Frequency-Warping in Speech

S. Umesh, L. Cohen and N. Marinovic
Hunter College of CUNY
695 Park Ave., NY 10021, USA

D. Nelson
Department of Defense
Ft. Meade, MD 20755, USA

ABSTRACT

In this paper we present results that indicate that the formant frequencies between different speakers scale differently at different frequencies. Based on our experiments on speech data, we then numerically compute a universal frequency-warping function, to make the scale-factor independent of frequency in the warped domain. The proposed warping function is found to be similar to the mel-scale, which has previously been derived from purely psycho-acoustic experiments. The motivation for the present experiments stems from our recently proposed use of scale-transform based cepstral coefficients [6] as acoustic features, since they provide superior separability of vowels than mel-cepstral coefficients.

1. Introduction

Recently, we have proposed the use of scale-transform based cepstral coefficients as acoustic features in speech recognition [6]. The scale-transform based cepstrum is motivated by speaker-normalization techniques. Such normalization techniques are necessary, since different speakers have different formant frequencies for the same vowel. One of the procedures for normalization is based on the assumption that the formant values of any given speaker are approximately a multiplicative scale factor times the formant values of any other speaker for a given vowel [3, 4]. Speaker normalization procedures, therefore, estimate the scale-factor and attempt to normalize the formant frequencies among different speakers. For example, Wakita [7] normalized the formant frequencies based on estimates of vocal tract length, assuming that the main source of interspeaker scatter is the scale of vocal tract organs. However, as we will show next, the scale-transform (see Cohen [1]) provides a useful tool to achieve speaker normalization *without* explicitly computing the speaker-specific scaling constant.

Briefly, the scale-transform of a function, $X(f)$, is given by,

$$D_X(c) = \int_0^\infty X(f) \frac{e^{-j2\pi c \ln f}}{\sqrt{f}} df, \qquad (1)$$

and inversely

$$X(f) = \int_{-\infty}^{\infty} D_X(c) \frac{e^{j2\pi c \ln f}}{\sqrt{f}} dc \qquad f \geq 0. \qquad (2)$$

One of the most important 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, since

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

The scale-factor $\sqrt{\alpha}$ is necessary to preserve the total energy of the signal. Note that the scaling constant α is a part of the phase expression and does not appear in the magnitude of the scale transform. Hence, we have $|D_X^\alpha(c)| = |D_X(c)|$. With respect to speaker-normalization, if we were to compute the magnitude of the scale-transform of the formant envelope, then all speaker-dependent scaling constants that appear in the phase term would be removed. Unlike the conventional speaker-normalization techniques, we do not have to explicitly compute the speaker-specific scaling constant.

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

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

Note that as a result of log-warping, i.e. forming $X(e^f)$, the speaker-specific scale constant, α, is purely a function of the translation parameter in the log-warped domain. This may be easily seen by considering

$$X_1(f) = X(e^f) \qquad (5)$$

$$X_2(f) = X(\alpha e^f) = X(e^{f+\log\alpha}) = X_1(f + \log\alpha) \qquad (6)$$

Therefore, if we have two formant envelopes that are related by a pure scaling constant, that is independent of frequency but is dependent on the pair of speakers, then in the log-warped domain, the envelopes are the same except for a translation factor dependent on α.

2. Frequency-Warping Function

In this paper, based on our experiments on speech data obtained from the TIMIT database, we provide evidence that the scale-factor is not independent of frequency. In such a case, the log-warping discussed in the previous section may not be the appropriate warping function. The focus of the paper, is therefore, to find a more appropriate warping function to account for this frequency-dependency of the scale-factor.

Our ultimate goal is to find a suitable warping function, such that in the warped domain the formant envelopes between different speakers are approximately translated versions of one and another for a given vowel. The motivation for designing such a warping function, is to separate the effects of speaker-independent features from speaker-specific features which would be reflected only in the translation parameter. If we obtain such a warping function, it would prove useful in the solution of two important problems. First, if the warped formants of different speakers are essentially translated versions of one and another, then the magnitude of their Fourier transforms would essentially be similar and may therefore prove useful as acoustic features in *speaker-independent recognition*. Secondly, since the warped formant envelopes of different speakers differ by translation factors, the translation factors which carry speaker-specific information may be useful in *speaker identification and verification*.

Numerous researchers have previously attempted to relate the formants between speakers by finding a suitable warping function (e.g. Neuburg [5]), or as a special case computing a uniform scaling constant between two speakers (e.g. Miller, Nearey [3, 4]). Our approach differs from the others in that we are looking for a universal warping function (independent of speakers), while the differences in the speakers manifest themselves as differences in translation factor in the warped domain. In the next section, we will obtain such a warping function based on experiments on speech obtained from the TIMIT database.

