人工知能学会第二種研究会資料
AIチャレンジ研究会
SIG-Challenge-068-09

話者情報を利用した距離ベース音声強調の改善

Improvement of Distance-based Speech Enhancement Using Speaker Information

糸山 克寿[1*]　畑 和也[2]　住田 直亮[2]　中臺 一博[1]
Katsutoshi Itoyama[1]　Kazuya Hata[2]　Naoaki Sumida[2]　Kazuhiro Nakadai[1]

[1] 東京科学大学　[2] 株式会社ホンダ・リサーチ・インスティチュート・ジャパン
[1] Institute of Science Tokyo　[2] Honda Research Institute Japan Co., Ltd.

Abstract: Abstract (English) comes here.....

1　はじめに

深層学習技術の進展により音声認識の性能は飛躍的に向上し、会議記録作成や音声アシスタントなど、多くの場面で実用化が進んでいる。これらのタスクでは、混合音の中から目的話者の発話を正確に抽出することが不可欠である。しかし一般的なデバイスに搭載されているのは単一マイクロホンであり、モノラル信号からは音源方向を推定するための位相差情報が得られない。そのため、複数話者が同時に存在する環境では、誰が発話したかを音響信号のみで適切に識別することができない。

近距離音声を強調する技術が実現できれば、たとえばスマートフォンに収録される音声は所有者の声のみとなり、会議などにおいて発話者と発話内容を自動的に対応付けられる。また、特別なマイクロホンアレイを必要とせず、普及率の高い一般デバイスだけで高精度な議事録作成や音声認識を実現できるという大きな利点がある。

本研究の目標は、**単一マイクロホンから得られるモノラル信号のみを用いて、マイク近傍の話者（主としてデバイス所有者）の音声を高精度に抽出すること**である。これにより、

- 会議など多人数環境でも、各デバイスに録音されるのはその所有者の音声のみとなり、発話内容と発話者を確実に対応付けられる
- スマートフォンやロボットといった日常的なデバイスの音声認識性能が、雑音や他者の発話の影響を受けにくくなる

といった効果が期待できる。

*連絡先：東京科学大学 工学院 システム制御系
〒 152-8552 東京都目黒区大岡山 2-12-1-W8-18
E-mail: itoyama@i.sc.eng.isct.ac.jp

この目的の達成に向けては、以下の技術的課題を解決する必要がある。

1. モノラル信号では方向情報が得られない単一マイクロホンでは位相差情報が得られないため、複数の近距離話者が存在した場合に誰の声かを識別できない。
2. 距離ベース手法 (DSE) の限界 DSE は直接音と反射音の比率を利用することで近距離音声を強調できるが、近距離に他者が存在すると誤って抽出してしまう、話者識別の機能を持たない、といった課題があり、単一マイクロホンでの確実な近距離音声抽出には不十分である。
3. 話者ベース手法 (TSE) の限界 TSE は話者特徴が事前登録された場合のみ有効であり、未登録話者の抽出ができない。そのため、近距離にある音声を抽出したいという今回の目的とは一致しない。

本稿では、上記の課題を踏まえて、単一マイクロホンを用いた近距離音声強調手法を提案する。提案手法では、

- 距離に依存して変化する音響的手がかり（直接音・反射音の比率、残響特性など）を利用する
- モノラル信号のみから、近距離音声を抽出し遠距離音を抑圧する
- 会議や日常環境において、デバイス所有者の音声のみを取得できる状況を実現する

といった点に重点を置く。提案手法は、特殊なマイクロホンアレイを必要とせず、一般的なスマートフォンなどに搭載された単一マイクロホンのみで利用可能である点が大きな特徴である。

2　関連研究

本節では、本研究に関連する研究について述べ、それらの課題を検討する。

2.1　時間周波数マスクによる音声強調

音声信号は時間とともに振幅が変化する一次元の波形として観測されるが、その性質をより詳細に扱うため、多くの音声強調手法では時間と周波数の両方の変化を表す時間周波数表現が用いられる。代表的な時間周波数表現がスペクトログラムであり、横軸を時間、縦軸を周波数、色の濃淡で各時間周波数成分の強度を表す。無音区間では全体的にエネルギーが低く、スペクトログラム上では暗い（あるいは青い）領域として表現される。

