発話を先読みしマイクミュート制御をする マルチモーダル発話検知技術

山田 仰[1] 瀧上 順也[1] 仲 信彦[1] 吉村 健[1] 太田 賢[1]

概要： 本研究では，音声通話における自動マイクミュート制御のために，従来技術を用いた場合の２つの課題である，話頭切れと周囲雑音の誤検知を防止するため，口唇の変動に基づき発話開始を先読みしてマイクを有効化し，口唇の変動と音声信号に基づき非発話中と推定されたときにマイクを無効化するミュート制御手法を提案する．また本研究では，共話を行っているオンラインコミュニケーションの収録動画を用いて提案技術の評価を行い，収録動画の全発話の内の 99.1%の発話を話頭切れを含む発話区間の欠損無く検出でき，共話のユースケースでも十分に実用性があることを示す．

Multi-modal Voice Activity Detection Using Utterance Prediction for Mute Control

Aogu YAMADA[1] Junya TAKIUE[1] Nobuhiko NAKA[1] Takeshi YOSHIMURA[1] Ken OHTA[1]

1. はじめに

現在の Web 会議では，発話時だけマイクが有効になるようにマイクの有効化と無効化を手動で切り替えることが通例となっており，共話 [1][2] と呼ばれるコミュニケーションを阻害する要因の一つとなっている．共話とは，会話中の一つの発話を必ずしも一人の話し手が完結させるのでなく，複数人で発話の積極的なオーバラップをしながら作っていくコミュニケーションの形態であり，会話の参加者が互いの信頼感を醸成し，コミュニケーションのアウトプットの質を高める効果が期待される．発話の積極的なオーバラップのためマイクを常時有効にしておくことが効果的であるが，そのためには周囲雑音への対策が必須である．

一般的に，ノイズキャンセラと有音無音検出（Voice Activity Detection）の組み合わせでは特に人間の音声に似た雑音の除去が困難である [3]．このような雑音に対しても話者の発話を検出する方法として，カメラ画像から取得された話者の口唇の変動パターンとマイクから取得された音声パターンに基づいて発話区間を推定するマルチモーダルな発話検知手法が提案されている [4][5]．しかしながら，従来の手法では発話中の音声と口唇の変動パターンを合わせた分析により発話検知を行うため，発話の検出は発話開始からある程度時間が経過した後にしかできず，音声の話頭切れが生じることが課題である．話頭切れを防ぐためにも，音声バッファリングすれば，遅延が増大する．

そこで本研究では，口唇の変動に基づき発話開始を先読みすることで発話の頭切れを避け，口唇の変動と音声信号に基づき非発話の状態を推定することで，周囲雑音の混入を防止する自動マイクミュート制御手法を提案する．提案手法では，マイクの有効化判断のために，発話者の発声時において音の発生よりも僅かに先行して生じる発話の予備動作としての口唇の変動（具体的には 1. 発話前の呼吸に基づくもの 2. 子音の発声のために予め唇の形を作る動きに基づくものなど）の検知を行う．さらに発話を伴わない口唇の動きに基づいてマイクが一度有効化された場合に不必要にマイクが有効化されてしまう時間を短縮するために，口唇の変動パターンに基づいて即時的に発話が開始される可能性が高いか判定を行う．即時的に発話が開始される可能性が高いにも関わらず，マイクから音声信号が一定時間取得されない場合に非発話中と推定し，マイクの無効化を行う．

本研究による貢献は以下の３点である．

- 発話開始の予備動作を検知し前もってミュート解除す

[1] NTT ドコモ サービスイノベーション部

るため，話頭切れなくリアルタイムでの音声通話を可能とする．

- 発話開始の判断のために音声バッファリングが不要なため，提案するミュート制御手法によって遅延が発生しない．
- 音声情報のみならず口唇情報を踏まえて非発話状態を推定しアンミュートを解除するため，周囲雑音に頑健な音声通話を可能とする．

2. 関連技術

雑音環境に頑健な発話検出手法として，カメラ画像から取得された話者の口唇の変動パターンとマイクから取得された音声パターンに基づいて発話区間を推定するマルチモーダルな発話検知手法が提案されている [4][5]．

