DOA-HMMに基づく移動音源の劣決定ブラインド音源分離*

☆樋口卓哉，高宗典玄，中村友彦，亀岡弘和 (東大情報理工)

1 はじめに

本稿では，移動音源を対象とした劣決定ブラインド音源分離の問題を扱う。ブラインド音源分離 (Blind Source Separation; BSS) とは，音源から観測信号までの伝達特性が未知である場合に，複数の音声が混合した信号から元の音声信号を分離する技術である。例えば，会議において複数の音声信号の混じった録音データから会議録を自動作成したり，ロボットに周囲の音環境を認識する機能を備えさせる用途への応用が期待されている。

BSS では観測信号から音源信号とその混合過程を推定する必要があるため，通常は音源やその混合過程に対して何らかの仮定を置き，その仮定により立てられる規準をもとに未知変数を推定する最適化問題として定式化される。例えば，BSS において観測信号数が音源数よりも多い優決定問題では，音源信号間の独立性を仮定して分離する独立成分分析 (Independent Component Analysis; ICA) が有用であることが知られており，音源信号間の独立性を最大化するように分離フィルタを推定することが目的となる [1]。しかし，ICA では観測信号数が音源数よりも少ない劣決定問題を扱うことはできず，この場合は独立性よりもさらに強い仮定が必要である。

音声を対象とした劣決定 BSS では，音声の時間周波数成分のスパース性を利用したアプローチが有効であることが知られている [2–9]。音声のスパース性とは，音声信号の時間周波数成分がほとんどの時間周波数点においてほぼ 0 となる性質である。この性質により，複数の音声が同時に発話された状況でも，各音声の優勢な時間周波数成分が互いにほとんど重なり合わないと仮定できる場合が多い。よって，チャネル間の位相や振幅の違い等を手掛かりとして各時間周波数点でどの音源が最も優勢らしいかを推定できれば，目的の音声信号のみを通過させる時間周波数マスクを設計することで分離信号を得ることができる。

以上の音声のスパース性を観測信号のモデルに組み込むためには，観測信号のモデルを時間周波数領域で定式化する必要がある。通常，各マイクロフォンの観測信号は音源信号の時間遅れを含む畳み込み混合で表されるが，音源からマイクロフォンまでのインパルス応答長に対して十分に長い時間窓をもつ時間周波数分解を用いると，畳み込み混合を近似的に瞬時混合で表すことができる。この観測信号モデルに基づく BSS は周波数領域 BSS と呼ばれ，時間領域の BSS に対して演算量の少ないアルゴリズムを実現できる点や，音声のスパース性を組み込める点など特徴がある一方で，周波数ごとに分離した信号を音源ごとにグルーピングするパーミュテーション整合と呼ぶ問題を解決する必要がある。

本研究の目的は，各音源が移動した場合にも音源位置を追跡しながら適切に音源分離を行える手法を実現することである。我々は以前，音源到来方向を離散値の潜在変数と扱い，その混合モデルにより各音源のステアリングベクトルを確率モデル化し，観測信号の生成モデルに組み込むことでパラメータ推論を通してパーミュテーション整合と周波数領域 BSS を同時に行うアプローチを提案した [8] (なお，ほぼ同時期に大塚らによっても類似したアプローチが提案されている [9])。本稿ではこれを拡張し，時間変化する各音源のステアリングベクトルを，離散化された各角度を状態とする隠れマルコフモデル (Hidden Markov Model; HMM) により確率モデル化し，観測信号の生成モデルに組み込み，パラメータ推論を通してパーミュテーション整合，各移動音源の到来方向追跡，周波数領域 BSS を同時に行う手法を提案する。

2 観測モデル

I 個の音源から到来する信号を M 個のマイクロフォンで観測する場合を考え，m 番目のマイクロフォンで観測される信号の時間周波数成分を $y_m(\omega_k, t_l)$，i 番目の音源信号の時間周波数成分を $s_i(\omega_k, t_l)$ とし，$\boldsymbol{y}(\omega_k, t_l) = (y_1(\omega_k, t_l), \ldots, y_M(\omega_k, t_l))^{\mathsf{T}} \in \mathbb{C}^M$，$\boldsymbol{s}(\omega_k, t_l) = (s_1(\omega_k, t_l), \ldots, y_I(\omega_k, t_l))^{\mathsf{T}} \in \mathbb{C}^I$ とする。ただし，$1 \leq k \leq K$, $1 \leq l \leq L$ は時間周波数領域においてそれぞれ周波数および時間に対応するインデックスである。先に述べた通り，時間周波数領域において観測信号 $\boldsymbol{y}(\omega_k, t_l)$ は近似的に

$$\boldsymbol{y}(\omega_k, t_l) = \sum_{i=1}^{I} \boldsymbol{a}_i(\omega_k) s_i(\omega_k, t_l) + \boldsymbol{n}(\omega_k, t_l) \quad (1)$$

