# EXPLOITING FREQUENCY-SCALING INVARIANCE PROPERTIES OF THE SCALE TRANSFORM FOR AUTOMATIC SPEECH RECOGNITION

*S. Umesh*[1] *Richard C. Rose*[2] *S. Parthasarathy*[2]

[1]Indian Institute of Technology, Kanpur, INDIA *
[2]AT&T Labs–Research, Florham Park, NJ, USA



6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

## ABSTRACT

An experimental study of the application of scale-transform to improve the performance of speaker independent continuous speech recognition, is presented in this paper. Three major results are described. First, a comparison was made between the scale-transform based magnitude cepstrum coefficients (STCC) and mel-scale filter bank cepstrum coefficients (MFCC) on a telephone based connected digit recognition task. It was shown that the STCC can obtain a performance that is close to that of the MFCC. Second, a simple frequency-normalization procedure was applied to the scale-transform representation that improved performance on the connected digit recognition task with respect to the MFCC. Finally, in a more controlled experimental setting using the TIMIT database, it was shown that the application of phone-specific frequency warpings improved phone classification performance over using a single speaker-specific warping. This last result may have general implications for all frequency warping based speaker normalization procedures.

## 1. INTRODUCTION

The scale transform (ST) is based on the Fourier transform of a frequency–warped version of the speech signal [1, 2]. One of the distinguishing characteristics of the scale transform representation is the insensitivity of the scale transform magnitude to linear warpings in the frequency domain [1, 2]. It has been suggested that this property has the potential to improve the speaker robustness of speaker independent automatic speech recognition (ASR) systems. This arises from a commonly held assumption that physiological differences among speakers can be approximated by a linear warping of the frequency axis. The purpose of this paper is to demonstrate the extent to which this property can have practical implications for speaker independent automatic speech recognition (ASR). An experimental study is presented applying the ST spectral representation to several automatic speech recognition tasks.

An introduction to the scale transform along with a discussion of its properties and a review of variants of the scale transform developed in previous work is given in Section 2. A description of a scale transform based cepstrum coefficient (STCC) feature representation and its use in a continuous density hidden Markov model (HMM) based ASR system is given in Section 3. A performance comparison between the STCC and mel–frequency filter–bank based cepstrum coefficient (MFCC) is described for a telephone based connected digit recognition task. In Section 4, a low complexity frequency normalization procedure that can be applied during STCC analysis is described and evaluated. Finally, in Section 5 a more controlled experimental study is presented on the TIMIT speech corpus demonstrating that ASR performance may be significantly improved through the use of phone specific warping factors.

## 2. REVIEW OF THE SCALE TRANSFORM

A commonly used approximation for the relationship between the spectral envelopes $F_A(\omega)$ and $F_B(\omega)$ of two speakers uttering phonetically similar steady state sounds is the uniform scale relationship

$$F_A(\omega) \approx F_B(\alpha_{AB}\omega), \tag{1}$$

where $\alpha_{AB}$ is the scaling constant that characterizes the difference between two speakers. In ASR, we would like a representation that is sensitive to phonetic differences but insensitive to speaker differences. An example of such a representation is the scale transform. The scale transform $D(c)$ of a function $F(\omega)$ is [1, 2]

$$D(c) = \frac{1}{\sqrt{2\pi}} \int_0^\infty F(\omega)\ \frac{e^{-jc\ln\omega}}{\sqrt{\omega}}\, d\omega \tag{2}$$

Using a change of variables, we can write Eq. 2 as

$$D(c) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} F(\omega = e^{\nu})\ e^{-jc\nu} e^{\nu/2}\, d\nu \tag{3}$$

