Unsupervised Single-Channel Separation of Nonstationary Signals Using Gammatone Filterbank and Itakura–Saito Nonnegative Matrix Two-Dimensional Factorizations

Bin Gao, *Member, IEEE*, W. L. Woo, *Senior Member, IEEE*, and S. S. Dlay

***Abstract*—A new unsupervised single-channel source separation method is presented. The proposed method does not require training knowledge and the separation system is based on nonuniform time–frequency (TF) analysis and feature extraction. Unlike conventional researches that concentrate on the use of spectrogram or its variants, we develop our separation algorithms using an alternative TF representation based on the gammatone filterbank. In particular, we show that the monaural mixed audio signal is considerably more separable in this nonuniform TF domain. We also provide the analysis of signal separability to verify this finding. In addition, we derive two new algorithms that extend the recently published Itakura–Saito nonnegative matrix factorization to the case of convolutive model for the nonstationary source signals. These formulations are based on the Quasi-EM framework and the multiplicative gradient descent (MGD) rule, respectively. Experimental tests have been conducted which show that the proposed method is efficient in extracting the sources' spectral-temporal features that are characterized by large dynamic range of energy, and thus leading to significant improvement in source separation performance.**

***Index Terms*—Gammatone filterbank, Itakura–Saito divergence, matrix factorization, nonstationary source separation.**

I. Introduction

UNSUPERVISED source separation (USS) of multiple sources from multiple sensors is a sophisticated research field with numerous applications in the fields of neural computation, encryption [1], security [2] and pattern detection [3]. A review of current literature shows that there are three main classifications of USS [4], [5]. These include linear and nonlinear, instantaneous and convolutive, overcomplete, and underdetermined. In the first classification, linear algorithms dominate the USS research field due to its simplicity in analysis and its explicit separability. Linear USS assumes that the mixture is represented by a linear combination of sources. Extension of USS for solving nonlinear mixtures has also been introduced [6]. This model takes nonlinear distorted signals into consideration offers a more accurate representation of a realistic environment. In the second classification, when the observed signals consist of combinations of multiple time-delayed versions of the original sources and/or mixed signals themselves, the system is referred as the convolutive mixture [5]. Otherwise, the absence of time delays results in the instantaneous mixture of observed signals. Finally, when the number of observed signals exceeds the number of sources, this refers to the overcomplete USS. Conversely, when the number of observed signals is less than the number of sources, this becomes the underdetermined USS. In general and for many practical applications, the challenging case for source separation is when only one monaural recording is available. This leads to the unsupervised single-channel source separation (USCSS) where the problem can be stated as one observation mixed with several unknown sources. In this work, we consider the case of two sources, namely,

$$y(t) = x_1(t) + x_2(t) \tag{1}$$

where $t = 1, 2, \ldots, T$ denotes time index and the goal is to estimate the two sources $x_1(t)$ and $x_2(t)$ given only the observation signal $y(t)$. Unlike conventional assumption used in USS where the sources are assumed to be statistical independent which is rather too restrictive; in this paper, the sources are characterized as non-stationary processes with time-varying spectra [7]. This assumption is practically justified since most signals encountered in applications are non-stationary with time-varying spectra.

Solutions to USCSS using nonnegative matrix factorization (NMF) [8] have recently gained popularity. They exploit an appropriate time-frequency (TF) analysis on the mono input recording, yielding a TF representation as

$$|\mathbf{Y}|^{\cdot 2} \approx \mathbf{DH} \tag{2}$$

where $|\mathbf{Y}|^{\cdot 2} \in \Re_+^{F \times T_s}$ is the power time–frequency (TF) representation of the mixture $y(t)$ which is factorized as the product of two nonnegative matrices, $\mathbf{D} \in \Re_+^{F \times I}$ and $\mathbf{H} \in \Re_+^{I \times T_s}$. The superscript "$\cdot$" represents element wise operation. F and T_s represent the total frequency units and time slots in the TF domain, respectively. If I is chosen to be $I = T_s$, no benefit is achieved in terms of representation. Thus, the idea is to determine $I < T_s$ so the matrix $\mathbf{D}$ can be compressed and reduced

Manuscript received September 02, 2011; revised November 28, 2011 and February 01, 2012; accepted February 28, 2012. Date of publication October 02, 2012; date of current version February 21, 2013. This paper was recommended by Associate Editor C.-C. Tseng.

The authors are with School of Electrical and Electronic Engineering, Newcastle University, Newcastle upon Tyne NE1 7RU, U.K. (e-mail: w.l.woo@ncl.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSI.2012.2215735

to its integral components such as it contains only a set of spectral basis vectors, and $\mathbf{H}$ is an encoding matrix which describes the amplitude of each basis vector at each time point. Because NMF gives a parts-based decomposition [8], [9], it has recently been proposed for separating drums from polyphonic music [10] and automatic transcription of polyphonic music [11]. Commonly used cost functions for NMF are the generalized Kullback–Leibler (KL) divergence and least square (LS) distance [8]. A sparseness constraint [12] can be added to these cost functions for optimizing $\mathbf{D}$ and $\mathbf{H}$. Other cost functions for audio spectrograms factorization have also been introduced such as that of [13] which assumes multiplicative gamma-distributed noise in power spectrograms, while [14] attempts to incorporate phase into the factorization by using a probabilistic phase model. Notwithstanding above, families of parameterized cost functions, such as the Beta divergence [15] and Csiszar's divergences [16] have also been presented for the source separation. However, they have some crucial limitations that explicitly use training knowledge of the sources [17]. As a consequence, these methods are only able to deal with a very specific set of signals and situations.

Model-based techniques have also been proposed for SCSS which usually require training a set of isolated recordings. The sources are trained by using a hidden Markov model (HMM) based on Gaussian mixture model (GMM) and they are combined in a factorial HMM to separate the mixture [18]. Good separation requires detailed source models that might use thousands of full spectral states, e.g., in [19], HMMs with 8000 states were required to accurately represent one person's speech for a source separation task. The large state space is required because it attempts to capture every possible instance of the signal. These model-based techniques, however, consume long time not only in training the prior parameters but also presenting many difficult challenges during the inference stage.

From above, existing solutions to USCSS are still practically limited and fall short of the success enjoyed in other areas of source separation. In this paper, a novel separation system is proposed and the contributions are summarized as follows:

1) Derivation of a separability analysis in the TF domain for USCSS and development a quantitative performance measure to evaluate the degree of "separateness" in the monaural mixed signal. In particular, we have identified the ideal condition when the sources are perfectly separable.
2) Novel development of a separation framework based on the gammatone filterbank. Unlike the spectrogram which deals only with uniform resolution, the gammatone filterbank produces nonuniform TF domain (termed as the cochleagram) whereby each TF unit has different resolution. We prove that the mixed signal is significantly more separable in the cochleagram than the spectrogram and the log-frequency spectrogram (constant-Q transform).
3) Novel development of two-dimensional NMF (NMF2D) signal model optimized under the Itakura–Saito (IS) divergence with Quasi-EM and MGD updates. We term this model as the IS-NMF2D. Two new algorithms have been developed to estimate the spectral and temporal features of the signal model. The first algorithm is founded on the framework of Quasi-EM (Expectation–Maximization) while the second algorithm is based on the multiplicative gradient decent (MGD) update rule. Both algorithms have the unique property of scale-invariant whereby the lower energy components in the TF domain can be treated with equal importance as the higher energy components. This property is highly desirable since it enables the spectral and temporal features of the non-stationary sources which are usually characterized by large dynamic range of energy to be estimated with significantly higher accuracy. This is to be contrasted with other methods based on LS distance [20] and KL divergence [21] which favor the high-energy components but neglect the low-energy components.

The paper is organized as follows: Section II introduces the TF matrix representation using the gammatone filterbank. Section III delves into the separability analysis of the single-channel mixture in the non-uniform TF domain. In Section IV, the two new algorithms are derived and the proposed separation system is developed. Experimental results and a series of performance comparison with methods are presented in Section V. Finally, Section VI concludes the paper.

II. Time-Frequency Representation

In the task of audio source separation, one critical decision is to choose a suitable TF domain to represent the time-varying contents of the signals. There are several types of TF representations and the most widely used ones are spectrogram [22] and log-frequency spectrogram (using constant-Q transform) [23]. This is documented over the last few years in the research of audio source separation [10]–[21]. In this work, however, we develop our separation algorithms using a TF representation based on the gammatone filterbank.

A. Gammatone Filterbank and Cochleagram

The Gammatone filterbank [24] is a cochlear filtering model which decomposes an input signal into the time–frequency domain using a set of gammatone filters. The impulse response of a gammatone filter centered at frequency f is mathematically expressed as

$$g(f,t)=\begin{cases} t^{h-1}e^{-2\pi v t}\cos(2\pi f t), & t\geq 0\\ 0, & \text{else}\end{cases} \tag{3}$$

where h denotes the order of filter, v represents the rectangular bandwidth which increases as the center frequency f increases. Considering a particular filter channel c with f_c as the center frequency, the filter output response $x(c,t)$ can be expressed as

$$x(c,t)=x(t)*g(f_c,t) \tag{4}$$

where "$*$" denotes time-domain convolution. The response is shifted backwards by $(h-1)/(2\pi v)$ to compensate for the filter delay. The output of each filter channel is divided into time frames with 50% overlap between consecutive frames. The time–frequency spectra of all the filter outputs are then constructed to form the cochleagram.

Fig. 1(a) shows the cochleagram of female speech mixed with jazz in which 128-channel gammatone filterbank is used

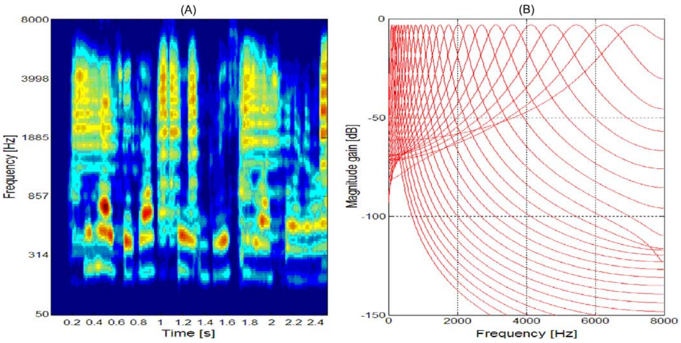

Fig. 1. (a) Cochleagram of a female utterance mixed with jazz music. (b) Frequency response of a gammatone filterbank.

over the frequency range from 50 to 8000 Hz. Also note that the time-varying spectrum of the signal is clearly visible. In [25] and [26], it was noted that some crucial differences exist in the TF representation of how sound is analyzed by the ear. In particular, the ear's frequency subbands get wider for higher frequencies whereas the classical spectrogram as computed by the short-time Fourier transform (STFT) has an equal-spaced bandwidth across all frequency channels. Since speech signals are characterized as highly nonstationary and nonperiodic whereas music changes continuously; therefore, application of the Fourier transform will produce errors when complicated transient phenomena such as the mixture of speech and music is contained in the analyzed signal. Unlike the spectrogram, the log-frequency spectrogram possesses nonuniform TF resolution. However, it does not exactly match to the nonlinear resolution of the cochlear since their center frequencies are distributed logarithmically along the frequency axis and all filters have constant-Q factor [23]. On a separate hand, the gammatone filters used in the cochlear model (3) are approximately *logarithmically* spaced with constant-Q for frequencies from $f_s/10$ to $f_s/2$ (f_s denotes the sampling frequency), and approximately *linearly* spaced for frequencies below $f_s/10$. Hence, this characteristic results in selective *non-uniform* resolution in the TF representation of the analyzed audio signal. Fig. 1(b) shows the frequency response of a general gammatone filterbank for $f_s = 16$ kHz. It is seen that the higher frequencies correspond to the wider frequency subbands which resemble closely to the human perception of frequencies [27]. Therefore, the cochleagram is developed as an alternative TF analysis tool for source separation to overcome the limitations associated with the Fourier transform approach.

