Agglomerative Hierarchical Clustering of Basis Vector for Monaural Sound Source Separation Based on NMF

Kentaro Murai*, Taiho Takeuchi†, and Yosuke Tatekura*
* Graduate School of Integrated Science and Technology, Shizuoka University, Shizuoka, Japan
E-mail: k-murai@spalab.eng.shizuoka.ac.jp Tel/Fax: +81-53-478-1139
† Graduate School of Science and Technology, Shizuoka University, Shizuoka, Japan

***Abstract*—This paper proposes a method of monaural sound source separation by clustering based on the similarity of basis vectors decomposed by Non-negative Matrix Factorization (NMF). In the proposed method, the basis vectors are clustered on the assumption that the similarity between the basis vectors constituting the target sound source is higher than the similarity with the basis vectors of the other sound sources. Hierarchical clustering, which forms clusters in descending order of feature similarity, is introduced. Since it is unnecessary to explicitly determine the number of clusters in hierarchical clustering, hierarchical clustering can be classified into an optional number of clusters according to the threshold. Therefore, the proposed method can separate to an optional number of sound sources. From the numerical evaluation result, it was found that the Signal to Distortion Ratio (SDR), which is an evaluation index of sound source separation, can be improved by approximately 6 to 10 dB. Undesirable cases in which most of the basis vectors are classified into the same cluster are also discussed. In addition, sound source separation with mixed three mixed sound sources was also evaluated, and it was confirmed that SDR can be improved by about 10 dB.**

I. INTRODUCTION

In recent years, acoustic event detection technology [1][2], which is used to detect the sounds of everyday life, has attracted attention. This technology is expected to be applied to a life log system, e.g., a security system that detects footsteps and analyzes the noise produced. In everyday life, it is rare for only the desired sound source to be observed; in most cases, sounds from a mixture of multiple sources are observed. If a mixture of sounds from multiple sources is used as input in an algorithm that detects sound from a single source, each sound source cannot be accurately recognized, and the detection accuracy is degraded. To resolve this, sound source separation is used a preprocessing technique to reduce the overlap of sound sources [3].

Blind Source Separation is a technique that extracts sound from a single source from sounds from a mixed source without explicitly using information about recording environments, mixed systems, sound source positions, etc. When the number of sound sources is less than the number of observation channels, the separation method based on statistical independence [4][5] is widely applied. In the method based on statistical independence, an inverse mixing system that separates the mixed sound source into each sound source is estimated under the assumption that each sound source is statistically independent. This method needs to know the number of sound sources beforehand. Therefore, it is necessary to appropriately change the algorithm and the number of microphones; however, the application of the algorithm is limited. When the number of observation channels is smaller than the number of sound sources, the separation method based on Non-negative Matrices Factorization (NMF) [6][7] is widely applied. In an NMF based method, the spectrogram obtained by the Short-Time Fourier Transform (STFT) [8] is regarded as an observation matrix and decomposes into non-negative basis vectors and activations. The sound source separation is performed by classifying the basis vectors of the mixed sound source and reconstructing the spectrogram to each sound source. Multichannel NMF (MNMF) [9] archives highly accurate separation by using the spatial differences between microphones for a mixed sound source observed by multiple microphones. However, MNMF requires a large calculation cost. For the monaural channel, a method of supervised and semi-supervised musical instrument separation [10][11], supervised speech separation [12], and supervised drum separation [13] have been proposed. However, since the sounds of everyday life are much more diverse than the sounds of musical instruments or speech, it is difficult to obtain the training data for all sound sources. Therefore, an unsupervised method that realizes monaural sound source separation without using prior information is required.

This paper proposes an unsupervised method of monaural sound source separation by clustering, based on the similarity of the basis vectors decomposed by NMF. The proposed method introduces Agglomerative Hierarchical Clustering [14], which forms clusters of the basis vectors in a descending order of similarity. The system overview of the proposed method is shown in Fig. 1. Clustering is performed on the feature vector extracted from the basis vector, and classified into clusters for each sound source. Sound source separation experiments were conducted to evaluate the proposed method.