The energy-normalized scaled version of $F(\omega)$ is $F_\alpha(\omega) = \sqrt{\alpha}F(\alpha\omega)$, where $\alpha$ is the magnification factor. The respective scale transforms of $F(\omega)$ and $F_\alpha(\omega)$, denoted by $D(c)$ and $D_\alpha(c)$, are related by $D_\alpha(c) = e^{jc\ln a} D(c)$. Hence, the scaling factor appears only in the phase factor and we have $|D_\alpha(c)| = |D(c)|$.[1] Thus the scale transform can be used to ascertain whether an arbitrary number of functions are scaled versions of each other without any *a priori* knowledge of the magnification factor. This is the

*This research was conducted while S. Umesh was a visiting researcher at AT&T Shannon Laboratory

[1]This is totally analogous to the Fourier transform case where the translation factor appears in the phase.

case even if magnification factors are all different. As seen from Eq. 3, what makes the magnitude of the scale transform insensitive to changes in scale is that it results from applying a Fourier transform to a frequency-warped representation of the signal. From Eq. 3 it is clear that the transformation from $\omega$ to $\nu$ is essentially log-warping of the frequency-axis. Hence, as shown below, in the log-warped domain the scaling $\alpha_{AB}$ appears as a translation factor.[2]

$$\begin{aligned} f_A(\nu) &= F_A(\omega = e^{\nu}) = F_B(\alpha_{AB}(e^{\nu})) \\ &= F_B(e^{(\nu + \ln \alpha_{AB})}) = f_B(\nu + \ln \alpha_{AB}) \end{aligned} \quad (4)$$

Since the magnitude of the Fourier transform is translation invariant, the scale invariance of the magnitude of the scale transform follows. Note that the uniform scaling parameter, $\alpha_{AB}$, appears as a shift parameter, $\ln \alpha_{AB}$ in the warped domain

In a subsequent work [3], we have shown that formant structures of different speakers are not related by an exact uniform scaling relationship. The scale relationship appears to be frequency dependent. Hence we need to find a warping function $\omega = g(\nu)$ (more appropriate than log-warping) such that

$$f_A(\nu) = F_A(\omega) = F_B(\alpha_{AB}(\omega)\omega) = f_B(\nu + \zeta_{AB}), \quad (5)$$

where $\zeta_{AB}$ is dependent on speakers A and B, but not on frequency. In [3], we have obtained a piece-wise approximation of the frequency-warping function, $\omega = g(\nu)$, for the frequency-dependent case, and found it to be similar to the Mel scale.

Finally, motivated by a desire to obtain a more parsimonious representation and to reduce the dynamic range of the spectral features, we use a log-transformation on the magnitude of the spectral features. Note that the log-warping affects only the magnitude, and functions that were frequency scaled version of each other, continue to remain so after the log-transformation. Using a form similar to Eq. 3, we compute the scale-transform based cepstral coefficients (STCC) as the magnitude of

$$D'(c) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \log(F(\omega = g(\nu))e^{\nu/2}) \; e^{-jc\nu} \, d\nu. \quad (6)$$

## 3. COMPARISON OF MFCC AND STCC FEATURES

In our first set of experiments, we compare the performance of the ASR system using the proposed scale-transform based features, STCC, and the mel-filter bank based features, MFCC. This is done by simply replacing the MFCC features with STCC features for a connected digit recognition task using a continuous density hidden Markov model based system. The state distributions were modeled using a mixture of Gaussian densities with diagonal covariance matrices. The training data consisted of 4781 digit string utterances from 3026 adult speakers collected over a public switched telephone network. The test data consisted of two databases, one made up of 2656 digit string utterances from 501 adults and another made up of 5767 digit utterances from children ranging in age from 10 years to 17 years.