III. Single-Channel Source Separability Analysis

For separation, one generates the TF mask corresponding to each source and applies the generated mask to the mixture to obtain the estimated source TF representation. In particular, when the sources do not overlap in the TF domain, an optimum mask $M_i^{opt}(f,t_s)$ exists which allows one to extract the $i^{\rm th}$ original source from the mixture as

$$X_i(f,t_s) = M_i^{opt}(f,t_s)Y(f,t_s). \tag{5}$$

Given any TF mask $M_i(f,t_s)$ such that $0 \leq M_i(f,t_s) \leq 1$ for all (f,t_s), we define the separability for the target source $x_i(t)$ in the presence of the interfering sources $p_i(t) = \sum_{j=1,j\neq i}^{N} x_j(t)$ as

$$S_{M_i}^{Y\to X_i,P_i} = \frac{\|M_i(f,t_s)X_i(f,t_s)\|_F^2}{\|X_i(f,t_s)\|_F^2} - \frac{\|M_i(f,t_s)P_i(f,t_s)\|_F^2}{\|X_i(f,t_s)\|_F^2} \tag{6}$$

where $X_i(f,t_s)$ and $P_i(f,t_s)$ are the TF representations of $x_i(t)$ and $p_i(t)$, respectively. $\|.\|_F$ is the Frobenius norm. We also define the separability of the mixture with respect to all the N sources as

$$S_{M_1,\ldots,M_N}^{Y\to X_1,\ldots,X_N} = \frac{1}{N}\sum_{i=1}^{N} S_{M_i}^{Y\to X_i,P_i}. \tag{7}$$

Equation (6) is equivalent to measuring the success of extracting the $i_{\rm th}$ source $X_i(f,t_s)$ from the mixture $Y(f,t_s)$ given the TF mask $M_i(f,t_s)$. Similarly, (7) measures the success of extracting all the N sources simultaneously from the mixture. To further analyze the separability, we invoke the followings: 1) Preserved signal ratio (PSR) which determines how well the mask preserves the source of interest, and 2) signal-to-interference ratio (SIR) which indicates how well the mask suppresses the interfering sources:

$$\begin{aligned} PSR_{M_i}^{X_i} &= \frac{\|M_i(f,t_s)X_i(f,t_s)\|_F^2}{\|X_i(f,t_s)\|_F^2} \\ SIR_{M_i}^{X_i} &= \frac{\|M_i(f,t_s)X_i(f,t_s)\|_F^2}{\|M_i(f,t_s)P_i(f,t_s)\|_F^2}. \end{aligned} \tag{8}$$

Using (8), it can be shown that (7) can be expressed as $S_{M_i}^{Y\to X_i,P_i} = PSR_{M_i}^{X_i} - PSR_{M_i}^{X_i}/SIR_{M_i}^{X_i}$. Analyzing the terms in (6), we have

$$\begin{aligned} PSR_{M_i}^{X_i} &:= \begin{cases} 1, & \text{if supp } M_i^{opt} = \text{supp} M_i \\ <1, & \text{if supp } M_i^{opt} \subset \text{supp} M_i \end{cases} \\ SIR_{M_i}^{X_i} &:= \begin{cases} \infty, & \text{if supp } [M_i X_i] \cap \text{supp} P_i = \varnothing \\ finite, & \text{if supp } [M_i X_i] \cap \text{supp} P_i \neq \varnothing \end{cases} \end{aligned} \tag{9}$$

where "supp" denotes the support. When $S_{M_i}^{Y\to X_i,P_i} = 1$ (i.e., $PSR_{M_i}^{X_i} = 1$ and $SIR_{M_i}^{X_i} = \infty$), this indicates that the mixture $y(t)$ is separable with respect to the $i^{\rm th}$ source $x_i(t)$. In other words, $X_i(f,t_s)$ does not overlap with $P_i(f,t_s)$ and the TF mask $M_i(f,t_s)$ has perfectly separated the $i^{\rm th}$ source $X_i(f,t_s)$ from the mixture $Y(f,t_s)$. This corresponds to $M_i(f,t_s) = M_i^{opt}(f,t_s)$ in (5). Hence, this is the maximum attainable $S_{M_i}^{Y\to X_i,P_i}$ value. For other cases of $PSR_{M_i}^{X_i}$ and $SIR_{M_i}^{X_i}$, we have $S_{M_i}^{Y\to X_i,P_i} < 1$. Using this concept, we can extend the analysis for the case of separating N sources. A mixture $y(t)$ is fully separable to all the N sources if and only if $S_{M_1,\ldots,M_N}^{Y\to X_1,\ldots,X_N} = 1$ in (7). For the case $S_{M_1,\ldots,M_N}^{Y\to X_1,\ldots,X_N} < 1$, this implies that some of the sources overlap with each other in the TF domain and therefore, they cannot be fully separated. Thus, $S_{M_1,\ldots,M_N}^{Y\to X_1,\ldots,X_N}$ provides the quantitative performance measure for evaluating how separable is the mixture in the TF domain. In our comparison, the following TF representations are used to test the mixture's separability: spectrogram, log-frequency spectrogram and cochleagram. In the log-frequency spectrogram, the frequency scale is set to logarithmic and grouped into 175 frequency bins in the range of 50 Hz to 8 kHz with

TABLE I
AVERAGED SEPARABILITY PERFORMANCE

Types of TF domain	Mixtures	PSR	SIR	$S^{Y\to X_1,X_2}_{M_1,M_2}$
Cochleagram	M and M	0.996	275.8	0.993
	M and S	0.995	186.8	0.989
	S and S	0.984	184.2	0.979
Log-frequency spectrogram	M and M	0.958	165.5	0.953
	M and S	0.942	118.5	0.947
	S and S	0.943	20.2	0.934
Spectrogram	M and M	0.885	55.8	0.869
	M and S	0.882	53.6	0.865
	S and S	0.871	50.83	0.854

TABLE II
SEPARABILITY UNDER DIFFERENT WINDOW LENGTH

Types of TF domain	Window Length	$S^{Y\to X_1,X_2}_{M_1,M_2}$
Cochleagram	20ms (320)	**0.985**
	32ms (512)	0.972
	64ms (1024)	0.965
	128ms (2048)	0.892
Log-frequency spectrogram	20ms (320)	0.813
	32ms (512)	0.874
	64ms (1024)	**0.948**
	128ms (2048)	0.912
Spectrogram	20ms (320)	0.801
	32ms (512)	0.834
	64ms (1024)	**0.864**
	128ms (2048)	0.842

24 bins per octave while the bandwidth follows the constant-Q rule [23]. To ensure fair comparison, we generate the ideal binary mask (IBM) [27] directly from the original sources. To reiterate our aim, the separability analysis is undertaken without recourse to any separation algorithms but utilizing only the energy of the sources to ascertain the degree of "separateness" of the mixture in different TF domains. These results have been tabulated in Table I. The symbols "M" and "S" denotes music and speech, respectively.

In Table I, three types of mixture have been used: 1) music mixed with music, 2) speech mixed with music, and 3) speech mixed with speech. The speech signals are selected from ten male and ten female speeches taken from TIMIT database and are normalized to unit energy. The ten music sources are selected from the RWC database [28] and also normalized to unit energy. Two sources are randomly chosen from the databases and the mixed signal is generated by adding the sources. All mixed signals are sampled at 16-kHz sampling rate. TF representation using different window length has also been investigated and the results are tabulated in Table II.

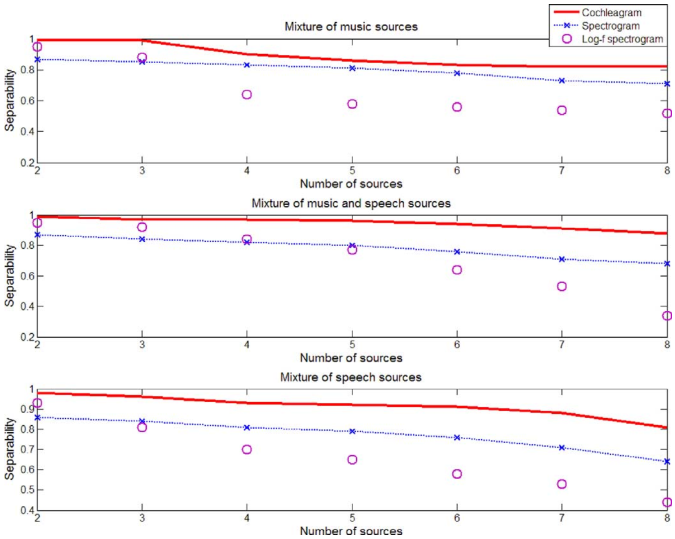

Fig. 2. Overall separability performance for each mixture type.

Table II shows the average separability results for all types of the mixture based on different window length. The bracketed number shows the number of data points corresponding to the particular window length. It is clear that, for both spectrogram and log-frequency spectrogram settings, the STFT with 1024-point window length is the best setting to analyze the separability performance. The results of PSR, SIR and separability for each TF domain are obtained by averaging over 300 realizations. Following the listening performance test proposed in [29], we conclude that $S^{Y\to X_i,P_i}_{M_i} > 0.8$ leads to acceptable separation performance. Therefore, all TF representations in Table I satisfy this condition. While this is true, the spectrogram gives only a mediocre level of separability with averaged $S^{Y\to X_1,X_2}_{M_1,M_2} \approx 0.86$ while the log-frequency spectrogram shows a better result with $S^{Y\to X_1,X_2}_{M_1,M_2} \approx 0.94$. Nevertheless, the cochleagram yields the best separability with $S^{Y\to X_1,X_2}_{M_1,M_2} \approx 0.98$. Notwithstanding this, it is also seen that the average SIR of the cochleagram exhibits a much higher value than those of spectrogram and log-frequency spectrogram. This implies that the amount of interference between any two sources is lesser in the cochleagram.

The analysis conducted above is based on a mixture of two sources. In the following, we extend the separability analysis by increasing the number of sources from 2 to 8. For mixture of music and speech sources, the number of music sources is selected equal to the number of speech sources (e.g., for mixture of 8 sources, 4 are drawn from music and another 4 from speech; for mixture of 7 sources, either 3 (or 4) are drawn from music and the remaining 4 (or 3) from speech). The result is shown in Fig. 2. Similar to above, the separability performance for each TF representation is obtained by averaging over 300 realizations. It is observed that for all number of sources, the cochleagram has shown the best separability performance across all different types of mixture. It is worth pointing out that the cochleagram always retain a high level of separability even when the number of sources increases. Also, the curve of separability decreases steadily as the number of sources increases. On the contrary, other TF representations fail to separate the mixture when large number of sources is present, e.g., for mixture of music and speech (8 sources mixed), $S^{Y\to X_1,\ldots,X_N}_{M_1,\ldots,M_N} \approx 0.65$ for spectrogram and $S^{Y\to X_1,\ldots,X_N}_{M_1,\ldots,M_N} \approx 0.3$ for log-frequency spectrogram. They are considerably below the acceptable level of separability. On the other hand, the cochleagram maintains at a good level with $S_{M_1,\ldots,M_N}^{Y \to X_1,\ldots,X_N} \approx 0.9$ which is well above the rest. It is noted that the curve of separability for the log-frequency spectrogram decreases very sharply as number of sources increases. In Table I, it is shown that the log-frequency spectrogram leads to better separability than the classic spectrogram. However, this is not always the case especially when the number of sources in the mixture is increased from four onwards. The curves in Fig. 2 indicate that the separability of the spectrogram degrades more gracefully as compared with the log-frequency spectrogram. Finally, of all the mixture types only the cochleagram preserves the separability larger than 0.8 over the range of eight sources. Therefore, based on this study, it can be concluded that the cochleagram is a better TF transform than spectrogram and log-frequency spectrogram.

