IMPROVED SCALE-CEPSTRAL ANALYSIS IN SPEECH

S. Umesh — Dept. of Electrical Engg., I.I.T., Kanpur-208016, India

L. Cohen — Hunter College of CUNY, New York, NY 10021, USA

D. Nelson — U.S. Dept. of Defense, Ft. Meade, MD 20755, USA

ABSTRACT

In this paper, we present improvements over the original scale-cepstrum proposed in [1]. The scale-cepstrum was motivated by a desire to normalize the first-order effects of differences in vocal-tract lengths for a given vowel. Our subsequent work [2] has shown that a more appropriate frequency-warping than the log-warping used in [1] is necessary to account for the frequency dependency of the scale-factor. Using this more appropriate frequency-warping and a modified method of computing the scale-cepstrum we have obtained improved features that provide better separability between vowels than before, and are also robust to noise.

1. INTRODUCTION

Recently, we have proposed the use of Scale-Transform based features as acoustic features in speech analysis [1]. The motivation for using Scale-Transform stems from a desire to normalize the first-order effects of differences in vocal-tract lengths. A simplified model for vocal-tract is a uniform tube of length, L, and the corresponding frequency spectrum is given by

$$F_n = \frac{(2n+1)c}{4L} \tag{1}$$

where c is the velocity of sound. Note that the uniform tube model is not the best model for vocal-tract, but to a first order effect such a linear scaling of frequency axis may be assumed [3, 4, 5]. Hence to a first-order approximation, the formant envelopes of different speakers are frequency-scaled versions of one and another for a given vowel, i.e.,

$$A(f) = B(\alpha_{AB} f). \tag{2}$$

where $\alpha_{AB} = \frac{L_A}{L_B}$. The Scale-Transform [6] is a useful tool to analyze such signals that are scaled versions of one and another. The Scale-Transform of a function, $X(f)$, is given by,

$$D_X(c) = \int_0^\infty X(f) \frac{e^{-j2\pi c \ln f}}{\sqrt{f}} df. \tag{3}$$

This work was supported in part by HBCU/MI Program and by the Indian AICTE Career Award for Young Teachers.

One of the interesting properties of the Scale-Transform is that the magnitude of the Scale Transform of a function, $X(f)$ and its scaled version, $\sqrt{\alpha}X(\alpha f)$ are the same. This property is exploited in [1] to normalize the first-order effects of linear frequency-scaling of formant envelopes for a given vowel.

The scale-transform may also be computed as the Fourier transform of the function $X(e^f)e^{f/2}$, i.e.

$$D_X(c) = \int_{-\infty}^{\infty} X(e^f) e^{f/2} e^{-j2\pi cf} df. \tag{4}$$

Note that as a result of log-warping, i.e. forming $X(e^f)$, the scale constant, α, is a function of the translation parameter in the log-warped domain.

In [1], we computed the scale-cepstrum of the formant envelope $X(f)$ using the formula,

$$D_S(c) = \int (\log|X(e^f)|) e^{\frac{f}{2}} \; e^{-j2\pi cf} df, \tag{5}$$

Note that the logarithm operation affects only the magnitude of the spectral components. Therefore, functions that are frequency-scaled versions of each other continue to remain so even after the logarithm is taken. The scale-cepstrum differs from the conventional cepstrum in that the function is frequency-warped before the logarithm operation and there is an additional emphasis factor $e^{\frac{f}{2}}$.

2. FREQUENCY-WARPING

In [1], the scale-factor α_{AB} in Equation 2 is assumed to be a constant independent of frequency. However, in our subsequent work based on experiments on actual speech data, we found evidence that the scale-factor is not independent of frequency [2]. In such a case, the log-warping may not be the appropriate warping function. Therefore, in [2], we address the problem of finding a more appropriate warping function to account for this frequency-dependency of the scale-factor. The frequency band between 100 Hz and 7000 Hz is

divided into logarithmically equal bands of [100,240) Hz, [240,550) Hz, [550,1280) Hz, [1280,3000) Hz and [3000,7000) Hz. This is done to obtain a piece-wise approximation to the warping function. The scale factor is assumed to be constant within each frequency band, but its value may vary across the different frequency bands, i.e.

$$A(f) = B(\alpha_{AB}^{(i)} f) \quad f \in i^{th} \text{ band}, \tag{6}$$

and $i = 1, 2, \cdots, 5$. Further we will assume $\alpha_{AB}^{(i)}$ to be of the form

$$\alpha_{AB}^{(i)} = \alpha_{AB}^{(1+\beta_i)} = \alpha_{AB} \cdot \alpha_{AB}^{\beta_i}. \tag{7}$$

Note that α_{AB} is a constant independent of i (the frequency band) and is dependent on the pair of speakers, while β_i depends only on the i^{th} frequency band and is independent of the pair of speakers. The modified warping function is $X(e^{(1+\beta_i)f})$. In [2], we have studied the use of such a warping operation on utterances spoken by many speakers and have found that they are essentially translated versions of one and another.

