ヴァーチャルマイクロフォンの外挿による マイクロフォン間隔の仮想的拡張 *

☆陣在遼河, 山岡洸瑛, 松本光雄, 山田武志, 牧野昭二 (筑波大)

1 はじめに

音像定位は空間的印象に関する聴覚事象である。両耳間時間差 (interaural time difference; ITD) 及び、両耳間レベル差 (interaural level difference; ILD) は、水平面内の音源について音像を特定する主要な手掛かりであり [1–3]、音源が聴取者の真横にある時、ITD は最大 (約 630 μs) となる。適切な ITD を持つ両耳信号を記録するために、ダミーヘッドを用いる方法が存在する。しかし、ITD は聴取者に依存しており、聴取者の正確なレプリカとなるダミーヘッドを用いることは非現実的である。

頭部伝達関数 (head related transfer function; HRTF) は、音源と外耳道入口との間の伝達関数として定義される。HRTF には空間的印象の知覚に関する多くの手掛かりが含まれているため、[4] で報告されているように、HRTF から ITD を抽出して、個人性の考慮をすることができる。また、測定された頭部伝達関数とシミュレートされた頭部伝達関数との差を用いて評価基準を作成するための技術として、頭部と胴体のシミュレータが [5] で提案されている。

近年、時間周波数領域における音響信号処理による音像定位が [6–9] で提案されている。これらの手法では、既に統合された信号から音源ごとに分離した信号を再度、再生時のチャネル間の振幅差によってパンニング (再配置) するが、この再生信号では聴取者は頭内に音像を知覚してしまう。

この問題に対し、マイクロフォンアレーを用いることができれば、マイクロホン信号間の時間差として、アレーからみた各音源の方向が記録される。これらの時間差は、音源の位置及びマイクロフォン間の距離に依存する。マイクロフォンアレーに使用されるマイクロフォンは、空間的エイリアシングを防ぐために隣接して配置される事が多く、その場合、マイクロフォン間の距離は、聴取者の両耳間距離とは異なることは明らかである。したがって、マイクロフォンアレーによって録音された音を聴取者が聴くと、音像はマイクロフォンから見た音源方向とは異なる方向に知覚される。前ら [10] は、ヴァーチャル多素子化によるアンラップ位相シフトを用いて、横方向に音像を定位させている。しかし、位相アンラップでは、複数の音像を異なる方向に移動させることはできない。

本研究では、ヴァーチャル多素子化 [11,12] によりマイクロフォンを仮想的に外挿し、マイクロフォン間距離を仮想的に聴取者の両耳間距離と等しくすることで ITD の復元を行う手法を提案する。本手法では、マイクロフォン間隔の仮想的拡張のために、音源数や到来方向の事前情報は不要であり、複数の音源が同時に存在していてもそれぞれの ITD を復元可能である。また、補間係数 α によって ITD の個人差を容易に調整できることが期待される。

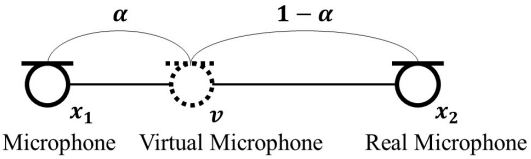

Fig. 1 Arrangement of real and virtual microphones in interpolation technique

2 ヴァーチャル多素子化と外挿への適用

2.1 ヴァーチャルマイクロフォンの内挿信号

ここでは、ヴァーチャル多素子化について紹介する。本手法では、時間周波数領域で 2 つの実マイクロフォンの観測信号 $x_i(\omega,t)$ からヴァーチャルマイクロフォン信号 $v(\omega,t,\alpha)$ を生成する。ここで $x_i(\omega,t)$ は、i 番目 $(i=1,2)$ のマイクロフォンにおける周波数ビン ω、時間フレーム t での複素信号を表す。α はヴァーチャルマイクロフォンの補間係数である。図 1 に実マイクロフォンとヴァーチャルマイクロフォンの関係を示す。

複数の音が異なる方向から到来する環境では、マイクロフォンの位置と波形の関係は複雑になり、補間が難しくなる。そこで、この手法では、観測信号の W-DO(W-disjoint orthogonality) を仮定する。W-DO は、信号スペクトルの強いスパース性であり、1 つの時間周波数ビンでは、ある 1 つの音源成分のみが支配的であるという仮定である。これにより、複数の音が到来した時であっても各時間周波数ビン内では単一の音とみなすことができ、ヴァーチャルマイクロフォン信号の補間を行うことができる。