IV. Proposed Algorithms

In this section, two new algorithms will be developed, namely the *Quasi-EM IS-NMF2D* and the *MGD IS-NMF2D*. The former algorithm optimizes the parameters of the signal model using the expectation–maximization approach whereas the latter is directly based on the multiplicative gradient descent. To facilitate the derivation of these algorithms, we first consider the signal model in terms of the power TF representation

A. Signal Models

Since the sources have time-varying spectra, it is befitting to adopt a model whose power spectra can be described separately in terms of time and frequency. Although conventional NMF model can still be used, it will need large number spectral components and requires a clustering step to group and assign each spectral component to the appropriate source. As a result, the NMF model may not always yield the optimal results. An alternative model is to use the two-dimensional NMF model (NMF2D) [33], [34]. This model extends the basic NMF to be a two-dimensional convolution of $\mathbf{D}$ and $\mathbf{H}$ i.e., $|\mathbf{Y}|^{\cdot 2} \approx \sum_{\tau,\phi} \overset{\downarrow\phi}{\mathbf{D}^{\tau}} \overset{\rightarrow\tau}{\mathbf{H}^{\phi}}$ where the vertical arrow in $\overset{\downarrow\phi}{\mathbf{D}^{\tau}}$ denotes the downward shift which moves each element in the matrix down by ϕ rows, and the horizontal arrow in $\overset{\rightarrow\tau}{\mathbf{H}^{\phi}}$ denotes the right shift operator which moves each element in the matrix to the right by τ columns. In scalar representation, the $(f,t_s)^{\text{th}}$ element in $|\mathbf{Y}|^{\cdot 2}$ is given by $|\mathbf{Y}_{f,t_s}|^2 \approx \sum_{i=1}^{I} \sum_{\tau=0}^{\tau_{\max}} \sum_{\phi=0}^{\phi_{\max}} \mathbf{D}_{f-\phi,i}^{\tau} \mathbf{H}_{i,t_s-\tau}^{\phi}$ where $\mathbf{D}_{f',i'}^{\tau'}$ is the $(f',\tau',i')^{\text{th}}$ element of $\mathbf{D}$ and $\mathbf{H}_{i',t_s'}^{\phi'}$ is the $(i',\phi',t_s')^{\text{th}}$ element of $\mathbf{H}$. In source separation, this model compactly represents the characteristics of the nonstationary sources by a time–frequency profile convolved in both time and frequency by a time–frequency weight matrix. $\mathbf{D}_i^{\tau}$ represents the spectral basis of i^{th} source in the TF domain and $\mathbf{H}_i^{\phi}$ represents the corresponding temporal code for each spectral basis.

The TF representation of the mixture in (1) is given by $Y(f,t_s) = X_1(f,t_s) + X_2(f,t_s)$ where $Y(f,t_s)$, $X_1(f,t_s)$ and $X_2(f,t_s)$ denote the TF components which are obtained by applying the gammatone filterbank to the mixture. The time slots are given by $t_s = 1,2,\ldots,T_s$ while frequencies by $f = 1,2,\ldots,F$. Since each component is a function of t_s and f, we represent this as a $F \times T_s$ matrix $\mathbf{Y} = [Y(f,t_s)]_{t_s=1,2,\ldots,T_s}^{f=1,2,\ldots,F}$ and $\mathbf{X}_i = [X_i(f,t_s)]_{t_s=1,2,\ldots,T_s}^{f=1,2,\ldots,F}$. It is shown in Section III that the sources are almost perfectly separable in the cochleagram. This therefore enable us to express the power TF representation as $|\mathbf{Y}|^{\cdot 2} \approx \sum_{i=1}^{I} |\mathbf{X}_i|^{\cdot 2}$ which we will model as $|\mathbf{Y}_{f,t_s}|^2 \approx \sum_{i=1}^{I} \sum_{\tau=0}^{\tau_{\max}} \sum_{\phi=0}^{\phi_{\max}} \mathbf{D}_{f-\phi,i}^{\tau} \mathbf{H}_{i,t_s-\tau}^{\phi}$. The source we seek to determine are $\{|X_i(f,t_s)|^{\cdot 2}\}_{i=1}^{I}$ and this will be obtained by using the matrix factorization as $|\tilde{X}_i(f,t_s)|^{\cdot 2} = \sum_{\tau=0}^{\tau_{\max}} \sum_{\phi=0}^{\phi_{\max}} \mathbf{D}_{f-\phi,i}^{\tau} \mathbf{H}_{i,t_s-\tau}^{\phi}$. In the following, we propose two novel algorithms to estimate $\mathbf{D}_{f,i}^{\tau}$ and $\mathbf{H}_{i,t_s}^{\phi}$ from the mixture signal.

B. Algorithm 1: Quasi-EM Formulation of IS-NMF2D (Quasi-EM IS-NMF2D)

We consider the following generative model defined as

$$\begin{aligned} \mathbf{y}_{t_s} &= \sum_{k=1}^{K} \mathbf{c}_{k,t_s}, \quad \forall t_s = 1,\ldots,T_s \\ \mathbf{c}_{k,t_s} &= [c_{k,1,t_s},\ldots,c_{k,F,t_s}]^{\mathbf{T}} \\ c_{k,f,t_s} &\sim N_c\left(0, \sum_{\tau,\phi} \mathbf{H}_{k,t_s-\tau}^{\phi} \mathbf{D}_{f-\phi,k}^{\tau}\right) \end{aligned} \tag{10}$$

where $\mathbf{y}_{t_s} \in \mathbb{C}^{F\times 1}$, $\mathbf{c}_{k,t_s} \in \mathbb{C}^{F\times 1}$ and $N_c(u,\Sigma)$ denotes the proper complex Gaussian distribution and the components $\mathbf{c}_{1,t_s},\ldots,\mathbf{c}_{K,t_s}$ are both mutually and individually independent. The EM framework is developed for the ML estimation of $\boldsymbol{\theta} = \{\mathbf{D}^{\tau},\mathbf{H}^{\phi}\}$. Due to the additive structure of the generative model (10), the parameters describing each component $\mathbf{C}_k = [\mathbf{c}_{k,1},\ldots,\mathbf{c}_{k,T_s}]$ can be updated separately. We now consider a partition of the parameter space $\boldsymbol{\theta} = \bigcup_{k=1}^{K} \boldsymbol{\theta}_k$ as $\boldsymbol{\theta}_k = \{\mathbf{D}_k^{\tau},\mathbf{H}_k^{\phi}\}$ where $\mathbf{D}_k^{\tau}$ is the k^{th} column of $\mathbf{D}^{\tau}$ and $\mathbf{H}_k^{\phi}$ is the k^{th} row of $\mathbf{H}^{\phi}$. The EM algorithm works by formulating the conditional expectation of the negative log likelihood of $\mathbf{C}_k$ as

$$Q_k^{ML}(\boldsymbol{\theta}_k|\boldsymbol{\theta}') = -\int_{\mathbf{C}_k} p\left(\mathbf{C}_k|\mathbf{Y},\boldsymbol{\theta}'\right) \log p(\mathbf{C}_k|\boldsymbol{\theta}_k) d\mathbf{C}_k \tag{11}$$

where $\boldsymbol{\theta}'$ always contains the most recent parameter values of $\{\mathbf{D}^{\tau},\mathbf{H}^{\phi}\}$.

1) Expressions of the E- and M-step: One iteration of the EM algorithm includes computing the E-step and maximizing the M-step $Q_k^{ML}(\boldsymbol{\theta}_k|\boldsymbol{\theta}')$ for $k = 1,\ldots,K$. The minus hidden-data log likelihood is defined as

$$\begin{aligned} -\log p(\mathbf{C}_k|\boldsymbol{\theta}_k) &= -\sum_{t_s=1}^{T_s} \sum_{f=1}^{F} \log N_c \\ &\quad \times \left(c_{k,f,t_s} \middle| 0, \sum_{\tau,\phi} \mathbf{D}_{f-\phi,k}^{\tau} \mathbf{H}_{k,t_s-\tau}^{\phi}\right) \\ &\doteq \sum_{t_s=1}^{T_s} \sum_{f=1}^{F} \log\left(\sum_{\tau,\phi} \mathbf{D}_{f-\phi,k}^{\tau} \mathbf{H}_{k,t_s-\tau}^{\phi}\right) \\ &\quad + \frac{|c_{k,f,t_s}|^2}{\sum_{\tau,\phi} \mathbf{D}_{f-\phi,k}^{\tau} \mathbf{H}_{k,t_s-\tau}^{\phi}} \end{aligned} \tag{12}$$

where "$\doteq$" in the second line denotes equality up to constant terms. Then, by virtue of (10), the hidden-data posterior also has a Gaussian form as $p(\mathbf{C}_k|\mathbf{Y},\boldsymbol{\theta}) = \prod_{t_s=1}^{T_s}\prod_{f=1}^{F} N_c(c_{k,f,t_s}|u^{post}_{k,f,t_s}\lambda^{post}_{k,f,t_s})$ where u^{post}_{k,f,t_s} and λ^{post}_{k,f,t_s} are the posterior mean and variance of c_{k,f,t_s} given as

$$\begin{aligned} u^{post}_{k,f,t_s} &= \frac{\sum\limits_{\tau,\phi} \mathbf{D}^{\tau}_{f-\phi,k}\mathbf{H}^{\phi}_{k,t_s-\tau}}{\sum\limits_{\tau,\phi,l} \mathbf{D}^{\tau}_{f-\phi,l}\mathbf{H}^{\phi}_{l,t_s-\tau}}\mathbf{Y}_{f,t_s} \\ \lambda^{post}_{k,f,t_s} &= \frac{\sum\limits_{\tau,\phi} \mathbf{D}^{\tau}_{f-\phi,k}\mathbf{H}^{\phi}_{k,t_s-\tau}}{\sum\limits_{\tau,\phi,l} \mathbf{D}^{\tau}_{f-\phi,l}\mathbf{H}^{\phi}_{l,t_s-\tau}} \sum_{\tau,\phi,l\neq k} \mathbf{D}^{\tau}_{f-\phi,l}\mathbf{H}^{\phi}_{l,t_s-\tau}. \end{aligned} \tag{13}$$

Thus, the E-step merely includes computing the posterior power $\mathbf{V}_k$ of component $\mathbf{C}_k$, defined as $[\mathbf{V}_k]_{f,t_s} = v_{k,f,t_s} = |u^{post}_{k,f,t_s}|^2 + \lambda^{post}_{k,f,t_s}$. The M-step can be treated as one-component NMF problem:

$$\begin{aligned} &Q^{ML}_k(\boldsymbol{\theta}_k|\boldsymbol{\theta}') \\ &\doteq \sum_{t_s=1}^{T_s}\sum_{f=1}^{F} \log\left(\sum_{\tau,\phi}\mathbf{D}^{\tau}_{f-\phi,k}\mathbf{H}^{\phi}_{k,t_s-\tau}\right) \\ &\quad + \frac{\left|u^{post'}_{k,f,t_s}\right|^2 + \lambda^{post'}_{k,f,t_s}}{\sum\limits_{\tau,\phi}\mathbf{D}^{\tau}_{f-\phi,k}\mathbf{H}^{\phi}_{k,t_s-\tau}} \\ &\doteq \sum_{t_s=1}^{T_s}\sum_{f=1}^{F} d_{IS} \\ &\quad \times \left(\left|u^{post'}_{k,f,t_s}\right|^2 + \lambda^{post'}_{k,f,t_s}\middle| \sum_{\tau,\phi}\mathbf{D}^{\tau}_{f-\phi,k}\mathbf{H}^{\phi}_{k,t_s-\tau}\right) \end{aligned} \tag{14}$$

where $d_{IS}(\cdot|\cdot)$ is the IS divergence [30] and is formally defined as $d_{IS}(a|b) = (a/b) - \log(a/b) - 1$. The IS divergence has the property of scale invariant, i.e., $d_{IS}(\kappa a|\kappa b) = d_{IS}(a|b)$ for any κ. This implies that any low energy components (a, b) will bear the same relative importance as the high energy ones $(\kappa a, \kappa b)$. This is particularly important to situations where $|\mathbf{Y}|^{\cdot 2}$ is characterized by large dynamic range such as the audio short-term spectra.

