
Multidimensional Systems and Signal Processing, 9, 465–473 (1998)

Scale and Translation Invariant Methods for Enhanced Time-Frequency Pattern Recognition

WILLIAM J. WILLIAMS* wjw@eees.umich.edu wjw@w.imap.itd.umich.edu
Electrical Engineering and Computer Science Dept., 423OC EECS Bldg., University of Michigan, Ann Arbor MI 48109

EUGENE J. ZALUBAS
Electrical Engineering and Computer Science Dept., 4313 EECS Bldg., University of Michigan, Ann Arbor MI 48109

ROBERT M. NICKEL
Electrical Engineering and Computer Science Dept., 4313 EECS Bldg., University of Michigan, Ann Arbor MI 48109

ALFRED O. HERO III
Electrical Engineering and Computer Science Dept., 42329 EECS Bldg., University of Michigan, Ann Arbor MI 48109

Received October 4, 1996; Revised February 12, 1997

Abstract. Time-frequency (t-f) analysis has clearly reached a certain maturity. One can now often provide striking visual representations of the joint time-frequency energy representation of signals. However, it has been difficult to take advantage of this rich source of information concerning the signal, especially for multidimensional signals. Properly constructed time-frequency distributions enjoy many desirable properties. Attempts to incorporate t-f analysis results into pattern recognition schemes have not been notably successful to date. Aided by Cohen's scale transform one may construct representations from the t-f results which are highly useful in pattern classification. Such methods can produce two dimensional representations which are invariant to time-shift, frequency-shift and scale changes. In addition, two dimensional objects such as images can be represented in a like manner in a four dimensional form. Even so, remaining extraneous variations often defeat the pattern classification approach. This paper presents a method based on noise subspace concepts. The noise subspace enhancement allows one to separate the desired invariant forms from extraneous variations, yielding much improved classification results. Examples from sound classification are discussed.

Key Words: time-frequency, scale, speech, pattern recognition

1. Introduction

Time-frequency (t-f) analysis is useful for signals which exhibit changes in frequency content with time. The well-known spectrogram often presents serious difficulties when it is used to analyze rapidly varying signals, however. If the analysis window is made short enough to capture rapid changes in the signal it becomes impossible to resolve signal components which are close in frequency within the analysis window duration. If the window is made long to resolve the frequencies of sinusoids, the time when sinusoidal components act becomes difficult to determine.

* Supported in part by ONR Grants nos. N00014-90-J-1654 and N000014-97-1-0072, NSF Grant BCS 9110571 and DOD Contract no. MDA904-95-C-2157

Until recently, there was one alternative t-f analysis technique which was widely believed to avoid some of the problems of the spectrogram. The well known Wigner distribution (WD) avoids the problems of windowing and enjoys many useful properties, but often produces an unacceptable amount of interference or cross-term activity between signal components when the signal consists of many components [3,5]. Despite its shortcomings, the Wigner distribution has been employed as an alternative to overcome the resolution shortcomings of the spectrogram. It also provides a high resolution representation in time and in frequency. The WD has many important and interesting properties.

Both the spectrogram and the WD are members of Cohen's Class of Distributions [2]. Cohen has provided a consistent set of definitions for a desirable set of t-f distributions which has been of great value in this area of research. Different members of Cohen's class can be obtained by using different kernels. In this framework, the WD has a unity valued kernel. Choi and Williams introduced the Exponential Distribution (ED), with kernel $\phi_{ED}(\theta, \tau) = e^{-\theta^2\tau^2/\sigma}$, where σ is a kernel parameter ($\sigma > 0$) [1]. The ED overcomes several drawbacks of the spectrogram and WD, providing high resolution with suppressed interferences. A recent comprehensive review by Cohen [3] provides an excellent overview of TFDs and recent results using them.

