sensors

MDPI

Article

Efficient Noisy Sound-Event Mixture Classification Using Adaptive-Sparse Complex-Valued Matrix Factorization and OvsO SVM

Phetcharat Parathai [1], Naruephorn Tengtrairat [1], Wai Lok Woo [2,*], Mohammed A. M. Abdullah [3], Gholamreza Rafiee [4] and Ossama Alshabrawy [2]

1 School of Software Engineering, Payap University, Chiang Mai 50000, Thailand; phetcharat@payap.ac.th (P.P.); naruephorn_t@payap.ac.th (N.T.)
2 Department of Computer and Information Sciences, Northumbria University, Newcastle upon Tyne NE1 8ST, UK; ossama.alshabrawy@northumbria.ac.uk
3 Computer and Information Engineering Department, Ninevah University, Mosul 41002, Iraq; mohammed.abdulmuttaleb@uoninevah.edu.iq
4 School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast BT9 5BN, UK; g.rafiee@qub.ac.uk
* Correspondence: wailok.woo@northumbria.ac.uk

Received: 20 June 2020; Accepted: 4 August 2020; Published: 5 August 2020

Abstract: This paper proposes a solution for events classification from a sole noisy mixture that consist of two major steps: a sound-event separation and a sound-event classification. The traditional complex nonnegative matrix factorization (CMF) is extended by cooperation with the optimal adaptive L_1 sparsity to decompose a noisy single-channel mixture. The proposed adaptive L_1 sparsity CMF algorithm encodes the spectra pattern and estimates the phase of the original signals in time-frequency representation. Their features enhance the temporal decomposition process efficiently. The support vector machine (SVM) based one versus one (OvsO) strategy was applied with a mean supervector to categorize the demixed sound into the matching sound-event class. The first step of the multi-class MSVM method is to segment the separated signal into blocks by sliding demixed signals, then encoding the three features of each block. Mel frequency cepstral coefficients, short-time energy, and short-time zero-crossing rate are learned with multi sound-event classes by the SVM based OvsO method. The mean supervector is encoded from the obtained features. The proposed method has been evaluated with both separation and classification scenarios using real-world single recorded signals and compared with the state-of-the-art separation method. Experimental results confirmed that the proposed method outperformed the state-of-the-art methods.

Keywords: audio signal processing; sound event classification; nonnegative matric factorization; blind signal separation; support vector machines

1. Introduction

Surveillance systems have become increasingly ubiquitous in our living environment. These systems have been used in various applications including CCTV in traffic and site monitoring, and navigation. Automated surveillance is currently based on video sensory modality and machine intelligence. Recently, intelligent audio analysis has been taken into account in surveillance to improve the monitoring system via detection, classification, and recognition sound in a scenario. However, in a real-world situation, background noise has interfered in both the image and sound of a surveillance system. This will hinder the performance of a surveillance system. Hence, an automatic signal separation and event classification algorithm was proposed to improve the surveillance system

by classifying the observed sound-event in noisy scenarios. The proposed noisy sound separation and event classification method is based on two approaches (i.e., blind signal separation and sound classification, which are introduced in the sections to follow, respectively).

The classical problem of blind source separation (BSS), the so-called "cocktail party problem", is a psycho-acoustic spectacle that alludes to the significant human-auditory capability to selectively focus on and identify the sound-source speaker from the scenarios. The interference is produced by competing speech sounds or a variety of noises that are often assumed to be independent of each other. In the case of only a single microphone being available, this reduces to the single channel blind source separation (SCBSS) [1–4]. The majority of SCBSS algorithms work in time-frequency domain, for example, binary masking [5–7] or nonnegative matrix factorization (NMF) [8–11]. NMF has been continuously developed with great success for decomposing underlying original signals when a sole sensor is available. NMF was developed using the multiplicative update (MU) algorithm to solve its parametrical optimization based on a cost function such as the Kullback–Liebler divergence and the least square distance. Later, other families of cost functions have been continuously proposed, for example, the Beta divergence [12], Csiszar's divergences, and Itakura–Saito divergence [13]. Additionally, iterative gradient update was presented where a sparsity constraint can be included into the optimizing function through regularization by minimizing penalized least squares [14] and using different sparsity constraints for dictionary and code [15]. The complex nonnegative matrix factorization (CMF) spreads the NMF model by combining a sparsity representation with the complex-spectrum domain to improve the audio separability. The CMF can extract the recurrent patterns of the phase estimates and magnitude spectra of constituent signals [16–18]. Nevertheless, the CMF lacks the generalized mechanics used for controlling the sparseness of the code. However, the sparsity parameter is manually determined for the above proposed methods. Approximate sparsity is an important consideration as they represent important information. Many sparse solutions have been proposed in the last decade [19–25]. Nonetheless, the optimal sparse solution remains an open issue.

Sound event classification (SEC) has vastly been exploited by many researchers. Sound can be categorized into speech, music, noise, environmental sound, or daily living sound [26]. Sound events are available in all classes, for example, car horn, traffic, walking, or knocking, etc. [27,28]. Sound-events contain significant information that can be used to describe what has happened or to predict what will happen next in the future. Most algorithms of the SEC methods are conveyed from sound classification approaches such as sparse coding, deep learning, and support vector machine (SVM). These approaches have been exploited to categorize a sound event in both indoor and outdoor scenarios. In recent years, the deep learning approach has been used to classify the sound-event. A deep learning framework can be established with two convolutional neural networks (CNNs) and a deep multi-layer perceptron (MLP) with rectified linear units (ReLU) as the activation function [29,30]. A Softmax function that is the final activation function is used to classify the sound into its corresponding class. The Softmax function is considered as the generalization of the logistic function, which aims to avoid overfitting. One of the advantages of deep learning is that it does not require feature extraction for the input sound. However, a deep neural network requires large training samples and despite a plethora of research, there is a general consensus that deep neural networks are still difficult to fine tune and generalize to test data. Moreover, it does not lend itself to the explanation as to why a certain decision is being made. Separate from the deep learning framework, another SEC approach is support vector machines [31,32], which has been practically presented to solve the classifier problem in various fields. The SVM algorithm relies on supervised learning by using the fundamental concept of statistic and risk minimization. The main process of the SVM is to draw the optimal separating hyperplane as the decision boundary located in such a way that the margin of separation between classes is maximized. The SVM approach is considered as supervised learning algorithm that is comprised of two sections: (1) a training section to model feature space and an optimal hyperplane, and (2) a testing section to use the SVM model for separating the observed data. The margin denotes the distance of the closest instance and the hyperplane. SVM has the desirable properties in that it requires only two

differentiating factors to categorize two classes and a hyperplane that can be constructed to suit for an individual problem, even in the nonlinear case by selecting a kernel. Second, SVM provides a unique solution, since it is a convex optimization problem.

The rest of this paper is organized as follows. Section 2 presents the proposed noisy sound separation and event classification method, respectively. Next, Section 3 demonstrates and analyzes the performance of the proposed method. Finally, conclusions are drawn up in Section 4.

2. Background

Noisy mixed signals observed via a recording device can be stated as: $y(t) = s_1(t) + s_2(t) + n(t)$ where $s_1(t)$ and $s_2(t)$ denote the original sounds, and $n(t)$ is noise. This research is focused on two sound events in a single recorded signal. The proposed method consists of two steps: noisy sound separation and sound event classification, which is illustrated in Figure 1, where $y(t)$ and $\mathbf{Y}(\omega, t)$ denote a sound-event mixture in the time domain and time-frequency domain, respectively. The terms $\mathbf{W}^k(\omega), \mathbf{H}^k(t)$, $\boldsymbol{\phi}^k(\omega, t)$ are spectral basis, temporal code or weight matrix, and phase information, respectively. The term $\lambda^k(t)$ represents sparsity and $\hat{s}_j(t)$ is an estimated sound event source. The abbreviations MFCC, STE, and STZCR stand for Mel frequency cepstral coefficients, short-time energy, and short-time zero-crossing rate, respectively. The proposed method is consecutively elaborated in the following parts.

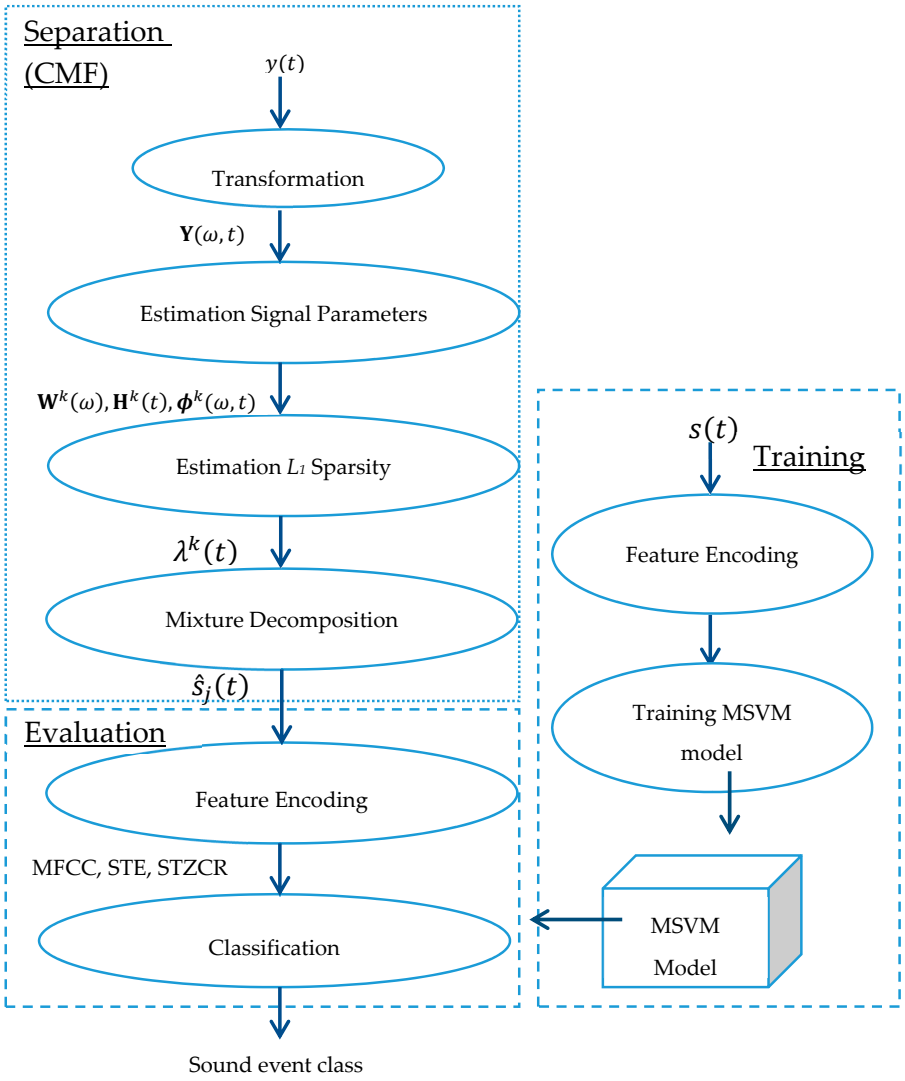


Figure 1. Signal flow of the proposed method.

2.1. Single-Channel Sound Event Separation

The problem formulation in time-frequency (TF) representation is given by an observed complex spectrum, $\mathbf{Y}_{f,t} \in \mathbb{C}$, to estimate the optimal parameters $\theta = \{\mathbf{W}, \mathbf{H}, \boldsymbol{\phi}\}$ of the model. A new factorization algorithm named as the adaptive L_1-sparse complex non-negative matrix factorization (adaptive L_1-SCMF) is derived in the following section. The generative model is given by

$$\mathbf{Y}(\omega,t) = \sum_{k=1}^{K} \mathbf{W}^k(\omega)\mathbf{H}^k(t)\mathbf{Z}^k(\omega,t) = \mathbf{X}(\omega,t) + \epsilon(\omega,t) \tag{1}$$

where $\mathbf{Z}^k(\omega,t) = e^{j\boldsymbol{\Phi}^k(\omega,t)}$ and the reconstruction error $\boldsymbol{\epsilon}(\omega,t) \sim \mathcal{N}_{\mathbb{C}}\left(0,\sigma^2\right)$ is assumed to be independently and identically distributed (i.i.d.) with white noise having zero mean and variance σ^2. The term $\boldsymbol{\epsilon}(\omega,t)$ is used to denote a modeling error for each source. The likelihood of $\theta = \{\mathbf{W}, \mathbf{H}, \boldsymbol{\phi}\}$ is thus written as

$$P(\mathbf{Y}|\theta) = \prod_{f,t} \frac{1}{\pi\sigma^2} \exp\left(-\frac{\left|\mathbf{Y}(\omega,t) - \mathbf{X}(\omega,t)\right|^2}{\sigma^2}\right) \tag{2}$$

It is assumed that the prior distributions for $\mathbf{W}$, $\mathbf{H}$, and $\boldsymbol{\phi}$ are independent, which yields

$$P(\theta|\boldsymbol{\lambda}) = P(\boldsymbol{W})P(\boldsymbol{H}|\boldsymbol{\lambda})P(\boldsymbol{\phi}) \tag{3}$$

The prior $P(\mathbf{H}|\boldsymbol{\lambda})$ corresponds to the sparsity cost, for which a natural choice is a generalized Gaussian prior. When $p = 1$, $P(\mathbf{H}|\boldsymbol{\lambda})$ promotes the L_1-norm sparsity. L_1-norm sparsity has been shown to be probabilistically equivalent to the pseudo-norm, L_0, which is the theoretically optimum sparsity [33,34]. However, L_0-norm is non-deterministic polynomial-time (NP) hard and is not useful in large datasets such as audio. Given Equation (3), the posterior density [35,36] is defined as the maximum a posteriori probability (MAP) estimation problem, which leads to minimizing the following optimization problem with respect to θ. Equations of Gaussian prior and maximum a posteriori probability (MAP) estimation are expressed in Appendix A.