[2]The Fourier transform of function, $f_A(\nu)$, in Eq. 4 has the same fundamental properties as the scale transform. The main difference being that the scale transform is derived from a Hermitian operator, and hence satisfies directly Parseval's theorem, and is *directly* reversible. However, one can also use the Fourier transform of function in Eq. 4 in which case the inverse uses the kernel $e^{jc \ln \omega}$.

| | | Adult | Child |
|---|---|---|---|
| MFCC Features | 6-Mix. | 97.42 | 82.73 |
| | 12-Mix | 98.07 | 84.54 |
| STCC Features | 6-Mix. | 89.39 | 75.51 |
| | 12-Mix | 91.41 | 77.46 |
| Decorrelated STCC Features | 6-Mix. | 94.97 | 79.02 |
| | 12-Mix | 96.32 | 79.72 |

**Table 1:** The Table shows the performance of MFCC and STCC features for a connected digit recognition task. The matched and mismatched results are in the columns labeled "adult" and "child" respectively

The results are shown in Table 1. The performance of STCC is much lower than MFCC under both matched and mismatched speaker conditions. On further analysis, we found that the STCC coefficients are highly correlated. The MFCC features, on the other hand, are approximately uncorrelated, since the DCT basis functions are the eigen functions of the spectral domain features. Since the STCC features are correlated, they may not be modeled accurately with a mixture of Gaussian densities with diagonal covariance matrices.

To test this hypothesis, we modified the STCC as follows. We first force-aligned the training speech data with the correct transcription. We then computed the covariance of STCC features for individual phone models, and computed a global covariance matrix $R$ as the weighted average of the phone-specific covariance matrices. Using $R$ we obtained the pre-whitening matrix, $D_R$, which is then multiplied with the STCC features to obtain $STCC_d$ which are approximately decorrelated. Using the decorrelated features, the previous digit recognition task was repeated. As seen from Table 1, the use of a simple global covariance matrix, $R$, significantly improves the performance of STCC features. This suggests that with a more appropriate model parameterization the performance of STCC features could be improved. Full covariance Gaussian distributions, distribution–dependent decorrelating transformations, and semi-tied covariance matrices, are examples of approaches for modeling correlations that allow for different trade–offs in the number of parameters required per Gaussian component.

We propose a simpler alternative, a normalization procedure, based on the idea of speaker-dependent shifts in the frequency-warped domain as discussed in Eq. 5. This is the subject of the next section.

## 4. A SIMPLE NORMALIZATION PROCEDURE

The computation of STCC and MFCC involves similar signal processing steps. The main difference is in the transformation of the frequency-warped spectral features to the cepstral coefficients. For the purposes of this discussion, we will assume that the frequency-warped spectral-domain features, $X[k]$, are the same for both MFCC and STCC. From our discussion in Section 1 (see Eq.5), speaker-dependent differences manifest themselves as translations of the frequency-warped spectral features, $X[k]$. Mathematically, the DCT of $X[k]$ denoted by $C[l]$ can be written as

$$C[l] = 2\Re[W_{2N}^{\frac{l}{2}} \underbrace{\sum_{k=0}^{N-1} X[k] W_{2N}^{kl}}_{|D_X[l]|e^{j\phi(l)} e^{-j\frac{2\pi}{2N} l k_0}}] \quad (7)$$

where $k_0$ is speaker related shift parameter with respect to a reference speaker. The MFCC retains unwanted "speaker-related" shift information, denoted by $k_0$, which introduces speaker-related variability. Alternately, as seen from Eq. 7, the STCC is insensitive to speaker-specific scale factor, $k_0$, because it uses only the magnitude $|D_X[l]|$. On the other hand, the STCC coefficients are correlated and are therefore not modeled adequately by a mixture of Gaussian densities with diagonal covariance matrices.

A reasonable approach would be to estimate and normalize the speaker-specific shift parameter, $k_0$, before computing the cepstral coefficients by applying DCT. From our discussion in Section 1, if we model speaker differences using constant frequency-scaling, then in the log-warped domain the spectral envelopes are translated versions of each other. This property holds even when the frequency dependent scaling is performed using a more appropriate warping function, $g(\nu)$. Most vocal-tract normalization approaches estimate linear scaling (which is related to our shift parameter, $k_0$, by a log transformation) by maximizing the likelihood of the data with respect to a target model acoustic model. Lee and Rose [4] provide an efficient procedure for such speaker normalization. However, most of these procedures are computationally expensive and require a target model. We propose a *simple* procedure to estimate the shift parameter, $k_0$ as a part of the cepstrum computation.