3. MODIFIED SCALE-CEPSTRUM

If we had two functions, $X_1(f) = A_1 X(\alpha_1 f)$ and $X_2(f) = A_2 X(\alpha_2 f)$, then the respective Scale-Cepstrums are,

$$\begin{aligned} D_{X1}(c) &= \int_{-\infty}^{\infty} \log(A_1 X(e^{f+\log\alpha_1}) e^{f/2} e^{-j2\pi cf} df \\ &= \log(A_1)\delta(c) + \frac{e^{j2\pi c \log\alpha_1}}{\sqrt{\alpha_1}} D_X(c), \end{aligned} \tag{8}$$

and,

$$D_{X2}(c) = \log(A_2)\delta(c) + \frac{e^{j2\pi c \log\alpha_2}}{\sqrt{\alpha_2}} D_X(c) \tag{9}$$

For $c \neq 0$ (i.e. for values other than scale-DC) the two scale-cepstrums differ only by the term $\frac{e^{j2\pi c \log\alpha_i}}{\sqrt{\alpha_i}}$ that depends on the respective scale-factor α_i. Therefore, if we take the magnitude and normalize to unit energy, we get identical terms. This was the procedure used in [1]

We now propose the following modification for computing the scale-cepstrum

$$D_X(c) = \int_{-\infty}^{\infty} \log|(X(e^f)|e^{\frac{f}{2}}) \; e^{-j2\pi cf} df, \tag{10}$$

It can be easily verified that $\sqrt{\alpha} X(\alpha f)$ will have the same modified scale-cepstrum except for the phase factor, since the log-operation affects only the amplitude and not the frequency-scaling. The modified scale-cepstrum for $X_1(f)$ and $X_2(f)$ are:

$$D_{X1}(c) = \log(\frac{A_1}{\sqrt{\alpha_1}})\delta(c) + e^{j2\pi c \log\alpha_1} D_X(c) \tag{11}$$

$$D_{X2}(c) = \log(\frac{A_2}{\sqrt{\alpha_2}})\delta(c) + e^{j2\pi c \log\alpha_2} D_X(c) \tag{12}$$

For values other than scale-DC the magnitude of the scale-cepstrum for the two functions are identical. There is no need to normalize their energy. Simulation results indicate that the use of such scale-cepstral features along with the modified warping function provide better separability of vowels than before, as seen in the Section 5.

4. DISCRETE IMPLEMENTATION

Since the sampling frequency of the TIMIT database is 16 KHz, for computations in this paper, we assume that the signal is bandlimited between 100 Hz and 7000 Hz. The *modified* scale-cepstrum of Equation 10 for log-warping may be digitally implemented using the Fast Fourier Transform (FFT), i.e.

$$\begin{aligned} D_S[\frac{k_c C_p}{N}] &= \sum_{m=0}^{K-1} \log(|S(e^{m\Delta\nu+\ln(100)})|e^{\frac{m\Delta\nu+\ln(100)}{2}}) \\ &\quad \times e^{-j2\pi\frac{k_c}{N}m} \quad k_c = 0, 1, \cdots, (N-1) \end{aligned} \tag{13}$$

where $\Delta\nu = \frac{\ln(7000)-\ln(100)}{K-1}$ and $C_p = \frac{1}{\Delta\nu}$. The phase term $e^{\frac{-j2\pi k_c C_p \ln(100)}{N}}$ can be ignored, since it does not contribute to the magnitude of $|D_S[\frac{k_c C_p}{N}]|$.

$S(e^{m\Delta\nu+\ln(100)})$ can be easily computed from the time-lag samples of the smoothed formant-envelope, $s[n]$ as

$$\begin{aligned} S(e^{m\Delta\nu+\ln(100)}) &= \sum_{n=0}^{L-1} s[n] e^{-j2\pi e^{|m\Delta\nu+\ln(100)|} n T_s}, \\ m &= 0, 1, \cdots, (K-1), \end{aligned} \tag{14}$$

where T_s is the sampling period in the time-lag domain. The procedure to obtain the smoothed envelope, $s[n]$ is described in [1], but a brief summary of the procedure is given in the next section. The above procedure may be used to implement log-warping.

