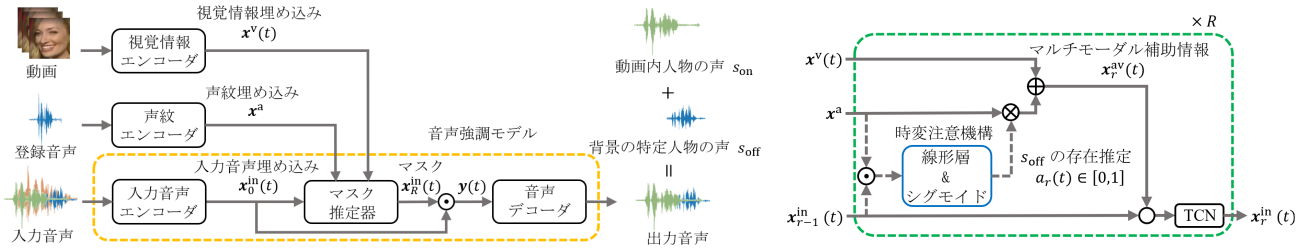
5S-06

動画内話者の音声強調における特定背景音声の透過

吉永 朋矢 † 田中 啓太郎 † 森島 繁生 ††
† 早稲田大学 †† 早稲田大学理工学術院総合研究所



(a) モデル全体の構造　　(b) マスク推定器内部の繰り返し構造

図 1: 動画内の人物と背景の特定人物の音声を同時に抽出するニューラルネットワーク

1. はじめに

本稿では，ノイズ環境で人が話している動画から，計算機によってその話者の音声と動画外の特定人物の音声を同時に抽出する問題を扱う．動画内の口の動きを補助に，同期する動画内話者の音声を抽出するタスクを audio-visual speech enhancement (AV-SE) [1] という．口の動きはノイズに影響されないため，AV-SE は様々なノイズに頑健な音声の抽出を可能にする．AV-SE の応用先は，補聴器や音声認識システムの前処理など多岐にわたる．

近年，深層学習を用いた AV-SE は高い推定精度を誇っている．AV-SE における多くの先行研究は，動画外の音を一様に抑制することを目的としてきた．しかし，補聴器を筆頭に実応用の場面では，動画外の音でも透過すべき特定人物の声が存在しうる．例えばユーザが子供の場合，家族や学校の先生の声は，ユーザの視線が発話者の顔を捉えていなくとも聞こえることが望ましい．また，駅の構内アナウンスは，安全上あらゆるユーザにとって抑制されるべきでない．このような状況への対処には，動画内の人物の音声（s_{on}）に加え，声の特徴を事前に与えられている動画外の特定人物の音声（s_{off}）も同時に抽出できる AV-SE の手法が必要となる．

この問題設定に取り組む第一のアプローチとして，s_{on} と s_{off} を独立に抽出した後，再合成する手法が考えられる．具体的には，s_{on} は AV-SE を用いて，s_{off} は事前に収録した話者の音声（登録音声）から得られる声の特徴を補助に，その話者の声を抽出する手法 [4] を用いて抽出する．しかし，このアプローチでは，個々の推定で入り込んだノイズやアーティファクトが蓄積し，推定精度が悪化してしまう．さらに，ノイズの混ざった同一の入力音声から個々の出力音声までのネットワーク（音声強調モデル）を別個に用意するため，その冗長性によりモデルサイズも増加してしまう．第二に別アプローチとして，人間の声以外の音の抑制 [2] や，事前に収録した音と同種類の音の抑制 [3] を行うノイズ抑制手法の適用も考えられる．しかしながら，ノイズには人間の声を含むあらゆる音が想定されるため，抑制対象に関する事前情報を与えるアプローチでは，事実上解決不能である．

Selective Off-Screen Speech Extraction for Audio-Visual Speech Enhancement: *Tomoya Yoshinaga†, Keitaro Tanaka†, and Shigeo Morishima†† (†Waseda University, ††Waseda Research Institute for Science and Engineering)*

