INCORPORATING FREQUENCY WARPING INTO SPARSE COMPONENT ANALYSIS

Saeid. Sanei, Savvas Constantinides, Clive Cheong Took, and Binyang Song

Centre for Digital Signal Processing, School of Engineering, Cardiff University, Cardiff, CF24 3AA, UK,
Phone +44(0)2920875946, email: {saneis, cheongc, songb}@cf.ac.uk and savvas_constantinides@hotmail.co.uk

ABSTRACT

Frequency warping using short-time Laguerre transform (STLT) has been employed here as an effective tool in increasing the efficiency of the sparse component analysis (SCA) for underdetermined blind source separation systems. An attempt has been made to maximise a measure of sparseness. There are three major advantages for such an application; 1. The psycho-acoustic features such as fundamental harmonics are well separated in frequency domain, 2. The permutation problem as the most troublesome effect in frequency domain blind source separation (FDBSS), is mitigated, and 3. In SCA the sparseness measured based on l_0-norm, increases and hence the performance of the SCA methods is improved.

1. INTRODUCTION

Frequency warping is an interesting processing effect in which the frequency axis is remapped to obtain a signal with desired characteristics [1]. Laguerre transform is effectively used in warping (for either expansion or compression) of the signals. Frequency warping has been traditionally employed in signal compression, speaker normalization [2], sound morphing, detuning the partials, pitch shifting, as an approach to wavelet transform [3], and many other applications. During the warping process, depending on the sign of the warping parameter the frequency components of the signals are shifted to the left and right resulting in either compression or expansion of the spectrum respectively. In BSS it is very favourable to expand the frequency axis to increase the sparseness of the signal in frequency domain. In the case of musical signal separation very often the number of sources is larger than the number of sensors (i.e. the system is underdetermined). Moreover, these signals are sparser in frequency domain than in the time domain. In such cases the warping process can be adapted to increase the sparseness of the signals in frequency domain.

As for the other BSS methods, there are ambiguities due to the change in sign, scale, and order of the output independent components. Techniques exist to overcome these problems, but no robust method has, as yet, been found to overcome the permutation problem. Most of the methods reported in the literature rely on estimation of the direction of arrival (DOA) and exploitation of psychoacoustic properties of the human auditory system [5] [6]. As a result of frequency warping these features are enhanced and the frequency components are better discriminated. In the following sections we explain how to set the warping parameter in order to have the maximum sparseness in frequency domain. An overall model for the system has been given in Figure 1.

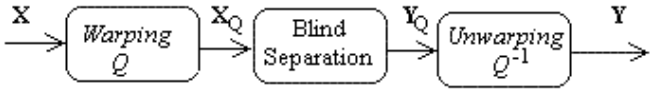

Fig. 1 The overall proposed warping-FDBSS system

2. SHORT-TIME FREQUENCY WARPING

For a block-based first-order Laguerre transform (LT) the transfer function is denoted by:

$$Q(z) = \frac{z^{-1} - b}{1 - bz^{-1}} \qquad (1)$$

where $-1<b<1$ is the LT parameter. If b < 0 the effect of warping will be compressing the signal in time while b > 0 expands the signal. In a real time processing, however, the STLT is computed by Laguerre transforming windowed frames of the signal $x(n)$. Practically, an iterative, non-causal scheme to compute the Laguerre coefficients is given by the diagram in Figure 2. In this diagram the second block is defined as:

$$\Lambda_0(z) = \frac{\sqrt{1-b^2}}{1 - bz^{-1}} \qquad (2)$$

This block is used to change the basis of the input signal. The third block is a dispersed delay line for which $Q(z)$ defined in (1) is tapped sequentially [3].

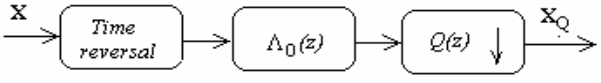

Fig. 2 Structure for computing the LT of the mixtures

The estimated sources are unwarped at the end using an inverse LT or more practically following the block diagram in Figure 3 [3]. Here the first block is an inverse operation to the third block in Figure 2. In the following sections we

examine the effect of warping on separation of the underdetermined mixtures.

