非同期スマートデバイスを用いた雑音に頑健な音声翻訳アプリケーションの検討

高島　遼一† 　孫　　慶華† 　住吉　貴志† 　戸上　真人†

† 日立製作所 研究開発グループ　〒 185–8601 東京都国分寺市東恋ケ窪 1–280
E-mail: †ryoichi.takashima.dh@hitachi.com

あらまし　本稿では，汎用スマートデバイスを用いた雑音に頑健な音声翻訳アプリケーションを検討する．提案する音声翻訳システムでは，二人のユーザがそれぞれスマートデバイスを持って対話を行う．各スマートデバイスで収録された音声は，雑音除去，音声認識，翻訳を行うサーバへ送信され，翻訳結果が返送される．雑音除去部においては，二個のスマートデバイスを擬似的なマイクロホンアレーとして活用することで，複数チャネルによる音源分離を行う．このとき，デバイス毎に録音開始時刻とサンプリングレートのミスマッチがあるため，相互相関関数による録音開始時刻の補正と，サンプリングレートのミスマッチに対して，位相差よりも頑健な音圧差を用いた音源分離方式を用いる．騒音レベル 70dB の雑音環境下において音声認識実験を実施した結果，雑音除去の無いシステムと比較して 21.2% の単語認識誤り率改善を確認した．

キーワード　音声翻訳，雑音除去，非同期マイクロホンアレー，スマートデバイス

Application of Noise-Robust Speech Translation Using Asynchronous Smart Devices

Ryoichi TAKASHIMA†, Qinghua SUN†, Takashi SUMIYOSHI†, and Masahito TOGAMI†

† Hitachi Ltd. Research and development group,　Higashi Koigakubo 1–280, Kokubunji-shi, Tokyo, 105–0123 Japan
E-mail: †ryoichi.takashima.dh@hitachi.com

Abstract　In this paper, we propose an application of noise-robust speech translation for general purpose smart devices. In the proposed speech translation system, two users have a conversation with their own smart devices installing speech translation applications. The recorded speech signals are sent to a server performing speech signal processing, speech recognition and translation, and translation results are returned to users' smart devices. In the speech signal processing part, a virtual microphone array is constructed from the microphones on users' smart devices, and a microphone-array-based noise reduction is performed. Then, because each smart device has different beginning time of recording and different sampling rate, we apply a cross-correlation-based compensation of beginning time and a signal separation method based on the difference of sound energy that is robust against the mismatch of sampling rate. We carried out a speech recognition experiment using two smart devices at a noisy environment in which noise level is 70dB, and as a result, the noise reduction process improved the word error rate of a speech recognition system without noise reduction by 21.2%.

Key words　speech translation application, noise reduction, asynchronous microphone array, smart device

1. 背　　景

音声翻訳 [1], [2] はユーザがある言語で発話した音声を認識し，別の言語に翻訳する技術である．音声翻訳システムは，公共交通機関や店舗などで外国人旅行客と係員の円滑なコミュニケーションや，外国人同士の会議など，さまざまな場面で活用が期待されている．これまで，スマートデバイス向けに様々な音声翻訳アプリケーションが開発されてきた．しかし，公共交通機関やショッピングセンターなどのシーンでは雑音により音声認識性能が低下するという課題があった．

従来の雑音除去技術として，スペクトルサブトラクション法 [3] やウィナーフィルタ [4]，minimum mean-square error short-term spectral amplitude (MMSE-STSA) [5]，optimally-modified log-spectral amplitude (OM-LSA) [6] などが単一マイクロホンによる雑音除去技術として研究されている．これらの手法は，時間毎に振幅が大きく変化しない定常な雑音は効果的に抑圧できるが，振幅が時々刻々変化する非定常な雑音は抑圧しきれないという課題がある．一方，目的音源と雑音源の方向の違いに着目し，マイクロホンアレーを用いて，マイク間の時間差情報により雑音除去を行う方式が，非定常な雑音にも有効な技術として研究されている [7]〜[11]．しかし従来のマイクロホンアレーによる雑音除去技術は，複数のマイクロホンが同期して音声を収録するための専用デバイスが必要であり，マイクが 1 個しかない汎用のスマートデバイスには適用できない．