本研究では，単一の音声強調モデルで出力混合音声を直接推定する枠組みを提案する．加えて，s_{off} に対する時変注意機構と，s_{on} または s_{off} を遮断する新たな学習方法により，さらなる推定精度向上を図る．評価実験を通じて，波形推定精度とモデルの軽量さの観点で本手法の有用性を確認する．

2. 提案手法

2.1 問題設定

本稿では，(1) ノイズ環境下の動画に対して，動画内の人物と背景の特定人物の音声の同時抽出を実現し，(2) 我々の直接推定アプローチが，二つの音声強調の出力の再合成を行うアプローチに対して，推定精度とモデルの軽量さの観点で優れていることを確認する．ベースライン手法として，s_{on} を既存の AV-SE モデル [1] で抽出し，s_{off} を既存の話者抽出モデル [4] で抽出した後，再合成を行うアプローチを採用する．これに対し提案手法では，動画から抽出した s_{on} の話者の口の動きと，登録音声から抽出した s_{off} の話者の声の特徴によって，一つの音声強調モデル [5] を同時に条件付ける．

2.2 提案手法

提案手法は五つの部分で構成される（図 1）．まず，入力音声エンコーダが入力音声を埋め込み表現 $\boldsymbol{X}_0^{in} = \{\boldsymbol{x}_0^{in}(1), \ldots, \boldsymbol{x}_0^{in}(T)\} \in \mathbb{R}^{T \times D^{in}}$ に変換する．ただし T は時間，D^{in} は特徴量次元数である．並行して，視覚情報エンコーダが s_{on} の話者の口の部分の動画から，口の動きを表す埋め込み表現 $\boldsymbol{X}^v = \{\boldsymbol{x}^v(1), \ldots, \boldsymbol{x}^v(T)\} \in \mathbb{R}^{T \times D^{aux}}$ を抽出 [1] し，声紋エンコーダが登録音声から，s_{off} の話者の声の特徴を表す時不変の埋め込み表現 $\boldsymbol{x}^a \in \mathbb{R}^{D^{aux}}$ を抽出 [4] する．ただし D^{aux} は特徴量次元数である．これらをもとに，$\boldsymbol{X}^v$，$\boldsymbol{x}^a$ によって条件付けられたマスク推定器が，$\boldsymbol{X}_0^{in}$ からそれ自身に適用するマスク $\boldsymbol{X}_R^{in} \in \mathbb{R}_0^{+\,T \times D^{in}}$ の推定を行う．具体的には，音声強調モデル [5] に基づき，連続した $R(=4)$ 個の temporal convolutional network (TCN) が段階的にマスクを推定する．r 番目の TCN は，各時刻で $\boldsymbol{X}^v$ と $\boldsymbol{x}^a$ の和をとることで得た $\boldsymbol{X}_r^{av} = \{\boldsymbol{x}_r^{av}(1), \ldots, \boldsymbol{x}_r^{av}(T)\} \in \mathbb{R}^{T \times D^{aux}}$（マルチモーダル補助情報）と $\boldsymbol{X}_{r-1}^{in}$ を結合した特徴量を入力として，$\boldsymbol{X}_r^{in} = \{\boldsymbol{x}_r^{in}(1), \ldots, \boldsymbol{x}_r^{in}(T)\} \in \mathbb{R}^{T \times D^{in}}$ を出力し，次の TCN へ送る．最後に，音声デコーダが $\boldsymbol{X}_R^{in}$

と $\boldsymbol{X}_0^{\mathrm{in}}$ の要素積であるマスク適用後の入力音声の埋め込み表現 $\boldsymbol{Y} = \{\boldsymbol{y}(1), \ldots, \boldsymbol{y}(T)\} \in \mathbb{R}^{T \times D^{\mathrm{in}}}$ から，$s_{\mathrm{on}} + s_{\mathrm{off}}$ の推定波形の復元を行う．学習時の損失関数には次式で計算される SNR ロス $\mathcal{L}^{\mathrm{sep}}$ を用いる．

$$\mathcal{L}^{\mathrm{sep}} = -10 \log_{10} \frac{\|s\|^2}{\|\hat{s} - s\|^2} \tag{1}$$

