言語非依存な口真似による効果音合成手法 PronounSEの評価

〇滝沢力(京産大院/産総研),平井重行(京産大),金崎朝子(東京科学大),須田仁志(産総研)

1. 背景

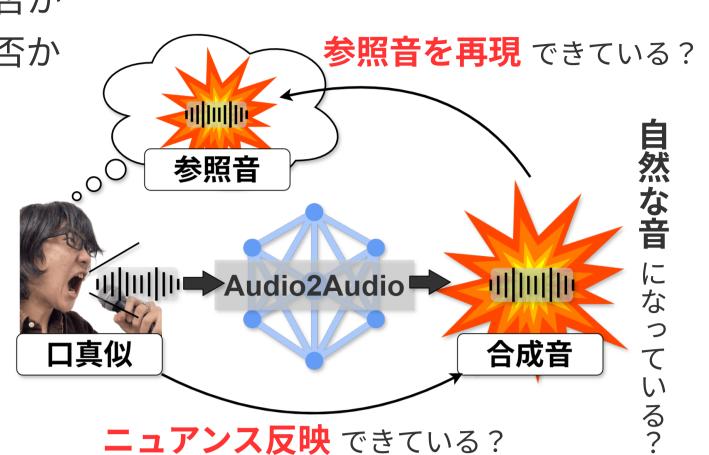
- ・環境音,効果音合成の研究事例の増加ノ
- ・口真似からの効果音合成手法 <u>PronounSEの提案</u>
- ▶ 爆発音用データセットにより,口真似ニュアンスを反映した**爆発音合成**を実現
- ・当該分野の評価手法の課題
- ▶ 合成音の品質,自然性に関するMOS
- ► FAD [1] やKLDによる客観評価
- ・扱う効果音は架空の音や環境音の
- デフォルメ音も含む → 自然性? 不十分?
- ・主観に沿わない場合あり

・Audio-to-Audioに適切な評価軸を策定する必要あり

▶ 策定する評価軸を基に主観・客観評価を実施!

2. 評価軸の策定

- ・Audio-to-Audioにおいて,入力と出力の妥当性が重要
- ▶ 想像した(望んだ)音が得られているか否か
- ▶ 入力口真似に応じた合成になっているか否か
- ・以下の観点で爆発音合成の評価を実施
- ① 所望する音との類似性
- ② 口真似ニュアンスの反映性
- ③ 合成音の自然性
- ・評価対象の手法
 PronounSE (Ours)
 T-Foley (TF) [2]
 Stable Audio 2.0 (SA2) [3]



3. 実験準備

・モデル準備

モデル	手法概要	学習データセット	エポック数 バッチサイズ	サンプリング周波数
Ours	・口真似と効果音のMelの変換を学習 ・Neural Vocoderで波形合成	1,748種類の爆発音と 3名による7,768個の口真似	20,000 64	22,050 Hz
TF [2]	・効果音(波形)の逆拡散をモデル化 ・効果音のRMSとラベルで条件付け	1,748種類の爆発音のみ (学習時に口真似不要)	5,000 8	22,050 Hz
SA2 [3]	・大規模な音楽・効果音データで学習 ・プロンプトと参照の音響信号を入力可能	大規模な音楽・効果音データ	不明	44,100 Hz

- ・構築した評価用データセット
- ▶ 100種類の未学習爆発音の収集と未学習話者6名(m-01~05, f-01)による600個の口真似

4. 所望する音との類似性の客観評価結果

- ・全ての音源を事前学習済みPANNs [4] で2048次元の埋め込みに変換
- ・実施する所望する音との類似性に関する客観評価

提案

- ▶ 参照音と合成音の埋め込みを用いたFAD [1]
 - → 参照音100音を評価セット,話者毎 モデル毎の合成音を検証セット
- ▶ 参照音と合成音の埋め込みのコサイン類似度
- → 合成音と対応する参照音のコサイン類似度を算出し,話者 モデル毎で平均

	FAD by PANNs ↓			Cosine similarity ± SD ↑		
Speaker ID	TF [2]	<u>Ours</u>	SA2 [3]	TF [2]	<u>Ours</u>	SA2 [3]
m-01	54.02	17.85	21.35	0.73 ± 0.07	0.85 ± 0.08	0.83 ± 0.06
m- 02	54.47	17.05	20.94	0.74 ± 0.07	0.86 ± 0.06	0.84 ± 0.06
m-03	48.85	17.28	20.95	0.75 ± 0.07	0.86 ± 0.07	0.84 ± 0.06
m-04	48.52	16.65	21.46	0.75 ± 0.08	0.85 ± 0.07	0.84 ± 0.06
m- 05	47.89	19.29	19.75	0.75 ± 0.08	0.86 ± 0.07	0.85 ± 0.07
f-01	56.94	19.28	22.01	0.72 ± 0.07	0.85 ± 0.07	0.85 ± 0.06
Whole	55.93	14.09	21.15	0.74 ± 0.07	0.85 ± 0.07	0.84 ± 0.06