マイクロホンアレーによる雑音除去技術を汎用のスマートデバイスを用いて実現するために，非同期マイクロホンアレーの研究がされている [12], [13]．これらの研究では，複数のスマートデバイスのマイク入力を用いて擬似的なマイクロホンアレーを構築し，雑音除去技術を適用する．このとき，汎用のスマートデバイスではサンプリング周波数が端末個体差により微小に異なるため，各マイク入力信号の時間差が時間毎に変化していく．そのため従来のビームフォーミング [8] や独立成分分析のような時間差を利用したマルチチャネル雑音除去技術を単純に適用することが困難であることが課題となる．これに対して文献 [13] ではサンプリング周波数のミスマッチを補完する手法を提案している．また文献 [12], [14] では，時間差情報に比べてサンプリング周波数のミスマッチに頑健な音圧差情報を用いて音源分離を行う手法を提案している．

提案する音声翻訳システムでは，二人のユーザがそれぞれスマートデバイスを持って対話を行う．各スマートデバイスで収録された音声は，雑音除去，音声認識，翻訳を行うサーバへ送信され，翻訳結果が返送される．雑音除去部においては，二個のスマートデバイスを擬似的なマイクロホンアレーとして活用することで，複数チャネルによる音源分離を行う．音源分離手法として文献 [12], [14] と同様に音圧差情報を用いた音源分離方式である時変ガウスモデリングに基づく方式 [11] を用いる．

騒音レベル 70dB の雑音環境下において音声認識実験を実施した結果，雑音除去の無いシステムと比較して 21.2% の単語認識誤り率改善を確認した．

2. 提案システムの概要

提案する音声翻訳システムの構成を図 1 に示す．提案システムでは，二人のユーザがそれぞれ汎用のスマートデバイスを用いて対話することを想定している．スマートデバイスの GUI では，ユーザごとにあらかじめ使用言語を設定しておく．ユーザ同士のスマートデバイスをペアリングすることで，各ユーザの使用言語間で音声翻訳が実施される．スマートデバイスで収録された音声はサーバへ送信され，サーバ上で，雑音除去，音声認識，音声認識結果テキストの翻訳が行われ，翻訳結果がサーバからタブレットへ送信，GUI に出力される．本システムにおいて音声認識，認識結果テキストの翻訳部は従来システムと同様である．ただし従来の単一スマートデバイスによるシステムと比べて，複数スマートデバイスを使用する本システムでは，従来のように発話毎にどの言語からどの言語への翻訳かを指定する必要がない．また，二つのスマートデバイスのマイク入力を用いることで，非定常雑音に有効な複数チャネル雑音除去を適用可能である．次章では，雑音除去手法について説明する．

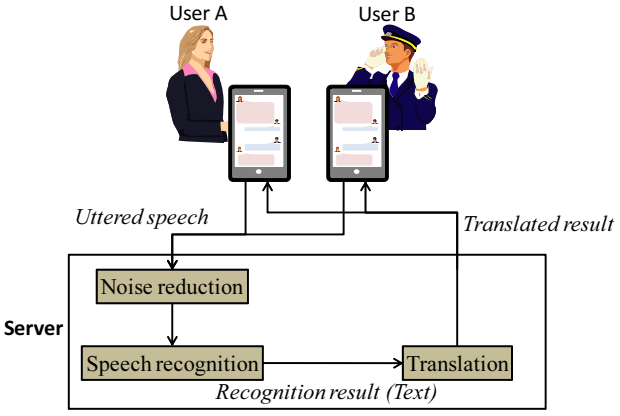

図 1　Proposed speech translation system

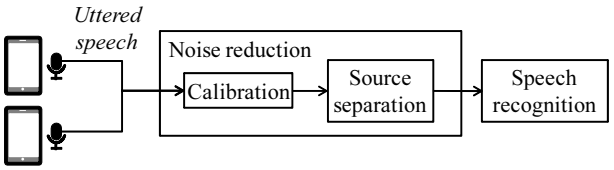

図 2　Overview of noise reduction part

3. 雑音除去手法

3.1 概　　要

非定常雑音除去に有効な，従来のマイクロホンアレーに基づく雑音除去方式は，マイクが 1 個しか無い汎用スマートデバイスには適用できない．これに対して提案システムにおいては，2 台のスマートデバイスが存在しているため，各スマートデバイスのマイクを用いて非同期なマイクロホンアレーを構築して，雑音除去を行う．2 台の非同期なスマートデバイスを用いる場合，各端末の音声取得タイミングが個体差によって微小にずれるため，従来の時間差を用いた雑音除去方式をそのまま適用することが困難である．これに対して本システムでは，音声取得タイミングずれによる影響が小さい，マイク間の音圧差情報を使った方式である，時変ガウスモデリングに基づく手法 [11] を用いることで，音声取得タイミングずれに頑健な雑音除去を行う．

