FREQUENCY-WARPING AND SPEAKER-NORMALIZATION

S. Umesh *L. Cohen* *D. Nelson*

Dept. of Electrical Engg. Hunter College of CUNY U.S. Dept. of Defense
I.I.T., Kanpur-208016, India New York, NY 10021, USA Ft. Meade, MD 20755, USA

ABSTRACT

Recently, we have proposed the use of scale-cepstral coefficients as features in speech recognition [1]. We have developed a corresponding frequency-warping function, such that, in the warped domain the formant envelopes of different speakers are approximately translated versions of one and another for any given vowel. These methods were motivated by a desire to achieve speaker-normalization. In this paper, we point out to very interesting parallels of the various steps in computing the scale-cepstrum, with those observed in computing features based on physiological models of the auditory system or psychoacoustic experiments. It may therefore be useful to have a better understanding of the need for the various signal-processing steps which may result in the development of more robust recognizers.

1. INTRODUCTION

Recently, we have proposed the use of scale-cepstral coefficients as features in speech recognition, [1], since preliminary results indicate that they provide better separation between the vowels than mel-cepstral coefficients. The scale-cepstrum was motivated by a desire to achieve speaker-normalization, and therefore reduce inter-speaker scatter. We assume that to a first order approximation the main source of the inter-speaker scatter is from the differences in the vocal tract length. Hence, the formant envelopes between two speakers, $F_A(f)$ and $F_B(f)$ may be assumed to be scaled versions of each other, i.e.

$$F_A(f) \approx F_B(\alpha_{AB} f). \tag{1}$$

Many conventional speaker-normalization techniques attempt to estimate the scale factor, α_{AB}, and compensate for it [2, 3]. The Scale Transform [4] is a useful tool to apply in this situation, since the magnitude of the Scale Transform (ST) in invariant to scaling, much like how the magnitude of the Fourier Transform is invariant to translation in the time-domain. It can be easily shown that the Scale Transform of a function $F_A(f)$ is the Fourier Transform of the related function $F_A(e^f)e^{\frac{f}{2}}$, i.e.

$$D_A(c) = S.T.\{F_A(f)\} = \int_{-\infty}^{\infty} F_A(e^f)e^{\frac{f}{2}}e^{-j2\pi fc}df, \tag{2}$$

where c is the scale parameter. Note that in the warped domain, $F_A(f)$ and $F_B(f)$ are translated versions of each other, i.e.

$$\begin{aligned} F_A'(f) &= F_A(e^f) = F_B(\alpha_{AB}e^f) \\ &= F_B(e^{f+\log\alpha_{AB}}) = F_B'(f+\log\alpha_{AB}). \end{aligned} \tag{3}$$

In the next section, we will discuss how our experiments indicate that the frequency-warping is not an exact log-warping, but that the desired warping is very similar to mel-scale obtained from psycho acoustic experiments and to Georg Von Békésy's experimental results relating the position of the maximum vibration on the cochlear partition to the actual frequency in Hertz.

2. FREQUENCY-WARPING FUNCTION

As we have discussed in the previous section, if the scaling factor α_{AB}, is a constant, independent of frequency, f, then in the log-warped domain, the formant envelopes are translated versions of one and another. However, our experiments [5] on actual speech data obtained from the TIMIT database indicate that the scaling constant α_{AB} between any two speakers is *not* a constant but is a function of frequency, f. We have then written this frequency dependent scaling function as $\alpha_{AB}(f) = \alpha_{AB}^{(1+\beta(f))} = \alpha_{AB}\alpha_{AB}^{\beta(f)}$. This is done to separate it into a purely frequency dependent function, $\beta(f)$, which is independent of the pair of speakers, A and B, and α_{AB}, a constant that just depends on the pair of speakers but is independent of frequency. The interested reader is referred to our previous work [5], where we have obtained a piece-wise approximation of $\beta(f)$. Interestingly, the frequency-warping obtained is very similar to the frequency-scales from psycho acoustics and auditory models.

Work supported by the HBCU/MI program and the PSC-CUNY Research Award Program.

