残響環境下音声認識に対する残響除去とシステム統合手法の有効性 REVERB チャレンジ*

○太刀岡 勇気, 成田 知宏 (三菱電機・情報総研), 渡部 晋治, Felix Weninger(MERL)

1 序論

REVERB チャレンジは、最近導入されたタスクで、残響環境下での音声認識タスクが含まれる [1]。本報では REVERB チャレンジのために、構築した音声強調・認識手法の有効性を検証する。音声強調は、提案の残響除去法 [2] と、到来方向推定を行った [3, 4] 後に、ビームフォーミング (BF) により行う。

既報 [5] において、騒音環境下で識別学習 [6] や特徴量変換手法 [7, 8] が有効であることを示した。識別的手法は、学習条件と評価条件の近いマッチ条件で有効であるが、ミスマッチ条件で識別的手法が有効であるかは検討の余地がある。REVERB チャレンジでは、8 種の異なる環境があり、環境のミスマッチで、識別的手法が有効でない可能性があるので検討する。

特徴量変換に関しては、線形判別分析 (LDA) [7]、最尤線形変換 (MLLT) [8] と識別的特徴量変換 [6] がある。LDA は長いコンテキストを扱うので、特徴量の動的特徴をモデル化でき、残響の影響を低減できると考えられる。未知の条件に適応させるには、モデル適応が有効なので、話者適応学習 (SAT) [9] と基底特徴量空間最尤線形回帰 (basis fMLLR) [10] を用いた。合わせて深層回路網 (DNN) の有効性も検証する。

さらに複数のシステムの認識結果の統合が有効であることが知られている。環境ごとに最適な音声認識システムが異なる際には、ROVER [11] などにより結果を統合することで、性能を向上させられる。上述の種々のシステムに加えて、識別学習の枠組みに依拠した、意図的に構築した補助システムも検証する [12]。

2 提案システムの概観

Fig. 1 に、提案システムの概要図を示す。提案システムは、音声強調・特徴量変換・音声認識の 3 つの要素から構成されている。音声強調は、到来方向推定と多 ch 遅延和 BF、残響時間予測に基づく残響除去手法、正規化最小 2 乗誤差法 (NLMS) により短時間の歪を除去する手法より成る。特徴量変換では、メル周波数ケプストラム係数 (MFCC) と知覚的線形予測 (PLP) の 2 種の特徴量を用いることで、システム統合に使う補助システムが異なる傾向の仮説を出力することを期待できる。音声認識では、識別学習 (相互情報量最大化法 (MMI)) により、3 種の音響モデル (ガウス混合モデル (GMM)、部分空間 GMM(SGMM[13]) と DNN) を構築した。これに、識別的に学習された補助システムも加え、ROVER により結果を統合した。

3 音声強調部

3.1 CSP 法による到来方向推定に基づく遅延和 BF

音源からの直接音を強調するために、遅延和 BF を適用した。強調されたスペクトル $\tilde{\boldsymbol{y}}_t$ は、m 番目のマイクにより観測された短時間フーリエ変換 (STFT) によるスペクトル $\boldsymbol{x}_t(m)$ の和として得られる。

$$\tilde{\boldsymbol{y}}_t = \sum_m \boldsymbol{x}_t(m) \odot \exp(-\jmath\boldsymbol{\omega}\tau_{1,m}) \tag{1}$$

t は現在フレームの番号、$\odot$ は要素ごとの積、$\boldsymbol{\omega}$ は角周波数の組である。1 番目のマイク基準の m 番目のマイクの到達時間遅れ $\tau_{1,m}$ は、到来方向に関連しており、2 マイク間のクロスパワースペクトルよりクロススペクトル位相 (CSP) 分析 [3] で推定できる。

$$\tau_{1,m} = \arg\max \mathcal{S}^{-1}\left[\frac{\boldsymbol{x}_t(1) \odot \boldsymbol{x}_t(m)^*}{|\boldsymbol{x}_t(1)||\boldsymbol{x}_t(m)|}\right] \tag{2}$$

$\mathcal{S}$ は、STFT 演算であり、*は複素共役を表す。ピークホールド処理 [14] とノイズ成分引き去り (推定 SNR が 0dB 以下の場合に、当該のクロスパワースペクトルを 0 にする処理)[4] を行い、ペアごとに CSP 係数を同期加算 [15] することで、騒音の影響を軽減した。