2) Estimation of the Spectral Basis and Temporal Code Using Quasi-EM Method: The spectral basis and temporal code can be obtained from (14). The derivative of a given element of $g_{k,f,t_s} = \sum_{\tau,\phi}\mathbf{D}^{\tau}_{f-\phi,k}\mathbf{H}^{\phi}_{k,t_s-\tau}$ with respect to $\mathbf{D}^{\tau}_{f,k}$ and $\mathbf{H}^{\phi}_{k,t_s}$ is given by

$$\begin{aligned} \frac{\partial g_{k,f,t_s}}{\partial \mathbf{D}^{\tau'}_{f',k'}} &= \frac{\partial \sum\limits_{\tau,\phi}\mathbf{D}^{\tau}_{f-\phi,k}\mathbf{H}^{\phi}_{k,t_s-\tau}}{\partial \mathbf{D}^{\tau'}_{f',k'}} = \mathbf{H}^{f-f'}_{k',t_s-\tau'} \\ \frac{\partial g_{k,f,t_s}}{\partial \mathbf{H}^{\phi'}_{k',t'_s}} &= \frac{\partial \sum\limits_{\tau,\phi}\mathbf{D}^{\tau}_{f-\phi,k}\mathbf{H}^{\phi}_{k,t_s-\tau}}{\partial \mathbf{H}^{\phi'}_{k',t'_s}} = \mathbf{D}^{t_s-t'_s}_{f-\phi',k'}. \end{aligned} \tag{15}$$

The derivatives of (14) corresponding to $\mathbf{D}^{\tau}_{f,k}$ and $\mathbf{H}^{\phi}_{k,t_s}$ is then obtained as

$$\begin{aligned} \frac{\partial Q^{ML}_k(\boldsymbol{\theta}_k|\boldsymbol{\theta}')}{\partial \mathbf{D}^{\tau'}_{f',k'}} &= \frac{\partial}{\partial \mathbf{D}^{\tau'}_{f',k'}} \sum_{f,t_s} \log\left(g_{k,f,t_s}\right) + \frac{v'_{k,f,t_s}}{g_{k,f,t_s}} \\ &= \sum_{\phi,t_s}\left(\frac{g_{k,f'+\phi,t_s} - v'_{k,f'+\phi,t_s}}{g^2_{k,f'+\phi,t_s}}\right)\mathbf{H}^{\phi}_{k',t_s-\tau'} \\ \frac{\partial Q^{ML}_k(\boldsymbol{\theta}_k|\boldsymbol{\theta}')}{\partial \mathbf{H}^{\phi'}_{k',t'_s}} &= \frac{\partial}{\partial \mathbf{H}^{\phi'}_{k',t'_s}} \sum_{f,t_s} \log\left(g_{k,f,t_s}\right) + \frac{v'_{k,f,t_s}}{g_{k,f,t_s}} \\ &= \sum_{\tau,f}\left(\frac{g_{k,f,t'_s+\tau} - v'_{k,f,t'_s+\tau}}{g^2_{k,f,t'_s+\tau}}\right) \\ &\quad \times \mathbf{D}^{\tau}_{f-\phi',k'}. \end{aligned} \tag{16}$$

Unlike the conventional EM algorithm, it is not possible to directly set $\partial Q^{ML}_k(\boldsymbol{\theta}_k|\boldsymbol{\theta}')/\mathbf{D}^{\tau'}_{f',k'} = 0$ and $\partial Q^{ML}_k(\boldsymbol{\theta}_k|\boldsymbol{\theta}')/\mathbf{H}^{\phi'}_{k',t'_s} = 0$ because of the nonlinear coupling between $\mathbf{D}^{\tau}_{f,k}$ and $\mathbf{H}^{\phi}_{k,t_s}$ via v'_{k,f,t_s}. Thus, closed form expressions for estimating $\mathbf{D}^{\tau}_{f,k}$ and $\mathbf{H}^{\phi}_{k,t_s}$ cannot be accomplished. To overcome this problem, we use the following update rules and unify it as part of the M-step:

$$\boldsymbol{\theta}_k \leftarrow \boldsymbol{\theta}_k \cdot \left(\frac{\left[\nabla Q^{ML}_k(\boldsymbol{\theta}_k|\boldsymbol{\theta}')\right]_-}{\left[\nabla Q^{ML}_k(\boldsymbol{\theta}_k|\boldsymbol{\theta}')\right]_+}\right) \tag{17}$$

where $\nabla Q^{ML}_k(\boldsymbol{\theta}_k|\boldsymbol{\theta}') = [\nabla Q^{ML}_k(\boldsymbol{\theta}_k|\boldsymbol{\theta}')]_+ - [\nabla Q^{ML}_k(\boldsymbol{\theta}_k|\boldsymbol{\theta}')]_-$. For each $\mathbf{D}^{\tau}_k$ and $\mathbf{H}^{\phi}_k$ variables, we have

$$\begin{aligned} \left[\nabla Q^{ML}_k(\boldsymbol{\theta}_k|\boldsymbol{\theta}')\right]^{\mathbf{D}}_- &= \sum_{\phi,t_s}\left(g_{k,f'+\phi,t_s}\right)^{-2} v'_{k,f'+\phi,t_s}\mathbf{H}^{\phi}_{k',t_s-\tau'} \\ \left[\nabla Q^{ML}_k(\boldsymbol{\theta}_k|\boldsymbol{\theta}')\right]^{\mathbf{D}}_+ &= \sum_{\phi,t_s}\left(g_{k,f'+\phi,t_s}\right)^{-1}\mathbf{H}^{\phi}_{k',t_s-\tau'} \end{aligned} \tag{18}$$

and

$$\begin{aligned} \left[\nabla Q^{ML}_k(\boldsymbol{\theta}_k|\boldsymbol{\theta}')\right]^{\mathbf{H}}_- &= \sum_{\tau,f}\mathbf{D}^{\tau}_{f-\phi',k'}\left(g_{k,f,t'_s+\tau}\right)^{-2} v'_{k,f,t'_s+\tau} \\ \left[\nabla Q^{ML}_k(\boldsymbol{\theta}_k|\boldsymbol{\theta}')\right]^{\mathbf{H}}_+ &= \sum_{\tau,f}\mathbf{D}^{\tau}_{f-\phi',k'}\left(g_{k,f,t'_s+\tau}\right)^{-1}. \end{aligned} \tag{19}$$

Inserting (18) and (19) into (17) leads to

$$\mathbf{D}^{\tau'}_{f',k'} \leftarrow \mathbf{D}^{\tau'}_{f',k'}\frac{\sum\limits_{\phi,t_s}\left(g_{k,f'+\phi,t_s}\right)^{-2} v'_{k,f'+\phi,t_s}\mathbf{H}^{\phi}_{k',t_s-\tau'}}{\sum\limits_{\phi,t_s}\left(g_{k,f'+\phi,t_s}\right)^{-1}\mathbf{H}^{\phi}_{k',t_s-\tau'}}. \tag{20}$$

Similarly, the update rules in $\mathbf{H}^{\phi'}_{k',t'_s}$ writes

$$\mathbf{H}^{\phi'}_{k',t'_s} \leftarrow \mathbf{H}^{\phi'}_{k',t'_s}\frac{\sum\limits_{\tau,f}\mathbf{D}^{\tau}_{f-\phi',k'}\left(g_{k,f,t'_s+\tau}\right)^{-2} v'_{k,f,t'_s+\tau}}{\sum\limits_{\tau,f}\mathbf{D}^{\tau}_{f-\phi',k'}\left(g_{k,f,t'_s+\tau}\right)^{-1}}. \tag{21}$$

It can be verified that the above update rules have an advantage of ensuring the nonnegativity constraints of $\mathbf{D}^{\tau}_{f,k}$ and $\mathbf{H}^{\phi}_{k,t_s}$ are always maintained during every iteration.

C. Algorithm 2: Multiplicative Gradient Descent Formulation of IS-NMF2D (MGD IS-NMF2D)

We consider the following generative model defined as

$$|\mathbf{Y}_{f,t_s}|^2 = \left(\sum_{i=1}^{I} \sum_{\tau=0}^{\tau_{\max}} \sum_{\phi=0}^{\phi_{\max}} \mathbf{D}^{\tau}_{f-\phi,i} \mathbf{H}^{\phi}_{i,t_s-\tau} \right) \bullet \mathbf{E}_{f,t_s} \tag{22}$$

where $\mathbf{E}_{f,t_s}$ is a scalar of multiplicative independent and identically-distributed (i.i.d.) Gamma noise with unit mean, i.e., $p(\mathbf{E}_{f,t_s}) = \xi(\mathbf{E}_{f,t_s}|\alpha,\beta)$ where $\xi(\mathbf{E}_{f,t_s}|\alpha,\beta)$ denotes the Gamma probability density function (pdf) defined as $\xi(\mathbf{E}_{f,t_s}|\alpha,\beta) = (\beta^{\alpha}/\Gamma(\alpha))(\mathbf{E}_{f,t_s})^{\alpha-1}\exp(-\beta\mathbf{E}_{f,t_s})$, $\mathbf{E}_{f,t_s} \geq 0$. Next, we define $\mathbf{D} = [\mathbf{D}^1\ \mathbf{D}^2 \cdots \mathbf{D}^{\tau_{\max}}]$ and $\mathbf{H} = [\mathbf{H}^1\ \mathbf{H}^2 \cdots \mathbf{H}^{\phi_{\max}}]$. Under the independent and identically distributed (i.i.d.) noise assumption, the term $-\log p(\mathbf{Y}|\mathbf{D},\mathbf{H})$ becomes

$$\begin{aligned} &-\log p(\mathbf{Y}|\mathbf{D},\mathbf{H}) \\ &= \frac{-\sum\limits_{t_s=1}^{Ts}\sum\limits_{f=1}^{F} \log \xi \left(\dfrac{|\mathbf{Y}|^{\cdot 2}_{f,t_s}}{\sum\limits_{i=1}^{I}\sum\limits_{\tau=0}^{\tau_{\max}}\sum\limits_{\phi=0}^{\phi_{\max}} \mathbf{D}^{\tau}_{f-\phi,i}\mathbf{H}^{\phi}_{i,t_s-\tau}} \middle| \alpha,\beta \right)}{\sum\limits_{i=1}^{I}\sum\limits_{\tau=0}^{\tau_{\max}}\sum\limits_{\phi=0}^{\phi_{\max}} \mathbf{D}^{\tau}_{f-\phi,i}\mathbf{H}^{\phi}_{i,t_s-\tau}} \\ &\doteq d_{IS}\left(|\mathbf{Y}|^{\cdot 2}_{f,t_s} \middle| \sum_{i=1}^{I}\sum_{\tau=0}^{\tau_{\max}}\sum_{\phi=0}^{\phi_{\max}} \mathbf{D}^{\tau}_{f-\phi,i}\mathbf{H}^{\phi}_{i,t_s-\tau} \right) \end{aligned} \tag{23}$$

where $\doteq$ in the second line denotes equality up to constant terms. Thus, the cost function is $C_{IS}^{NMF2D} = -\log p(\mathbf{Y}|\mathbf{D},\mathbf{H})$. The derivatives of (23) corresponding to $\mathbf{D}^{\tau}$ and $\mathbf{H}^{\phi}$ are given by

$$\begin{aligned} \frac{\partial C_{IS}^{NMF2D}}{\partial \mathbf{D}^{\tau'}_{f',i'}} &= \frac{\partial}{\partial \mathbf{D}^{\tau'}_{f',i'}} \sum_{f,t_s} \left(\frac{|\mathbf{Y}|^2_{f,t_s}}{\mathbf{Z}_{f,t_s}} - \log\frac{|\mathbf{Y}|^2_{f,t_s}}{\mathbf{Z}_{f,t_s}} - 1 \right) \\ &= -\sum_{\phi,t_s}\Big(\left(\mathbf{Z}_{f'+\phi,t_s}\right)^{-2} \left(|\mathbf{Y}|^2_{f'+\phi,t_s} - \mathbf{Z}_{f'+\phi,t_s}\right)\Big) \\ &\quad \times \mathbf{H}^{\phi}_{i',t_s-\tau'} \end{aligned} \tag{24}$$

$$\begin{aligned} \frac{\partial C_{IS}^{NMF2D}}{\partial \mathbf{H}^{\phi'}_{i',t'_s}} &= \sum_{f,t_s} \mathbf{D}^{t_s-t'_s}_{f-\phi',i'} \Big(\left(\mathbf{Z}_{f,t_s}\right)^{-2} \left(\mathbf{Z}_{f,t_s} - |\mathbf{Y}|^2_{f,t_s}\right)\Big) \\ &= -\sum_{\tau,f} \mathbf{D}^{\tau}_{f-\phi',i'} \\ &\quad \times \Big(\left(\mathbf{Z}_{f,t'_s+\tau}\right)^{-2} \left(|\mathbf{Y}|^2_{f,t'_s+\tau} - \mathbf{Z}_{f,t'_s+\tau}\right)\Big) \end{aligned} \tag{25}$$

where $\mathbf{Z} = \sum_{\tau}\sum_{\phi} \overset{\downarrow\phi}{\mathbf{D}^{\tau}} \overset{\rightarrow\tau}{\mathbf{H}^{\phi}}$. The standard gradient decent approach gives

$$\begin{aligned} \mathbf{D}^{\tau'}_{f',i'} &\leftarrow \mathbf{D}^{\tau'}_{f',i'} - \eta_D \frac{\partial Cost_{IS}^{NMF2D}}{\partial \mathbf{D}^{\tau'}_{f',i'}} \quad \text{and} \\ \mathbf{H}^{\phi'}_{i',t'_s} &\leftarrow \mathbf{H}^{\phi'}_{i',t'_s} - \eta_H \frac{\partial Cost_{IS}^{NMF2D}}{\partial \mathbf{H}^{\phi'}_{i',t'_s}} \end{aligned} \tag{26}$$

where η_D and η_H are positive learning rates can be obtained as

$$\begin{aligned} \eta_D &= \frac{\mathbf{D}^{\tau'}_{f',i'}}{\sum\limits_{\phi,t_s} \left(\mathbf{Z}_{f'+\phi,t_s}\right)^{-1} \mathbf{H}^{\phi}_{i',t_s-\tau'}} \quad \text{and} \\ \eta_H &= \frac{\mathbf{H}^{\phi'}_{i',t'_s}}{\sum\limits_{\tau,f} \mathbf{D}^{\tau}_{f-\phi',i'} \left(\mathbf{Z}_{f,t'_s+\tau}\right)^{-1}}. \end{aligned} \tag{27}$$

