Cocktail Machine Speech Chain: 重複あり音声を用いた音声認識・音声合成モデルの統一的学習 *

☆松永 裕太，高道 慎之介 (東大)，上乃 聖 (名工大)，猿渡 洋 (東大)

1 はじめに

対話音声の認識・合成技術 [1, 2] が近年発展している．それらの多くは各話者が交互に重複なく話す状況を想定しているが，実際の人間の対話では話者間の発話の重複が頻繁に発生する．対話での重複あり音声には，相槌や，笑いなどの非言語発話が含まれ，それらは聞き手の理解を示し [3]，対話の緊張を和らげる [4] 効果を持つ．そのような役割から，より自然な人間と計算機の音声インタラクションを実現するには，対話における重複あり音声を認識・合成する技術が必要である．

そこで我々は，重複あり音声を用いて，Automatic Speech Recognition (ASR), Text-To-Speech (TTS) モデルを同時に学習する手法の実現を目指す．これまでに，重複なし音声を用いて ASR, TTS モデルを学習する Machine Speech Chain [5] が提案されている．我々は，この手法をカクテルパーティのような実際の対話状況で発生する複数話者が重複して話す音声に拡張した Cocktail Machine Speech Chain を提案する (Fig. 1)．提案手法では，テキストと重複あり音声の対データを用いた ASR,TTS モデルの学習と，テキストのみの非対データを用いた ASR モデルの学習を同時に行う．これにより，Web 上などから容易に入手可能な膨大な言語資源を活用して ASR モデルの性能を改善することが可能となる．さらに今後の発展として，第 5 節で述べるように，重複あり音声のみの非対データを用いた音声合成モデルの学習にも拡張可能である．

本稿では，提案手法の詳細について述べ，ASR の性能改善の観点から客観評価を行う．さらに，話者の音声が完全に重複せず一部のみ重複するというより現実的な状況を想定し，話者間の音声の重複率と ASR の性能の関係を調査する．最後に，重複あり音声のみの非対データを用いた音声合成モデルの学習への拡張について議論する．

2 関連研究

2.1 重複あり対話音声の認識・生成

複数話者が重複して話す対話音声を認識する技術 [1, 6, 7, 8, 9, 10, 11, 12] が広く研究されており，Transformer の導入などにより高性能な認識が可能となっている [12]．単一チャネル [10]，あるいは複数チャネル [11] の重複あり音声データを用いる 2 種類のアプローチが存在する．本稿では，Web 上のデータ資源活用を目指し，単一チャネルの重複あり音声データを用いるアプローチを採用する．また，これらの研究では，テキストと重複あり音声の対データを含むデータセットを用いてモデルを学習している．この方法では，データセットに含まれるテキスト内容の多様性やテキストドメインの制約により，性能が制限される可能性がある．これに対し本稿では，Web 上から容易に収集可能な膨大なテキスト資源を活用することで，更なる性能向上を目指す．

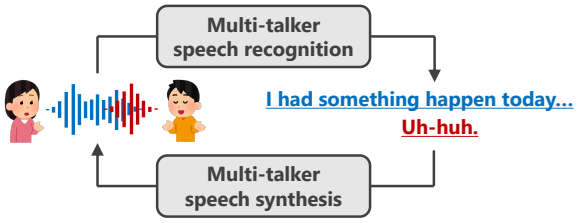

Fig. 1 Cocktail Machine Speech Chain の概要．

対話音声生成の技術が広く研究されている [2, 13, 14]．対話音声をテキストを介さず直接モデリングする話し言葉音声言語モデル dGSLM [2] が提案されている．CHATS [13] は，話し手のテキスト・音響特徴量を入力として利用可能にし，高品質な話し手の音声合成と聞き手の相槌の生成を同時に実現している．これらは，学習データが各話者の音声を別チャネルに保存していることを前提にするため，Web 上に存在するような，全音声を単一チャネルに保存したデータを学習に利用できない．これに対し我々の提案手法は，単一チャネルの重複あり音声での TTS モデルの学習にも発展可能であり，今後の研究対象である．

2.2 Machine Speech Chain

Machine Speech Chain [5] では，話し手が発話し，聞き手が発話内容を知覚するという Speech Chain の過程を計算機的に再現することで，ASR,TTS モデルの同時学習を可能にした．この手法では，1) テキスト・音声の対データを用いて計算される損失と，2) テキストデータのみから計算される損失，3) 音声データのみから計算される損失により，ASR,TTS モデルを学習する．2), 3) は，TTS を用いて ASR（あるいはその逆）のデータを増強する意図がある．学習には重複なし音声を用いる．我々の提案手法では，この手法を複数話者の重複あり音声を用いた学習に拡張する．

---

* Cocktail Machine Speech Chain: Unified training of speech recognition and synthesis models using overlapped speech by MATSUNAGA, Yuta, TAKAMICHI, Shinnosuke (The University of Tokyo), UENO, Sei (Nagoya Institute of Technology), and SARUWATARI, Hiroshi (The University of Tokyo).

