FREQUENCY-BAND SPECIFICATION IN CEPSTRAL DISTANCE COMPUTATION

Frantz Clermont and Parham Mokhtari

Department of Computer Science,
University College, University of New South Wales,
Australian Defence Force Academy,
Australia

ABSTRACT - Distances derived from the all-pole, Linear-Prediction (LP) cepstrum are known for their ability to capture important spectral differences between speech sounds with relatively small computational complexity, and hence are widely used for computer speech and speaker recognition. However, these so-called cepstral distances have todate been formulated in such a way as to yield similarity measures, which are integrated over the *entire* spectral range defined between zero Hertz and half the sampling frequency. While this time-honoured formulation can be very effective, it is limited by the fact that arbitrary frequency bands within the available spectral range cannot be isolated or emphasised in the distance computation itself. In this paper we show that the existing mathematical framework for deriving LP-cepstrum distances is amenable to one which permits direct frequency-band specification. In particular, the quefrency-weighted cepstral distance, also known as a spectral slope distance, is re-formulated as a parametric function of frequency, and then illustrated using directly selected frequency bands of a pair of speech spectra.

INTRODUCTION

A relevant issue in the search for robust methods of computer speech and speaker recognition is the choice of a suitable distance measure. Much progress in this area has been prompted by the development of the all-pole, Linear Prediction (LP) model of speech (Atal and Schroeder, 1967; Markel and Gray, 1976), which has led to distance formulations more closely related to the spectral properties of speech sounds.

The unweighted Euclidean distance based on the low-order LP-cepstrum, for example, is particularly well-known for its computational simplicity, and for its strong correlation with the log spectral distance used as a benchmark by Gray and Markel (1976). However, this cepstral distance naturally tends to weight differences equally across the frequency spectrum, and thus is expected to be sensitive not only to peak frequency but also to peak amplitude and spectral tilt to some degree. This lack of localised sensitivity is generally regarded as an impediment to robust speech recognition, and consequently more specialised forms of cepstral distances have been developed.

The proposal of a cepstral distance sensitive mainly to differences around spectral peaks, appears to have first emerged from Yegnanarayana and Reddy's (1979) study of the consequences of the peak-sensitivity property of the first derivative of the LP-phase spectrum. Yegnanarayana and Reddy showed this property to be indeed embodied in the *index-weighted or quefrency-weighted cepstral* (QCEP) distance, which they derived from the squared Euclidean distance between the first derivatives or equivalently between the respective local slopes of two LP-phase spectra. Only a few years later, some strong evidence supporting the likely importance of spectral slope in computer speech recognition, arose from Klatt's (1982) finding that only deviations in spectral peak frequency contribute significantly to human perception of phonetic change in vowels and fricatives.

Since then, a series of studies (e.g., Hanson and Wakita, 1986; Juang *et al.*, 1986; Tohkura, 1986; Itakura and Umezaki, 1987) have sought to enhance the performance of cepstral distances, by designing weighting windows (referred to as lifters) aimed at reducing the effects of certain LP-cepstrum coefficients. The results therein reported indicate that speech or speaker recognition accuracy can be improved by applying cepstral lifters: (1) which give more weight to spectral slope differences; and (2) which help to control "non information-bearing cepstral variabilities" (Rabiner and Juang, 1993: p. 169) caused by analysis window positions, transmission line characteristics, or by certain inter-speaker differences. Although the liftering approach represents a significant step towards obtaining robust cepstral distance measures, it does not offer direct control of arbitrary frequency bands within the available spectral range.

In this paper, we endeavour therefore to show that LP-cepstrum distances, defined traditionally over the entire available spectral range, can be directly computed within any desirable frequency band. In particular, the quefrency-weighted cepstral distance is re-defined as a parametric function of frequency, and then illustrated within selected frequency bands of a pair of speech spectra.