Inserting (27) into (26) gives the multiplicative gradient decent rules

$$\mathbf{D}^{\tau'}_{f',i'} \leftarrow \mathbf{D}^{\tau'}_{f',i'} \frac{\sum\limits_{\phi,t_s} \left(\mathbf{Z}_{f'+\phi,t_s}\right)^{-2} |\mathbf{Y}|^2_{f'+\phi,t_s} \mathbf{H}^{\phi}_{i',t_s-\tau'}}{\sum\limits_{\phi,t_s} \left(\mathbf{Z}_{f'+\phi,t_s}\right)^{-1} \mathbf{H}^{\phi}_{i',t_s-\tau'}} \tag{28}$$

and

$$\mathbf{H}^{\phi'}_{i',t'_s} \leftarrow \mathbf{H}^{\phi'}_{i',t'_s} \frac{\sum\limits_{\phi,t_s} \left(\mathbf{Z}_{f,t'_s+\tau}\right)^{-2} |\mathbf{Y}|^2_{f,t'_s+\tau} \mathbf{D}^{\tau}_{f-\phi',i'}}{\sum\limits_{\tau,f} \mathbf{D}^{\tau}_{f-\phi',i'} \left(\mathbf{Z}_{f,t'_s+\tau}\right)^{-1}}. \tag{29}$$

The key difference between both algorithms is that the Quasi-EM IS-NMF2D algorithm prevents zeros in the factors, i.e., $\mathbf{D}^{\tau}$ and $\mathbf{H}^{\phi}$ cannot take entries equal to zero. On the contrary, this is not a feature shared by the MGD IS-NMF2D algorithm since zero coefficients are invariant under MGD updates. If the MGD IS-NMF2D algorithm attains a fixed point solution with zero entries, then it cannot be determined since the limit point is a stationary point [31]. Consequently, the resulting factorizations rendered by these algorithms are not equivalent. For this reason, the Quasi-EM IS-NMF2D algorithm can be considered more reliable for updating $\mathbf{D}^{\tau}$ and $\mathbf{H}^{\phi}$. We have summarized both proposed algorithms in Table III. Details of the source separation performance between these algorithms will be shown in Section V-A where $\psi = 10^{-6}$ is the threshold for ascertaining the convergence.

D. Estimation of Sources

The two matrices which we seek to separate from $|\mathbf{Y}_{f,t_s}|^2$ are $|\tilde{X}_1(f,t_s)|^{\cdot 2}$ and $|\tilde{X}_2(f,t_s)|^{\cdot 2}$. These matrices are estimated as $|\tilde{X}_1(f,t_s)|^{\cdot 2} = \sum_{\tau=0}^{\tau_{\max}}\sum_{\phi=0}^{\phi_{\max}} \mathbf{D}^{\tau}_{f-\phi,1}\mathbf{H}^{\phi}_{1,t_s-\tau}$ and $|\tilde{X}_2(f,t_s)|^{\cdot 2} = \sum_{\tau=0}^{\tau_{\max}}\sum_{\phi=0}^{\phi_{\max}} \mathbf{D}^{\tau}_{f-\phi,2}\mathbf{H}^{\phi}_{2,t_s-\tau}$ [29] which are then used to generate the binary mask as $\mathbf{mask}_i(f,t_s) = 1$ if $|\tilde{X}_i(f,t_s)|^{\cdot 2} > |\tilde{X}_j(f,t_s)|^{\cdot 2}$ and zero otherwise. Finally, the estimated time-domain sources are obtained as $\tilde{\mathbf{x}}_i = \mathrm{Resynthesize}(\mathbf{mask}_i \bullet \mathbf{Y})$ for $i = 1,2$ where $\tilde{\mathbf{x}}_i = [\tilde{x}_i(1),\ldots,\tilde{x}_i(T)]^{\mathbf{T}}$ denotes the ith estimated source. The time-domain estimated sources are resynthesized using the approach in [22] by weighting the mixture cochleagram by the mask and correcting phase shifts introduced during the gammatone filtering.

V. Experimental Results and Analysis

The proposed separation system is tested on recorded audio signals. All recordings and processing are conducted using a PC with Intel Core 2 CPU 6600 @ 2.4 GHz and 2 GB RAM. For mixture generation, three types of mixtures are used, i.e., mixture of music and speech; mixture of different kinds of music and mixture of different kinds of speech. The speech sources (male and female) are selected from the TIMIT speech database while the music sources (jazz and piano) from the RWC database [28]. All mixtures are sampled at 16 kHz sampling rate. In all cases, the sources are mixed with equal average power over the duration of the signals. As for our proposed algorithms, the convolutive components are selected as follows:

1) For jazz and speech mixture, $\tau = \{0,\ldots,4\}$ and $\phi = \{0,\ldots,4\}$.
2) For jazz and piano mixture, $\tau = \{0,\ldots,6\}$ and $\phi = \{0,\ldots,9\}$.
3) For piano and speech mixture, $\tau = \{0,\ldots,6\}$ and $\phi = \{0,\ldots,9\}$.
4) For speech and speech mixture, $\tau = \{0,1\}$ and $\phi = \{0,1,2\}$

These parameters are selected after conducting Monte-Carlo tests over 100 realizations of audio mixture. We have evaluated our separation performance in terms of the signal-to-distortion ratio (SDR) which unifies the signal-to-interference ratio (SIR) and signal-to-artifacts ratio (SAR). MATLAB routines for computing these criteria are obtained from the SiSEC'08 webpage [32].

A. Comparison Between Quasi-EM IS-NMF2D and MGD IS-NMF2D

In this section, we compare the performance of source separation using the above algorithms. In our comparison, we have included the IS-NMF [31] combined with clustering algorithm as the benchmark. This method is selected on two grounds: First, it has been proven to work for music analysis. Second, there are no reliable NMF methods currently available for automatic estimation of the number of components. This issue is sidestepped in [31] which proceeds firstly by factorizing the mixture signal into $K = 2,4,\ldots,10$ components by using the IS-NMF. A grouping method is then used to cluster the K components to each source. Given the K different configurations, the SDR performance is then determined for each case and the best value is eventually retained for comparison. To obtain an objective evaluation, we have also included the separation results using the ideal binary mask (IBM). Note that since the IBM is derived directly from the source signals, its separation performance represents the ideal case. Table IV shows their performance under various audio mixtures.

Remark: The work in [31] splits the separation stage into two steps, i.e., factorization and clustering. Our proposed methods, on the other hand, do not require any components clustering and therefore bypass the need for post-processing clustering step which is one of the main advantages over [31]. Referring to Table IV, it is generally noted that the SDR performance varies significantly depending on the algorithms used for separation. For all type of mixtures, the IS-NMF algorithm delivers an average SDR of 2.78 dB; the MGD IS-NMF2D algorithm with an average SDR of 6.7 dB and, the Quasi-EM IS-NMF2D algorithm with an average SDR of 8 dB. The results obtained using the NMF with convolutive factors outperform the method that does not use the convolutive factors. It is also noted that both MGD IS-NMF2D and Quasi-EM IS-NMF2D algorithms exhibit relatively high SDR performance. However, the performance of Quasi-EM IS-NMF2D is generally better than the MGD IS-NMF2D. This confirms our analysis the MGD IS-NMF2D could be trapped in local minima as explained in Section IV-C. Additionally, the Quasi-EM IS-NMF2D algorithm has outperformed all the above algorithms at every type of audio mixture. More precisely, the Quasi-EM IS-NMF2D algorithm leads to an average SDR improvement close to 1.3 dB per source across all the different type of mixtures as compared to the MGD IS-NMF2D algorithm. To further analyze the performance, we have plotted the cochleagram of the original sources and mixed signal in Fig. 6 and each recovered source in Fig. 3. Panels (A)–(B), (C)–(D) and (E)–(F) denote the recovered cochleagram of the female speech and jazz music by using the IS-NMF, MGD IS-NMF2D and Quasi-EM IS-NMF2D algorithms, respectively. In particular, panels (A)–(B) show that the IS-NMF algorithm can only reconstruct the sources partially. Many spectral-temporal components have been mislabeled as indicated by the red box marked area. On the other hand, it is noted that both MGD IS-NMF2D and Quasi-EM IS-NMF2D algorithms exhibit good reconstruction of the female speech and the jazz music. However, the MGD IS-NMF2D algorithm fails to identify several missing components as indicated in the red box marked area of panels (C)–(D). Hence, the estimation of the jazz music and speech is less accurate compared with the Quasi-EM IS-NMF2D algorithm which has successfully estimated both sources. It is relatively difficult to separate speech sources by using single channel USS methods. Currently, to our best knowledge, there are no single channel USS methods that can obtain reliable separation results of speech mixtures. The reason is because the spectral bases for speech sources are too similar to separate. This issue only can be handled by using the supervised source separation method such as model based technique which we have summarized in the Introduction. Despite the low SDR performance reported above, our proposed Quasi-EM using cochleagram still yields better separation than others.

B. Separation Performance Under Different TF Representations

In Section II, the separability analysis was undertaken by using the IBM to determine the "separateness" of the mixture without recourse to the separation algorithms. In this section, the impact of separation algorithm is analyzed. Instead of using the IBM, the Quasi-EM IS-NMF2D algorithm is now used to estimate the mask according to Section IV-D. In this situation, we are investigating the performance of mixture separation (rather than mixture separability). Speech signals and music are used to generate the monoaural mixture recording. The separation performance is evaluated by using three types of TF representation: 1) spectrogram (STFT with 1024-point Hamming windowed FFT and 50% overlap), 2) log-frequency spectrogram (as described in Section III with 1024-point Hamming windowed FFT) and 3) cochleagram based on Gammatone filterbank of 128 channels, filter order of 4 (i.e., $h = 4$ in (4)), and each filter output is divided into 20-ms time frame with 50% overlap. To validate the parameters setting of cochleagram (e.g., h and v), we have constructed an experiment based on three speech

TABLE III
PSEUDOCODES FOR QUASI-EM IS-NMF2D AND IS-NMF2D (MGD) ALGORITHMS

Quasi-EM IS-NMF2D algorithm

```
Input: $|\mathbf{Y}|^2$, random nonnegative matrix $\mathbf{D}^\tau$ and $\mathbf{H}^\phi$, $\phi$, $\tau$
Output: $\mathbf{D}^\tau$ and $\mathbf{H}^\phi$
Procedure:
Compute initialize cost value $Cost(1)$ using (12)
for n=1: max number of iterations
  for k=1:K
  (E-step): Compute $v_{k,f,t_s} = \left|u^{post}_{k,f,t_s}\right|^2 + \lambda^{post}_{k,f,t_s}$ using (13).
  (M-step): Iterate the following until convergence is
            achieved.
              - Update $\mathbf{D}^{\tau'}_{f',k'}$ using (20) for all $\tau$, $\phi$.
                Normalize $\mathbf{D}^{\tau'}_{f',k'}$.
              - Update $\mathbf{H}^{\phi'}_{k',t_s'}$ using (21) for all $\tau$, $\phi$.
                Normalize $\mathbf{H}^{\phi'}_{k',t_s'}$.
  end
end
stopping criterion: $\frac{Cost(n-1)-Cost(n)}{Cost(n)} < \psi$
```

