TR-SLT-0025

クラスタリングによる音声対話データに対する自動構造化の検討

Automatic Finding of Structure in Spoken Dialogues by Clustering

山下 洋一 Yoichi YAMASHITA
谷 智洋 Tomohiro TANI
松井 知子 Tomoko MATSUI

2002.8.31

音声対話における構造の自動抽出を実現するための要素技術として、対話における発話のクラスタリングおよび会議音声における話者セグメンテーションについて述べる。音声対話における発話は、質問・応答などの発話タイプや対話における話題などの対話の状況によって、言語的な特徴が異なっている。クラスタリングによって類似した発話をまとめ、複数の言語モデルを作成することによって音声認識の性能が改善することを示す。合わせて、対話における典型的発話系列を自動抽出するための基本的アイディアについて述べる。次に、会議音声の話者セグメンテーションを行うために、多数話者との類似度によって発話の話者性を表現して発話を話者ごとに分類する手法について述べる。

（株）国際電気通信基礎技術研究所
音声言語コミュニケーション研究所
〒619-0288 「けいはんな学研都市」光台二丁目2番地2 TEL: 0774-95-1301

Advanced Telecommunication Research Institute International
Spoken Language Translation Research Laboratories
2-2-2 Hikaridai "Keihanna Science City" 619-0288, Japan
Telephone: +81 774 95 1308
Fax: +81 774 95 1308

目次

1 はじめに

二名の話者による対話や多数の話者による会議、あるいは一人の話者による講演やナレーションなど、ある程度の長さを持つ音声データには様々な構造が内在している。例えば、対話における質問と応答なら成るやり取りの構造、意味的な話のまとまりとしての話題の構造、会議における話者の交替、講演における起承転結のような全体としての談話の構造などである。

大量の音声データが容易に蓄積できるようになり、コーパスに基づいた問題解決手法の提案や、大量データを対象とする情報検索などのアプリケーションの研究開発が精力的に行われている。このような大量データ指向の方法論を進める上での重要な問題として、データの構造化が挙げられる。音声データには様々な構造が見られることから、構造化を行う視点も多岐に渡る。二名の話者による対話音声に対しては、発話単位の同定、発話タイプの分類スキームの検討、発話タイプの自動認識、あるいは話題のセグメンテーションなどが行われている [1]。また、講演音声に対する重要文の同定、会議音声に対する話者セグメンテーションなどが試みられている [2, 6, 7]。このような構造化を行うことによって、必要なデータのみを取り出すことやデータの属性に応じた処理などが可能になる。

データを自動的に構造化するための手法の一つとしてクラスタリングが挙げられる。クラスタリングは、あらかじめ決められた尺度に基づいて全対象データをサブセットに分割することから、類似したデータセットやセグメンテーション結果を取り出すことができる。本報告では、クラスタリングに基づいた、対話音声における文クラスタリングによる発話分類、会議音声の自動話者セグメンテーションについて述べる。

2 対話音声データに対する文クラスタリング

音声対話における発話は、質問・応答などの発話タイプや対話における話題などの対話の状況によって、言語的な特徴が異なっている。このような対話の状況に応じた言語モデルを作成できれば、対話音声認識での性能改善が期待できる。また、目的指向の対話では、それぞれのタスクにおいて、典型的な対話の進行パターンがあると考えられ、そのようなパターンやパターンの構成要素を同定できれば、VoiceXML などによる音声対話の記述に非常に有用なものとなる。本章では、類似した発話文のセットを得るためにクラスタリングを利用する手法について述べる。

2.1 発話文クラスタリング

(2.1.1) 距離尺度

発話文のクラスタリングを行うには、まず、文間の距離を定義する必要がある。

情報検索の分野などでは、一般に、単語の出現数を並べた高次元ベクトルで文章の特徴を表現する、いわゆるベクトル空間モデル (あるいは “bag-of-word(BOW)” モデル) がよく使われる [3, 4]。すなわち、注目する単語数を M、単語を $w_m(m = 1, \cdots, M)$ として、単語 w_m の文章 D_i での出現回数を d_{im} としたとき、文章中での単語の出現順序を無視して単語の出現数のみを考慮することにより、文章 D_i は、文書ベクトル

$$d_i = (d_{i1}, d_{i2}, \cdots, d_{iM}) \tag{1}$$

によって表現される。さらに、文章 D_i と D_j の距離 $d_D(i,j)$ を

$$d_D(i,j) = \frac{d_i \cdot d_j}{\|d_i\| \cdot \|d_j\|} \tag{2}$$

$$= \frac{\sum_{m=1}^{M} d_{im} d_{jm}}{\sum_{m=1}^{M} d_{im}^2 \sum_{m=1}^{M} d_{jm}^2} \tag{3}$$