のように $s_1, \ldots, s_I$ の瞬時混合の形で表すことができる。ここで，$\boldsymbol{a}_i(\omega_k)$ は音源 i のステアリング (方向) ベクトルを表し，これを並べた行列 $\boldsymbol{A}(\omega_k) =$

---

*Underdetermined blind separation of moving sound sources based on DOA-HMM. by HIGUCHI Takuya, TAKAMUNE Norihiro, NAKAMURA Tomohiko, KAMEOKA Hirokazu (Graduate School of Information Science and Technology, The University of Tokyo)

$(\boldsymbol{a}_1(\omega_k), \ldots, \boldsymbol{a}_I(\omega_k)) \in \mathbb{C}^{M\times I}$ を混合行列と呼ぶ。$\boldsymbol{n}(\omega, t)$ は背景雑音やフレーム長を超える残響成分などである。音声のスパース性を仮定し，各時間周波数点 (ω_k, t_l) においてアクティブである音源のインデックスを $z_{k,l} \in \{1, \ldots, I\}$ と表すと，式 (1) は

$$\boldsymbol{y}(\omega_k, t_l) = \boldsymbol{a}_{z_{k,l}}(\omega_k) s(\omega_k, t_l) + \boldsymbol{n}(\omega_k, t_l) \quad (2)$$

のように書き直せる。この観測モデルおいては，各時間周波数点において $z_{k,l}$ 番目の音源以外の成分はすべて 0 と仮定されたことになる。従って各時間周波数で音源成分を表す変数は $z_{k,l}$ のみで十分であり，このため上式では $s_i(\omega_k, t_l)$ のインデックス i を省いている。すなわち $s(\omega_k, t_l)$ は各時間周波数点においてアクティブないずれかの音源の成分を表す変数となる。以後紙面のスペースの節約のため，ω_k と t_l を下付き添え字 k, l で表記することにする。

3 生成モデル

3.1 観測信号の生成プロセス

観測モデルをもとに，観測信号が生成されるプロセスを生成モデルにより記述する。

まず，雑音成分 $\boldsymbol{n}_{k,l}$ が，平均が 0，共分散が $\Sigma_k^{(n)}$ の複素正規分布に従うと仮定すると，もし $\boldsymbol{a}_{1:I,k} = \{\boldsymbol{a}_{1,k}, \ldots, \boldsymbol{a}_{I,k}\}$,$s_{k,l}$ および $z_{k,l}$ が既知であれば，式 (2) より $\boldsymbol{y}_{k,l}$ は

$$\boldsymbol{y}_{k,l} | \boldsymbol{a}_{1:I,k,l}, s_{k,l}, z_{k,l} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{a}_{z_{k,l},k} s_{k,l}, \Sigma_k^{(n)}) \quad (3)$$

により生成される。ここで，$z_{k,l}$ を離散値の潜在変数と見なせば，$\boldsymbol{y}_{k,l}$ の確率分布は混合正規分布となる [6,7]。和泉らは，この確率モデルに基づき，Expectation-Maximization (EM) アルゴリズムにより最尤の時間周波数マスクを推定するアプローチを提案している [6]。

3.2 混合 DOA モデル [8]

本節ではまず音源位置が固定の場合を考え，次節で音源が移動する場合を考える。これまで各音源の伝達周波数特性 $\boldsymbol{a}_{i,k}$ を周波数インデックス k ごとに独立な変数であるかのように扱っていたが，もし各音源が単一方向から平面波として到来すると仮定できるならば，例えばマイクロフォン数が 2 の場合，伝達周波数特性 $\boldsymbol{a}_{i,k}$ は，到来方向 (Direction-of-Arrival; DOA)θ の関数として

$$\boldsymbol{h}(\theta, \omega) = \begin{bmatrix} 1 \\ e^{\mathrm{J}\omega B \cos\theta / C} \end{bmatrix} \quad (4)$$

