: Scatter plots of Δ-cepstral vs. Δ-sepstral distances

$c_n(m)$ we show a scatter plot of the cepstral distance d_{cep} versus the sepstral distance d_{sep} for a sequence of speech frames. These distances are computed from $c_n(m)$ and $s_n(m)$ (compared in a consecutive manner) as follows:

$$d_{cep}(m) = \sum_n (c_n(m) - c_n(m-1))^2. \quad (11)$$

$$d_{sep}(m) = \sum_n (s_n(m) - s_n(m-1))^2. \quad (12)$$

Figure (2) shows a scatter plot of $d_{cep}(m)$ versus $d_{sep}(m)$. The normalized correlation coefficient between these two distances is found to be 0.55. Since both cepstral and sepstral features carry similar information, the value of 0.55 is not high. This based on the argument of Soong and Rosenberg [7] where they stated that a value of 0.6 for the correlation between cepstral and differential cepstral (Δ-cepstrum) distances is relatively low.

The differential sepstral features (Δ-sepstrum) were also examined and found to be almost completely uncorrelated to the Δ-cepstral features. The normalized correlation coefficient between the Δ-cepstral distance and the Δ-sepstral distance is found to be less the 0.1. The scatter plot between these two distances is shown in figure (3).

3. PRELIMINARY EXPERIMENTAL RESULTS

A text independent speaker identification experiment is conducted on the San Diego speakers of the narrowband portion of the King database. For the results reported in this paper, training is done on one session (session 1), while testing is done on each of the other nine. Due to the division of the data, testing on sessions among 2 to 5 is denoted by the "**within the great divide**" experiment, whereas testing on sessions among 6 to 10 is denoted by the "**across the great divide**" experiment. The classifier used here is a VQ classifier [4]. During training, a codebook of 46 codewords is constructed to model each speaker. Upon identifying an unknown speaker, each test vector is compared to the codebook of each speaker. The codebook entries which are closest to the test vectors are found using full search, and the corresponding distances are recorded. The distances are accumulated for each codebook and the unknown speaker's identity is chosen as the one corresponding to the codebook associated with the minimum accumulated distance.

Sepstral and cepstral features are combined using the same method utilized for combining instantaneous and transitional cepstral features [7].

Table 1 shows the identification results using cepstral coefficients and using combined cepstral & sepstral coefficients. The "within the great divide" experiments shows no improvement due to the combining of the cepstral and sepstral features. However, the "across the great divide" shows over 8% improvement in the identification percentage over using cepstral features alone. This suggests that the benefit of combining cepstral and sepstral features is manifested when there is a mismatch between the training and the testing data. The improvement could be attributed to the fact that sepstral coefficients filter out the real poles which only contribute to overall spectral slope. It is well-known that the overall spectral slope is greatly affected by mismatched channels and microphones between training and testing data.

test session	cepstrum	cepstrum & sepstrum
2	23/26	23/26
3	17/26	16/26
4	16/26	17/26
5	16/26	18/26
average	69.23%	71.15%

"within the great divide"

test session	cepstrum	cepstrum & sepstrum
6	11/26	12/26
7	12/26	13/26
8	13/26	15/26
9	12/26	15/26
10	12/26	16/26
average	46.3%	54.6%

"across the great divide"

Table 1: Incorporating sepstral coefficients with cepstral coefficients

4. SUMMARY AND CONCLUSION

In this paper we have introduced a new set of features associated with the imaginary components of the LP poles. These features are referred to as the sepstral features. The sepstral features are found to be relatively uncorrelated to the cepstral features. Therefore both cepstral and sepstral features can be combined to yield improved recognition rates for different applications. In a preliminary speaker identification experiment performed on a portion of the King database the combining of the cepstral and sepstral features is found to be advantageous especially when there is a mismatch between the training and the testing data. It should be noted that this experiment is by no means conclusive, and additional experiments need to conducted. Also, the suggested features are not specific for speaker identification. Hence, they can be tested for other applications such as speech recognition. The differential sepstral features are found to be almost completely uncorrelated with the differential cepstral features. This finding was not used in the provided experiments since differential features in general were not found to be helpful in this particular application [8].

5. ACKNOWLEDGMENT

The author gratefully acknowledges the support of Rome Laboratories, Contract No. F30602-91-C-0120. Also, the author would like to thank Professor Richard Mammone of Rutgers University for his useful comments.

6. REFERENCES

[1] B. Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *Journal of the Acoustical Society of America*, 55:1304–1312, June 1974.

[2] S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE IEEE Trans. Acoust., Speech, Signal Process.*, ASSP-29:254–272, April 1981.

[3] J.P. Campbell. *Features and measures for speaker recognition.* Ph.D. thesis, Oklahoma State University, December 1992.

[4] F.K. Soong, A.E. Rosenberg, L.R. Rabiner, and B.H. Juang. A vector quantization approach to speaker recognition. In *Proc. Intl. Conf. Acoust., Speech, Signal Process.*, pages 387–390, 1985.

[5] A.V. Oppenheim and R.W. Schafer. Homomorphic analysis of speech. *IEEE Transactions on Audio and Electroacoustics*, AU-16:221–226, June 1968.

[6] M. R. Schroeder. Direct (nonrecursive) relations between cepstrum and and predictor coefficients. *IEEE Trans. Acoust., Speech, Signal Process.*, 29:297–301, Apr. 1981.

[7] F. K. Soong and A. E. Rosenberg. On the use of instantaneous and transitional spectral information in speaker recognition. *IEEE Trans. Acoust., Speech, Signal Process.*, ASSP-36:871–879, June 1988.

[8] K. T. Assaleh and R. J. Mammone. New LP-derived features for speaker identification. *IEEE Trans. Speech and Audio Processing.*, October 1994.