Vol.2023-SLP-146 No.12
2023/2/28

実環境下日本語話し言葉音声コーパスの構築と音声認識ベンチマーク

三村 正人[1]　井上 昂治[1]　河原 達也[1]　中村 友彦[2]　猿渡 洋[2]

概要：実環境下において低遅延かつ高精度で動作する話し言葉のための遠隔音声認識システムは、対話ロボットとの円滑なコミュニケーションを実現する上で必須の技術である。本研究では、多数の雑音源が存在する４つのロケーションにおいて様々なトピックについてのプレゼンテーション音声を遠隔マイクで収録し、遠隔話し言葉音声認識の評価を行うためのコーパスを構築する。また、種々の音声強調・音声発話区分化・音声認識手法を用いた本コーパスの音声認識ベンチマーク結果について報告する。音声強調については、特に未知の環境下で頑健に動作する教師なし音源分離に基づく手法に焦点を当てる。既存の音声・雑音データセット上で学習したオンライン音声発話区分化およびバックエンド音声認識モデルを用いた音声認識実験において、オフライン音声強調で平均文字誤り率 15.0%、ストリーミング音声強調で 16.2%の音声認識精度を達成した。

1. はじめに

本研究では、対話ロボットの現実的なユースケースに則した条件下において、実用に耐える音声強調、区分化及び遠隔音声認識システムの開発を行っている。対話ロボットは展示会場やショッピングモールなど強いバブルノイズや BGM が存在する環境で低遅延かつ頑健に動作する必要がある一方、受付や案内など典型的なシナリオでは対話ロボットとやり取りを行うユーザは通常に一名であり、ユーザは自発的ではあるが明瞭な話し言葉で質問やリクエストを行うことが想定される。

既存の遠隔音声認識ベンチマーク用のデータセット（表 1）では、読み上げスタイルの入力か、他の話者からの重複音声が主要な雑音源であるような会話音声が主である。読み上げ音声の認識は残響・雑音が存在する条件でもすでに非常に高い精度で実現可能である一方 [1]、実環境下の会話音声はオーバーラップ音声の分離が困難であることに加えて、バックエンド音声認識モデルの学習のための書き起こしつきデータが十分に存在しないことから、実用的な水準には至っていない [2]。

本研究では、言語的には既存の話し言葉コーパスでカバーできる独話音声、具体的にはプレゼンテーションスタイルの音声を様々な雑音源が存在するストリートや博物館ホールのような実環境下で収録し、上記のロボット対話タスクの評価を目的とした新たなコーパスを構築する。本コーパス (”noisy-CSJ”と呼ぶ) の位置づけを図 1 に示す。

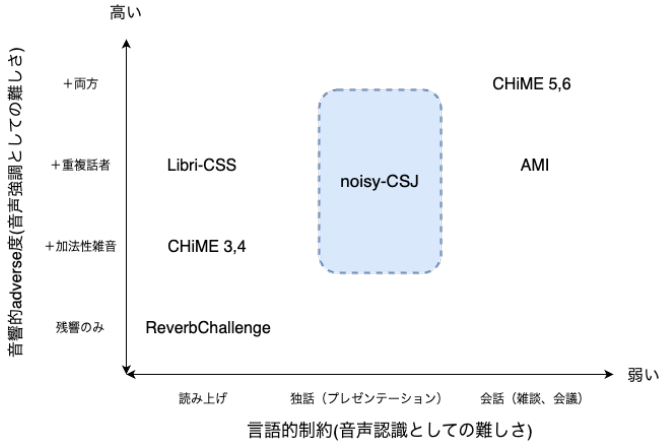


図 1: 本コーパスの位置づけ

2. コーパスの概要

これまでに４箇所の異なるロケーションにおいて、男性 17 名、女性 3 名の計 20 名の話者による 67 セッションの音声データを収録した。音声収録には、4 チャネルマイクアレイ（Seeed Studio 社製 ReSpeaker USB Mic Array）を用いた。すべてのセッションを通じて、マイクアレイと話者の距離は 1 メートルとした。また、話者の方向 (direction

[1] 京都大学 情報学研究科
Kyoto University, Sakyo-ku, Kyoto 606–8501, Japan
[2] 東京大学 情報理工学系研究科
The University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo 113–0033, Japan