MATHEMATICAL DERIVATION OF THE QUEFRENCY-WEIGHTED CEPSTRAL DISTANCE

Closed-form representation: $[D_1(0,\pi)]^2$

The well-known method of linear prediction of speech samples yields an equivalent linear model for speech production, which can be represented by an all-pole filter of transfer function $H(z)$ defined as follows:

$$H(z) = \frac{\sigma}{A(z)} = \frac{\sigma}{1 + \sum_{k=1}^{M} a_k z^{-k}} \quad , \tag{1}$$

where σ is a gain factor, a_k are the linear-prediction or autoregressive coefficients of the inverse filter polynomial $A(z)$, and M is the polynomial order. By evaluating $A(z)$ on the unit circle (i.e., at $z = e^{j\theta}$), the inverse filter's frequency response $A(e^{j\theta})$ can be decomposed into its magnitude $|A(e^{j\theta})|$ and phase $-\phi(e^{j\theta})$ components. The following equation describes these components with the *normalised* frequency $\theta = \pi$ at half the sampling frequency f_s:

$$\begin{aligned} ln[A(e^{j\theta})] &= ln[|A(e^{j\theta})|\, e^{-j\phi(e^{j\theta})}] \\ &= ln\,|A(e^{j\theta})| - j\phi(e^{j\theta}). \end{aligned} \tag{2}$$

Assuming a stable all-pole filter with its roots lying inside the unit circle, the logarithm of $A(e^{j\theta})$ can further be expanded as a Taylor series of the so-called LP-cepstrum coefficients C_k:

$$\begin{aligned} ln[A(e^{j\theta})] &= -\sum_{k=1}^{\infty} C_k\, e^{-jk\theta} \\ &= -\sum_{k=1}^{\infty} C_k\, cos(k\theta) + j \sum_{k=1}^{\infty} C_k\, sin(k\theta). \end{aligned} \tag{3}$$

Of special interest in this study is the Negative Derivative of the linear-prediction Phase Spectrum (NDPS), which can easily be related to the LP-cepstrum by equating the imaginary parts of Equations 2 and 3, and then taking the negative of the first derivative with respect to the normalised frequency θ, the result of which is as follows:

$$-\frac{d\phi(e^{j\theta})}{d\theta} = \sum_{k=1}^{\infty} kC_k\, cos(k\theta). \tag{4}$$

It is the squared Euclidean distance between an NDPS pair that Yegnanarayana and Reddy (1979) used as a basis for deriving the quefrency-weighted cepstral distance. The mathematical steps followed by these authors are presented below:

$$(D_1)^2 = \frac{1}{\pi} \int_0^{\pi} \left[\left(-\frac{d\phi(e^{j\theta})}{d\theta}\right) - \left(-\frac{d\phi'(e^{j\theta})}{d\theta}\right) \right]^2 d\theta \tag{5}$$

$$= \frac{1}{2} \sum_{k=1}^{\infty} [k(C_k - C'_k)]^2 \, , \tag{6}$$

where $\phi(e^{j\theta})$ and $\phi'(e^{j\theta})$ are the LP-phase spectra of two speech frames to be compared with each other, C_k and C'_k are the LP-cepstrum coefficients for the two frames, and k represents the *quefrency* of each coefficient. Equation 6 then describes the so-called quefrency-weighted cepstral distance which is readily obtained by: (1) substituting for Equation 4 into Equation 5; and (2) then using *Parseval's identity* which in this case eliminates the need to carry out the integration in Equation 5. Note that, in practice, the summation in Equation 6 is truncated to a small number of terms, which is usually set equal to or moderately greater than the order of the LP-filter (Gray and Markel, 1976; Yegnanarayana and Reddy, 1979), in order to retain in the distance measure essential features of the pair of resulting *cepstrally-smoothed* spectra.