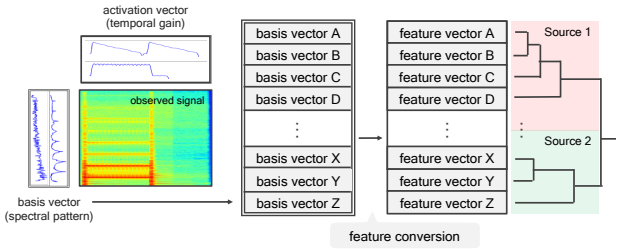

Fig. 1. Overview of the proposed method.

II. OUTLINE OF SOUND SOURCE SEPARATION BY NON-NEGATIVE MATRIX FACTORIZATION

NMF is an algorithm that decomposes a non-negative observed matrix $\boldsymbol{X} \in \mathbb{R}_{\geq 0}^{I \times J}$ into two non-negative matrices:

$$\boldsymbol{X} \approx \boldsymbol{T}\boldsymbol{V}, \tag{1}$$

where $\boldsymbol{T} \in \mathbb{R}_{\geq 0}^{I \times K}$ is the basis matrix, which represents the spectral pattern, and $\boldsymbol{V} \in \mathbb{R}_{\geq 0}^{K \times J}$ is the activation matrix, which represents the temporal gain. Also, I is the number of frequency bins, J is the number of time frames, and K is the number of basis vectors.

The decomposition is performed by minimizing the distance between $\boldsymbol{X}$ and $\boldsymbol{T}\boldsymbol{V}$:

$$\min_{\boldsymbol{T},\boldsymbol{V}} \mathcal{D}(\boldsymbol{X} \| \boldsymbol{T}\boldsymbol{V}), \tag{2}$$

where $\mathcal{D}(\cdot\|\cdot)$ represents the divergence between the observation and a model. Kullback–Leibler (KL) divergence is one of the NMF cost functions and is defined as follows:

$$\mathcal{D}_{\mathrm{KL}} = \sum_i \sum_j \left(x_{ij} \log \frac{x_{ij}}{\sum_k t_{ik} v_{kj}} - x_{ij} + \sum_k t_{ik} v_{kj} \right), \tag{3}$$

where x_{ij}, t_{ik} and v_{kj} are the non-negative elements of matrices $\boldsymbol{X}$, $\boldsymbol{T}$, and $\boldsymbol{V}$, respectively; $i = 0, \cdots, I$, is the frequency index; $j = 0, \cdots, J$ is the time index, and $k = 1, \cdots, K$ is the basis index. Minimization of the cost function based on KL divergence is performed using the multiplicative updating rules, which can be given by

$$t_{ik} \leftarrow t_{ik} \frac{\sum_j x_{ij} (\sum_{k'} t_{ik'} v_{k'j})^{-1} v_{ik}}{\sum_j v_{kj}}, \tag{4}$$

$$v_{kj} \leftarrow v_{kj} \frac{\sum_i t_{ik} x_{ij} (\sum_{k'} t_{ik'} v_{k'j})^{-1}}{\sum_i t_{ik}}. \tag{5}$$

Decomposition is performed by repeatedly applying these expressions to the matrices $\boldsymbol{T}$ and $\boldsymbol{V}$, which have been given random initial values. In general, NMF algorithm is not essentially a sound source separation method, but an approximation expression obtained by minimizing the divergence between $\boldsymbol{X}$ and $\boldsymbol{T}\boldsymbol{V}$. The mixed sound source is modeled as a sum of spectral parts. There is no guarantee that spectral parts represent one sound source; it is possible for spectral parts to represent the components of one or more sound sources. However, in many cases, the basis vector represents a distinctive component of a single sound source. If the decomposed basis vectors are properly clustered, the overlap of the sound sources is reduced.

III. NMF BASIS VECTOR CLUSTERING

A hierarchical clustering method based on the similarity of the spectral pattern which classified the basis vectors of each sound source is introduced.

In the proposed method, the basis vectors are clustered under the assumption that the similarity of the basis vectors constituting the target sound is higher than that of the basis vectors constituting other sound sources. In the hierarchical clustering method, the distances between clusters are calculated, and a new cluster is formed by coupling the clusters closest in distance.