The Reduced Interference Distribution (RID), which is a generalization of the ED, shares many of the desirable properties of the WD, but also has the important reduced interference property. RID is discussed in a recent book chapter [13] and a design procedure for RID kernels has been developed [7]. One may start with a primitive function, $h(t)$, which has certain simple constraints, and evolve a full-fledged RID kernel from it. The RID kernel retains a unity value along the θ and τ axes in the ambiguity plane, generally providing good time-frequency resolution and auto-term preservation, but attenuates strongly elsewhere for good cross-term suppression.

1.1. The Scale Transform

The scale transform has been described by Cohen [4] to be:

$$D(c) = \frac{1}{\sqrt{2\pi}} \int_0^\infty x(t) \frac{e^{-jc\ln t}}{\sqrt{t}} dt \tag{1}$$

The scale transform has an analogy to the Fourier transform. The Fourier transform of a signal, $x(t)$ and the Fourier transform of a shifted version of that signal, $x(t - t_o)$ differ only by a phase factor.

$$F[x(t - t_o)] = X_o(\omega) = X(\omega)e^{-j\omega t_o} \tag{2}$$

so that

$$|X(\omega)| = |X_o(\omega)|. \tag{3}$$

In a like manner, the scale transform of $\sqrt{|a|}x(at)$ differs from the scale transform of $x(t)$

only by a phase factor, so that the magnitudes of the scale transform of $x(t)$ and $\sqrt{|a|}x(at)$ are identical.

$$|D(c)| = |D_a(c)| \tag{4}$$

We have developed discrete forms of the scale transform [14,16] which can be computed efficiently. One might question the use of the scale transform rather than the more well-known Mellin transform. There are several reasons for using the scale transform. One reason is that the standard Mellin transform weights signal components in lower time more than in higher time. A second reason is the relationship of scale to wavelet concepts and the insights it brings in this light.

2. Acoustic Signals

Two types of acoustic signals were used to test the effectiveness of these methods. These were human speech and marine mammal sounds.

2.1. Marine Mammal Sounds

Marine mammal sounds are well characterized using the RID and overcome some of the shortcomings of the SP as described by Watkins [12]. RID clearly reveals both the tonal structure in the whistles and the temporal structure of clicks which are simultaneously produced by these animals. It appears that the clicks of marine mammals such as whales and dolphins may have a distinctive structure based on the individual animal and may be useful in nonintrusive tagging and tracking of these animals. Our new TFD methods provide a powerful means of representing the complex sounds produced by marine mammals.

One can now readily design TFDs which represent the joint energy of a signal as a function of time and frequency or space-frequency distributions which represent the joint energy of images as space-spatial frequency distributions (two spatial variables x and y and two spatial frequency variables Ω_X and Ω_Y). Furthermore, with careful design, these joint distributions can exhibit proper covariances with time, frequency or spatial shifts such that the representation shifts in accordance with these shifts but does not change in its configuration [14]. The well-known spectrogram has been extensively used in speech analysis and it has these useful properties. A shift in time or a shift in frequency of the signal[1] will shift the representation appropriately in time and frequency. However, the spectrogram does not exhibit the proper characteristics in response to *scale* changes in the signal. That is if $x(t)$ becomes $x(at)$, the Fourier transform of $x(t)$ changes from $X(\omega)$ to $\frac{1}{a}X(\frac{\omega}{a})$. This is illustrated in Figure 1.

2.2. Speech Processing

The spectrogram has long been a widely used tool in speech analysis. Other TFDs have been investigated in speech analysis, but none have yet provided a strong advance beyond