なお，s と $\hat{s}$ はそれぞれ正解波形と推定波形である．

この枠組みに加え，本研究ではさらなる性能向上のため，二つの機構を導入する．一つは，s_{off} に対する時変注意機構（attention mechanism, AM）である．s_{off} が存在しない時間に不要となる $\boldsymbol{x}^{\mathrm{a}}$ を $\boldsymbol{X}^{\mathrm{v}}$ と同時に参照することは，s_{on} のみの抽出を妨げる可能性がある．本研究では，音声検出機構 [4] により，各時刻 t $(t = 1, \ldots, T)$ において $\boldsymbol{x}_{r-1}^{\mathrm{in}}(t)$ と $\boldsymbol{x}^{\mathrm{a}}$ から s_{off} の存在の信頼の度合 $a_r(t)$ を推定する．$a_r(t)$ が $\boldsymbol{x}^{\mathrm{a}}$ の参照の度合となるように，各時刻におけるマルチモーダル補助情報を次式で計算する．

$$\boldsymbol{x}_r^{\mathrm{av}}(t) = \boldsymbol{x}^{\mathrm{v}}(t) + a_r(t)\boldsymbol{x}^{\mathrm{a}} \tag{2}$$

AM の適用時は，以下の損失関数 $\mathcal{L}^{\mathrm{total}}$ を用いる．

$$\mathcal{L}^{\mathrm{total}} = \mathcal{L}^{\mathrm{sep}} + \mathcal{L}^{\mathrm{att}} \tag{3}$$

ただし，$\mathcal{L}^{\mathrm{att}}$ は $a_r(t)$ と s_{off} の有無（存在する時間は 1，しなければ 0）との間のクロスエントロピーである．

もう一つは，新たな学習方法（muting strategy，MS）である．提案モデルでは，二つの補助情報を同時に参照して s_{on} と s_{off} の混合音を出力する．ここで，各音声と補助情報の対応関係は陽に取り扱われないため，正確な波形の推定に必要な補助情報の抽出方法を明示的に学習する余地がある．本研究では，訓練時に s_{on} か s_{off} を一定確率で消すことで，モデルに片方の音声のみを抽出させる．これにより，モデルに正しい対応関係の学習を促す．

3. 評価実験

3.1 実験条件

s_{on}，s_{off}，ノイズのために，VoxCeleb2 [6] の動画付き音声，WSJ0 [7] の音声，AudioSet [8] の音を用いた．AudioSet のノイズは人間の声，音楽，生活音のような幅広い種類の環境音である．訓練と検証には VoxCeleb2 の 800 人の 25,000 個の音声，WSJ0 の 101 人の 12,776 個の音声，AudioSet の 18,870 個の音を用い，各々のデータの 80%を訓練に，20%を検証に用いた．評価には VoxCeleb2 の 118 人の 3,000 個の音声，WSJ0 の 18 人の 1,857 個の音声，AudioSet の 3,000 個の音を用いた．ただし，評価用集合の音声の人物は訓練と検証用集合には存在しない．ランダムな組み合わせで s_{on} に s_{off} とノイズを -2.5dB から 2.5 dB のランダムな SNR で混合することで，訓練，検証，評価のためにそれぞれ 20,000，5,000，3,000 個の 4 秒間の動画付き入力音声を作成した．音声のサンプリング周波数は 16kHz，動画のフレームレートは 25fps である．混合に際し，s_{off} は訓練と検証時は 2 から 4 秒，評価時は 0 から 4 秒のランダムな長さで入力音声のランダムな時間位置に配置した．登録音声は s_{off} と同じ話者の異なる音声を用いた．