The CMF parameters have been upgraded by using an efficient auxiliary function for an iterative process. The auxiliary function for $f(\theta)$ can be expressed as the following: for any auxiliary variables with $\sum_k \overline{\mathbf{Y}}^k(\omega,t) = \mathbf{Y}(\omega,t)$, for any $\beta^k(\omega,t) > 0$, $\sum_k \beta^k(\omega,t) = 1$, for any $\mathbf{H}^k(t) \in \mathcal{R}$, $\overline{\mathbf{H}}^k(t) \in \mathcal{R}$, and $p = 1$. The term $f(\theta) \leq f^+\left(\theta,\overline{\theta}\right)$ with an auxiliary function was defined as:

$$f^+\left(\theta,\overline{\theta}\right) \equiv \sum_{f,k,t} \frac{\left|\overline{\mathbf{Y}}^k(\omega,t) - \mathbf{W}^k(\omega)\mathbf{H}^k(t)\cdot e^{j\boldsymbol{\phi}^k(\omega,t)}\right|^2}{\beta^k(\omega,t)} + \sum_{k,t}\left[\left(\lambda^k(t)\right)^p\left(p\left|\overline{\mathbf{H}}^k(t)\right|^{p-2}\mathbf{H}^k(t)^2 + (2-p)\left|\overline{\mathbf{H}}^k(t)\right|^p\right) - \log\lambda^k(t)\right] \tag{4}$$

where $\overline{\theta} = \left\{\overline{\mathbf{Y}}^k(\omega,t), \overline{\mathbf{H}}^k(t) \middle| 1 \leq f \leq F,\ 1 \leq t \leq T,\ 1 \leq k \leq K\right\}$. The function $f^+\left(\theta,\overline{\theta}\right)$ is minimized w.r.t. $\overline{\theta}$ when

$$\overline{\mathbf{Y}}^k(\omega,t) = \mathbf{W}^k(\omega)\overline{\mathbf{H}}^k(t)\cdot e^{j\boldsymbol{\phi}^k(\omega,t)} + \beta^k(\omega,t)(\mathbf{Y}(\omega,t) - \mathbf{X}(\omega,t)) \tag{5}$$

$$\overline{\mathbf{H}}^k(t) = \mathbf{H}^k(t) \tag{6}$$

2.2. Formulation of Proposed CMF Based Adaptive Variable Regularization Sparsity

2.2.1. Estimation of the Spectral Basis and Temporal Code

In Equation (4), the update rule for θ is derived by differentiating $f^{+}\left(\theta,\overline{\theta}\right)$ partially w.r.t. $\mathbf{W}^{k}(\omega)$ and $\mathbf{H}^{k}(t)$, and setting them to zero, which yields the following:

$$\mathbf{W}^{k}(\omega) = \frac{\sum_{t}\frac{\mathbf{H}^{k}(t)}{\beta^{k}(\omega,t)}Re\left[\overline{\mathbf{Y}}^{k}(\omega,t)^{*}\cdot e^{j\phi^{k}(\omega,t)}\right]}{\sum_{t}\frac{\mathbf{H}^{k}(t)^{2}}{\beta^{k}(\omega,t)}} \tag{7}$$

$$\mathbf{H}^{k}(t) = \frac{\sum_{f}\frac{\mathbf{W}^{k}(\omega)}{\beta^{k}(\omega,t)}Re\left[\overline{\mathbf{Y}}^{k}(\omega,t)^{*}\cdot e^{j\phi^{k}(\omega,t)}\right]}{\sum_{f}\frac{\mathbf{W}^{k}(\omega)^{2}}{\beta^{k}(\omega,t)} + \left(\lambda^{k}(t)\right)^{p} p\left|\overline{\mathbf{H}}^{k}(t)\right|^{p-2}} \tag{8}$$

The update rule for the phase, $\phi^{k}(\omega,t)$, can be derived by reformulating Equation (4) as follows:

$$\begin{aligned} f^{+}\left(\theta,\overline{\theta}\right) &= \sum_{k,f,t}\frac{\left|\overline{\mathbf{Y}}^{k}(\omega,t)\right|^{2} - 2\mathbf{W}^{k}(\omega)\mathbf{H}^{k}(t)\mathrm{Re}\left[\overline{\mathbf{Y}}^{k}(\omega,t)\cdot e^{-j\phi^{k}(\omega,t)}\right] + \mathbf{W}^{k}(\omega)^{2}\mathbf{H}^{k}(t)^{2}}{\beta^{k}(\omega,t)} + \sum_{k,t}\lambda^{k}(t)\left(\left|\overline{\mathbf{H}}^{k}(t)\right|^{-1}\mathbf{H}^{k}(t)^{2} - \overline{\mathbf{H}}^{k}(t)\right) - \sum_{k,t}\log\lambda^{k}(t) \\ &= A - 2\sum_{k,f,t}\left|\mathbf{B}^{k}(\omega,t)\right|\cos\left(\phi^{k}(\omega,t) - \mathbf{\Omega}^{k}(\omega,t)\right) \end{aligned} \tag{9}$$

where A denotes the terms that are irrelevant with $\phi^{k}(\omega,t)$, $\mathbf{B}^{k}(\omega,t) = \frac{\mathbf{W}^{k}(\omega)\mathbf{H}^{k}(t)\overline{\mathbf{Y}}^{k}(\omega,t)}{\beta^{k}(\omega,t)}$, $\cos\Omega^{k}(\omega,t) = \frac{Re\left[\overline{\mathbf{Y}}^{k}(\omega,t)\right]}{\left|\overline{\mathbf{Y}}^{k}(\omega,t)\right|}$, and $\sin\mathbf{\Omega}^{k}(\omega,t) = \frac{\mathrm{Im}\left[\overline{\mathbf{Y}}^{k}(\omega,t)\right]}{\left|\overline{\mathbf{Y}}^{k}(\omega,t)\right|}$. Derivation of (9) is elucidated in Appendix B. The auxiliary function, $f^{+}\left(\theta,\overline{\theta}\right)$ in Equation (4) is minimized when $\cos\left(\phi^{k}(\omega,t) - \Omega^{k}(\omega,t)\right) = \cos\phi^{k}(\omega,t)\cos\Omega^{k}(\omega,t) + \sin\phi^{k}(\omega,t)\sin\Omega^{k}(\omega,t) = 1$, namely, $\cos\phi^{k}(\omega,t) = \cos\Omega^{k}(\omega,t)$ and $\sin\phi^{k}(\omega,t) = \sin\Omega^{k}(\omega,t)$. The update formula for $e^{j\phi^{k}(\omega,t)}$ eventually leads to

$$\begin{aligned} e^{j\phi^{k}(\omega,t)} &= \cos\phi^{k}(\omega,t) + j\sin\phi^{k}(\omega,t) \\ &= \frac{\overline{\mathbf{Y}}^{k}(\omega,t)}{\left|\overline{\mathbf{Y}}^{k}(\omega,t)\right|} \end{aligned} \tag{10}$$

The update formula for $\beta^{k}(\omega,t)$ and $\mathbf{H}^{k}(t)$ for projection onto the constraint space is set to

$$\beta^{k}(\omega,t) = \frac{\mathbf{W}^{k}(\omega)\mathbf{H}^{k}(t)}{\sum_{k}\mathbf{W}^{k}(\omega)\mathbf{H}^{k}(t)} \tag{11}$$

$$\mathbf{H}^{k}(t) \leftarrow \frac{\mathbf{H}^{k}(t)}{\sum_{k}\mathbf{H}^{k}(t)} \tag{12}$$

2.2.2. Estimation of L_1-Optimal Sparsity Parameter $\lambda^{k}(t)$

This section aims to facilitate spectral dictionaries with adaptive sparse coding. First, the CMF model is defined as the following terms:

$$\overline{\mathbf{W}} = \left[\mathbf{I}\otimes\mathbf{W}^1(\omega)\vdots\mathbf{I}\otimes\mathbf{W}^2(\omega)\vdots\cdots\vdots\mathbf{I}\otimes\mathbf{W}^K(\omega)\right],$$
$$e^{j\overline{\underline{\mathbf{\Phi}}}(t)} = \left[e^{j\underline{\mathbf{\Phi}}^1(t)}\vdots\cdots\vdots e^{j\underline{\mathbf{\Phi}}^K(t)}\right]$$
$$\underline{\mathbf{y}} = \mathrm{vec}(\mathbf{Y}) = \begin{bmatrix}\underline{\mathbf{Y}}^1(:)\\ \cdots\\ \underline{\mathbf{Y}}^2(:)\\ \cdots\\ \vdots\\ \cdots\\ \underline{\mathbf{Y}}^K(:)\end{bmatrix}, \underline{\mathbf{h}} = \begin{bmatrix}\mathbf{H}^1(t)\\ \cdots\\ \mathbf{H}^2(t)\\ \cdots\\ \vdots\\ \cdots\\ \mathbf{H}^K(t)\end{bmatrix}, \underline{\boldsymbol{\lambda}} = \begin{bmatrix}\lambda^1(t)\\ \cdots\\ \lambda^2(t)\\ \cdots\\ \vdots\\ \cdots\\ \lambda^K(t)\end{bmatrix}, \underline{\mathbf{\Phi}} = \begin{bmatrix}\mathbf{\Phi}^1(:,t)\\ \cdots\\ \mathbf{\Phi}^2(:,t)\\ \cdots\\ \vdots\\ \cdots\\ \mathbf{\Phi}^K(:,t)\end{bmatrix} \quad (13)$$
$$\overline{\mathbf{A}} = \begin{bmatrix}\overline{\mathbf{W}}\circ e^{j\overline{\underline{\mathbf{\Phi}}}(t)} & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{0} & \overline{\mathbf{W}}\circ e^{j\overline{\underline{\mathbf{\Phi}}}(t)} & \mathbf{0} & \vdots\\ \vdots & \mathbf{0} & \overline{\mathbf{W}}\circ e^{j\overline{\underline{\mathbf{\Phi}}}(t)} & \mathbf{0}\\ \mathbf{0} & \cdots & \mathbf{0} & \overline{\mathbf{W}}\circ e^{j\overline{\underline{\mathbf{\Phi}}}(t)_t}\end{bmatrix}$$

where "$\otimes$" and "$\circ$" are the Kronecker product and the Hadamard product, respectively. The term vec$(\cdot)$ denotes the column vectorization and the term $\mathbf{I}$ is the identity matrix. The goal is then set to compute the regularization parameter $\lambda^k(t)$ related to each $\mathbf{H}^k(t)$. To achieve the goal, the parameter p in Equation (4) is set to 1 to acquire a linear expression (in $\lambda^k(t)$). In consideration of the noise variance σ^2, Equation (4) can concisely be rewritten as:

$$F(\underline{\mathbf{h}},\underline{\boldsymbol{\lambda}}) = \frac{1}{2\sigma^2}\parallel\underline{\mathbf{y}} - \overline{\mathbf{A}}\underline{\mathbf{h}}\parallel_F^2 + \underline{\boldsymbol{\lambda}}^{\mathbf{T}}\underline{\mathbf{h}} - (\log\underline{\boldsymbol{\lambda}})^{\mathbf{T}}\underline{\mathbf{1}} \quad (14)$$

where the $\underline{\mathbf{h}}$ and $\underline{\boldsymbol{\lambda}}$ terms indicate vectors of dimension $R\times 1$ (i.e., $R = F\times T\times K$), and the superscript '$\mathbf{T}$' is used to denote complex Hermitian transpose (i.e., vector (or matrix) transpose followed by complex conjugate). The Expectation–Maximization (EM) algorithm will be used to determine $\underline{\boldsymbol{\lambda}}$ and $\underline{\mathbf{h}}$ is the hidden variable where the log-likelihood function can be optimized with respect to $\underline{\boldsymbol{\lambda}}$. The log-likelihood function satisfies the following [12]:

$$\ln p\left(\underline{\mathbf{y}}\left|\underline{\boldsymbol{\lambda}},\overline{\mathbf{A}},\sigma^2\right.\right) \geq \int Q(\underline{\mathbf{h}})\ln\left(\frac{p\left(\underline{\mathbf{y}},\underline{\mathbf{h}}\left|\underline{\boldsymbol{\lambda}},\overline{\mathbf{A}},\sigma^2\right.\right)}{Q(\underline{\mathbf{h}})}\right)d\underline{\mathbf{h}} \quad (15)$$

by applying the Jensen's inequality for any distribution $Q(\underline{\mathbf{h}})$. The distribution can simply verify the posterior distribution of $\underline{\mathbf{h}}$, which maximizes the right-hand side of Equation (15), is given by $Q(\underline{\mathbf{h}}) = p\left(\underline{\mathbf{h}}\left|\underline{\mathbf{y}},\underline{\boldsymbol{\lambda}},\overline{\mathbf{A}},\sigma^2\right.\right)$. The posterior distribution in the form of the Gibbs distribution is proposed as follows:

$$Q(\underline{\mathbf{h}}) = \frac{1}{Z_h}\exp[-F(\underline{\mathbf{h}})] \text{ where } Z_h = \int\exp[-F(\underline{\mathbf{h}})]d\underline{\mathbf{h}} \quad (16)$$

The term $F(\underline{\mathbf{h}})$ in Equation (16) as the function of the Gibbs distribution is essential for simplifying the adaptive optimization of $\underline{\boldsymbol{\lambda}}$. The maximum-likelihood (ML) estimation of $\underline{\boldsymbol{\lambda}}$ can be decomposed as follows:

$$\underline{\boldsymbol{\lambda}}^{ML} = \arg\max_{\underline{\boldsymbol{\lambda}}}\int Q(\underline{\mathbf{h}})\ln p(\underline{\mathbf{h}}|\underline{\boldsymbol{\lambda}})d\underline{\mathbf{h}} \quad (17)$$

In the same way,

$$\sigma^2{}_{ML} = \arg\max_{\sigma^2}\int Q(\underline{\mathbf{h}})\ln p\left(\underline{\mathbf{y}}\left|\underline{\mathbf{h}},\overline{\mathbf{A}},\sigma^2\right.\right)d\underline{\mathbf{h}} \quad (18)$$

Individual element of $\mathbf{H}$ is required to be exponentially distributed with independent decay parameters that delivers $p(\underline{\mathbf{h}}|\underline{\lambda}) = \prod_g \lambda_g \exp\left(-\lambda_g h_g\right)$, thus Equation (17) obtains

$$\underline{\lambda}^{ML} = \arg\max_{\underline{\lambda}} \int Q(\underline{\mathbf{h}})\left(\ln \lambda_g - \lambda_g h_g\right) d\underline{\mathbf{h}} \tag{19}$$