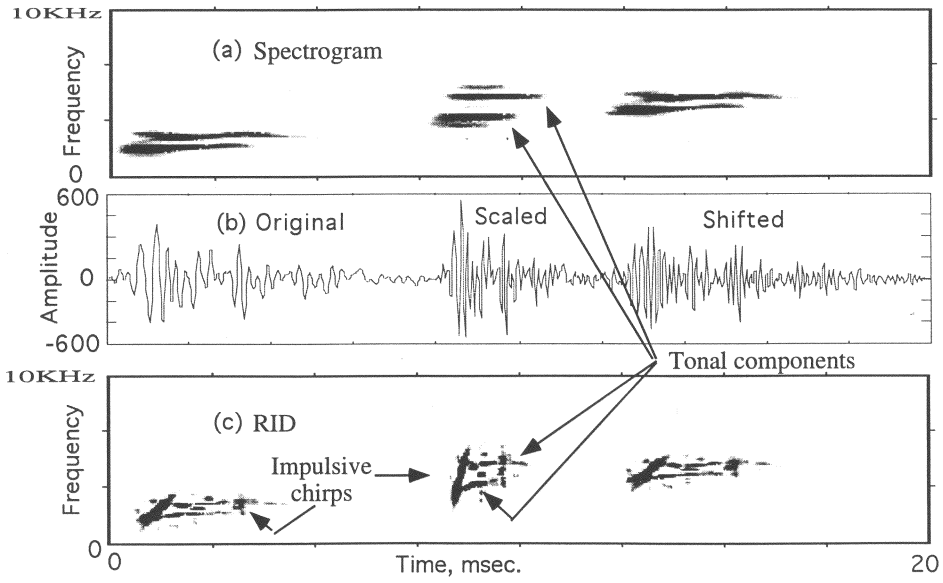

Figure 1. TFD results for time shifted, frequency shifted and scaled dolphin click. a. Spectrogram. b. Original click, scaled and time shifted click, time shifted and frequency shifted click. c. Reduced Interference Distribution (RID) click results for the same time shifts, scaling and frequency shift. (From: W. J. Williams, Reduced Interference Distributions: Biological Applications and Interpretations, *IEEE Proc.*, Vol 84, 1996, pp 1264–1280.)

the level of analysis provided by the spectrogram [9]. It is generally believed that the potential exists for significant advances in speech analysis using recently developed TFD tools, however.

3. Classification and Detection of Acoustic Signals

Acoustic signals may vary in time of occurrence, frequency and scale. The Doppler effect manifests itself as scaling. The "scalogram" [10], which is an analog of the spectrogram with frequency replaced by scale, might be effective in analyzing scaled signals. However, it, as does the wavelet transform, lacks the frequency covariance property. One might like to have invariant representations under time-shift, frequency-shift and scale. The techniques described in this paper may be assembled to achieve all of these invariances. First, the sound or a segment of the sound to be analyzed must be isolated. Next, the RID is computed. Then, the autocorrelation[2] along time is performed as

$$A_{RID}(m,k) = \sum_n RID(n,k)RID(n-m,k) \tag{5}$$

where n is the time sample, k is the frequency sample and m is the autocorrelation lag sample. This removes absolute time and produces a centered autocorrelation of each frequency slice.

Next, one has the choice of also removing absolute frequency by performing a similar autocorrelation along k. Finally, one may scale transform the resulting representation along the time and frequency directions. This serves to produce a representation that is invariant to time, frequency and scale shifts.

It may not be desirable to remove all variation. These variations may serve to classify or detect the signal. For example, if frequency shift is an important indicator of the identity of a signal, one may bypass that step. It is important to note that even though these techniques serve to make the representations invariant, true frequency shifts and scale shifts may be retained in the phases of the requisite transforms.

3.1. Classification and Detection of Images

Recognizing characters or spotting words in bitmapped documents (images) has been of particular interest to us [14]. One may convert an image into a four dimensional representation in a manner analogous to the conversion of a one dimensional signal into a t-f representation. Here, the two spatial dimensions (x, y) are retained and joined by the spatial frequencies (ω_x, ω_y). Four dimensional kernels may be applied in a manner similar to t-f analysis. We have developed software which accomplishes this and it is clear that these 4-D representations are rich in detail. The 4-D structure does not prevent the application of the invariance producing transformations along all four dimensions, however. In order to reduce the complexity of the representations, we have reduced the image computations to two dimensional autocorrelations along x and y with the idea of expanding to the full four dimensional forms after working out details in the simplified formulations. Noise subspace and higher order statistical moments have been gainfully applied to this problem [6].