The algorithm for normalization is based on the idea that if two positive functions are translated versions of each other, then their center of gravities differ by the amount of translation. This is illustrated in Fig. 2. We define the center of gravity as

$$< m_A >= \frac{\sum_{i=0}^{N-1} i \cdot f(\nu_i)}{\sum_{i=0}^{N-1} f(\nu_i)} \quad (8)$$

where we assume a discrete variable function.

Hence, if the spectral envelopes are translated versions of each other in the frequency-warped domain, (translated by say $k_0$), then their center of gravities would differ by

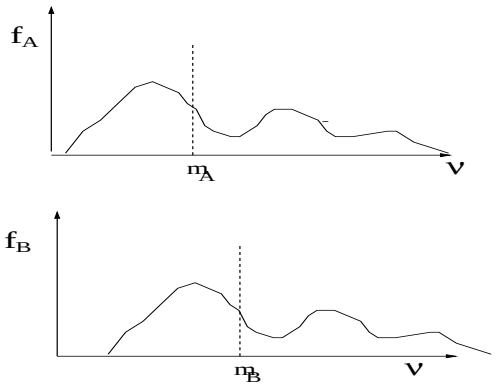


Figure 1: This figure schematically shows that if in the frequency-warped domain, $\nu$, two functions are shifted versions of each other, then their center of gravities, $m_A$ and $m_B$ would differ by an amount equal to the shift.

| | | Adult | Child |
|---|---|---|---|
| MFCC Features | 6-Mix. | 97.42 | 82.73 |
| | 12-Mix | 98.07 | 84.54 |
| Features from DCT *after* shift-normalization | 6-Mix. | 97.79 | 85.04 |
| | 12-Mix | 98.16 | 86.32 |

**Table 2:** The Table shows the performance improvement using the proposed speaker normalization procedure.

$k_0$. In our proposed algorithm, we find the average center of gravity over all frames in an utterance, and compensate (or normalize) the frequency-warped spectral envelope, before computing the normalized cepstral coefficients using DCT. Thus we obtain normalized features that are *also* uncorrelated. The computational cost for such features is extremely small, the only additional computation is the estimation of the center of gravity.

Since the amplitude variations of the spectral envelope affect the estimation of the center of gravity, we divide the frequency region into three bands and estimate the center of gravity independently in each band. These estimates are combined to obtain a robust estimate of the overall center of gravity. The estimates from a given band are used for estimating the overall center of gravity only if the energy in the band exceeds a threshold. This improves the reliability of the estimate. The speaker-specific scaling constant, $\hat{k_0}$, is obtained by computing the difference between the center of gravity of the frequency-warped spectral envelope and an arbitrary reference. The shift-normalized features are computed by multiplying the cepstral features by $e^{+j2\pi\frac{l\hat{k_0}}{2N}}$ before computing the real part as in Eq. 7. The results of a recognition experiment using the above normalization procedure is shown in Table 2. These experiments are performed on the *same* database and task described in Section 3. It can be seen that there is a 12% reduction in the recognition error when the shift-normalized features are used relative to the error obtained using MFCC coefficients.

| | Male | Female | Girls | Boys |
|---|---|---|---|---|
| MFCC | 85.9 | 48.7 | 34.6 | 34.4 |
| Shift-normalized features | 84.01 | 75.98 | 73.05 | 70.56 |

**Table 3:** Results of vowel classification using 2-mixture GMM models trained on male speech (Hillenbrand data).

## 5. VOWEL CLASSIFICATION USING FREQUENCY-SCALING NORMALIZATION

From the previous section, it is clear that shift-normalization does improve recognition accuracy. However, there is still considerable difference in performance between matched and mismatched speaker conditions in spite of the normalization. This leads to the question whether there are factors other than frequency scaling that significantly contribute to speaker differences.