For discrete implementation of the proposed piecewise warping function, we need to compute $B(e^{(1+\beta_i)f})$ for $e^{(1+\beta_i)f} \epsilon [U_i, L_i]$, where U_i and L_i are upper and lower frequency limits of the i^{th} frequency band. We discretize by computing $B(e^{(1+\beta_i)f})$ at M_i equally spaced

Band 1	*Band 2*	*Band 3*	*Band 4*	*Band 5*
5.0	3.3869	1.4629	0.4616	0

Table 1: The β_i are estimated as described in [2].

intervals in the region $\log(L_i)$ to $\log(U_i)$. The sampling period is therefore

$$\triangle\nu_i = \frac{\log(U_i) - \log(L_i)}{(1+\beta_i)M_i}. \quad (15)$$

Recall, that we have chosen frequency bands that are equally spaced on the logarithm scale, hence, we have

$$\log(U_i) - \log(L_i) = \log(U_j) - \log(L_j). \quad (16)$$

To have equally spaced samples in the warped domain, we require that $\triangle\nu_i$ be a constant independent of i. This is satisfied if,

$$(1+\beta_k)M_k \;=\; (1+\beta_j)M_j. \quad (17)$$

The method used to estimate β_i is described in [2], and the estimates are shown in Table 1 Once we have estimates of β_i we may appropriately choose the M_i's to satisfy Equation 17. For the special case of $\beta_i = 0$ and all the M_i's equal we have log-warping.

5. COMPARISON OF FEATURES

In this section, we will compare the separability of vowels classes when improved scale-cepstral and mel-cepstral coefficients are used as features. We point out that when we refer to scale-cepstral coefficients as features, we are using the *magnitude* of $D_S[\frac{k_c C_p}{N}]$ in the feature vector. In comparing the separability afforded by the different cepstral features, a generalized F-ratio method is used [7, 8]. In deriving the F-ratio separability, let M_i and R_i denote the mean feature vector and sample covariance matrix , respectively, of the i^{th} phoneme class. Let $M_0 = \frac{1}{I}\sum_{i=1}^{I} M_i$, where I denotes the number of phoneme classes being compared. We then compute the within-class and between-class scatter matrices, S_w and S_b respectively as

$$S_w = \frac{1}{I}\sum_{i=1}^{I} R_i \quad \text{and} \quad S_b = \frac{1}{I}\sum_{i=1}^{I}(M_i - M_0)(M_i - M_0)^T. \quad (18)$$

The separability criterion is then given by

$$J = tr(S_w^{-1} S_b). \quad (19)$$

The data used in comparing the features consist of utterances of each vowel spoken by different speakers from dialect region 7 of the TIMIT training set data. /aa/,/ao/,/ae/,/ax/,/eh/,/er/,/ey/,/iy,/ih/ and /ow/ are the ten vowels that are considered for comparison of the different cepstra. Each utterance is so chosen that the corresponding phoneme is relatively stationary over at least 768 samples, and the middle 512 samples are used in the computation of the different cepstra. The improved scale-cepstral and mel-cepstral coefficients of clean and noisy utterances at 15 dB SNR are computed. The noisy utterance is simulated by adding artificially generated white Gaussian noise. The signal-to-noise (SNR) ratio is defined as the ratio of energy in the utterance to the noise energy.

The modified scale-cepstrum is computed using the following values of the parameters: $K = 128$, $L = 512$, $N = 256$ and $T_s = \frac{1}{16E3}$. The number of samples in each of the five frequency bands is given by $M_1 = 9, M_2 = 12, M_3 = 21, M_4 = 35, M_5 = 51$. The smoothed formant estimate, $s[n]$, is obtained using the method described in [9, 1]. Briefly, each frame of speech is segmented into Q overlapping subframes, and each subframe is hamming windowed. We have chosen the subframes to be 96 samples long, and the overlap between the subframes is 64 samples, resulting in 14 subframes. We estimate the sample autocorrelation function for each subframe and average over the available Q subframes. This averaged autocorrelation estimate is then hamming windowed and Fourier transformed to obtain an estimate of the formant-spectral envelope, $s[n]$.

The Mel-cepstrum is implemented using the program in the Signal Processing Information Base [10].

In all cepstra the zero-th coefficient is not used since this is roughly a measure of the spectral energy. Figure 1 and Figure 2 show the separability measure, J, as a function of the number of coefficients for clean and noisy speech. Note that in the scale-cepstrum proposed in [1], since the magnitude of scale-cepstrum was a function of the unknown scale-factor α (see Equation 9), we had to normalize to unit energy. The modified scale-cepstrum and the mel-cepstrum do not need such normalization. Figures 1 show the separability where all the three types of features have been normalized to unit energy, while Figures 2 show the separability when the features are *not* normalized to unit energy.