3.2 残響時間推定に基づく 1ch 残響除去法 [2]

観測パワースペクトル $|\boldsymbol{x}|^2$ は、音源のパワースペクトル $|\hat{\boldsymbol{y}}|^2$ の重み付き和でモデル化でき、ノイズのパワースペクトル $|\boldsymbol{n}|^2$ が定常な場合、

$$|\boldsymbol{x}_t|^2 = \sum_{\mu=0}^{t} w_\mu |\hat{\boldsymbol{y}}_{t-\mu}|^2 + |\boldsymbol{n}|^2 \tag{3}$$

のようになる。μ と w_μ は、遅れフレームと重み係数で、$|\hat{\boldsymbol{y}}|^2$ は、$|\boldsymbol{x}|^2$ と関連付けられる。

$$|\hat{\boldsymbol{y}}_{t-\mu}|^2 = \eta(T_r)|\boldsymbol{x}_{t-\mu}|^2 - |\boldsymbol{n}|^2 \tag{4}$$

η は全エネルギーの内の直接音が占める割合であり、残響時間 T_r の減少関数である。w_0 は 1 と仮定すると、

$$|\hat{\boldsymbol{y}}_t|^2 = |\boldsymbol{x}_t|^2 - \sum_{\mu=1}^{t} w_\mu \left[\eta(T_r)|\boldsymbol{x}_{t-\mu}|^2 - |\boldsymbol{n}|^2\right] - |\boldsymbol{n}|^2 \tag{5}$$

が導かれる。残響は、閾値 D(フレーム) により、初期と後期の 2 つに分けられる。音声認識性能に悪影

*Effectiveness of dereverberation techniques and system combination approach for various reverberant environments: REVERB challenge, by TACHIOKA, Yuuki, NARITA, Tomohiro (Mitsubishi Electric Corporation), WATANABE, Shinji, WENINGER, Felix (Mitsubishi Electric Research Laboratories).

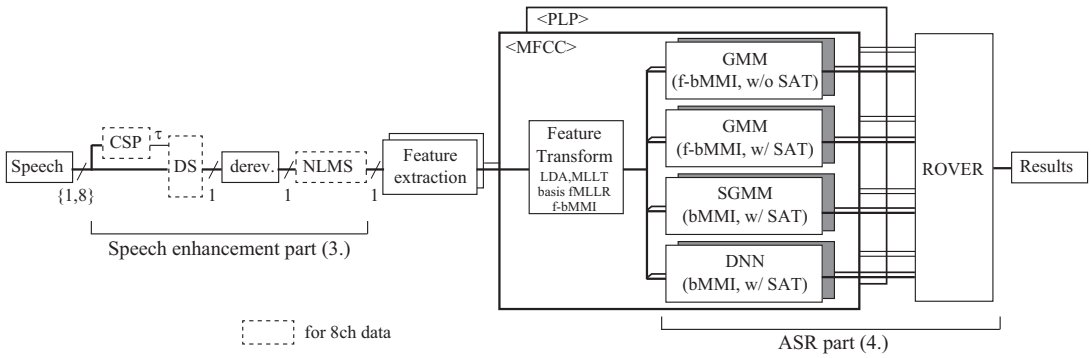

Fig. 1 Schematic diagram of the proposed system. (CSP: cross spectrum phase analysis, DS: delay-and-sum beamformer, derev.: proposed dereverberation method, NLMS: normalized least-mean-squares algorithm, gray blocks are complementary systems for each system type)

響を与えるのは主に後期で、初期は無視できる。後期では、音響エネルギー密度が指数的に減衰し、w_μ は

$$w_\mu = \begin{cases} 0 & (1 \le \mu \le D) \\ \frac{\alpha_s}{\eta(T_r)} e^{-2\Delta\varphi\mu} & (D < \mu) \end{cases} \quad (6)$$

のようにモデル化できる。φ はフレームシフト、α_s は引き去り係数である。上段・下段は各々、初期・後期残響に対応している。η が定数と仮定すると、式 (5) は、スペクトルサブトラクション (SS) と類似である。引き去られたパワースペクトル $|\hat{\boldsymbol{y}}|^2$ が $\beta|\boldsymbol{x}|^2$ より小さい時は、$\beta|\boldsymbol{x}|^2$ で置換する処理をフロアリングと呼ぶ。(β は定数) フロアリング率 r を、時間周波数ビンの内、フロアリングした数の比率とする。