The term $\underline{\mathbf{h}}$ denotes the dependent variable of the distribution $Q(\underline{\mathbf{h}})$, whereas other parameters are assumed to be constant. As such, the $\underline{\lambda}$ optimization in Equation (19) is derived by differentiating the parameters within the integral with respect to $\underline{\mathbf{h}}$. As a result, the functional optimization [37] of $\underline{\lambda}$ then obtains

$$\lambda_g = \frac{1}{\int h_g Q(\underline{\mathbf{h}}) d\underline{\mathbf{h}}} \tag{20}$$

where $g = 1, 2, \ldots, R$, λ_g denotes the g^{th} element of $\underline{\lambda}$. Notice that the solution $\underline{\mathbf{h}}$ naturally splits its elements into distinct subsets $\underline{\mathbf{h}}_M$ and $\underline{\mathbf{h}}_P$, consisting of components $\forall_m \in M$ so that $h_m > 0$ and components $\forall_p \in P$ so that $h_P = 0$. The sparsity parameter is then obtained as presented in Equation (21):

$$\lambda_g = \begin{cases} \frac{1}{\int h_g Q_M(\underline{\mathbf{h}}_M) d\underline{\mathbf{h}}_M} = \frac{1}{h_g^{\text{MAP}}} \text{ if } g \in M \\ \frac{1}{\int h_g \hat{Q}_P(\underline{\mathbf{h}}_P) d\underline{\mathbf{h}}_P} = \frac{1}{u_g} \text{ if } g \in P \end{cases} \tag{21}$$

and its covariance X is given by

$$X_{ab} = \begin{cases} \left(\overline{\mathbf{C}}_P^{-1}\right)_{ab}, \text{ if } a, b \in M \\ u_p^2 \delta_{ab}, \text{ Otherwise.} \end{cases} \tag{22}$$

where $\hat{Q}_P\left(\underline{\mathbf{h}}_P \geq 0\right) = \prod_{p \in P} \frac{1}{u_p} \exp\left(\frac{-h_p}{u_p}\right)$, $\overline{\mathbf{C}}_P = \frac{1}{\sigma^2} \overline{\mathbf{A}}_P^{\mathbf{T}} \overline{\mathbf{A}}_P$ and $u_p \leftarrow u_p \frac{-\hat{b}_p + \sqrt{\hat{b}_p^2 + 4 \frac{(\hat{\mathbf{C}}\underline{\mathbf{u}})_p}{u_p}}}{2(\hat{\mathbf{C}}\underline{\mathbf{u}})_p}$. The function $Q_M\left(\underline{\mathbf{h}}_M\right)$ will be expressed as the unconstrained Gaussion with mean $\underline{\mathbf{h}}_M^{\text{MAP}}$ and covariance $\overline{\mathbf{C}}_M^{-1}$ based on a multivariate Gaussian distribution. Similarly, the inference for σ^2 can be computed as

$$\sigma^2 = \frac{1}{N_0} \int Q(\underline{\mathbf{h}})\left(\|\underline{\mathbf{y}} - \overline{\mathbf{A}}\underline{\mathbf{h}}\|^2\right) d\underline{\mathbf{h}} \tag{23}$$

where

$$\hat{h}_g = \begin{cases} h_g^{\text{MAP}} \text{ if } g \in M \\ u_g \text{ if } g \in P \end{cases}$$

The core procedure of the proposed CMF method is based on L_1-optimal sparsity parameters. The estimated sources are discovered by multiplying the respective rows of the $W^k(\omega)$ components with the corresponding columns of the $H^k(t)$ weights and time-varying phrase spectrum $e^{j\phi^k(\omega,t)}$. The separated source $\hat{s}_j(t)$ is obtained by converting the time-frequency represented sources into the time domain. Derivation of L_1-optimal sparsity parameter, is elucidated in the Appendix C.

2.3. Sound Event Classification

Once the separated sound signal is obtained, the next step is to identify the sound event. A multiclass support vector machine (MSVM) is employed to achieve the goal. The MSVM is comprised of two phases: the learning phase and the evaluation phase. The MSVM is based on one versus one strategy (OvsO) that splits observed c classes into $\frac{c(c-1)}{2}$ binary classification sub-problems. To train the $\sqsupseteq^{th}$ MSVM model, the MSVM will construct hyperplanes for discriminating each observed data into its corresponding class by executing the series of binary classification. Starting from the learning phase, sound signatures are extracted from the training dataset in the time-frequency domain. The sound signatures that were studied in this research were the Mel frequency cepstral coefficients (MFCC: MF), short-time energy (STE: $E(t)$), and short term zero-crossing rate (STZCR: $STZ(t)$), which can be orderly expressed as: $MF = 2525 \times \log[1 + (f/7)]$, $E(t) = \sum_{\tau=-\infty}^{\infty} [y(t) \cdot f_w(t-\tau)]^2$, $Z(t) = \sum_{\tau=-\infty}^{\infty} \Big| sgn[s(\tau)] - sgn[s(\tau-1)] \Big| \cdot f_w(t-\tau)$ where $f_w(t)$ denotes the windowing function. The training signals are segmented into small blocks, then the individual block is extracted to the three signature parameters. The mean supervector is then computed as an average of individual feature of all blocks for each sound event input. Thus, the mean feature supervector (O) with a corresponding sound-event-label vector ((w)) is paired together (i.e., ($\psi(O, w)$)) and supplied to the MSVM model. The discriminant formula can be expressed as:

$$\begin{aligned} \{\hat{\sqsupseteq}, \beta\} &= \arg\max_{\sqsupseteq} \Big\{ \alpha_{\sqsupseteq}^T \psi(O, w; \beta) \Big\} \\ &= \arg\max_{\sqsupseteq} \Bigg\{ \max_{\beta} \sum_{i=1}^{|w|} \alpha_{\sqsupseteq}^T \psi\Big(O_{i|\beta}, w_i\Big) \Bigg\} \end{aligned} \tag{24}$$

where $\Big(O_{i|\beta}, w_i\Big), i = 1, \ldots, M$ represents the i^{th} separated sound signals; the weight vector $\alpha_{\sqsupseteq}$ is employed for individual class $\sqsupseteq$ to compute a discriminant score for the O; the i term is the index of the block order (β); and the function $\alpha_{\sqsupseteq}^T \psi(O, w; \beta)$ measures a linear discriminant distance of the hyperplane with the extracted feature vector from the observed data. The MSVM based OvsO strategy for class $\sqsupseteq^{th}$ and other, the hyperplane, can be maximized as $\alpha_{\sqsupseteq}^T \psi(O, w; \beta) + b_{\sqsupseteq}$ and can then be learned via the following equation as

$$\min_{\alpha_{\sqsupseteq}, \xi_{\sqsupseteq}} \frac{1}{2} \|\alpha_{\sqsupseteq}\|^2 + C \sum_{i=1}^{M} \xi_i^{\sqsupseteq} \tag{25}$$

where $\xi_i^{\sqsupseteq} \geq 0$, $b_{\sqsupseteq}$ is a constant. The term $\sum_{i=1}^{M} \xi_i^{\sqsupseteq}$ denotes a penalty function for tradeoff between a large margin and a small error penalty. The optimal hyperplane can be determined by minimizing $\frac{1}{2}\|\alpha_{\sqsupseteq}\|^2$ to maximize the condition (i.e., $\alpha_{\sqsupseteq}^T \psi(O, w; \beta) + b_{\sqsupseteq} \geq 1 - \xi_i^{\sqsupseteq}$). If the conditional term is greater than $1 - \xi_i^{\sqsupseteq}$, then the estimated sound event belongs to the $\sqsupseteq^{th}$ class. Otherwise, the estimated sound event classifies into other classes.

The overview of the proposed algorithm is presented in the following table as Algorithm 1.

Algorithm 1 Overview of the proposed algorithm.

(1) Compute $\mathbf{Y}(\omega,t) = \mathrm{STFT}(y(t))$ from the noisy single-channel mixture $y(t)$.

(2) Initial values $\mathbf{W}^k(\omega), \mathbf{H}^k(t),\ \beta^k(\omega,t)$, fixing the value of $\boldsymbol{\phi}^k(\omega,t)$ at $e^{j\boldsymbol{\Phi}^k(\omega,t)} = \frac{\mathbf{Y}(\omega,t)}{|\mathbf{Y}(\omega,t)|}$, and calculate $\lambda^k(t)$ and σ^2 .

(3) Update $\overline{\theta} = \left\{\overline{\mathbf{X}}, \overline{\mathbf{H}}\right\}, \theta = \left\{\mathbf{W}^k(\omega), \mathbf{H}^k(t),\ \boldsymbol{\phi}^k(\omega,t)\right\}, \beta^k(\omega,t)$.

(4) Update parameters (21) and (23) until convergence is reached as determined by the rate of change of the parameters update falling within a pre-determined threshold.

(5) Estimation of each source by multiplying the respective rows of the spectral components $\boldsymbol{W}^k(\omega)$ with the corresponding columns of the mixture weights $\boldsymbol{H}^k(t)$ and time-varying phrase spectrum $e^{j\boldsymbol{\phi}^k(\omega,t)}$. (i.e., $\left|\hat{\mathbf{S}}_i\right|^{.2} = \sum_{k=1}^{K_i} \mathbf{W}_{if}^k \mathbf{H}_{it}^k \cdot e^{j\phi_{if,t}^{(k)}}$ and construct the binary TF mask for the i^{th} source

$M_i(f,t_s) := \begin{cases} 1, & \text{if } \left|\hat{\mathbf{S}}_i(f,t_s)\right|^{.2} > \left|\hat{\mathbf{S}}_j(f,t_s)\right|^{.2},\ i \neq j \\ 0, & otherwise \end{cases}$).

(6) Convert the time-frequency represented sources into time domain to obtain the separated sources $\hat{s}_j(t)$ i.e., $\hat{s}_j(t) = STFT^{-1}\left(\left|\hat{\mathbf{S}}_i\right|^2\right)$.

(7) Classify the $\sqsupseteq^{th}$ sound event by computing the optimal hyperplane $\alpha_{\sqsupseteq}^T \psi(O,w;\beta) + b_{\sqsupseteq}$ by minimizing the following equation: $\min_{\alpha_{\sqsupseteq},\xi_{\sqsupseteq}} \frac{1}{2}\|\alpha_{\sqsupseteq}\|^2 + C\sum_{i=1}^{M} \xi_i^{\sqsupseteq}$.

3. Experimental Results and Analysis

The performance was evaluated on recorded sound-event signals in a low noisy environment at 20 signal-to-noise ratios (SNRs). The sound-event database had a total of 500 recorded signals containing four event classes: speech (SP), door open (DO), door knocking (DK), and footsteps (FS). An overview of the experimental setup is given as the following: all signals had a 16-bit resolution and a sampling frequency of 44.1 KHz. A 2048 length of Hanning window with 50% overlap was used for signal processing. Nonlinear SVM with a Gaussian RBF kernel was used for constructing the MSVM learning model. Other kernels such as polynomials, sigmoid, and even linear function were tested, but the best performance was delivered by the Gaussian kernel. A 4-fold cross-validation strategy was used in the training phase for tuning the classifier parameters when using 80% of the recorded signals (n = 400) from the sound-event database.

The performance of the proposed noisy sound separation and event classification (NSSEC) method was demonstrated and presented into the following two sections: (1) the separating performance, and (2) the MSVM classifier.

3.1. Sound-Event Separation and Classification Performance

Event mixtures consist of two sound-event signals in low noisy environment at 20 dB SNRs. A hundred sound-event signals of four classes were randomly selected and then mixed to generate 120 mixtures of six types (i.e., DO + DK, DO + FS, DO + SP, DK + FS, DK + SP, and FS + SP). The separation performance measured the signal-to-distortion ratio (SDR) (i.e., $SDR = 10\ log_{10}\left(\|s_{target}\|^2 / \|e_{interf} + e_{noise} + e_{artif}\|^2\right)$ where e_{interf}, e_{noise}, and e_{artif}). The SDR represents the ratio of the magnitude distortion of the original signal by the interference from other sources. The proposed separation method was compared with the state-of-the-art NMF approach (i.e., CMF [38], NMF-ISD [14,39], and SNMF [40–42] methods). The cost function was the least squares with 500 maximum number of iterations.

3.1.1. Variational Sparsity Versus Fixed Sparsity

In this implementation, several experiments were conducted to investigate the effect of sparsity regularization on source separation performance. The proposed separation method was evaluated by variational sparsity in the case of (1) uniform constant sparsity with low sparseness e.g., $\lambda_t^k = 0.01$ and (2) uniform constant sparsity with high sparseness (e.g., $\lambda_t^k = 100$). The hypothesis is that the proposed variational sparsity will significantly yield improvement of the audio source separation when compared with fixed sparsity.

To investigate the impact of uniform sparsity parameter, the set of sparsity regularization values from 0 to 10 with a 0.5 interval were determined for each experiment of 60 mixtures of six types. Results of the uniform regularization given by various sparsity (i.e., $\lambda_t^k = 0, 0.5, \ldots, 10$) is illustrated in Figure 2.

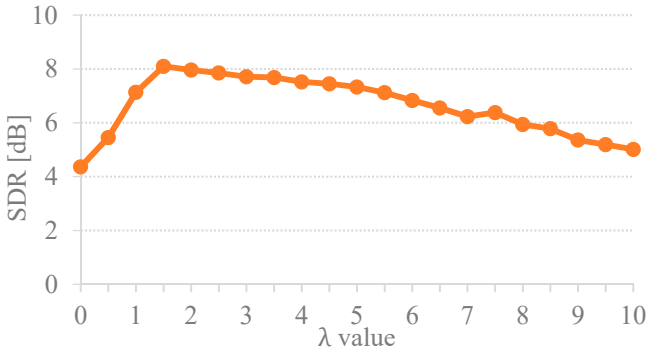


Figure 2. Separation results of the proposed method by using different uniform regularization.