3. Piece-wise Approximation of Warping Function.

For speech obtained from TIMIT database (sampled at 16 KHz), the frequency region of interest is between 100 Hz and 7000 Hz. We divide this into logarithmically equal bands of [100,240) Hz, [240,550) Hz, [550,1280) Hz, [1280,3000) Hz and [3000,7000) Hz. We assume that in each such frequency band, the formant envelope of any two speakers are scaled versions of each other. (Although the ear may be processing speech with a smooth warping function, for convenience, we assume a piece-wise log-warped function. Such a piece-wise approximation may provide us with sufficient insight in determining the true functional form of the warping function.) While the scaling factor is assumed to be constant within each frequency band, its value may vary across the different frequency bands. In other words, the formant envelopes of two speakers for the same utterance are assumed to be of the form

$$A(f) = B(\alpha_{AB}^{(i)} f) \quad f \ \epsilon \ i^{th} \text{ band}, \tag{7}$$

and $i = 1, 2, \cdots, 5$. Note that $\alpha_{AB}^{(i)}$ may be different for each of the five frequency bands. Further we will rewrite Equation 7 as

$$A(f) = B(\alpha_{AB}^{(1+\beta_i)} f) \quad f \ \epsilon \ i^{th} \text{ band}, \tag{8}$$

where

$$\alpha_{AB}^{(i)} = \alpha_{AB}^{(1+\beta_i)} = \alpha_{AB} \cdot \alpha_{AB}^{\beta_i}. \tag{9}$$

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 advantage of such a formulation is that we have separated the speaker-specific characteristics from the warping function that is purely determined by β_i. The β_i are assumed to be constant over the i^{th} frequency band. Therefore, over any given frequency band, i, the formant envelopes of any two speakers are assumed to be scaled versions of each other.

4. Discrete-Implementation of Warping Function.

We, briefly discuss the discrete implementation of the piece-wise warping function. We need to compute $B(e^f)$ for $e^f \epsilon [U_i, L_i]$, where U_i and L_i are upper and lower frequency limits of the i^{th} frequency band. We discretize by computing $B(e^f)$ at M_i equally spaced intervals in the region $\log(L_i)$ to $\log(U_i)$. Let,

$$\triangle\nu_i = \frac{\log(U_i) - \log(L_i)}{M_i}. \tag{10}$$

Then, the exponentially spaced samples in the i^{th} frequency region are $B(e^{m_i \triangle\nu_i + \log(L_i)})$ for $m_i = 0, 1, \cdots, (M_i - 1)$. Let $A(f) = B(\alpha_{AB}^{(i)} f)$ be a scaled version of $B(f)$, where $\alpha_{AB}^{(i)}$ is the scaling factor in the i^{th} frequency band. If we exponentially sample $A(f)$ at $\triangle\nu_i$ spacing in the i^{th} frequency band we have,

$$\begin{aligned} A(e^{m_i \triangle\nu_i + \log(L_i)}) &= B(e^{m_i \triangle\nu_i + \log(\alpha_{AB}^{(i)}) + \log(L_i)}) \\ & \quad m_i = 0, 1, \cdots, (M_i - 1) \end{aligned} \tag{11}$$

We can rewrite Equation 11 as,

$$A(e^{m_i \triangle\nu_i + \log(L_i)}) = B(e^{(m_i + \frac{\log(\alpha_{AB}^{(i)})}{\triangle\nu_i}) \triangle\nu_i + \log(L_i)}). \tag{12}$$

Hence,

$$A[m_i] = B[m_i + \frac{\log(\alpha_{AB}^{(i)})}{\triangle\nu_i}], \tag{13}$$

and is translated by $\frac{\log(\alpha_{AB}^{(i)})}{\triangle\nu_i}$ samples. From Equation 9, we have

$$\frac{\log(\alpha_{AB}^{(i)})}{\triangle\nu_i} = \frac{\log(\alpha_{AB}^{1+\beta_i})}{\triangle\nu_i} = (1 + \beta_i) \frac{\log(\alpha_{AB})}{\triangle\nu_i}. \tag{14}$$

If,

$$\frac{\triangle\nu_k}{1+\beta_k} = \frac{\triangle\nu_j}{1+\beta_j} = \triangle\nu', \tag{15}$$

then, we have equal translation of $\frac{\log(\alpha_{AB})}{\triangle\nu'}$ in all frequency bands. Recall, that we have chosen frequency bands that are equally spaced on the logarithm scale, hence, we have

$$\log(\frac{U_k}{L_k}) = \log(\frac{U_j}{L_j}). \tag{16}$$

Hence, Equation 15 will be satisfied if,

$$(1+\beta_k)M_k \;=\; (1+\beta_j)M_j. \tag{17}$$

Once we have estimated all β_i (which is described in the next section), we may appropriately choose the M_i's to satisfy Equation 17. Therefore, we use different sampling rates in different frequency bands to achieve warping, and the resulting sequences are translated versions of each other. For the special case of $\beta_i = 0$ and all the M_i's equal we have log-warping.