ヴァーチャル多素子化では、位相と振幅が個別に補間される。実マイクロフォンでの観測信号 $x_i(\omega,t)$ の位相と振幅は次のようにして表される。

$$\phi_i = \angle x_i(\omega,t) = \tan^{-1}\frac{\mathrm{Im}(x_i(\omega,t))}{\mathrm{Re}(x_i(\omega,t))} \quad (1)$$

$$A_i = |x_i(\omega,t)| \quad (2)$$

平面波の到来を仮定したとき、ヴァーチャルマイクロフォンの位相 ϕ_v は線形補間によって次のように表すことができる。

$$\begin{aligned}\phi_\mathrm{v} &= \phi_1 + \alpha(\phi_2-\phi_1)\\ &= (1-\alpha)\phi_1 + \alpha\phi_2\end{aligned} \quad (3)$$

ここで、α はヴァーチャルマイクロフォンの補間係数であり、図 1 に示すように、2 つの実マイクロフォン間を $\alpha:(1-\alpha)$ に内分する位置にヴァーチャルマイクロフォンが補間されていることを示す。位相の値は n を自然数として $\phi_i \pm 2n\pi$ の任意性が生じることから、ここでは、信号の位相差が π を超えていないことを仮定して補間を行う。

$$|\phi_1-\phi_2| \le \phi \quad (4)$$

---

* Microphone position realignmnet by extrapolation of virtual microphone. by Ryoga JINZAI, Kouei YAMAOKA, Mitsuo MATSUMOTO, Takeshi YAMADA, Shoji MAKINO (University of Tsukuba)

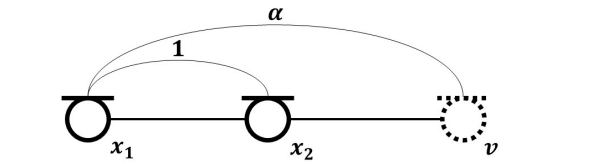

Fig. 2 Arrangement of real and virtual microphones in extrapolation technique

ヴァーチャルマイクロフォンの振幅の補間は、到来方向や残響音のような多くの条件に依存する。そのため、実際の振幅減衰を忠実にモデル化することは困難である。したがって、物理的なモデルの代わりに β ダイバージェンスを用いて振幅 A_{v} を補間する。

$$A_{\mathrm{v}} = \begin{cases} \exp\left((1-\alpha)\log A_1 + \alpha \log A_2\right) & (\beta = 1) \\ \left((1-\alpha)A_1^{\beta-1} + \alpha A_2^{\beta-1}\right)^{\frac{1}{\beta-1}} & (\text{otherwise}) \end{cases} \tag{5}$$

β の値によって、2 つの実マイクロホン間の振幅を非線形に補間する事ができる。これらより、ヴァーチャルマイクロフォン信号 $v(\omega, t, \alpha)$ は以下のように表すことができる。

$$v(\omega, t, \alpha) = A_{\mathrm{v}} \exp\left(j\phi_{\mathrm{v}}\right) \tag{6}$$

[11,12] では、ヴァーチャル多素子化を SN 比最大化ビームフォーマによる音声強調に用いている。

2.2 ヴァーチャルマイクロフォンの外挿信号

本稿では、音像定位のために、ヴァーチャルマイクロフォンの内挿を外挿へ適用する。図 2 に実マイクロフォンとヴァーチャルマイクロフォンの関係を示す。

ヴァーチャルマイクロフォンの外挿のために、前の節で述べたヴァーチャルマイクロフォンの生成方法の妥当性を確かめる必要がある。位相の外挿については、内挿における位相補間と同じ式を使用することができる (式 3)。一方、先述の通り内挿における振幅補間は困難であり、外挿ではさらに複雑になる。内挿と同じように式 5 を用いて外挿を行おうとすると、複素振幅や正の無限大への発散といった非現実的な振幅を生成してしまうことがあり、このような補間は避けなければならない。本論文では 1.5 kHz 以下の周波数帯域の信号において ILD は ITD に比べて音像の定位に大きく寄与しないため、ヴァーチャルマイクロフォンの位置に最も近い実マイクロフォンの振幅を外挿したヴァーチャルマイクロフォンの振幅として用いた。

$$A_{\mathrm{v}} = \begin{cases} A_1 & (\alpha < 0) \\ A_2 & (\alpha > 1) \end{cases} \tag{7}$$