Here, the basis vector clustering procedure of the proposed method is described in detail. As a preparatory step, a feature vector extracted from each basis vector is defined as an initial cluster. The number of initial clusters corresponds to the number of basis vectors K. The number of samples included in the cluster is defined as the cluster size n. Since nothing is coupled in the initial cluster, the cluster contains only one sample. Thus, the size of the initial cluster n is equal to 1. After the initial clusters are defined, clustering is performed as follows:

1) The distance between the initial clusters are calculated based on the cosine similarity as follows:

$$d_{\cos}(\boldsymbol{a}, \boldsymbol{b}) = 1 - \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{|\boldsymbol{a}||\boldsymbol{b}|}, \tag{6}$$

where $\boldsymbol{a}$ and $\boldsymbol{b}$ are feature vectors for evaluating similarity.

2) New clusters are generated by coupling clusters with minimal intercluster distance. The size of the new cluster is given by the sum of the sizes of each cluster.

3) Intercluster distance is calculated again. In the case of $n \geq 2$, the intercluster distance $d_{\mathcal{C}_1\mathcal{C}_2}$ is calculated based on the group average method as follows:

$$d_{\mathcal{C}_1\mathcal{C}_2} = \frac{1}{n_1 n_2} \sum_{x_1 \in \mathcal{C}_1} \sum_{x_2 \in \mathcal{C}_2} d(x_1, x_2), \tag{7}$$

where, n_1 and n_2 are the sizes of the clusters $\mathcal{C}_1$ and $\mathcal{C}_2$, respectively, and $d(x_1, x_2)$ is the distance between the samples x_1 and x_2.

Cluster coupling is repeated until a single cluster that includes all the initial clusters is generated. The final resultant cluster is divided at an appropriate distance to obtain basis vectors separated for each sound source.

An example of simple clustering is shown in Fig. 2. Consider the case of clustering four samples of A–D in the feature space shown in Fig.2 (a). A–D are defined as initial clusters. The size of each initial cluster is set to $n = 1$. In the first coupling, A and B, which have the smallest distance

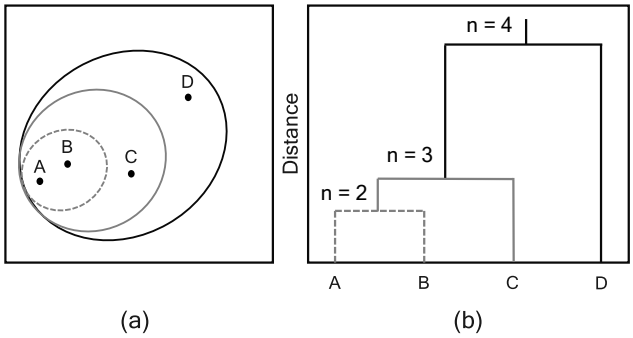

Fig. 2. An example of hierarchical clustering and dendrogram representing the clustering process.

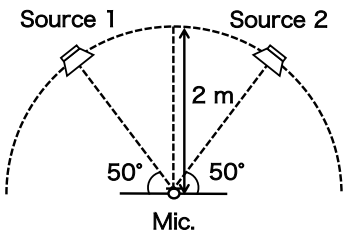

Fig. 3. Recording the E2A impulse response condition.

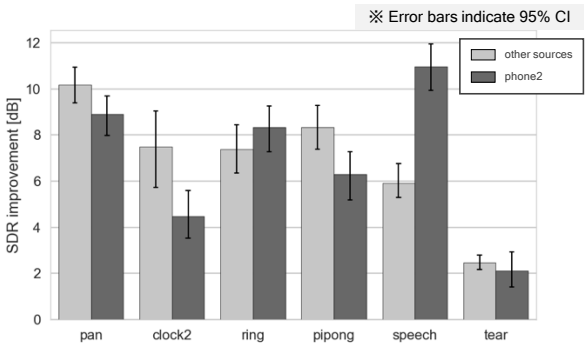

Fig. 4. SDR improvement by clustering.

to each other in the feature space, are coupled to generate a new cluster AB. Next, the distance between the clusters is recalculated. Since the size of the new cluster is $n = 2$, the distance between AB and C is calculated as the average of the distances A–C and B–C. Similarly, the distances between AB and D is calculated as the average of the distances A–D and B–D. In this example, AB and C, which are closet to each other than AB and D, are coupled to generate a new cluster with a size of $n = 3$. Finally, the distance between ABC and D is generated by coupling ABC and D. The size of the new cluster is $n = 4$. This cluster contains all the initial clusters, so the clustering process is terminated. As shown in Fig. 2 (b), in the hierarchical clustering method, it is possible to represent the process in which clusters are coupled according to the distance by using the dendrogram. The height of the bars representing cluster coupling corresponds to the intercluster distance. In the proposed method, clusters coupled at a distance less than the threshold are regarded as clusters one sound source. Therefore, the proposed method can separate the optional number of sound sources. In this study, the threshold was set to $0.7 \times d_{\max}$ using an implementation of Scipy, where, $d_{\max}$ is the maximum value of the distance.

