4ZA-05

Gesture-To-Talk による サイネージ端末用の音声区間検出の改善

藤岡　和紘†　　小松　美幸†　市川　治†

†滋賀大学データサイエンス学部

1　はじめに

大学に訪れる人々を利用者として想定し，質問の音声を入力とし，回答を合成音声で返す質問応答のサイネージ端末を作成した．

しかし音声を使用した通常の音声区間検出(VAD)だけでは多数の人々が来場する環境での雑踏の音声にサイネージ端末が反応してしまい，ひとりでに話し始めてしまうことが問題となっている．

2　従来手法

従来の手法では，発話者の音声を正確に取得するためにシングルチャンネルの VAD(Voice Activity Detection)処理を施すことが行われてきた．VAD には，音声のパワーを基準に判定する Power ベースの VAD と，GMM などの音声モデルの尤度を使用する Model ベースの VAD がある．

Power VAD は入力音声が対象の発話のみであれば上手く動作するが，他者の発話や環境音などが混在してしまうと，音声区間の始端・終端を誤検出したり，環境音を音声区間であると誤って判定したりしてしまうという問題があった．そのため，この手法は雑音の多い環境では上手く動作しない傾向にある．近年，スマートフォンなどの音声認識では，特定の単語を認識した時点を音声区間の始端とし，音声の終端をパワーなどにより終端として判定する Attention-word を用いた VAD が主流となっている．しかし，公共の場に設置されるサイネージ端末では，この方式も Attention-word の認識と発話終端の判定に雑音の影響を多く受けてしまう．

確実な音声区間を取得するためには，音声以外のキューを利用することが有効である．タッチパネルなどでトークボタンを実装できるのであれば，発声している間ボタンを押し続ける Push-To-Talk (PTT) 方式が，最も有効である．次善の策は発話の開始のみをボタンで知らせる Push-To-Activate (PTA) である．しかし，コロナ禍の現在，不特定の人が接触するトークボタンの実装は推奨されない．

Gesture-to-talk for Voice Activity Detection of public service machine.

Kazuhiro FUJIOKA†, Miyuki KOMATSU†, Osamu ICHIKAWA†

† Faculty of Data Science, Shiga University

その他の有効なキューとしては Visual cue がある．カメラを用いて話しかける姿勢[1]や口唇画像を認識するものである．しかし，話しかける姿勢だけでは十分な判定精度は期待できず，また，口唇画像の利用は，コロナ禍でマスク装着が当たり前になっている現在の状況では，正常な動作を期待できない．

3　提案手法

コロナ禍の時代に適合した音声認識インターフェイスとして，特定のジェスチャーをトークボタンの代わりとして使用する Gesture VAD を提案する．トークボタンの PTT 方式と同様に，ユーザーがそのジェスチャーを行っている間だけを発話区間とみなす仕組みを Gesture-To-Talk (GTT) 方式と呼ぶことにする．また，PTA 方式と同様に，開始のキューだけをジェスチャーで与える仕組みを Gesture-To-Active (GTA) 方式と呼ぶ．

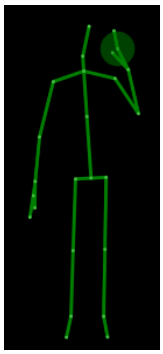

図１　規定ジェスチャー

本報告では，図１のように発話のキューとして，開いた手を口元にあてて人に話しかける形のポーズを規定のジェスチャーとして定義する．今回の実装では，骨格座標の取得に Kinect V2 センサを使用した．今回は簡便な方法として，発話者の右手が体の中心より右側に存在し，右肩の高さを超した時点を音声区間の始端とした．音声区間の終端は，GTT 方式では骨格座標が上記の条件を満たせなくなった時点とし，GTA 方式では Power VAD により与える．

GTT 方式では，発話者は発話をしている間ずっとジェスチャーを維持していなければならないが，GTA 方式では，発話の開始時のみジェスチャーを行えばよく，発話者の負担は少ない．

4　評価実験

複数人の大学生に受験時の疑問に関する調査を行い，1293 文の質問文を作成した．その質問文からランダムに 50 文抽出し，6 名の被験者(男性 4 名，女性 2 名)に 2 m 離れたマイクに向かっ