3.2. Applying the Scale Transform

One of the problems in applying the scale transform is finding the zero reference. Unlike the Fourier transform, the scale transform exhibits strong non-stationary characteristics. However, the process of autocorrelation provides an unequivocal zero reference for time. Likewise, the frequency dimension of the RID has an unequivocal zero frequency reference, so that the scale transform may be easily applied. Due to symmetries, the 2-D scale transform may only need to be applied to unique quadrants of the autocorrelated RID representation or the 2-D autocorrelated images. Previous results in using the 2-D scale transform to render 2-D autocorrelated images invariant to displacement and scale are very encouraging [14].

3.3. Sound Classification using the Invariant Representations

Starting with a suitable TFD[3], almost all of the undesired variation due to time shift, frequency shift and scale may be squeezed out of the final invariant form. There may still be some residual effects due to discretization and computation. The next task is to design a classifier. Suppose that the invariant form is characterized by a two dimensional

representation $\Delta(p, q)$. This 2-D representation may be decomposed using eigensystem techniques as

$$\Delta(p, q) = \sum_j a_j \beta_j(p, q) \tag{6}$$

where the $\beta_j(p, q)$ are eigenimages and the a_j are the eigenvalues of the decomposition. The eigensystem decomposition is carried out on a collection of $\Delta(p, q)$ examples coming from the classes of objects (signals or images) that are of interest. The eigensystem decomposition then provides an ordered set of eigenimages ordered according to their eigenvalues. Although the eventual goal is to use true two dimensional eigenimage analysis, suitable algorithms to accomplish this have not been identified. One may utilize a simpler one dimensional approach which lends itself to readily available algorithms.

The 2-D $N \times M$ invariant forms may be converted into vectors of length $N \times M$ by either concatenating the rows or columns. Then, readily available Singular Value Decomposition (SVD) techniques may be applied to the vectorized set of images. Suppose there are several different extraneous variations in the supposedly invariant representations caused by a variety of factors. For example, the same person may not say the same word exactly the same way each time or the same whale or dolphin may click slightly differently each time. Such extraneous variations often confound the invariant representations so that effective detection or classification of a specific signal or image is rendered impossible. A new and very effective method using noise subspace concepts has been developed to overcome these problems.

4. Noise Subspace Methods

The $N \times M$ vectorized 2-D forms have a large number of elements. Usually, for classification methods to work, one wishes to have a considerably greater number of representations of the signal vectors than there are elements in those representations. Here, we have exactly the opposite. There are many more elements in the vectorized 2-D forms than there are vectorized 2-D forms. This is usually a statistical nightmare. However, suppose there are K sound examples ($K << N \times M$). Then the SVD produces $N \times M$ orthogonal eigenvectors, the first K of which form a complete orthonormal basis for the vectorized invariant forms. The remaining SVD eigenvectors (the noise eigenvectors) must be orthogonal to all of the original vectorized invariant forms. Suppose that we now obtain a new signal. Convert it into the TFD, then to the 2-D invariant form and finally, vectorize the 2-D invariant form. If it belongs to the set of vectorized 2-D invariant forms used to produce the SVD results, then it should be **orthogonal** to all of the noise eigenvectors produced by the SVD. Therefore, its projection on any of the noise eigenvectors should be zero. If we have carried out the whole process through the SVD for a number of different sets of signals, we should find the projection of the vectorized 2-D invariant form of the unknown signal on the noise eigenvectors of each set of signals. The smallest result will be theoretically obtained when this is done using the noise eigenvectors of the set to which the signal belongs.