ここで、適当な残響時間 T_a を仮定すると、r は T_a の単調増加関数となる。(線形な関係でモデル化し、傾きを Δ_r とする) また、T_a が同一の時、r は T_r の増加関数となる。実際の $\eta(T_r)$ は T_r とともに減少するので、η が定数して残響除去したパワースペクトルは、T_r が長いほどフロアリングしやすい。(2 つの定数 a と b を用いて、$T_r = a\Delta_r - b$ とする) よって、実際の残響時間 T_r は、フロアリング率 r を計算し、いくつか仮定した残響時間 T_a に対して最小 2 乗法で傾き Δ_r を算出することで推定される。

4 音声認識部

4.1 音響モデルの MMI 識別学習

MMI 識別学習は、正解ラベルと認識仮説の相互情報量を最大化する教師有り学習である。MMI 学習の改良版であるブーステッド MMI (bMMI) 学習 [6] では、音素正解率によって学習データの重みを変化させる目的で増幅係数 $b(\ge 0)$ を導入し、評価関数は

$$\mathcal{F}(\lambda) = \sum_r \ln \frac{p_\lambda\left(\boldsymbol{x}^r|\mathcal{H}_{s_r}\right)^\kappa p_L(s_r)}{\sum_s p_\lambda\left(\boldsymbol{x}^r|\mathcal{H}_s\right)^\kappa p_L(s)\mathrm{e}^{-bA(s,s_r)}} \quad (7)$$

のようにあらわされる。$\boldsymbol{x}^r$ は、r 番目の発話の特徴量系列である。音響モデルパラメータ λ は、拡張バウム・ウェルチ法により最適化される。$\mathcal{H}_{s_r}$ と $\mathcal{H}_s$ は、それぞれ、正解ラベル s_r と仮説 s に対する HMM の系列である。p_λ は音響モデル尤度、κ は音響スケール、p_L は言語モデル尤度であり、$A(s, s_r)$ は s の s_r に対する音素正解率である。本報では、GMM と SGMM の bMMI モデルの性能を最尤 (ML) モデルと比較する。

4.2 識別的特徴量変換

識別学習の特徴量変換への拡張は、特徴量空間識別学習 [6] と呼ばれる。この方法では、以下のように高次元な特徴量 $\boldsymbol{h}_t$ を低次元な特徴量空間に写像する行列 $I \times J$ の行列 $\boldsymbol{M}$ を、識別的基準により推定する。

$$\boldsymbol{y}_t = \boldsymbol{x}_t + \boldsymbol{M}\boldsymbol{h}_t \quad (8)$$

$\boldsymbol{x}_t$ は、t フレームにおける元の I 次元特徴量、$\boldsymbol{y}_t$ は変換された同じく I 次元の特徴量、$\boldsymbol{h}_t$ は $J(\gg I)$ 次元の補助的な特徴量である。通例、$\boldsymbol{h}_t$ としては、universal background model(UBM) のガウス事後確率を使うことが多い。評価関数は、式 (7) の $\boldsymbol{x}^r$ を、r 番目の発話の変換特徴量系列 $\boldsymbol{y}^r$ で置き換えることで得られ、行列 $\boldsymbol{M}$ は、評価関数を最大化するように最適化される。ここでは、特徴量空間ブーステッド MMI (f-bMMI) の有効性に関して検討する。

4.3 DNN の識別学習

DNN-HMM のシステムにおいて、通常のクロスエントロピー (CE) 学習に加えて、(b)MMI 基準 (7) に基づく系列の識別学習法 [16] が提案されている。DNN は HMM 状態 j に対する事後確率を出力する。音響モデル尤度 p_θ は、疑似的な尤度

$$p_\theta\left(\boldsymbol{x}^r|j\right) = \frac{p_\theta\left(j|\boldsymbol{x}^r\right)}{p_0\left(j\right)} \quad (9)$$

によって置き換えられる。$p_0\left(j\right)$ は状態 j に対する事前確率である。それぞれの HMM の状態に対して、モデル θ は soft-max の活性化関数

$$p_\theta(j|\boldsymbol{x}^r) = \frac{\exp a_j(\boldsymbol{x}^r)}{\sum_{j'} \exp a_{j'}(\boldsymbol{x}^r)} \quad (10)$$