文献 [4] では，発話する音声を含む外部音から音声特徴量を抽出して音声発話確率を出力するとともに，発話者の口の動きを含む口唇画像から口唇特徴量を抽出して画像発話確率を出力する．そして，これらの音声確率及び発話確率を統合した結果から，発話区間を検知する．しかしこの手法を音声通話に用いる場合，音声情報と映像情報をある程度取得した後に発話開始の検知を行うことになるため，発話検知の判定時間待ちによる遅延増加より，音声通話のリアルタイム性が損なわれる．

文献 [5] では，リアルタイムな発話検知を行うために，音声情報と口唇情報を入力とした時系列ニューラルネットワークモデルによるマルチモーダルな発話検知手法が提案されている．しかしながら，実際の音声通話におけるミュート制御にかかる処理時間等を考慮すると，ある程度前もって発話開始を先読みし，ミュートを解除することが望ましい．

3. 提案手法

本研究では，オンラインコミュニケーションにおいて，口唇の変動に基づき発話開始を先読みすることで，従来技術の課題であった発話の頭切れを避けるミュート制御技術を提案する．本技術は図 1 の構成のように，PC やスマートフォンのマイクとカメラから取得したユーザの音声と映像を入力としてユーザの発話音声を出力とする機能により実現される．提案技術のマルチモーダル発話検知機能は，図 2 のように，以下の 4 つの機能ブロックにより構成される．

(1) ユーザの顔を撮像した顔映像に基づいて発話予備動作を検知する口唇動作検知部
(2) マイクから取得されるユーザの音声信号に基づいて、音声を検出する音声検出部
(3) 口唇動作検知部と音声検知部からの検知信号に基づいて，音声信号を有効化または無効化することを判定する発話判定部
(4) 音声信号の有効化及び無効化を制御する音声制御部

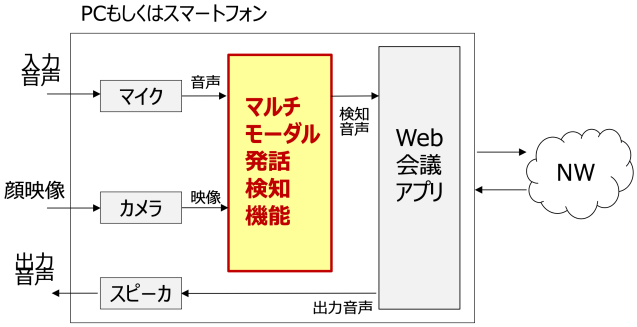

図 1 提案するマルチモーダル発話検知技術によるコミュニケーションシステム

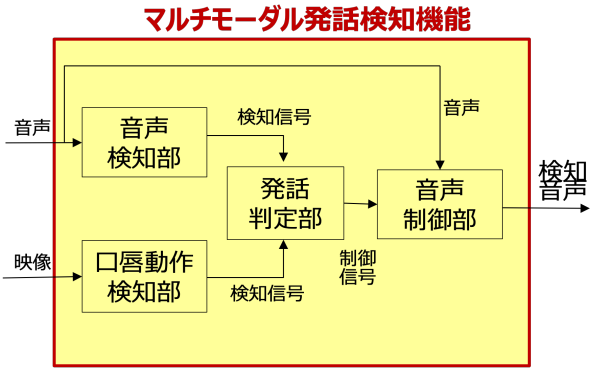

図 2 提案するマルチモーダル発話検知機能のブロック図．

上記の構成により，発話時において口唇の変動が発話よりも僅かに先行して生じるところ、顔画像に基づいて口唇の動きとして規定された発話予備動作が検知された場合に音声信号が有効化されるので、冒頭部分が切れることなく発話音声を取得できる。さらに発話検知部では，不必要にマイクが有効化されてしまう時間を短縮するために，音声情報と口唇情報に基づいて非発話状態の推定を行う．これにより発話を伴わない口唇の動きに基づいて音声信号が一旦有効化された場合であっても、非発話状態推定時にマイク入力が無効化されるので、ハウリング及びエコーの発生並びに雑音の混入等が防止される．