表 1: 遠隔音声認識システムの学習・評価のための代表的データセット

コーパス名	発話スタイル	収録場所	実環境下データを含む	連続音声	主な雑音源
AMI [3]	会話	室内	√	√	重複話者、残響
CHiME 3, 4 [4]	読み上げ	バス、路上、カフェ	√		雑踏、交通雑音、バブル、残響
CHiME 5, 6 [2]	会話	室内	√	√	生活雑音、重複話者、残響
Reverb Challenge [5]	読み上げ	室内	√		残響
Libri-CSS [6]	読み上げ	室内		√	重複話者、残響
SMS-WSJ [7]	読み上げ	室内			重複話者、残響

表 2: 各ロケーションにおける収録音声の概要

				データ量 (時間)	
ロケーション	トピック	主な雑音源	セッション数	読み上げ	話し言葉
ポスター発表会場	研究紹介	バブル、残響	7	0.07	0.24
博物館ホール	展示内容の紹介	バブル、BGM、残響	20	0.19	0.63
ストリート	大学生活について	交通雑音、バブル	20	0.20	0.64
カフェテリア	大学周辺の食事処について	バブル、残響	20	0.19	0.61
計			67 (異なり話者数 20)	0.66	2.12

of arrival = DOA) は、マイクアレイの正面に固定した。各セッションは、ロケーションによって異なるトピックについてのプレゼンテーション音声と、JNAS 評価セットから選択した 5 文の読み上げ音声から構成される。プレゼンテーション音声の収録では、話者とアイクロフォンアレイをはさんで正対する位置に聞き手を配し、必要に応じて頷き等のリアクションを返すことにより円滑な発話を促した。

各ロケーションにおける収録音声の概要を表 2 に示す。また、実際の各ロケーションにおけるデータ収録の様子を図 2 に示す。(d) ストリート以外のロケーションにおける主な雑音源は、離れた位置の背景話者によるバブルノイズであったが、目的話者以外の音声や笑い声も強いエネルギーで重複することが頻繁に見られた。また、(b) カフェテリアと (c) 博物館ホールでは残響も主な雑音源となった。(b) カフェテリアおよび (c) 博物館ホールでは、他の収録参加者がマイクアレイから 2〜3 メートルの距離で雑談することで方向性の雑音源となるようにした。また (c) 博物館ホールでは、ショッピングモールなどのバブルノイズと BGM を模擬するために、話者の後方に設置した 2 つのスピーカにより DEMAND コーパス [8] の cafeteria 雑音および著作権フリーの音楽を再生した。(d) ストリートでは、車道を移動するバス、乗用車、二輪車等による非常に大きな交通雑音が主であった。

3. 音声認識ベンチマーク

3.1 比較手法

本研究では、遠隔音声認識システムを独立した 3 つのコンポーネント、すなわちフロントエンド音声強調、音声発話区分化、バックエンド音声認識のカスケード接続として構成した。その上で、それぞれのコンポーネントにおいて、以下のように複数の手法を比較した。

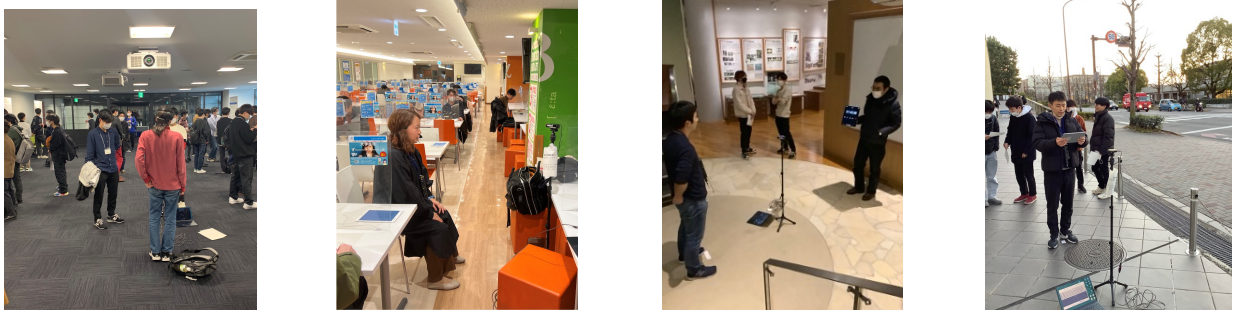
3.1.1 フロントエンド音声強調

残響除去

オフライン版の WPE(= weighted prediction error) 法 [9] によりマルチチャネル残響除去を行った。実装は Nara-WPE (`https://github.com/fgnt/nara_wpe`) を用いた。