で定義して文章間の非類似度を計算しクラスタリングが行われる。情報検索では、単語数は通常数万にも及び文書ベクトルの空間が非常にスパースになってしまう。このためメモリ節約やノイズ除去の目的から潜在的意味インデキシング (latent semantic indexing; LSI) などの手法によって次元圧縮が用いられることが多い。

対話中の発話文を単位としたクラスタリングにおいても、同様にベクトル空間モデルによって、文の特徴および文間距離を定義することはできる。しかし、対話での発話は「はい」「ありがとうございます」「○○です」など非常にすくない単語で構成される文も多く、情報検索での文章に比べ、特徴ベクトルはさらにスパースなベクトルとなってしまう。そこで、発話文

間の距離を単語間の距離に基づいて以下のように定義した。発話 U_i に現れる語彙中の単語数を n_i、そのうちの p 番目の単語を w_{ip}、単語 w_{ip} と単語 w_{jq} 間の距離を $d_W(w_{ip}, w_{jq})$ とするとき、発話 U_i と U_j の距離 $d_U(i,j)$ を

$$d_U(i,j) = \frac{1}{n_i + n_j} \left\{ \sum_{p=1}^{n_i} \min_q d_W(w_{ip}, w_{jq}) + \sum_{q=1}^{n_j} \min_p d_W(w_{ip}, w_{jq}) \right\} \tag{4}$$

で求める。これは、発話中の個々の単語に対して、比較相手の発話中の単語のうち最も似ている単語との距離を求め、それを平均した値となっている。

単語間の距離を求める方法としては、シソーラスなどの人手によって作成された情報をもとに算出する手法もある。本報告では、先見的知識に依らずコーパスからの学習によってクラスタリング結果を得ることを目指して、発話内における単語の共起情報に注目して単語間距離を定義する。先の定義と同様に、注目する単語数を M、単語を $w_m(m = 1, \cdots, M)$ とする。単語 w_p がコーパス中の一発話内において単語 w_m と共起した回数を a_p^m としたとき、単語 w_p の単語共起ベクトル c_p は、

$$c_p = \frac{1}{\sum_{m=1}^{M} a_p^m} \left\{ a_p^1, a_p^2, \cdots, a_p^M \right\} \tag{5}$$

$$= \left\{ c_p^1, c_p^2, \cdots, c_p^M \right\} \tag{6}$$

で与えられ、単語 w_p と w_q との単語間距離 $d_W(w_p, w_q)$ は、

$$d_W(w_p, w_q) = \sqrt{\sum_{m=1}^{M} (c_p^m - c_q^m)^2} \tag{7}$$

のユークリッド距離として定義した。

(2.1.2) クラスタリング手法

クラスタリングは、式 (4) によって定義された距離尺度を用いて k-means (k- 平均) 法で行う。通常の k-means 法では、クラスタ内のデータの特徴ベクトルを平均することによってクラスタ中心を求める。本手法では、式 (4) にように、出現単語ごとの距離計算に基づいて発話間の距離を計算するため、平均操作によってクラスタ中心の (非整数の出現回数も認めるような) 出現単語リストを求めると、非ゼロの単語が急増して大量データのクラスタリングにおいて膨大な計算時間を要することになる。そこで、平均操作によって中心を求める代わりに、クラスタ内におけるデータのうち、自分以外のデータとの距離の総和が最小となるデータが最も中心に近い位置に位置すると考えて、これをクラスタ中心とすることにした。これによって、あらかじめ全ての発話間距離を計算し記憶しておくことによって、発話間距離の計算はテーブル参照のみでクラスタリング処理を実装できる。

2.2 複数の言語モデルの同時利用に基づく音声認識

(2.2.1) 認識手法

一つの対話はいくつかの話題や局面から構成されており、それぞれの話題や局面ごとに用いられる内容語や文末表現がある程度限定されていると考えることができる。そのような対話の状態に合った言語モデルを作成しておけば、発話の認識率の向上が期待できる。これまでにも対話の状態の変化をトレースしながら言語モデルを切り替えて認識を行う手法が提案されている。この手法では、