近距離音声と遠距離音声が混合している場合、それぞれは時間周波数平面上に異なるパターンとして現れる。例えば、ある区間では近距離話者の発話が強く現れ、その後に遠距離話者の発話が主となるような場合、混合音声のスペクトログラムは、近距離音声に対応する強い成分と、遠距離音声に対応する弱い成分が重畳した形となる。

時間周波数マスクは、この混合スペクトログラムに対して要素ごとに乗算されるフィルタであり、モノラル音響信号に対する音声強調手法として広く用いられている。混合信号のスペクトログラムを $Y(t,f)$、マスクを $M(t,f)$ とすると、強調後のスペクトログラム $\hat{S}(t,f)$ は

$$\hat{S}(t,f) = M(t,f) \odot Y(t,f) \tag{1}$$

のように要素積で表される。ここで、$M(t,f)$ は 0 から 1 の範囲の値であり、1 に近いほどその時間周波数成分をそのまま通過させ、0 に近いほど抑圧することを意味する。近距離音声強調の文脈では、「近距離音声に由来すると推定される成分のマスク値を 1 に近づけ、それ以外（遠距離音声や雑音）のマスク値を 0 に近づける」ことで、近距離音声のみから構成されるスペクトログラムを得ることを目指す。

近年の深層学習ベースの音声強調・音源分離手法では、この時間周波数マスクをニューラルネットワークによって直接推定する枠組みが主流となっている。本研究で扱う距離ベース音声強調（DSE）およびターゲット話者抽出（TSE）も、いずれも時間周波数マスキングの枠組みの上に構成されている。

2.2　距離ベース音声強調

距離ベース音声強調（Distance-based Speech Enhancement, DSE）は、音源とマイクロホンの距離に依存する音響的性質を利用して、近距離音声を強調する手法である。室内環境においてマイクで観測される音は、音源から直接届く「直接音」と、壁や天井等で反射した「残響音」の和として捉えることができる。直接音のエネルギーは音源との距離の二乗に反比例して減衰する一方、残響音は空間内で比較的均一なエネルギー分布を持ち、距離による減衰が緩やかである。その結果、近距離音源ほど Direct-to-Reverberant Ratio（DRR）が高くなる傾向がある。

DSE は、この DRR を主な手がかりとして、時間周波数マスクを用いて近距離成分を強調し、遠距離成分を抑制する。具体的には、混合音声のスペクトログラム上の各時間周波数ビンに対して、そのビンが近距離音声に由来するか遠距離音声に由来するかを推定し、近距離と推定されるビンのマスク値を高く、遠距離と推定されるビンのマスク値を低く設定する。

距離ベース音源分離という枠組みは、Patterson らによって初めて提案された。彼らは、近距離音声と遠距離音声を混合したモノラル信号のスペクトログラムを入力とし、畳み込みニューラルネットワーク（CNN）と再帰型ニューラルネットワーク（LSTM）を組み合わせたモデルにより、近距離音声を通過させる時間周波数マスクを推定する手法を提案した。この研究は、単一チャネル（モノラル）音声信号を用いて距離ベースの音声強調が可能であることを示した点で重要である。しかし、Patterson らの手法では、混合音声が近距離音声と遠距離音声のみから構成されており、それ以外の背景雑音が考慮されていない。また、仮想空間に配置した仮想マイクロホン・音源に基づくシミュレーション環境でのみ評価が行われており、実環境の録音に対する検証が十分ではなかった。

石井らは、Patterson らの手法を発展させ、より実環境に近い条件での距離ベース音声強調を実現した。具体的には、背景雑音を加えた上で、各音源に対して実環境で収録したインパルス応答を畳み込むことで学習データを生成し、Patterson らの CNN を表現力の高い ResNet に、LSTM を高速な GRU に置き換えたネットワークアーキテクチャを構築した。これにより、距離ベース音声強調の性能向上と計算効率の改善が報告されている。