Parametric representation: $[D_2(\theta_1, \theta_2)]^2$

An important property embodied in Equation 6 is the closed-form representation of $[D_1(0, \pi)]^2$ in the spectral domain. That is, the distance $(D_1)^2$ between a pair of speech frames will yield a measure of spectral similarity, which is integrated over the entire spectral range defined between zero Hertz and half the sampling frequency of the corresponding speech signals. Although it is possible to reduce the contribution of certain frequency bands within that range by using appropriate cepstral lifters or a non-linear scaling of the spectrum (e.g., the mel scale), these approaches do not afford the flexibility of directly selecting a particular band in the distance computation. In order to achieve this flexibility, we re-consider Equation 5 with a view to deriving a parametric form of the QCEP distance, which depends on the lower and upper limits $[\theta_1, \theta_2]$ of the selected frequency band.

Assuming cepstral series truncated to M terms, and assuming θ_1 and θ_2 to be the *normalised* limits of any frequency band of interest within the range $[0, \pi]$, the substitution for Equation 4 into Equation 5 yields the following expressions:

$$\begin{aligned}(D_2)^2 &= \frac{1}{(\theta_2 - \theta_1)} \int_{\theta_1}^{\theta_2} \left[(-\frac{d\phi(e^{j\theta})}{d\theta}) - (-\frac{d\phi'(e^{j\theta})}{d\theta}) \right]^2 d\theta \\ &\simeq \frac{1}{(\theta_2 - \theta_1)} \int_{\theta_1}^{\theta_2} \left[\sum_{k=1}^{M} k\,(C_k - C'_k)\,cos(k\theta) \right]^2 d\theta. \end{aligned} \tag{7}$$

Parseval's theorem cannot be applied in this case, and thus the squared summation in Equation 7 is first decomposed and then integrated, according to the following equation:

$$(D_2)^2 = \frac{1}{(\theta_2 - \theta_1)} [A(\theta_1, \theta_2) + B(\theta_1, \theta_2)], \tag{8}$$

where:

$$A(\theta_1, \theta_2) = \int_{\theta_1}^{\theta_2} \sum_{k=1}^{M} \left[k\,(C_k - C'_k)\,cos(k\theta) \right]^2 d\theta. \tag{9}$$

$$B(\theta_1, \theta_2) = \int_{\theta_1}^{\theta_2} \left[2 \sum_{k=1}^{M-1} \sum_{\ell=k+1}^{M} k\ell(C_k - C'_k)(C_\ell - C'_\ell) cos(k\theta) cos(\ell\theta) \right] d\theta. \tag{10}$$

The additive components $A(\theta_1, \theta_2)$ and $B(\theta_1, \theta_2)$ of the QCEP distance $(D_2)^2$ re-formulated in Equation 8, are themselves obtained by carrying out the integrations specified in Equations 9 and 10, the results of which are summarised below:

$$A(\theta_1, \theta_2) = \sum_{k=1}^{M} \left[k(C_k - C'_k) \right]^2 \cdot \alpha_k, \tag{11}$$

where:

$$\alpha_k(\theta_1, \theta_2) = \frac{(\theta_2 - \theta_1)}{2} + \frac{sin(2k\theta_2) - sin(2k\theta_1)}{4k}, \tag{12}$$

and

$$B(\theta_1, \theta_2) = \sum_{k=1}^{M-1} \sum_{\ell=k+1}^{M} \left[k\ell(C_k - C'_k)(C_\ell - C'_\ell) \right] \cdot \beta_{k\ell}, \tag{13}$$

where:

$$\beta_{k\ell}(\theta_1, \theta_2) = \frac{sin(k - \ell)\theta_2 - sin(k - \ell)\theta_1}{(k - \ell)} + \frac{sin(k + \ell)\theta_2 - sin(k + \ell)\theta_1}{(k + \ell)}. \tag{14}$$

Equations 8, and 11 through 14 describe a new, parametric formulation $[D_2(\theta_1, \theta_2)]^2$ of the QCEP distance, which not only embodies the spectral-slope sensitivity of the NDPS, but which can now be computed over any frequency band $[\theta_1, \theta_2]$ of interest within the available spectral range $[0, \pi]$.