を含む。a_j は、出力層の j 番目のユニットの活性化関数であり、これらは、bMMI 基準に基づき識別的に学習される。bMMI の評価関数は、単純に λ を θ を置き換えるだけで、式 (7) と全く同一である。

4.4 システム統合のための補助システムを構築するための一般的な枠組み [12]

補助システムを意図的に構築する枠組みについて述べる。補助システムは、初期モデルから始めて識別学習により構築される。Q 個の元となるシステムが既に構築されているときに、識別学習の評価関数 $\mathcal{F}$ は、識別学習の原理を拡張すると、

$$\mathcal{F}^c(\varphi) = (1+\alpha_c)\mathcal{F}(\varphi) - \frac{\alpha_c}{Q}\sum_{q=1}^{Q}\mathcal{F}(\varphi) \qquad (11)$$

のように、一般化できる。これは、正解ラベル s_r に関連する元の評価関数から、q 番目の元のシステムの 1 位の仮説 $s_{q,1}$ に関連する評価関数を引き去ったものである。φ は最適化されるべき補助システムのモデルパラメータの組 (例: λ、$\boldsymbol{M}$ や θ)、α_c はスケーリング係数である。識別的基準 $\mathcal{F}$ には、bMMI や f-bMMI が選択できる。もし α_c が零ならば、この評価関数は元々の $\mathcal{F}$ に一致する。式 (11) の第 1 項は、識別学習の基準に従って性能を向上させる効果が、第 2 項は今構築しようとしているシステムが、元のシステムと異なる傾向の仮説を出力するようにさせる効果がある。この手順は、4.1〜4.3 節のいずれにも適用できる。

5 実験

5.1 音声認識タスク

発話内容は Wall Street Journal で (WSJCAM0)、以下の 2 種がある。“SimData” は、残響時間が異なる 3 室 (Room 1〜3) で、音源・マイク間距離が 0.5 m (near)、2 m (far) の計 6 つの室内伝達関数を畳み込み、騒音を重畳したデータである。“RealData” は、比較的定常的な騒音が存在する室の実測データである。8ch マイクが半径 0.1 m の円状に配されている。学習セット (**tr**)、開発セット (**dev**)、評価セット (**eva**) が提供され、音響モデルは **tr** により学習し、言語モデル重みといったパラメータは、**dev** の単語誤り率 (WER) で調整した。語彙は 5 k で、tri-gram 言語モデルを使った。全て「発話単位の一括処理」である。

5.2 音声強調

チャレンジでは、1、2、8ch のデータが提供されているが、1ch と 8ch の場合について検討した。1ch の場合には、残響除去法のみを用いた。パラメータは、($D = 9$, $\alpha = 5$, $\beta = 0.05$, $a = 0.005$, $b = 0.6$) のように設定した。8ch の場合には、残響除去前に、全ペアのマイクを用いて推定した到来方向情報に基づく遅延和 BF を行った。残響除去後に、200 タップの NLMS により、短時間の歪を除去した。

5.3 特徴量抽出および特徴量変換、話者適応

音響特徴量と特徴量変換の設定に関して述べる [5]。音響特徴量は、13 次元の MFCC、PLP とその動的特徴量 (Δ, $\Delta\Delta$) である。9 連続フレームの静的 MFCC を結合した 117 次元の特徴量を、LDA を用いて 40 次元に圧縮した。LDA のクラスは、HMM の状態 (2,500 状態) とした。これに加え、MLLT を用いた。

音響モデル適応には、適応の速い基底 fMLLR [10] を用いた。さらに、話者間の多様性に対処するために、SAT [9] による音響モデル学習を行った。

Table 1 Average WER [%] on the REVERB Challenge (**dev**) using single channel data. (MFCC)

	Feature	Type	Sim Avg	Real Avg
baseline	MFCC	ML	22.05	47.95
derev.			19.75	45.81
GMM	+LDA+MLLT	ML	15.87	40.33
	+basis fMLLR		13.79	34.27
		f-bMMI	11.18	**29.82**
		f-bMMI$_c$	11.71	30.37
	+SAT	ML	14.11	36.15
		f-bMMI	10.15	33.18
		f-bMMI$_c$	10.32	33.83
SGMM		ML	11.80	34.03
		bMMI	9.91	32.43
		bMMI$_c$	10.01	32.01
DNN		CE	11.34	33.35
		bMMI	9.25	30.56
		bMMI$_c$	**8.92**	30.91

Table 2 Average WER [%] on **dev** using eight channel data.