IV. EXPERIMENTS

A. Experimental setup

The following were used as sound source: phone ringing (phone2), frying pan tapping (pan), alarm clock ringing (clock2), bell ringing (ring), correct answer ping (pipong), sound of paper tearing (tear) from the RWCP real environment voice / sound database [15], and the sound of female speech (speech) from the JSUT dataset [16]. The mixed signal was generated by convolving E2A impulse response contained in the dataset. The sampling frequency of the impulse responses and sound sources was 48000 Hz, and the reverberation time in the room was 300 ms. The recording environment of the E2A impulse response is shown in Fig. 3.

The observed signals were transformed into a spectrogram by STFT with a window length of 85 ms and shift length of 43 ms. In order to perform auditory and objective evaluation, it was necessary to prioritize the representation accuracy of the sound source, and the number of basis vectors K was set to 60. The number of iterations was set to 1000 times. The obtained basis vectors were converted into 16 dimensions of Mel-Frequency Cepstrum Coefficients (MFCC).

B. Evaluation

The separation results were estimated by improving the Signal to Distortion Ratio (SDR) [17], which evaluates the accuracy of the sound source separation, e.g., distortion caused by sound source separation, interference of non-target sound, and noise.

The SDR improvement was calculated to the following procedure. The separated signal $\hat{s}(t)$ was decomposed into four elements:

$$\hat{s}(t) = s_{\text{true}}(t) + e_{\text{interf}}(t) + e_{\text{noise}}(t) + e_{\text{artif}}(t), \quad (8)$$

where $s_{\text{true}}(t)$ is the target source signal; $e_{\text{interf}}(t)$ is the interference of non-target sound; $e_{\text{noise}}(t)$ is noise, and $E_{\text{artif}}(t)$ represents artifacts. SDR is defined by the following energy ratio:

$$\text{SDR} = 10 \log_{10} \frac{\sum_t s_{\text{true}}(t)^2}{\sum_t (e_{\text{interf}}(t) + e_{\text{noise}}(t) + e_{\text{artif}}(t))^2}. \quad (9)$$

The improved SDR, SDR_{imp}, was calculated as follows:

$$\mathrm{SDR}_{\mathrm{imp}} = \mathrm{SDR}_{\mathrm{after}} - \mathrm{SDR}_{\mathrm{before}}, \tag{10}$$

where $\mathrm{SDR}_{\mathrm{before}}$ is the SDR before clustering, and $\mathrm{SDR}_{\mathrm{after}}$ is the SDR after clustering.

The correspondence to each target sound source was manually confirmed in the evaluation as the proposed clustering method is an unsupervised method, and the resulting cluster are exclusively groups of basis vectors with high similarity.

C. Results

The results of sound source separation for each mixture pattern are shown in Fig. 4. Here the separation result in the case of mixed sound sources is shown. Error bars represent 95 % of the average confidence intervals. A significant improvement in SDR can be found in the pan mixture case, clock2 mixture case, ring mixture case, pipong mixture case, and speech mixture case. SDR improvements could not be obtained in the tear mixture case, which means the separation was insufficient. The dendrograms of the pan and tear mixture cases are shown in Fig. 5 and Fig. 6, respectively. In Fig. 5, it can be confirmed that the basis vectors with a high similarity form clusters hierarchically according to the proposed method. It can also be confirmed that clustering was properly performed based on similarity for all the sound sources, except in the tear mixture case.