て質問文の内容を発話してもらった．ただし，冒頭に Attention word を付加している．収録は，16 kHz のサンプリングレートで行った．この雑音なし音声データを Clean data と呼ぶ．また，同時に Kinect V2 を使用して被験者の骨格座標情報を取得した．その後，Clean data を適切な閾値の Power VAD にかけ音声区間の始端と終端の情報を取得した．Power VAD の閾値は最適なものを選択している．この音声区間情報は，後述の実験のための正解データとみなすことができる．次に，雑踏の状況を模擬するために，トピックに無関係な発声を重畳した．雑音源の発声データは CENSREC-3[2]のアイドリング時の女性の遠隔マイク発声である．雑音重畳の目標 SN 比は 10 dB とした．この雑音あり音声データをテストデータ (Noisy data) と呼ぶ．

4.1 音声区間推定精度実験

テストデータ(Noisy data)に対して以下の VAD を行い，音声区間を推定した．

1. Power VAD
2. GTT 方式
3. GTA 方式
4. Attention-word 方式

表１に雑音環境下での音声区間の誤り率を示す．Ins， Del はそれぞれフレーム単位での音声区間の挿入誤り率と削除誤り率を算出している．Err は２つの合計値で音声区間の誤り率を算出し，この値が小さいほど良い手法であるといえる．提案法(GTT)は，従来法(Attention-word)に比べ音声区間誤り率が減少した．内訳としては挿入誤り率が増加し，削除誤り率が減少した．従来法は Attention-Word 自身の認識率が低く、Power VAD によって音声区間の終端をとっているため，雑音 の影響で削除誤り率が顕著に高い．

4.2 音声認識実験

4.1 の 4 種の手法で推定された音声区間を用いて，テストデータ(Noisy data)を切り出し，音声認識を行った．

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．誤り率が低いものが良い手法である．音声認識システムは IBM Watson Speech-To-Text を使用した．ただし，言語モデルは，Attention-word を含む質問文テキストによって適応化している．

表 2 に音声認識の文字誤り率を示す．Ins ，Del ，Sub はそれぞれ文字単位での挿入誤り率，削除誤り率，置換誤り率を表している．そして Err がそれらを合計した文字誤り率を表している．この値が小さいほど良い手法であるといえる．また比較のための参考値として，正解とみなされる音声区間で切り出しを行い，音声認識を実行した．表２の Clean は Clean data について行った結果，Oracle は Noisy data について行った結果である．

提案法(GTT)は，従来法(Attention-word)に比べ文字誤り率が顕著に低く，Oracle ケースとほぼ同じ精度に至った．内訳としては，提案法は従来法と比べ削除誤り率が大幅に減少した．また，GTA は従来手法と比べ文字誤り率は低いが提案手法には及ばなかった．これは音声区間の終端を Power VAD でとっているため，雑音の影響を強く受けたためだと考えられる．

５　おわりに

本報告では，ジェスチャーを用いた VAD によって，対象話者以外の発声がある環境下での音声認識が，現在主流の Attention-word 方式よりも改善することを示した．また，コロナ禍の状況に適したインターフェイスとなっている．

謝辞

サイネージ端末用の対話システムは名工大の MMD-Agent を用いて作成した．本研究は科研費（19K02999）の助成を受けた．

表１　雑音環境下での音声区間誤り率
(話者ごとの平均)

	Err	Ins	Del
Power VAD	129.1%	129.1%	0.1%
GTT	23.7%	23.3%	0.4%
GTA	30.0%	14.3%	15.7%
Attention-word	70.8%	3.8%	67.0%

表２　音声認識文字誤り率(話者ごとの平均)

	Err	Ins	Del	Sub
Power VAD	49.1%	19.4%	13.0%	16.7%
GTT	25.4%	1.1%	19.7%	4.6%
GTA	51.1%	1.2%	45.0%	5.0%
Attention-word	61.4%	0.8%	58.1%	2.6%
Oracle	26.8%	0.4%	22.8%	3.5%
Clean	17.2%	0.2%	14.9%	2.2%

参考文献

[1] M. Cristani *et al.*, Workshop on Interactive Human Behavior Analysis in Open or Public Spaces, 2011

[2] M.Fujimoto *et al.*, IEICE Tras., Vol.E89-D, No.11, pp. 2783-2793, 2006