3.1 口唇動作検知

口唇動作検知は，ユーザの顔を撮像した顔画像に基づいて、発話時の口唇の動きとして予め規定された発話予備動作を検知する [6]．発話予備動作の検知のための顔画像の例を図 3 に示す．口唇動作検知は，まずユーザの顔画像 f から口唇部分 lp を抽出し，抽出した口唇部分 lp の上唇（口唇部分 lp における上方向座標が最も大きい点）及び下唇（口唇部分 lp における上方向座標が最も小さい点）に標識点 mu, ml を取得すると同時に，顔画像 f から顔の長さ fl をさらに取得する．顔の長さ fl に対する標識点 mu と標識点 ml とを結んだ線分の長さの割合が予め設定された閾値 θ_1 を超えた場合，もしくはこの割合の単位時間あたりの変動量が閾値 θ_2 を超えた場合を発話動作と判断する．こ

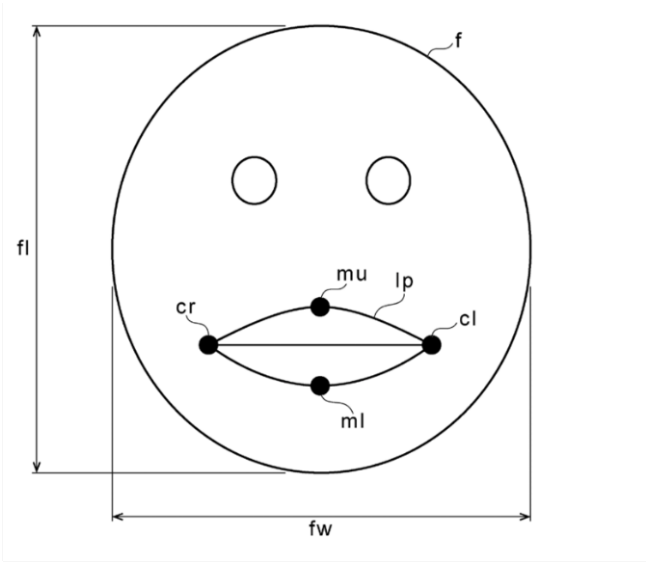

図 3　発話動作の検知に用いる顔画像パラメータ.

れにより口唇部分の大きさまたはその変動量が所定の量以上である場合には発話状態である可能性が高いことに鑑みて，発話予備動作を適切に検知できる.

3.2　音声の検出

マイクより取得された音声信号の振幅がしきい値 θ_3 以上であるとき，音声が検出されたと判定する.

3.3　音声の有効化及び無効化の判定

発話予備動作が検知された場合，音声信号が有効化される．音声信号の有効化後，予め設定された無効化判定時間 T 内に音声が検出されなかった場合に音声信号を無効化される.

図 4 に，3.3 節のフローチャートを示す．発話検出処理が開始されると，カメラによる画像取得処理，マイクによる音声の取得処理が，継続的に実施される.

- ステップ S1 において，音声信号が有効化されているか否か（マイクがアンミュートかミュートか）を判定し場合分けに応じて処理を行う.
- ステップＳ 2 において，口唇動作検知からの結果に基づき，発話動作が検知されたか否かを判定する．発話動作が検知されない場合はミュート状態を維持し，検知された場合は続いてのステップＳ 3 において，マイクを有効化（アンミュート）する.
- ステップＳ 4 において，無効化判定時間 T の時間だけ待機し，続いてのステップＳ 5 において，この T の間に音声が検出されたか判定を行う．音声が検出された場合は，発話予備動作に引き続いて実際にユーザが発話した場合に該当するため引き続きマイクを有効化する．検出され場合は，発話動作が検知されたものの、その動作が実際の発話を伴わない場合等に該当するため引き続いてのステップ S6 にてマイクを無効化する.
- ステップＳ 7 において，口唇動作検知からの結果に基づき、発話予備動作が検知されたか否かを判定する．発話動作が検知されたと判定された場合には，ステップＳ 4 に進む．一方，発話動作が検知されたと判定されなかった場合には，ステップＳ 6 に進む。