として陽に表される。ただし，$0 \le \theta \le 2\pi$，B をマイクロフォンの間隔 (m)，C を音速 (m/s) とする。実際には残響や時間周波数領域の瞬時混合近似などにより，$\boldsymbol{a}_{i,k}$ は上記の理論式から逸脱することが予想される。そこで，音源 i の到来方向 θ_i が既知のとき，$\boldsymbol{a}_{i,k}$ は $\boldsymbol{h}(\theta_i, \omega_k)$ を平均とした複素正規分布より生成されると仮定する。しかし当然ながら到来方向 θ_i は実際には観測することができないため，これを潜在変数見なすことにすると，$\boldsymbol{a}_{i,k}$ の生成モデルは DOA を潜在変数とした混合モデルとなる。これを 3.1 節の生成モデルに組み込み，生成モデル全体のパラメータ推論を行うことは，パーミュテーション整合，各音源の DOA 推定，周波数ごとの音源分離を協調的に行うことに相当する [8]。

まず，$\vartheta_1, \ldots, \vartheta_D$(すべて定数) からなる D 個の DOA 候補の集合を用意する。例えば 180 度を D 等分した角度 $\vartheta_d = (d-1)\pi/D, (d = 1, \ldots, D)$ の集合を考える。各音源の DOA がこの DOA 候補値の中から決定されると仮定すると，音源 i の到来方向 θ_i が生成されるプロセスは以下のように記述できる。

$$c_i | \boldsymbol{\rho}_i \sim \mathrm{Categorical}(c_i; \boldsymbol{\rho}_i) \quad (5)$$

$$\theta_i = \vartheta_{c_i} \quad (6)$$

ただし，$\boldsymbol{y} = (y_1, \ldots, y_D)$，$\sum_d y_d = 1$ とすると，$\mathrm{Categorical}(x; \boldsymbol{y}) \propto y_x$ である。また，$\boldsymbol{\rho}_i = (\rho_{i,1}, \ldots, \rho_{i,D})$ である。$c_i \in \{1, \ldots, D\}$ は i 番目の音源にどの DOA 候補値が割り当てられるかを表すインジケータ変数であり，上式はこれが離散分布 (各確率値が $\rho_{i,1}, \ldots, \rho_{i,D}$) から生成されることを意味している。このプロセスにより各音源の DOA が決定され，伝達周波数特性 $\boldsymbol{a}_{i,k}$ は

$$\boldsymbol{a}_{i,k} | c_i \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{a}_{i,k}; \boldsymbol{h}(\vartheta_{c_i}, \omega_k), \Sigma_k^{(a)}) \quad (7)$$

により生成される。

3.3 DOA-HMM

音源が移動する場合，時刻ごとにステアリングベクトルが変化してしまうため，移動音源を扱えるようにするためには $\boldsymbol{a}_{i,k}$ を時刻 l に依存する変数 $\boldsymbol{a}_{i,k,l}$ に拡張する必要がある。このとき，式 (2) は

$$\boldsymbol{y}_{k,l} = \boldsymbol{a}_{z_{k,l},k,l} s_{k,l} + \boldsymbol{n}_{k,l} \quad (8)$$

と書き直せる。

ここで，3.2 節の自然な拡張として，各音源の DOA インデックス c_i を時刻 l に依存する変数 $c_{i,l}$ に拡張し，$c_{i,1}, \ldots, c_{i,L}$ を状態系列とした HMM によりステアリングベクトル系列 $\boldsymbol{a}_{i,k,1}, \ldots, \boldsymbol{a}_{i,k,L}$ を確率モデル化することを考える。このとき，音源 i の時刻 l における DOA$\theta_{i,l}$ の生成プロセスは，

$$c_{i,l} | c_{i,l-1} \sim \mathrm{Categorical}(c_{i,l}; \boldsymbol{\rho}_{c_{i,l-1}}) \quad (9)$$

$$\theta_{i,l} = \vartheta_{c_{i,l}} \quad (10)$$

と表せる。$\boldsymbol{\rho}_d = (\rho_{d,1}, \ldots, \rho_{d,D})$ は状態 d から状態 $1, \ldots, D$ への遷移確率を表し，$\rho_{d,d'}$ を要素とする $D \times D$ 行列 $\boldsymbol{\rho} = (\rho_{d,d'})_{D \times D}$ を遷移行列という。実際の移動音源は，十分短い時間の間に大きく到来方向を変える可能性は低いと考えられるので，隣接する状態への遷移確率を高めに設定すれば良い。

