6Q-08

咽喉マイクを用いた発話区間検出に基づく多人数会話音声認識

大高 祥裕　　綱川 隆司　　西田 昌史　　西村 雅史
静岡大学大学院　総合科学技術研究科　情報学専攻

1. はじめに

多人数会話においては，他話者の発話もマイク音声に混入するため，正確に音声認識を行うことは難しい．このため，多人数会話におけるダイアライゼーションの研究も多くなされている[1].

本研究では，咽喉マイクとピンマイクといった特性の異なるマイクを併用し，各マイクに対し異なる発話区間推定法を用いてその結果を統合することにより，他話者の発話による誤検出を低減し，推定性能の向上を目指す．性能評価として，自由会話において各話者に装着したマイク音声に対して提案手法での区間推定を実施し，従来用いられるパワー情報を用いた VAD と比較して音声認識時の性能にどのような影響を与えるかについて評価した．

2. 提案手法

本研究では，多人数会話の実施時に咽喉マイクとピンマイクの同時集音を行う．咽喉マイクは首の咽喉周辺に装着するマイクであり，外部からの騒音や，自分以外の話者の発話は収録されにくい．その特性を活かし発話区間推定に用いる先行研究も行われている[2]．先に我々は，この咽喉マイクと従来個人発話の集音に用いられるピンマイクの両方を用いることで発話区間検出の精度を向上できることを確認した[3]．本稿では，その手法を改良し，大語彙音声認識の前段の発話区間推定に用いる．

提案手法の処理フローを図 1 に示す．本手法では，学習データの咽喉マイク音声を用いた発話及び非発話の 64 混合の GMM(Gaussian Mixture Model)と，テストデータのピンマイク音声を用いた発話及び非発話の 16 混合の GMM の，2 セット計 4 種類のモデルを使用する．特徴量抽出には，窓サイズ 25ms，シフト幅 10ms，低次から 13 次元及び Δ，ΔΔ を含めた計 39 次元の MFCC (Mel-Frequency Cepstrum Coefficient)を用いた．

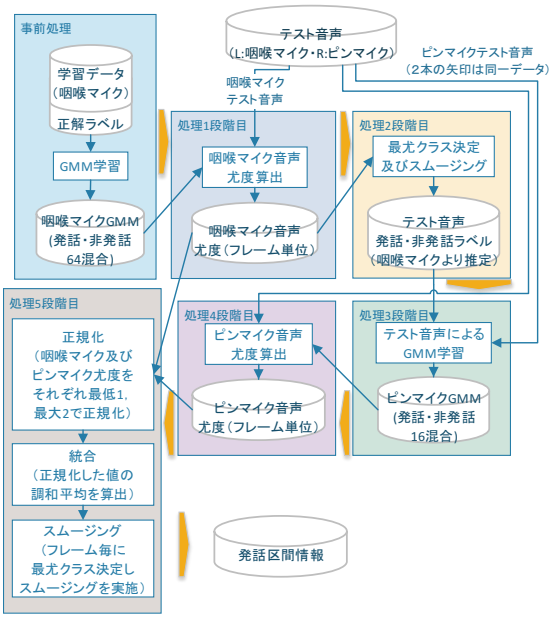

図 1 : 咽喉マイクとピンマイクを
併用した VAD 処理フロー

まず，咽喉マイク音声の学習データを用いて， 発話及び非発話の GMM を用意する．このモデルを用いて，テスト音声の咽喉マイク側の音声に対してフレーム毎の尤度を求める．ここで得られた尤度を，咽喉マイク側の尤度として後の統合に使用する．

次に，この尤度を用いて，フレーム毎に尤度の高いクラスを所属クラスとし，後述のスムージングを行い，テスト音声の咽喉マイク音声から暫定的な発話区間を得る．これを教師ラベルとし，テスト音声のピンマイク側音声を用いて発話及び非発話の GMM を作成する．このモデルに対して，テスト音声のピンマイク側音声の尤度を求め，統合に使用する．データクローズの推定となるため，汎化性能を高める目的で混合数は 16 混合としている．

咽喉マイク及びピンマイクから得られた尤度を利用するため，調和平均を求めることで両マイク音声の情報を統合する．この際，調和平均を使用するために，各音声から得られた発話・非発話モデルに対するフレームごとの尤度を最低 0，最大 1 のデータへと正規化し， 1 のオフセットを加える．得られた調和平均後の値を用いて，フレーム毎の発話，非発話をより値の高い方に決定する．

Multi-Party Conversation Speech Recognition Based on VAD Using Throat Microphone
Yoshihiro Otaka, Takashi Tsunakawa, Masafumi Nishida, Masafumi Nishimura
Department of Informatics, Graduate School of Integrated Science and Technology, Shizuoka University 432-8011, Hamamatsu, Japan

