様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ６ 月 １５ 日現在

機関番号：１３８０１
研究種目：基盤研究(C)（一般）
研究期間：2013～2015
課題番号：２５３３０２１０

研究課題名（和文）多人数における多様な会話形態に頑健な話者ダイアライゼーションに関する研究

研究課題名（英文）A Study on Robust Speaker Diarization to Various Speaking Styles for Multi-party Conversations

研究代表者

西田　昌史（Nishida, Masafumi）

静岡大学・情報学部・准教授

研究者番号：８０３６１４４２

交付決定額（研究期間全体）：（直接経費）　　2,200,000円

研究成果の概要（和文）：本研究では，多様な発話形式に頑健な話者クラスタリングを実現するために，主成分分析による空間分離手法を用いて発話ごとに音韻性と話者性の分離を行い，さらに，発話内分散に応じて最適な話者空間の次元数を設定することで，音韻性を抑制した話者空間を構築する話者クラスタリング手法を提案した．従来のBICを用いた手法とGMMを用いたCLRによる話者クラスタリング手法との比較実験を行った結果，提案手法が最も高いクラスタリング精度を実現した．

研究成果の概要（英文）：We proposed a speaker clustering method using Gaussian mixture model in flexibly selected speaker subspace based on variance of intra-utterance in order to realize a robust speaker clustering to various speaking style. We carried out speaker clustering experiments compared with conventional methods based on Bayesian information criterion and Gaussian mixture model in an observation space. The experimental results showed that the proposed method can achieve higher clustering accuracy than conventional methods.

研究分野：音声情報処理

キーワード：多人数会話　話者ダイアライゼーション　発話形式　音韻性　話者性　話者内分散　話者間分散

１．研究開始当初の背景

近年、コンピュータとネットワーク技術の進歩に伴い、ミーティングや講演、対話における音声などのマルチメディア情報をクラウドに蓄積することが容易になりつつあり、これらのマルチメディア情報から重要な情報を自動的に取得する技術が望まれている。

これを踏まえて、2012 年の 5 月に情報処理学会音声言語情報処理研究会にて、東京工業大を中心に国立情報学研究所や日本電信電話（株）といった産・官・学の研究者により「音声・音響クラウドワーキンググループ」が立ち上がり、研究代表者もメンバーとして参加している。その中で、本研究では多人数による多様な会話形態の音声を対象にした話者ダイアライゼーションに着目した。

話者ダイアライゼーションは、会話中から無音区間を取り除き発話区間のみを抽出し、発話区間中にいつ話者が交替したかを判別する話者交替の判別と、得られた発話区間を同一話者ごとにクラスタリングする話者分類の処理から構成される。

このような話者ダイアライゼーションを多人数による会話音声に適用することで、いつ話者が交替しているか、どの発話同士が同一話者であるか、どの発話がオーバーラップしているかといった情報が抽出され、議事録の作成や会話のマイニング、特定話者の発話検索などに利用することが期待される。

２．研究の目的

実際の環境では、ミーティング内の話者が知り合いの場合と知り合いでない場合があり、同じメンバーで異なる話題についてミーティングが行われることも多い。また、はじめは知り合いでないメンバーであるが、会話を繰り返していくうちに親密度が高まることも想定される。このような状況では、知り合いかどうかや話題に応じて発話内容や話し方が大きく異なると考えられることから、これらの影響について分析を行い多様な会話形態に頑健な話者ダイアライゼーション技術が必要となる。また、これまで日本語での多人数会話を多数収録したコーパスは存在しない。

そこで、本研究では話題と対人関係の違いに着目した多人数における多様な会話形態の音声コーパスを構築し、話題と対人関係の違いが会話時における発話の音響的・言語的な変動ならびに会話の振る舞い方に与える影響を解明する。また、多人数における多様な会話形態の音声を対象に、どのタイミングで話者が交替しているかの判別ならびに、どの発話同士が同一話者であるかを自動的に分類する処理といった頑健な話者ダイアライゼーション技術の実現を目指す。

３．研究の方法

(1)初年度は、多人数会話において話題内容の違いによる発話動作の影響について分析を行うため、日本人学生 3 名による日本語と英語の会話音声を収録する。その際、趣味などの自由な話題に関する会話（自由会話）と 3 名で強調して一定の結論に達する会話（目的会話）の 2 種類のデータを収集する。

