4Q-01

授業グループワークの音声認識精度改善のための マルチチャネルＶＡＤ

中野 魁人† 中山 隆弘‡ 白水 始‡ 市川 治†
滋賀大学データサイエンス学部† 東京大学高大接続研究開発センター‡

１ はじめに

新しい教育手法として生徒が自ら考え学ぶというアクティブラーニングが注目されている．東京大学CoREFが推進する「知識構成型ジグソー法」では，クラス内の生徒を3～4名の小グループに分け，グループ内外で話し合うことで発見的に学びを進める．この学びの過程を可視化するために，生徒ひとりひとりに接話マイクロフォンを装着してもらい，音声データを音声認識によりテキスト化することが行われる．この記録を教師が授業後に見ることで，生徒が理解を深める過程を追跡することができる．

このような目的では，隣接話者の音声や教室の雑音をなるべく拾わないようにするために，接話マイクロフォンを使用することが多い．しかし，雑音に強いとされる接話マイクロフォンをもってしても，出力の音声データに，隣接する生徒の発声が誤って混入してしまうことが少なからずあり，問題となっている．

２ 従来手法

従来は，隣接話者の音声を排除するために，対象話者のシングルチャネルの VAD (Voice Activity Detection)処理を施すことが行われてきた．これには，パワーベースの方式と，モデルベースの方式があるが，いずれにしろ，発話区間推定の性能は閾値の設定に大きく依存し，隣接話者の音声が対象話者の音声と誤って判断されることが頻発する．

一方，単独のマイクロフォンの出力のみを使うのではなく，近辺で使われている複数のマイクロフォンの出力を総合的に利用して，対象話者の音声のみを取り出すという分散マイクロフォンアレイの手法も存在する[1]．この手法では，テーブルに分散的に配置した遠隔マイクロフォンを使用し， NMF 法により対象話者の音声のみを抽出する．この方式は，VAD のようなざっくりとした手法ではなく，基底ベクトルと活性化ベクトルを推定する精密な論理に基づいた手法なので，話者とマイクロフォンの固定的な位置関係に依存しており，今回対象とするような，接話マイクロフォンをつけたまま，いろいろな隣接話者に顔を向けるケースに適用できない．

また，咽喉部に装着した特別なマイクロフォンを併用するという手法も存在する[2]．喉の震えを検知することにより，対象話者の発声か隣接話者の発声かを区別することができる．しかし，特別なマイクロフォンを使用するという追加の機器コストと，喉元にマイクロフォンを貼るという心理的・肉体的な負担が実用化を妨げている．

３ 提案手法

提案法は，基本的にパワーベースの VAD であるが，グループ内の複数の接話マイクロフォンの出力を総合して，対象話者の発声区間の判定を行う．

図１に従来のパワーベースのシングルチャネル VAD による発声区間の判定ロジックを示す．横軸は対数の音声パワーである．ただし，背景雑音のパワーからの上昇分（以下 Local SNR と呼ぶ）を表示している．すなわち，雑音の準定常的な成分を背景雑音とみなし，その音声パワーをゼロ，すなわち原点にとっている．対象話者や隣接話者の発声や突発性の雑音があれば，横軸の右側に値がシフトする．十分に音声パワーが高くなり，閾値 A を超えたときに，対象話者の発声区間として判定される．

いま，別の閾値 B を閾値 A よりも大きな値として設定する．閾値 B を超えた音声は，十分な確信をもって対象話者の発声区間として判定されるべきであろう．一方で，閾値 A と閾値 B の間（以下，中間領域と呼ぶ）は，対象話者の発声区間であるかもしれないが，隣接話者の発声区間である可能性も十分に高い．提案法では，この中間領域の判定をより正確に行うために，対象話者のマイクロフォン出力だけでなく，複数の隣接話者のマイクロフォン出力も判断材料にする．例えば，4 人のグループで討議を行っていたとすれば，3 個の隣接話者のマイクロフォン出力が存在するので，順番に一つずつ選び，3通りのペアの比較を行う．

Multi-channel VAD for improved transcription of group work in active learning classroom.
Kaito NAKANO†, Takahiro NAKAYAMA‡, Hajime SHIROUZU‡, Osamu ICHIKAWA†
† Faculty of Data Science, Shiga University
‡ The University of Tokyo - Center for Research and Development on Transition from Secondary to Higher Education

図2に提案法による発声区間判定の概念を示す．横軸は対象話者，縦軸は隣接話者のマイクロフォンの対数音声パワー（Local SNR）に相当する．ここでは，中間領域に斜めの直線を配置した．音声が対象話者からの発声であれば，そのフレームの音声のパワーは，この斜めの線よりも，下側に位置することが期待できる．また，音声パワーが，この斜めの線よりも，上側に位置した場合には，縦軸に対応する隣接話者からの発声であることが期待できる．