提案する雑音除去システムの概要を図 2 に示す．雑音除去は大きくキャリブレーション，信号分離（それぞれ図中 Calibration, Source separation）の処理に分かれる．キャリブレー

ションはスマートデバイスのペアリング時に 1 度だけ実行される．信号分離は，雑音除去用のバッファにマイク入力信号が 0.5 秒たまるごとにリアルタイムで実行される．分離されたそれぞれのユーザの発話は，各ユーザが事前に設定した言語によって音声認識される．

3.2 キャリブレーション

キャリブレーションはスマートデバイスのペアリング時に 1 度だけ実行される．キャリブレーションの目的は，各スマートデバイスの収録開始時間の差を推定し，補正することである．キャリブレーションは下記の手順で行われる．2 台のスマートデバイスをそれぞれスマートデバイス A，B とする．

（1） 各スマートデバイスの収録音声のバッファをクリア．

（2） どちらかのスマートデバイスのスピーカから参照信号を出力する．

（3） スマートデバイス A とスマートデバイス B の収録時間差を推定・補正する．

各スマートデバイスでそれぞれアプリケーションが起動し，収録を開始すると収録音声が内部バッファに格納され始める．このとき，それぞれのデバイスで収録を開始する時間は異なるため，各デバイスの内部バッファの時間同期が取られていない．そこで処理 1 では，収録音声のバッファをクリアすることで，各スマートデバイスでアプリケーションが起動して収録を開始した時間の差をリセットする．ただし，サーバ側から同時にバッファクリア命令を送ったとしても，命令伝送時間が異なるため，処理 1 だけではスマートデバイス間の収録時間差は完全には無くならない．そこで処理 2，3 により収録時間差をサンプル点レベルで補正する．

まず参照信号をスマートデバイス A のスピーカから出力させる．出力された参照信号は，スマートデバイス A，B それぞれのマイクによって収録される．このとき，参照信号と各マイクの収録信号との時間差を推定する．推定方法として，音源方向推定に用いられている手法である GCC-PHAT (Generalized Cross-Correlation PHAse Transform) 法 [15] を用いる．スマートデバイス A，B のマイク収録信号のスペクトルをそれぞれ $x_A(f)$，$x_B(f)$，とする．f は周波数ビンのインデクスを表す．GCC-PHAT 法は，クロススペクトルに逆フーリエ変換を行って相互相関関数を算出する際に，パワーで正規化することによって信号間の時間差を強調させる手法で，相互相関関数を用いるよりも時間差の推定精度が高いことが知られている．スマートデバイス A，B のマイク間の GCC-PHAT $G_{(x_A,x_B)}(f)$ で定義される．

$$G_{(x_A,x_B)}(f) = \frac{x_A(f)x_B^*(f)}{|x_A(f)||x_B(f)|} \tag{1}$$

$*$ は共役複素数を表す．GCC-PHAT $G_{(x_A,x_B)}(f)$ を逆フーリエ変換したものを $G_{(x_A,x_B)}(t)$ とすると，時間差の推定値 $\hat{t}_{(x_A,x_B)}$ は次式によって求められる．

$$\hat{t}_{(x_A,x_B)} = \underset{t}{\mathrm{argmax}}\, G_{(x_A,x_B)}(t) \tag{2}$$

式 (2) において，$G_{(x_A,x_B)}(t)$ の最大値は t の全探索により特定される．処理 1 により，バッファがクリアされることで収録時間差が縮められるため，探索領域を狭めることが可能となる．上記によって推定した時間差を用いて，信号間の時間差を補正する．具体的には，各信号のバッファ内で推定した時間差分だけデータをずらすことで補正を行う．

3.3 信号分離

従来のマイクロホンアレーを用いた信号分離手法に対して，本研究では，2 台の汎用スマートデバイスのマイク（それぞれマイク A，マイク B と呼ぶ）を用いて非同期のマイクロホンアレーを構築して信号分離を行う．本研究では，時間差情報と比べてサンプリング周波数のミスマッチに影響されにくい，マイク間の音圧差情報を使った方式として，時変ガウスモデリングに基づく信号分離手法 [11] (LGM; Local Gaussian Model) を用いる．