ビームフォーマ

教師なしブラインドビームフォーマとして BeamformIt(`https://github.com/xanguera/BeamformIt`) を用いた。また、教師あり音声強調としてニューラルマスク推定に基づく MVDR および GEV ビームフォーマを実装し、評価した。マスク推定モデルは、クリーンな CSJ と、CSJ に残響・雑音を重畳したマルチコンディションデータ (”マルチコンディション CSJ”と呼ぶ) をペアデータとして、観測音から ideal binary mask (IBM) を予測する LSTM6 層からなるリカレントニューラルネットワークとして学習した。マルチコンディション CSJ は、CSJ のクリーン発話に Python の Pyroomacoustics ライブラリ (`https://github.com/LCAV/pyroomacoustics`) により生成した多様なインパルス応答を畳み込んだ上で、DEMAND [8]、MUSAN [10]、CHiME4 [4] の各データセットの雑音データを 0dB〜10dB の一様分布からサンプリングした SNR により加算することで構築した。推定したマスクを用いて音声・雑音の空間共分散行列を計算し、GEV および MVDR ビームフォーマの線形フィルタを求めた。



(a) ポスター発表会場　(b) カフェテリア　(c) 博物館ホール　(d) ストリート

図 2: 各ロケーションにおけるデータ収録の様子

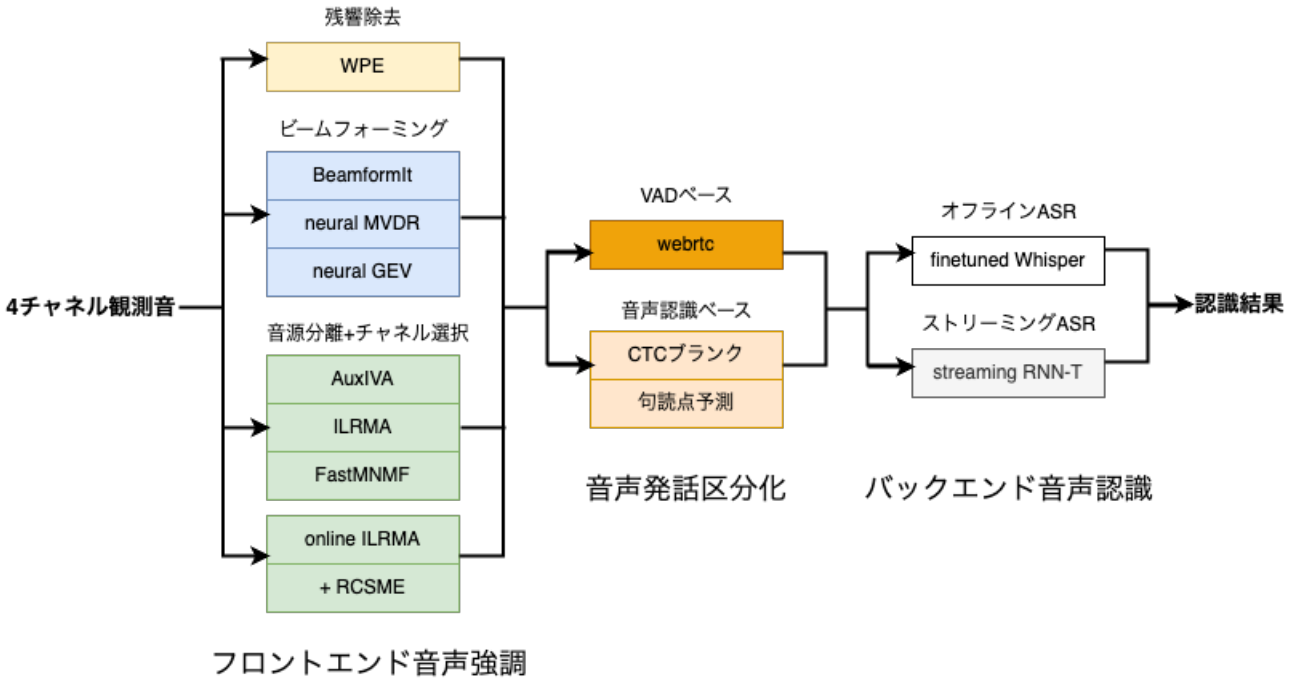


図 3: 音声認識ベンチマークにおける遠隔音声認識システムの構成

オフライン教師なし音源分離

AuxIVA [11]、ILRMA [12] および FastMNMF [13] の 3 つの手法による音源分離を行い、パワー最大のチャネルを抽出することで教師なし音声強調を行った。これらの音源分離には、Pyroomacoustics ライブラリを用いた。

ストリーミング教師なし音源分離

ストリーミング遠隔音声認識の実現には、フロントエンド音声強調も低遅延のオンライン手法で行う必要がある。ここでは、オンライン版の ILRMA と、ILRMA の分離音に基づいて拡散性雑音抑制を行う音声強調手法 RCSME(= rank-constrained spatial covariance matrix estimation) [14] を評価した。すべての音声認識実験を通じて、文献 [14] の式 (13) において、アルゴリズムのハイパーパラメータ α を 0.3 とした。

3.1.2 音声発話区分化