Matrix formulation of $[D_2(\theta_1, \theta_2)]^2$

For the sake of economy in formulaic description and efficiency in computer implementation, a matrix notation is sought for the parametric QCEP distance $[D_2(\theta_1, \theta_2)]^2$ derived earlier. Using Equations 11 and 13, Equation 8 can be re-written as follows:

$$[D_2(\theta_1, \theta_2)]^2 = \frac{1}{(\theta_2 - \theta_1)} \left[\underbrace{\sum_{k=1}^{M} (\delta_k)^2 \cdot \alpha_k}_{} + \underbrace{\sum_{k=1}^{M-1} \sum_{\ell=k+1}^{M} (\delta_k \delta_\ell) \cdot \beta_{k\ell}}_{} \right] , \tag{15}$$

where:

$$\delta_k = [k(C_k - C'_k)] \ , \ \delta_\ell = [\ell(C_\ell - C'_\ell)] \ .$$

Next, the first and second terms highlighted on the right-hand side of Equation 15 are, respectively, expressed in matrix form as $[\underline{d}^T \mathsf{A} \underline{d}]$ and $[\underline{d}^T \mathsf{B} \underline{d}]$, where A is a diagonal matrix, B is a strictly triangular matrix, $\underline{d}$ is a column vector and $\underline{d}^T$ is the transpose of $\underline{d}$, as shown below:

$$\mathsf{A} = \begin{bmatrix} \alpha_1 & 0 & \dots & 0 \\ 0 & \alpha_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \alpha_M \end{bmatrix} , \ \mathsf{B} = \begin{bmatrix} 0 & \beta_{12} & \beta_{13} & \dots & \beta_{1,M} \\ \vdots & 0 & \beta_{23} & \dots & \beta_{2,M} \\ \vdots & & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \beta_{(M-1),M} \\ 0 & \dots & \dots & \dots & 0 \end{bmatrix} , \ \underline{d}^T = (\delta_1 \ \delta_2 \ \cdots \ \delta_M).$$

Using A, B, $\underline{d}$ and $\underline{d}^T$ defined above, Equation 15 can finally be expressed using the following matrix notation:

$$[D_2(\theta_1, \theta_2)]^2 = \frac{1}{(\theta_2 - \theta_1)} \left[\underline{d}^T \mathsf{W} \underline{d} \right] , \tag{16}$$

where W = [A + B] is a triangular matrix the elements of which (α_k and $\beta_{k\ell}$) are independent of the cepstral coefficients, but are dependent on the frequency-band parameters θ_1 and θ_2 as shown in Equations 12 and 14.

ILLUSTRATION OF $[D_2(\theta_1, \theta_2)]^2$: SPECTRAL SLOPE and FREQUENCY-BAND SELECTION

We now offer an illustration of the parametric QCEP distance formulated above, as it operates on directly selected frequency bands of a pair of speech spectra. For this purpose, two frames 60-msec distant in time from each other were extracted from the diphthong nucleus of the monosyllabic word "hoy" spoken by an adult, male speaker of Australian English.

The log-magnitudes of the LP-spectra of the two frames are superimposed in Figure 1(a), and appear to be markedly shifted with respect to each other in certain spectral regions. Indeed, the respective first, second and fourth formants (F_1, F_2 and F_4) differ in amplitude but only slightly in frequency, the respective third formants (F_3) do differ primarily in frequency, while the respective fifth formants (F_5) do differ both in amplitude and frequency. There are also notable differences in the valleys intermediate between formant peaks. However, by virtue of the peak-enhancement property of the NDPS described by Fuchi and Ohta (1977) and by Yegnanarayana (1978), it can be observed that only the regions of maximum slope located in Figure 1(a) near spectral peaks are emphasised in the corresponding NDPS representations shown in Figure 1(b).

