E-004

多人数会話におけるホットスポットの自動推定
Automatic Estimation of Hotspot for Multi-party Conversation

大高 祥裕† 綱川 隆司† 西田 昌史† 西村 雅史†
Yoshihiro Otaka Takashi Tsunakawa Masafumi Nishida Masafumi Nishimura

1. はじめに

現代社会において，会議やグループワークといった多人数会話は日常的に行われており，上司や教師等の実施管理者は録音した音声から会話の状況を把握することがある[1]. 近年では認知症予防を目的とした共想法[2]のように，高齢者に対して実施されることもある．本研究は，これらの多人数会話の録音音声からの状況分析を支援することを目的とする．

多人数会話においては，相槌や笑いといった他話者の発話に対する反応が多く見られる．これらが生起する場所の付近には重要な内容が含まれていることが多く，これらはホットスポットと呼ばれている[3].

一方，多人数会話において，咽喉マイクを用いた発話区間推定が有効であることを以前の研究で示した[4]．これを利用し，多人数会話における相槌や笑いを咽喉マイクの音声から検出し，その情報を用いてホットスポットを自動推定するシステムを検討している．

本論文では，ホットスポット検出を行う上で重要な情報となる，「笑い」と「相槌」という２種類の音イベントの自動検出・分類方法について述べる．

2. 発話区間検出

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，本人の発話，及び本人の発する咳や嚥下音などの生体音を記録することができる．なお，この咽喉マイクでは外部からの騒音や，自分以外の話者の発話は収録されにくく，従来用いられるピンマイクとは異なった特性を持っている．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[4]．これを利用して，今回対象となる多人数会話に対しても，2ch での録音を行い，[4]の手法を使用して発話区間検出を行った．(図 1)

事前調査として，**3.3 分類性能実験**と同様の条件下で録音した音声(話者 5 名による 1 時間の自由会話)に対して発話区間検出を実施し，その内相槌及び笑いを合計した他話者に対する反応の再現率を調査した．結果，0.94 と高い精度で区間検出を実施できていることがわかった．（表 1）

表 1　発話区間検出時点での再現率

分類	総イベント数	検出数	再現率
発話	1731	1596	0.92
反応	793	748	0.94

† 静岡大学, Shizuoka University

図 1　咽喉マイクとピンマイクを使った２ch 収録

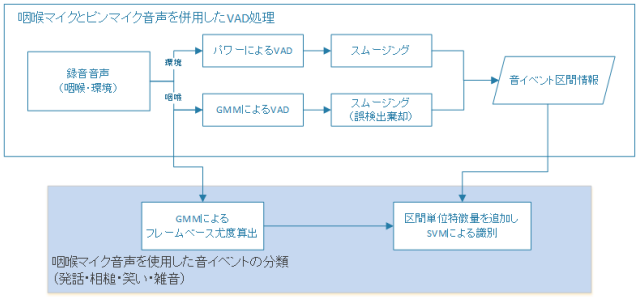

図 2　音イベント識別フロー

3. 音イベント分類

前項で得られた発話区間検出結果に従って切り分けられた音イベントに対して，ホットスポットを推定するための音イベント分類を行う．相槌及び笑いの区間を検出することで，ホットスポット位置の推定に役立てる．

3.1 ホットスポット

ホットスポットとは会話中における重要な意味が含まれる発言であり，会話に参与している他者の反応を呼び起こす働きをする．音声会話においては聞き手が発言者の発話に対して反応を示すことで会話が成立し，特に大きな興味や関心を持った発話に対して，大きな反応を示すことが多々ある．これを利用して，会話中のホットスポットを自動推定する試みが多数行われている[5][6]．本研究においても，相槌や笑いといった他話者に対する反応を検出することが，ホットスポット検出を行う上で重要な情報となるとして，音イベント分類にて相槌及び笑いを検出することを試みる．

3.2 提案手法

本研究では音イベント分類の性能を確認するため，次の２つの異なる手法を提案し比較する．方法１）音イベント区間内の各フレーム単位で GMM(Gaussian Mixture Model)によってモデルを作成し，区間内の累積尤度を用いて区間全体の種類を推定する手法，方法 2）1 つ目の手法で得られたフレーム単位の尤度を音イベント区間で平均化し，区間内の，話者ごとに正規化された基本周波数(F0)の平均値

及び最大値を加えて SVM により分類する手法である．(図 2)

音イベント区間の各フレームの累積尤度を使用する方法では，識別器に GMM を使用し，発話，相槌，笑い，雑音の 4 クラスの GMM(混合数 32)を作成する．フレームごとに各 GMM との尤度を比較し，区間内の累積尤度によって所属クラスを推定する．使用する特徴量はパワーを含めた MFCC(Mel Frequency Cepstrum Coefficient)13 次元と Δ 及び ΔΔ の計 39 次元を使用する．