システムへの入力はマイクアレイからの連続した音声ストリームまたはその強調音声であるため、バックエンドにおいて音声認識を行うために、適切な長さのセグメントに分割する必要がある。VAD(voice activity detection) に基づく区分化と、音声認識モデルに基づく手法を比較する。VAD による区分化は、webrtcvad ツールキット (`https://github.com/wiseman/py-webrtcvad`) を用いた。音声認識モデルに基づく手法では、CTC ブランクがしきい値以上連続する箇所で分割する手法 [15] と、より意味的・統語的にまとまった区間へ分割することを指向した句読点検出に基づく手法 [16] を比較する。これらの手法は、自己回帰的エンコーダを持つストリーミング CTC 音声認識モデルとして上記のマルチコンディション CSJ を用いて構築した。また、句読点つきの疑似ラベルを Whisper punctuator ツールキット (`https://github.com/jumon/whisper-punctuator`) を用いて生成した。エンコーダは 16 層のカーネルサイズ 15、

表 3: 音声強調手法の比較。音声発話区分化は人手分割により行い、音声認識には finetuned Whisper を用いた (文字誤り率 (%))

			ロケーション				
音声強調手法	教師あり	ストリーミング可	ポスター	カフェテリア	博物館	ストリート	平均
観測音		√	25.2	29.1	35.0	22.9	28.6
offline WPE			25.9	27.9	31.4	22.5	27.1
BeamformIt			23.1	24.3	25.4	20.9	23.5
neural MVDR	√		15.1	19.5	20.1	14.4	17.6
neural GEV	√		10.3	19.5	22.9	15.5	18.3
AuxIVA			10.7	14.7	16.4	13.5	14.4
ILRMA			10.2	14.7	16.3	13.3	14.3
FastMNMF			10.5	15.0	16.6	13.6	14.5
online ILRMA (DOA-informed)		√	13.3	16.4	19.1	16.1	16.7
+ RCSCME		√	12.5	15.3	18.6	16.8	16.4

表 4: 音声発話区分化手法の比較。音声認識は finetuned Whisper を用いた (文字誤り率 (%))

入力	webrtcvad	CTC(blank のみ)	CTC(blank と句読点)	人手
ヘッドセット	28.9	12.9	**12.0**	-
マイクアレイ観測音	37.6	34.0	**27.1**	33.8
ILRMA	18.6	17.0	**14.3**	14.3
online ILRMA + RCSCME	23.5	20.6	**16.2**	16.4