LGM では，複数マイクで収録した音声信号の分散共分散行列を時刻毎に推定し，推定した分散共分散行列を用いて分離フィルタを求める．周波数ビン f，フレーム τ における 2ch の短時間複素スペクトルを $\boldsymbol{x}(f,\tau) = [x_A(f,\tau), x_B(f,\tau)]^T$ とする．T は行列またはベクトルの転置を表す演算子である．本研究では，入力音声を二人の話者と背景雑音の 3 音源に分離する．$\boldsymbol{x}(f,\tau)$ を 3 音源の音の混合信号と仮定し，次式のようにモデル化する．

$$\begin{aligned}\boldsymbol{x}(f,\tau) &= \sum_{n=0,1,2} \boldsymbol{c}_n(f,\tau) \\ &= \sum_{n=0,1,2} s_n(f,\tau)\boldsymbol{a}_n(f,\tau)\end{aligned} \tag{3}$$

n $(n = 0,1,2)$ は音源のインデクスを表す．$\boldsymbol{c}_n(f,\tau) = s_n(f,\tau)\boldsymbol{a}_n(f,\tau)$ は音源毎の 2ch 信号を表す．また $s_n(f,\tau)$ は，各音源の原信号を，$\boldsymbol{a}_n(f,\tau)$ を各音源からマイクまでの伝達関数をそれぞれ表す．ここで，音源毎の 2ch 信号 $\boldsymbol{c}_n(f,\tau)$ を，平均 0，共分散行列 $v_n(f,\tau)\boldsymbol{R}_n(f)$ の多次元正規分布に従う確率変数してモデル化する．この場合，マイク入力信号の時間周波数毎の確率分布は，同様に平均 0，共分散行列 $\boldsymbol{R}_x(f,\tau)$ の多次元正規分布としてモデル化でき，$\boldsymbol{R}_x(f,\tau)$ は次式のようにモデル化できる．

$$\boldsymbol{R}_x(f,\tau) = \sum_{n=0,1,2} v_n(f,\tau)\boldsymbol{R}_n(f) \tag{4}$$

$\boldsymbol{R}_n(f)$ は音源毎の空間相関行列，$v_n(f,\tau)$ は各音源の周波数・フレーム毎のアクティビティを表す．

LGM による音源分離手法では，EM アルゴリズム [16] により分散共分散行列の推定と音源分離を行う．以下，EM アルゴリズムによる推定アルゴリズムの i ステップ目の手順を記す．

E ステップ

分離信号の十分統計量を以下のステップ求める．まず，n 番目の音源の信号 $\boldsymbol{c}_n(f,\tau)$ を分離するための分離フィルタ $\boldsymbol{W}_n(f,\tau)$ を時間周波数毎に次式で求める．

$$\boldsymbol{R}_{c_n}(f,\tau) = v_n(f,\tau)\boldsymbol{R}_n(f)^{(i-1)} \tag{5}$$

$$\boldsymbol{R}_x(f,\tau) = \sum_{n=0,1,2} \boldsymbol{R}_{c_n}(f,\tau) \tag{6}$$

$$\boldsymbol{W}_n(f,\tau) = \boldsymbol{R}_{c_n}(f,\tau)\boldsymbol{R}_x(f,\tau)^{-1} \quad (7)$$

求めた分離フィルタ $\boldsymbol{W}_n(f,\tau)$ を用いて，音源毎の音声の推定解 $\hat{\boldsymbol{c}}_n(f,\tau)$ を次式で求める．

$$\hat{\boldsymbol{c}}_n(f,\tau) = \boldsymbol{W}_n(f,\tau)\boldsymbol{x}(f,\tau) \quad (8)$$

更に，M ステップでのパラメータ更新に必要な十分統計量を次式で求める．

$$\hat{\boldsymbol{R}}_{c_n}(f,\tau) = \hat{\boldsymbol{c}}_n(f,\tau)\hat{\boldsymbol{c}}_n(f,\tau)^H + (\boldsymbol{I} - \boldsymbol{W}_n(f,\tau))\boldsymbol{R}_{c_n}(f,\tau) \quad (9)$$

ここで，$\boldsymbol{I}$ は単位行列，H は行列またはベクトルの共役転置を表す演算子である．

M ステップ

E ステップで得られた，分離信号の十分統計量を用いて，アクティビティおよび空間相関行列を以下のように更新する．

$$v_n^{(i)}(f,\tau) = \frac{1}{3}tr(\boldsymbol{R}_n^{-1}(f)^{(i-1)}\hat{\boldsymbol{R}}_{c_n}(f,\tau)) \quad (10)$$

$$\boldsymbol{R}_n(f)^{(i)} = \frac{1}{\sum_\tau v_n(f,\tau)}\sum_\tau \hat{\boldsymbol{R}}_{c_n}(f,\tau) \quad (11)$$