Figure 2 illustrates that the best performance of the unsupervised CMF was in a range of 1.5–3, which yielded the highest SDR of over 8dB. When the term λ_t^k was set too high, the low spectral values of sound-event signals were overly sparse. This overfitting sparsity $\boldsymbol{H}^k(t)$ caused the separation performance toward a tendency to degrade. Conversely, the underfitting sparsity $\boldsymbol{H}^k(t)$ occurred when the term λ_t^k was set too low. The coding parameter $\boldsymbol{H}^k(t)$ could not distinguish between the two sound-event signals. It was also noticed that if the factorization is non-regularized, this will cause the separation results to contain a mixed sound. According to the uniform sparsity results in Figure 2, the separation performance of the proposed method varies depending on the assigned sparsity values. Thus, it is challenging to find a solution for the indistinctness among the sound-event sources in the TF representation to determine the optimal value of sparseness. Thus, this introduces the importance of determining the optimal λ for separation. Table 1 presents the essential sparsity value on the separation performance by comparing the proposed method given by variational sparsity against the uniform sparsity scheme. The average performance improvement of the proposed adaptive CMF method against the uniform constant sparsity was 1.32 dB SDR. The SDR results clearly indicate that the adaptive sparsity yielded the surpass separation performance over the constant sparsity scheme. Hence, the proposed variational sparsity improves the performance of the discovered original sound-event signals by adaptively selecting the appropriate sparsity parameters to be individually adapted for each element code (i.e., $\lambda_g = \begin{cases} \frac{1}{\int h_g Q_M(\underline{\mathbf{h}}_M) d\underline{\mathbf{h}}_M} = \frac{1}{h_g^{\mathrm{MAP}}} \text{ if } g \in M \\ \frac{1}{\int h_g \hat{Q}_P(\underline{\mathbf{h}}_P) d\underline{\mathbf{h}}_P} = \frac{1}{u_g} \text{ if } g \in P \end{cases}$ and $\sigma^2 = \frac{1}{N_0} \int Q(\underline{\mathbf{h}}) \left(\|\underline{\mathbf{y}} - \overline{\mathbf{A}}\underline{\mathbf{h}}\|^2 \right) d\underline{\mathbf{h}}$ where $\hat{h}_g = \begin{cases} h_g^{\mathrm{MAP}} \text{ if } g \in M \\ u_g \text{ if } g \in P \end{cases}$). Consequently, the optimal sparsity facilitates the estimated spectral dictionary via the estimated temporal code. The quantitative measures of separation performance were performed to assess the proposed single-channel sound event separation method. The overall average signal-to-distortion ratio (SDR) was 8.62 dB as illustrated in Figure 3.

Table 1. Comparison of average SDR performance on three types of mixtures between uniform regularization methods and the proposed method.

Mixtures	Methods	SDR
DO + DK	Proposed method	7.63
	(Best) Uniform regularization sparsity	6.59
DO + FS	Proposed method	9.06
	(Best) Uniform regularization sparsity	8.74
DO + SP	Proposed method	8.45
	(Best) Uniform regularization sparsity	6.91
DK + FS	Proposed method	7.04
	(Best) Uniform regularization sparsity	6.35
DK + SP	Proposed method	9.72
	(Best) Uniform regularization sparsity	7.78
FS + SP	Proposed method	9.81
	(Best) Uniform regularization sparsity	7.42

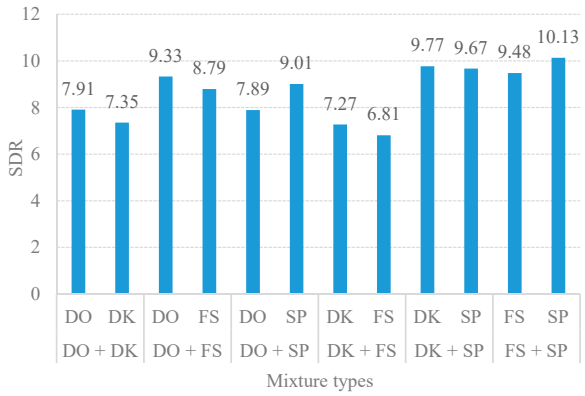


Figure 3. Average SDR results of six-mixture types.

Each sound-event signal has its own temporal pattern that can be clearly noticed in TF representation. Examples of sound-event signals in the TF domain are illustrated in Figure 4. Through the adaptive L_1-SCMF method, the proposed single-channel separation method can generate complex temporal patterns such as speech. Thus, the separation results clearly indicate that the performances of noisy source separation perform with high SDR values.

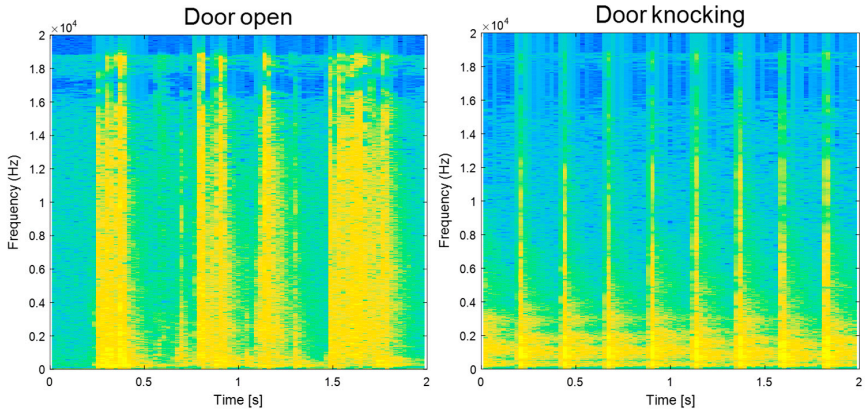


Figure 4. *Cont.*

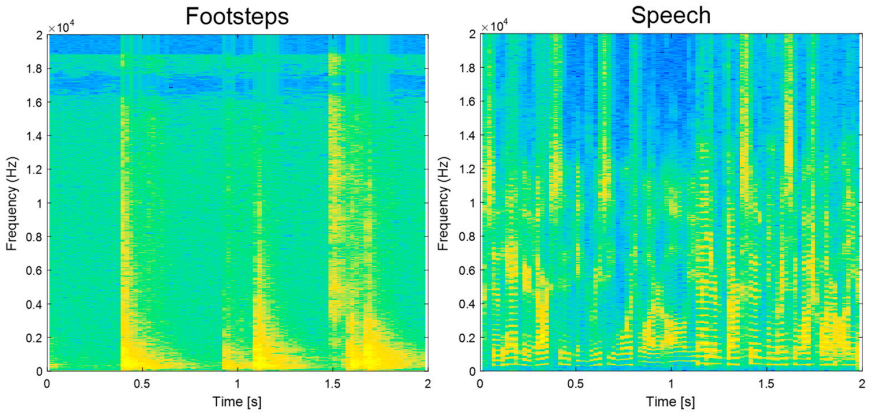


Figure 4. Example of time-frequency representation of four sound event classes.

3.1.2. Comparison of the Proposed Adaptive CMF with Other SCBSS Methods Based on NMF

This section presents the adaptive CMF separating performance against the state-of-the-art NMF methods (i.e., CMF, SNMF, and NMF-ISD). In the compared methods, the experimental variables such as the normalizing time-frequency domain were computed by using the short-time Fourier transform (i.e., 1024-point Hanning window with 50% overlap). The number of factors was two, with a sparsity weight of 1.5. One hundred random realizations of twenty second-event mixtures were executed. As a result, the average SDRs are presented in Table 2. The proposed adaptive CMF method yielded the best separating performance over the CMF, SNMF, and NMF_ISD methods with the average improvement SDR at 2.13 dB. The estimated door open signals obtained the highest SDR among the four event categories.

Table 2. Comparison of average SDR and SIR performance on three types of mixtures between SCICA, NMF-ISD, SNMF, CMF, and the proposed method.

Mixtures	Methods	SDR
Door Open	Proposed method	8.38
	CMF	7.11
	SNMF	6.23
	NMF-ISD	6.17
Door Knocking	Proposed method	8.13
	CMF	7.06
	SNMF	6.52
	NMF-ISD	6.55
Footsteps	Proposed method	8.36
	CMF	7.89
	SNMF	6.62
	NMF-ISD	6.06
Speech	Proposed method	9.60
	CMF	6.73
	SNMF	5.61
	NMF-ISD	5.32

The sparsity parameter was carefully adapted using the proposed adaptive L_1-SCMF method exploiting the phase information and temporal code of the sources, which is inherently ignored by SNMF and NMF-ISD and has led to an improved performance of about 2 dB in SDR. On the other hand, the parts decomposed by the CMF, SNMF, and NMF-ISD methods were unable to capture the phase spectra and the temporal dependency of the frequency patterns within the audio signal.

Additionally, the CMF and NMF-ISD are unique when the signal adequately spans the positive octant. Thus, the rotation of **W** and opposite **H** can obtain the same results. The CMF method can easily be over or under sparse resolution of the factorization due to manually determining the sparsity value.

3.2. *Performance of Event Classification Based on MSVM Algorithm*

This section elucidates the features and performance of the MSVM-learning model. The MSVM-learning model was investigated to obtain the optimal size of the sliding window and then determine the significant features that led to the classification performance. Finally, the efficiency of the MSVM model was evaluated. These topics are presented in order in the following parts.

3.2.1. Determination Optimal Window Length for Feature Encoding

For the MSVM method, sound-event signals are segmented into small blocks for encoding feature parameters by using a fixed-length of the sliding window. The sets of feature vectors are computed using the mean supervector and then loaded to the MSVM model for learning and constructing the hyperplane. The size of blocks can affect the information of the feature vectors, which leads to the classifier performance. The block's size will affect the $\alpha_{\sqsupseteq}$, hence modifying the block size will mark the learning efficiency of the MSVM model. Therefore, in order to obtain the optimal value of $\alpha_{\sqsupseteq}$, the optimal block size was exploited by training the MSVM model given various lengths of window sizes (i.e., 0.5, 1, 1.5, and 2.0 s) to learn the 400 noisy sound-event signals of four event classes with cross-validation.

The experimental results are plotted in Figure 5, where the block size varies from 0.5 to 2.0 with 0.5 increments. The MSVM model of the 1.5 s block size yielded the best sound-event classification at 100% accuracy. The sliding window function benefits from SVM to learn an unknown sound event by generating the set of blocks from the observed event, regarded as a number of observed events. As a result, a set of sound event characteristics were computed for each block (i.e., $O_{i|\beta}, w_i$ in Equation (24)).

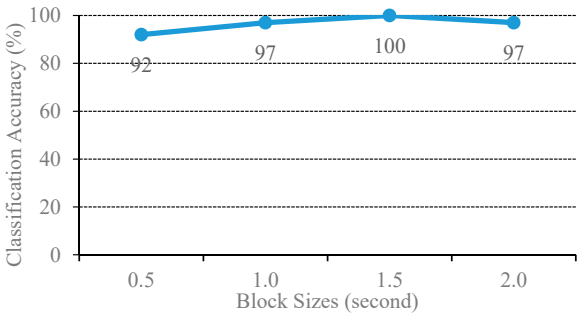


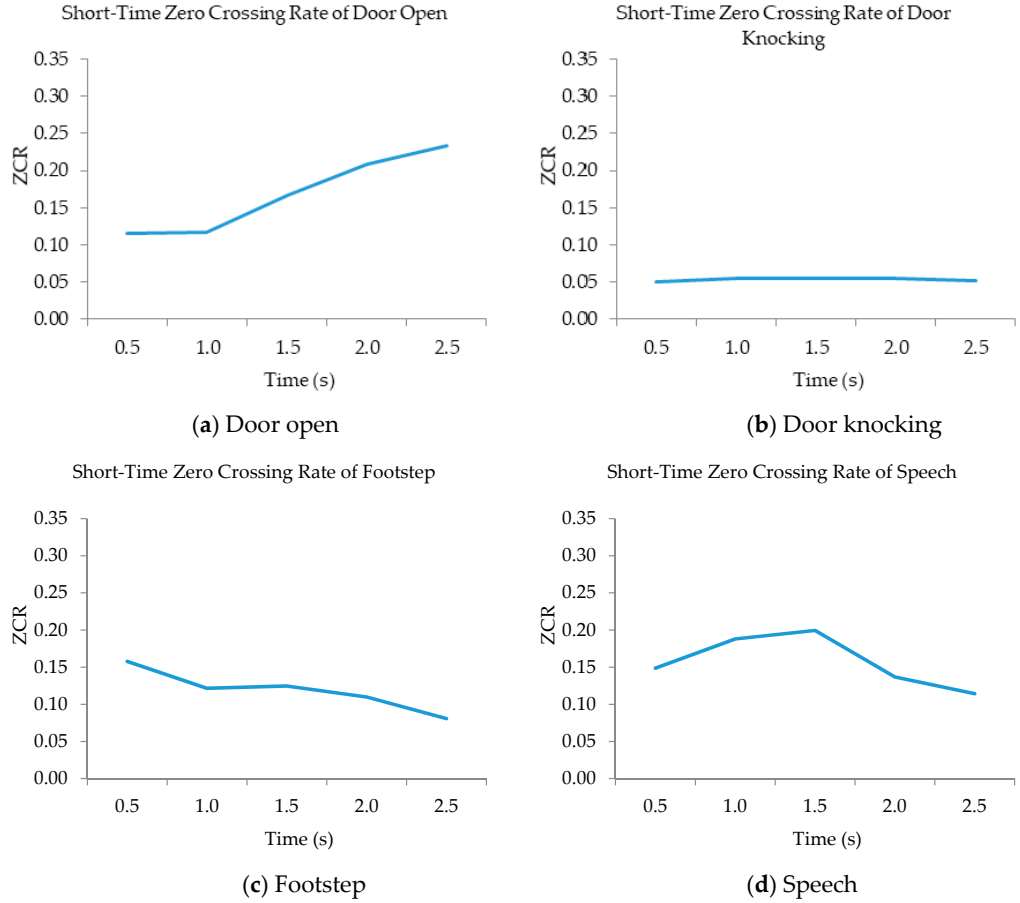
Figure 5. Classification performance of the original and combination source MSVM with various block sizes.

