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

大高　祥裕　　西田 昌史　　西村　雅史

静岡大学情報学部

Voice Activity Detection Using Throat Microphone for Multi-Party Conversation

Yoshihiro Otaka　　Masafumi Nishida　　Masafumi Nishimura

Faculty of Informatics, Shizuoka University

1.　はじめに

グループディスカッション等の多人数会話の分析において，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[1][2]．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，話者毎にピンマイク等を装着し，多チャンネル収録を行うことも多いが，それでも，周囲話者の発話の混入は避けられず，対象話者の発話区間の正確な同定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．

本研究では，多人数会話環境でも周囲の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と考えられる咽喉マイクを発話区間推定（VAD：Voice Activity Detection）に利用することを検討する．話者が個々にマイクを装着する状況を想定し，咽喉マイクとピンマイクを同時装着して会話を収録した[3]．発話区間推定に咽喉マイクの収録音を用いる場合と，従来のピンマイクを用いる場合とで発話区間推定性能の比較を行ったので報告する．

2.　実験機器

咽喉マイクは首の周囲に装着することで集音が可能なマイクであり，咽喉周辺の皮膚の振動を音として記録することが出来る．今回実験に使用した咽喉マイクの写真を Fig.1 に示す．このマイクの特性として，咽喉から得られる音以外の外部ノイズや騒音の影響を受けにくいこと，嚥下や咳，比較的声の小さな発話や笑いといった，生体音を含む本人から発せられる音を記録しやすいことが挙げられる[3]．この中でも，今回は外部騒音の影響を受けにくい点，比較的声が小さな発話でも記録が可能な点に着目し，咽喉マイクを使用した．

今回の研究では，咽喉マイクおよびピンマイクの音を記録するために IC レコーダを用いた．IC レコーダは外部マイクでの録音ができる録音端子を備え，ステレオで録音が可能なものを選択した（L: 咽喉マイク，R: ピンマイク）．この IC レコーダによって，ビットレート 128Kbps，サンプリングレート 44.1KHz の MP3 形式で録音を行った．

3.　発話区間推定

3.1.　GMM

発話区間推定(VAD)には，様々な手法が既に提案されている[4]．本研究では VAD を，GMM(Gaussian Mixture Model)を用いて行う．

咽喉マイクを使用した VAD の先行研究では， Energy ベースで VAD を行っている研究がある[5]．嚥下音，咳などの生体音は比較的大きなレベルで録音されることがわかっているため，Energy ベースでは発話と判断されることが多い．しかし，先行研究では数字のみを対象とした音声認識を行うため，咳や嚥下音で発話区間とされても，誤認識の可能性は低いため大きな問題ではなかった．一方，本研究では大語彙音声認識を視野に入れているため，GMM を用いて，音声区間だけの正確な VAD を試みる．

Fig.1：咽喉マイク

まず，GMM の学習データと評価データの音声データに対して特徴量抽出を行う．学習用の音声データに対して，人手で波形及び聞き取りの双方の観点から，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無音区間として sil のラベルを付与し，当該区間のデータを用いてそれぞれ無音区間の GMM と発話区間の GMM を学習した（混合数 32）．

評価データに対して，フレームごとにそれぞれの GMM との尤度を算出することで無音区間か発話区間であるかを判定し，尤度が高かった GMM の区間を判定結果とすることで発話区間の推定を行う．

特徴量抽出には，窓サイズ 25msec，シフト幅 10msec，39 次元の MFCC （Mel-Frequency Cepstrum Coefficient)を用いた．MFCC は音声情報処理の分野にて主に用いられる特徴量であり，フーリエ変換によって求めたスペクトル情報に対して，人間の聴覚特性（低い周波数では細かく，高い周波数では荒い分解能をもつ）に合わせてフィルタ群を使い，人間の声道特性を示すスペクトル包絡を表すケプストラムを抽出したものである．メルフィルタバンクは 24 チャネルのものを用い，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．

3.2.　スムージング

GMM による VAD ではフレーム毎に逐一発話区間と無音区間の尤度判定を行うため，発話と発話の間が少しあいていたり，声の音量の小さな揺れが起こったりすることにより，ごく短い区間で発話区間と無音区間が交互に検出されること