以上のステアリングベクトル系列の確率モデルを 3.1 節のモデル (の時変版) に組み込み，全体のパラメータ推論 (後述) を通してパーミュテーション整合，移動音源の追従，周波数ごとの音源分離を同時に行おうというのが提案手法の要点である。

4　変分推論アルゴリズム

観測信号 $\boldsymbol{Y} = \boldsymbol{y}_{1:K,1:L}$ が与えられたもとで，以上の生成モデルのパラメータ $\boldsymbol{A} = \boldsymbol{a}_{1:I,1:K,1:L}$，$\boldsymbol{S} = s_{1:K,1:L}$，$\boldsymbol{Z} = Z_{1:K,1:L}$，$\boldsymbol{C} = c_{1:K,1:L}$ の事後分布 $p(\boldsymbol{A}, \boldsymbol{S}, \boldsymbol{Z}, \boldsymbol{C}|\boldsymbol{Y})$ を求めたい。この事後分布を解析的に得ることは難しいが，変分推論法に基づき近似分布を反復計算により得ることができる。以下では，ρ, $\Sigma^{(n)}_{1:K}, \Sigma^{(a)}_{1:K}$ は実験的に定める定数とする。

変分推論は事後分布 $p(\boldsymbol{A}, \boldsymbol{S}, \boldsymbol{Z}, \boldsymbol{C}|\boldsymbol{Y})$ と，

$$\int \cdots \int q(\boldsymbol{A}, \boldsymbol{S}, \boldsymbol{Z}, \boldsymbol{C}) \mathrm{d}\boldsymbol{A} \cdots \mathrm{d}\boldsymbol{C} = 1 \tag{11}$$

を満たす非負の変関数 $q(\boldsymbol{A}, \boldsymbol{S}, \boldsymbol{Z}, \boldsymbol{C})$ との間の Kullback-Leibler ダイバージェンス

$$\mathcal{F}[q] = \left\langle \log \frac{p(\boldsymbol{A}, \boldsymbol{S}, \boldsymbol{Z}, \boldsymbol{C}|\boldsymbol{Y})}{q(\boldsymbol{A}, \boldsymbol{S}, \boldsymbol{Z}, \boldsymbol{C})} \right\rangle_{q(\boldsymbol{A}, \boldsymbol{S}, \boldsymbol{Z}, \boldsymbol{C})} \tag{12}$$

を q に関して最小化することが目的となる。ただし $\langle f(x) \rangle_{q(x)}$ は $\int q(x) f(x) dx$ を表す。無論，$\mathcal{F}[q]$ は $p = q$ のとき最小となるが，q に関して

$$q(\boldsymbol{A}, \boldsymbol{S}, \boldsymbol{Z}, \boldsymbol{C}) = q(\boldsymbol{A}) q(\boldsymbol{S}) q(\boldsymbol{Z}) q(\boldsymbol{C}) \tag{13}$$

となるような分布クラスを考え，$\mathcal{F}[q]$ を $q(\boldsymbol{A})$, $q(\boldsymbol{S})$, $q(\boldsymbol{Z})$, $q(\boldsymbol{C})$ について交互に最小化するステップを繰り返すことで，当該分布クラスの中で $p(\boldsymbol{A}, \boldsymbol{S}, \boldsymbol{Z}, \boldsymbol{C}|\boldsymbol{Y})$ を最も良く近似する分布を得ようというのが変分推論法の基本的な考え方である。

導出は省略するが，式 (12) を式 (11) の拘束の下で最小化する各 q は解析的に以下の形として求まる。

$$\hat{q}(\boldsymbol{A}) = \prod_{i,k,l} \mathcal{N}_{\mathbb{C}}(\boldsymbol{a}_{i,k,l}; m_{i,k,l}, \Gamma_{i,k,l}) \tag{14}$$

$$\hat{q}(\boldsymbol{S}) = \prod_{k,l} \mathcal{N}_{\mathbb{C}}(s_{k,l}; \mu_{k,l}, \sigma_{k,l}) \tag{15}$$

$$\hat{q}(\boldsymbol{Z}) = \prod_{k,l} \hat{q}(z_{k,l}), \hat{q}(z_{k,l} = i) = \phi_{i,k,l} \tag{16}$$