The optimal length of the window size can capture the signature of the sound event. If the window length is too short, the encoded features will then deviate from the character of the sound event. In addition, the mean supervector is computed from the set of features of all blocks, which can be regarded as the mean of the probability distribution of the features. This mean supervector advantages the MSVM to reduce misclassifications when compared to the conventional SVM. Hence, the STFT of all experiments set the window function at 1.5 s.

3.2.2. Determination of Sound-Event Features

Each sound-event signal was encoded with three features: Mel frequency cepstral coefficients (MFCCs), short-time energy (STE), and short-time zero-crossing rate (STZCR). MFCCs are represented as a frequency domain feature that is evaluated in a similar assembly to the human ear (i.e., logarithmic frequency perception). STE is the total spectrum power of an observed event.

The STZCR denotes the number of times that the signal amplitude interval satisfies the condition (i.e., $STZCR = (1/T-1)\sum_{t=1}^{T-1}[[\{s_t s_{t-1} < 0\}$ where $[[\{s_t s_{t-1} < 0\}$ is 1 if the condition is true and 0 otherwise). The STZCR features of four sound-event classes are illustrated in Figure 6.

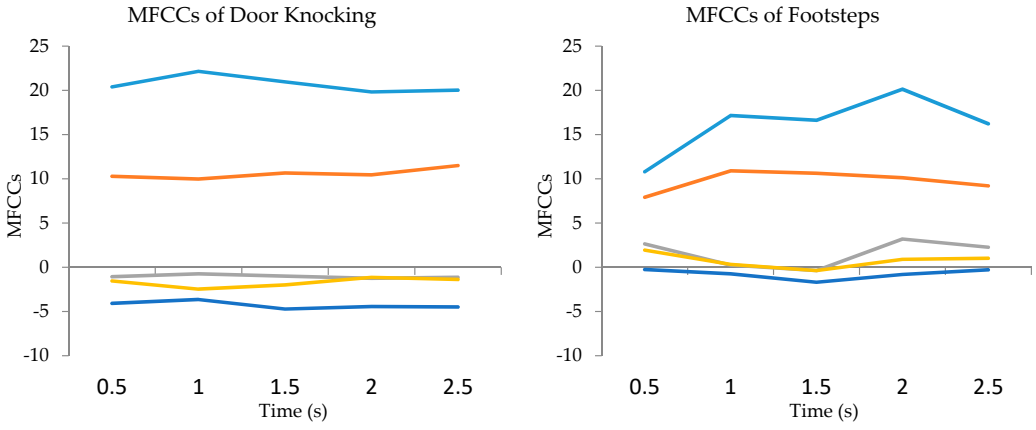


(**a**) Door open (**b**) Door knocking

(**c**) Footstep (**d**) Speech

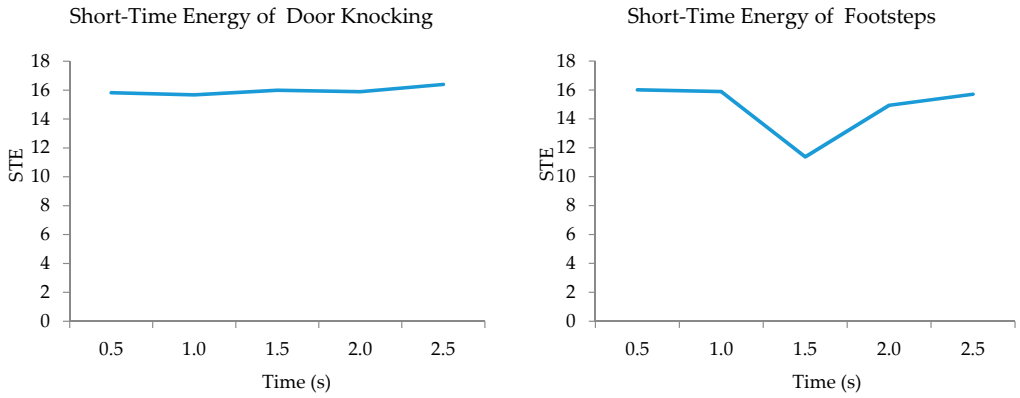
Figure 6. STZCR patterns of four sound-events (**a**–**d**).

The STZCR feature represents unique patterns of four sound-event classes. The four sound-event patterns are different in shape and data range. Similarly, the MCFFs and STE features extract distinctive patterns of all event classes, except for the patterns between door knocking and footstep, as illustrated in Figure 7.



(**a**) MFCCs patterns

Figure 7. *Cont.*



(**b**) STE patterns

Figure 7. MFCCs (**a**) and STE (**b**) patterns of door knocking and footstep.

Figure 7 aims to compare the characteristics of similar sound events such as door knocking and footsteps. Thus, MFCCs and STE features were used to illustrate the patterns of sound events. Figure 7a represents the five orders of MFCC features to compare patterns between door knocking and walking while the STE features are shown in Figure 7b.

The proposed method separated the six categories of mixtures, then classified each estimated sound event signal into its corresponding class. Classified results of the six categories are presented as confusion matrixes below:

	Actual											
Predict	DO	DK		DO	FS		DO	SP		DK	FS	
DO	19	3	DO	12	8	DO	19	5	DK	12	4	
DK	3	15	FS	4	16	SP	3	13	FS	9	15	
	DK	FS		DK	SP		FS	SP				
DK	12	4	DK	16	2	FS	14	6				
FS	9	15	SP	5	17	SP	3	17				

The classification of the proposed method was measured by Precision = TP/(TP + FP), Recall = TP/(TP + FN), and F1-score = 2 × (Precision × Recall)/(Precision + Recall). The TP and TN terms refer to the true positive and true negative, while the FP and FN terms mean false positive and false negative. The scores of Precision, Recall, and F1-score were 0.7667, 0.7731, and 0.7699, respectively.

Each feature represents unique characteristics of an individual sound event. Thus, features were matched into seven cases for exploiting their influence on the MSVM classifiers (i.e., {(MFCC), (STE), (STZCR), (MFCC, STE), (MFCC, STZCR), (STE, STZCR), (MFCC, STE, STZCR)}).

As shown in Figure 8, the MSVM model given by MFCCs and STZCR yielded the best classified accuracy at 100%, with less deviation among the other cases. Therefore, the separated signals were then classified by the proposed MSVM method given by the MFCC and STZCR vectors and the 1.5 s window function. The computational complexity of the proposed method was analyzed by two steps. First, the adaptive L1-SCMF method was NP-hard. Big-O of the adaptive L_1-SCMF method consists of spectral basis (m), temporal code (n), and phase information that rely on components (k). Thus, Big-O of the separation step is $(mn)^{O(k^2)}$. For MSVM steps, it is a polynomial algorithm where Big-O is $O(n^3)$. Therefore, the computational complexity of the proposed method based on Big-O is $(mn)^{O(k^2)}$. All experiments were performed by a PC with Intel® Core™ i7-4510U CPU2.00 GHz and 8 GB RAM. MATLAB was used as the programming platform.

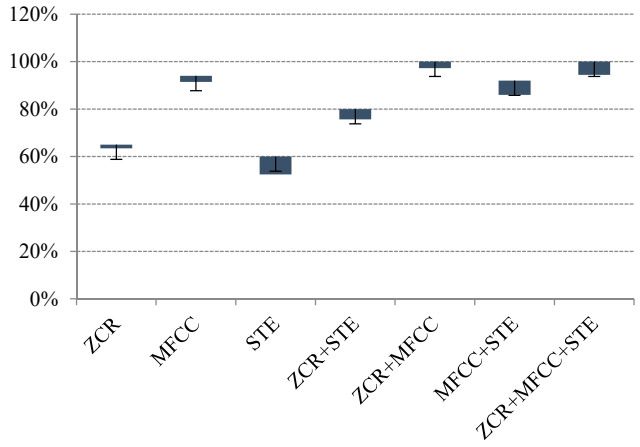


Figure 8. Classification performances of multi-class MSVM of various sets of features and length of event signal.

3.2.3. Performance of MSVM Classifier

The MSVM-classifier performance is presented in terms of percentage of the corrected sound-event classification. The 240 separated signals of four classes from the proposed separation method were individually identified by the MSVM classifier.

Figure 9 compares the classification performance on the four classes of individual sound events. The best classification accuracy was door open, followed by footstep, door knocking, and speech. On the other hand, the classification results based on the mixed sound events are illustrated in Figure 10. The MSVM model delivered the highest performance of the door-open event with 84% accuracy.

From the above experiments, the proposed method yields an average classification accuracy of 76.67%. The MSVM method can well discriminate and classify the mixed event signals with high classification accuracy (i.e., the mixture of door open with door knocking and door knocking with speech were correctly classified with above 80% accuracy). Due to the MFCC and STZCR features in the individual event, these signals had obvious distinguishable patterns, as shown in the example of STZCR plots in Figure 6. Despite the SDR scores of the separated signals between door open and door knocking being relatively low (as given in Figure 3), the MSVM yielded the highest classification accuracy for the door open with door knocking mixture (DO + DK). This is attributed to the fact that interference remaining in the separated event signals causes the extracted MFCC and STZCR vectors to deviate from their original sound event vectors.

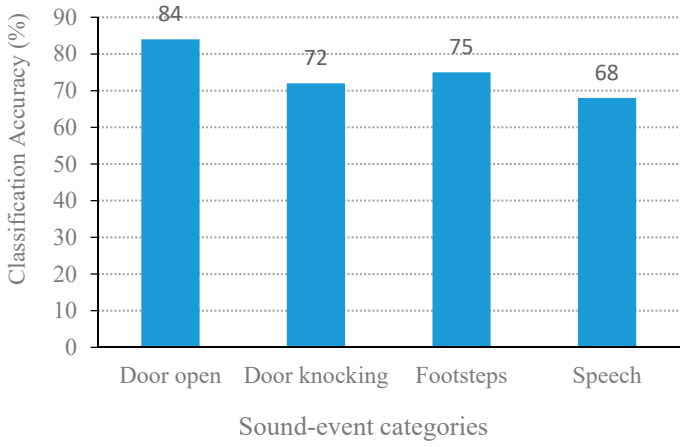


Figure 9. Average percentage of classification accuracy from the perspective of event group of the proposed NSSEC method.

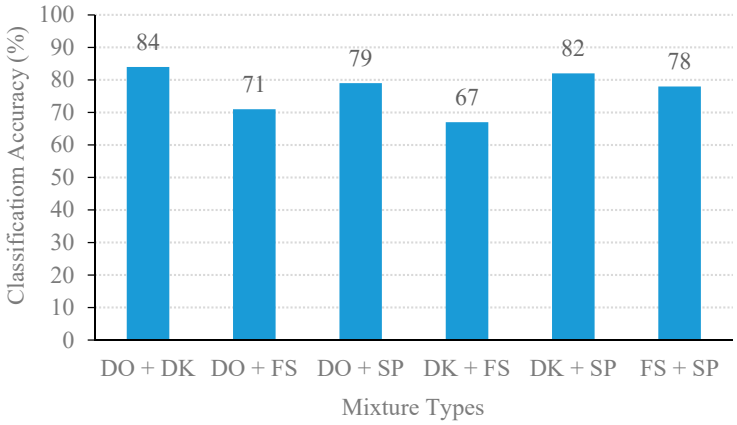


Figure 10. Classification performance of NSSEC model with 1.5 s block size.

4. Conclusions

A novel solution for classification of the noisy mixtures using a single microphone was presented. The complex matrix factorization was proposed and extended by adaptively tuning the sparse regularization. Thus, the desired L_1-optimal sparse decomposition was obtained. In addition, the phase estimates of the CMF could extract the recurrent pattern of the magnitude spectra. The updated equation was derived through an auxiliary function. For classification, the multiclass support vector was used as the mean supervector for encoding the sound-event signatures. The proposed noisy sound separation and event classification method was demonstrated by using four sets of noisy sound-event mixtures, which were door open, door knocking, footsteps, and speech. Based on the experimental results, first, the optimal window length of STFT was found where 1.5 s of the sliding window yielded the best separation performance. The second was two significant features that were ZCR and MFCCs. These parameters were set for examining the proposed method. The proposed method achieved outstanding results in both separation and classification. In future work, the proposed method will be evaluated on a public dataset such as the DCASE 2016, alongside the comparison with other machine learning algorithms.

Author Contributions: Conceptualization, P.P. and W.L.W.; Methodology, P.P. and N.T.; Software, P.P.; Validation, N.T. and W.L.W.; Investigation, P.P. and N.T.; Writing—original draft preparation, P.P. and W.L.W.; Writing—review and editing, N.T., M.A.M.A., O.A., and G.R.; Visualization, M.A.M.A.; Supervision, W.L.W.; Project administration, N.T., M.A.M.A., and O.A.; Funding acquisition, W.L.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the UK Global Challenge Research Fund, the National Natural Science Foundation of China (No. 61971093, No. 61401071, No. 61527803), and supported by the NSAF (Grant No. U1430115) and EPSRC IAA Phase 2 funded project: "3D super-fast and portable eddy current pulsed thermography for railway inspection (EP/K503885/1).

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Single-Channel Sound Event Separation

The prior $P(\mathbf{H}|\boldsymbol{\lambda})$ corresponds to the sparsity cost, for which a natural choice is a generalized Gaussian prior:

$$P(\mathbf{H}|\boldsymbol{\lambda}) = \prod_{k,t} \frac{p\lambda^k(t)}{2\Gamma(1/p)} \exp\left(-\left(\lambda^k(t)\right)^p \left|\mathbf{H}^k(t)\right|^p\right) \quad \text{(A1)}$$

where $\lambda^k(t)$ and p are the shape parameters of the distribution. When $p = 1$, $P(\mathbf{H}|\boldsymbol{\lambda})$ promotes the L_1-norm sparsity. L_1-norm sparsity has been shown to be probabilistically equivalent to the pseudo-norm, L_0, which is the theoretically optimum sparsity [29,30]. However, L_0-norm is non-deterministic polynomial-time (NP) hard and is not useful in large datasets such as audio. Given Equation (3), the posterior density is defined as

$$P(\theta|\mathbf{Y},\boldsymbol{\lambda}) \propto P(\mathbf{Y}|\boldsymbol{\theta})P(\mathbf{H}|\boldsymbol{\lambda}) \tag{A2}$$

The maximum a posteriori probability (MAP) estimation problem leads to minimizing the following optimization problem with respect to θ:

$$f(\theta) = \sum_{f,t}\left|\mathbf{Y}(\omega,t) - \mathbf{X}(\omega,t)\right|^2 + \sum_{k,t}\left[\left(\lambda^k(t)\right)^p\left|\mathbf{H}^k(t)\right|^p - \log\lambda^k(t)\right] \tag{A3}$$

subject to $\sum_f \mathbf{W}^k(\omega) = 1$ $(k = 1,\ldots,K)$.