収録した会話データに対して総発話時間、平均発話長の母語である日本語と第二言語である英語間での違い及び話題内容の違いによる影響について分析を行う。

総発話時間および平均発話長に関して、使用言語と話題内容を被験者内要因とし、被験者の相対的な英語能力を表す Rank を被験者間要因として ANOVA 分析を行う。

(2)二年度目は、従来では、読み上げ音声などを対象として、主に個人性を考慮して話者モデルを学習する事で認識を行ってきたが、発話形式の違いによる認識率への影響は十分に検討されていない。そこで、講演音声を対象に発話形式の違いによる認識率への影響を話者内分散と話者間分散の観点から分析を行い、発話形式に頑健な話者認識手法について検討を行う。

具体的には、UBM(Universal Background Model)の特徴空間に判別分析を適用し、得られた空間に音声データを射影して UBM-MAP(Maximum A Posteriori)による話者モデルの学習を行う手法を提案する。UBM 特徴空間に対して判別分析を適用する場合、UBM の学習に用いた 1 話者分の学習データを 1 クラスとして判別分析を適用して固有値行列を求める。固有値行列を用いて UBM 学習データを射影し、UBM を再学習する．固有値行列で射影した話者モデルの学習データに対して、再学習した UBM を MAP 推定して話者モデルを学習する。同様に、認識用データも射影して話者認識を行う。提案手法である UBM 特徴空間に対する判別分析の適用の流れを図 1 に示す。

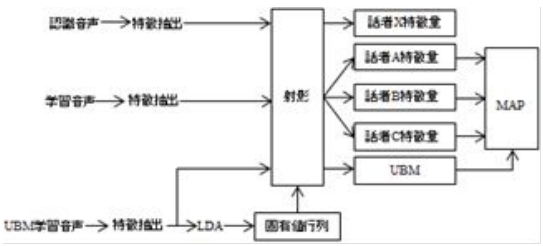

図 1 UBM 特徴空間に対する判別分析の適用の流れ

(3)三年度目は、多様な発話形式に頑健な話者クラスタリングを実現するために、音声データに含まれる音韻性と話者性に着目した。音声データから音韻性と話者性を分離することができれば、音韻性を抑制することで話者性をより強調することができると考えられる。

話者識別と話者照合においては、主成分分析により得られる分散が大きい空間は音韻

性、分散が小さい空間は話者性を表しているとみなして、音韻性を抑制した話者空間に音声データを射影し、話者空間上でGMM(Gaussian Mixture Model)を学習する手法が提案され、有効性が示されている。音韻ベクトルと話者ベクトルの分離の様子を図 2 に示す。しかしながら、従来の話者クラスタリング手法では音韻性と話者性の分離という観点で処理されてこなかった。また、多人数会話では発話ごとに発話時間が異なるため、発話に含まれる音韻のばらつきが話者モデルを構築する際に影響を与えると考えられる。

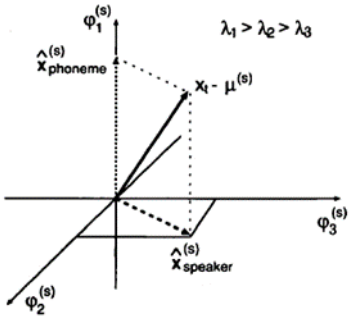

図 2　音韻ベクトルと話者ベクトル

そこで、発話ごとに発話内分散を考慮した話者空間を構築し、統計的手法である GMM を学習することで音韻による影響を抑えた話者クラスタリング手法を提案する。さらに、固有値の累積寄与率に着目した発話内分散に応じて最適な話者空間の次元数を設定することで、音韻性を抑制した話者空間を構築する手法を提案する。

４．研究成果

(1)初年度は、多人数会話において話題内容の違いによる発話動作の影響について分析を行うため、日本人学生 3 名による日本語と英語の会話音声を収録した。その際に、趣味などの自由な話題に関する会話（自由会話）と 3 名で協調して一定の結論に達する会話（目的会話）の 2 種類のデータを収集した。

収録した会話データに対して総発話時間、平均発話長の母語である日本語と第二言語である英語間での違いおよび話題内容の違いによる影響について分析を行った。その際、グループごとに TOEIC スコアを用いて降順に Rank1、Rank2、Rank3 と順位付けを行うことで参加者の英語能力の違いを表現した。