隠れノード数 256、FNN のノード数 256 の因果的 Conformer [17] で構成した。

3.1.3 バックエンド音声認識

フルコンテクスト sequence-to-sequence モデル

低遅延性より認識精度を重視したオフラインの音声認識実験では、OpenAI が公開する Whisper [18] を用いた。Whisper は 68 万時間の弱教師ありデータを用いて学習された Transformer に基づく sequence-to-sequence モデルの一種であり、zeroshot でも様々な音声認識タスクにおいて優れた性能を示すことが知られている。本研究では、zeroshot の Whisper とともに、上記の日本語マルチコンディションデータで finetune したモデルも評価した。

ストリーミング RNN-Transducer モデル

デコード速度を重視したストリーミング音声認識では、日本語マルチコンディションデータでスクラッチ学習した RNN-T [19] 型のモデルを用いた。RNN-T のエンコーダは CTC 音声発話区分化モデルと同じパラメータ数の因果的 Conformer を用いた。また、プレディクターは 2 層の単方向 LSTM を用いて実装した。

3.2 フロントエンド音声強調手法の比較

まず、人手による音声発話区分化とマルチコンディション CSJ で finetune した Whisper を用いて、音声強調手法の音声認識性能を比較した。結果を表 3 に示す。

offline WPE を用いた残響除去では、残響が強いカフェテリアと博物館で観測音よりやや精度が改善した。一方、教師なしビームフォーマ (BeamformIt) では、すべてのロケーションで誤り率が低下し、平均で相対 17.8%の大きな改善が得られた。さらに、マルチコンディション CSJ を用いて学習した neural MVDR では、観測音に比べて相対 31.4%、BeamformIt に比べても相対 25.1%誤り率が低下した。このことから、音声・雑音ともに未知の環境下でも、教師ありマスク推定に基づくビームフォーマがある程度効果的であることがわかる。

しかし、教師なし音源分離手法に基づく音声強調では（AuxIVA、ILRMA、FastMNMF）、さらに大きな改善が見られた。ILRMA に基づく音声強調では、教師ありビームフォーマより相対で 18.8%誤り率が低下した。このことから、学習・評価時の音響的ミスマッチの影響を受けない教師なし音源分離手法は、実環境下で非常に頑健に動作することがわかる。ただし、認識精度の面でこれらの手法間に大きな差は見られなかった。

オンラインの ILRMA では、オフライン手法に比べて相対 14.3%、音声認識性能が低下したが、教師ありビームフォーマより低い誤り率を維持する結果となった。後段の RCSCME による音声強調では、交通雑音が主であるストリートで認識性能の改善が見られず、平均では 1.7%の改善に留まった。

表 5: オフライン音声認識モデル (Whisper) の finetuning の効果。音声発話区分化は人手分割により行った (文字誤り率 (%))

	ポスター		カフェテリア		博物館		ストリート		平均	
finetuning	なし	あり	なし	あり	なし	あり	なし	あり	なし	あり
観測音	31.7	**25.2**	35.9	**29.1**	41.1	**35.0**	31.4	**22.9**	35.7	**28.6**
ILRMA	23.5	**10.2**	25.0	**14.7**	26.9	**16.3**	24.1	**13.3**	25.1	**14.3**
online RCSCME	23.9	**12.5**	25.9	**15.3**	27.9	**18.6**	26.6	**16.8**	26.5	**16.4**

表 6: 読み上げ・話し言葉音声の比較。音声発話区分化は句読点予測に基づくモデルを用い、バックエンド音声認識は finetuned Whisper を用いた (文字誤り率 (%))

	ポスター		カフェテリア		博物館		ストリート		平均	
	読み上げ	自発	読み上げ	自発	読み上げ	自発	読み上げ	自発	読み上げ	自発
ヘッドセット	9.7	7.5	6.4	11.8	13.1	14.6	7.8	11.2	9.2	12.0
ILRMA	24.0	10.2	24.5	14.7	35.6	16.9	25.8	12.6	28.1	14.3
online RCSCME	39.1	13.1	28.0	15.8	33.8	18.5	27.2	15.4	30.7	16.2

3.3 音声発話区分化手法の比較

次に、VAD および音声認識に基づく音声発話区分化手法の比較を行った。バックエンド音声認識は finetuned Whisper を用いた。結果を表 4 に示す。

