E-020

咽喉マイクとピンマイクの同時集音に基づく多人数会話における発話区間推定

Voice Activity Detection Using Throat Microphone and Lavalier Microphone for Multi-party Conversations

大高祥裕† 西田昌史†‡ 綱川隆司†‡ 西村雅史†‡
Yoshihiro Otaka Masafumi Nishida Takashi Tsunakawa Masafumi Nisimura

1. はじめに

企業では情報の共有や新たな発案を目的とした会議が日々行われ，また，教育現場では複数人が同じテーマで課題を達成するグループワークを実施することが多い．これらの多人数会話を分析するには録音音声の聞き直しや，テープおこしと言った作業が必要であった．これらの処理を自動化し，作業を軽減することを目的とした研究が行われている[1] が，その中でも，話者および発話区間の正確な同定は重要な課題となっている[2][3]．発話の分離をより正確に行うため，多人数会話の音声収録では，話者毎にヘッドセットマイク等を装着し，多チャンネル収録を行うことも多い．それでも，周囲話者の発話が混入することは避けられず，対象話者の発話区間を正確に同定することは容易ではなかった．特に，発話が重畳することの多い，相槌や同調の区間を検出することは難しい[4].

先に我々は，多人数会話環境でも周囲発話の影響を受けにくく，対象話者の発話のみを安定して収録できる手段として咽喉マイクの利用を提案し，その有効性を検証した[5].

本稿では，咽喉マイクとピンマイクの 2ch 録音を利用した発話区間推定手法を新たに提案する．評価実験の結果，提案手法による発話区間推定性能の改善が見られたので報告する．

2. 発話区間推定

先に我々が提案した方法[5]では咽喉マイク単体の収録音に対し，発話区間推定（VAD）を行っていた．咽喉マイクの収録音は対象話者の発話の検出に大変有効であったが，嚥下などの生体音，服とマイクが干渉する際の衣擦れ音などが発話区間の誤検出の原因となることも分かった．Table 1 に，特に結果が悪かった被験者 2 名に対する発話誤検出の内訳を示す．他話者の発話以外は，咽喉マイク特有の雑音であり，誤検出原因の大部分を占めていた．逆にこれらの雑音は通常のマイク収録音にはほとんど影響を与えないので，ピンマイクで同時収録された音を使えば，誤検出を低減できる可能性がある．一方，被験者の発話は両方のマイクに同時収録されるので，これらの雑音との区別がさらに容易になると考えた.

今回咽喉マイクに対しては GMM に基づく VAD を行う．これはスペクトル情報を基に推定を行うことにより，パワーの小さい発話でも検出できるようにするためである．一方，ピンマイクにて収録される多種多様な音を十分に表現できる GMM を事前に学習することは困難であるので，ピンマイク側の VAD にはパワー情報による VAD を用いることにした.

Table 1 咽喉マイクを用いた VAD の誤検出の内訳

	分類	検出数	割合
話者A	嚥下	12	14%
	衣擦れ	46	55%
	肌との干渉	8	10%
	呼吸	5	6%
	他話者の発話	13	15%
	合計	84	
話者B	嚥下	13	15%
	衣擦れ	48	55%
	肌との干渉	12	14%
	呼吸	8	9%
	他話者の発話	7	8%
	合計	88	

これらの２つの収録音の検出結果を統合することで発話区間推定性能のさらなる向上を図った．各処理の詳細を以下に示す.

2.1 咽喉マイク収録音に対する VAD

学習用の音声データに対して，発話区間には speech，それ以外の特に嚥下や咳等のイベントが起こっていない区間を非発話区間として sil のラベルを人手で付与し，当該区間のデータを用いてそれぞれの GMM を学習した（いずれも混合分布数 32）．特徴量には，窓サイズ 25msec，シフト幅 10msec で抽出した MFCC (Mel-Frequency Cepstrum Coefficient)を用い， 0 番目のケプストラム係数を含めた低次から 13 次元，そのΔ，ΔΔの計 39 次元のパラメータを使用した.

なお，嚥下や衣摺れといった雑音を事前学習することでより正確に発話区間の検出が可能になると考えられるが，今回は学習データ量が限られていたこともあり，むしろ性能低下が見られたので雑音は学習には含めないこととした.

一方，評価データに対しては，フレームごとに各 GMM の尤度を比較することで非発話区間か発話区間であるかの判定を行う.

2.2 ピンマイク収録音に対する VAD

ピンマイクによって収録される環境音に対してはパワー情報による VAD を行う．フレームサイズ 40ms，シフト幅 20ms のフレームを設定し，フレーム毎の対数パワー情報に対して，閾値処理を行う.