これらより、外挿されたヴァーチャルマイクロフォン信号 $v(\omega, t, \alpha)$ は内挿と同様に以下のようにして表すことができる。

$$v(\omega, t, \alpha) = A_{\mathrm{v}} \exp\left(j\phi_{\mathrm{v}}\right) \tag{8}$$

ヴァーチャルマイクロフォンの推定信号は、時間周波数領域の信号を逆短時間フーリエ変換によって、時間領域の信号に変換することによって得られる。

本研究では、近接に配置された 2 つの実マイクロフォンに対してヴァーチャルマイクロフォンの外挿を適用し、マイクロフォン間隔を仮想的に拡張することで聴取者の両耳間距離と等しくする。これにより、実マイクロフォンとヴァーチャルマイクロフォンの信号は、聴取者が適切な音像を定位するのに必要な ITD を有した 1 対の両耳信号とみなすことができる。

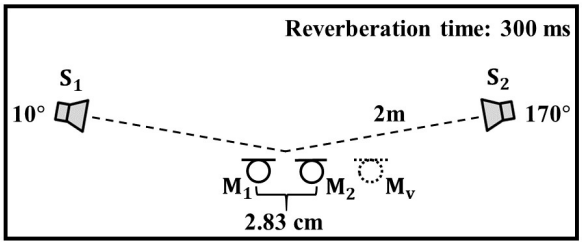

Fig. 3 Layout of sound sources and microphones in experiment

Table 1 Experimental conditions

Sampling rate	8 kHz
Real microphone interval	2.83 cm
Reverberation time	300 ms
FFT frame length / shift	1024 / 256 samples
Sound source 1	Female Japanese speech
Sound source 2	Male English speech

3 実験

本実験では、異なる方向から到来する 2 つの信号を用いて、提案手法が 2 つの実マイクロフォンの観測信号から 2 つの音源信号の位相を別々にシフトし、それぞれの信号の ITD が正しく復元されることを調べた。

3.1 実験条件

推定したヴァーチャルマイクロフォンの信号から 2 つの音源信号を抽出し、実マイクロフォンで観測される 2 つの音源信号と比較するためのシミュレーション実験を行った。

音源と実マイクロフォンの配置を図 3 に示し、その他の実験条件を表 1 に示す。M_1,M_2 はそれぞれ左右の実マイクロフォンであり、M_{v} はヴァーチャルマイクロフォンである。また、S_1,S_2 はそれぞれ左右から到来する音源信号である。M_1,M_2 での観測信号 x_1,x_2 は、音源信号 S_1,S_2 に対してインパルス応答を畳み込むことによって作成した。本研究で用いたインパルス応答は、RWCP 実環境音声・音響データベース [13] 内のものを使用した。実験に用いたインパルス応答は、残響時間 300 ms、音源距離 2 m、音源方向は 10° 及び 170° である。音源信号には女性の日本語音声と男性の英語音声を用いた。

ヴァーチャルマイクロフォンの外挿を利用して、x_1,x_2 から M_{v} における推定信号 v を推定する。評価を行うためには、ヴァーチャルマイクロフォンの時間周波数ビン内の信号が S_1 と S_2 のどちらのものかを見分ける必要がある。そのために、各時間周波数ビン内の S_1 及び S_2 のパワーを比較し、パワーの大きい方を選択するバイナリマスクを構成した。このバイナリマスクを v に適用することにより、S_1 及び S_2 それぞれの Mv での推定観測信号が抽出される。S_1 及び S_2 の M_1 と M_{v} での観測信号の位相を比較することにより、異なる方向から到来する 2 つの音が混合していても、それぞれの信号の ITD を個別に復元できることを確認する。そのために、本研究では両耳間相互相関 (interaural cross-correlation; IACC) を用いる。IACC は、聴取者が x_1 を左耳、v を右耳で聴いたときの音像の方向を示す。IACC に基づいて