	Feature	Type	Sim Avg	Real Avg
BF+derev.	MFCC	ML	13.77	41.60
+NLMS			14.40	39.56
GMM	+LDA+MLLT	ML	11.83	35.62
	+basis fMLLR		10.57	28.77
		f-bMMI	7.99	**24.42**
		f-bMMI$_c$	8.03	25.06
	+SAT	ML	10.03	30.88
		f-bMMI	7.01	26.94
		f-bMMI$_c$	7.22	27.00
SGMM		ML	8.36	27.81
		bMMI	**7.00**	27.36
		bMMI$_c$	7.11	27.20
DNN		CE	8.74	27.30
		bMMI	7.25	26.06
		bMMI$_c$	7.05	25.58

5.4 識別的手法と音響モデル

識別的特徴量変換 (4.2 節) では、400 ガウス分布が使われ、40 次元のオフセット特徴量を連続 9 フレームでコンテキスト拡張したものから計算される。(f-)bMMI における増幅係数は 0.1 とした。補助システムを構築するためのパラメータは、式 (11) の第 2 項に付加される増幅係数が 0.3、α_c は 0.75 とした。

DNN は、Kaldi [17] の Povey の実装により学習した。隠れ層 2 層でパラメータの総数は 2,000,000 である。実験のパラメータは、基本的に Kaldi に付属の WSJ のチュートリアル (s6) のものを流用した。

GMM システムには、f-bMMI を、SGMM と DNN[16] に関しては bMMI を用いた。各々のシステムに対して、組となる補助システムを提案法により構築した。これらは、MFCC と PLP 特徴量、双方に対して構築されているので、構築したシステムの総数は 16 となる。

Table 3 WER [%] on the REVERB Challenge (**eva**). All systems except ROVER are single systems. MFCC feature was used for single system; MFCC and PLP features were used for ROVER.

		SimData							RealData		
		Room 1		Room 2		Room 3		Avg	Room 1		Avg
		near	far	near	far	near	far		near	far	
1ch	Kaldi baseline	13.23	14.13	15.54	29.69	20.06	37.44	21.68	50.62	45.98	48.30
	derev.	12.50	13.43	14.61	24.71	17.09	32.62	19.16	44.75	43.32	44.04
	GMM+f-bMMI	7.27	8.17	8.82	14.11	10.54	18.76	11.28	28.65	29.54	29.10
	SAT-GMM+f-bMMI	6.44	7.22	7.57	13.97	9.52	18.44	10.53	28.87	29.78	29.33
	SGMM+bMMI	5.81	6.54	7.22	13.84	8.70	18.17	10.05	27.75	28.36	28.06
	DNN+bMMI	5.90	6.84	7.35	12.57	9.40	16.55	**9.77**	25.97	25.69	**25.83**
	ROVER	**5.30**	**5.61**	**6.30**	**11.16**	**7.76**	**14.95**	**8.51**	**23.79**	**23.60**	**23.70**
8ch	CSP+BF+derev.	10.94	11.69	10.98	16.33	12.79	21.39	14.02	34.33	36.93	35.63
	+NLMS	10.94	12.32	11.38	17.59	13.46	22.96	14.78	35.32	35.28	35.30
	GMM+f-bMMI	6.57	6.93	6.80	9.93	7.47	12.76	8.41	20.22	23.19	21.71
	SAT-GMM+f-bMMI	6.17	6.64	6.51	10.13	7.40	13.15	8.33	20.63	23.67	22.15
	SGMM+bMMI	5.86	6.44	6.29	9.23	6.96	12.83	7.94	20.66	23.50	22.08
	DNN+bMMI	5.64	6.18	6.16	9.29	7.08	12.40	**7.79**	19.35	22.28	**20.82**
	ROVER	**4.96**	5.62	**5.58**	8.18	**5.73**	**10.47**	**6.76**	**16.90**	**20.29**	**18.60**

6 結果と考察

6.1 ベースラインと音声強調手法

Table 1(1ch) と 2(8ch) は、開発セット (**dev**) の平均の WER である。Table 1 の “Kaldi baseline” は、残響音声より学習された音響モデルで、音声強調手法なしの場合の WER である。“derev.” は、提案の残響除去法である。残響の短い Room 1 では、残響除去法が効果的でない場合も見られたが、他の場合や平均では、性能が向上した。8ch の場合 (Table 2) は、BF と “derev.” の併用で、認識性能が大幅に改善した。“NLMS” は、環境ごとに効果に参差が見られたが、悪影響の方が少なかったので採用した。