この際，閾値は学習データに対して推定した対数パワー

† 静岡大学大学院 総合科学技術研究科
‡ 静岡大学 情報学部

のヒストグラムから算出し，定常雑音以外の殆どの音イベントが検出されるように設定した．

2.3 推定結果のスムージング

GMM 及びパワー情報による VAD では，フレーム毎に逐一発話区間と非発話区間の判定を行うため，ごく短い区間で発話区間と非発話区間が交互に検出されることが多い．これを防ぐため，推定発話区間に対してスムージングを行う．微小時間の判定揺れや嚥下などの雑音を発話とした区間を消去し，かつ実際の発話区間に影響がないよう，暫定的に発話区間に関する閾値を 0.2[sec]，非発話区間に関する閾値を 0.3[sec]と定め，閾値以下の区間を削除し，前後区間の結合を行った．この処理は各マイクに対する発話区間推定結果それぞれに行い，また推定結果を統合した後にも再度行っている．

2.4 2ch 音声における推定発話区間の統合

各マイクで推定された発話区間に対し，統合処理を行う．この際，両方の音から発話区間と推定された区間のみを推定発話区間とした（Fig.1）．

咽喉マイクの音に対して GMM で推定した区間が，ピンマイクのパワー情報により推定した区間内に無い場合はその区間を削除し（Fig.1，A.），ピンマイクの音から他人の発話が発話区間と推定されても，咽喉マイクで発話区間として検出されていなければ削除できる（Fig.1，B.）．この手法により，ピンマイクで発話区間と推定される他人の声や外部雑音，咽喉マイクで発話区間と推定される雑音の両方を推定区間から削除することができる．さらに，両マイクでの誤推定区間が重複してしまった場合でも，重複区間が短ければ統合処理において取り除くことができる（Fig.1，C.）．

3. 評価実験

評価実験として，男子大学生 3 人組による 5 分間の自由会話を 1 セッションとし，計 5 セッションを対象に発話区間の推定を行った．テスト対象となった発話数は，計 887 発話である．対象となる音声に対して人手で正解区間を付与し，検出された区間の開始及び終了に対して計 0.5 秒以上の誤差がある場合には不正解とした．一方，GMM の学習には，評価実験とは異なる被験者グループによる会話音声（計 138 発話）を用いた．発話区間推定結果に対し，再現率(recall)，適合率(precision)，及び F 値(F-measure)を算出した．再現率は正解ラベルの内どれだけ正しく検出できたか，適合率は検出されたラベルの内どれだけ正しく検出できたかを示す．F 値は適合率と再現率の調和平均であり，正確性と網羅性を総合的に評価する値である．

Table 2 は，咽喉マイクの音声単体を対象に GMM による VAD を行ったものと，2ch 音声を利用した提案手法の比較である．いずれのセッションにおいても，F 値において少なくとも 0.05 以上の向上が見受けられる． F 値向上の要因として，咽喉マイクと衣服の衣擦れ音や肌との干渉音などの雑音を排除できたことによる適合率の大幅な改善が挙げられる．

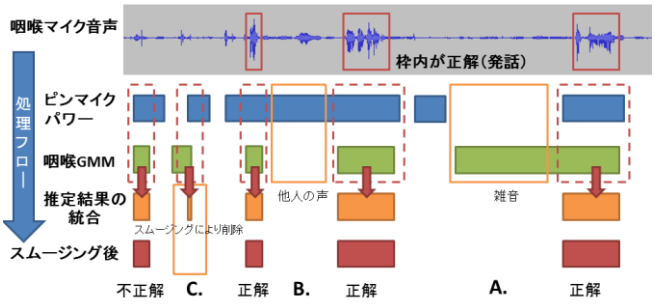

Fig.1　2ch 音声における推定発話区間の統合フロー

Table 2　各セッションにおける発話区間推定性能

	検出数	正解数	再現率	適合率	F値
第1セッション		214			
咽喉マイク	278	196	0.92	0.71	0.80
2ch	215	196	0.92	0.91	0.91
第2セッション		164			
咽喉マイク	207	155	0.95	0.75	0.84
2ch	181	154	0.94	0.85	0.89
第3セッション		181			
咽喉マイク	217	167	0.92	0.77	0.84
2ch	190	165	0.91	0.87	0.89
第4セッション		168			
咽喉マイク	219	154	0.92	0.70	0.80
2ch	178	154	0.92	0.87	0.89
第5セッション		160			
咽喉マイク	196	149	0.93	0.76	0.84
2ch	174	148	0.93	0.85	0.89
全セッション合計		887			
咽喉マイク	1117	821	0.93	0.74	0.82
2ch	938	817	0.92	0.87	0.90

4. おわりに

咽喉マイクの収録音に加え，ピンマイクの収録音の情報を利用することで，多人数会話において，より正確な発話区間推定が可能になることを示した．

今後はさらに多人数の活発な会話を対象とするなど，本手法の実用性を高める研究を行う．また，コミュニケーションの可視化も検討する予定である．

謝辞

本研究の一部は科研費補助金（16H01817），科研費助成金（16K13028，16K01543）の交付を受けた．また，電気通信普及財団の研究調査助成を受けた．

参考文献

[1] 藤本雅清，”音声区間検出の基礎と最近の研究動向”，電子情報通信学会技術研究報告.SP，110(81)，pp.7-12，(2010).
[2] 坊農真弓，高梨克也，”多人数インタラクションの分析手法”，オーム社，(2009)
[3] 荒木章子，藤本雅清，石塚健太郎，澤田宏，牧野昭二，”音声区間検出と方向情報を用いた会議音声話者識別システムとその評価”，日本音響学会 2008 年春季研究発表会，pp1-4，(2008).
[4] 石塚健太郎，荒木章子，藤本雅清，瀬戸口久雄，高梨克也，河原達也，”ポスター会話 に対する発話区間検出と話者識別の検討”，情報処理学会研究報告.SLP，69，pp.217-222，(2007).
[5] 大高祥裕，西田昌史，西村雅史，”咽喉マイクを利用した多人数会話における発話区間推定”,WiNF2015 P2-11,pp.104-106,(2015).