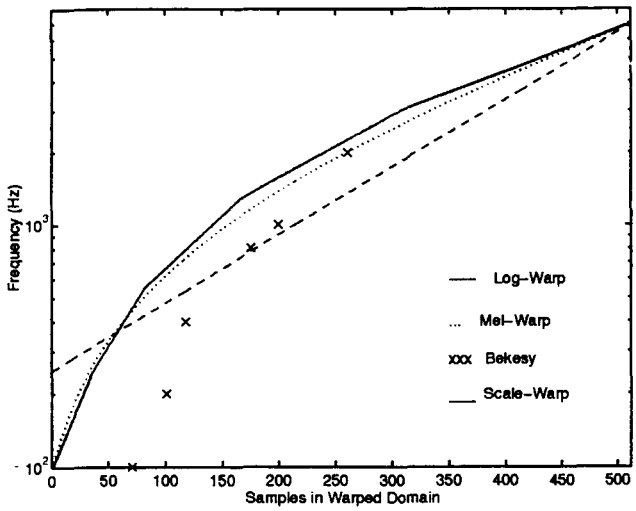

Figure 1: The different warping schemes are compared. Note that the experimental piece-wise scale-warped function, the mel frequency-scale and the mapping of frequency to maximum response on basilar membrane obtained by Von Békésy are all very similar. Also shown is the log-warping used by Shamma *et al* for the computation of the auditory spectrum.

One of the popular front-ends for speech recognition is based on mel-scale, ever since experiments by Davis and Mermelstein [6] indicated better recognition rates when compared to conventional methods of analysis. The mel-scale is obtained from psycho acoustic experiments and relates the real frequency (in Hertz) to that perceived by the ear. Physiological models of the auditory system also indicate a mapping of the real frequency to the maximum response position on the basilar membrane that approximately follows the mel-scale. Much of the current knowledge of the acoustic-mechanical properties of the human basilar membrane is due almost exclusively to the the efforts of Georg Von Békésy [7]. In his experiments on human ears, the stapes were vibrated sinusoidally with constant amplitude and the displacement of various points along the cochlear partition were examined under the microscope as a function of the real frequency. The plot of the distance from stapes to the point of maximum displacement versus frequency, has the general form of the mel-scale and our proposed frequency-warping (scale-warping) as shown in the Figure 1.

3. SCALE-CEPSTRUM

We will now briefly describe the important steps involved in the computation of the scale-cepstrum. In the next section, we will compare each of these steps with those involved in the computation of mel-cepstrum [6]and auditory spectrum [8].

To obtain an estimate of the formant envelope, we use a variation of the averaged periodogram technique [9], proposed by Nelson. Each frame of speech corresponding to 512 samples (sampled at 16 KHz) is segmented into 14 overlapping sub-frames and each sub-frame is hamming windowed. The sub-frames are chosen to be 96 samples long and the overlap between sub-frames is 64 samples. We estimate the sample autocorrelation function for each sub-frame and average over the available 14 sub-frames. This averaged autocorrelation estimate is then hamming windowed and is denoted as $s[l]$. It may easily be shown that $S(f)$, the Fourier Transform of $s[l]$, is the convolution in frequency of the signal power spectral density with a smoothing function $W(f)$.

We approximate the frequency warping function by a set of five piecewise log-warped frequency bands. The five logarithmically equal bands are [100,240) Hz, [240, 550) Hz, [550,1280) Hz, [1280,3000) Hz and [3000, 7000) Hz. Let L_i and U_i be the lower and upper frequency of the i^{th} frequency band, then the frequency-spacing of each sample in that band is given by,

$$\Delta\nu_i = \frac{\log(U_i) - \log(L_i)}{M_i}. \tag{4}$$

The number of samples, M_i, for each of the five frequency bands are $M1 = 8, M2 = 12, M3 = 21, M4 = 35, M5 = 52$. Note that if all the M_i's are equal, then we get log-warping. The frequency warped spectral samples in each of the five frequency bands can be easily computed from the smoothed autocorrelation estimate, $s[l]$, i.e.

$$S(e^{m_i\Delta\nu_i + ln(L_i)}) = \sum_{l=0}^{L-1} s[l]\exp(-j2\pi e^{[m_i\Delta\nu_i + ln(L_i)]lT_s})$$
$$m_i = 0, 1, 2, \cdots, M_i - 1. \tag{5}$$