上記の結果は、MFCC 特徴量を用いた場合であり、PLP 特徴量を使った場合は、これよりも若干性能が低かった。しかし、それらの誤り傾向は相当異なっていたので、システム統合に組み入れた。

6.2 特徴量変換と識別学習、SGMM と DNN

LDA と MLLT による特徴量変換により、WER は大幅に改善した。さらに、識別学習が有効に機能していることがわかる。提案の補助システムの性能は、元のシステムの性能よりも若干低い程度なので、システム統合に適っている。SGMM による音響モデルは、SimData では GMM の結果を上回ったが、RealData では GMM よりも性能が低かった。DNN モデルは、SimData あるいは全体の平均で、最良の性能を得た。DNN における系列の識別学習は、他手法同様に有効であった。

6.3 評価セット (eva) とシステム統合

Table 3 には、評価セットの結果を示す。識別学習した DNN は、単一のシステムの中では、最良の性能を得た。これは DNN の未知条件に対する頑健性を示すものといえる。さらに、システム統合により WER が、それぞれ SimData と RealData に対して、1.26%、2.13%(1ch)、1.03%、2.22%(8ch) 改善した。

7 結論

音声強調および特徴量変換と識別学習が、残響音声認識に有効であることを示した。またシステム統合手法により、環境の多様性に対する頑健性が向上した。さらに提案のシステム統合法により、性能が向上した。

参考文献

[1] K. Kinoshita *et al.*, “The REVERB Challenge: A common evaluation framework for dereverberation and recognition of reverberant speech,” WASPAA, (2013).

[2] Y. Tachioka, T. Hanazawa, and T. Iwasaki, “Dereverberation method with reverberation time estimation using floored ratio of spectral subtraction,” *Acoust Sci & Tech*, **34**, 212–215 (2013).

[3] C. Knapp and G. Carter, “The generalized correlation method for estimation of time delay,” *IEEE Trans on ASSP*, **24**, 320–327, (1976).

[4] Y. Tachioka, T. Narita, and T. Iwasaki, “Direction of arrival estimation by cross-power spectrum phase analysis using prior distributions and voice activity detection information,” *Acoust Sci & Tech*, **33**, 68–71, (2012).

[5] Y. Tachioka et al., “Discriminative methods for noise robust speech recognition: A CHiME challenge benchmark,” 2nd CHiME Workshop, pp.19–24, (2013).

[6] D. Povey *et al.*, “Boosted MMI for model and feature-space discriminative training,” ICASSP, pp.4057–4060 (2008).

[7] R. Haeb-Umbach and H. Ney, “Linear discriminant analysis for improved large vocabulary continuous speech recognition,” ICASSP, pp.13–16 (1992).

[8] R. Gopinath, “Maximum likelihood modeling with Gaussian distributions for classification,” ICASSP, pp.661–664 (1998).

[9] T. Anastasakos *et al.*, “A compact model for speaker-adaptive training,” ICSLP, pp.1137–1140 (1996).

[10] D. Povey and K. Yao, “A basis representation of constrained MLLR transforms for robust adaptation,” *Computer Speech and Language*, **26**, 35–51 (2012).

[11] J. Fiscus, “A post-processing system to yield reduced error word rates: Recognizer output voting error reduction (ROVER),” ASRU, pp.347–354, (1997).

[12] Y. Tachioka *et al.*, “A generalized framework of discriminative training for system combination,” ASRU, (2013).

[13] D. Povey *et al.*, “The subspace Gaussian mixture model –A structured model for speech recognition,” *Computer Speech and Language*, **25**, 404–439, (2011).

[14] T. Suzuki and Y. Kaneda, “Sound source direction estimation based on subband peak-hold processing,” *The Journal of the Acoust Soc of Jpn*, **65**, 513–522, (2009).

[15] T. Nishiura *et al.*, “Localization of multiple sound sources based on a CSP analysis with a microphone array,” ICASSP, vol.2, **2**, pp.1053–1056 (2000).

[16] K. Veselý *et al.*, “Sequence-discriminative training of deep neural networks,” INTERSPEECH, (2013).

[17] D. Povey *et al.*, “The Kaldi speech recognition toolkit,” ASRU, pp.1–4 (2011).