MGD IS-NMF2D algorithm

```
Input: $|\mathbf{Y}|^2$, random nonnegative matrix $\mathbf{D}^\tau$ and $\mathbf{H}^\phi$, $\phi$, $\tau$
Output: $\mathbf{D}^\tau$ and $\mathbf{H}^\phi$
Procedure:
Compute initialize cost value $Cost(1)$ using (23)
for n=1: max number of iterations
  Compute $\mathbf{Z} = \sum_\tau \sum_\phi \mathbf{D}^\tau_{f-\phi} \mathbf{H}^\phi_{t_s-\tau}$.
    - Update $\mathbf{D}^{\tau'}_{f',i'}$ using (28) for all $\tau$, $\phi$.
      Normalize $\mathbf{D}^{\tau'}_{f',i'}$.
  Compute $\mathbf{Z} = \sum_\tau \sum_\phi \mathbf{D}^\tau_{f-\phi} \mathbf{H}^\phi_{t_s-\tau}$.
    - Update $\mathbf{H}^{\phi'}_{i',t_s'}$ using (29) for all $\tau$, $\phi$.
      Normalize $\mathbf{H}^{\phi'}_{i',t_s'}$.
    Compute cost value using (23)
end
stopping criterion: $\frac{Cost(n-1)-Cost(n)}{Cost(n)} < \psi$
```

TABLE IV
SEPARATION RESULTS USING MATRIX FACTORIZATION METHODS

Mixtures	Algorithms	SDR
jazz and male	IS-NMF with clustering	4.14
	MGD IS-NMF2D	7.45
	Quasi-EM IS-NMF2D	8.87
	IBM	12.51
jazz and female	IS-NMF with clustering	4.51
	MGD IS-NMF2D	7.67
	Quasi-EM IS-NMF2D	9.34
	IBM	11.23
piano and male	IS-NMF with clustering	-0.70
	MGD IS-NMF2D	5.84
	Quasi-EM IS-NMF2D	7.16
	IBM	10.65
piano and female	IS-NMF with clustering	2.59
	MGD IS-NMF2D	6.36
	Quasi-EM IS-NMF2D	7.44
	IBM	11.23
jazz and piano	IS-NMF with clustering	3.37
	MGD IS-NMF2D	6.18
	Quasi-EM IS-NMF2D	7.21
	IBM	11.42
speech and speech	IS-NMF with clustering	-0.4
	MGD IS-NMF2D	-0.2
	Quasi-EM IS-NMF2D	0.5
	IBM	11.3

sources and tested the result by fixing the parameter h in (3) to unity.

The experiment is then repeated by progressively increasing h from 2 to 10. Over this range, the optimal separability is obtained when $h = 4$. The parameter v determines the rate of decay of the impulse response of the gammatone filters. In most audio processing tasks, it is set to $v(f) = 1.019ERB(f)$ where $ERB(f) = 24.7 + 0.108f$ is the equivalent rectangular bandwidth of the filter with the center frequency f. A range of values for v has been tested, i.e., $v(f) = (1.019 + c)ERB(f)$ where c ranges from -0.5 to 0.5 with increment of 0.1. The obtained result indicates that the optimal separability is obtained by setting $c = 0$. As c moves away from 0, the separability result progressively deteriorates. This confirms the validity of setting $v(f) = 1.019ERB(f)$ for the cochleagram.

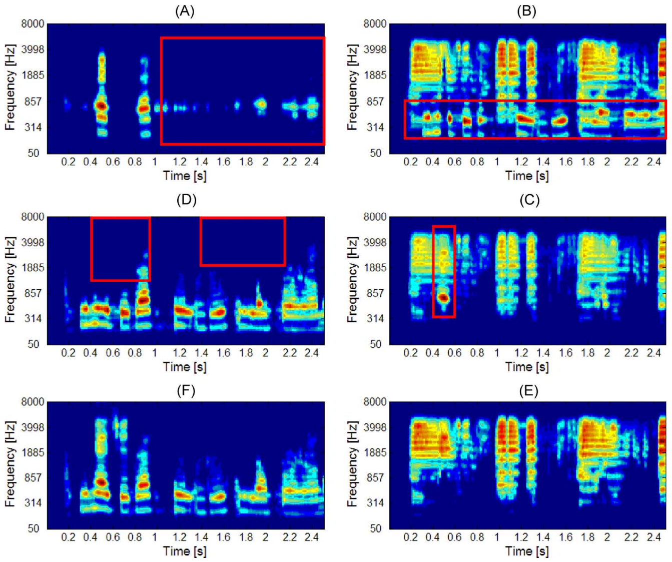

Fig. 3. Separation results using the following factorization methods: (A)–(B) IS-NMF. (C)–(D) MGD IS-NMF2D. (E)–(F) Quasi-EM IS-NMF2D.

Table V shows the comparison of our proposed algorithm based on the spectrogram, log-frequency spectrogram and cochleagram under various audio mixtures. The separation results for all mixture types based on the spectrogram gives an average SDR of 0.51 dB while the log-frequency spectrogram an average SDR of 2.8 dB. However, a significantly higher performance is attained by the cochleagram with an average SDR of 8 dB. This leads to a substantial improvement gain of 7.5 dB and 5.2 dB, respectively. The major reason for the large

TABLE V
SEPARATION RESULTS USING DIFFERENT TF REPRESENTATIONS

Mixtures	TF methods	SDR
jazz and male	spectrogram	3.47
	log-frequency spectrogram	6.54
	cochleagram	**8.87**
jazz and female	spectrogram	-1.41
	log-frequency spectrogram	3.97
	cochleagram	**9.34**
piano and male	spectrogram	2.10
	log-frequency spectrogram	2.31
	cochleagram	**7.16**
piano and female	spectrogram	-1.01
	log-frequency spectrogram	0.27
	cochleagram	**7.44**
jazz and piano	spectrogram	-0.59
	log-frequency spectrogram	1.21
	cochleagram	**7.21**
speech and speech	spectrogram	-0.6
	log-frequency spectrogram	-0.3
	cochleagram	**0.5**

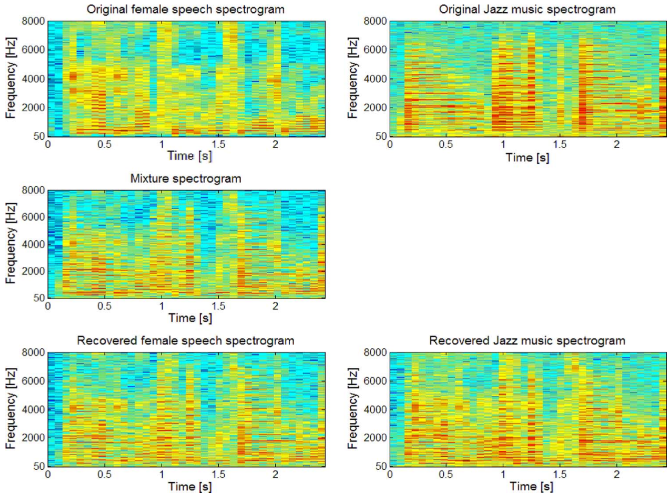

Fig. 4. Separation results in spectrogram.

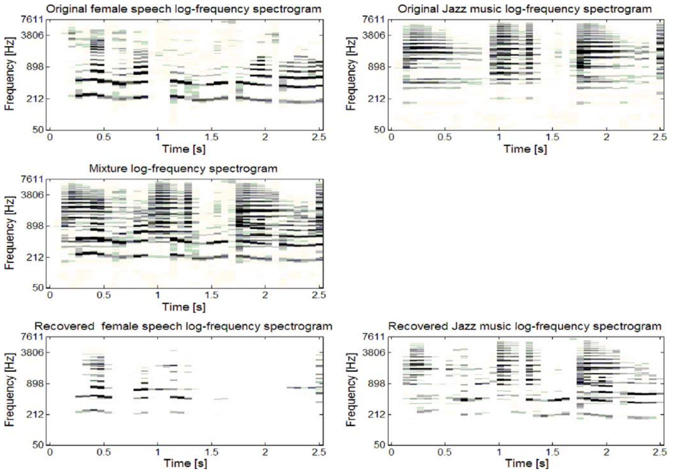

Fig. 5. Separation results in log-frequency spectrogram.

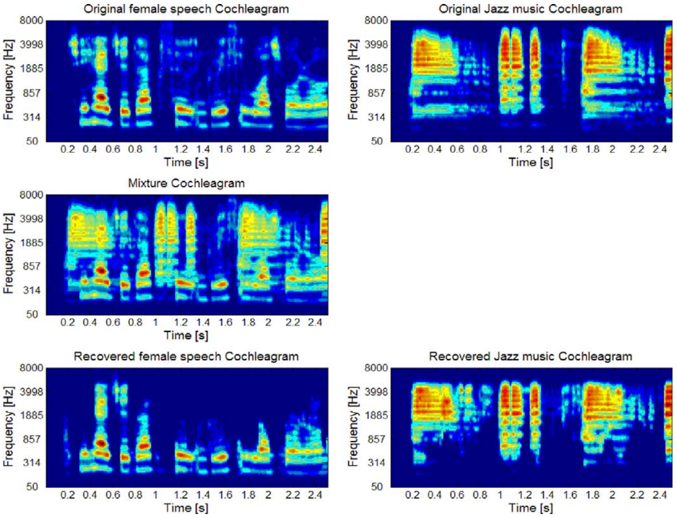

Fig. 6. Separation results in cochleagram.

discrepancy is due to the mixing ambiguity between $|\mathbf{X}_1|^{\cdot 2}$ and $|\mathbf{X}_2|^{\cdot 2}$. The larger the mixing ambiguity between $|\mathbf{X}_1|^{\cdot 2}$ and $|\mathbf{X}_2|^{\cdot 2}$, the more TF units will be ambiguous which subsequently decreases the probability of correct assignment of each unit to the sources and eventually results in poorer separation performance. To validate this, Fig. 4 shows the spectrogram of the original sources, the mixed signal, and the estimated sources using the proposed Quasi-EM IS-NMF2D algorithm. Both figures indicate that the STFT lacks provision for further low-level information of a TF unit and therefore, the resulting spectrogram fails to infer the dominating source. This leads to high degree of ambiguity in TF domain and causes lack of uniqueness in extracting the spectral-temporal features of the sources.

Similar to above, Fig. 5 shows the separation results based on the log-frequency spectrogram. Comparing with spectrogram, the separation performance is better since log-frequency spectrogram has the propensity of nonuniform time–frequency resolution. However, the transform operation used by the log-frequency spectrogram is still based on the Fourier transform which may not be an optimal option. On the other hand, the results of separation in the cochleagram have led to significant SDR improvement. The cochleagram enables the mixed signal to be more separable and thus reduces the mixing ambiguity between $|\mathbf{X}_1|^{\cdot 2}$ and $|\mathbf{X}_2|^{\cdot 2}$. This explains the average performance of separating mixture jazz music and female utterance is highest among all the mixtures because both sources have very distinguishable TF patterns in the cochleagram. This is evident in Fig. 6 which shows the separation results in the cochleagram. The plot clearly shows that the spectral energy of the two audio sources has been clustered at different frequencies in the cochleagram due to their different fundamental frequencies. These prominent features have been separated using our proposed Quasi-EM IS-NMF2D algorithm.

The performance of source separation also depends on how accurate the spectral bases are estimated. Given the different types of TF representation, a question arises as to which set of estimated spectral bases have yielded better approximation to the respective original sources' spectral bases. Fig. 7 shows the results of the original and the estimated spectral basis $\mathbf{D}_i^{\tau}$ for the above mixture when the factorization is performed in the cochleagram. In Fig. 7, panels (A)–(B) refer to the original

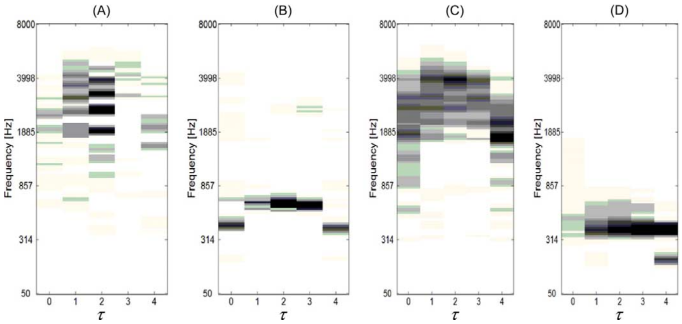

Fig. 7. (A)–(B) Original spectral bases of jazz music and female utterance in the cochleagram. (C)–(D) The corresponding estimated spectral bases.