ノイズに様々な音を想定するため，環境音ノイズ下での実験 (A) に加え，VoxCeleb2 または WSJ0 の s_{on}，s_{off} と異なる話者の音声をノイズとして用いた実験 (B) も行った．提案手法の波形推定精度をベースライン手法と比較評価し，評価尺度には，音声内のノイズの小ささを表す SI-SDR の推定前後での変化量 SI-SDRi を用いた．

表 1: ベースライン手法と提案手法との比較

手法	SI-SDRi (dB) ↑ 実験 (A)	SI-SDRi (dB) ↑ 実験 (B)	#Params ↓
ベースライン	7.35	7.46	29.8M
提案	**8.06**	**8.73**	**25.1M**

表 2: 二種類の提案機構の効果検証

手法	AM	MS	SI-SDRi (dB) ↑	#Params ↓
ベースライン	-	-	7.35	29.8M
提案	-	-	7.56	**25.1M**
	-	✓	7.67	**25.1M**
	✓	-	7.77	**25.1M**
	✓	✓	**8.06**	**25.1M**

3.2 実験結果

提案手法はベースライン手法と比較して，16%少ないモデルパラメータ数でより高い推定精度を実現した（表 1）．実験 (A) の環境音ノイズだけでなく，実験 (B) の音声ノイズ下でも，提案手法は推定精度の観点で有効である．これは，提案手法がさまざまなノイズ環境に適用できることと，視覚情報と事前情報を参照して所望の音声を選択的に抽出できることを示している．表 2 に示した時変注意機構と新たな学習方法の両方がない場合の実験結果から，直接推定の枠組み自体の推定精度における有効性が確認できる．また，表 2 の実験結果は二つの機構が精度向上に貢献することも示している．提案モデルは軽量であるが，これは音声強調モデルが一つのみ存在し，冗長性が削られたためである．時変注意機構は一つの線形層で実現可能であり，新たな学習方法はパラメータ数を増加させないため，モデルの軽量さを保つことができる．

4. おわりに

本稿では，動画内話者の音声強調における特定背景音声の透過の手法を提案した．評価実験の結果，ベースライン手法と比較して，動画内の音声と背景の特定人物の音声の混合音の抽出精度とモデルの軽量さの観点で，提案手法の有効性が確認された．今後は，透過する特定の背景の音をサイレンや警笛のような人間の声以外の音へ拡張することを目指す．また，透過する背景の音が複数種類存在する場合に対応可能な枠組みへの拡張も目指す．

謝辞 本研究は，JSPS 科研費 (19H04137，21H05054，22J22424) の補助を受けた．

参考文献

[1] J. Wu *et al.*: “Time domain audio visual speech separation,” *IEEE ASRU*, 667–673, 2019.

[2] S. Pascual *et al.*: “SEGAN: Speech enhancement generative adversarial network,” *Interspeech*, 3642–3646, 2017.

[3] S. Liu *et al.*: “N-HANS: A neural network-based toolkit for in-the-wild audio enhancement,” *Multimedia Tools and Applications*, 28365–28389, 2021.

[4] Y. Hao *et al.*: “Wase: Learning when to attend for speaker extraction in cocktail party environments,” *IEEE ICASSP*, 6104–6108, 2021.

[5] Y. Luo *et al.*: “Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation,” *IEEE/ACM TASLP*, 1256–1266, 2019.

[6] J. S. Chung *et al.*: “VoxCeleb2: Deep speaker recognition,” *Interspeech*, 1086–1090, 2018.

[7] J. Garofolo *et al.*: “CSR-I (WSJ0) Complete LDC93S6A,” *Philadelphia: Linguistic Data Consortium*, 1993.

[8] J. F. Gemmeke *et al.*: “Audio set: An ontology and human-labeled dataset for audio events,” *IEEE ICASSP*, 776–780, 2017.