VAD のみに基づく webrtcvad による音声発話区分化は、人手による分割に比べてどの入力に対しても顕著に低い認識精度となった。一方、ストリーミング CTC 音声認識の連続ブランクを手がかりとした区分化では、webrtcvad に比べてヘッドセットで相対 55.5%、RCSCME で相対 12.3% と大幅に精度が改善した。句読点予測に基づく手法では、さらにすべての入力に対して性能が向上し、人手分割と同等の誤り率を達成した。CTC ブランクのみを用いた手法とは、特に観測音やオンライン音声強調で差が大きかった。これらの結果から、実環境下ではショートポーズだけでなく言語的な手がかりを用いた分割が特に効果的であることがわかる。

3.4 バックエンド音声認識手法の比較

最後に、バックエンド音声認識手法の比較を行う。まず、オフラインモデルにおけるマルチコンディションデータによる finetuning の効果を見るために、zeroshot Whisper と finetuned Whisper を比較した。結果を表 5 に示す。

Whisper を finetuning することにより、すべてのロケーション、すべての入力に対して顕著に性能が向上し、例えばオンライン RCSCME では、平均で相対 38.1%誤り率が低下した。ただし、zerhoshot Whisper は字幕スタイルの弱教師ありラベルで学習されているため、忠実な書き起こしの正解と異なるスタイルのテキストを出力する傾向があり、フィラーの有無などのミスマッチも誤り率に含まれる。

表 7 に、オフライン音声認識（finetuned Whisper）と因果的 Conformer を用いたストリーミング RNN-Transducer の認識結果の比較を示す。マルチコンディション CSJ のみを用いて学習したストリーミングモデルは、フルコンテクストモデルに比べて大幅に低い認識性能となり、特に観測音やオンライン手法を用いた強調音声では差が大きかった。YouTube 等の実データを用いた学習データの拡張や、オフラインモデルからの知識蒸留、遅延最小化学習などの学習アルゴリズム上の改善が課題である。

表 7: ストリーミング音声認識の評価 (文字誤り率 (%))

音声認識モデル	finetuned Whisper		streaming RNN-T	
音声発話区分化	webrtc	句読点	webrtc	句読点
ヘッドセット	28.9	**12.0**	41.9	**22.2**
マイクアレイ観測音	37.6	**27.1**	54.1	**43.8**
ILRMA	18.6	**14.3**	33.2	**25.2**
online RCSCME	23.5	**16.2**	38.1	**29.8**

表 7 に、JNAS 評価セットの文を用いた読み上げ音声と話し言葉音声の認識性能を示す。ヘッドセットマイクの音声ではすべてのロケーションで読み上げ音声が話し言葉音声を上回ったが、アレイ入力を用いた強調音では話し言葉より読み上げ音声が顕著に高い誤り率となった。これは、いくつかのセッションで読み上げ音声の SNR がきわめて低く、意味のある認識結果が得られなかったためである。

4. おわりに

対話ロボットの現実的なユースケースに則した条件下で新たな遠隔音声認識システムの評価用コーパスを構築し、現代的なアルゴリズムに基づく音声強調、音声発話区分化、バックエンド音声認識の評価を行った。バックエンドにフ

ルコンテクストモデルを用いた認識実験では、フロントエンドでストリーミング音声強調を用いた場合も、おおむね実用レベルの認識性能が得られた。また、句読点検出に基づく手法により、オンラインで人手分割に遜色ない性能の音声発話区分化が実現できた。一方、バックエンドにストリーミング RNN-Transducer を用いた音声認識は低い水準に留まった。真に低遅延かつ頑健な遠隔音声認識システムを構築するためには、バックエンドのストリーミング音声認識の性能を向上させることが最も重要な課題である。

謝辞 本研究は、JST ムーンショット型研究開発事業、JPMJMS2011 の支援を受けて行われた。

参考文献

[1] Yoshioka, T., Ito, N., Delcroix, M., Ogawa, A., Kinoshita, K., Fujimoto, M., Yu, C., Fabian, W. J., Espi, M., Higuchi, T., Araki, S. and Nakatani, T.: THE NTT CHIME-3 SYSTEM: ADVANCES IN SPEECH ENHANCEMENT AND RECOGNITION FOR MOBILE MULTI-MICROPHONE DEVICES, *Proc. ASRU*, pp. 436–443 (2015).