- 対話状態の同定誤りが起こると、それ以後の音声認識にも悪影響を残す。
- 対話の状態を人手で決定する必要がある。
- 状態変化のモデルを作成しておく必要がある。

などの問題点がある。複数の言語モデルの利用を考える時、選択的に一つのモデルを用いるのではなく、同時に複数のモデルを用いる手法も考えられる。異なる言語モデルを用いる複数の音声認識システムへ音声がそれぞれ入力され、尤度最大の結果を選択し、最終的な認識結果とすれば良い。この手法でも、発話ごとに用いられる言語モデルの関係を対話モデルとして利用することもできるが、必ずしもそのような対話モデルがなくても動作する。複数の言語モデルを作成するには、何らかの基準で類似した文セットを複数作成する必要がある。対話コーパスに対して人手で話題などのタグを付与して分類することもできるが、タグ付けのコストが高くつく上に、タグセット設計の妥当性、タグ付けの一貫性の問題も解決しなければならない。一方、文クラスタリングによる文の分割では、このような問題を避けることができる。

本研究では、クラスタリングによって得られた個々のクラスタ内の発話文で一般的な言語モデルを適応化し、図1に示すように、ベースラインの言語モデルと合わせて複数の言語モデルを並列的に同時使用することにより音声を認識する手法を試みる。

(2.2.2) 対話データ

クラスタリングに用いた発話文は、旅行対話タスクに関する対話データで、ATR の V8 の ANS 形式のデータのうち、`train.list` に入っているデータから、後述する対象単語の出現頻度の多いものを抜粋して用いた。例として図 2(a) に

`/RR/Recognition/ResearchJ/ldata/19990817/ANS/SDC71001.A.ANS`

の内容を示す。図 2(b) は、辞書

`/RR/Recognition/ResearchJ/lmodel/19990817/LEX.W`

を参照することにより、内容がわかりやすくなるように通し番号と単語表記を付加したものである。`tarin.list` 中のデータ量を表 1にまとめておく。このうち、対象単語が 2 回以上出現する 83,211 文をクラスタリング対象データとした。