・今回の客観評価結果ではOurs の再現性が他2つより高いことを確認

5. 主観評価結果

・100 Hz以下のノイズや定常ノイズ発生

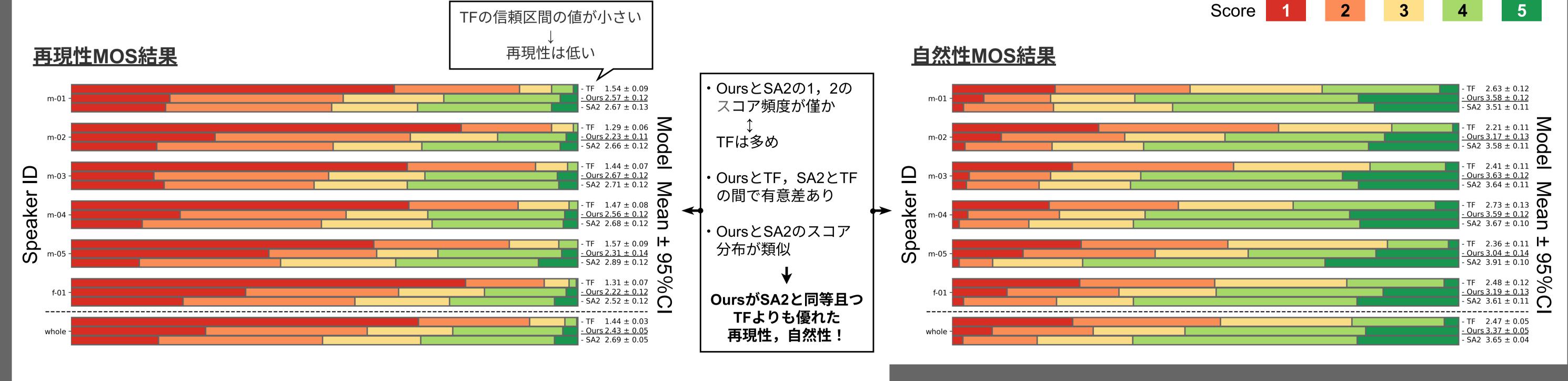
・10種類の参照爆発音と口真似,合成音を使用して350名によるクラウドソーシング上での主観評価を実施

・横縞ノイズがしばしば発生

- ▶ 参照音に対する合成音の再現性主観評価(**再現性MOS**)
- :爆発音と合成音を使用,1.完全に異なる ~ 5.とても似ている
- ▶ 合成音への口真似ニュアンス反映性主観評価(ニュアンス反映性MOS)
- :口真似と合成音を使用,1.全く反映されていない ~ 5.とても反映されている

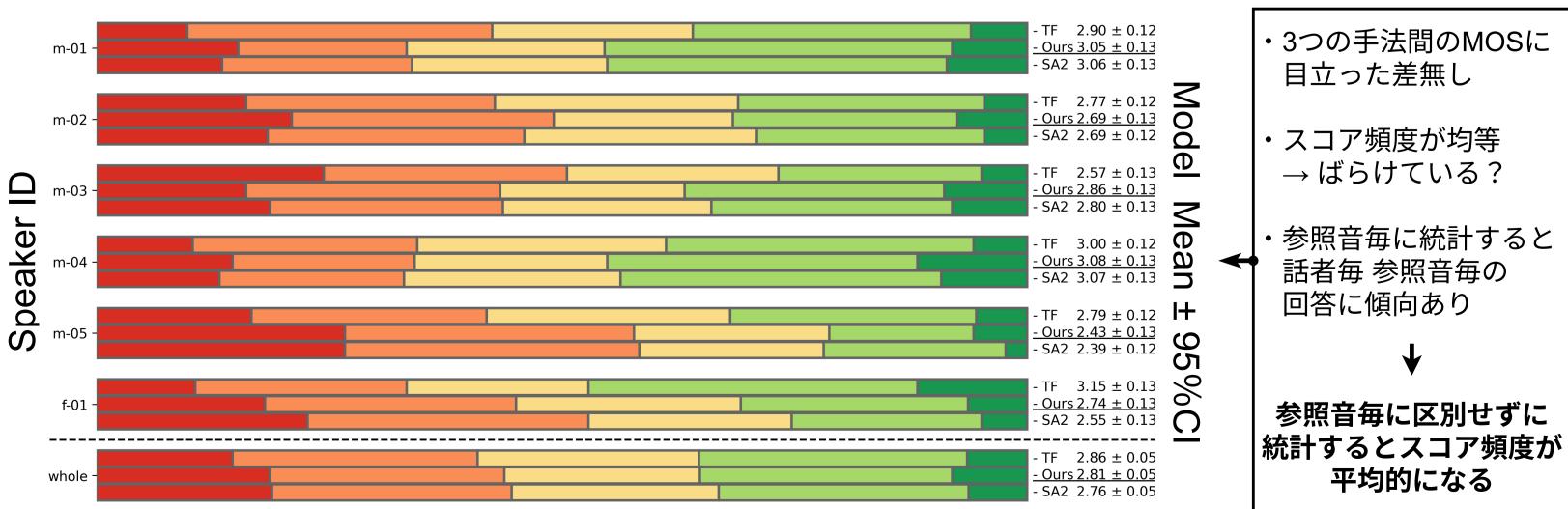
▶ 合成音の自然性主観評価(**自然性MOS**)

:合成音のみ使用 ,1.かなり不自然 ~ 5.とても自然



・破壊音のような合成音が多い

ニュアンス反映性MOS結果



SIZ

6. まとめと課題

- ・評価結果に関して
- ▶ 再現性,自然性でPronounSEがT-Foleyよりも優れたスコア
 - → 口真似と効果音を対応付けた学習の有効性を示唆
- ▶ PronounSEの横縞模様の雑音により,自然性低下
- ▶ ニュアンス反映性MOSで,参照音や話者毎に傾向が異なることを確認
- ・ 今後の課題
- ▶ 自然性,品質向上に向けた横縞模様の雑音の改善
- ▶ ニュアンス反映性MOSに関する参照音毎 話者毎の詳細な分析
- ▶ 効果音制作ツールとしての新規性,有効性を検証するユーザ評価
- ▶ PronounSEにおける爆発音以外での合成