T2R2 東京科学大学 リサーチリポジトリ
Science Tokyo Research Repository

論文 / 著書情報
Article / Book Information

論題(和文)	発声様態依存モデルを用いた話者認識
Title(English)	
著者(和文)	小塚俊来, 岩野公司, 篠田浩一
Authors(English)	Toshiki Kozuka, Koji Iwano, Koichi Shinoda
出典(和文)	日本音響学会講演論文集, , , pp. 185-188
Citation(English)	, , , pp. 185-188
発行日 / Pub. date	2013, 3

Powered by T2R2 (Science Tokyo Research Repository)

発声様態依存モデルを用いた話者認識*

☆小塚俊来 (東工大), 岩野公司 (都市大), 篠田浩一 (東工大)

1　はじめに

現在，複数話者による会話，特に会議音声の分析を行う研究が盛んに行われている [1, 2, 3]．これは会議の議事録の自動作成や対話システムの構築などを目的としている．この応用では発声内容だけでなく，いつ誰が話したかという情報も重要である．そのような情報を抽出する処理は話者ダイアライゼーション [4, 5] と呼ばれている．

会議音声には書き起こしの対象となる音韻情報以外にも笑い声，咳，フィラーなども含まれている．これらは音韻とは異なる特徴も持っている [6, 7]．そのためこれらも含めて話者ダイアライゼーションを行うとその精度が低下する可能性がある．

そこで本研究では，音韻情報を含まないこれらの発声のうち特に笑い声を対象とし，それを含んでいる会議音声における話者ダイアライゼーションを考える．特に発声区間が既知の仮定のもとで，誰が話したかを判別する話者認識について，性能を向上させるための手法を提案する．

2　データベース

日本人の男子学生 4 名による合計 40 分程度の会議音声データを収録した．Fig. 1 に示すように長方形の机の中央にバウンダリーマイクを置いて録音した．また，各話者はピンマイクを付けている．データは 2 回に分けて収録した．録音時間は各々19 分 26 秒と 20 分 14 秒である．前者をデータ A, 後者をデータ B と呼ぶ．データ A のトピックは “大学への要望”，データ B のトピックは “世界旅行の行き先” である．

これらに対し，日本語話し言葉コーパス (CSJ) の形式に準拠して人手でラベルを付けた．各話者が付けていたピンマイクで収録した音声を用いて，各話者毎にその発声区間のラベルを付ける．ここで発声区間とは，0.2 秒以上のポーズによって挟まれた音声の範囲のことを指す．なお，全発声時間のうちデータ A は 50.6%，データ B は 46.7%が複数話者の音声が重畳した区間となっている．これは発話スタイルが極めてカジュアルであることを示している．

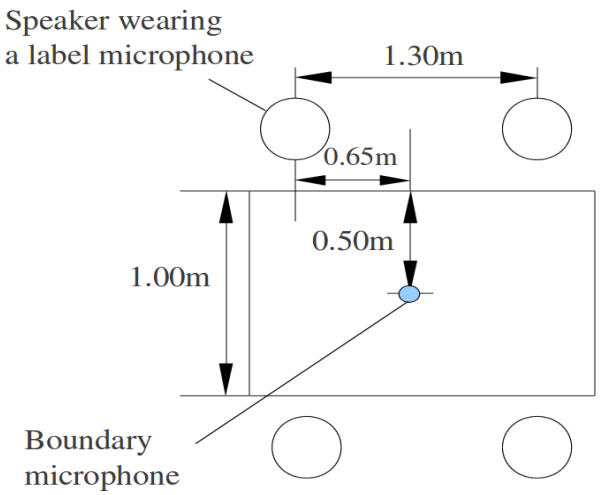

Fig. 1　Position of speakers and microphones

ラベル付けの例を Table 1 に示す．音韻情報以外について，笑い声を示す L，笑いながらの発声を M と表記し，その開始地点を示す <M と終了地点を示す M> をつけた．それ以外に咳を表す C，フィラーを示す F，語断片を示す D，曖昧な発声を示す W，相槌を示す B，感動詞を示す E のラベルを付けた．

本研究では笑い声を対象とし，C，F，B，E のみが付与されている発声区間は除外する．以降では，笑い声ラベル L を有する区間を “L”，笑いながら発声のラベル M を有する区間を “M”，それ以外の区間を “N” と表す．各話者について，L，M，N に分類された各発声区間の総時間長を Table 2 に示す．発声時間の割合は，L 10.2%，M 31.3 %，N 52.2%である．