5. Experimental Calculation of Warping Parameters

In this section, we will describe our experiments to estimate the parameters $\alpha_{AB}^{(i)}$ and β_i. The speech data consisted of `sa1` and `sa2` sentences spoken by two male and two female speakers from dialect region 7 of the TIMIT training set data. The four speakers considered were *fksr0, fvkb0, mtlc0,* and *mgsl0.* In our experiments, we chose vowels that were relatively stationary over 768 samples and the middle 256 samples were used in the computation of the formant envelopes. The procedure used to compute the formant envelopes are described in detail in [6]. Briefly, each frame of speech (consisting of 256 samples) 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 nine subframes. We estimate the sample autocorrelation function for each subframe and average over the available $Q(=9)$ subframes. This averaged autocorrelation estimate is then hamming windowed and Fourier transformed to obtain an estimate of the formant-spectral envelope.

5.1. Estimation of $\alpha_{AB}^{(i)}$

We consider two speakers at a time, and consider all vowels that are spoken in the same context in sentences `sa1` and `sa2`. We visually inspect the formant envelope of each such context-dependent vowel of the two speakers. The peaks in the formant envelopes are attributed to the individual formant frequencies. We consider only those pair of formants that lie within the same frequency band (preferably towards the middle of the band), and compute the scaling factor as $\alpha_{AB}^{(i)} = \frac{F_A^{(i)}}{F_B^{(i)}}$. $F_A^{(i)}$ corresponds to the formant of speaker A that lies in the frequency band i. We compute $\alpha_{AB}^{(i)}$ for different frequency bands and different vowels. Table 1 shows the average scaling parameter $\alpha_{AB}^{(i)}$ for each of the frequency band for a given pair of speakers. From the Table, it is seen that the scale factor is not constant across different bands. The trend is to decrease with increasing frequency.

Speakers	*Band 2*	*Band 3*	*Band 4*	*Band 5*
ksr-vkb	1.1564	1.0944	1.0717	1.0509
ksr-tlc	1.6022	1.3991	1.2847	1.2552
ksr-gsl	1.2960	1.3431	1.2134	1.1642
vkb-tlc	1.4301	1.2123	1.1404	1.0842
vkb-gsl	1.3143	1.2767	1.0979	1.0611
tlc-gsl	0.7954	.9363	0.9618	0.9786

Table 1: The estimates of $\alpha_{AB}^{(i)}$ for different pair of speakers. Note that these are not constant over different frequency bands. The trend is to have larger compression/dilation at lower frequencies.

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

Table 2: The β_i are computed from the estimates of $\alpha_{AB}^{(i)}$ using Equation 18. The estimates shown here are the average values computed over each frequency band. Since, we had no data for Band 1 we have assumed $\beta_1 = 5$.These estimates provide us with sufficient insight into the nature of the warping function as seen from Figure 1.

5.2. Estimation of β_i

We estimate β_i from the estimates of $\alpha_{AB}^{(i)}$. We make the simplifying assumption that $\beta_5 = 0$. Therefore, $\alpha_{AB}^{(5)} = \alpha_{AB}^{1+0} = \alpha_{AB}$, is purely dependent on the pair of speakers. This is a reasonable assumption, since the fifth frequency band corresponds to the higher order formants and these are mostly affected by the length of the pharyngeal-oral tract of the speaker. The other β_i can be computed as,

$$\beta_i = \frac{\log(\alpha_{AB}^{(i)})}{\log(\alpha_{AB}^{(5)})} - 1 \quad for\ i = 1, 2, 3 \text{ and } 4. \tag{18}$$

We average the β_i's obtained from all pairs of speakers. Note that the β_i show considerable variation due to the continuous nature of speech and due to the lack of a large sample of data because of the manual method of estimating formants. However, the available estimates provide us with sufficient insight into the nature of the warping function.

Figure 1 compares the proposed warping function with the mel-warping and log-warping. The proposed scale-warping is plotted by concatenating the samples in the different frequency bands, with $M_1 = 34, M_2 = 47, M_3 = 84, M_4 = 141, M_5 = 206$. We experimented with a number of closed-

form expressions to approximate the proposed warping. We found the following expression to match the experimental warping, i.e.

$$F_{formula} = \frac{1500}{\log(2)} \log(\frac{F_{Hz}}{1500} + 1). \qquad (19)$$

The proposed formula is similar to the Technical-Mel [2] except that the warped frequency is made to match the real frequency at 1500 Hz instead of 1000 Hz. Further studies are required to understand the significance of such a warping function.