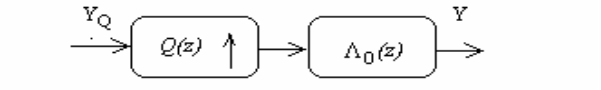

Fig. 3 Unwarping the separated components

In many applications such as sound or scene effects, a time varying frequency warping is more adequate. In such cases by defining

$$Q_r(z) = \begin{cases} 1 & ,r = 0 \\ \prod_{k=1}^{r} \dfrac{z^{-1} - b_k}{1 - b_k z^{-1}} & ,r > 0 \end{cases} \qquad (3)$$

the warped mixture signal after recurrent r is

$$x_{qr} = \sum_{k=-\infty}^{\infty} x(k) q_r(k) \qquad (4)$$

where $Q_r(z)$ is the z-transform of $q_r(k)$. Similarly the independent components can be unwarped by calculating

$$y(k) = \sum_{r=0}^{\infty} y_{qr} \psi_r(k) \qquad (5)$$

where

$$\Psi_r(z) = \begin{cases} \dfrac{1}{1 - b_1 z^{-1}} & ,r = 0 \\ \prod_{k=1}^{r} \dfrac{1 - b_r b_{r+1}}{(1 - b_r z^{-1})(1 - b_{r+1} z^{-1})} Q_r(z) & ,r > 0 \end{cases}$$

where $\Psi_r(z)$ is the z-transform of $\psi_r(k)$. However, the energy preservation property of the orthogonal warping is no longer valid.

3. MITIGATING THE PERMUTATION PROBLEM

Frequency-domain BSS (FD-BSS) algorithms are sensitive to the permutation of the separated frequency bin signals. The scaling ambiguity can cause the scaling of every frequency band to be different resulting in spectral deformation of the original sources. As suggested in [7], the scaling problem can be remedied by forcing the determinant of the separating matrices to unity. This prevents alteration of the spectral envelope, while preserving the separation. But permutation indeterminacy remains as an open problem. In regions where there is no severe spectral deformation and the number of sources is low, the uniformity of the spectrum may be exploited in readjusting the weights of the separating matrix to alleviate the problem. However, a systematic approach to the problem is required where the number of sources is high. The existing methods for solving the problem are: (1) constraints on the filter models in the frequency domain [7] [1]; (2) exploiting the continuity of the spectra of the recovered signals [8]; (3) co-modulation of different frequency bins [9]; (4) using a time-frequency source model [7]; (5) using a beamforming view and measurement of the direction of arrival (DOA) [5], and finally a recent scheme for realigning the permuted components based on a coupled HMM [6]. For separation of musical signals the beamforming methods are suitable since the position of the players and the microphones are often fixed within a reasonable distance from the instruments. However, the DOAs for different sources are not well separated in angular positions for when the environment is highly reverberant i.e. the mixing filter is long. The angle of arrival, θ, is proportional to the phase difference between the mixtures, φ, i.e. for two mixtures for each time point k, $\theta(\omega,k) = \cos^{-1} \dfrac{\varphi(\omega,k)c}{\omega d}$, where c is the speed of sound and d is the microphone spacing.