が多い．これを防ぐため，推定された発話区間に対してスムージングを行う．具体的には，先述した微小時間の判定揺れや咳や嚥下といった音声イベントを発話とした区間を消去し，かつ実際に発話した区間に影響がない時間設定であることを踏まえ，暫定的に発話区間に関する閾値を 0.2[sec]，無音区間に関する閾値を 0.3[sec]と定め，閾値以下の時間長の区間を削除し，前後区間の結合を行った．

4. 評価実験

4.1. 実験条件

本研究で提案した咽喉マイクによる発話区間推定の評価を行うため，自由会話における発話区間の推定実験を行った．本実験ではピンマイクによる集音及び学習，推定と，咽喉マイクによる集音及び学習，推定の精度比較を行う．

被験者は男子大学生 5 名で，雑音の少ない環境下にて，2 人組及び 3 人組による自由会話を 5 分間記録した．内訳として，2 人組会話のセッションでは話者 A と話者 B 組み合わせで 1 回，3 人組会話のセッションでは話者 C，話者 D，話者 E の組み合わせで 1 回行った．一方，学習用には話者 A~E とは異なる男子大学生 2 名の 5 分間の自由会話を録音し，その内の 1 人の音声データに正解ラベルを付与して GMM の学習を行った．その GMM を元に話者 A～E のデータに対して GMM による VAD を行い，その結果を正解ラベルと比較した．正解ラベルはスムージングの処理の閾値と同様，発話と発話の間の無音が 0.3 秒未満であれば，まとめて 1 つの発話とした．スムージングと違う点は，発話に関して 0.2 秒以下の時間長であるように見える発話でも，発話であると判断した区間は正解ラベルを付与している．また，正解ラベル上は 2 つの発話区間であっても，推定されたラベルが 1 つの発話区間として推定されることがある．その際には正解ラベルの 2 つの発話区間の間にある無音区間における中間点で推定ラベルを分割し，正解かどうかの判断を行った．判断基準として，検出された発話開始時刻及び発話終了時刻に対して，正解ラベルとそれぞれ比較し，計 0.5 秒以上の誤差が認められた場合は不正解とした．これらの結果に対し，再現率(precision)，適合率(recall)，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．

4.2. 実験結果

実験の結果得られた 2 人組会話における各話者の再現率，適合率，F 値を Table 1 に，3 人組会話における各話者の再現率，適合率，F 値を Table 2 に示す．Table 1 に示したデータからわかるように，2 人組会話におけるデータでは，咽喉マイクを使用した発話区間推定はピンマイク単体を使用した発話区間推定よりも再現率において 0.20,適合率に関して 0.14 の上昇がみられた．特に，再現率に関しては 0.97 と高い水準を示しており，従来よりも推定漏れが少ないことがわかる．Table.2 より，3 人組会話に関しても再現率に関して 0.39，適合率に関して 0.38 の向上がみられ，再現率が 0.97 を維持していることから，更に人数が増加しても，咽喉マイクはベースラインであるピンマイクよりも発話区間推定に有効であることがわかる．

各話者のデータに着目すると，話者によって結果の傾向が異なる状況が見られた．3 人組の会話における話者 D では，咽喉マイクにおける再現率が 0.97，ピンマイクにおける再現率が 0.48 と結果の差が他話者と比べ大きく開いた.他話者よりも発話の音量が小さく，ピンマイクにおいては対面する他話者の発話音量とほぼ差が無いことが原因と推察される．このように，音量の小さい発話でも咽喉マイクではより正確に発話区間を推定することが出来る．一方，話者 B において，他話者に比べ再現率は高いが，適合率が低いという結果が得られた．これは， 3 人組の会話における話者 C でも同様に見られた．

話者 B 及び話者 C に関して，適合率の低さの原因を調査する為,不正解ラベルの内容毎に集計した.対象となる区間は，話者 B における不正解区間 84 区間,及び話者 C における不正解区間 88 区間である．区別した種類は，嚥下音，衣擦れのような音，クリックノイズ，呼吸音，他話者の発話音声の 5 種類である．クリックノイズは咽喉マイクに触れた場合などに突発的に発生する雑音で，ごく短い区間で発生する加法性ノイズは全てクリックノイズと分類した．集計結果を Table 3 に示す．結果として，不正解区間のうち最も多かったのは衣擦れのような音であり，話者 B，話者 C 共に 55%であった．この衣擦れのような音は首の周りにある衣服との干渉によって起こると考えており，話者の衣服や所作によってノイズの数が大きく変わると推測できる．また，話者 B は話者 C と比較して，他話者の発話音声を誤って発話区間と推定する割合