このように，E ステップでは信号の分離を，M ステップでは分散共分散行列の推定を交互に行うことで，音源分離を行う．

LGM では，使用者の音声に関する事前情報を使わず，ブラインドで信号を分離している．そのため，LGM により分離された 3 個の音声信号に対して，どれがスマートデバイス A，あるいは B の使用者の音声成分か，あるいは背景雑音成分かを，時間・周波数毎に自動的に割り当てる必要がある．本研究では，各ユーザの音声は，自身のスマートデバイスのマイクには大きく収録され，相手のスマートデバイスのマイクには小さく収録され，さらに背景雑音は各スマートデバイスにほぼ同じ音圧で収録されると仮定し，次式のように，分離されたマルチチャネル信号の音圧比を用いて割り当てを行う．

$$n(f,\tau) = \begin{cases} A & (n = \underset{n=0,1,2}{\operatorname{argmax}} \frac{|c_{n,A}(f,\tau)|^2}{|c_{n,B}(f,\tau)|^2}) \\ B & (n = \underset{n=0,1,2}{\operatorname{argmax}} \frac{|c_{n,B}(f,\tau)|^2}{|c_{n,A}(f,\tau)|^2}) \\ C & (otherwise) \end{cases} \quad (12)$$

4. 評価実験

4.1 実験環境

提案システムの性能を評価するため，2 台の汎用タブレットを用いて音声認識実験を行った．実験環境を図 3 に示す．実験では，各ユーザがタブレットを手に持って対話するシーンを想定して，2 台のタブレットを 50cm 離した状態で斜めに向けて固定した．うち 1 台のタブレットから 20cm 離した場所をユーザの口元位置と想定してスピーカを固定し，音声を再生した．また，2 台のタブレットの垂直線上 150cm 離した場所にもう 1 台スピーカを固定し，雑音を再生した．

音声は，日本語会話文章 100 文を読み上げたものを，あらかじめ接話マイクで収録し，スピーカで再生した．雑音として，The 3rd CHiME Challenge [17] の雑音データベースより街頭雑音をスピーカで再生した．音声用スピーカに近い方のタブレットのマイク付近において，音声の騒音レベルは平均約 75dB，雑音の騒音レベルは平均約 70dB となるようにした．

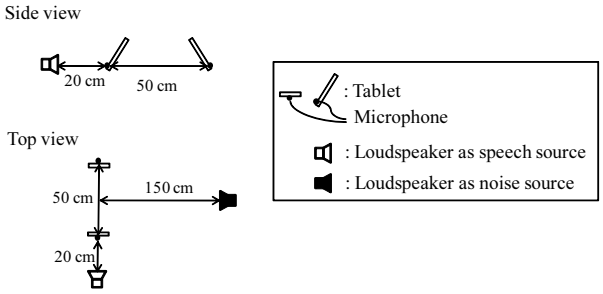

図 3 Experimental environment

サンプリングレートは 16,000 Hz である．雑音除去においては，窓幅は 512 ポイント，フレームシフトは 256 ポイントとした．音声認識においては，窓幅は 20 msec，フレームシフトは 10 msec とし，12 次元の MFCC と，ΔMFCC，Δ 対数パワーの 25 次元を特徴量として隠れマルコフモデルによる文章認識を実施した．

4.2 実験結果

実験条件として，下記の 3 条件を評価した．

（1） Clean speech condition: 雑音を再生せずに音声のみを収録．雑音除去を実施しない．

（2） w/o noise reduction: 雑音を再生して音声を収録．雑音除去を実施しない．

（3） w/ noise reduction: 雑音を再生して音声を収録．雑音除去を実施する．

条件 1,2 はそれぞれクリーン音声環境，雑音環境における従来システムの評価を，条件 3 は雑音環境における提案システムの評価を目的としている．

各実験条件における word error rate (WER) と sentence error rate (SER) を図 4 に示す．図 4 より，雑音除去の無いシステムでは雑音により WER，SER がそれぞれ 21.4%，36.5% 低下するが，雑音除去を行うことで，WER と SER をそれぞれ 21.2%，31.6% 改善することを確認した．