一方で、これら DSE 系手法には依然として課題が残る。近距離・遠距離という距離情報のみに基づいてマスクを推定するため、近距離に複数の話者が存在する場合、それらを区別することができない。また、背景雑音の抑圧が過剰になった結果として音声成分が削られたり、逆に雑音が取り残されたりすることがあり、これらが後段の音声認識に悪影響を与えることが指摘されている。

2.3 ターゲット話者抽出

ターゲット話者抽出（Target Speaker Extraction, TSE）は、混合音声の中から特定の話者の音声のみを選択的に分離・強調する技術である。DSE が距離という物理的指標を利用するのに対し、TSE は話者固有の声質や話し方などの「話者性」を利用する点に特徴がある。

TSE では、分離対象としたい話者の参照音声を事前に用意し、その音声から抽出した話者埋め込みをネットワークへの条件情報として与える。モデルは、この話者埋め込みをもとに、混合音声のスペクトログラム上でターゲット話者に対応する時間周波数ビンを通過させるマスクを推定する。これにより、特定の人物の音声を高精度に抽出することが可能となる。

しかし、TSE もまた根本的な制約を持つ。参照音声によって指定された話者のみが抽出対象となるため、事前に音声を登録していない未知話者の音声を抽出することができない。会議への新規参加者や、不特定多数が存在する環境での利用を想定した場合、この柔軟性の欠如は大きな課題となる。

3 提案手法

本節では、混合音声スペクトログラムから近距離話者の音声スペクトログラムを推定するための、提案手法の構造について述べる。提案モデルは、ResNet ベースのエンコーダ、GRU を用いたデコーダ、および wav2vec 2.0 によって抽出された話者特徴量を用いる構成となっている。図 3.1 に提案アーキテクチャの概要を示す。

本モデルでは、観測された混合信号のスペクトログラムを $X \in \mathbb{C}^{T\times F}$ とし、これが近距離話者の音声スペクトログラム X_{near} と遠距離話者の音声スペクトログラム X_{far} から構成されるものと仮定する。

$$X = X_{\text{near}} + X_{\text{far}} \tag{2}$$

本研究の目的は、混合信号 X から近距離音声スペクトログラム X_{near} を推定することであり、時間周波数マスク推定に基づく音声強調を行う。

3.1 ResNet による混合音特徴抽出

本モデルのエンコーダ部には、深層残差ネットワーク（ResNet）を採用する。ResNet は残差接続による勾配消失の抑制を通じて深いネットワーク構造を可能にし、通常の CNN よりも高い表現能力を持つ。

音声スペクトログラムは時間 × 周波数の 2 次元構造を持つため、画像と同様に畳み込み処理が適用できる。本研究では、入力スペクトログラム Y を ResNet に入力し、時間・周波数構造を保持したまま高次特徴量 $H \in \mathbb{R}^{C\times T\times F}$ を抽出する。ここで、全ての畳み込み層のストライドを 1 に設定し、画像認識用途で一般的なダウンサンプリングを行わない。これにより、時間方向・周波数方向の解像度を保ったまま、後段のマスク推定に必要な詳細な特徴表現を得ることができる。

3.2 GRU によるマスク推定

デコーダには、再帰型ニューラルネットワークの一種である GRU（Gated Recurrent Unit）を用いる。GRU は LSTM と比較してゲート構造が簡潔であり、計算効率に優れる一方、長期依存関係の獲得能力はほぼ同等とされているため、本研究の音声強調タスクに適している。

ResNet で得られた特徴量マップ $H \in \mathbb{R}^{C\times T\times F}$ を周波数ビンごとに $H_f \in C \times T$ に分割し、各ビンに対して独立な GRU を適用する。これにより、周波数ごとに時間方向の依存関係をモデル化できる。

GRU の出力は全結合層（FC）を通じて時間周波数マスク値へと変換される。具体的には、次の処理が各周波数ビンに対して独立に行われる。

1. 全結合層（入力次元: 30 → 出力次元: 16）
2. ReLU
3. 全結合層（入力次元: 16 → 出力次元: 1）
4. ReLU（マスク値を非負に制限）