The CMF parameters has been upgraded by using an efficient auxiliary function for an iterative process. The auxiliary function for $f(\theta)$ can be expressed as the following: for any auxiliary variables with $\sum_k \overline{\mathbf{Y}}^k(\omega,t) = \mathbf{Y}(\omega,t)$, for any $\beta^k(\omega,t) > 0$, $\sum_k \beta^k(\omega,t) = 1$, for any $\mathbf{H}^k(t) \in \mathcal{R}$, $\overline{\mathbf{H}}^k(t) \in \mathcal{R}$, and $p = 1$. The term $f(\theta) \le f^+\left(\theta,\overline{\theta}\right)$ with an auxiliary function was defined as

$$f^+\left(\theta,\overline{\theta}\right) \equiv \sum_{f,k,t}\frac{\left|\overline{\mathbf{Y}}^k(\omega,t) - \mathbf{W}^k(\omega)\mathbf{H}^k(t)\cdot e^{j\boldsymbol{\phi}^k(\omega,t)}\right|^2}{\beta^k(\omega,t)} + \sum_{k,t}\left[\left(\lambda^k(t)\right)^p\left(p\left|\overline{\mathbf{H}}^k(t)\right|^{p-2}\mathbf{H}^k(t)^2 + (2-p)\left|\overline{\mathbf{H}}^k(t)\right|^p\right) - \log\lambda^k(t)\right] \tag{A4}$$

where $\overline{\theta} = \left\{\overline{\mathbf{Y}}^k(\omega,t), \overline{\mathbf{H}}^k(t)\middle|\, 1 \le f \le F,\ 1 \le t \le T,\ 1 \le k \le K\right\}$. The function $f^+\left(\theta,\overline{\theta}\right)$ is minimized w.r.t. $\overline{\theta}$ when

$$\overline{\mathbf{Y}}^k(\omega,t) = \mathbf{W}^k(\omega)\overline{\mathbf{H}}^k(t)\cdot e^{j\boldsymbol{\phi}^k(\omega,t)} + \beta^k(\omega,t)(\mathbf{Y}(\omega,t) - \mathbf{X}(\omega,t)) \tag{A5}$$

$$\overline{\mathbf{H}}^k(t) = \mathbf{H}^k(t) \tag{A6}$$

Appendix B. Estimation of the Spectral Basis and Temporal Code

In Equation (4), the update rule for θ is derived by differentiating $f^+\left(\theta,\overline{\theta}\right)$. partially w.r.t. $\mathbf{W}^k(\omega)$ and $\mathbf{H}^k(t)$, and setting them to zero, which yields the following:

$$\mathbf{W}^k(\omega) = \frac{\sum_t \frac{\mathbf{H}^k(t)}{\beta^k(\omega,t)} Re\left[\overline{\mathbf{Y}}^k(\omega,t)^*\cdot e^{j\boldsymbol{\phi}^k(\omega,t)}\right]}{\sum_t \frac{\mathbf{H}^k(t)^2}{\beta^k(\omega,t)}} \tag{A7}$$

$$\mathbf{H}^k(t) = \frac{\sum_f \frac{\mathbf{W}^k(\omega)}{\beta^k(\omega,t)} Re\left[\overline{\mathbf{Y}}^k(\omega,t)^*\cdot e^{j\boldsymbol{\phi}^k(\omega,t)}\right]}{\sum_f \frac{\mathbf{W}^k(\omega)^2}{\beta^k(\omega,t)} + \left(\lambda^k(t)\right)^p p\left|\overline{\mathbf{H}}^k(t)\right|^{p-2}} \tag{A8}$$

The update rule for the phase, $\boldsymbol{\phi}^k(\omega,t)$, can be derived by reformulating Equation (A1) as follows:

$$\begin{aligned}
&f^+\left(\theta,\overline{\theta}\right)\\
&= \sum_{k,f,t}\frac{\left|\overline{\mathbf{Y}}^k(\omega,t)\right|^2 - 2\mathbf{W}^k(\omega)\mathbf{H}^k(t)\mathrm{Re}\left[\overline{\mathbf{Y}}^k(\omega,t)\cdot e^{-j\boldsymbol{\phi}^k(\omega,t)}\right] + \mathbf{W}^k(\omega)^2\mathbf{H}^k(t)^2}{\beta^k(\omega,t)} + \sum_{k,t}\lambda^k(t)\left(\left|\overline{\mathbf{H}}^k(t)\right|^{-1}\mathbf{H}^k(t)^2 - \overline{\mathbf{H}}^k(t)\right) - \sum_{k,t}\log\lambda^k(t)\\
&= A - 2\sum_{k,f,t}\frac{\mathbf{W}^k(\omega)\mathbf{H}^k(t)\left|\overline{\mathbf{Y}}^k(\omega,t)\right|}{\beta^k(\omega,t)}\left(\frac{Re\left[\overline{\mathbf{Y}}^k(\omega,t)\cdot e^{-j\boldsymbol{\phi}^k(\omega,t)}\right]}{\left|\overline{\mathbf{Y}}^k(\omega,t)\right|}\right)\\
&= A - 2\sum_{k,f,t}\left|\mathbf{B}^k(\omega,t)\right|\frac{Re\left[\left(\overline{\mathbf{Y}}^k(\omega,t)^{(r)} + j\overline{\mathbf{Y}}^k(\omega,t)^{(i)}\right)\left(\cos\boldsymbol{\phi}^k(\omega,t) - j\sin\phi^k(\omega,t)\right)\right]}{\left|\overline{\mathbf{Y}}^k(\omega,t)\right|}\\
&= A - 2\sum_{k,f,t}\left|\mathbf{B}^k(\omega,t)\right|\cos\boldsymbol{\phi}^k(\omega,t)\cos\Omega^k(\omega,t) + \sin\boldsymbol{\phi}^k(\omega,t)\sin\Omega^k(\omega,t)\\
&= A - 2\sum_{k,f,t}\left|\mathbf{B}^k(\omega,t)\right|\cos\left(\boldsymbol{\phi}^k(\omega,t) - \Omega^k(\omega,t)\right)
\end{aligned} \tag{A9}$$

where A denotes the terms that are irrelevant with $\boldsymbol{\phi}^k(\omega,t)$, $\mathbf{B}^k(\omega,t) = \frac{\mathbf{W}^k(\omega)\mathbf{H}^k(t)\overline{\mathbf{Y}}^k(\omega,t)}{\beta^k(\omega,t)}$, $\cos\Omega^k(\omega,t) = \frac{Re\left[\overline{\mathbf{Y}}^k(\omega,t)\right]}{\left|\overline{\mathbf{Y}}^k(\omega,t)\right|}$, and $\sin\Omega^k(\omega,t) = \frac{\mathrm{Im}\left[\overline{\mathbf{Y}}^k(\omega,t)\right]}{\left|\overline{\mathbf{Y}}^k(\omega,t)\right|}$. The auxiliary function, $f^+\left(\theta,\overline{\theta}\right)$ in (A4) is minimized when $\cos\left(\boldsymbol{\phi}^k(\omega,t) - \Omega^k(\omega,t)\right) = \cos\boldsymbol{\phi}^k(\omega,t)\cos^k(\omega,t) + \sin\phi^k(\omega,t)\sin\Omega^k(\omega,t) = 1$, namely, $\cos\boldsymbol{\phi}^k(\omega,t) = \cos\Omega^k(\omega,t)$ and $\sin\boldsymbol{\phi}^k(\omega,t) = \sin\Omega^k(\omega,t)$. The update formula for $e^{j\boldsymbol{\phi}^k(\omega,t)}$ eventually leads to

$$\begin{aligned} e^{j\boldsymbol{\phi}^k(\omega,t)} &= \cos\boldsymbol{\phi}^k(\omega,t) + j\sin\boldsymbol{\phi}^k(\omega,t) \\ &= \frac{Re\left[\overline{\mathbf{Y}}^k(\omega,t)\right] + \mathrm{Im}\left[\overline{\mathbf{Y}}^k(\omega,t)\right]}{\left|\overline{\mathbf{Y}}^k(\omega,t)\right|} \\ &= \frac{\overline{\mathbf{Y}}^k(\omega,t)}{\left|\overline{\mathbf{Y}}^k(\omega,t)\right|} \end{aligned} \tag{A10}$$

The update formula for $\beta^k(\omega,t)$ and $\mathbf{H}^k(t)$ for projection onto the constraint space is set to

$$\beta^k(\omega,t) = \frac{\mathbf{W}^k(\omega)\mathbf{H}^k(t)}{\sum_k \mathbf{W}^k(\omega)\mathbf{H}^k(t)} \tag{A11}$$

$$\mathbf{H}^k(t) \leftarrow \frac{\mathbf{H}^k(t)}{\sum_k \mathbf{H}^k(t)} \tag{A12}$$

Appendix C. Estimation of L_1-Optimal Sparsity Parameter $\lambda^k(t)$

This section aims to facilitate spectral dictionaries with adaptive sparse coding. First, the CMF model is defined as the following terms:

$$\begin{gathered} \overline{\mathbf{W}} = \left[\mathbf{I}\otimes\mathbf{W}^1(\omega)\vdots\mathbf{I}\otimes\mathbf{W}^2(\omega)\vdots\cdots\vdots\mathbf{I}\otimes\mathbf{W}^K(\omega)\right], \\ e^{j\overline{\underline{\mathbf{\Phi}}}(t)} = \left[e^{j\underline{\mathbf{\Phi}}^1(t)}\vdots\cdots\vdots e^{j\underline{\mathbf{\Phi}}^K(t)}\right] \\ \underline{\mathbf{y}} = \mathrm{vec}(\mathbf{Y}) = \begin{bmatrix} \underline{\mathbf{Y}}^1(:) \\ \cdots \\ \underline{\mathbf{Y}}^2(:) \\ \cdots \\ \vdots \\ \cdots \\ \underline{\mathbf{Y}}^K(:) \end{bmatrix}, \underline{\mathbf{h}} = \begin{bmatrix} \mathbf{H}^1(t) \\ \cdots \\ \mathbf{H}^2(t) \\ \cdots \\ \vdots \\ \cdots \\ \mathbf{H}^K(t) \end{bmatrix}, \underline{\boldsymbol{\lambda}} = \begin{bmatrix} \lambda^1(t) \\ \cdots \\ \lambda^2(t) \\ \cdots \\ \vdots \\ \cdots \\ \lambda^K(t) \end{bmatrix}, \underline{\mathbf{\Phi}} = \begin{bmatrix} \boldsymbol{\phi}^1(:,t) \\ \cdots \\ \boldsymbol{\phi}^2(:,t) \\ \cdots \\ \vdots \\ \cdots \\ \boldsymbol{\phi}^K(:,t) \end{bmatrix} \\ \overline{\mathbf{A}} = \begin{bmatrix} \overline{\mathbf{W}}\circ e^{j\overline{\underline{\mathbf{\Phi}}}(t)} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \overline{\mathbf{W}}\circ e^{j\overline{\underline{\mathbf{\Phi}}}(t)} & \mathbf{0} & \vdots \\ \vdots & \mathbf{0} & \overline{\mathbf{W}}\circ e^{j\overline{\underline{\mathbf{\Phi}}}(t)} & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \overline{\mathbf{W}}\circ e^{j\overline{\underline{\mathbf{\Phi}}}(t)_t} \end{bmatrix} \end{gathered} \tag{A13}$$

where "$\otimes$" and "$\circ$" are the Kronecker product and the Hadamard product, respectively. The term vec$(\cdot)$ denotes the column vectorization and the term $\mathbf{I}$ is the identity matrix. The goal is then set to compute the regularization parameter $\lambda^k(t)$ related to each $\mathbf{H}^k(t)$. To achieve the goal, the parameter p in Equation (A3) was set at 1 to acquire a linear expression (in $\lambda^k(t)$). In consideration of the noise variance σ^2, Equation (A3) can concisely be rewritten as:

$$F(\underline{\mathbf{h}},\underline{\boldsymbol{\lambda}}) = \frac{1}{2\sigma^2}\underline{\mathbf{y}} - \overline{\mathbf{A}}\underline{\mathbf{h}}_F^2 + \underline{\boldsymbol{\lambda}}^{\mathbf{T}}\underline{\mathbf{h}} - (\log\underline{\boldsymbol{\lambda}})^{\mathbf{T}}\underline{\mathbf{1}} \tag{A14}$$

where the $\underline{\mathbf{h}}$ and $\underline{\boldsymbol{\lambda}}$ terms indicate vectors of dimension $R \times 1$ (i.e., $R = F \times T \times K$), and the superscript '**T**' is used to denote complex Hermitian transpose (i.e., vector (or matrix) transpose), followed by complex conjugate. The Expectation–Maximization (EM) algorithm is used to determine $\underline{\boldsymbol{\lambda}}$ and $\underline{\mathbf{h}}$ is the hidden variable, where the log-likelihood function can be optimized with respect to $\underline{\boldsymbol{\lambda}}$. The log-likelihood function satisfies the following [12]:

$$\ln p\left(\underline{\mathbf{y}} \left| \underline{\boldsymbol{\lambda}}, \overline{\mathbf{A}}, \sigma^2\right.\right) \geq \int Q(\underline{\mathbf{h}}) \ln\left(\frac{p\left(\underline{\mathbf{y}}, \underline{\mathbf{h}} \left| \underline{\boldsymbol{\lambda}}, \overline{\mathbf{A}}, \sigma^2\right.\right)}{Q(\underline{\mathbf{h}})}\right) d\underline{\mathbf{h}} \tag{A15}$$

by applying the Jensen's inequality for any distribution $Q(\underline{\mathbf{h}})$. The distribution can simply verify the posterior distribution of $\underline{\mathbf{h}}$ that maximizes the right-hand side of Equation (A19) is given by $Q(\underline{\mathbf{h}}) = p\left(\underline{\mathbf{h}} \left| \underline{\mathbf{y}}, \underline{\boldsymbol{\lambda}}, \overline{\mathbf{A}}, \sigma^2\right.\right)$. The posterior distribution in the form of the Gibbs distribution is proposed as follows:

$$Q(\underline{\mathbf{h}}) = \frac{1}{Z_h} \exp[-F(\underline{\mathbf{h}})] \text{ where } Z_h = \int \exp[-F(\underline{\mathbf{h}})] d\underline{\mathbf{h}} \tag{A16}$$