4. SCA AND WARPING

The most important property of warping is an increase in sparseness of the mixtures (in frequency domain) when the number of sources is larger than the number of sensors. This is extremely important in the separation of music originated from different (or even similar) instruments. In this case, the standard ICA cannot be utilised since the mixing matrix is not invertible. In these cases transforming the signal from time to frequency domain by itself increases the sparseness [11]. Assuming no major overlap between (more than two of) the source signals in a real room recordings the mixtures are convolutive and the mixed signals are represented as $x_j(t) = \sum_{i=1}^{N} h_{ji} * s_i(t)$, where * is convolution sign and h is the mixing medium. The problem here is to determine belong to which signal each sample in the time-frequency domain is. In order to do that the phase difference between each two observed signals, $\varphi(\omega,k)$, is measured as $\varphi(\omega,k) = \angle \dfrac{X_1(\omega,k)}{X_2(\omega,k)}$. The direction or angle of arrival (DOA) $\theta(\omega,k)$ is then proportional to the phase difference between the mixtures, φ, i.e. for two mixtures for each discrete-time point k, $\theta(\omega,k) = \cos^{-1} \dfrac{\varphi(\omega,k)c}{\omega d}$, where c is the speed of sound and d is the microphone spacing. After frequency warping, $\varphi(\omega,k)$ will change to

$$\varphi(\omega,k) = \angle \frac{X_{q_1}(\omega,k)}{X_{q_2}(\omega,k)}$$ where $X_{q_1}(\omega,k)$ and $X_{q_2}(\omega,k)$ are the warped mixtures. Finally the peaks are classified using an unsupervised method such as k-mean algorithm. This is done by plotting $\theta(\omega,k)$ and using a binary mask defined as

$$M_\gamma(\omega,k) = \begin{cases} 1 & \hat{\theta}_\gamma - \Delta \le \theta(\omega,k) \le \hat{\theta}_\gamma + \Delta \\ 0 & elsewhere \end{cases} \quad (7)$$

where Δ is the range parameter and $\hat{\theta}_\gamma$ is the estimated DOA for source γ. Each signal is then extracted by calculating $X_j^c(\omega,k) = M_\gamma(\omega,k) X_j(\omega,k),\ j = 1,2$. For smaller value of Δ better separation but a higher distortion results. When Δ increases, the musical noise reduces but the separation performance deteriorates. After one source is separated the same process may be continued to extract the second source or an ICA-based BSS can be applied for solving the two-source-two-microphone problem [10]. However, often there are overlaps between the classes or the peaks are very close to each other. Therefore a binary decision making process results in musical noise. By expanding the signals in frequency domain through warping we can considerably mitigate this problem making the classes well away from each other and the peaks split. However, since by expanding the signals the length of the signal increases there is a compromise between the amount of expansion. This results in an increase in complexity, time, and the degree of sparseness. In the literature l_0 and l_1 norms are often used as measures of sparseness, i.e. for l_0 -norm, $\min \sum_{i=1}^{m} \sum_{j=1}^{k} |s_{ij}|^0$, *subject to* $\mathbf{HS} = \mathbf{X}$, where k is the number of sources and m is the signal duration. Here, the received samples are distributed around a number of distinct DOAs proportional to the number of sources. In such (convolutive) cases the value of peak divided by variance can best represent the sparseness. Normally the peak amplitudes are rather small. On the other hand, if the warping parameter b is largely negative, the variance will be high. Therefore we choose b (which in this application remains the same for all the observed signals) in order to have

$$\max(sparseness) = \max \frac{\sum_{i=1}^{m} peak_i(dist(\theta))}{\text{var}(dist(\theta))} \quad (8)$$

where *dist* refers to distribution. Figure 4 shows the histogram for the values of θ when we have two mixtures and three sources set in 30^0 and 90^0 and 135^0 positions; (a) is for before warping and (b) is for after warping the mixtures. The mixtures are recorded in a real room environment. Figure 5 shows the effect of warping on the spectrum (DFT points=512).

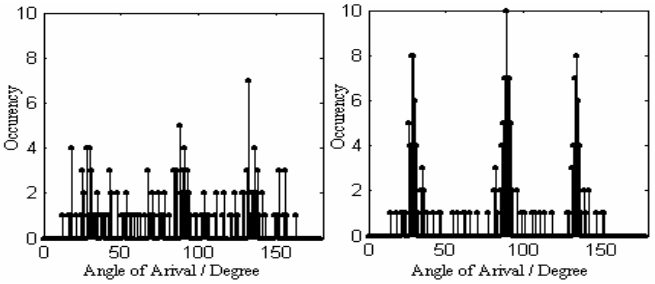

Fig. 4. Effect of warping on the DOA for convolutive mixtures; (a) before warping and (b) after warping (b=0.5).