3 Cocktail Machine Speech Chain

複数話者の重複あり音声を用いて ASR, TTS モデルを同時に学習する Cocktail Machine Speech Chain を提案する．Fig. 2 に提案法のモデル構成と学習法を示す．本稿では特に，テキストのみを用いた ASR モデルの精度改善に着目し，音声のみを用いた TTS モデルの品質改善 (第 5 節) は今後の研究対象とする．

3.1 モデル構成

提案手法では，複数話者の重複あり音声から各話者の発話内容を予測する multi-talker ASR（MT-ASR）モデルと，各話者のテキストから複数話者の重複あり音声を合成する multi-talker-mixing TTS（MT-TTS）モデルを用いる．MT-ASR モデルは，複数話者の重複あり音声を各話者の特徴量に分離する talker demixer と，各話者の特徴量から発話内容を予測する single-talker speech recognizer から構成され，これらは neural network (NN) を用いて実装される．MT-TTS モデルは，各話者について発話内容テキストから音声を合成する single-talker speech synthesizer と，各話者の合成音声からそれらを重複させた音声を出力する talker mixer から構成される．Single-talker speech synthesizer は，一般的な NN を用いた話者制御機能付き TTS と同様に構成され，テキストに加え話者ベクトルを入力とする．Talker mixer は，各合成音声と各々の発話開始時刻に基づき音声波形を単純に加算する．重複あり音声に含まれる話者数は既知（Fig. 2 では 2 名）とする．

3.2 モデル学習

提案手法では，以下の 2 種類の損失関数を用いてモデルを学習する．まず，1) 各話者のテキスト・複数話者の重複あり音声の対データを用いて計算される損失である．各モデルにより予測されたテキスト（あるいは重複あり音声）と真のテキスト（あるいは重複あり音声）の間で損失を計算する．次に，2) テキストのみを用いて計算される損失である．MT-TTS モデルにより重複あり音声を合成し，その音声から MT-ASR モデルにより各話者のテキストを認識する．認識されたテキストと入力テキストとの間で損失を計算する．損失関数は以下のように定式化される:

$$L = \alpha_1 L_{\text{TTS}} + \alpha_2 L_{\text{ASR}} + \alpha_3 L_{\text{TTS2ASR}}, \quad (1)$$

L_{TTS} は MT-TTS により合成された重複あり音声と真の重複あり音声の L1 loss である．L_{ASR} は MT-ASR により予測された各話者のテキストと真のテキストの損失で，先行研究 [12] の損失関数の式に従って計算される．L_{TTS2ASR} は MT-TTS により合成された重複あり音声から MT-ASR により予測された各話者のテキストと真のテキストの損失であり，L_{ASR} と同様の方法で計算される．

L_{TTS} の計算には，対データに加え，真の話者ベクトルと真の発話開始時刻のデータ（例えば，CALLHOME [15] のように話者 ID と発話開始時刻が付与

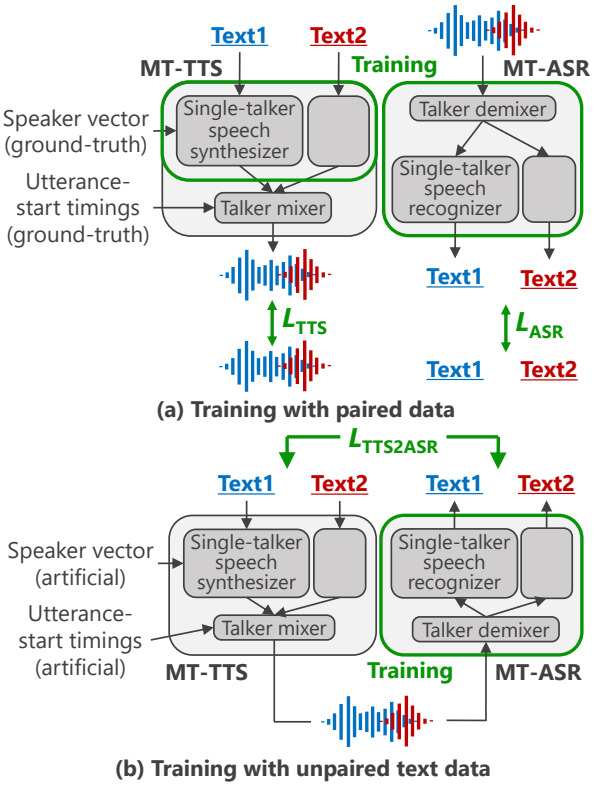

Fig. 2 (a) テキスト・重複あり音声の対データを用いた MT-TTS,MT-ASR の学習，および (b) 非対テキストデータを用いた MT-ASR の学習．