Figures 1(a) and 1(b) illustrate "exact" spectral representations as they are obtained using an FFT of the LP-autoregressive coefficients. For the sake of completeness, Figure 1(c) shows, for the same pair of speech frames, the corresponding *cepstrally-smoothed* NDPS which are *implicitly* compared in the cepstral distance computation. These spectra are generated from a cosine expansion of a *finite* number (e.g., *M = 14*) of quefrency-weighted cepstral coefficents (i.e., a practical implementation of the right-hand side of Equation 4). As a result, the features of the exact NDPS representations become somewhat blurred in Figure 1(c), but retained in their essence.

In Figure 1(d), we illustrate the flexibility of our new parametric formulation and highlight the spectral slope sensitivity of the QCEP distance $[D_2(\theta_1, \theta_2)]^2$, as it is applied within 6 consecutive, but independent frequency bands of 600-Hz width. The very small measure obtained within the lowest frequency

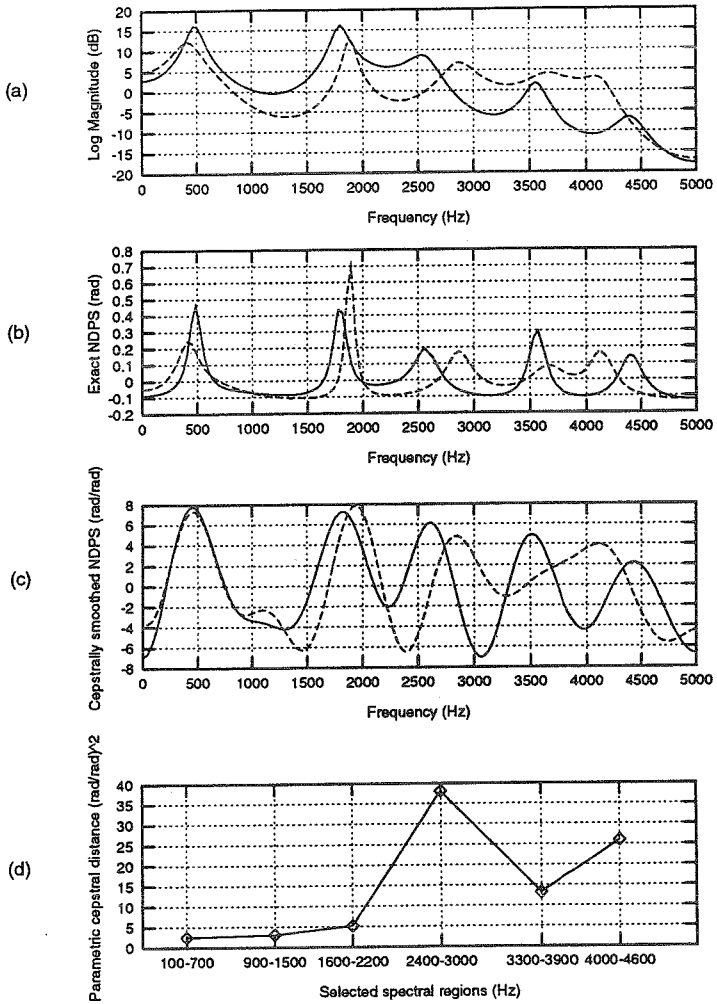

Figure 1: *Illustration of* $[D_2(\theta_1, \theta_2)]^2$ between a pair of frames chosen 60-msec apart from each other in the diphthong nucleus of the monosyllabic word "hoy" spoken by an adult male, speaker of Australian English. *Analysis conditions*: Sampling frequency f_s is 10 kHz; LP-order is set to 14; M (number of LP-cepstrum coefficients) is also set to 14; Hamming window size is 25.6 msecs; frame advance is 10 msecs; FFT size is 256 points. *Graph (a)*: Superimposed, log-magnitude of LP-spectra of the two frames. *Graph (b)*: Superimposed, exact NDPS of the two frames. *Graph (c)*: Superimposed, cepstrally-smoothed NDPS of the two frames. *Graph (d)*: Profile of distance measures computed within 6 consecutive, but non-contiguous frequency bands. The Hertz values (f) shown for the lower and upper limits of each band are first converted into radians (θ) via $\theta = f \cdot \frac{2\pi}{f_s}$, and then substituted into $[D_2(\theta_1, \theta_2)]^2$.