5. Sound Classification Results

Two different experiments were carried out. Invariant forms as described were derived from the signals. In one experiment the clicks from two sperm whales were considered. In the second experiment, speaker identification was attempted. Ten speakers said 'Michigan' ten times. Half the data sets were used to develop the classifier and the technique was tested using the remaining data sets. In both experiments, classification success of individuals was quite high. Some results are shown in Figure 2. Half of the responses for each individual were used to develop a classifier and the other half to test it. Only autocorrelation along time was applied to the RID result in the human speaker study. It is believed that variations in the frequency direction should be retained, since they may serve to identify individual speakers. The methodology reported in this paper is one of two new techniques we have developed recently. The other (moment-based) method [17] is also quite effective in sound

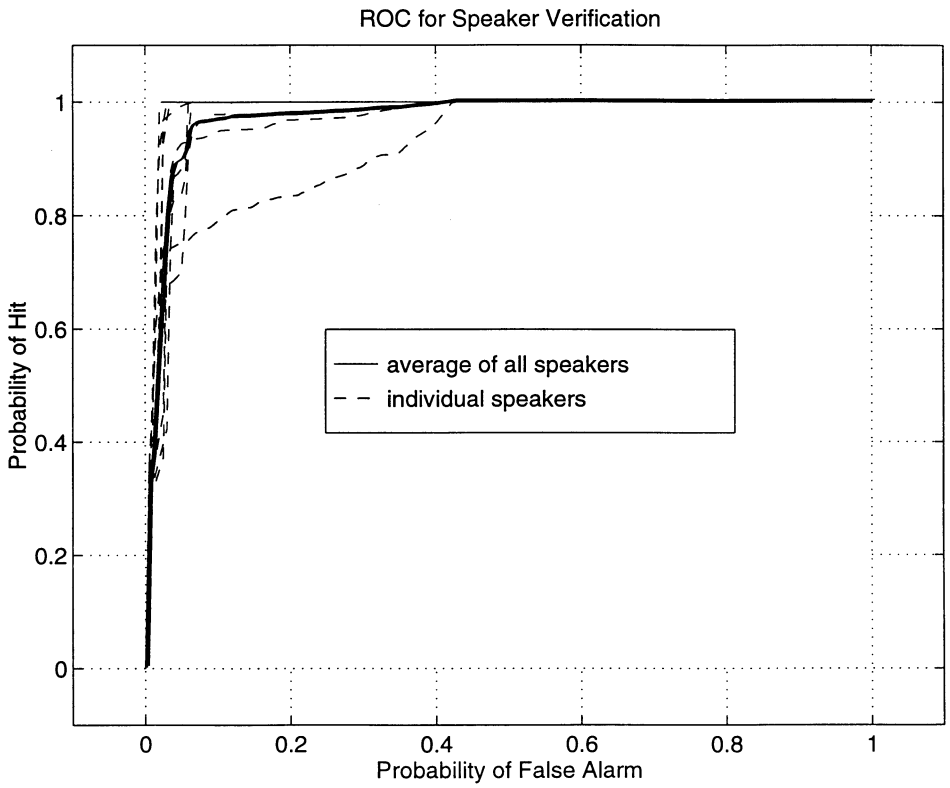

Figure 2. Individual ROC results for ten speakers and the average ROC for all speakers.

pattern recognition. Tacer and Loughlin [11] have had success in untilizing joint TFD moments in classifying postural sway.

6. Conclusions

The methods outlined in this paper seem to work very well. At this point the ideas have been applied in a straightforward way with very little "tuning" of the various components of the technique. At this time the approach appears to be competitive, at least, with alternative methods of speaker identification where complex and highly developed alignment algorithms and time warping algorithms have been applied together with commonly used pattern recognition engines. Further refinements of our technique may provide significant improvements over present results. It is believed the technique could be applied to a wide variety of sounds, other signals and images in terms of specific identification of distinct classes of signals and images. There are a number of trade-offs to be considered and we have a number of improvements in mind. The computational burden is high, but with dedicated hardware and fast algorithms it is believed that very reliable real time detection and classification of sounds could be achieved. The methods need to be tested with large databases and the classifications need to be extensively tested using new data sets which have not participated in the classifier design. Methods based on joint moments also appear to be promising, but at this writing the methods presented herein appear to have an advantage.