[2] Barker, J., Watanabe, S., Vincent, E. and Trmal, J.: The fifth 'CHiME' Speech Separatioand Recognition Challenge: Dataset, task and baselines, *Proc. Interspeech* (2018).

[3] S.Renals, T.Hain and H.Bourlard: Recognition and Understanding of Meetings: The AMI and AMIDA Projects, *Proc. IEEE Workshop Automatic Speech Recognition & Understanding* (2007).

[4] J.Barker, R.Marxer, E.Vincent and S.Watanabe: The third CHiME speech separation and recognition challenge: Dataset, task and baselines, *Proc. ASRU* (2015).

[5] K.Kinoshita, M.Delcroix, T.Yoshioka, T.Nakatani, E.Habets, R.Haeb-Umbach, V.Leutnant, A.Sehr, W.Kellermann, R.Maas, S.Gannot and B.Raj: The REVERB Challenge: A Common Evaluation Framework for Dereverberation and Recognition of Reverberant Speech, *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA-13)* (2013).

[6] Chen, Z., Yoshioka, T., Lu, L., Zhou, T., Meng, Z., Luo, Y. and Jian Wu and, Xiong Xia and, J. L.: CONTINUOUS SPEECH SEPARATION: DATASET AND ANALYSIS, *ICASSP*, pp. 7284–7288 (2020).

[7] Drude, L., Heitkaemper, J., Boeddeker, C. and Haeb-Umbach, R.: SMS-WSJ: Database, performance measures, and baseline recipe for multi-channel source separation and recognition, *arXiv preprint arXiv:1910.13934* (2019).

[8] Thiemann, J., Ito, N. and Vincent., E.: The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings, *21st International Congress on Acoustics, Acoustical Society of America* (2013).

[9] T.Nakatani, T.Yoshioka, K.Kinoshita, M.Miyoshi and B.H.Juang: Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation, *ICASSP*, pp. 85–88 (2008).

[10] Snyder, D., Chen, G. and Povey, D.: MUSAN: A Music, Speech, and Noise Corpus, *arXiv preprint arXiv:1510.08484v1* (2015).

[11] Ono, N.: Stable and fast update rules for independent vector analysis based on auxiliary function technique, *WASPAA*, pp. 189–192 (2011).

[12] Kitamura, D., Ono, N., Sawada, H., Kameoka, H. and Saruwatari, H.: Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 24, pp. 1626–1641 (2016).

[13] Sekiguchi, K., Nugraha, A. A., Bando, Y. and Yoshii, K.: Fast Multichannel Source Separation Based on Jointly Diagonalizable Spatial Covariance Matrices, *EUSIPCO* (2019).

[14] Kubo, Y., Takamune, N., Kitamura, D. and Saruwatari, H.: Blind Speech Extraction Based on Rank-Constrained Spatial Covariance Matrix Estimation With Multivariate Generalized Gaussian Distribution, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 28, pp. 1948–1963 (2020).

[15] Yoshimura, T., Hayashi, T., Takeda, K. and Watanabe, S.: End-to-end Automatic Speech Recognition Integrated with CTC-based Voice Activity Detection, *ICASSP*, pp. 6999–7003 (2020).

[16] 三村正人，河原達也：国会会議録のための音声から書き言葉への end-to-end 変換，自然言語処理，Vol. 30, No. 1 (2023).

[17] Li, B., Gulati, A., Yu, J., Sainath, T. N., Chiu, C.-C., Narayanan, A., Chang, S.-Y., Pang, R., He, Y., Qin, J., Han, W., Liang, Q., Zhang, Y., Strohman, T. and Wu, Y.: A BETTER AND FASTER END-TO-END MODEL FOR STREAMING ASR, *ICASSP*, pp. 5619–5623 (2021).

[18] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C. and Sutskever, I.: Robust Speech Recognition via Large-Scale Weak Supervision, *arXiv preprint arXiv:2212.04356* (2022).

[19] Graves, A.: Sequence transduction with recurrent neural networks, *LCML*, pp. 4945–4949 (2012).