されたコーパス）を使用する．L_{ASR} はパーミュテーション不変学習に基づくため，通常の MT-ASR と同様に対データのみで計算できる．L_{TTS2ASR} の計算には L_{TTS} と同様に話者ベクトルと発話開始時刻を使用するが，L_{TTS} と異なり任意の設定（例えばランダムな話者対や発話開始時刻）を使用できる．これには，MT-TTS を用いて重複あり音声を疑似的に作成し，MT-ASR にデータ拡張効果をもたらす意図がある．

4 実験的評価

4.1 実験条件

提案手法の有効性検証のため，HiFiTTS [16]（読み上げ TTS コーパス）の異なる話者の音声を重複させることで，複数話者の重複あり音声のデータセットを疑似的に作成した．重複する話者数は 2 名とし，女性/女性の話者対を作成した．TTS, ASR モデルの事前学習に使用する LibriTTS_R, LibriMix と重複しないように話者を選択した．音声の重複に関する具体的な手順は，先行研究 [10] に従った．各話者対の重複あり音声発話のうち，約 6000, 100, 200 発話をそれぞれ学習，検証，テストセットとして使用した．

非対テキストデータによる学習で使用するテキストを用意するため，Project Gutenberg から 10 冊の英語の書籍をダウンロードした．HiFiTTS のテキストは Project Gutenberg の書籍を基にしているため，

Table 1　提案法の学習に用いる HiFiTTS の話者．F は女性を示す

Model	Speaker		
	Training(paird)	Training(unpaired text)	Test
Baseline	9136(F)/11614(F)	–	9136(F)/11614(F)
Proposed	9136(F)/11614(F)	9136(F)/11614(F)	9136(F)/11614(F)

重複しないように書籍を選択した．さらに，書籍のテキスト中から本文のみを抽出して使用した．学習では，約 8000 組のテキストデータを使用した．

Single-talker speech synthesizer では，入力を文字とし，モデルに FastSpeech2 [17] と HiFiGAN [18] を使用した．FastSpeech2 は，16 kHz にダウンサンプリングした LibriTTS_R [19] のサブセット train-clean-360 で事前学習した．学習のハイパーパラメータは公開実装[1]に従った．HiFiGAN には 16 kHz 音声での学習済みモデル[2]を使用した．MT-ASR には，モデル構成として transformer を用いた end-to-end の Hybrid CTC/attention モデル [12] を採用した．入力には 80 次元の対数メルフィルタバンク特徴量，出力には文字を用いた．事前学習済みモデルとして，LibriMix [20] で学習されたモデル[3]を使用した．話者ベクトルには，事前学習済みモデル[4]により抽出した x-vector を使用した．

学習では，学習率を 10^{-3} とし，Adam [21] オプティマイザと Noam の学習率スケジューラ [22] を使用した．Baseline, Proposed の学習では，バッチサイズを 8 とし，10 k ステップ学習させた．式 1 の $\alpha_1, \alpha_2, \alpha_3$ を，Baseline ではそれぞれ $0.0, 1.0, 0.0$ とし，Proposed ではそれぞれ $10^4, 1.0, 1.0$ とした．$L_{\mathrm{TTS}}, L_{\mathrm{TTS2ASR}}$ の計算で用いる各話者の話者ベクトルは，事前に HiFiTTS コーパスの各発話から計算した話者ベクトルの平均とした．L_{TTS} の計算では，事前にアライメントにより取得した真の文字継続長を使用した．L_{TTS2ASR} の計算では，Tab. 1 に示した話者の平均話者ベクトルと，ランダムな発話開始時刻を使用した．

MT-ASR モデルの推論では，ビームサイズを 4 としてビームサーチを使用し，LibriMix のテキストデータで学習された言語モデルを用いた shallow fusion による推論を行った．その他の設定は前述の事前学習済みモデルに従った．