In Fig. 6, most of the basis vectors in the tear mixture case are classified as the same cluster: "phone2." An example of the basis vector which represents a part of the tear is shown in Fig. 7 (a). This basis vector has a flat power in the entire frequency band. An example of the basis vector classified as "phone2" is shown in Fig. 7 (b). As shown in this figure, the basis vector also has a flat power in the entire frequency band, but two peaks can be observed around 2900 Hz and 5600 Hz. These peaks represent a part of phone2, and the basis vectors are considered to contain both the tear and phone2 components. The same phenomenon can be confirmed in the speech mixture case. The spectrogram of phone2 in the speech mixture case is shown in Fig. 8. SDR improvement is good, and the reduced overlap of the sound sources can be confirmed. However, a clear overlap of the sound source can be observed at approximately 1.1 sec. This is the sound of /s/ in the speech, which is like a tear with power in the wide band, so it cannot properly represent overlap with phone2. This is because the NMF algorithm does not explicitly perform the sound source separation. In the case of a sound source with power in all frequency bands, such as a tear, it is difficult to represent the overlap of sound sources, which leads to undesirable results.

The results of the situation with three mixed sound sources is shown in Fig. 9. An SDR improvement of about 10 dB is also obtained in this case. The dendrogram of the situation with three mixed sound sources is shown in Fig. 10. In order to ensure a sufficient number of bases, K is set to 180. Highly similar basis vectors were classified into each cluster. From the above, it can be found that the proposed clustering method can perform even for three sound sources mixed by a monaural microphone.

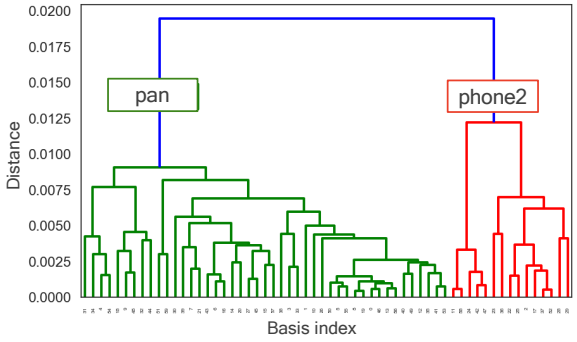

Fig. 5. Dendrogram for the pan mixture case.

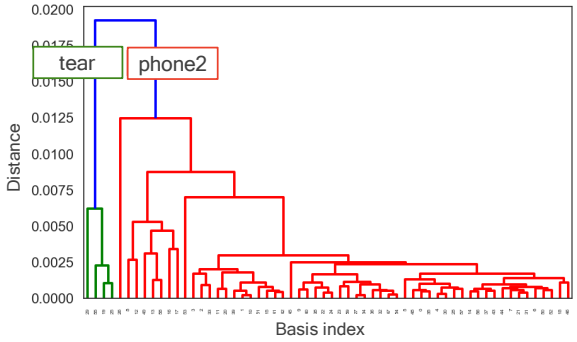

Fig. 6. Dendrogram for the tear mixture case.

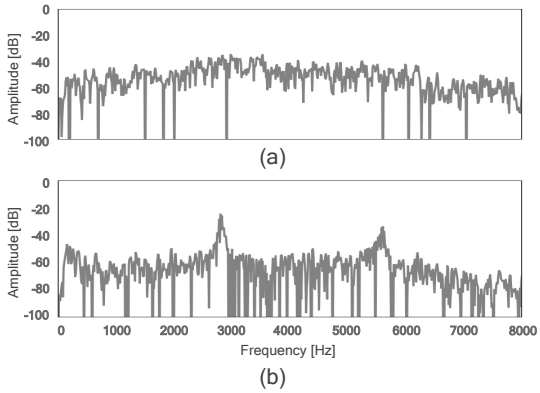

Fig. 7. Basis vectors represent the (a) tear and (b) cluster for "phone2."