これにより、各時間フレームに対して 1 次元のマスク値が生成され、全周波数ビンに対してまとめることで最終的な時間周波数マスク $M \in \mathbb{R}^{T\times F}$ が得られる。

3.3 wav2vec 2.0 による話者特徴量抽出

提案手法では、話者識別に有効であることが示されている wav2vec 2.0 を用いて、近距離話者の特徴量を取得する。wav2vec 2.0 に参照音声を入力し、その出力特徴マップを時間方向に平均することで、固定長の話者ベクトル $F_{\text{near}} \in \mathbb{R}^S$ を得る。本研究では特徴量次元として $S = 256$ を採用する。

図 3.2 に示すように、wav2vec 2.0 により抽出された特徴量は話者ごとに分布が明確に分離しており、未学習話者に対しても適切な話者特徴量が得られることが確認されている。このように、wav2vec 2.0 は従来手法では利用されていなかった新たな構成要素であり、提案手法における話者特徴量抽出を担う重要なブロックである。

4　評価実験

本節では、提案手法の評価のための実験について述べる。実験の目的は、提案手法が従来手法よりも優れた近距離音声抽出性能を実現できるかどうか、また、抽出された近距離音声が正しく音声認識されるかどうかを検証することである。近距離音抽出性能の評価には、音声強調でよく用いられる Scale-Invariant Signal-to-Distortion Ratio improvement (SI-SDRi) を、音声認識性能の評価には、認識結果の文字誤り率 (Character Error Rate, CER) を、それぞれ指標として用いた。音声認識には、End-to-End 音声処理ツールキットである ESPnet を、日本語話し言葉コーパス (CSJ) で訓練したモデルを用いた。

本実験では、話者特徴を用いる提案手法と、話者特徴を用いない石井らの距離ベース音声強調手法を比較した。石井らの手法は、提案手法から話者特徴抽出部分と FiLM による特徴統合部分を取り除いたものに相当する。

4.1　データセット

本実験における各手法の訓練および評価のために構築した、摸擬対話音声データセットについて述べる。

摸擬対話音声は、近距離音声、遠距離音声、背景雑音の 3 つの信号の組からなる。二人の話者が対話している状況を想定し、近距離音声と遠距離音声は、それぞれの話者が交互に発話するように構成する。実際の対話ではしばしば発話のオーバーラップが生じるため、オーバーラップ割合を定め、対話全体の長さに対してその割合の区間が、近距離音声と遠距離音声がオーバーラップするように構成する。

具体的には以下の手順で近距離音声と遠距離音声をそれぞれ構築した。

1. 無残響環境で収録された音声コーパスから、それぞれ異なる話者の音声を 2 つ抽出する。
2. 各音声を、音声コーパスに付与された区間情報に基づいて、数秒程度の発話区間に分割する。
3. 時間軸に各話者の発話区間を交互に配置し、対話音声を構築する。この際、事前に定めたオーバーラップ割合に基づいて、発話の一部がオーバーラップするように、区間の開始・終了時刻を調整する。

背景雑音は、事前に用意した背景雑音データベースからランダムに選択したものを組み合わせて用いた。

近距離音声と遠距離音声の作成に用いたインパルス応答について述べる。会議や打ち合わせでの使用を想定した評価を行うため、大きさが 3m x 6m、幅 150 cm の会議机が 5 台設置された会議室でインパルス応答を収録した。一般的なオフィス環境であり、防音設備はないものの、室内にはエアコンや OA 機器等の雑音源が存在する。

この会議室で、2 名が別々の机に着席して打ち合わせを行っている場面を想定した。この設定を模倣するため、以下のように機材を配置した。

- マイクロホン（スマートフォン）
  - 使用機種：iPhone XR, iPhone SE2, iPhone SE3 のいずれか
  - 机上、ノート PC の左側に設置
- スピーカー
  - 使用機種：Genelec 8020D スタジオ・モニター
  - 机上 40cm の高さ、机の手前側にスタンドで設置
- ノート PC
  - 机上、スピーカーの正面に設置
  - 現実の会議における音の反射・回折の効果を再現する目的で、ディスプレイを開いた状態と閉じた状態に設定