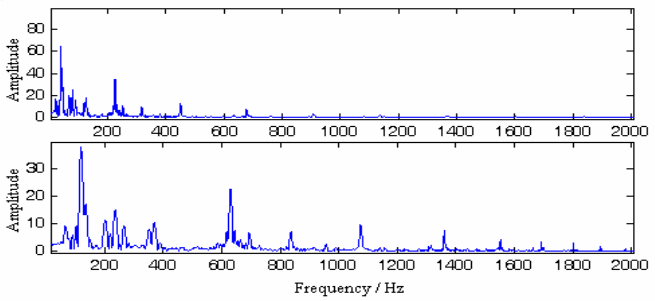

Fig. 5. The effect of warping on the spectrum; top plot refers to the case without warping and bottom one shows the case with warping (b=0.5).

5. EXPERIMENTAL RESULTS

5.1. FDBSS and the Permutation Problem:

To see the effect of frequency warping on mitigation of permutation problem the source signals are downloaded from the website http://medi.uni-oldenburg.de. Both signals are sampled at 12 kHz. The samples are 16-bit 2's complement in little endian format. The blocks are considered to be 512 samples. In the first experiment the sources are mixed using

$$H = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} \quad (9)$$

where $H_{11}(z) = 1+1.0\,z^{-1} - 0.75\,z^{-2}$, $H_{21}(z) = -0.7\,z^{-5} - 0.3\,z^{-6} + 0.2\,z^{-7}$, $H_{12}(z) = 0.5\,z^{-5} + 0.3\,z^{-6}\,0.2\,z^{-7}$, $H_{22}(z) = 0.8 - 0.1\,z^{-1}$. The frame lengths are set to 512 samples. The weights are initialised as $W_0(\omega)=I$, and μ =1, η =0 and λ = 0.01. The results are analysed, by comparing the error, $\varepsilon^2 = E[\|\mathrm{y}-\mathrm{s}\|^2]$, for two different trials: (I) when the FDBSS is applied without warping, (II) the same as (I) but with warping the mixtures (b=-0.5), (III) when FDBSS followed by mitigation of the permutation and estimation of the DOA is performed, and (IV) the same as (III) but with warping the mixtures. Table 1 illustrates the results

respectively. Clearly, warping followed by DOA-based method mitigates the permutation problem more effectively.

Table 1. Estimation error for the conventional BSS and for when the signals are warped before BSS:

ε^2	No Warping	With Warping (b=-0.5)
BSS	-18.1 dB	-19.8 dB
BSS and DOA	-20.3 dB	-22.5 dB

For the real room recording the microphone music sounds are downloaded from http://www.esp.ele.tue.nl/. The room size was a $3.4 \times 3.8 \times 5.2$ m^3, and the microphones spaced 58 cm apart. The sampling frequency and the bitrate were 12 kHz and 16 bits/sample respectively. In a subjective comparison, eight out of ten trained listeners verified the improvement achieved as a result of application of the warping to reduce the permutation problem.

5.2. Effect of Warping on SCA

Here we assume b to be different for different observed signal, i.e. $\mathbf{b}=[b_1, b_2, \ldots, b_n]^T$ where n is equal to the number of mixtures and ^{T} denotes transpose operation. The objective is then minimising the l_0-norm. In this part we considered three musical signals (playing the same piece of music) mixed using

$$H = \begin{bmatrix} H_{11} & H_{12} & H_{13} \\ H_{21} & H_{22} & H_{23} \end{bmatrix} \qquad (10)$$