band (100, 700 Hz) chosen to span the F_1 peaks, is indicative of insensitivity to changes in formant amplitude. The second band (900, 1500 Hz) is chosen to cover the spectral valley between F_1 and F_2. In spite of the large differences in this spectral region which are clearly evident in the log-magnitude LP-spectra shown in Figure 1(a), the resulting distance measure remains relatively small. Similarly, in the frequency bands [(1600, 2200 Hz), (3300, 3900 Hz)] spanning the F_2 and F_4 peaks respectively, the distance measures obtained are still relatively small, and thus changes in amplitude are definitely not very influential. In sharp contrast, the maximum measure obtained in the frequency band (2400, 3000 Hz) encompassing the F_3 peaks, clearly illustrates the higher degree of sensitivity of the QCEP distance to deviations in formant frequency. A similar observation can be made for the frequency band inclusive of the fifth formant regions (4000, 4600 Hz).

CONCLUSION

We have derived and illustrated a new formulation of the quefrency-weighted cepstral distance, which allows direct selection of any frequency band $[\theta_1, \theta_2]$ within the available spectral range $[0, \pi]$. This flexibility in cepstral distance computation represents a further step towards gaining more control of spectral information using the low-order LP-cepstrum. As a result, we conjecture that the relative importance of various frequency bands for computer speech and speaker recognition could be more easily studied.

REFERENCES

Atal, B.S. and Schroeder, M.R. (1967), "Predictive coding of speech signals", Proceedings of the Conference on Speech Communication and Processing, pp. 360-361.

Fuchi, K. and Ohta, K. (1977), "Observation on group delay characteristics of connected vowels", *Electrotechnical Laboratory, Speech Processing Section: Progress Report on Speech Research*, pp. 44-47.

Gray, A.H., Jr. and Markel, J.D. (1976), "Distance measures for speech processing", IEEE Transactions on Acoustics, Speech and Signal Processing *24(5)*, pp. 380-391.

Hanson, B.A. and Wakita, H. (1986), "Spectral slope distortion measures for all-pole models of speech", Proceedings of the International Conference on Acoustics, Speech and Signal Processing, pp. 757-760.

Itakura, F. and Umezaki, T. (1987), "Distance measure for speech recognition based on the smoothed group delay spectrum", Proceedings of the International Conference on Acoustics, Speech and Signal Processing, pp. 1257-1260.

Juang, B.H., Rabiner, L.R. and Wilpon, J.G. (1986), "On the use of bandpass liftering in speech recognition", Proceedings of the International Conference on Acoustics, Speech and Signal Processing, pp. 765-768.

Klatt, D.H. (1982), "Prediction of perceived phonetic distance from critical-band spectra: A first step", Proceedings of the International Conference on Acoustics, Speech and Signal Processing, pp. 1278-1281.

Markel, J.D. and Gray, A.H., Jr. (1976), *Linear Prediction of Speech* (Springer-Verlag: Berlin).

Rabiner, L.R. and Juang, B.H. (1993), *Fundamentals of Speech Recognition* (Prentice Hall: New Jersey).

Tohkura, Y. (1986), "A weighted cepstral distance measure for speech recognition", Proceedings of the International Conference on Acoustics, Speech and Signal Processing, pp. 761-764.

Yegnanarayana, B. (1978), "Formant extraction from linear-prediction phase spectra", Journal of the Acoustical Society of America *63*, pp. 1638-1640.

Yegnanarayana, B. and Reddy, R. (1979), "A distance measure derived from the first derivative of linear prediction phase spectrum", Proceedings of the International Conference on Acoustics, Speech and Signal Processing, pp. 744-747.