総発話時間および平均発話長に関して、使用言語と話題内容を被験者内要因とし、被験者の相対的な英語能力を表す Rank を被験者間要因として ANOVA 分析を行った。その結果、総発話時間に関しては使用言語の主効果と話題内容の主効果で有意となり、また使用言語と英語能力を表す Rank の交互作用で有意傾向であることから、使用言語と英語能力、話題内容の違いがそれぞれ総発話時間に影響することが示された。さらに、平均発話長に関しては使用言語の主効果で有意となり、使用言語の違いが発話の長さに影響することが示された。

会話の活発さを示す総発話時間での評価では、会話テーマに関わらず第二言語である英語よりも母語である日本語での会話でより多く話す傾向が示された。

話題内容に関しては、3 人で協調して結論を出す目的会話の方が趣味などの自由会話よりも発話時間が短くなる傾向が示された。なお、英語能力の最も低い話者の発話時間が英語発話では短くなる傾向が示唆された。

(2)二年度目は、日本語話し言葉コーパスの学会講演と模擬講演を対象に分析を行った結果、模擬講演は話者内分散が小さく話者間分散が大きいため認識率が高く、学会講演は話者内分散が大きく話者間分散が小さいため認識率が低い傾向があり、発話形式の違いが話者認識に影響を与えることが明らかになった。

これらの結果を踏まえて、学会講演と模擬講演からなる多人数の音声データで学習した UBM の学習データに対して判別分析を行うことで、発話形式の違いを考慮した特徴空間を構成する手法の評価を行った。評価結果を表 1 に示す。

表 1 UBM128 分布時の話者認識結果(%)

学会講演					
従来手法	79.7				
UBM 分割	**79.9**				
次元数	20	21	22	23	24
UBM 分布 LDA	76.3	76.5	76.6	77.4	75.8
UBM 特徴空間 LDA	76.2	76.2	76.9	77.6	**81.3**
模擬講演					
従来手法	94.3				
UBM 分割	**94.4**				
次元数	20	21	22	23	24
UBM 分布 LDA	92.7	92.5	92.5	92.9	92.9
UBM 特徴空間 LDA	92.5	92.6	92.9	93.2	**94.8**

学会講演と模擬講演合わせて 200 名による話者認識実験を行った結果、従来の UBM-MAP 手法では学会講演にて 79.7%、模擬講演にて

94.3%、提案手法では学会講演にて 81.3%、模擬講演にて 94.8%の認識率が得られ、いずれの講演音声に対しても認識精度の改善が得られた。したがって、提案手法により講演の発話形式を考慮した話者認識を実現することができた。

また、収録した多人数会話データに対して、音声や視線の動きなどの会話動作情報のタグ付けを行った。聞き手が話し手を見ている割合に着目して分析を行った結果、自由な会話と課題を達成するための目的会話においてどちらも話し手を見ている割合に違いがないことが明らかになった。また、会話中の沈黙の時間が自由会話に比べて目的会話の方が長いことが明らかになった。これらの結果を踏まえて、会話形式が異なる場合に音響情報のみならず沈黙などの発話動作を話者交替の検出時に考慮することの有効性を検討する必要があると考えられる。

(3)三年度目は、日本語話し言葉コーパスに含まれる講演音声を用いて、任意の長さの無音区間で音声を区切り、複数名の話者の発話順がランダムになるように音声データを作成し、1 セットあたり 5 名と 10 名からなる疑似的な討論音声データを作成した。

これらの疑似的な討論音声データを用いて話者クラスタリングの評価実験を行った。評価結果を表 2,3 に示す。評価には話者クラスタリングにおいて一般的に用いられている DER(Diarization Error Rate)とクラスタ中の発話の純度を表す Purity を使用した。