The term $F(\underline{\mathbf{h}})$ in Equation (A16) as the function of the Gibbs distribution is essential for simplifying the adaptive optimization of $\underline{\boldsymbol{\lambda}}$. The maximum-likelihood (ML) estimation of $\underline{\boldsymbol{\lambda}}$ can be decomposed as follows:

$$\begin{aligned} \underline{\boldsymbol{\lambda}}^{ML} &= \arg\max_{\underline{\boldsymbol{\lambda}}} \ln p\left(\underline{\mathbf{y}} \left| \underline{\boldsymbol{\lambda}}, \overline{\mathbf{A}}, \sigma^2\right.\right) \\ &= \arg\max_{\underline{\boldsymbol{\lambda}}} \int Q(\underline{\mathbf{h}}) \left(\ln p\left(\underline{\mathbf{y}} \left| \underline{\mathbf{h}}, \overline{\mathbf{A}}, \sigma^2\right.\right) + \ln p(\underline{\mathbf{h}} | \underline{\boldsymbol{\lambda}})\right) d\underline{\mathbf{h}} \\ &= \arg\max_{\underline{\boldsymbol{\lambda}}} \int Q(\underline{\mathbf{h}}) \ln p(\underline{\mathbf{h}} | \underline{\boldsymbol{\lambda}}) d\underline{\mathbf{h}} \end{aligned} \tag{A17}$$

In the same way,

$$\begin{aligned} {\sigma^2}_{ML} &= \arg\max_{\sigma^2} \ln p\left(\underline{\mathbf{y}} \left| \underline{\boldsymbol{\lambda}}, \overline{\mathbf{A}}, \sigma^2\right.\right) \\ &= \arg\max_{\sigma^2} \int Q(\underline{\mathbf{h}}) \left(\ln p\left(\underline{\mathbf{y}} \left| \underline{\mathbf{h}}, \overline{\mathbf{A}}, \sigma^2\right.\right) + \ln p(\underline{\mathbf{h}} | \underline{\boldsymbol{\lambda}})\right) d\underline{\mathbf{h}} \\ &= \arg\max_{\sigma^2} \int Q(\underline{\mathbf{h}}) \ln p\left(\underline{\mathbf{y}} \left| \underline{\mathbf{h}}, \overline{\mathbf{A}}, \sigma^2\right.\right) d\underline{\mathbf{h}} \end{aligned} \tag{A18}$$

Individual element of $\mathbf{H}$ is required to be exponentially distributed with independent decay parameters that delivers $p(\underline{\mathbf{h}} | \underline{\boldsymbol{\lambda}}) = \prod_g \lambda_g \exp\left(-\lambda_g h_g\right)$, thus Equation (20) obtains

$$\underline{\boldsymbol{\lambda}}^{ML} = \arg\max_{\underline{\boldsymbol{\lambda}}} \int Q(\underline{\mathbf{h}}) \left(\ln \lambda_g - \lambda_g h_g\right) d\underline{\mathbf{h}} \tag{A19}$$

The term $\underline{\mathbf{h}}$ denotes the dependent variable of the distribution $Q(\underline{\mathbf{h}})$ whereas other parameters are assumed to be constant. As such, the $\underline{\boldsymbol{\lambda}}$ optimization in (A19) is derived by differentiating the parameters within the integral with respect to $\underline{\mathbf{h}}$. As a result, the functional optimization of $\underline{\boldsymbol{\lambda}}$ then obtains

$$\lambda_g = \frac{1}{\int h_g Q(\underline{\mathbf{h}}) d\underline{\mathbf{h}}} \tag{A20}$$

where $g = 1, 2, \ldots, R$, λ_g denotes the g^{th} element of $\underline{\boldsymbol{\lambda}}$. The iterative update for σ^2_{ML} is given by

$$\begin{aligned} \sigma^2_{ML} &= \arg\max_{\sigma^2} \int Q(\underline{\mathbf{h}}) \left(\tfrac{-N_0}{2} \ln\left(\pi\sigma^2\right) - \tfrac{1}{2\sigma^2} \|\underline{\mathbf{y}} - \overline{\mathbf{A}}\underline{\mathbf{h}}\|^2\right) d\underline{\mathbf{h}} \\ &= \tfrac{1}{N_0} \int Q(\underline{\mathbf{h}}) \left(\|\underline{\mathbf{y}} - \overline{\mathbf{A}}\underline{\mathbf{h}}\|^2\right) d\underline{\mathbf{h}} \end{aligned} \tag{A21}$$

where $p\left(\underline{\mathbf{y}} \middle| \underline{\mathbf{h}}, \overline{\mathbf{A}}, \sigma^2\right) = \left(\pi\sigma^2\right)^{-N_0/2} \exp\left(-\left(1/2\sigma^2\right)\|\underline{\mathbf{y}} - \overline{\mathbf{A}}\underline{\mathbf{h}}\|^2\right)$ and $N_0 = K \times T$. However, the integral forms in Equations (A20) and (A21) are complex to compute and analyzed analytically. Thus, an approximation to $Q(\underline{\mathbf{h}})$ is exploited. Notice that the solution $\underline{\mathbf{h}}$ naturally splits its elements into distinct subsets $\underline{\mathbf{h}}_M$ and $\underline{\mathbf{h}}_P$ consisting of components $\forall_m \in M$ such that $h_m > 0$ and components $\forall_p \in P$ such that $h_P = 0$. Hence, this can be derived as follows:

$$F(\underline{\mathbf{h}}, \underline{\boldsymbol{\lambda}}) = F\left(\underline{\mathbf{h}}_M, \underline{\boldsymbol{\lambda}}_M\right) + F\left(\underline{\mathbf{h}}_P, \underline{\boldsymbol{\lambda}}_P\right) + G \tag{A22}$$

Defined $F\left(\underline{\mathbf{h}}_M, \underline{\boldsymbol{\lambda}}_M\right) = \frac{1}{2\sigma^2}\|\underline{\mathbf{y}} - \overline{\mathbf{A}}_M\underline{\mathbf{h}}_M\|^2 + \underline{\boldsymbol{\lambda}}_M^{\mathbf{T}}\underline{\mathbf{h}}_M - (\log\underline{\boldsymbol{\lambda}})_M^{\mathbf{T}}\underline{\mathbf{1}}M$, $F\left(\underline{\mathbf{h}}_P, \underline{\boldsymbol{\lambda}}_P\right) = \frac{1}{2\sigma^2}\|\underline{\mathbf{y}} - \overline{\mathbf{A}}_P\underline{\mathbf{h}}_P\|^2 + \underline{\boldsymbol{\lambda}}_P^{\mathbf{T}}\underline{\mathbf{h}}_P - (\log\underline{\boldsymbol{\lambda}})_P^{\mathbf{T}}\underline{\mathbf{1}}P$, and $G = \frac{1}{2\sigma^2}\left[2\left(\overline{\mathbf{A}}_M\underline{\mathbf{h}}_M\right)^{\mathbf{T}}\left(\overline{\mathbf{A}}_P\underline{\mathbf{h}}_P\right) - \|\underline{\mathbf{y}}\|^2\right]$. Here, the term $\|\underline{\mathbf{y}}\|^2$ is a constant and the cross-term $\left(\overline{\mathbf{A}}_M\underline{\mathbf{h}}_M\right)^{\mathbf{T}}\left(\overline{\mathbf{A}}_P\underline{\mathbf{h}}_P\right)$ measures the orthogonality between $\overline{\mathbf{A}}_M\underline{\mathbf{h}}_M$ and $\overline{\mathbf{A}}_P\underline{\mathbf{h}}_P$, where $\overline{\mathbf{A}}_M$ and $\overline{\mathbf{A}}_P$ denote the sub-matrix of $\overline{\mathbf{A}}$ that corresponds to $\underline{\mathbf{h}}_M$ and $\underline{\mathbf{h}}_P$. To obtain a simplified expression in Equation (A22), the $F(\underline{\mathbf{h}})$ function can be approximated as $F(\underline{\mathbf{h}}, \underline{\boldsymbol{\lambda}}) \approx F\left(\underline{\mathbf{h}}_M, \underline{\boldsymbol{\lambda}}_M\right) + F\left(\underline{\mathbf{h}}_P, \underline{\boldsymbol{\lambda}}_P\right)$ and the G can be safely discounted since its value is typically much smaller than $F\left(\underline{\mathbf{h}}_M, \underline{\boldsymbol{\lambda}}_M\right)$ and $F\left(\underline{\mathbf{h}}_P, \underline{\lambda}_P\right)$. Thus, the approximation of $Q(\underline{\mathbf{h}})$ can be expressed as

$$\begin{array}{ll} Q(\underline{\mathbf{h}}, \underline{\boldsymbol{\lambda}}) & = \frac{1}{Z_h}\exp[-F(\underline{\mathbf{h}}, \underline{\boldsymbol{\lambda}})] \\ & \approx \frac{1}{Z_h}\exp\left[-\left(F\left(\underline{\mathbf{h}}_M, \underline{\boldsymbol{\lambda}}_M\right) + F\left(\underline{\mathbf{h}}_P, \underline{\boldsymbol{\lambda}}_P\right)\right)\right] \\ & = \frac{1}{Z_M}\exp\left[-F\left(\underline{\mathbf{h}}_M, \underline{\boldsymbol{\lambda}}_M\right)\right] \\ \frac{1}{Z_P}\exp[- & F\left(\underline{\mathbf{h}}_P, \underline{\boldsymbol{\lambda}}_P\right)] \\ & = Q_M\left(\underline{\mathbf{h}}_M\right)Q_P\left(\underline{\mathbf{h}}_P\right) \end{array} \tag{A23}$$

Defining $Z_M = \int \exp\left[-F\left(\underline{\mathbf{h}}_M, \underline{\boldsymbol{\lambda}}_M\right)\right]d\underline{\mathbf{h}}_M$ and $Z_P = \int \exp\left[-F\left(\underline{\mathbf{h}}_P, \underline{\boldsymbol{\lambda}}_P\right)\right] d\underline{\mathbf{h}}_P$. With the purpose of characterizing $Q_P\left(\underline{\mathbf{h}}_P\right)$, some positive deviation to $\underline{\mathbf{h}}_P$ is needed to be allowed for, whereas the $\underline{\mathbf{h}}_P$ values will reject all negative values due to CMF only accepting zero and positive values. Thus, $\underline{\mathbf{h}}_P$ admits zero and positive values in $Q_P\left(\underline{\mathbf{h}}_P\right)$. The approximation of the distribution $Q_P\left(\underline{\mathbf{h}}_P\right)$ is then utilized in the Taylor expansion as the *maximum a posterior probability* (MAP) estimate. Therefore, with $\underline{\mathbf{h}}^{\mathrm{MAP}}$, one obtains

$$\begin{array}{rl} Q_P\left(\underline{\mathbf{h}}_P \geq 0\right) & \propto \exp\left\{-\left[\left(\frac{\partial F}{\partial \underline{\mathbf{h}}}\right)\Big|_{\underline{\mathbf{h}}^{\mathrm{MAP}}}\right]_P^{\mathbf{T}}\underline{\mathbf{h}}_P - \frac{1}{2}\underline{\mathbf{h}}_P^{\mathbf{T}}\overline{\mathbf{C}}_P\underline{\mathbf{h}}_P\right\} \\ & = \exp\left[-\left(\overline{\mathbf{C}}\underline{\mathbf{h}}^{\mathrm{MAP}} - \frac{1}{\sigma^2}\overline{\mathbf{A}}^{\mathbf{T}}\underline{\mathbf{y}} + \underline{\boldsymbol{\lambda}}\right)_P^{\mathbf{T}}\underline{\mathbf{h}}_P - \frac{1}{2}\underline{\mathbf{h}}_P^{\mathbf{T}}\overline{\mathbf{C}}_P\underline{\mathbf{h}}_P\right] \end{array} \tag{A24}$$

where $\overline{\mathbf{C}}_P = \frac{1}{\sigma^2}\overline{\mathbf{A}}_P^{\mathbf{T}}\overline{\mathbf{A}}_P$ and $\overline{\mathbf{C}} = \frac{1}{\sigma^2}\overline{\mathbf{A}}^{\mathbf{T}}\overline{\mathbf{A}}$. The integration of the term $Q_P\left(\underline{\mathbf{h}}_P\right)$ in Equation (A24) is hard to derive in its closed form expression for analytical evaluation, which subsequently prohibits inference of the sparsity parameters. A fixed form distribution is employed for computing variational approximate $Q_P\left(\underline{\mathbf{h}}_P\right)$. As a result, the closed form expression is obtained. Subsequently, the term $\underline{\mathbf{h}}_P$ only takes on nonnegative values, so a suitable fixed form distribution is to use the factorized exponential distribution given by

$$\hat{Q}_P\left(\underline{\mathbf{h}}_P \geq 0\right) = \prod_{p \in P}\frac{1}{u_p}\exp\left(\frac{-h_p}{u_p}\right) \tag{A25}$$