4.2　音声認識精度改善に関する客観評価

非対テキストデータを用いた学習による ASR の精度改善効果を検証する．1) 事前学習済みモデル (Pretrained), 2) 対データのみを用いて学習したモデル (Baseline), 3) 対データと非対テキストデータを用いて学習したモデル (Proposed)，を比較する．各モデルの学習で用いる話者を Tab. 1 に示す．非対テキストデータを用いた学習では，表中に記載の話者の平均話者ベクトルを用いて MT-TTS により重複あり音声を合成した．評価には，4.1 節で述べた HiFiTTS を用いた疑似重複あり音声のテストセットを使用した．Word error rate (WER)/character error rate (CER) を基準に評価した．

Table 2　提案手法により学習された ASR モデルの精度に関する客観評価結果

Model	CER(%)	WER(%)
Pretrained	19.0	26.9
Baseline	16.8	22.3
Proposed	22.2	29.8

Table 3　提案手法により学習された ASR モデルの誤り種類ごとの WER(%)．SUB, DEL, INS はそれぞれ置換，削除，挿入誤りを示す

Model	WER (%)		
	SUB	DEL	INS
Pretrained	14.9	1.9	10.1
Baseline	10.1	0.7	11.4
Proposed	12.3	0.5	17.0

結果を Tab. 2 に示す．まず，Baseline は Pretrained に比べて誤り率が低く，評価データと同一の話者の対データを用いて学習したことにより精度が向上したことを示している．次に，Proposed は Baseline に比べて誤り率が高く，非対テキストデータを用いた学習により精度が低下するという結果となった．これは，MT-TTS の性能が低く対象話者の音声に近い音声を合成できていないため，データ拡張効果が得られなかったと考えられる．

置換・削除・挿入の各誤りに関する WER を，Tab. 3 に示す．Proposed では，Baseline に比べて特に挿入誤りが増加しており，真のテキストに存在しない単語を余分に挿入するような予測を行っていることを示している．

4.3　重複率と音声認識精度の関係の評価

次に，重複あり音声の重複度と認識精度の関係を調査する．疑似重複あり音声は，発話の重複率を考慮せず，ランダムな発話開始位置を用いて作成した．一方で，実際の対話音声は，音声の一部のみが低い割合で重複することが多いなど，重複率に偏りがある．そこで，重複あり音声の重複率と認識精度の関係を調査する．前述の疑似重複あり音声のテストセットと同一の発話を使用し，複数の重複率を基に重複あり音声を作成した．各重複率の評価データの発話数は 60 と

[1] https://github.com/ming024/FastSpeech2/tree/d4e79eb52e8b01d24703b2dfc0385544092958f3

[2] https://huggingface.co/speechbrain/tts-hifigan-libritts-16kHz

[3] https://huggingface.co/espnet/simpleoier_librimix_asr_train_asr_transformer_multispkr_raw_en_char_sp

[4] https://huggingface.co/speechbrain/spkrec-xvect-voxceleb

Table 4　重複あり音声のオーバーラップ率と提案手法により学習された ASR モデルの精度の関係．0S, 0L はそれぞれ，発話間に 0.0-0.5, 3.0 秒の無音区間を含む 0%のオーバーラップ率を意味する

(a) **CER(%)**

Model	Overlap ratio in %					
	0S	0L	10	20	30	40
Pretrained	24.9	41.9	18.7	14.6	16.7	18.7
Baseline	17.1	16.2	16.2	16.6	17.2	17.1
Proposed	19.6	19.3	17.7	19.8	17.5	21.9

(b) **WER(%)**

Model	Overlap ratio in %					
	0S	0L	10	20	30	40
Pretrained	31.8	51.9	25.7	22.3	24.6	26.2
Baseline	21.4	20.3	21.1	22.3	22.5	23.6
Proposed	25.5	26.1	23.8	26.9	25.0	29.4

した．先行研究 [23] に倣い，重複率を 0L, 0S, 10 %, 20 %, 30 %, 40 %と変更し，それらの音声での音声認識精度を比較した．

結果を Tab. 4 に示す．Baseline では，Pretrained と比べ，特に重複なしの音声で誤り率が低く，精度が改善していることがわかる．これは，評価対象の話者の対データでの学習により重複なし音声での精度が特に改善することを示している．また，Proposed では，Baseline と比べ，重複率によらず全体的に誤り率が上がり，精度が低下している．この結果は， 4.2 節と同様であり，今後は MT-TTS の品質改善により重複率によらず全体的な精度を改善する必要がある．