2 つ目の手法では，音イベント区間内の各モデルに対するフレーム尤度を区間単位に平均した値(4 次元)に加え，Wrede ら[5]の調査により相槌や笑いの検出に有効とされた，話者ごとに Z スコアによって正規化された音イベント区間内の F0 の平均及び最大(2 次元)を加えた計 6 次元を特徴として，RBF カーネルの SVM(Support Vector Machine)を使用して所属クラスを識別する．

これらの手法を用いて，咽喉マイクの音声とピンマイクの音声を利用した場合それぞれの精度の確認・比較を行った．

3.3 分類性能実験

音イベント分類の評価実験として，高齢者 5 人組による 4 時間の会話の中から，特に会話の活発であった 1 時間を対象とし，発話区間推定を実施し切り分けられた音イベントに対して音イベント分類を行った．テスト対象となった音イベント数は，計 1590 区間である．対象となる音イベントに対して人手で発話，相槌，笑い，雑音のいずれかのラベルを付与した． 本実験は交差検証にて実施し，1 名のテスト話者を決定した後，残りの 4 話者で GMM の学習を行った．これを話者毎に計 5 回実施した．分類結果に対し，再現率(recall)，適合率(precision)，及び F 値(F-measure)を算出した．結果を表 2 に示す．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．

結果として，1 つ目の提案手法である，フレーム単位の特徴量で累積尤度を用いて推定する方法よりも，若干ではあるが，F0 を加えた手法において各クラスに対する F 値が向上した．相槌の検出においては特に顕著であり，F0 を使用しない手法において，発話に分類されていた幾つかのサンプルを正しく推定できている．また，雑音も本研究の目的ではないが検出性能が向上している．これは，F0 の抽出範囲を 60Hz~800Hz と定めているため，F0 が抽出できなかった区間が雑音クラスのデータに多数見られたことが関係していると考える．

なお，従来使用されるピンマイク側の音声を使用した場合，いずれの手法においても精度は低下した．これは，ピンマイク側に混入する他者発話の影響や，小さな声で発言された相槌が録音できていなかったことに起因すると考えられる．

表 2　相槌・笑い分類性能比較実験

クラス	総データ数	識別数	正解数	再現率	適合率	F値
GMM(F0不使用)：咽喉マイク						
発話	1050	1033	896	0.85	0.87	0.86
相槌	326	272	216	0.66	0.79	0.72
笑い	156	233	115	0.74	0.49	0.59
雑音	58	52	5	0.09	0.10	0.09
GMM(F0不使用)：ピンマイク						
発話	1050	1089	902	0.86	0.83	0.84
相槌	326	266	170	0.52	0.64	0.57
笑い	156	181	76	0.49	0.42	0.45
雑音	58	54	9	0.16	0.17	0.16
GMM+SVM(F0使用)：咽喉マイク						
発話	1050	1035	907	0.86	0.88	0.87
相槌	326	290	236	0.72	0.81	0.77
笑い	156	205	111	0.71	0.54	0.61
雑音	58	60	28	0.48	0.47	0.47
GMM+SVM(F0使用)：ピンマイク						
発話	1050	1087	891	0.85	0.82	0.83
相槌	326	261	181	0.56	0.69	0.62
笑い	156	185	82	0.53	0.44	0.48
雑音	58	57	13	0.22	0.23	0.23

4. おわりに

本実験では多人数会話におけるホットスポットの自動推定に関して，推定に利用する相槌や笑いを検出する際，咽喉マイクの音声を用いて推定を行うことにより一定の精度が得られることがわかった．また，F0 を特徴量に加えることによって多少ながら精度が改善する見込みがあることを確認できた．

今後の予定として，更なる音イベント分類の精度向上を目指すとともに，推定されたホットスポット位置の妥当性を検証する．また，発話区間情報から抽出できる発話区間長，発話交代数，話者交代数を用いてホットスポットの重み付けを実施する方法も検討していく．

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] 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] 大武美保子,”認知症予防回復支援サービスの開発と忘却の科学”, 2007 年度人工知能学会全国大会論文集, 1H2-1, (2007).
[3] 河原達也, 須見康平, 緒方淳, 後藤真孝, “音声会話コンテンツにおける聴衆の反応に基づく音響イベントとホットスポットの検出”, 情報処理学会論文誌, Vol.52, No.12, 3363-3373, (2011).
[4] 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).
[5] Britta Wrede, Elizabeth Shriberg, “Spotting “Hot Spots” in Meetings: Human Judgments and Prosodic Cues”, Proc. of Eurospeech, 2805-2808, (2003)
[6] 嶋田和孝, 楠本章裕, 横山貴彦, 遠藤勉, “複数人談話における笑いの情報を考慮した盛り上がり判定”, 信学技報 n vol.112, no.111, 25-30, (2012).