to make two mixtures, where $H_{11}(z) = 1+1.0\, z^{-1} - 0.75\, z^{-2}$, $H_{21}(z) = -0.7\, z^{-5} - 0.3\, z^{-6} + 0.2\, z^{-7}$, $H_{12}(z) = 0.5\, z^{-5} + 0.3\, z^{-6}\, 0.2\, z^{-7}$, $H_{22}(z) = 0.8 - 0.1\, z^{-1}$, $H_{13}(z) = 0.4\, z^{-5} - 0.3\, z^{-6} + 0.2\, z^{-7}$, $H_{23}(z) = 0.7 - 0.2\, z^{-1} + 0.1 z^{-3}$. In the first trial the method explained in part 4 has been used for separation of the first source and the ICA-based BSS method given in [4] has been used for separation of the other two sources. The extraction range parameter, Δ, has been set to 6^0. In the second trial **b** has been iteratively computed to have minimum l_0-norm and the mixtures were warped. Then the method in the first trial was applied. The results are given in Table 2. For the real room recording two mixtures of three music signals have been used. Similar sampling frequency, number of bits/sample, and b has been used. In a subjective comparison, nine out of ten trained listeners verified the improvement achieved as a result of application of the warping to increase the sparseness.

6. SUMMARY AND CONCLUSIONS

In this paper the effect of warping has been investigated specially in the context of SCA and we demonstrated that the performance of the system is enhanced when the signals are warped before the separation process. In fact, in SCA the warping parameters can be adaptively calculated to have the maximum sparseness. Also, the inherent permutation problem has been significantly mitigated since the angle of arrivals is better estimated. It has been illustrated that frequency warping increases the sparseness of the mixtures in the underdetermined cases. As a result, separation of the signals become more accurate and the musical noise decreases. Optimization of the algorithm is dependent on the adaptation process for computation of the warping parameters.

Table 2. The estimation error for the conventional SCA and when the signals are warped before SCA:

ε^2	Without Warping	With Warping (**b** computed iteratively)
SCA	-15.4 dB	-17.8 dB

REFERENCES

[1] C. Braccini and A. V. Oppenheim, "Unequal bandwidth spectral analysis using digital frequency warping," *IEEE Trans. on ASSP, vol. ASSP-22, pp. 236-244, 1974.*

[2] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," *IEEE Trans. SAP, vol. 6, No. 1, pp. 49-59, 1998.*

[3] G. Evangelista and S. Cavaliere, "Frequency-warped filter banks and wavelet transforms: a discrete-time approach via Laguerre expansion," *IEEE Trans. on SP, vol. 46, No. 10, pp. 2638-2649, 1998.*

[4] W. Wang, J. A. Chambers, and S. Sanei, "A joint diagonalization method for convolutive blind separation of nonstationary sources in the frequency domain," *Proc. of ICA2003, Nara, Japan, pp. 939-944, 2003.*

[5] H. Sawada et al., "A robust and precise method for solving the permutation problem of frequency domain blind source separation," *IEEE Trans. on SAP, vol. 12, no. 5, pp. 530-538,Sept.2004.*

[6] S. Sanei, W. Wang, and J. Chambers, "A coupled HMM for solving the permutation problem in frequency domain BSS," *Proc. IEEE, ICASSP, V, pp. 565-568, 2004.*

[7] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing, vol. 22, pp. 21-34, 1998.*

[8] V. Capdevielle, C. Serviere, and J. L. Lacoume, "Blind separation of wide-band sources in the frequency domain," *Proc. ICASSP95, pp. 2080-2083, 1995.*

[9] J. Anemuller and B. Kollmeier, "Amplitude modulation decorrelation for convolutive blind source separation," *Proc. ICA2000, Helsinki, Finland, pp. 215-220, June 2000.*

[10] S. Araki et al, "Underdetermined blind separation for speech in real environments with sparseness and ICA," *Proc. IEEE, ICASSP, vol. III, pp. 881-884, 2004.*

[11] P. Bofil and M. Zibulevsky, "Underdetermined blind source separation using sparse representations," *Signal Processing, 81, pp. 2353-2362, 2001.*