In this section, we describe a set of experiments that were performed to test the effect of phone-specific normalization on classification rates of vowels. The results may have general implications for all frequency-warping normalization procedures. In our first experiment, we address the question whether normalization by appropriate shifting after frequency-warping as described in Section 1 reduces most of speaker variability. Since the aim of the experiment is to test the idea that speaker variability is almost completely characterized by the shift parameter, we assume complete knowledge of the speaker and the vowel being spoken.

A convenient database for this controlled study is the Hillenbrand data, which consist of vowels spoken in the same hVd context. The speaker population consist of men, women, boys and girls. The speech data consist of 9 vowels spoken in the same context and is sampled at $16KHz$. For each vowel, we estimate the shift parameter for each speaker and vowel by correlating the frequency-warped spectral envelopes for an utterance from that speaker with that of the same vowel spoken by a reference set of speakers. The reference speakers were chosen arbitrarily. Normalized features for the given utterance is computed using this estimate of the shift factor. A 2-mixture GMM is built using half the adult male data. The results of the vowel classification experiment under matched and mismatched conditions is shown in Table 3. For comparison, we also repeated the same experiment using MFCC features. It is clear that shift-normalization significantly reduces the degradation in performance under mismatched speaker conditions.

In Section 3, for the digit recognition task, there is considerable degradation in performance between matched and mismatched speaker conditions *even* after the normalization procedure. The normalization is done by a single speaker-specific scale factor for the entire utterance. On the other hand for vowel classification task spoken in the *same* context, the results indicate that with appropriate shift-normalization for *each* vowel and *each* speaker, there is only a small difference in performance between matched and mismatched speaker conditions. We now investigate the effect of normalization at different levels – one normalization factor for each speaker for all vowels, vowel-specific normalization for each speaker, and finally context-specific phone normalization for each speaker. These experiments are similar to that done with Hillenbrand data, but here we use TIMIT Dr5 data, where the vowels are spoken in different context. We build 2-mixture GMM models for vowels using male data. Table 4 shows the results when normalization is done at different levels. As a reference, we show the classification results for MFCC features also. From Table 4, it is clear that while a speaker-specific warping (i.e. one normalization factor per speaker) improves vowel classification performance in the mismatched case, using phone class specific warping provides additional improvement. Finally, when this is taken a step further so that separate warping factors are computed for phones as they appear in separate contexts, we obtain the best performance. Classification results for matched and mismatched speaker cases are almost the same for this scenario.

| | Male Model | | Female Model | |
|---|---|---|---|---|
| | Male | Female | Male | Female |
| MFCC | 53.0 | 40.0 | 33.5 | 48.5 |
| One Normalization per speaker for all vowels | 54.5 | 50.5 | 44.5 | 46.5 |
| Vowel-level Normalization | 55.0 | 54.0 | 49.0 | 52.0 |
| Context-specific vowel Normalization | 58.5 | 55.5 | 51.5 | 55.0 |

**Table 4:** Results of vowel classification accuracy (% accuracy) for different levels of normalization.

## 6. CONCLUSION

In summary, the results obtained from this work suggest that the scale-transform, coupled with the associated low-complexity frequency warping and normalization procedure, represents a robust and practical alternative to existing feature analysis techniques. In addition our experiments on vowel classification suggests that using context-specific normalization provides performance that is close to the matched speaker case.

## 7. REFERENCES


1. L. Cohen, "The Scale Representation," *IEEE Trans. Signal Proc.*, pp. 3275–3292, Dec. 1993.
2. S. Umesh, L. Cohen, N. Marinovic, and D. Nelson, "Scale Transform In Speech Analysis," *IEEE Trans. on Speech and Audio Processing*, Jan.1999.
3. S. Umesh, L. Cohen, N. Marinovic, and D. Nelson in *Proc. International Conference on Spoken Language Processing*, (Philadelphia, USA), 1996.
4. Li Lee and R. C. Rose, "A Simple Frequency-Warping Approach to Speaker-Normalization" *IEEE Trans. on Speech and Audio Processing*, Jan.1998.