Table1：2 人組間会話における話者毎の発話区間推定性能

	使用マイク	発話数	検出数	再現率	適合率	F値
話者A	咽喉マイク	91	108	1.00	0.84	0.91
	ピンマイク	85	139	0.93	0.61	0.74
話者B	咽喉マイク	75	159	0.94	0.47	0.63
	ピンマイク	46	160	0.58	0.29	0.38
合計	咽喉マイク	166	395	0.97	0.42	0.59
	ピンマイク	131	471	0.77	0.28	0.41

Table2：3 人組間会話における話者毎の発話区間推定性能

	使用マイク	発話数	検出数	再現率	適合率	F値
話者C	咽喉マイク	34	122	0.97	0.28	0.43
	ピンマイク	19	182	0.54	0.10	0.18
話者D	咽喉マイク	64	90	0.97	0.71	0.82
	ピンマイク	32	176	0.48	0.18	0.26
話者E	咽喉マイク	40	49	0.95	0.82	0.88
	ピンマイク	32	182	0.76	0.18	0.29
合計	咽喉マイク	138	261	0.97	0.53	0.68
	ピンマイク	83	540	0.58	0.15	0.24

Table3：話者 B 及び C の不正解ラベル内容別分類

	分類	検出数	割合
話者B	嚥下	12	14%
	衣擦れ	46	55%
	クリックノイズ	8	10%
	呼吸	5	6%
	他話者の発話	13	15%
	合計	84	
話者C	嚥下	13	15%
	衣擦れ	48	55%
	クリックノイズ	12	14%
	呼吸	8	9%
	他話者の発話	7	8%
	合計	88	

が多かった．話者 B の録音データには，他の話者に比べて本人ではない他話者の発話が咽喉マイクにごく小さな音量で，はっきりと記録されていた．

5. おわりに

本研究では，発話者の発話区間の推定に従来のピンマイクでなく，咽喉マイクを用いることを提案した．評価実験によって，ピンマイク単体による発話区間推定よりも，提案手法である咽喉マイクによる発話区間推定が，より正確であり有効な手段であることがわかった．

今回は 1 話者の発話のみを学習データとしたが，今後は学習データを増やし，より正確に区間推定を行える GMM の作成を目指す．また，今後は挿入誤りの原因であった咽喉マイクに記録される金属音のようなノイズ，僅かに入ってしまう外部騒音，咽喉マイクから通常得られる嚥下や咳といった生体音と発話の識別も検討する．これら発話以外の誤検出に関して，発話ほどの時間長が得られていないことから，3.2.の項で記述したような推定ラベルのスムージング処理に関しても改良が可能であることが考えられる．

なお，音声認識の対象としては咽喉マイクの収録音は帯域などに制限があり，必ずしも望ましいものではない．これについてはピンマイク側の収録音の活用を検討したい[6]．

今後は会話の流れの推定及び会話内容の分析に関しても研究をすすめていく予定である．

参考文献

[1] 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)

[2] 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,(2008)

[3] 西村 雅史, 小林 悠一, 桐山 伸也, 峰野 博史:” 生体音と環境音の同時収録による高齢者の行動および身体状態認識に関する検討”, 音響学会講演論文集, 2-4-9, pp. 1309-1310, (2015)

[4] 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, 音声 110(81), 7-12,(2010)

[5] Tomas Dekens, Werner Verhelst, Francois Capman, Frédéric Beaugendre:“Improved speech recognition in noisy environments by using a throat microphone for accurate voicing detection”, 18th European Signal Processing Conference (EUSIPCO-2010), pp. 23-27,(2010)

[6] Stéphane Dupont, Christophe Ris, Damien Bachelart :”Combined use of close-talk and throat microphones for improved speech recognition under non-stationary background noise”, Proceedings of Robust 2004 (Workshop (ITRW) on Robustness Issues in Conversational Interaction),(2004)