From the figures, it is clear that the improved scale-cepstrum provides better separability than the mel-cepstrum. Further, it is also seen to be robust to noise.

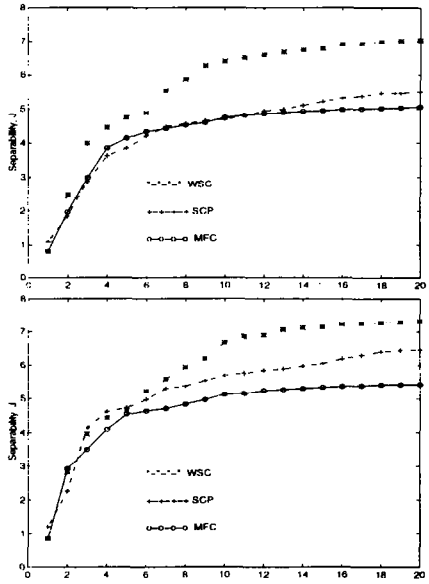

Figure 1: Separability between phonemes using scale (indicated by "+-+-+"), mel-cepstral coefficients (indicated by "o-o-o") and proposed improvements of scale-cepstrum (indicated by "*-*-*") for (a) clean and (b) noisy speech at 15 dB SNR. The feature vectors have been unit-normalized.

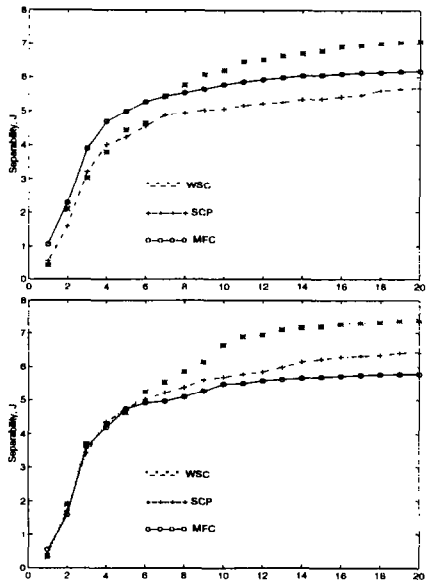

Figure 2: In this figure, we have shown the separability as in Figure 1, when the feature vectors have *not* been normalized. The data consist of 10 vowels spoken by different speakers and is taken from TIMIT database.

6. DISCUSSION

It is very interesting to note, as we have done before [11], that the various signal processing steps done to compute the improve scale-cepstrum is very similar to those used in computing mel-based features or auditory-model based features. From the simulation results, it is seen that the improved scale-cepstral features provide better separability than mel-cepstral features for vowels, and are robust to noise. The improved scale-cepstral features may, therefore, prove useful as acoustic features in speech processing.

7. REFERENCES

[1] S. Umesh, L. Cohen, N. Marinovic, and D. Nelson, "Scale Transform In Speech Analysis," *IEEE Trans. on Speech and Audio Processing*, 1996. Submitted April 1996, Revised June 1997.

[2] S. Umesh, L. Cohen, N. Marinovic, and D. Nelson, "Frequency-Warping in Speech," in *Proc. ICSLP'96*, (Philadelphia, USA), 1996.

[3] L. Cohen, N. Marinovic, S. Umesh, and D. Nelson, "Scale-Invariant Speech Analysis," in *Proc. SPIE*, (San Diego, USA), 1995.

[4] E. Eide and H. Gish, "A Parametric Approach to Vocal Tract Length Normalization," in *Proc. ICASSP'96*, (Atlanta), pp. 346–349, May 1996.

[5] T. Kamm, G. Andreou, and J. Cohen, "Vocal Tract Normalization in Speech Recognition: Compensating for Systematic Speaker Variability," in *Proc. of the 15th Annual Speech Research Symposium*, (Johns Hopkins University, Baltimore), pp. 175–178, June 1995.

[6] L. Cohen, "The Scale Representation," *IEEE Trans. Signal Proc.*, pp. 3275–3292, Dec. 1993.

[7] K. Fukunaga, *Introduction to Statistical Pattern Recognition.* San Diego: Academic Press, 1990.

[8] T. Parsons, *Voice and Speech Processing.* New York: McGraw Hill, 1987.

[9] D. Nelson, "Correlation Based Speech Formant Recovery," in *ICASSP'97*, (Munich), April 1997.

[10] D. H. Johnson and P. N. Shami, "The Signal Processing Information Base," *IEEE Signal Processing Magazine*, pp. 36–42, Oct. 1993.

[11] S. Umesh, L. Cohen, and D. Nelson, "Frequency-Warping and Speaker Normalization," in *Proc. IEEE ICASSP'97*, (Munich), April 1997.