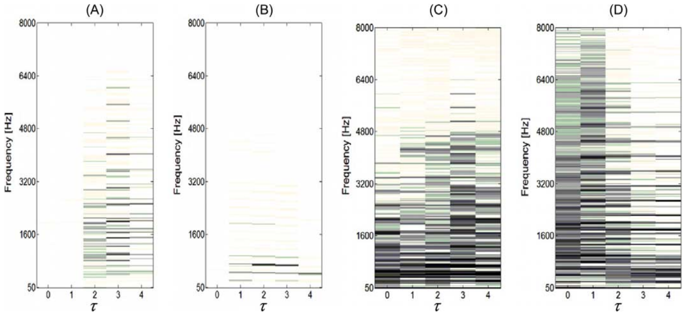

Fig. 8. (A)–(B) Original spectral bases of jazz music and female utterance in the spectrogram. (C)–(D) The corresponding estimated spectral bases.

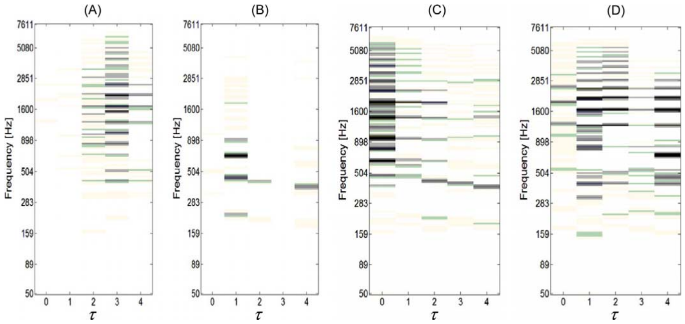

Fig. 9. (A)–(B) Original spectral bases of jazz music and female utterance in the log-frequency spectrogram. (C)–(D) The corresponding estimated spectral bases.

spectral bases of the jazz music and female utterance, respectively. Panels (C)–(D) refer to the estimated spectral bases. In comparison, we have also included similar factorization results of the same mixture in the spectrogram and log-frequency spectrogram. These are shown in Figs. 8 and 9, respectively. In sharp contrast with Fig. 7, it is noted that the estimated spectral bases in Figs. 8 and 9 are quite dissimilar to the original spectral bases. Thus, the construction of the separating mask will inevitably introduce errors in assigning the TF units to the respective sources. Therefore, the recovered sources are very coarse with very low values of SDR in Table V.

C. Comparison Between Different Cost Functions

In the following, experiments are conducted to evaluate the efficiency of the proposed algorithm under different cost functions. Here, we consider the least square (LS) distance and Kullback–Leibler (KL) divergence. Table VI shows the separation results in the cochleagram based on LS, KL and IS cost functions. In Table VI, it is noted that Quasi-EM IS-NMF2D algorithm outperforms those of LS distance and KL divergence with an average SDR of 3.1 dB, and 1.8 dB, respectively. This is evidenced by the fact that the IS divergence holds a desirable property of scale invariant so that low energy components can be precisely estimated and they bear the same relative importance as the high energy ones. On the contrary, factorizations obtained with LS distance and KL divergence tend to favor the high energy components at the expense of disregarding the low energy ones. In the cochleagram, the dynamic range of the mixture signal can be considerably large such that the dominating signal at a particular TF unit can manifest either as low or high energy components. In addition, these components tend to exist as clusters. As such, when either LS distance- or KL divergence is used, clusters with low energy tend to be ignored in favor of the high energy ones. This leads to mixing ambiguities especially for low energy ones in which case when they are subsumed together leads to significant lost of spectral-temporal information of the sources. Fig. 10 shows how different cost functions have impacted the separation performance. It is clearly seen that the LS-NMF2D algorithm fails to determine the correct TF components of each source. Panels (A)–(B) show a considerable level of mixing ambiguities (red box marked area) which have not been accurately resolved by the LS-NMF2D algorithm. The KL-NMF2D exhibits better performance but ignores some low energy TF components in the red box marked area of (C). On the other hand, the proposed algorithm has successfully extracted the low energy components for both female speech and jazz music with high accuracy.

TABLE VI
SEPARATION RESULTS WITH DIFFERENT COST FUNCTIONS

Mixtures	Algorithms	SDR
jazz and male	LS-NMF2D	6.15
	KL-NMF2D	7.24
	Quasi-EM IS-NMF2D	8.87
jazz and female	LS-NMF2D	4.69
	KL-NMF2D	7.35
	Quasi-EM IS-NMF2D	9.34
piano and male	LS-NMF2D	5.11
	KL-NMF2D	5.42
	Quasi-EM IS-NMF2D	7.16
piano and female	LS-NMF2D	4.21
	KL-NMF2D	5.38
	Quasi-EM IS-NMF2D	7.44
jazz and piano	LS-NMF2D	4.61
	KL-NMF2D	5.86
	Quasi-EM IS-NMF2D	7.21
speech and speech	LS-NMF2D	-0.4
	KL-NMF2D	-0.5
	Quasi-EM IS-NMF2D	0.5

D. Determining the Number of Sources

The issue of determining number of sources from single mixture channel is an open problem. One possible approach is to use the model order selection method based on Akaike information criteria (AIC) to predict the source numbers. The expression of AIC is given by $AIC = 2I - 2\ln(L)$ where L is maximized value of the likelihood function. In the proposed method, the $-\ln(L)$ is given by (12). Note that the term $-\ln(L)$ is the final converged value for calculating AIC and the optimal number of

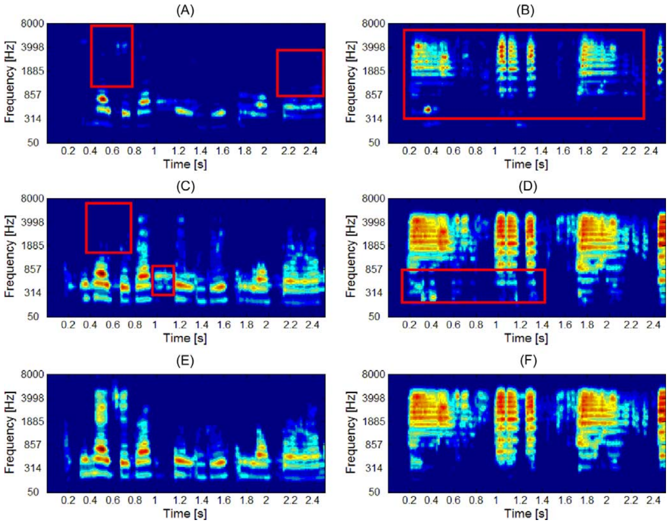

Fig. 10. Separation results: (A)–(B), (C)–(D), and (E)–(F) denote the recovered female speech and jazz music in the cochleagram by using the algorithms with different cost function.

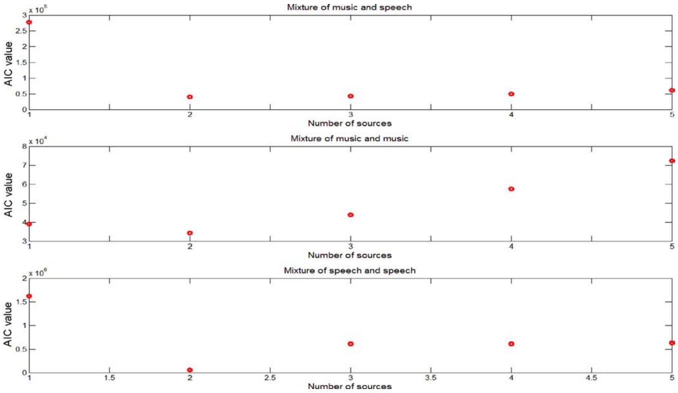

Fig. 11. AIC results for each type of mixture.

sources can be predicted at the point when AIC is the smallest. We have conducted the experiments on two sources mixture where the following figure summarizes the prediction results of the number of sources. Fig. 11 show that the AIC has successfully predicted the number of sources.

E. Separating Noisy Mixture

We have added the results testing on noise mixture. Two types of the noise have been tested: 1). White noise and 2). Street noise. For each type noise, we have tested the mixture with noise power with $\mathrm{SNR} = 0$ dB, 5 dB, and 10 dB. In general, the results show that the proposed method still can separate the target sources when corrupted with noise. Tables VII and VIII show the separation results using the proposed algorithms under different level of SNR with white noise and street noise, respectively. For separating mixture of music and music, and mixture of music and speech that are contaminated with white noise $(\mathrm{SNR} = 0 \text{ dB})$, the MGD IS-NMF2D algorithm gives an average SDR of -0.3 dB while the Quasi-EM IS-NMF2D algorithm with an average SDR of 0.3 dB. On the other hand, the separation of noisy mixture speech and speech delivers an average SDR of -2.4 dB for MGD IS-NMF2D and -1.9 dB for Quasi-EM IS-NMF2D. For all types of mixture, the SDR for both algorithms increase coherently when SNR increase. The separation performance of mixture contaminated with street noise is better than those of white noise. It delivers a high SDR performance even the mixtures are corrupted with high power noise.

TABLE VII
SEPARATION RESULTS WITH WHITE NOISE

Mixtures	Algorithms	SDR		
		SNR 0dB	SNR 5dB	SNR 10dB
jazz and male	MGD IS-NMF2D	-0.2	3.5	5.3
	Quasi-EM IS-NMF2D	0.7	4.2	6.2
jazz and female	MGD IS-NMF2D	-0.7	3.1	4.5
	Quasi-EM IS-NMF2D	0.2	4.6	6.9
piano and male	MGD IS-NMF2D	-0.4	1.8	2.7
	Quasi-EM IS-NMF2D	0.3	2.8	4.6
piano and female	MGD IS-NMF2D	-0.6	0.9	2.1
	Quasi-EM IS-NMF2D	0.1	2.2	3.9
jazz and piano	MGD IS-NMF2D	-0.6	0.6	2.1
	Quasi-EM IS-NMF2D	0.2	1.8	3.7
speech and speech	MGD IS-NMF2D	-2.4	-1.6	-1.2
	Quasi-EM IS-NMF2D	-1.9	-1.2	-0.6

F. Comparing With Different USCSS Methods

We have compared recently published USCSS method EMD SCSS [35] which first decomposes the given signal into spectrally independent modes using EMD algorithm, and then, ICA is applied to extract statistically independent sources. All above single-channel USS methods will be tested across all types of mixture and compared in terms of SDR. The Table IX summarizes the comparison results. In comparison, the Quasi-EM IS-NMF2D with cochleagram leads to the best separation performance for all types of the mixture. The Method EMD SCSS also performs the relative acceptable results as compared with Quasi-EM IS-NMF2D. However, it is interesting to point out that the advantage of using Quasi-EM IS-NMF2D with cochleagram is that this method is less complex than the EMD SCSS and simultaneously it retains a higher level of the separation performance.