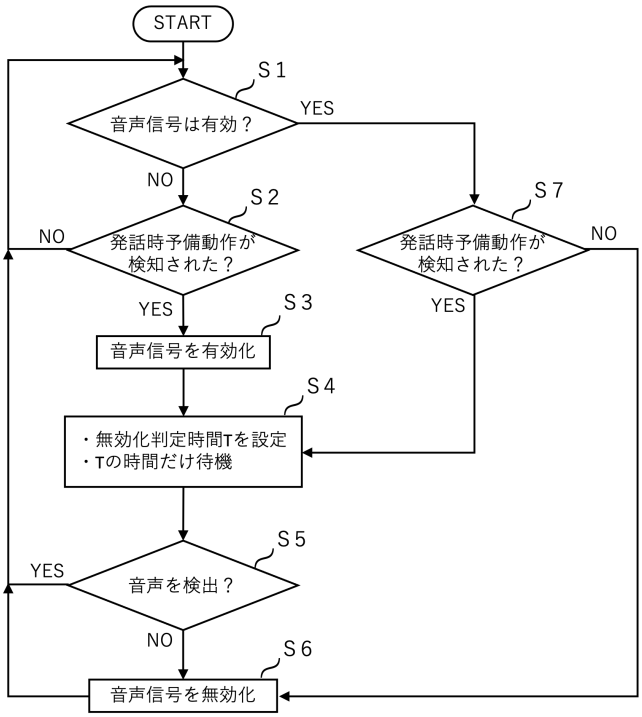

図 4　フローチャート

図 4 に示すフローチャートの動作によれば，音声検出部がユーザの発話ではない雑音等の音声が検出され続けている場合においても，口唇の変動が検出されない場合には音声が有効化されない，あるいは即時に無効化されるため，雑音などの音声が不必要に有効化された状態が維持されることを防止することができる.

3.4　音声制御部

音声制御部では，判定部による音声信号の有効化または無効化の判定に応じて，音声信号の有効化及び無効化を制御する．発話判定部により音声信号を有効化することが判定された場合，発話時には口唇の変動が発話よりも僅かに先行して生じるところ、顔画像に基づいて口唇の動きとして規定された発話予備動作が検知された場合に前もって音声信号が有効化されるため，相手方の通話システムに話頭切れが生じない．一方，音声信号のレベルを同時に監視しており，発話を伴わない口唇の動きがある場合に，音声を無効化することができる．これは，例えば話し終わりの部分で呼吸などにより，完全に口が閉じない場合にも音声を無効化でき，望まない音を通話相手に送信することを抑制できる.

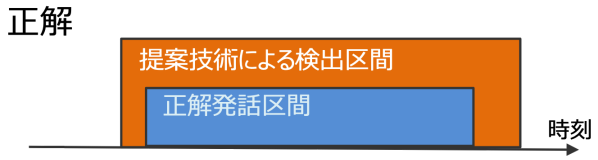

図 5　発話検知の正解例

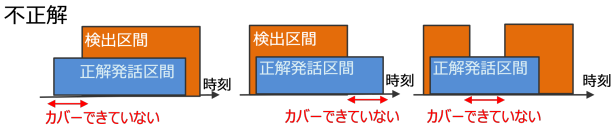

図 6　発話検知の正解でない例

4. シミュレーション評価

本研究では，共話を行っているオンラインコミュニケーションの収録動画データセットを用いたシミュレーションにより提案技術の評価を行う．

4.1 評価指標

本研究では，二つの指標を用いて提案する発話検知技術の性能を評価する．

4.1.1 発話無欠損検出率

一つ目の評価指標は，収録動画データセットの全発話の中で，欠損なく発話区間が検知された発話の割合を示す発話無欠損検出率である．本研究はオンラインコミュニケーションにおいて，話頭切れを含めて，欠損無く発話区間を検出し音声を通話相手に疎通することを目指すものであるため，この発話無欠損検出率を可能な限り 1 に近づけることを目指す．発話無欠損検出率 R_{correct} は以下の式 (1) で表される．

$$R_{\text{correct}} = \frac{N_{\text{correct}}}{N_{\text{utterance}}} \tag{1}$$

ただし $N_{\text{utterance}}$ は，収録動画データセットに含まれる全発話の数である．N_{correct} は，収録動画データセットに含まれる全発話の中で，欠損無く発話区間を検出された発話の数である．N_{correct} に含まれる発話の例を図 5 に示す．正解発話区間は，人手により収録動画の音声を聞いて記録した，各発話の発話開始時刻から終了時刻までの時間区間である．図 5 のように，発話検出区間が正解発話区間を漏れなく含んでいた場合を正しく検知できた発話として考える．逆に正解以外の例を図 6 に示す．