この条件に基づき、合計で 640 個のインパルス応答を収録した。

混合音声は以下のように生成した。

1. オーバーラップ割合に基づいて構成した 2 つの対話音声に対して、一方に近距離インパルス応答を、もう一方に遠距離インパルス応答を畳み込み、近距離音声と遠距離音声を作成する。この近距離音声が出力信号となる。
2. 雑音データベースから雑音を抽出し、近距離音声、遠距離音声、雑音を加算した混合音を作成する。この混合音が入力信号となる。
3. 近距離音声を話者特徴抽出器に入力し、話者特徴ベクトルを生成する。

上記の手順に基づいて、訓練データセットと評価データセットを作成した。摸擬対話音声作成のためのデータセットには日本語話し言葉コーパス CSJ を用いた。訓練データセットと評価データセットの分離のため、eval2 サブセットを評価データセットに、それ以外のサブセットを訓練データセットに用いた。畳み込むインパルス応答も同様に、訓練データセット用と評価データセット用のものが重複しないように分割した。訓練

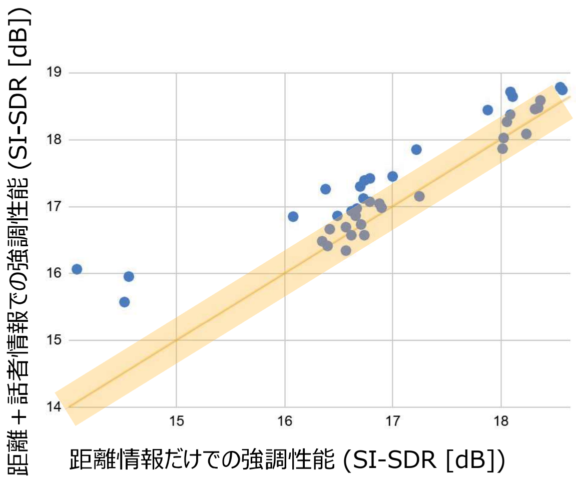

図 1: 音声強調性能評価結果。各点は、各評価データを提案手法と比較手法とで強調した結果の SI-SDR を示す。対角線は提案手法と比較手法の強調性能が同等であるような領域であり、この対角線よりも上方向に点が存在すれば、提案手法で音声強調性能が向上したことを示している。

データセットとして長さ 30 秒の摸擬対話音声を 1000 サンプル、評価データセットとして長さ 5 分の摸擬対話音声を 5 サンプル生成して用いた。

4.2 実験結果

図 1 に音声強調性能評価結果を示す。各点は、各評価データを提案手法と比較手法とで強調した結果の SI-SDR を示す。図中の対角線は、提案手法の音声強調性能（縦軸）と比較手法の強調性能（横軸）とが同等であるような領域であり、この対角線よりも上方向に点が存在すれば、提案手法で音声強調性能が向上したことを示している。すべての評価データが、対角線よりも上の領域にプロットされていることから、話者特徴の導入は音声強調性能の改善に寄与することが分かる。図の左側に対角線より上方にプロットされた点がいくつかあることから、話者情報なしでは音声強調性能が低いような評価データに対しては、話者情報の導入により大きく性能向上が可能であることが示唆される。図 2 に音声認識性能評価結果を示す。強調性能の評価結果と同様に、話者特徴の導入により、音声強調性能も改善していることが分かる。

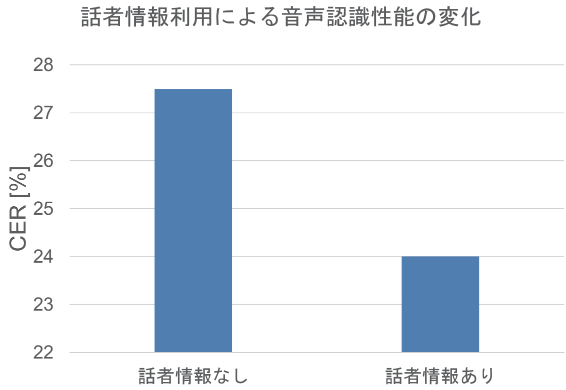

図 2: 音声認識性能評価結果

5 まとめ

謝辞

□□....

参考文献