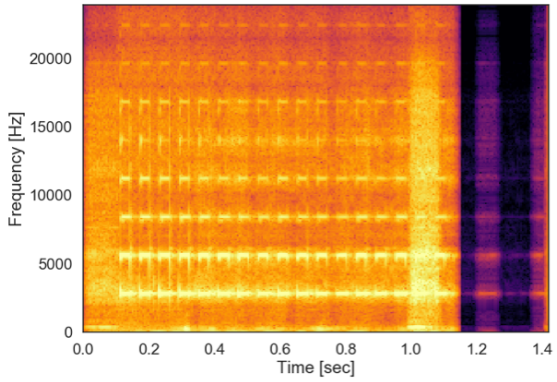

Fig. 8. Spectrogram of phone2 from speech mixture sound

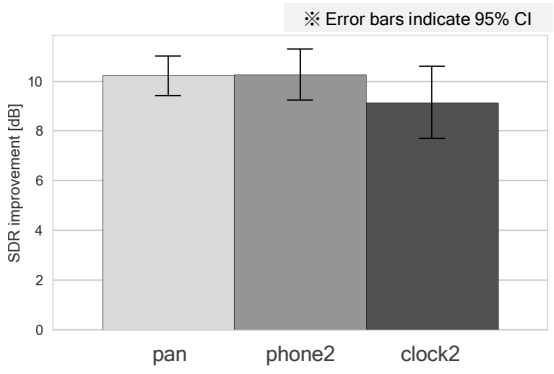

Fig. 9. SDR improvement by clustering the three sound sources.

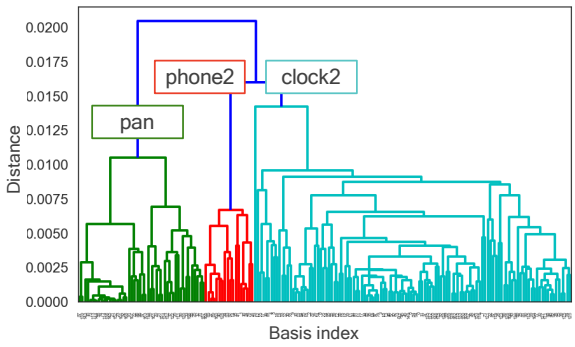

Fig. 10. Dendrogram of the three sound sources.

V. CONCLUSIONS

This paper proposed an unsupervised method of monaural sound source separation by clustering based on the similarity of basis vectors decomposed by NMF. Numerical experiments were conducted, and SDR improvements of approximately 6 dB to 10 dB were obtained. An experiment in which sounds from three sources were mixed by a monaural microphone was conducted, and an SDR improvement of about 10 dB was obtained. From this result, it can be concluded that the proposed clustering method can perform correctly for sounds from up to three sources mixed by a monaural microphone.

REFERENCES

[1] CV. Cotton, *et al*, "Spectral vs. spectro-temporal features for acoustic event detection," *In WASPAA*, 2011.
[2] T. Komatsu, *et al*, "Acoustic event detection method using semi-supervised non-negative matrix factorization with a mixture of local dictionaries," *In DCASE*, 2016.
[3] T. Heittola, *et al*, "Supervised model training for overlapping sound events based on unsupervised source separation," *In ICASSP*, pp. 8677–8681, 2013.
[4] N. Ono, "Auxiliary-function-based independent vector analysis with power of vector-norm type weighting functions," *in Proc. of APSIPA ASC*, 2012.
[5] D. Kitamura, *et al*, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE Trans. ASLP*, vol. 24, no. 9, pp. 1626–1641, 2016.
[6] D. D. Lee, *et al*, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.
[7] D. D. Lee, *et al*, "Algorithms for non-negative matrix factorization," *Proc. NIPS*, vol. 13, pp. 556–562, 2001.
[8] D. Griffin, *et al*, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. ASSP*, vol. 32, no. 2, pp. 236–243, 1984.
[9] H. Sawada, *et al*, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Trans. ASLP*, vol. 21, no. 5, pp. 971–982, 2013.
[10] P. Smaragdis, *et al*, "Supervised and semi-supervised separation of sounds from single-channel mixtures," *In Proc. of ICA*, pp. 414–421, 2007.
[11] D. Kitamura, *et al*, "Music signal separation based on supervised non-negative matrix factorization with orthogonality and maximum-divergence penalties," *IEICE Trans.*, Vol. 97, no. 5, pp. 1113–1118, 2014.
[12] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Trans. ASLP*, Vol. 15, no. 1, pp. 1–12, 2007.
[13] M. Helen, *et al*, "Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine,"*IEEE EUSIPCO.*, 13th, pp. 1–4, 2005.
[14] D. Mullner, "Modern hierarchical, agglomerative clustering algorithms," *arXiv:1109.2378*, 2011.
[15] S. Nakamura, *et al*, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," *Proc. LREC*, pp.965–968, 2000.
[16] R. Sonobe, *et al*, "JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis," *arXiv:1711.00354*, 2017.
[17] E. Vincent, *et al*, "Performance measurement in blind audio source separation," *IEEE Trans. ALSP*, vol.14, no.4, pp. 1462–1469, 2016.