4.1.2 発話過剰検出率

二つ目の指標は，式 (2) で示される発話過剰検出率 R_{error} である．

$$R_{\text{error}} = \frac{T_{\text{detected}} - T_{\text{correct}}}{T_{\text{dataset}}} \tag{2}$$

ただし T_{dataset} は全収録動画データセットの再生時間，T_{detected} は全収録動画データセットの再生時間の内の発話区間と判定された時間の総和であり，T_{correct} は全収録動画データセットの再生時間の内の正解発話区間の総和である．発話過剰検出率 R_{error} は，不必要に発話区間として判定されてしまった時間区間の割合であり，可能な限り 0 に近い値であることが望ましい．

表 1　パラメータ

$N_{\text{utterance}}$	3107
T_{dataset}	450 minuets
θ_1	0.02
θ_2	0.01
T	2 seconds

4.2 評価データセット

本研究では，共話を行っているオンラインコミュニケーションの収録動画を用いて提案するマルチモーダル発話検知技術の評価を行った．収録では，全 15 人の話者が 3 名ずつの合計 5 グループに分かれ，各グループごと 3 人の話者が別室でオンラインコミュニケーションツールを通じて 7〜8 分の模擬会話を 4 回（4 タスク分）行った．4 タスクの内訳は，ディベート 1 タスク，ブレーンストリーミング 1 タスク，会話を通じてあらかじめ決められた言葉を相手に言わせる NG ワードゲーム [7]1 タスク， 会話を通じて少数派を見つけ出すワードウルフ [8] がそれぞれ 1 タスクずつである．各タスクが共話となることを目的とし，複数人で発話の積極的なオーバラップが生じるよう下記のようなインストラクションの下，模擬会話を行なった．

- ディベートでは，一人の話者がテーマに沿った主張を行う中，二人目の話者がその主張の発話に反論の割り込みを行い，3 人目の話者が一人目と二人目の発話に対し補足や言い換え等の割り込みを行うインストラクションのもと模擬会話を実施．
- ブレストでは，全話者 3 名に予め定められたテーマに沿って制限時間内に五月雨的にできるだけ多くのアイディアを上げ，相互の発話に対し補足や言い換え等の割り込みを行うインストラクションのもと模擬会話を実施．
- NG ワードゲーム [7]，ワードウルフ [8] では，可能な限り他の話者への相槌やリアクションを積極的に行うインストラクションのもと模擬会話を実施

4.3 パラメータ

データセット及び提案手法の設定パラメータを表 1 に示す．

4.4 評価結果

提案するマルチモーダル発話検知技術の評価結果を表 2 に示す．

発話無欠損検出率は 99.1%であり，3107 発話中 3079 発

表 2　評価結果

正解無欠損検出率	過剰検知率
99.1%	47.5%

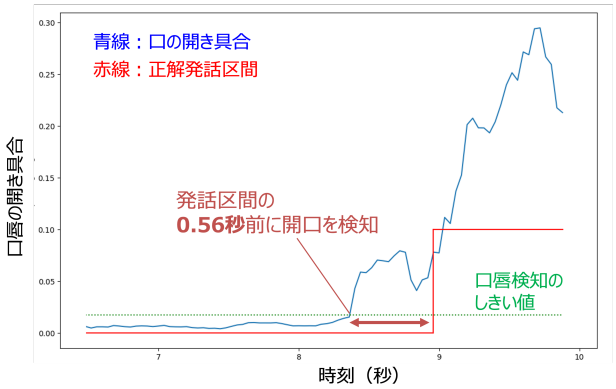

図 7　発話検知の例

話を発話区間の抜け漏れなく正しく検知できた．正しく検出ができなかった 28 発話は，目視で確認した限りいずれも口が開いていないように見えた相槌の発話であり，以下の通りであった．