得られた発話，非発話ラベルに対して2段階のスムージングを実施する．1段階目として，i番目のフレームの前後nフレームの推定結果に対して最頻値をとり，iフレームの結果とする．2段階目として，1段階目の結果から前後n/2フレームの最頻値を取り，その区間全て(n+1)の推定結果とする．この際，シフト幅はnとなる．これにより推定揺れやごく短い誤推定の削除を行い，推定結果とする．

3. 発話区間推定実験

提案手法が先行研究[3]の手法と比較するため，先行研究と同様の条件下で発話区間推定の性能評価実験を実施した．先行研究では本稿の提案手法とは異なり，ピンマイク側の音声に対してはパワー情報を用いた区間推定を行い，咽喉マイク音声のGMMによる区間推定との重複区間を発話区間としている．話者5名の自由会話音声1時間（総発話数2524発話）に対し，発話区間単位で集計し，前後誤差0.3秒以内を正解とした．話者ごとに交差検証を行っている．スムージング時のシフト幅はテストデータと同様の環境下で録音された音声で調整し，暫定的にn=40と定めた．

結果を表1に示す．先行研究と比較して，再現率を維持したまま適合率を高めることが出来た．これは，咽喉マイクの雑音と同時に発生した外部雑音や他話者の発話に対して，先行研究ではピンマイク側のパワーVADにて発話区間と誤推定していた区間を，GMMを用いることで音響的に棄却出来ていることが要因と考えられる．

4. 音声認識実験

提案手法が音声認識時の性能にどのような変化を与えるかを調査するため，音声認識実験を実施した．テスト音声は話者3名による15分程度のアクティブラーニング音声を対象とする．マイクは音声認識精度を高めるためヘッドセット型の接話マイクを使用し，咽喉マイクと同時装着して集音した．なお，咽喉マイクのGMM作成には異なる環境下で録音した，テスト音声とは異なる話者5名による1時間の雑談音声(計5時間)を使用している．スムージング時のシフト幅は区間推定実験と同様n=40とした．

テスト音声に対して，VAD無し，接話マイクのパワー情報を用いたVAD，提案手法，人手による区間切り出しの4つを実施し，発話と推定された音声区間のみを1つの音声へ結合し音声認識を実施した．パワー情報を用いたVADはJulius[4]等で一般的に利用されており，今回は閾値をテスト音声全体のパワー情報に対して判別分析法を用いて定めている．事前に書き起こしを実施し，認識結果と比較してCERを算出した．認識エンジンにはAttila[5]を使用している．

表1　各手法における発話区間推定性能(2524発話)

手法	検出数	正解数	再現率	適合率	F値
咽喉のみ	3230	2344	0.93	0.73	0.81
先行研究	2994	2331	0.92	0.78	0.84
提案手法	2685	2334	0.92	0.87	0.90

表2　各手法における音声認識誤り率(%)

手法	置換誤り	脱落誤り	挿入誤り	CER
VAD無し	24.5	16.5	17.7	58.6
パワー	22.4	14.2	6.9	43.5
提案手法	19.7	16.5	2.8	39.0
人手	20.8	14.3	2.8	37.8

実験結果を表2に示す．提案手法において，ベースラインと考えられるパワーVADから4.5%改善し，人手による切り出しに対して1.2%差まで性能を向上出来た．

5. おわりに

本実験では多人数会話における音声認識に関して，咽喉マイクの音声を併用することで前段の発話区間推定を改善し，認識精度の改善につながることが確認できた．

今後の予定としては，より騒音のある現実的な環境下で録音された音声に対する検証を実施したい．

謝辞

本研究の一部はJSPS科研費(16H01817, 16K13028, 16K01543)の交付を受けた．

参考文献

[1] 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)

[2] T.Dekens, W.Verhelst, F.Capman, D.Beaugendre,"Improved speech recognition in noisy environments by using a throat microphone for accurate voicing detection", 18th European Signal Processing Conference (EUSIPCO-2010), Aalborg, Denmark, August 23-27, (2010)

[3] 大高祥裕, 綱川隆司, 西田昌史, 西村雅史, "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に関する研究", 信学技報, vol. 116, no. 279, SP2016-43, pp. 15-20, (2016).

[4] LEE A.," Julius-An Open Source Real-Time Large Vocabulary Recognition Engine", Proc. of 7th European Conference on Speech Communication and Technology (EUROSPEECH2001), (2001)

[5] H.Soltau, G.Saon, B.Kingsbury, "The IBM Attila speech recognition toolkit", Spoken Language Technology Workshop (SLT), (2010)