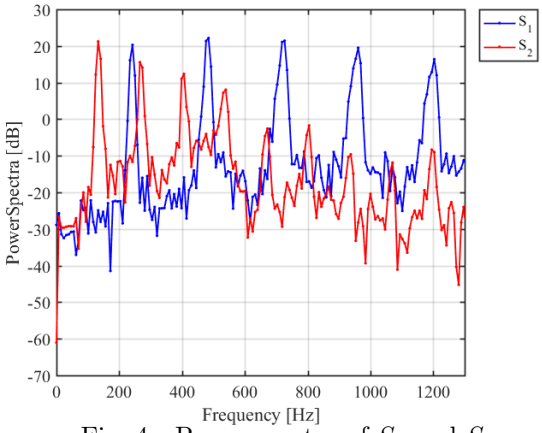

Fig. 4　Power spectra of S_1 and S_2

ITD を評価するために、x_1,v からバイナリマスクを用いて S_1,S_2 を抽出し、逆短時間フーリエ変換によって時間領域信号に変換した。M_1 と M_v で観測される S_1 と S_2 の信号波形から IACC を計算し、IACC が最大となる時間差が ITD となる。

$\alpha = 1$ の時、M_v は右側の実マイクロフォン M_2 に相当する。$|\alpha|$=8 の時、M_1,M_v 間の距離は人の両耳間距離にほぼ等しい。$\alpha < 0$ の時、M_v は M_1 の左側に外挿され、$\alpha > 1$ の時、M_2 の右側に外挿される。

3.2　実験結果

ある時間フレームにおける S_1,S_2 の信号のパワースペクトルを図 4 に示す。$\alpha = 1, 8, -8$ での実験結果をそれぞれ図 5〜7 に示す。それぞれの図における (a) は S_1 及び S_2 の M_1,M_v 間の位相差を表す。(b) 及び (c) はそれぞれ S_1,S_2 の M_1,M_v における観測波形を表し、(d) 及び (e) は S_1,S_2 の M_1,M_v 間の IACC を表す。

図 4 より、S_1 である女性日本語音声と S_2 である男性英語音声の周波数特性にはあまり重なりが見られず、スパースであることが分かる。図 5(a) より、S_1 は M_v よりも M_1 に近いため、M_1 における位相は M_v における位相よりも進んでいる (青線)。対照的に、S_2 については、M_v よりも M_1 の方が遠いため、M_1 における位相は M_v における位相よりも遅れている（赤線)。

M_1 と M_v における S_1 の信号波形を図 5（b）に示す。 M_1 は、M_v よりも S_1 に近いため、M_v よりも M_1 での観測波形の方がわずかに早い。一方、図 5（c）では、M_1 は M_v よりも S_2 から遠いため、M_1 への信号の到達はわずかに遅れている。

図 5(a) と図 6(a) を比較すると、図 6(a) の位相差は補間係数 α によって、図 5(a) の位相差の 8 倍となっている。図 5,6 の (b),(c) を比較しても時間差は 8 倍になっており、これは補間係数 α によって M_v が M_1,M_2 間距離の 8 倍の位置に外挿されたことを示している。

図 5(d)(e) より、IACC は $\tau = 0$ ms で 1 に近い。これは、M_1,M_2 の信号を聴取者が聴いたとき、S_1,S_2 の音像はほぼ正面に知覚されることを示唆している。それに対し、図 6(d) では、IACC は $\tau = -500\mu s$ で最大であり、これは聴取者が図 6(b) の信号を聴くとき、左端に音像を知覚すること示唆している。対照的に図 6(e) では、IACC が $\tau = 500\mu s$ で最大であり、こちらは右端に知覚されることを示唆している。

図 7 では補間係数 α が負であるため、IACC は図 6 の結果とは逆になっている。この場合に知覚される音像は、S_1,S_2 の位置が入れ替わって知覚される。