By minimizing the Kullback–Leibler divergence between Q_P and $\hat{Q}_P$, the variational parameters $\underline{\mathbf{u}} = \left\{u_p\right\}$ where $\forall p \in P$ can be derived as:

$$\begin{array}{ll} \underline{\mathbf{u}} & = \arg = \min_{\underline{\mathbf{u}}}\hat{Q}_P\left(\underline{\mathbf{h}}_P\right)\ln\frac{\hat{Q}_P\left(\underline{\mathbf{h}}_P\right)}{Q_P\left(\underline{\mathbf{h}}_P\right)}d\underline{\mathbf{h}}_P \\ & = \arg = \min_{\underline{\mathbf{u}}}\hat{Q}_P\left(\underline{\mathbf{h}}_P\right)\left[\ln\hat{Q}_P\left(\underline{\mathbf{h}}_P\right) - \ln Q_P\left(\underline{\mathbf{h}}_P\right)\right]d\underline{\mathbf{h}}_P \end{array} \tag{A26}$$

Solving Equation (A26) for u_p leads to the following update [37]:

$$u_p \leftarrow u_p \frac{-\hat{b}_p + \sqrt{\hat{b}_p^2 + 4\frac{(\hat{\mathbf{C}}\underline{\mathbf{u}})_p}{\widetilde{u}_p}}}{2\left(\hat{\mathbf{C}}\underline{\mathbf{u}}\right)_p} \tag{A27}$$

The approximate distribution for components $\underline{\mathbf{h}}_M$ can be obtained by substituting $F\left(\underline{\mathbf{h}}_M, \underline{\boldsymbol{\lambda}}_M\right)$ into $Q_M\left(\underline{\mathbf{h}}_M\right)$ as follows:

$$\begin{array}{rl} Q_M\left(\underline{\mathbf{h}}_M\right) & = \frac{1}{Z_M}\exp\left[-F\left(\underline{\mathbf{h}}_M, \underline{\boldsymbol{\lambda}}_M\right)\right] \\ & \propto \exp\left[-\left(\frac{1}{2}\underline{\mathbf{h}}_M^{\mathrm{T}}\overline{\mathbf{C}}_M\underline{\mathbf{h}}_M - \frac{1}{\sigma^2}\underline{\mathbf{y}}^{\mathrm{T}}\overline{\mathbf{A}}_M\underline{\mathbf{h}}_M + \underline{\boldsymbol{\lambda}}_M\underline{\mathbf{h}}_M\right)\right] \end{array} \tag{A28}$$

In Equation (A28), the function $Q_M\left(\underline{\mathbf{h}}_M\right)$ will be expressed as the unconstrained Gaussion with mean $\underline{\mathbf{h}}_M^{\mathrm{MAP}}$ and covariance $\overline{\mathbf{C}}_M^{-1}$ based on a multivariate Gaussian distribution. The term $\overline{\mathbf{C}}_M$ denotes the sub-matrix of $\overline{\mathbf{C}}$. The sparsity parameter is then obtained by substituting Equations (A24), (A25), and (A28) into Equation (A20) as presented in Equation (A29):

$$\lambda_g = \begin{cases} \frac{1}{\int h_g Q_M(\underline{\mathbf{h}}_M) d\underline{\mathbf{h}}_M} = \frac{1}{h_g^{\mathrm{MAP}}} \text{ if } g \in M \\ \frac{1}{\int h_g \hat{Q}_P(\underline{\mathbf{h}}_P) d\underline{\mathbf{h}}_P} = \frac{1}{u_g} \text{ if } g \in P \end{cases} \tag{A29}$$

and its covariance X is given by

$$X_{ab} = \begin{cases} \left(\overline{\mathbf{C}}_P^{-1}\right)_{ab}, \text{ if } a, b \in M \\ u_p^2\delta_{ab}, \text{ Otherwise.} \end{cases} \tag{A30}$$

Similarly, the inference for σ^2 can be computed from Equation (24) as

$$\sigma^2 = \frac{1}{N_0}\int Q(\underline{\mathbf{h}})\left(\|\underline{\mathbf{y}} - \overline{\mathbf{A}}\underline{\mathbf{h}}\|^2\right) d\underline{\mathbf{h}} \tag{A31}$$

where

$$\hat{h}_g = \begin{cases} h_g^{\mathrm{MAP}} \text{ if } g \in M \\ u_g \text{ if } g \in P \end{cases}$$

The core procedure of the proposed CMF method is based on L_1-optimal sparsity parameters. The estimated sources are discovered by multiplying the respective rows of the $\boldsymbol{W}^k(\omega)$ components with the corresponding columns of the $\boldsymbol{H}^k(t)$ weights and time-varying phrase spectrum $e^{j\Phi^k(\omega,t)}$. The separated sources $\hat{s}_j(t)$ are obtained by converting the time-frequency represented sources into time domain.

References

1. Wang, Q.; Woo, W.L.; Dlay, S. Informed single-channel speech separation using hmm–gmm user-generated exemplar source. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 2087–2100. [CrossRef]
2. Gao, B.; Bai, L.; Woo, W.L.; Tian, G.; Cheng, Y. Automatic defect identification of eddy current pulsed thermography using single channel blind source separation. *IEEE Trans. Instrum. Meas.* **2013**, *63*, 913–922. [CrossRef]
3. Yin, A.; Gao, B.; Tian, G.; Woo, W.L.; Li, K. Physical interpretation and separation of eddy current pulsed thermography. *J. Appl. Phys.* **2013**, *113*, 64101. [CrossRef]

4. Cheng, L.; Gao, B.; Tian, G.; Woo, W.L.; Berthiau, G. Impact damage detection and identification using eddy current pulsed thermography through integration of PCA and ICA. *IEEE Sens. J.* **2014**, *14*, 1655–1663. [CrossRef]
5. Cholnam, O.; Chongil, G.; Chol, R.K.; Gwak, C.; Rim, K.C. Blind signal separation method and relationship between source separation and source localisation in the TF plane. *IET Signal Process.* **2018**, *12*, 1115–1122. [CrossRef]
6. Tengtrairat, N.; Woo, W.L.; Dlay, S.S.; Gao, B. Online noisy single-channel blind separation by spectrum amplitude estimator and masking. *IEEE Trans. Signal Process* **2016**, *64*, 1881–1895. [CrossRef]
7. Tengtrairat, N.; Gao, B.; Woo, W.L.; Dlay, S.S. Single-Channel Blind Separation Using Pseudo-Stereo Mixture and Complex 2-D Histogram. *IEEE Trans. Neural Netw. Learn. Syst.* **2013**, *24*, 1722–1735. [CrossRef]
8. Koundinya, S.; Karmakar, A. Homotopy optimisation based NMF for audio source separation. *IET Signal Process.* **2018**, *12*, 1099–1106. [CrossRef]
9. Kim, M.; Smaragdis, P. Single channel source separation using smooth Nonnegative Matrix Factorization with Markov Random Fields. In Proceedings of the 2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP), Southapmton, UK, 22–25 September 2013; pp. 1–6.
10. Yoshii, K.; Itoyama, K.; Goto, M. Student's T nonnegative matrix factorization and positive semidefinite tensor factorization for single-channel audio source separation. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 51–55.
11. Al-Tmeme, A.; Woo, W.L.; Dlay, S.; Gao, B. Underdetermined convolutive source separation using gem-mu with variational approximated optimum model order NMF2D. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* **2016**, *25*, 35–49. [CrossRef]
12. Woo, W.L.; Gao, B.; Bouridane, A.; Ling, B.W.-K.; Chin, C.S. Unsupervised learning for monaural source separation using maximization–minimization algorithm with time–frequency deconvolution. *Sensors* **2018**, *18*, 1371. [CrossRef]
13. Gao, B.; Woo, W.L.; Dlay, S.S. Unsupervised single channel separation of non-stationary signals using Gammatone filterbank and Itakura-Saito nonnegative matrix two-dimensional factorizations. *IEEE Trans. Circuits Syst. I* **2013**, *60*, 662–675. [CrossRef]
14. Févotte, C.; Bertin, N.; Durrieu, J.-L. Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis. *Neural Comput.* **2009**, *21*, 793–830. [CrossRef] [PubMed]
15. Pu, X.; Yi, Z.; Zheng, Z.; Zhou, W.; Ye, M. Face recognition using fisher non-negative matrix factorization with sparseness constraints. *Comput. Vis.* **2005**, *3497*, 112–117. [CrossRef]
16. Magron, P.; Virtanen, T. Towards complex nonnegative matrix factorization with the beta-divergence. In Proceedings of the 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC), Tokyo, Japan, 17–20 September 2018; pp. 156–160.
17. King, B. New Methods of Complex Matrix Factorization for Single-Channel Source Separation and Analysis. Ph.D. Thesis, University of Washington, Seattle, WA, USA, 2012.
18. Parathai, P.; Tengtrairat, N.; Woo, W.L.; Gao, B. Single-channel signal separation using spectral basis correlation with sparse nonnegative tensor factorization. *Circuits Syst. Signal Process.* **2019**, *38*, 5786–5816. [CrossRef]
19. Woo, W.L.; Dlay, S.; Al-Tmeme, A.; Gao, B. Reverberant signal separation using optimized complex sparse nonnegative tensor deconvolution on spectral covariance matrix. *Digit. Signal Process.* **2018**, *83*, 9–23. [CrossRef]
20. Tengtrairat, N.; Parathai, P.; Woo, W.L. Blind 2D signal direction for limited-sensor space using maximum likelihood estimation. *Asia-Pac. J. Sci. Technol.* **2017**, *22*, 42–49.
21. Gao, B.; Woo, W.L.; Tian, G.Y.; Zhang, H. Unsupervised diagnostic and monitoring of defects using waveguide imaging with adaptive sparse representation. *IEEE Trans. Ind. Inform.* **2016**, *12*, 405–416. [CrossRef]
22. Gao, B.; Woo, W.L.; He, Y.; Tian, G.Y. Unsupervised sparse pattern diagnostic of defects with inductive thermography imaging system. *IEEE Trans. Ind. Inform.* **2016**, *12*, 371–383. [CrossRef]
23. Tengtrairat, N.; Woo, W.L. Single-channel separation using underdetermined blind autoregressive model and lest absolute deviation. *Neurocomputing* **2015**, *147*, 412–425. [CrossRef]
24. Gao, B.; Woo, W.; Ling, B.W.-K. Machine learning source separation using maximum a posteriori nonnegative matrix factorization. *IEEE Trans. Cybern.* **2013**, *44*, 1169–1179. [CrossRef]

25. Tengtrairat, N.; Woo, W. Extension of DUET to single-channel mixing model and separability analysis. *Signal Process.* **2014**, *96*, 261–265. [CrossRef]
26. Zhou, Q.; Feng, Z.; Benetos, E. Adaptive noise reduction for sound event detection using subband-weighted NMF. *Sensors* **2019**, *19*, 3206. [CrossRef] [PubMed]
27. Yan, L.; Zhang, Y.; He, Y.; Gao, S.; Zhu, D.; Ran, B.; Wu, Q. Hazardous traffic event detection using markov blanket and sequential minimal optimization (MB-SMO). *Sensors* **2016**, *16*, 1084. [CrossRef] [PubMed]
28. Chen, Y.-L.; Chiang, H.-H.; Chiang, C.-Y.; Liu, J.; Yuan, S.-M.; Wang, J.-H. A vision-based driver nighttime assistance and surveillance system based on intelligent image sensing techniques and a heterogamous dual-core embedded system architecture. *Sensors* **2012**, *12*, 2373–2399. [CrossRef] [PubMed]
29. McLoughlin, I.V.; Zhang, H.; Xie, Z.; Song, Y.; Xiao, W. Robust sound event classification using deep neural networks. *IEEE/ACM Trans. Audio. Speech. Lang. Process.* **2015**, *23*, 540–552. [CrossRef]
30. Noh, K.; Chang, J.-H. Joint optimization of deep neural network-based dereverberation and beamforming for sound event detection in multi-channel environments. *Sensors* **2020**, *20*, 1883. [CrossRef]
31. Hsu, C.W.; Lin, C.J. A comparison of methods for multi-class support vector machines. *IEEE Trans. Neural Netw.* **2002**, *13*, 415–425.
32. Martin-Morato, I.; Cobos, M.; Ferri, F.J. A case study on feature sensitivity for audio event classification using support vector machines. In Proceedings of the 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP), Salerno, Italy, 13–16 September 2016; pp. 1–6.
33. Candès, E.J.; Romberg, J.K.; Tao, T. Stable signal recovery from incomplete and inaccurate measurements. *Commun. Pure Appl. Math.* **2006**, *59*, 1207–1223. [CrossRef]
34. Selesnick, I. Resonance-based signal decomposition: A new sparsity-enabled signal analysis method. *Signal Process.* **2011**, *91*, 2793–2809. [CrossRef]
35. Al-Tmeme, A.; Woo, W.L.; Dlay, S.; Gao, B. Single channel informed signal separation using artificial-stereophonic mixtures and exemplar-guided matrix factor deconvolution. *Int. J. Adapt. Control. Signal Process.* **2018**, *32*, 1259–1281. [CrossRef]
36. Gao, B.; Woo, W.L.; Dlay, S.S. Single channel blind source separation using EMD-subband variable regularized sparse features. *IEEE Trans. Audio. Speech Lang. Process.* **2011**, *19*, 961–976. [CrossRef]
37. Bertsekas, D.P. *Nonlinear Programming*, 2nd ed.; Athena Scientific: Belmont, MA, USA, 1999.
38. Kameoka, H.; Ono, N.; Kashino, K.; Sagayama, S. Complex NMF: A new sparse representation for acoustic signals. In Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan, 19–24 April 2009; pp. 3437–3440. [CrossRef]
39. Parathai, P.; Woo, W.L.; Dlay, S.; Gao, B. Single-channel blind separation using L1-sparse complex non-negative matrix factorization for acoustic signals. *J. Acoust. Soc. Am.* **2015**, *137*, 124–129. [CrossRef] [PubMed]
40. Zdunek, R.; Cichocki, A. Nonnegative matrix factorization with constrained second-order optimization. *Signal Process.* **2007**, *87*, 1904–1916. [CrossRef]
41. Yu, K.; Woo, W.L.; Dlay, S. Variational regularized two-dimensional nonnegative matrix factorization. *IEEE Trans. Neural Netw. Learn. Syst.* **2012**, *23*, 703–716.
42. Gao, B.; Woo, W.L.; Dlay, S. Adaptive sparsity non-negative matrix factorization for single-channel source separation. *IEEE J. Sel. Top. Signal Process.* **2011**, *5*, 989–1001. [CrossRef]