表 2　話者 5 名のテストセットに対するクラスタリング結果

	DER(%)	Purity(%)
BIC	8.8	90.5
GMM	10.1	89.4
提案手法	6.5	93.2

表 3 話者 10 名のテストセットに対するクラスタリング結果

	DER(%)	Purity(%)
BIC	10.8	87.9
GMM	12.8	86.4
提案手法	6.4	93.1

評価結果から、従来手法である BIC を用いた手法では 5 名のテストセットに対する DER は 8.8%，GMM を用いた CLR(Cross Likelihood Ratio)による手法では 10.8%であったのに対し、提案手法では 6.5%であった。また、10 名のテストセットに対する BIC を用いた手法の DER は 10.8%、CLR による GMM を用いた手法は 12.8%であったのに対し、提案手法は 6.4%であった。

提案手法において、各発話の話者空間における選択された次元数とテストセット内の割合を表 4 と表 5 に示す。例えば、表 4 の話者空間次元数 2-21 とは第 2 軸から第 21 軸までで構成された話者空間を示している。また、話者 5 名のテストセットの中で 263 発話が話者空間 2-21 であり、全体の 92.6%であったことを示している。

表 4 と表 5 の結果から、提案手法は発話毎に異なる話者空間次元数が選択されており、話者空間次元数は話者 5 名のテストセットのとき 2-21、話者 10 名のテストセットのときは 2-20 が最も多く選択されていることが分かった。また、話者 10 名のテストセットでは低次の軸が 3 次元目から選択されている発話はほとんどなく、話者 5 名のテストセットと比べより低次の軸が選択されていた。これは、話者 10 名のテストセットの方が各発話の時間長が短く、得られた固有値が小さかったためであると考えられる。

表 4　話者 5 名のテストセットにおいて選択された話者空間次元数

話者空間次元数	発話数	割合(%)
2-21	263	92.6
2-22	5	1.8
3-21	1	0.4
3-22	15	5.3

表 5　話者 10 名のテストセットにおいて選択された話者空間次元数

話者空間次元数	発話数	割合(%)
2-19	65	7.6
2-20	728	85.0
2-21	60	7.0
3-21	3	0.4

従来の BIC(Bayesian Information Criterion)に基づく手法ならびに通常の GMM に基づく手法に比べて、提案手法によりクラスタリング性能が改善され、話者数が 5 名ならびに 10 名のいずれにおいても評価尺度 purity において 90%以上と高い精度を得ることができた。さらに話者空間は累積寄与率に基づく閾値の設定により，発話内分散を考慮して発話毎に話者空間の次元数を動的に設定する柔軟な構成法となっており、頑健な話者クラスタリング手法を実現できた。したがって、提案手法により多様な発話形式に頑健な話者クラスタリングを実現することがで

きた。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 4 件）

① X. Wang, S. Yamamoto, ”Speech Recognition of English by Japanese using Lexicon Represented by Multiple Reduced Phoneme Sets”, Trans. IEICE, Vol.E98-D, No. 12, pp. 2271－2279, 2015. 査読有

② S. Yamamoto, K. Taguchi, K. Ijuin, I. Umata, and M. Nishida, “Multimodal Corpus of Multiparty Conversations in L1 and L2 Languages and Findings Obtained from it”, Language Resources & Evaluation, 2015. 査読有
DOI: 10.1007/s10579-015-9299-2

③ K. Taguchi, A. Finch, S. Yamamoto, E. Sumita, “Automatic Induction of Romanization Systems from Bilingual Corpora”, 電子情報通信学会論文誌, Vol.J98-D, pp. 381-393, 2015. 査読有
DOI: 10.1587/transinf.2014EDP7236

④ K. Jokinen, H. Furukawa, M. Nishida, and S. Yamamoto, “Gaze and Turn-Taking Behavior in Casual Conversational Interactions”, ACM Transactions on Interactive Intelligent Systems, Issue 3, pp.1-30,2013. 査読有
DOI: 10.1145/2499474.2499481

〔学会発表〕（計 20 件）

① 林 升柯, ”多元的音情報に基づく頑健な音声認識に関する研究”, 日本音響学会春季研究発表会, 2016 年 3 月 9 日～11 日, 横浜桐蔭大学（神奈川県横浜市）.

② 安藤 純平, ”非侵襲簡易型身体状況認識システムに関する研究”, 日本音響学会春季研究発表, 2016 年 3 月 9 日～11 日, 横浜桐蔭大学（神奈川県横浜市）.

③ M. Nishida, “Daily Activity Recognition Based on Acoustic Signals and Acceleration Signals Estimated with Gaussian Process”, APSIPA, 2015 年 12 月 16 日～19 日, Hong Kong (China).