5　議論: 重複あり音声のみの非対データを用いた音声合成モデルの学習への拡張

本稿では，大規模なテキスト資源を用いた ASR の性能改善に着目し，非対テキストデータを用いた学習に焦点を当てた．この手法は，非対音声データを用いた学習にも拡張可能である．具体的には，疑似的に作成した重複あり音声を MT-ASR モデルに入力し，各話者のテキストを予測する．それらを，MT-TTS モデルに入力し，重複あり音声を合成する．そして，合成された重複あり音声と，入力の重複あり音声の間で計算した損失を基準に，MT-TTS モデルを学習する．これにより，重複あり音声のみの非対データを用いた TTS モデルの学習が可能となる．これが実現されれば，実際の重複を含む対話音声を用いて，書き起こしを必要とせず，TTS モデルを学習できる．

6　おわりに

本稿では，対話において重要な役割を果たす重複あり音声を認識・合成する技術の確立を目指し，重複あり音声を用いて ASR, TTS モデルを同時に学習する手法を提案し，音声認識精度の改善に関する有効性を評価した．今後は，音声認識精度の改善を目指し，重複あり音声を合成する TTS モデルの品質改善に取り組む．さらに，提案手法を重複あり音声のみを用いた TTS モデルの学習に拡張する．

謝辞　本研究は，JST 次世代研究者挑戦的研究プログラム JPMJSP2108，ムーンショット JPMJPS2011，JST 創発的研究支援事業 JP23KJ0828，科研費 21H05054, 22H03639，23H03418 の支援と，東京大学の齋藤佑樹博士, 佐伯高明氏の協力を受け実施したものです．

参考文献

[1] S. Watanabe et al., “CHiME-6 Challenge: Tackling Multispeaker Speech Recognition for Unsegmented Recordings,” in *Proc. CHiME 2020*, 2020.

[2] T. A. Nguyen et al., “Generative Spoken Dialogue Language Modeling,” *TACL*, 2023.

[3] Y. V. H., “On getting a word in edgewise,” *Chicago Linguistics Society, 6th Meeting*, 1970.

[4] V. Adelswärd, “Laughter and Dialogue: The Social Significance of Laughter in Institutional Discourse,” *Nordic Journal of Linguistics*, 1989.

[5] A. Tjandra et al., “Machine Speech Chain,” *TASLP*, 2020.

[6] A. Tripathi et al., “End-To-End Multi-Talker Overlapping Speech Recognition,” in *Proc. ICASSP*, 2020.

[7] N. Kanda et al., “Serialized Output Training for End-to-End Overlapped Speech Recognition,” in *Proc. Interspeech*, 2020.

[8] L. Lu et al., “Streaming End-to-End Multi-Talker Speech Recognition,” *IEEE SPL*, 2021.

[9] D. Raj et al., “Integration of Speech Separation, Diarization, and Recognition for Multi-Speaker Meetings: System Description, Comparison, and Analysis,” in *Proc. SLT*, 2021.

[10] H. Seki et al., “A Purely End-to-end System for Multi-speaker Speech Recognition,” *arXiv:1805.05826*, 2018.

[11] X. Chang et al., “MIMO-SPEECH: End-to-End Multi-Channel Multi-Speaker Speech Recognition,” *arXiv:1910.06522*, 2019.

[12] ——, “End-To-End Multi-Speaker Speech Recognition With Transformer,” in *Proc. ICASSP*, 2020.

[13] K. Mitsui et al., “Towards human-like spoken dialogue generation between AI agents from written dialogue,” *arXiv:2310.01088*, 2023.

[14] Z. Borsos et al., “SoundStorm: Efficient Parallel Audio Generation,” *arXiv:2305.09636*, 2023.

[15] A. Canavan et al., “CALLHOME American English Speech LDC97S42,” *Philadelphia: Linguistic Data Consortium*, 1997.

[16] E. Bakhturina et al., “Hi-Fi Multi-Speaker English TTS Dataset,” *arXiv:2104.01497*, 2021.

[17] Y. Ren et al., “FastSpeech 2: Fast and High-Quality End-to-End Text to Speech,” in *Proc. ICLR*, 2021.

[18] J. Kong et al., “HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis,” *arXiv:2010.05646*, 2020.

[19] Y. Koizumi et al., “LibriTTS-R: A Restored Multi-Speaker Text-to-Speech Corpus,” in *Proc. Interspeech*, 2023.

[20] J. Cosentino et al., “LibriMix: An Open-Source Dataset for Generalizable Speech Separation,” *arXiv:2005.11262*, 2020.

[21] D. P. Kingma, J. Ba, “Adam: A Method for Stochastic Optimization,” *arXiv:1412.6980*, 2014.

[22] A. Vaswani et al., “Attention is All you Need,” in *Proc. NeurIPS*, 2017.

[23] Z. Chen et al., “Continuous Speech Separation: Dataset and Analysis,” in *Proc. ICASSP*, 2020.