なお，以上の更新則は [8] と同様である。また得られた分布 $\hat{q}(\boldsymbol{A})$ によって $\boldsymbol{A}$ の期待値を計算し $\boldsymbol{\rho}$ を用いて Forward-Backward アルゴリズムを行うことで $\hat{q}(\boldsymbol{C})$ を求めることができる。

以上の変分推論アルゴリズムによって推定された $s_{k,l}$ の平均値 $\mu_{k,l}$ に確率値 $\phi_{k,l}$ を乗じることで，音源 i の推定信号を得ることができる。

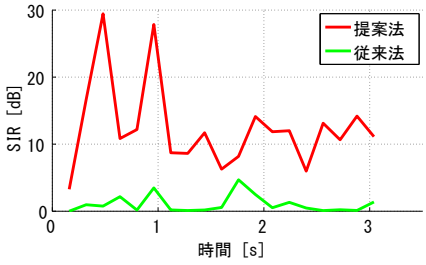

Fig. 1　提案法と従来法における移動音源 A に対する SIR の時間変化

5　複数移動音源の分離実験

提案法の有効性を示すため，移動音源に対して音源分離と到来方向推定性能の検証を行った。移動音源として移動音源データベース [9] の男性話者の音声信号 2 つを (移動音源 A，B)，固定音源として音声データベース [10] の女性話者の音声信号に室内インパルス応答を畳み込み加算したもの 1 つを用い，それらを人工的に混合したものを観測信号とした。残響時間は 0 ms である。移動音源を変えることで，10 通りの混合音声データセットを作成し，実験した。標本化周波数は 16 kHz とした。短時間フーリエ変換 (フレーム長は 64 ms，フレームシフトは 16 ms) により算出した。$\Sigma^{(n)}_k$ と $\Sigma^{(a)}_k$ はそれぞれ $\boldsymbol{I}$，$10^{1.5} \times \boldsymbol{I}$ とした。また角度の分割数は $M = 180$ とした。4 章の反復アルゴリズムの実行後，音源成分の推定値 $\mu_{k,n}$ に，音源 i が時間周波数点でどれだけアクティブらしいかを表す確率値 $\phi_{i,k,n}$ を乗じたものを，音源 i の推定時間周波数成分とした。音源分離性能の評価基準として，式 (17)〜(19) により導出される Signal-to-Interference-Ratio (SIR) [12] を用いた。SIR の計算には，3 つの音源のうち一番短い長さの音源が終了する約 3.1 s までを用いた。

$$\mathrm{SIR}_i[l] = \mathrm{OutputSIR}_i[l] - \mathrm{InputSIR}_i[l] \tag{17}$$

$$\mathrm{OutputSIR}_i[l] = 10 \log_{10} \frac{\sum_k \hat{s}_{i,k,l}}{\sum_{i' \neq i} \sum_k \hat{s}_{i',k,l}} [\mathrm{dB}] \tag{18}$$

$$\mathrm{InputSIR}_i[l] = 10 \log_{10} \frac{\sum_i \sum_k s_{i,k,l}}{\sum_{i' \neq i} \sum_k s_{i',k,l}} [\mathrm{dB}] \tag{19}$$