---

*Speaker recognition by models which depend on speaking style. by Toshiki Kozuka (Tokyo Institute of Technology), Koji Iwano (Tokyo City University) and Koichi Shinoda (Tokyo Institute of Technology)

Table 1　Example of speaking. L:Laugh, <M:Start of speaking with laughing, M>:End of speaking with laughing, C:Cough, F:Filler, D:Fragment of a word, B:Backchennel, E:Interjection,

Speaking ID	Start-End(sec)	Speaker ID	What speaker said
0142	198.5-199.2	2	(L)
0105	160.7-163.5	4	(<M) そういう訳でもないと思うんだけど (M>).
0205	327.5-330.2	3	観光客というか,(<M) 日本人観光客だけ (M>).
0775	993.7-994.2	1	(C)
0523	860.9-861.6	3	(F ん) まず, これ借りる.
0629	1056.3-1057.2	4	(D だ) 大学生だな.
0430	542.3-542.6	2	(W しりょ).
0521	658.0-658.4	4	(B うん).
0480	601.8-602.5	2	(E おー).
0428	540.1-541.8	1	なんか白黒の写真に持ってるやつ.

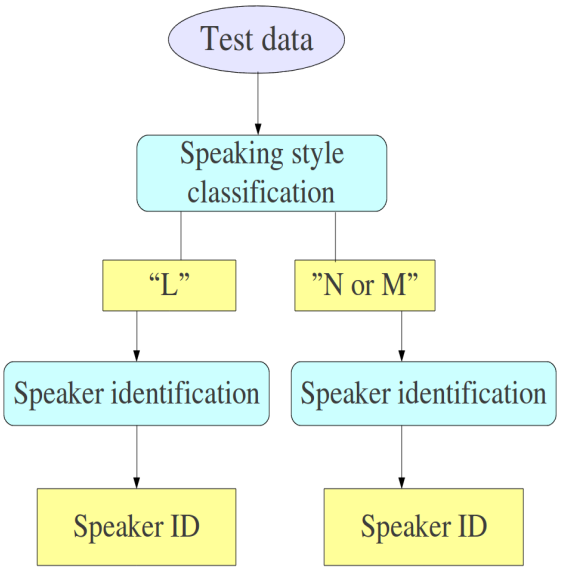

Fig. 2　Proposing two step method

3　発声様態の分類を含む話者認識

提案手法のフローを Fig.2 に示す．まず，入力された発声区間を特徴量 (Super Vector) に変換する．そして，“L” または “N または M” のカテゴリに分類する．その後，このカテゴリ毎に話者を認識して入力発声がどの話者のものか判定する．M を N に含めたのは，L に比べ時間長が大きく，笑い声よりもそれ以外の発声が多く含まれていると考えたからである．

Table 2　Database(sec)

Speaker ID	L	M	N	Total
1	56.9	118.0	316.1	491.1
2	109.0	347.1	624.9	1080.9
3	96.6	267.8	375.7	740.1
4	12.8	108.5	254.5	375.8
Total	275.3	841.3	1571.2	2687.8

発声様態の分類及び話者認識には，GMM と Support Vector Machine (SVM)[8] を用いる．Universal Background Model (UBM) と呼ばれる多数話者の音声から学習した GMM を初期モデルとして，GMM の適応を行う．そして，GMM の各混合成分における平均ベクトルを全成分に渡って連結した Supervector を作り，それを SVM の入力とする．SVM のカーネルとして線形カーネルを用いた．

Table 3　Accuracy of speaker recognition(%)

	Total
Proposed	67.3
Baseline	65.8

Table 4　Accuracy of speaking styles classification(%)

	L	N or M	Average
Proposed	83.3	92.7	88.0

4　実験

4.1　実験条件

GMM の学習と適応化に用いる音響特徴量には，12 次元の MFCC 及びそのパワーと，その 1 次差分と 2 次差分を利用する．その際，Hamming 窓を用い，分析窓長は 25 ms，フレーム間隔は 10 ms とした．UBM の混合数は 32 である．正しく認識できた区関数を全区間数で割った値を評価基準とする．ベースラインは，L，M，N のカテゴリを区別せずに学習した SVM で話者認識を行う手法とした．なお，学習データには，データ A と B からランダムに取り出された L 130.8 秒，M 403.1 秒，N 875.3 秒の発声区間を使用し，残りは評価データに使用した．