G. Separating More Than Two Sources

The proposed method can be extended to the case when $i > 2$ sources. If more than two sources mixed in a single channel, this requires specifying the number of sources to be separated. Since the method is unsupervised, the separability of the complex mixture depends highly on how accurate the spectral bases $\mathbf{D}_i^\tau$ can be estimated from the TF mixture. Consequently, a set of distinguishable spectral basis of each source for a generic case is a necessary condition to achieve good separation performance. Thus, we adopt three different types of sources, e.g., jazz, piano and trumpet to generate a complex mixture. The convolutive components in the proposed algorithm are selected as $\tau = \{0, \ldots, 3\}$ and $\phi = \{0, \ldots, 31\}$. Table X shows the overall separation results. It is seen that mixtures generated by all music sources have been recovered quite successfully. It is noted that the separation performance has deteriorated when the number of sources increases from two. Increased number of sources will mean that there exists more interference in separating every

TABLE VIII
SEPARATION RESULTS WITH STREET NOISE

Mixtures	**Algorithms**	**SDR**		
		SNR 0dB	**SNR 5dB**	**SNR 10dB**
jazz and male	MGD IS-NMF2D	4.8	6.2	7.3
	Quasi-EM IS-NMF2D	5.6	6.8	7.5
jazz and female	MGD IS-NMF2D	5.4	6.4	7.2
	Quasi-EM IS-NMF2D	5.8	7.1	7.9
piano and male	MGD IS-NMF2D	3.6	4.1	4.7
	Quasi-EM IS-NMF2D	4.6	5.2	5.8
piano and female	MGD IS-NMF2D	2.9	3.2	4.1
	Quasi-EM IS-NMF2D	3.5	4.6	5.3
jazz and piano	MGD IS-NMF2D	2.7	3.8	4.6
	Quasi-EM IS-NMF2D	3.5	4.3	5.1
speech and speech	MGD IS-NMF2D	-1.5	-1.1	-0.8
	Quasi-EM IS-NMF2D	-0.8	-0.3	0.2

TABLE IX
SEPARATION RESULTS USING DIFFERENT SCBSS METHODS

Mixtures	**Method**	**SDR**
jazz and male	EMD SCSS	6.3
	Quasi-EM IS-NMF2D	8.8
jazz and female	EMD SCSS	5.2
	Quasi-EM IS-NMF2D	9.3
piano and male	EMD SCSS	5.2
	Quasi-EM IS-NMF2D	7.1
piano and female	EMD SCSS	6.6
	Quasi-EM IS-NMF2D	7.4
jazz and piano	EMD SCSS	6.6
	Quasi-EM IS-NMF2D	8.5
speech and speech	EMD SCSS	0.4
	Quasi-EM IS-NMF2D	0.5

TABLE X
SEPARATION RESULTS OF THREE SOURCES

Mixtures: $y = x_1 + x_2 + x_3$			SDR of $\hat{x}_1$	SDR of $\hat{x}_2$	SDR of $\hat{x}_3$
x_1	x_2	x_3			
jazz	piano	trumpet	6.51	5.61	5.65
male	jazz	piano	5.23	5.73	4.13
male	jazz	trumpet	5.18	5.65	5.21
male	piano	trumpet	5.20	4.09	4.53
female	jazz	piano	5.36	5.47	4.24
female	jazz	trumpet	5.02	5.51	5.10
female	piano	trumpet	5.02	4.32	4.28
male	female	male	-0.8	1.3	-1.6

target source and hence results in higher probability of incurring an error. Comparing the results in the table, mixtures containing speech somehow results in slightly poorer performance than mixtures of music sources only. One reason is the seemingly more overlaps in the TF domain between the speech and music sources. It is observed from Fig. 6 that music pitches tend to jump discretely while speech pitches do not. Consequently, this leads to less efficiency in the estimation of the spectral basis $\mathbf{D}_i^{\tau}$ from the mixture signal. In addition, we have tested the performance of the proposed method on recordings mixed with $i > 3$ sources. We have found that the proposed method works well for mixture of music sources that are characterized with distinguishable spectral basis. However, the performance shows degradation when separating mixture contains speech sources.

VI. CONCLUSION

In this paper, a novel method to solve the single channel audio source separation has been proposed. In addition, two algorithms for nonnegative matrix two-dimensional factorization optimized using the Itakura–Saito divergence have been presented: Quasi-EM IS-NMF2D and MGD IS-NMF2D. Coupled with the theoretical support of signal separability in the TF domain, the separation system using the gammatone filterbank with these algorithms have shown to yield considerable success. The proposed method enjoys at least three significant advantages: First, it avoids strong constraints of separating sources without training knowledge. Second, the cochleagram rendered by the gammatone filterbank has nonuniform TF resolution which enables the mixed signal to be more separable and thus improves the efficiency of source separation. Finally, the method holds a desirable property of scale invariant which enables low energy components in the cochleagram bear the same relative importance as the high energy ones. The proposed cochleagram-based IS-NMF2D method in particular using the Quasi-EM algorithm has yielded significant improvements in source separation when compared with other nonnegative matrix factorizations.

REFERENCES

[1] Q. H. Lin, F. L. Yin, T. M. Mei, and H. L. Liang, "A blind source separation based method for speech encryption," *IEEE Trans. Circuit Syst. I*, vol. 53, no. 6, pp. 1320–1328, Jun. 2006.

[2] S. Li, C. Li, K.-T. Lo, and G. Chen, "Cryptanalyzing an encryption scheme based on blind source separation," *IEEE Trans. Circuits Syst. I*, vol. 55, no. 4, pp. 1055–1063, May 2008.

[3] R. Schachtner, G. Poppel, and E. W. Lang, "A nonnegative blind source separation model for binary test data," *IEEE Trans. Circuits Syst. I*, vol. 57, no. 7, pp. 1439–1448, Jul. 2010.

[4] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis and Blind Source Separation*. New York: Wiley, 2001, pp. 20–60.

[5] T. Kim, "Real-time independent vector analysis for convolutive blind source separation," *IEEE Trans. Circuits Syst. I*, vol. 7, no. 7, pp. 1431–1438, Jul. 2010.

[6] W. L. Woo and S. S. Dlay, "Neural network approach to blind signal separation of mono-nonlinearly mixed sources," *IEEE Trans. Circuits Syst. I*, vol. 52, no. 6, pp. 1236–1247, Jun. 2005.

[7] M. B. Priestley, "Evolutionary spectra and non-stationary processes," *J. R. Statist. Soc. Series B (Methodological)*, vol. 27, no. 2, pp. 204–237, 1965.

[8] D. Lee and H. Seung, "Learning the parts of objects by nonnegative matrix factorisation," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[9] S. Xie, Z. Y. Yang, and Y. L. Fu, "Nonnegative matrix factorization applied to nonlinear speech and image cryptosystems," *IEEE Trans. Circuits Syst. I*, vol. 55, no. 8, pp. 2356–2367, Sep. 2008.

[10] M. Helén and T. Virtanen, "Separation of drums from polyphonic music using nonnegative matrix factorization and support vector machine," in *Proc. 13th Eur. Signal Process.*, 2005.

[11] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, 2003, pp. 177–180.

[12] S. Rickard and A. Cichocki, "When is non-negative matrix decomposition unique?," in *Proc. 42nd Annu. Conf. Inf. Sci. Syst. CISS '08*, Mar. 2008, pp. 1091–1092.

[13] S. A. Abdallah and M. D. Plumbley, "Polyphonic transcription by non-negative sparse coding of power spectra," in *Proc. 5th Conf. Music Inf. Retrieval (ISMIR '04)*, Spain, Oct. 2004, pp. 318–325.

[14] R. M. Parry and I. Essa, "Incorporating phase information for source separation via spectrogram factorization," in *Proc. Conf. Acoust., Speech, Signal Process. (ICASSP'07)*, 2007, pp. 661–664.

[15] R. Kompass, "A generalized divergence measure for nonnegative matrix factorization," *Neural Comput.*, vol. 19, no. 3, pp. 780–791, Mar. 2007.

[16] A. Cichocki, R. Zdunek, and S.-I. Amari, "Csisz'ar's divergences for non-negative matrix factorization: Family of new algorithms," in *Proc. 6th Intl. Conf. Ind. Compon. Anal. Signal Separat. (ICA '06)*, Charleston, SC, Mar. 2006, pp. 32–39.

[17] T. Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1066–1074, Mar. 2007.

[18] M. H. Radfa and R. M. Dansereau, "Single-channel speech separation using soft mask filtering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2299–2310, Nov. 2007.

[19] S. Roweis, "One microphone source separation," in *Proc. Neural Inf. Process*, 2000, pp. 793–799.

[20] M. Morup and M. N. Schmidt, Sparse "Non-negative matrix factor 2-D deconvolution," 2006, Tech. Rep.

[21] M. N. Schmidt and M. Morup, "Nonnegative matrix factor 2-D deconvolution for blind single channel source separation," in *Proc. 6th Int. Conf. Ind. Compon. Anal. Signal Separat. (ICA '06)*, Charleston, SC, Mar. 2006, pp. 700–707.

[22] K. Gröchenig, *Foundations of Time-Frequency Analysis*. Boston, MA: Birkhäuser, 2001.

[23] J. C. Brown, "Calculation of a constant Q spectral transform," *J. Acoust. Soc. Amer.*, vol. 89, no. 1, pp. 425–434, 1991.

[24] G. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1135–1150, Sep. 2004.

[25] R. Curtis *et al., The Computer Music Tutorial*. Cambridge, MA: MIT Press, 1996.

[26] S. Schulz and T. Herfet, "Binaural source separation in non-ideal reverberant environments," in *Proc. 10th Int. Conf. Digital Audio Effects (DAFx-07)*, Bordeaux, France, Sep. 2007, pp. 10–15.

[27] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Norwell, MA: Kluwer, 2005, pp. 181–197.

[28] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Music genre database and musical instrument sound database," in *Proc. Int. Symp. Music Inf. Retrieval (ISMIR)*, Baltimore, MD, Oct. 2003, pp. 229–230.

[29] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[30] F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelihood method," in *Proc 6th Int. Congr. Acoust.*, Tokyo, Japan, Aug. 1968, pp. C-17–C-20.

[31] C. Fevotte, N. Bertin, and J. L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Comput.*, vol. 21, no. 3, pp. 793–830, 2009.

[32] Signal separation evaluation campaign (SiSEC 2008), 2008 [Online]. Available: http://sisec.wiki.irisa.fr

[33] B. Gao, W. L. Woo, and S. S. Dlay, "Single channel source separation using EMD-subband variable regularized sparse features," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 961–976, May 2011.

[34] B. Gao, W. L. Woo, and S. S. Dlay, "Adaptive sparsity non-negative matrix factorization for single channel source separation," *IEEE J. Sel. Topics Signal Process.*, vol. 5, pp. 1932–4553, Jun. 2011.

[35] B. Mijović *et al.*, "Source separation from single-channel recordings by combining empirical-mode decomposition and independent component analysis," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 9, pp. 2188–2196, Sep. 2010.

Bin Gao (M'12) received the B.S. degree in communications and signal processing from Southwest Jiao Tong University, Chengdu, China, in 2005, the M.Sc. degree with distinction in communications and signal processing from Newcastle University, Newcastle upon Tyne, U.K., in 2007, and the Ph.D. degree from Newcastle University in 2011, and his research topic is single-channel blind source separation under the supervision of Prof. Woo and Prof. Dlay.

Currently, he is a Research Associate at Newcastle University. His research interests include audio and image processing, machine learning, structured probabilistic modeling on audio applications such as audio source separation, feature extraction and denoising.

W. L. Woo (M'11–SM'12) was born in Malaysia. He received the B.Eng. degree (First Class Honors) in electrical and electronics engineering and the Ph.D. degree from the Newcastle University, Newcastle upon Tyne, U.K.

He is currently a Senior Lecturer with the School of Electrical, Electronics and Computer Engineering, Newcastle University. His major research is in the mathematical theory and algorithms for nonlinear signal and image processing. This includes areas of blind source separation, machine learning, multidimensional signal processing, and signal/image deconvolution and restoration. He has an extensive portfolio of relevant research supported by a variety of funding agencies. Prior to joining the School, he worked on source separation techniques supported by QinetiQ on signal processing-based applications. He has published over 200 papers on these topics on various journals and international conference proceedings. Currently, he serves on the editorial board of the many international signal processing journals. He actively participate in international conferences and workshops, and serves on their organizing and technical committees. In addition, he acts as a consultant to a number of industrial companies that involve the use of statistical signal and image processing techniques.

Dr. Woo was awarded the IEE Prize and the British Scholarship in 1998 to continue his research work. He is also a member of the Institution Engineering Technology (IET).

S. S Dlay received the B.Sc. (Honsors) degree in electrical and electronic engineering and the Ph.D. degree in VLSI design from Newcastle University, Newcastle upon Tyne, U.K., in 1979 and 1983, respectively.

During this time, he held a Scholarship from the Engineering and Physical Science Research Council (EPSRC) and the Charles Hertzmann Award. In 1984, he was appointed as a Post Doctoral Research Associate at Newcastle University and helped to establish an Integrated Circuit Design Centre, funded by the EPSRC. In November 1984, he was appointed as a Lecturer in the Department of Electronic Systems Engineering at the University of Essex. In 1986, he rejoined Newcastle University as a Lecturer in the School of Electrical, Electronic, and Computer Engineering, then in 2001 he was promoted to Senior Lecturer. In recognition of his major achievements he has been appointed to a Personal Chair in Signal Processing Analysis. He is currently Head of the Signal Processing theme. He has published over 250 research papers and his research interests lie in the mathematical advancement and application of modern signal processing theory to biometrics and security, biomedical signal processing and implementation of signal processing architectures. He serves on many editorial boards and has played an active role in numerous international conferences in terms of serving on technical and advisory committees as well as organizing special sessions.

Prof. Dlay is a College Member of the EPSRC.