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

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

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.,

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

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,

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

Note that as a result of log-warping, i.e. forming $X(e^f)$, the scale constant, $\alpha$, 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,

$$D_S(c) = \int (\log |X(e^j)|)^c e^j e^{-j2\pi cf} df,$$

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 $e^j$.

2. FREQUENCY-WARPING

In [1], the scale-factor $\alpha_{AB}$ 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.

\[ A(f) = B(\alpha^{(i)}_{AB}f) \quad f \in \text{ith band}, \] (6)

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

\[ \alpha^{(i)}_{AB} = \alpha^{(1+\beta_i)}_{AB} = \alpha_{AB} \cdot \alpha^{\beta_i}_{AB}. \] (7)

Note that \( \alpha_{AB} \) is a constant independent of \( i \) (the frequency band) and is dependent on the pair of speakers, while \( \beta_i \) 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, \( X_1(f) = A_1X(\alpha_1f) \) and \( X_2(f) = A_2X(\alpha_2f) \), then the respective Scale-Cepstrums are,

\[ D_{X1}(c) = \int_{-\infty}^{\infty} \log(A_1X(e^{\delta +\alpha_1}f)e^{j2\pi cf}e^{-j2\pi cf}df \]

\[ = \log(A_1)\delta(c) + \frac{e^{j2\pi c\log \alpha_1}}{\sqrt{\alpha_1}} D_X(c), \] (8)

and,

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

For \( c \neq 0 \) (i.e. for values other than scale-DC) the two scale-cepstrums differ only by the term \( \frac{e^{j2\pi c\log \alpha_1}}{\sqrt{\alpha_1}} \) that depends on the respective scale-factor \( \alpha_i \). 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

\[ D_X(c) = \int_{-\infty}^{\infty} \log |(X(e^{\delta})e^{\frac{f}{T}})| e^{-j2\pi cf}df, \] (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:

\[ D_{X1}(c) = \log\left(\frac{A_1}{\sqrt{\alpha_1}}\right)\delta(c) + e^{j2\pi c\log \alpha_1} D_X(c) \] (11)

\[ D_{X2}(c) = \log\left(\frac{A_2}{\sqrt{\alpha_2}}\right)\delta(c) + e^{j2\pi c\log \alpha_2} D_X(c) \] (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.

\[ D_X[k\frac{k_cC_p}{N}] = \sum_{m=0}^{K-1} \log(|S(e^{m\Delta\nu+\ln(100)})|e^{m\Delta\nu+\ln(100)}) \times e^{-j2\pi \frac{k_cC_p}{N}m} \]

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

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

\[ S(e^{m\Delta\nu+\ln(100)}) = \sum_{n=0}^{L-1} s[n]e^{-j2\pi e^{im\Delta\nu+\ln(100)}nT_s}, \] (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 piece-wise 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 \( M_t \) equally spaced
<table>
<thead>
<tr>
<th>Band 1</th>
<th>Band 2</th>
<th>Band 3</th>
<th>Band 4</th>
<th>Band 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.0</td>
<td>3.3869</td>
<td>1.4629</td>
<td>0.4616</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Table 1: The $\beta_i$ are estimated as described in [2].

intervals in the region $\log(L_1)$ to $\log(U_i)$. The sampling period is therefore

$$\Delta \nu_i = \frac{\log(U_i) - \log(L_1)}{(1 + \beta_i)M_i}.$$  \hspace{1cm} (15)

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

$$\log(U_i) - \log(L_i) = \log(U_j) - \log(L_j).$$  \hspace{1cm} (16)

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

$$(1 + \beta_k)M_k = (1 + \beta_j)M_j.$$  \hspace{1cm} (17)

The method used to estimate $\beta_i$ is described in [2], and the estimates are shown in Table 1. Once we have estimates of $\beta_i$, 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 vowel 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 $D_s[n]$ 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

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

The separability criterion is then given by

$$J = \text{tr}(S_w^{-1}S_b).$$  \hspace{1cm} (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/ /aw/ /ch/ /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 $T_s = \frac{1}{16888}$. 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, 11]. 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 $\alpha$ (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.
6. DISCUSSION

It is very interesting to note, as we have done before [11], that the various signal processing steps done to compute the improved 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