④ 大高 祥裕, “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”, 第 13 回情報学ワークショップ, 2015 年 12 月 5 日, 名城大学（愛知県名古屋市）.

⑤ I. Umata, ”Quantitative analyses of Gaze Activity during Silence: Comparison between Native-language and Second-language Conversations”, EAP Cogsci, 2015 年 9 月 25 日～27 日, Torino (Italy).

⑥ K. Ijuin, “Eye Gaze Analyses in L1 and L2 Conversations: Difference in Interaction Structure”, TSD, 2015 年 9 月 14 日～17 日, Plzen (Czech).

⑦ T. Hayashi, “Daily Activity Recognition Based on DNN Using Environmental Sound and Acceleration Signals”, EUSIPCO, 2015 年 8 月 31 日～9 月 4 日, Nice (France).

⑧ 阿部将和, ”日本語母語話者による第二言語音声を対象にした話者認識”, 音響学会 2015 年春季研究発表会, 2015 年 3 月 16 日～18 日, 中央大学（東京都文京区）.

⑨ 伊集院幸輝, ”第二言語での少人数会話における聞き手の視線動作の分析”, 電子情報通信学会総合大会, 2015 年 3 月 10 日～13 日, 立命館大学（滋賀県草津市）.

⑩ 荒木智彰, “複数人会話における対話者への視線の自動推定”, 電子情報通信学会総合大会, 2015 年 3 月 10 日～13 日, 立命館大学（滋賀県草津市）.

⑪ 堀内保大, ”第二言語習熟度による母語と第二言語間の視線動作の相違分析”, 電子情報通信学会総合大会, 2015 年 3 月 10 日～13 日, 立命館大学（滋賀県草津市）.

⑫ 中辻康太, “講演音声における発話形式を考慮した話者認識手法の検討”, 第 16 回音声言語シンポジウム, 2014 年 12 月 15 日～16 日, 東京工業大学（神奈川県横浜市）.

⑬ K. Ijuin, “Eye Gaze Analyses in L1 and L2 Conversations: From the Perspective of Listeners’ Eye Gaze Activity”, UMMMI-ICMI, 2014 年 11 月 16 日, Istanbul (Turkey).

⑭ K. Taguchi, “Multimodal Japanese Corpus of Multi-party Conversation on Two Different Topic Types”, Oriental COCOSDA, 2014 年 9 月 10 日～12 日, Phuket (Thailand).

⑮ 馬田一郎, “マルチモーダルコーパスを用いた母語と第二言語の沈黙時の視線行動の相違分析”, 第 13 回情報科学技術フ

ォーラム，2014 年 9 月 3 日～5 日，筑波大学（茨城県つくば市）．

⑯ 伊集院幸輝，”少人数会話での注視対象に関する比較分析”，第 13 回情報科学技術フォーラム，2014 年 9 月 3 日～5 日，筑波大学（茨城県つくば市）．

⑰ 野本沙斗子，”第二言語による少人数会話での話題内容の発話動作への影響分析”，電子情報通信学会総合大会，2014 年 3 月 18 日～21 日，新潟大学（新潟県新潟市）．

⑱ M. Nishida, “Gaze and Turn-Taking Behavior in Casual Conversational Interactions”, International Conference on Intelligent User Interfaces”,2014 年 2 月 24 日～27 日, Haifa(Israel).

⑲ I. Umata, “Effects of Language Proficiency on Eye-gaze in Second Language Conversations: Toward Supporting Second Language Collaboration”, International Conference on Multimodal Interaction, 2013 年 12 月 9 日～13 日, Sydney(Australia).

⑳ K. Taguchi, “Differences in Interactional Attitudes in Native and Second Language Conversations: Quantitative Analyses of Multimodal Three-Party Corpus”, Annual Meeting of the Cognitive Science Society, 2013 年 7 月 31 日～8 月 3 日, Berlin(Germany).

〔図書〕（計 0 件）

〔産業財産権〕
○出願状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
西田 昌史（NISHIDA MASAFUMI）
静岡大学・情報学部・准教授
研究者番号：80361442

(2)研究分担者
山本 誠一（YAMAMOTO SEIICHI）
同志社大学・理工学部・教授
研究者番号：20374100