The corresponding frequency dependent scaling for each such sample is given by $E_{m_i} = \exp(\frac{m_i\Delta\nu_i + \ln(L_i)}{2})$. The frequency-warped spectral sample and the corresponding scaling factor from each of the frequency band can be concatenated to obtain $M = M_1 + M_2 + M_3 + M_4 + M_5$ samples. The Scale Transform can then be computed using the Fast Fourier Transform (FFT) as

$$D_s[k_c] = \sum_{m=0}^{M-1} (\log|S_m|E_m)e^{-j2\pi\frac{k_c}{K}m}$$
$$k_c = 0, 1, 2, \cdots, K - 1. \tag{6}$$

The values of the various parameters are: $M = 128$, $L = (96x2)$, $K = 256$ and $T_s = \frac{1}{16E3}$. The magnitude

of $D_s[k_c]$, are the scale-cepstral coefficients and these may be used as features in speech recognition.

4. SCALE-CEPSTRUM VERSUS OTHER ACOUSTIC FEATURES

Our main motivation for developing the scale-cepstrum was to have speaker-normalized features as input to speech recognizers in order to improve their robustness to inter-speaker variations. A block diagram of the main steps in computing the scale-cepstrum is shown in Figure 2, and the reader is referred to our paper [1] for details regarding implementation. We will discuss each stage of the computation of the scale-cepstrum and compare with remarkably similar steps in other methods based on psycho acoustics and on auditory models.

4.1. Frequency-Warped Spectrum

We first compute the warped spectrum of speech where the warping function is discussed in Section 3. The justification for using the warping function is that in the warped domain the spectral envelopes between two speakers are essentially translated versions of one and another. In auditory models also the frequency analysis is done on a similar warped scale since the mapping of place along basilar membrane onto real frequency is approximately logarithmic. The mel-cepstrum which is motivated by the mel-scale obtained from psycho acoustic experiments also has a similar warping along the frequency axis. Thus, all the methods compute a warped spectrum, and the warping functions are similar as seen from Figure 1.

4.2. Filter-bank Analysis

Since our intended application was in speech recognition, we were essentially interested in extracting the spectral envelope information. Therefore, we have a preliminary smoothing stage where the effects of pitch are essentially smoothed out. This is necessary since in the scale-warped frequency domain, the pitch harmonics are no longer periodic. Hence, the transform into the scale-cepstral domain will contain these pitch components which will interfere with the spectral envelope components. This smoothing operation is a variation of the averaged periodogram spectral analysis method [9]. The smoothed spectrum can therefore be interpreted as the output of a filter bank that are made up of constant bandwidth filters. Similarly mel-cepstrum analysis use filters that are constant-Q above 1000 Hz and are constant bandwidth below 1000 Hz, since better recognition rates have been obtained using such a filter bank. Auditory models use filters that are constant-Q, since responses of successive points on the basilar membrane are roughly constant-Q in nature. Note that at low frequencies, since these filters are narrow-band the fine spectral detail is also passed through. However, in Wang and Shamma's [8] "linear" auditory spectrum, the output is that of the narrow-band differential cochlear filter, and hence it can be essentially thought of as the frequency-warped spectrum of the original signal. In other words the auditory spectrum still retains information about the fine spectral detail also.

4.3. Scaling of Filter-Bank output

As seen from Equation 2, the computation of scale-cepstrum require the multiplication of the scale-warped spectrum by a frequency-dependent but a signal independent scaling factor $e^{\frac{f}{2}}$. The physical significance of this high frequency emphasis is still not clear to us. Although, in the mel-cepstrum, there is no such scaling, the higher bandwidths at higher frequencies can be thought of as providing the high frequency emphasis. In the auditory spectrum of Wang and Shamma, there is a normalization factor that is neither uniform nor predetermined but is driven by the energy distribution of the signal.

4.4. Second Fourier Transform

In our scale-cepstral analysis, the formant envelopes between different speakers are translated versions of one and another, in the frequency-warped domain. Hence, we need to compute a second Fourier Transform and calculate its magnitude to extract the speaker independent information. (Note that the phase term carries information about the translation factor and hence about the relative scaling between the speakers.) Recall, that these steps are essentially equivalent to computing the magnitude of the Scale-Transform. Hence, at the end of the second Fourier transform we are in the scale-variable domain. Interestingly, Wang and Shamma have recently proposed a model for spectral shape analysis in the central auditory system [10] that is equivalent to computing the second Fourier-like Transform. They refer to this as ripple analysis and the ripple frequency (measured in cycles/octave) is same as our scale-variable, c. In Mel-cepstral analysis, the second Fourier-transform is motivated by a similar step in the original cepstral analysis, which is done to separate the slowly-varying formant information from the pitch information.

