A novel Non-negative Matrix Factorization (NMF) based subband decomposition in frequency-spatial domain for acoustic source localization using a microphone array. The proposed method decomposes source and noise subbands and emphasizes source dominant frequency bins for more accurate source representation. By employing NMF, we extract Delay Basis Vectors (DBVs) and their subband information in frequency-spatial domain for each frame. The proposed algorithm is evaluated in both simulated noise and real noise with a speech corpus database. Experimental results clearly indicate that the algorithm performs more accurately than other conventional algorithms under both reverberant and noisy acoustic environments.

Introduction: Acoustic source localization has been an active research area with applications in a variety of fields and it has become an important topic in acoustic based applications. Time Delay Estimation (TDE) between two or more microphone signals can be used as a mean for source localization. Generalized Cross-Correlation (GCC) is the most commonly used TDE approach.

In this paper, we propose a decomposition of signal and noise subbands based on Non-negative Matrix Factorization (NMF) and GCC. Using the decomposed signal subband information, the source dominant frequency bins can be emphasized by spectral weighting. A TDE algorithm based on the proposed subband decomposition approach outperforms conventional GCC algorithms and other TDE algorithms such as Adaptive Eigenvalue Decomposition (AED) [1] and other spectral weighting methods such as Cross-Power Spectrum (CPS) [2] and local-Peak-Weighted (LPW) [3] in reverberant and noisy environments. The proposed method exhibits conceptual similarity to the Multiple Signal Classification (MUSIC) algorithm [4]. It decomposes the cross-correlation matrix of the multichannel signals into signal and noise subspaces using eigenvalue decomposition. It was developed originally as a direction-of-arrival (DOA) estimation technique for narrowband signals, and there are many variants. Although there are subspace techniques, such as the MUSIC method, that are applicable to wideband signals, they cannot be used for coherent source localization such as acoustic environment with reverberations.

Proposed subband decomposition: Consider that the Mth channel microphone input signal is $x_m(t)$ and its Short Time Fourier Transform (STFT) is $X_m(t)$, then the GCC with PHAse Transform (PHAT) of the $l^\text{th}$ and the $q^\text{th}$ microphone signal is

$$R_{lq}(\theta) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \Psi_{ql}(\omega) X_l(\omega)X_q(\omega)e^{j\omega \cdot \theta} d\omega$$

(1)

where $\Psi_{ql}$ denotes a PHAT weight function as $\Psi_{ql}(\omega) = 1/|X_l(\omega)X_q(\omega)|$.

Note that $\theta$ is azimuth when the Time Delay Of Arrival of the $l^\text{th}$ and the $q^\text{th}$ microphones is $\tau \approx \sin^{-1}(\lambda f / d)$ where $d$ is the distance between the $l^\text{th}$ and the $q^\text{th}$ microphones, and $\gamma$ is the speed of sound.

Since STFT is designed for a discrete signal, frequency $\omega_0$ should be a discrete value, i.e., $\omega_0 = 2\pi k / N$, where $N$ is the length of the frame and $k$ denotes the frequency bin index. Therefore, for calculating GCC-PHAT corresponding to each frequency bin, (1) can be rewritten as

$$R_{lq}(\theta, o_0) = \frac{1}{K} \sum_{c=1}^{K} \Psi_{ql}(\omega_0) X_{l,c}(\omega_0) X_{q,c}(\omega_0) e^{j\omega_0 o_0}$$

(2)

Using (2), we show some examples of source localization in a single frame. For clean signals, we can see clear large amplitudes to source directions in all frequency bins as in Fig. 1 (a). However, when there is noise, the source signal is corrupted as shown in Fig. 1 (b). For more accurate and robust source localization, we utilize the NMF theory to decompose the source and noise subbands and accentuate the source dominant frequency bins.

Source DBV selection: For applying the subband information in TDE as (7), we must determine which basis vector $w_c, c=1,2,\ldots,C$, is the source DBV. If we use a source DBV index $c$, $R_{lq}^{\text{SNR}}$ amplitude will show the highest peak at the source’s existing azimuth, i.e. time delay, because the frequency bins that have the same time delay become emphasized. Hence, we can use a noise DBV index $c$, $R_{lq}^{\text{SNR}}$ will not show clearly high peak because the noise signal has no coherent time delay.
For example, Fig. 2 shows an effect of the proposed algorithm with the same acoustic frame applied to that in Fig. 1. We first obtained the TDE result with a conventional algorithm under a clean environment as in Fig. 2 (a). Next, we obtained the TDE results of both conventional and proposed algorithms under SNR -5dB high noise environment as shown in Fig. 2 (b)-(e). It can be easily seen that the result using $h_1$ is most similar to the clean environment result as in Fig. 2(a), while the others show significantly erroneous peaks. Therefore, determining the index of the source DBV can be done by finding the highest peak after applying all C subband information as follows. The azimuth candidate $\theta_c$ corresponding to each DBV $c$ is

$$\theta_c = \arg \max_{\theta, \tau} R_{\theta, \tau}^{(c)} (\theta).$$

(8)

Fig. 2 TDE Examples
a Conventional GCC-PHAT result when noise is absent
b Conventional GCC-PHAT result when SNR= -5dB
c Proposed result using (7) and $h_2$ when SNR= -5dB
d Proposed result using (7) and $h_2$ when SNR= -5dB
e Proposed result using (7) and $h_1$ when SNR= -5dB

Fig. 3 TDE Performance evaluation using RMSE
a With varying number $C$ of DBV on 20dB SNR with no reverb
b For varying SNR from -5dB to 20dB (RT60=100ms)
c For varying SNR from -5dB to 20dB (RT60=500ms)
d For real noise when SNR= -5dB

Then, we can find the azimuth of the source as

$$\theta_{source} = \max_{c} \{ R_{\theta, \tau}^{(c)} (\theta) \}$$

(9)

The source DBV index $\beta$ is found by

$$\beta = \arg \max_{c} \{ R_{\theta, \tau}^{(c)} (\theta) \}.$$