Notes

1. Within reasonable bounds that do not induce aliasing or some other undesirable effect.
2. One can also carry out this computation in the frequency domain as well, using FFTs along the n dimension and obtaining the magnitude of the resulting image.
3. TFDs other than RID may be suitable [8].

References

1. H. I. Choi, and W. J. Williams, "Improved Time-Frequency Representation of Multi-component Signals Using Exponential Kernels," *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. 37, no. 6, 1989, pp. 862–871.
2. L. Cohen, "Generalized Phase-Space Distribution Functions," *J. of Math. Phys.*, vol. 7, 1966, pp. 781–786.
3. L. Cohen, "Time-Frequency Distributions—A Review," *Proc. IEEE*, vol. 77, no. 7, 1989, pp. 941–981.
4. L. Cohen, "The Scale Representation," *IEEE Trans. Signal Processing*, vol. 41, 1993, pp. 3275–3292.
5. L. Cohen, *Time-Frequency Signal Analysis*, Prentice Hall, 1995.
6. A. O. Hero, III, J. C. O'Neill, and W. J. Williams, "Pattern Classification via Higher Order Moments and Signal Subspace Projections," CSPL Technical Report No. 20-96 EECS Department, University of Michigan, 1996.
7. J. Jeong W. J. and Williams, "Kernel Design for Reduced Interference Distributions," *IEEE Trans. Sig. Proc.*, vol. 40, no. 2, 1992, pp. 402–412.
8. P. J. Loughlin, "Comments on Scale Invariance of Time-Frequency Distributions," *IEEE-Signal Processing Letters.*, vol. 2, 1995, pp. 4–6.

9. J. W. Pitton, K. Wang, and B.-H. Juang, "Time-Frequency Analysis and Auditory Modeling for Automatic Recogition of Speech," *IEEE Proc.*, vol. 84, 1996, pp. 1199–1215.

10. O. Rioul, and P. Flandrin, "Time-Scale Energy Distributions: A General Class Extending Wavelet Transforms," *IEEE Trans. Sig. Proc.*, vol. 40, no. 7, 1992, pp. 1746–1757.

11. B. Tacer, and P. J. Loughlin,"Time-frequency Based Classification," *Advanced Signal Processing Algorithms, Architectures and Implimentions, VI*, SPIE vol. 2846, 1996, pp. 186–192.

12. W. A. Watkins, "The Harmonic Interval: Fact or Artefact in Spectral Analysis of Pulse Trains," *Marine Bio-acoustics*, vol. 2, 1966, pp. 15–43.

13. W. J. Williams, and J. Jeong, "Reduced Interference Time-Frequency Distributions," B. Boashash Ed., *Time-Frequency Signal Analysis*, Longman and Cheshire—J. W. Wiley, 1992.

14. W. J. Williams, E. J. Zalubas, and A. O. Hero, III, "Separating Desired Image and Signal Invariant Components from Extraneous Variations," *Advanced Signal Processing Algorithms, Architectures, and Implementations, SPIE Proceedings*, vol. 2846, 1996

15. W. J. Williams, "Reduced Interference Distributions: Biological Applications and Interpretations," *IEEE Proc.*, vol. 84, 1996, pp. 1264–1280.

16. E. J. Zalubas, and W. J. Williams, "Discrete Scale Transform for Signal Analysis," *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, vol. 3, 1995, pp. 1557–1561.

17. E. J. Zalubas, J. C. O'Neill, W. J. Williams, and A. O. Hero III, "Shift and Scale Invariant Detection," *IEEE Int. Conf. Acous., Speech, Sig. Proc.*, vol. 5, 1997, pp. 3637–3640.