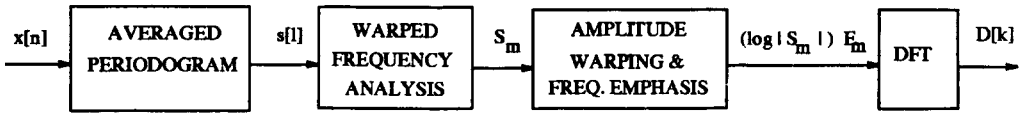

Figure 2: The block diagram shows the important steps in the computation of our proposed scale-cepstrum. A brief description of the important mathematical steps is given in Section 3. In Section 4 we compare these individual steps with similar ones used in the computation of the mel-cepstrum and the auditory spectrum. It is interesting to note that while the mathematical steps are similar the motivation in developing each of these representations are very different as discussed in Section 5.

5. DISCUSSION

It is very interesting to note from Figure 1 that the proposed frequency-warping function is quite similar to the mel-frequency scale and the auditory-model based frequency scales. Similarly, the subsequent signal processing steps done to compute the scale-cepstrum are very similar to those operations done in computing the mel-cepstrum and the auditory spectrum. However, the one important difference is that while for the scale-cepstrum we understand the necessity of each of these steps, in the other methods the various steps were rather heuristically derived in an attempt to either obtain better recognition accuracy or to mimic the ear as closely as possible. Although the signal processing steps in each of these methods are similar, the motivations for each of these methods are very different. The scale-cepstrum was designed with a desire to remove speaker-specific information to enable more robust recognition, while the mel-cepstrum is based on the psycho acoustic based mel-scale and has received a lot of attention recently because of the reported improvements in recognition accuracy [6, 11]. The auditory spectrum is motivated by a desire to mimic the auditory system with the hope that these will lead to new representations of sound signals which might prove useful in speech recognition applications. That all these methods have similar signal processing steps leads us to believe that a more in depth analysis of the various signal processing steps discussed above, would provide us with a better understanding of speech production, auditory physiology and psychoacoustic behavior, and may help us to bring the three into harmony.

6. 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 in April.

[2] H. Wakita, "Normalization of vowels by vocaltract length and its application to vowel identification," *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. ASSP-25, pp. 183–192, April 1977.

[3] T. M. Nearey, "Phonetic feature systems for vowels," tech. rep., Indiana Univ. Linguistics Club, Dec. 1978.

[4] L. Cohen, "The scale representation," *IEEE Trans. Sig. Proc.*, vol. ASSP-41, pp. 3275–3292, Dec. 1993.

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

[6] S. B. Davis and P. Mermelestein, "Comparison of parametric representation for monosyllabic word recognition in continiuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 357–366, Aug. 1980.

[7] G. v. Békésy and W. A. Rosenblith in *Handbook of Experimental Psychology* (S. S. Stevens, ed.), (New York), pp. 985–1039, John Wiley, 1951.

[8] K. Wang and S. Shamma, "S-elf Normalization and Noise-Robustness in Early Auditory Representations," *IEEE Trans. Speech and Audio Proc.*, vol. 2, pp. 421–335, July 1994.

[9] A. H. Nuttall and G. C. Carter, "Spectral Estimation using Combined Time and Lag Weighting," *Proc. of the IEEE*, vol. 70, pp. 1115–1125, Sept. 1982.

[10] K. Wang and S. A. Shamma, "Spectral Shape Analysis in Central Auditory System," *IEEE Trans. Speech and Audio Proc.*, vol. 3, pp. 382–394, September 1995.

[11] C. R. Jankowski, H.-D. H. Vo, and R. P. Lippmann, "A Comparison of Signal Processing Front Ends for Automatic Word Recognition," *IEEE Trans. Speech Audio Proc.*, pp. 286–293, July 1995.