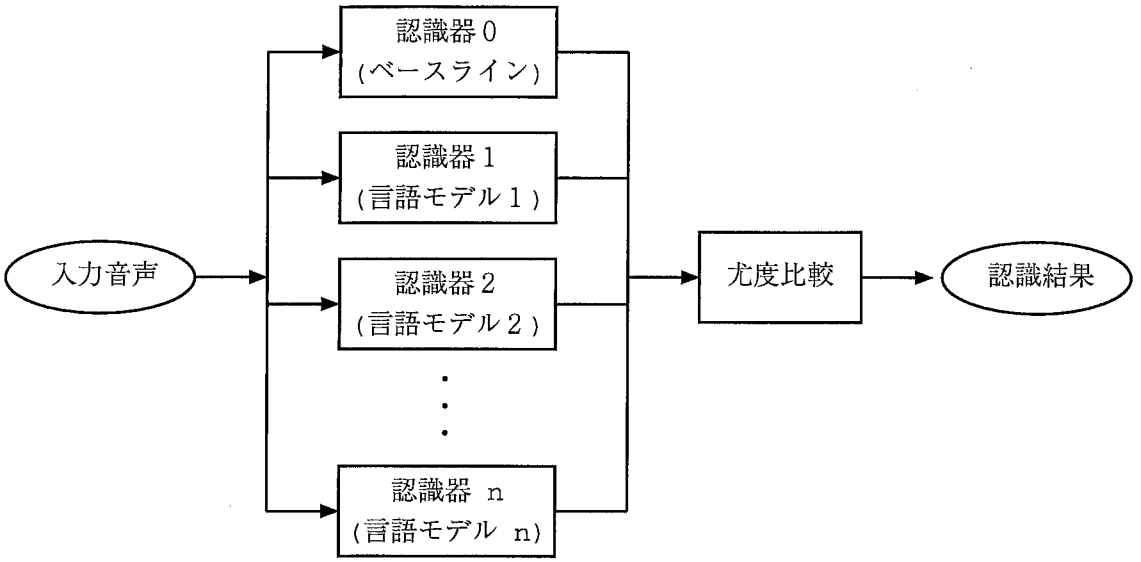

図 1: 複数言語モデルを用いた認識システムの概要

表 3は、対話コーパス中の各品詞の出現頻度情報をまとめたもので、品詞名に * が付いている 8 種類の名詞、本動詞、日時、数詞、数字の 12 品詞のうち、出現回数が 6 回以上の単語を (2.1.1)で述べた単語共起および発話間距離の計算に用いる対象単語とし、単語セットを構成した。ただし、日時に関しては、単語の違いを無視して 1 単語として扱い、数詞と数字に関しても、単語の違いを無視して 2 品詞で 1 単語として扱うこととした。この結果、単語セットは 3,710 単語から成っている。品詞ごとの内訳を表 2に示す。

(2.2.3) 音声認識における評価

音声認識エンジンは、当研究所で開発された ATRSPREC を用いた。ベースラインとなる言語モデルとして、表 1の全データで学習した多重クラス複合 N-gram を用いた。音響モデルには、窓長 20msec、フレームシフト 10msec で抽出した 25 次元の特徴ベクトル (12 次 MFCC、12 次 Δ MFCC、Δ power) によりモデル化した男女別、5 混合ガウス分布の状態共有化 HMM (1400 状態) を用いた。単語辞書サイズは、21,750 単語である。評価データには、学習データと同一タスクの男女 42 名、470 発話を用いた。

学習コーパスを本手法でクラスタリングし、それぞれのクラスタごとにベースラインの言語モデルを適応化することによって言語モデルを作成した。なお、クラスタ数は 2, 4, 8, 16, 32 を試みた。言語モデルの適応化は、各クラスタ中の文に対するカウント数を定数倍させて行い、クラスタ数が 8 の場合の各クラスタの文数、言語モデル適応化のための倍数を表 4に示す。認識結果を表 5に示す。比較のため、クラスタ内の文による言語モデルのエントロピーの総和が最小になるようにクラスタリングを行った場合 [5] (エントロピー基準) の結果も合わせて示

表 1: 対話コーパス (train.list) 中のデータ量

カテゴリ	量
ファイル数	7,195
総単語数	1,321,373
異なり単語数	16,332
総文数	158,488
対象単語が 1 回以上出現する文数	110,893
対象単語が 2 回以上出現する文数	83,211
対象単語が 3 回以上出現する文数	62,425

表 2: 対象単語の品詞ごとの内訳

品詞	単語数
普通名詞	1,495
外来普通名詞	724
形容名詞	112
外来形容名詞	7
サ変名詞	341
外来サ変名詞	84
サ変形容名詞	11
外来サ変形容名詞	2
本動詞	932
日時	1
数詞 & 数字	1
合計	3,710

5
10013
10014
10015
10014
11805
10014
10016
10017
10023
10143
10094
10019
10020
10021
10316
6
5
21435
10021
6
5
13350&26655
10014
10045&10107&10043
10021
6

(a) 単語 ID のみ

1 5 UTT-START
2 10013 そちら
3 10014 の
4 10015 ホテル
5 10014 の
6 11805 会議室
7 10014 の
8 10016 予約
9 10017 を
10 10023 お
11 10143 願い
12 10094 し
13 10019 たい
14 10020 の
15 10021 です
16 10316 けれども
17 6 UTT-END
18 5 UTT-START
19 21435 西日本鉄道株式会社
20 10021 です
21 6 UTT-END
22 5 UTT-START
23 13350&26655 七 & 月
24 10014 の
25 10045&10107&10043 十 & 四 & 日
26 10021 です
27 6 UTT-END

(b) 通し番号 + 単語 ID + 単語表記

図 2: 対話データ例 (SDC71001.A)

表 3: 学習コーパス中の各品詞の出現頻度

品詞名	異なり単語数	総単語数
普通名詞 *	4140	139971
外来普通名詞 *	2596	37040
形式名詞 *	10	9887
外来形容名詞 *	44	174
サ変名詞 *	842	31001
外来サ変名詞 *	197	5332
サ変形容名詞 *	23	804
外来サ変形容名詞 *	2	38
本動詞 *	2809	110010
日時 *	232	28559
数詞 *	18	23874
数字 *	14	19403
ローマ字	25	3210
会社名	637	4239
外姓	129	1889
外名	124	1909
格助詞	23	119581
感動詞	90	46304
間投詞	415	47673
係助詞	8	36459
形容詞	347	18841
形容名詞	273	11101
桁数詞	10	12262
固有名詞	507	2216
終助詞	14	49082
準数詞	2	4