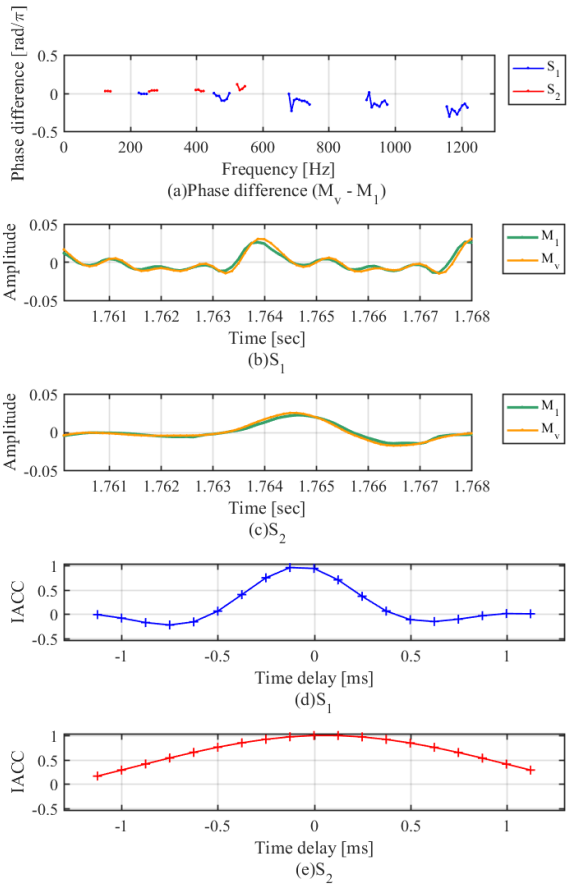

Fig. 5　Experimental result of $\alpha = 1$

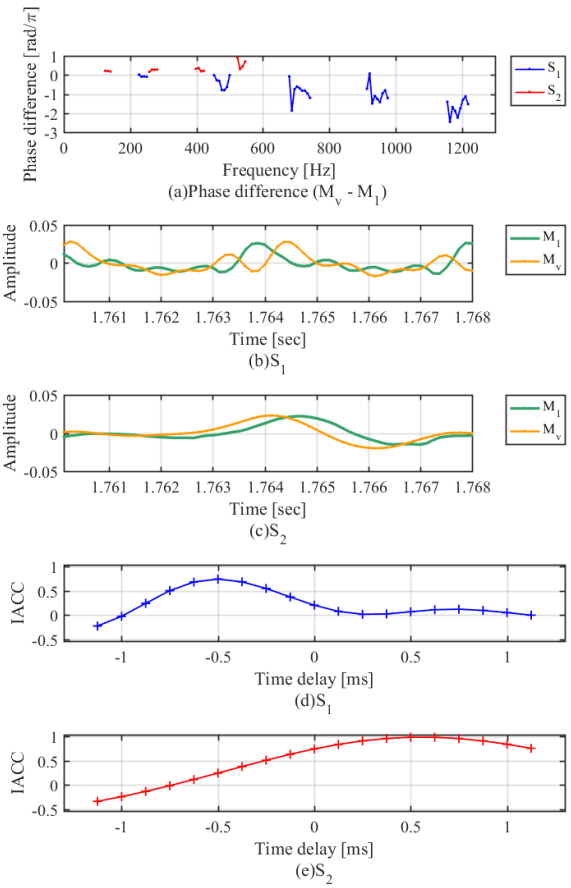

Fig. 6　Experimental result of $\alpha = 8$

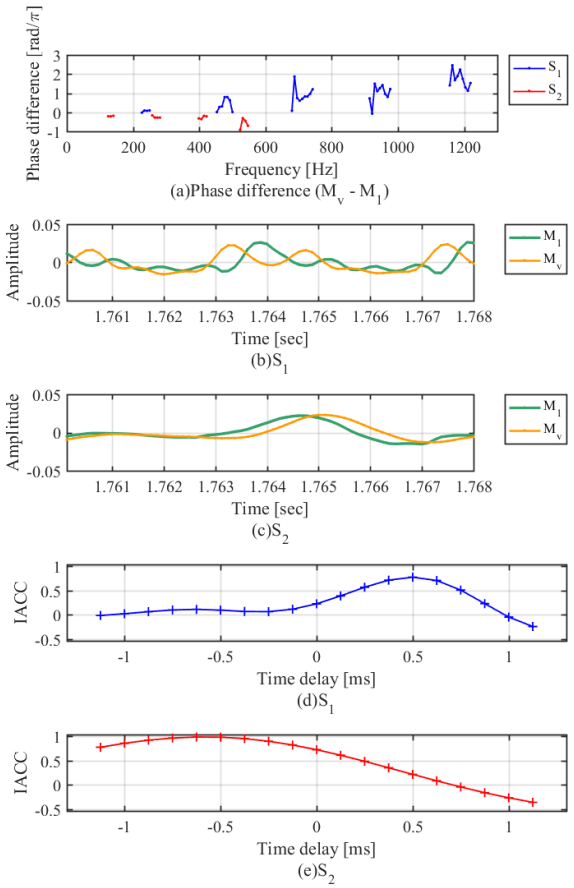

Fig. 7　Experimental result of $\alpha = -8$

4　結論

本論文ではヴァーチャルマイクロフォンの外挿を用いて、両耳間距離より狭い間隔で配置されたマイクロフォン間隔を仮想的に拡張することで、音像定位に適切な ITD の復元を行った。ヴァーチャルマイクロフォンを 2 つの実マイクロフォン信号を用いて外挿し、両耳間距離に等しい位置に配置する。これにより、2 つの実マイクロフォンのうちの一方と、ヴァーチャルマイクロフォンを用いて 1 対の適切な ITD を有する両耳信号とすることが可能である。