4.2　実験結果

提案手法とベースラインの比較結果を Table 3 に示す．平均して提案法のほうが 1.5 ポイント話者認識率が良く，提案法の効果が確認された．次に，第 1 段階の発声様態分類の正解率を Table 4 に示す．最後に発声様態に分類する際の SVM の閾値を変更した場合の結果を Fig. 3 に示す．これより閾値を適切に選ぶことが重要であることがわかる．

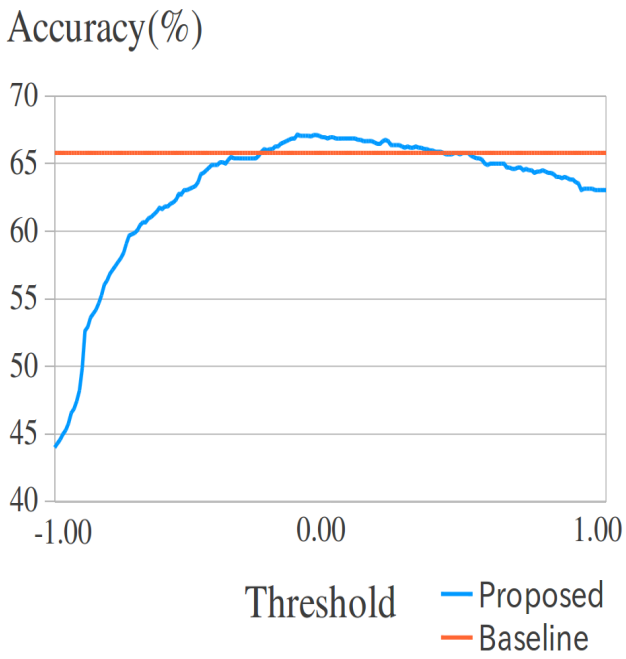

Fig. 3　Performance change of the proposed speaker recogniton method according to the threshold

5　おわりに

会議音声の話者認識に発声様態依存のモデルを用いることを提案し，用いない場合に比べて 1.5 ポイントの改善を得た．

今回発声様態として “笑い声” に着目したが，この他の発声様態についても，分類の後に話者認識を行う場合の性能を検討する必要がある．特に，“相槌” については自動検出が比較的良好に行われることが示されているため [9, 10]，その分類による検証をする必要がある．本稿では人手で区切られた発声区間の情報を用いて実験を行った．しかし，発声区間を自動的に求めた場合の評価を行う必要がある．また，今回の実験では他者の声が重畳する部分を特に除外せず，話者認識性能の評価を行っている．その影響を評価するため，今後は重畳区間を除いた場合の性能調査を行う予定である．

参考文献

[1] 横山　諒 他, “相互スペクトル減算と振幅スペクトル相関を用いた会議音声の重畳区間検出,” 日本音響学会春季発表会講演論文集, pp.13–14, 2012.

[2] X. Anguera, J. Bonastre, “Fast speaker diarization based on binary keys,” ICASSP, pp.4428–4431, 2011.

[3] Y. Nasu *et al.*, “Cross-channel spectral subtraction for meeting speech recognition,” ICASSP, pp.4812–4815, 2011.

[4] O. Vinyals, G. Friedland, “Modulation spectrogram features for improved speaker diarization,” INTERSPEECH, pp.630–633, 2008.

[5] D. Vijayasenan, F. Valente, “Speaker diarization of meetings based on large TDOA feature vectors,” ICASSP, pp.4173–4176, 2012.

[6] K. Laskowski *et al.*, “Contrasting emotion-bearing laughter types in multiparticipant vocal activity detection for meetings,” ICASSP, pp.4765–4768, 2009.

[7] 田中宏季 他, “自閉症児支援に向けた自然対話音声の笑いの種類分析,” 日本音響学会秋季発表会講演論文集, pp.375–378, 2011.

[8] W. Campbell *et al.*, “SVM based speaker verification using a GMM supervector kernel and NAP variability compensation,” ICASSP, pp.97–100, 2006.

[9] 児島宏明 他, “認知症者を対象とした情報支援ロボットとの対話における相槌の認識,” 日本音響学会秋季発表会講演論文集, pp.437–438, 2011.

[10] 榎本美香, 石本祐一, “「うん」の音響的系譜 ～ 応答・証人・相槌の自動検出に向けて～,” 情報処理学会研究報告, vol.2009-SLP-77, no.23, pp.1–6, 2009.