品詞名	異なり単語数	総単語数
準体助詞	1	22343
準体助動詞	15	16088
助数詞	20	2342
助動詞	337	131031
人称接尾辞	7	8480
数代名詞	2	1644
接続詞	118	20789
接続助詞	26	50737
接続副詞	4	1346
接頭辞	5	47347
接尾辞	3	648
代名詞	43	25079
地名	893	10209
日姓	563	4724
日名	242	1596
判定詞	10	58806
副詞	306	28321
副助詞	27	10205
並立助詞	3	4970
補助動詞	113	41504
洋数詞	6	103
連数詞	3	8
連体形容詞	3	305
連体詞	31	8273
連体助詞	17	60732
和数詞	3	2880

表 4: クラスタごとの文数と言語モデル適応化のための倍数

クラスタ番号	文数	倍数
1	788	203.05
2	12,618	12.68
3	2,679	59.72
4	9,503	16.84
5	8,894	17.99
6	7,832	20.43
7	25,907	10.00
8	14,990	10.67

表 5: クラスタ数と認識率 (WA%) の関係

クラスタ数	単語共起情報	エントロピー基準
1	86.0	
2	86.1	86.0
4	86.1	86.1
8	86.9	86.6
16	86.7	86.7
32	86.8	87.1

す。表5より、本手法では、クラスタ数 8 において最も高い認識率を示し、同一クラスタ数のエントロピー基準を用いた場合の結果を上回った。エントロピー基準を用いた場合には、クラスタ数を増加させるほど認識率は向上し、32 クラスタの場合には 提案手法の 8 クラスタでの結果を上回った。

さらに、提案手法、エントロピー基準で得られた両方のクラスタからそれぞれ言語モデルを作成し、音声を認識した場合の結果を表6に示す。両者の結果から得られた言語モデルを合わせて利用することにより、それぞれ単独で用いた場合よりも認識率が向上した。エントロピー基準でのクラスタリングは N-gram 確率に基づいて行われるため、文内の離れた単語間の関係は考慮されない。一方、提案手法では、文中の単語間距離に基づいて文間の距離が定義されるため、離れた単語の関係も文クラスタリングに反映されると考えられる。提案手法とエントロピー基準のクラスタリングは、このような異なった性質に基づいて文セットクラスタを構成しており、それから学習された言語モデルを合わせて利用することにより、相補的な効果により

表 6: 二種類のクラスタリング手法を組み合わせた場合の認識率 (WA%)

クラスタ数	単語共起情報 + エントロピー基準
2+2	86.3
4+4	86.4
8+8	86.9
16+16	87.3
32+32	87.3

表 7: テスト文に対する最尤言語モデルの分布

テスト文	Baseline	L1	L2	L3	L4	L5	L6	L7	L8	適合率 [%]
C1	0	3	0	0	0	0	0	0	0	100.0
C2	2	0	38	3	1	3	1	1	3	73.1
C3	0	0	0	2	0	0	0	0	0	100.1
C4	0	0	0	1	30	0	5	2	0	78.9
C5	1	1	4	0	2	9	1	1	1	45.0
C6	0	0	1	0	2	0	5	0	0	62.5
C7	5	4	5	5	1	5	2	58	4	65.2
C8	3	0	4	0	1	1	1	3	39	100.1

性能が向上したと考えられる。

表 7は、クラスタ数が 8 の場合に、評価に用いたテスト文が所属するクラスタ (C1 ～ C8) と、認識時に選択された言語モデル (L1 ～ L8) との関係を調べた結果である。表中の数字は該当文数である。テスト文が所属するクラスタの言語モデルが選択された割合 (適合率) は平均で 75.0% であった。

2.3 対話における典型的発話系列パターンの自動抽出

目的指向の対話は、いくつかの話題によって構成される。例えば、図 3の対話例では、

```
|-- 0010-0010  挨拶
|-- 0020-0380  宿泊予約
|    |-- 0030-0040  人数
|    |-- 0050-0060  名前
|    |-- 0070-0080  宿泊日
```

```
|    |-- 0090-0100   確認
|    |-- 0110-0140   部屋の種類
|    |-- 0150-0190   料金
|    |-- 0200-0250   朝食
|    |-- 0260-0330   支払い方法
|    |-- 0340-0380   連絡先
|-- 0390-0390   挨拶
```