実験では 2 つの音源からの音が 2 つの近接に配置された実マイクロホンに到来する環境を仮定し、シミュレーション実験を行った。実マイクロフォン間隔の 8 倍の位置にヴァーチャルマイクロフォンを配置することで、実マイクロフォンの一方とヴァーチャルマイクロフォン間の距離は人の両耳間距離にほぼ等しくなり、これにより ITD が復元されることを、これら 2 つのマイクロフォンの信号波形から評価を行った。信号の IACC より、ITD が復元されることを確認した。これはマイクロフォンから見た音源方向に音像が知覚されることを示唆している。この手法は、音源数や音源方向に関する事前情報を必要とせず、様々な方向からの複数の音源が同時に存在していても ITD の復元が可能である。また、補間係数 α によりヴァーチャルマイクロフォンの位置を制御できるため、聴取者ごとに ITD の個人化に対応できると考えられる。

謝辞　本研究は、科研費 16H01735,SECOM 科学技術振興財団の助成を受けたため謝意を表する。

参考文献

[1] B. C. J. Moore, “An Introduction to the Psychology of Hearing,” 4^{th} Edition, Academic Press, 1997
[2] J. Blauert, “Spatial Hearing. The Psychology of Human Sound Localization,” MIT Press 1996
[3] S. Busson *et al.* “Subjective Investigations of the Interaural Time Difference in the Horizontal Plane,” in Proc. Audio Engineering Society Convention, convention paper 6324, May 2005
[4] M. Aussal *et al.* ”ITD Interpolation and Personalization for Binaural Synthesis using Spherical Harmonic,” AES Conference, 04-1-04-10 2012
[5] F. Brinkmann *et al.* “A High Resolution and Full-Spherical Head-Related Transfer Function Database for Different Head-Above-Torso Orientations,” J. Audio Eng. Soc., Vol.65, No. 10, pp. 841 - 848, October 2017
[6] A. Härmä and C Faller, “Spatial Decomposition of Time-Frequency Regions: Subbands or Sinusoids,” in Proc. Audio Engineering Society Convention, convention paper 6061, May 2004
[7] C. Avendano and J-M Jot, “A Frequency-Domain Approach to Multichannel Upmix,” J. Audio Eng. Soc., Vol. 52, No. 7/8, pp. 740-749, Jul/Aug 2002
[8] D. Barry *et al.* “Real-Time Sound Source Separation: Azimuth Discrimination and Resynthesis,” in Proc. Audio Engineering Society Convention, convention paper 6258, Oct 2004
[9] M. Combos and J. Lopez, “Interactive Enhancement of Stereo Recording Using Time-Frequency Selective Panning,” AES International Conference, pp. 1 - 12, Oct 2010
[10] N. Mae *et al.* “Ego Noise Reduction and Sound Localization Adapted to Human Ears using Hose-shaped Rescue Robot,” in Proc. International Workshop on Nonlinear Circuits, Communications and Signal Processing, pp. 371-374, March 2018
[11] H. Katahira *et al.* “Nonlinear Speech Enhancement by Virtual increase of Channels and Maximum SNR Beamformer”, EURASIP Journal on Advances in Signal Processing, vol. 2016, no. 1, pp. 1-8, Jan. 2016
[12] K. Yamaoka *et al.* ”Performance Evaluation of Nonlinear Speech Enhancement Based on Virtual Increase of Channels in Reverberant Environments,” in Proc. EUSIPCO, pp. 2388-2392, Aug. 2017
[13] S. Nakamura *et al.* “Design and Collection of Acoustic Sound Data for Hands-Free Speech Recognition and Sound Scene Understanding,” in Proc. ICME2002, Vol. 2, pp. 161-164, Aug. 2002