この判定ロジックは隣接話者の数だけ実行されるが，一つのケースでも隣接話者からの発声という判定となった場合には，対象話者からの発声ではない，と判断する．すなわち対象話者のマイクロフォンは「全勝」しなければならない．

この斜めの直線，すなわち決定境界は，次のように求められる．話者①と話者②のペアのケースであれば，従来技術であるシングルチャネル VAD を使用して，話者①の発声区間と話者②の発声区間を求めておく．もちろんこれは正確ではなく，多くのフレームが両方の発声区間に分類されてしまうが，第1次近似として，それを採用する．提案法を適用して，この第1次近似をさらに改善した第2次近似を作成し，再び提案法を適用するという繰り返し処理を行う．図2を使用して説明すると，図中の赤丸と藍丸が，繰り返し処理のその時点で判定されている話者①の発声と話者②の発声のフレーム（点）ということになる．決定境界の直線は，これらを正解ラベル付きの学習データとみなし，機械学習により求められる．正規分布や SVM などの一般的な分類問題の手法を適用すればよい．簡便な例としては，それぞれのグループの要素の重心を求め，2つの重心を結んだ線分の中間位置を通り，その線分と直交する直線を決定境界とすればよい．

ここまで説明したロジックは音声フレーム単位の判定である．通常の VAD と同様に，時間方向の連続性の調整を行い，発話区間が最終決定される．

また，生の対数音声パワーを用いるのではなく，Local SNR を使用する点が大変重要である．背景雑音のパワーを基準にすることで，マイクロフォンごとのゲインの違いが補正される．

4　評価実験

実際の中学生を対象に，数学，理科，国語の3教科について，グループワーク（ジグソー法）を1回ずつ行った．その1回の中に Expert 活動と Jigsaw 活動という2セッションのグループワークがある．グループワークは，1チーム3～4名の6～7チーム構成である．生徒ひとりひとりに装着したマイクロフォンの出力を

1. VAD 無し（ベースライン）
2. シングルチャネル VAD（従来法）有り
3. マルチチャネル VAD（提案法）有り

の3種類で，前処理を行い，音声認識を実行し，その認識テキストを正解のテキストと比較することで，音声認識の文字誤り率を測定した．誤り率が小さい手法が良い手法である．ただし，本報告の実験で用いた提案法は「繰り返し処理無し」「重心による分離」であるため，性能は改善の余地がある．

表1に実験結果を示す．提案法(mvad15) は，従来法(svad15) と比較して，文字誤り率が顕著に下がった．内訳としては，挿入誤りが激減し，置換誤りも減少し，削除誤りが増加した．従来法(svad15) は，ベースライン（none）よりは文字誤り率が低いものの，提案法には及ばない．

5　おわりに

本報告では，グループ内の全員の音声トラックを参照することで，対象となる話者の発話区間をより高精度に推定するマルチチャネル VAD を提案し，音声認識の湧き出し誤りが効果的に減少することを示した．

謝辞

本研究は科研費（17H06107）の助成を受けた．

参考文献

[1] 小野，AI チャレンジ研究会，pp.33-38，2014
[2] 大高他，FIT2016 予稿集,vol.2,pp.149-150,2016

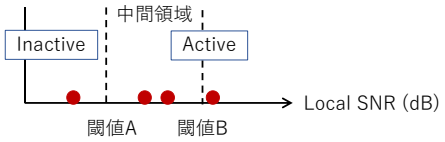

図１　シングルチャネル VAD

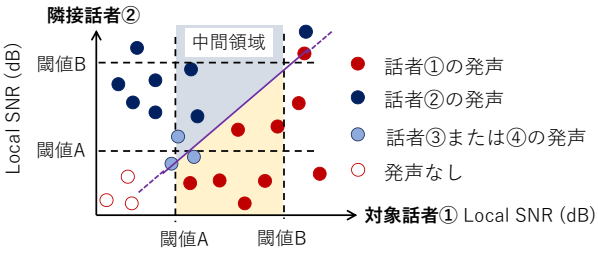

図２ マルチチャネル VAD の概念

表１　音声認識誤り率（%）

		挿入誤り率	削除誤り率	置換誤り率	文字誤り率
Exp	none	69.0	12.5	39.6	121.1
	svad15	34.9	19.5	34.1	88.5
	mvad15	8.8	25.4	29.9	64.1
Jig	none	66.8	12.1	38.5	117.3
	svad15	27.6	19.2	33.1	79.8
	mvad15	6.4	24.8	28.7	59.9