5. まとめ

本稿では，汎用のスマートデバイス向けの雑音に頑健な音声翻訳アプリケーションの検討を行った．提案システムでは，二人のユーザがそれぞれのスマートデバイスを持って会話を行う．そして，二つのスマートデバイスで収録された音声を用いてサーバ上で複数マイクに基づく雑音除去を実施する．このとき，収録開始時刻のずれは相互相関関数に基づいて補正をし，またサンプリングレートのミスマッチに頑健な，マイク間の音圧差を利用した音源分離を適用した．70dB の雑音環境におけ

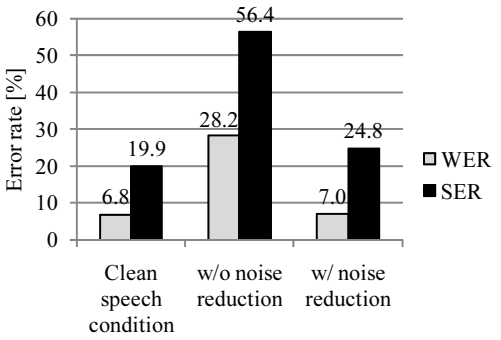

図 4 Word error rate (WER) [%] and Sentence error rate (SER) [%] of each condition

る音声認識実験の結果，雑音除去処理により，雑音除去の無いシステムと比較して WER が 21.2%改善された．今後は，ユーザが 3 人以上存在するときの構成や，デバイスを持つ手のぶれにより話者方向が時間変化する場合について検討を行う．

文　献

[1] W. Wahlster, “Verbmobil: Foundations of speech-to-speech translation,” Springer, 2000.
[2] X. He and L. Deng, “Speech recognition, machine translation and speech translation - a unified discriminative learning paradigm,” IEEE Signal Proc. Mag, vol. 27, pp. 126-133, 2011.
[3] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Trans. ASSP, vol. 27, no. 2, pp. 113-120, 1979.
[4] P. C. Loizou, “Speech Enhancement: Theory And Practice,” Signal Processing and Communications, Crc Pr Llc, 2007/6.
[5] Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” IEEE Trans. Acoust. Speech Signal Process., vol. ASSP-32, no. 6, pp. 1109-1121, 1984.
[6] I. Cohen and B. Berdugo, “Speech enhancement for non-stationary noise environments,” Signal Processing, vol. 81, no. 11, pp. 2403-2418, 2001.
[7] D. Johnson and D. Dudgeon, “Array Signal Processing,” Upper Saddle River, NJ, USA: Prentice Hall, 1996.
[8] O. L. F. III, “An algorithm for linearly constrained adaptive array processing,” in Proc. IEEE, vol. 60(8), 1972, pp. 926-935.
[9] E. O. A. Hyvärinen, J. Karhunen, “Independent component analysis,” John Wiley & Sons, 2001.
[10] A. Ozerov and C. Févotte, “Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation,” IEEE Trans. ASLP, vol. 18, no. 3, pp. 550-563, 2010.
[11] N. Duong, E. Vincent, and R. Gribonval, “Under-determined reverberant audio source separation using a full-rank spatial covariance model,” IEEE Trans. ASLP, vol. 18, no. 7, pp. 1830-1840, 2010.
[12] H. Chiba et al., "Amplitude-based speech enhancement with nonnegative matrix factorization for asynchronous distributed recording," Proc. IWAENC, pp. 204-208, Sept. 2014.
[13] S. Miyabe, N. Ono, and S. Makino, “Blind compensation of interchannel sampling frequency mismatch for ad hoc microphone array based on maximum likelihood estimation,” Elsevier Signal Processing, vol. 107, pp. 185–196, 2015.
[14] 戸上真人，川口洋平，小窪浩明，大淵康成, “音源のチャンネル間振幅差を基底ベクトルとする音源分離,” 音講論集, pp. 803-804, 2010.
[15] C. Knapp and G. Carter, “Time delay estimation by generalized cross correlation methods,” IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. 24-4, pp. 320-327, 1976.
[16] A.P. Dempster et al., “Maximum likelihood from incomplete data via the EM algorithm,” *J. of the Royal Statistic Society*, Series B 39(1),pp. 1–38, 1977.
[17] J. Barker et al., “The third ’CHiME’ Speech Separation and Recognition Challenge: Dataset, task and baselines,” Proc. ASRU, pp. 504-511, 2015.