- 相槌「うん」が 20 発話
- 笑い「ふふっ」が 4 発話
- 唸り声「んー」が 4 発話

相槌以外の発話が全て正しく検知でき，全発話の 4 割程度を占めた相槌の内の大多数が正しく検知できたため，提案技術は共話のユースケースでも十分に実用性があると言える．

図 7 に発話検知の一例を示す．発話の予備動作としての口唇の変動を検知することで，図 7 のように発話による音の発生に先立って発話開始を検知できた．発話の予備動作としては，以下の動作が観察された．

- 発話前の一呼吸
- 子音の発声のために，唇の形を作る動作
- 言葉が見つかり次第，喋ろうと口が開く動作
- 相手の発話中に喋ろうとして，間を掴むときに口が開く動作

過剰検知率は 47.5%であり，これは無音時間区間（正解発話区間ではない時間区間）の半分近くが，本技術によりミュート解除されるべきと判定された結果となった．この 47.5%の内訳は下記となる．

- 27.5%分は，上記で述べた発話の先読みのための予備動作によって検知された時間区間に該当した．口が動いているが発話音声が出ていない状況であるため，現状の提案技術の発話先読み検知の処理において，口唇情報に加えて音声情報を考慮することによるこの時間区間の短縮が課題である．また口唇の開き具合のパラメータ θ_1，θ_2 の個人適応についても今後の課題である．
- 20%分は，正解発話区間終了後の発話検知区間であり，発話動作検知後の無効化判定時間 T にの時間区間に該当した．この T は，会話において他の人の発話が終わったと判断するだけの間の開き方であり，今回の評価での T の値は発話無欠損検出率を可能な限り 1 に近づけるために，十分大きな値として $T = 2s$ を設定した．T の個人適応によって小さな値に抑えることが今後の課題である．
- 口が開いておらず，かつ口が動いていないが，発話区間と誤検出した割合が 1 %未満あった．これは口唇動作検出の技術的課題であるが割合は少なかった．具体的には，30 度以上のお辞儀して口がうまく映らない場合や，唇を舐める場合が観察された．

5. おわりに

本研究では，共話を行っているオンラインコミュニケーションの収録動画を用いて提案技術の評価を行った．収録動画の全発話の内，正解発話区間を漏れ（頭切れを含む）なく発話区間として判定できた発話の割合は 99.1%であった．相槌以外の発話が全て正しく検知でき，相槌についても大多数が正しく検知できたため，提案技術は共話のユースケースでも十分に実用性があると言える．また無音区間にも関わらず発話状態と判定してしまった時間区間の割合は 47.5%であり，この値を可能な限り低減するために，口が動いているが発話音声が出ていない時間区間の発話誤検出の低減が今後の課題である．また各種パラメータの個人適応等による発話開始前の過剰な発話検知区間と発話終了後の過剰な発話検知区間の低減が今後の課題である．

本研究の発表では，提案する技術を用いたミュート制御の有効性の体験のために，PC を用いた実機でのデモンストレーションを実施する．

参考文献

[1] 水谷信子：あいづち論，日本語学，Vol. 7, No. 13, pp. 4–11 (1988).
[2] 水谷信子：あいづちと応答，No. 37-44, 筑摩書房 (1983).
[3] 笹岡直人，伊藤良生：騒音抑圧技術—基礎とその応用—，電子情報通信学会基礎・境界ソサイエティ，Vol. 5, No. 2, pp. 136–145 (2011).
[4] 森藤健：発話区間検知装置、音声認識装置、発話区間検知システム、発話区間検知方法及び発話区間検知プログラム，特開 2021-162685，グローリー株式会社 (2021).
[5] Sharma, T. et al.: Real Time Online Visual End Point Detection Using Unidirectional LSTM, *INTERSPEECH 2019* (2019).
[6] 島田敬輔：音声認識装置、ロボット、音声認識方法及び記録媒体，特開 2019-113820，カシオ計算機株式会社 (2019).
[7] さくらツーリスト株式会社：NG ワードゲーム，宴会王国（オンライン），入手先 ⟨https://ryokou-ya.co.jp/game/ng-word/⟩（参照 2022-05-23）.
[8] はとまめ：ワードウルフ，僕とボドゲ（オンライン），入手先 ⟨https://boku-boardgame.net/wordwolf⟩（参照 2022-05-23）.