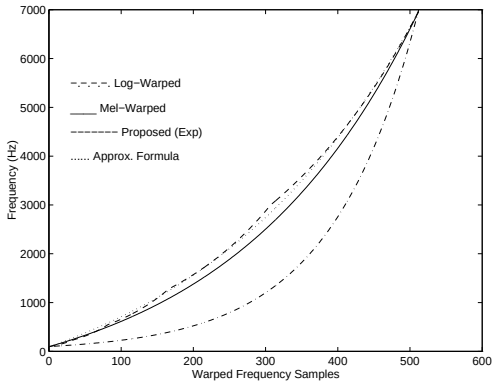

Figure 1: The different warping schemes are compared. Note that the experimental piece-wise log-warped function and the warping function given by the formula in Equation 19 match very well.

6. Discussion

It is very interesting to note from Figure 1 that the proposed warping function is quite similar to the mel-warping. The mel-scale is obtained from psycho-acoustic experiments done on the human auditory system. The mel-warping relates the real frequency scale (Hz) to the perceived frequency (mels) by the human ear. The proposed warping on the other hand, was obtained from actual speech data in an attempt to separate the speaker-specific characteristics and the speaker-independent features in speech. That the two warping functions are similar suggests that the ear may also be using the mel-scale to achieve this separation between speaker-specific and speaker-independent features. This additional insight may provide us with important clues to understand the nature of speech and the way it is processed by the human ear. We conjecture that the signal processing operation done in the ear may be similar to our recently proposed scale-cepstrum, with the log-warping being replaced by the proposed piece-wise log-warping. Further experiments will be performed to examine this conjecture.

Finally, we present an example, where we use the proposed frequency-warping on the formant envelopes of the vowel /ae/ in the word "that" for 4 speakers. The frequency-warped envelopes are shown in Figure 2(a). Figure 2(b) shows the envelopes after they have been aligned by suitable translation. Note the excellent alignment of the frequency-warped envelopes for the 4 different speakers. This indicates that using the proposed frequency-warping, the formant envelopes of different speakers are essentially translated versions of each other.

Acknowledgement: This work was supported by the HBCU/MI Program.

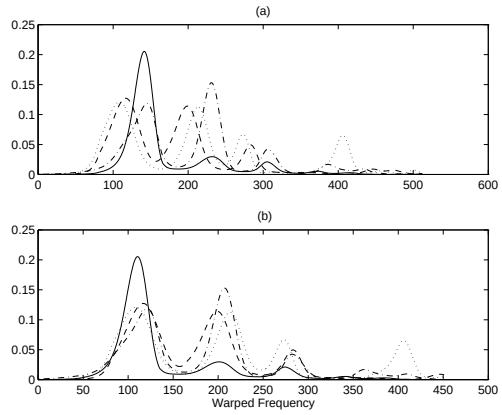

Figure 2: The frequency-warped formant envelopes of the vowel /ae/ in the word "that" for 4 speakers. In (b) the formant envelopes have been appropriately translated so that they approximately align with each other. Note that the alignment is excellent, indicating that in the frequency-warped domain the formant envelopes are essentially translated versions of one and another.

7. REFERENCES

1. L. Cohen. The scale representation. *IEEE Trans. Signal Processing*, ASSP-41:3275–3292, Dec. 1993.
2. C. G. M. Fant. Acoustic description and classification of phonetic units. In *Ericcson Technics*, number 1, 1959. Reprinted in *Speech Sound and Features*, MIT Press, Cambridge, 1973.
3. James D. Miller. Auditory-Perceptual Interpretation of the Vowel. *J. of Acoust. Soc. Am.*, 85(5):2114–2134, May 1989.
4. T. M. Nearey. Phonetic feature systems for vowels. Technical report, Indiana Univ. Linguistics Club, Dec. 1978.
5. E. P. Neuburg. Frequency axis warping to improve automatic word recognition. In *Proc. IEEE ICASSP'80*, pages 166–168, 1980.
6. S. Umesh, L. Cohen, N. Marinovic, and D. Nelson. Scale transform in speech analysis. *IEEE Trans. on Speech and Audio Processing*, 1996. Submitted.
7. H. Wakita. Normalization of vowels by vocal-tract length and its application to vowel identification. *IEEE Trans. Acoust., Speech, Signal Proc.*, ASSP-25(2):183–192, April 1977.