ただし $\hat{s}_{i,k,n}$ は音源 i の推定信号 $\phi_{i,k,n} \mu_{k,n}$ に含まれる音源 i の信号成分である。

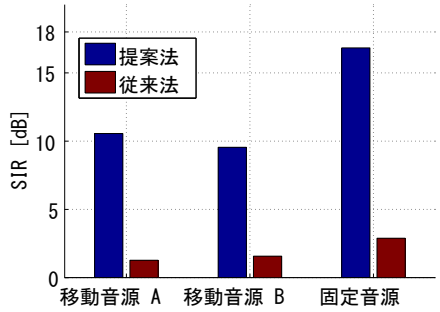

Fig. 2　提案法と従来法における音源ごとの SIR の平均値

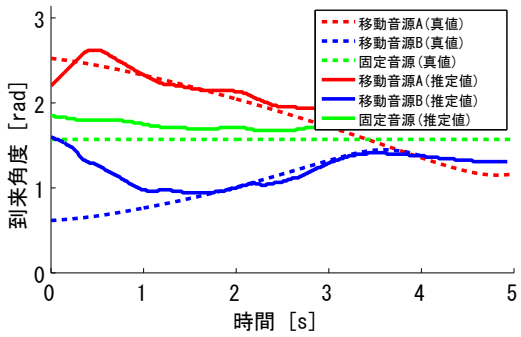

Fig. 3　各音源における到来角度の真値と推定値

また各時刻の到来角度の推定値には，推定された到来角度の確率分布から各時刻において最も確率値の高い角度を用いた。

さらに音源の移動を仮定しない従来法が，移動音源に対して良い分離性能をもたないことを示すため，[8] の手法を用いて同様の音源分離実験を行った場合の結果と提案法の結果を比較した。

Fig. 1 に提案法と従来法による，1 つの分離実験における移動音源 A の SIR の時間変化を示す。従来法では，音源分離がうまく行えておらず，SIR が多くの時刻で 0 に近い値である (input と output で SIR が改善されていない) のに対して，提案法では多くの時刻で SIR が改善されているのが見て取れる。Fig. 2 に各音源ごとにデータセットと時刻で平均をとった SIR の値を示す。3 つの音源すべてにおいて，従来法では SIR が低く音源分離が行えていないのに対して，提案法では SIR が 10 dB から 17 dB 程度の値を示しているのがわかる。3 つの音源における SIR の平均値は，従来法で 1.91 dB，提案法で 12.31 dB であった。

次に，1 つの分離実験における到来角度推定の結果を Fig. 3 に示す。実際の到来角度と比べて，1 s 付近から音源同士の到来角度が重なり，かつ音声の終了する 3 s 付近までは，おおむね正しく推定されていることが分かる。最初の約 1 s の間で到来方向推定の精度が良くないのは，生成モデルに組み込まれた，到来角度が急に変化しにくいという仮定により，音声の入っていない初期時刻付近のデータに対して推定された到来角度から滑らかにつなぐように到来角度が推定されてしまうからであると考えられる。固定音源の推定角度にバイアスがのっているのは，理想的なステアリングベクトルと実際のステアリングベクトルとの誤差からくる推定誤差である可能性だけでなく，[11] のデータベース作成時のマイクロフォンの角度誤差である可能性も考えられる。

6　おわりに

本稿では，音源が移動することで混合過程が変化する場合においても安定して動作する BSS アルゴリズムの実現を目指した。音声の時間周波数成分のスパース性に基づく周波数領域の劣決定 BSS モデルをベイズ的に記述し，音源の移動を，離散化した到来角度を状態とする隠れマルコフモデルとして表現し，短い時間において音源の到来角度が大きく変化する確率は小さいという仮定を遷移確率として観測信号の生成モデルに組み込み，混合 DOA モデルと組み合わせることで，音源分離と周波数ごと，時間ごとのパーミュテーション整合を同時に実現した。これにより，時間変化する到来角度の推定と音源分離を一挙に行えることが，提案法の主要な特徴である。

参考文献

[1] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001.

[2] Ö. Yılmaz & S. Rickard, *IEEE Trans. SP*, 52(7), pp. 1830–1847, 2004.

[3] Y. Mori *et al.*, in *Proc. IWAENC '05*, pp. 229–232, 2005.

[4] M. I. Mandel *et al.*, in *Adv. NIPS*, pp. 953–960, 2006.

[5] S. Araki *et al.*, *Signal Process.*, 87(8), pp. 1833–1847, 2007.

[6] 和泉他, 音講論 (春), 2-1-5, pp. 555–556, 2007.

[7] H. Sawada *et al.*, *IEEE Trans. ASLP*, 19(3), pp. 516–527, 2010.

[8] 亀岡他, 音講論 (春), 1-1-19, pp. 713–716, 2012.

[9] T. Otsuka *et al.*, in *Proc. AAAI-12* pp. 2038–2045, 2012.

[10] A. Kurematsu *et al.*, *Speech Communication*, pp. 357–363, 1990.

[11] S. Nakamura *et al.*, in *Proc. LREC '00*, pp. 965–968, 2000.

[12] E. Vincent *et al.*, *IEEE Trans. ASLP*, pp. 1462–1469, 2006.