のような話題構成となっている。このように話題は大きな話題の中により詳細な複数の話題が含まれる、入れ子構造をとることもある。対話コーパスにおける話題の構造が決定できれば、

- 話題に応じた言語モデルの作成、およびそれを用いた音声認識
- 話題を利用した対話モデルの作成
- 話題を利用した対話音声合成
- 話題を利用した発話理解

などの研究が進展することが期待できる。人手での話題決定はコストや信頼性の点からの問題点があることから、自動的な話題決定ができればその意義は大きい。

一つの話題は意味的に関連あるいは類似した発話から成っており、同じ話題は同じような発話系列のパターンから構成されていると考えられる。このような発話系列の典型的パターンを対話コーパスから自動的に抽出できれば、それを話題に関する一種のライブラリとすることで、例えば VoiceXML などによる対話システムの構築において非常に有用なものとなる。

文クラスタリングを適用することによって対話コーパスからの典型的発話系列パターンの自動抽出を行うことを考える。2.2節では、発話を単位として文クラスタリングを行ったが、ここではいくつかの発話からなる発話系列を対象としてクラスタリングすることを考える。(2.1.1) では、発話 U_i の特徴を U_i に出現した単語のみによって表現したが、ここでは U_i の特徴を $U_{i-L}, U_{i-L+1}, \cdots, U_{i-1}, U_i, U_{i+1}, \cdots, U_{i+L-1}, U_{i+L}$ に出現した単語によって表現し、式 (4) によって距離尺度を定義してクラスタリングを行う。このように L 発話に渡って、特徴を平滑化した距離尺度を定義することによって、近接した発話間の距離が小さくなり、意味的に関連した発話のまとまりを自動的に検出することが期待できる。

0010 通訳者: ありがとうございます、ニューヨークシティホテルでございます。
0020 申込者: ［あのー］ホテル予約したいんですけど。空いてますか。
0030 通訳者: かしこまりました。何名様でございましょうか。
0040 申込者: 大人一人に子供二人です。
0050 通訳者: はい、失礼ですが、お客様のお名前は。
0060 申込者: 田中武司です。
0070 通訳者: はい、田中様。いつご予約をご希望ですか。
0080 申込者: 八月十五日の月曜日から八月十六日の火曜日、一泊二日です。
0090 通訳者: はい、では、一泊で大人が一名様、子供が二名様でよろしいですね。
0100 申込者: はい、で、部屋はツインで、できれば安い部屋がいいんですけど。
0110 通訳者: はい、［えー、そうですね］十五日はツインルームがお一部屋、ご用意できます。
0120 － ：［え］ですから、お子様用に簡易ベッドをお入れすることができますが。
0130 申込者: そうですか。一人はちょっと大きいんですよ。
0140 通訳者: ［あ］それなら大丈夫ですよ。簡易ベッドのほうはかなり大きめになっておりますので、大人の方でも十分です。
0150 申込者: そうですか。で、料金のほうはおいくらですか。
0160 通訳者: はい。ツインルームは百三十ドルとなっております。これにサービス料と税金が付きます。
0170 申込者: そうですか。それより安い部屋はないですね。
0180 通訳者: 申し訳ございませんが、ダブルルームのほうは全室満室をちょうだいしております。シングルのほうもです。ま、シングルでしたらどちらにしてもお客様のほうは無理かと思いますが。
0190 申込者: そうですか、分かりました。じゃあ、その（ツイン）ツインでお願いします。
0200 通訳者: はい、かしこまりました。お客様、朝食をお付けしましょうか。
0210 申込者: 朝食はどういったものがあるんですか。
0220 通訳者: はい、コンチネンタル式が十ドル、英国式が十二ドルとなっております。
0230 申込者: そうですか。じゃあ、コンチネンタル式のほうをお願いします。
0240 通訳者: はい、［えー］こちらのほうは、コンチネンタル式の朝食を一名分ですか、それともお子様の分もご用意しましょうか。
0250 申込者: ［そうですね］（二名分）二人分でお願いします。
0260 通訳者: かしこまりました。お客様、現金かカード、どちらでお支払いになりますか。
0270 申込者: カードでお願いします。
0280 通訳者: はい。では、カードナンバーをお願いします。
0290 申込者: ビザカードです。
0300 － ：ナンバー、四九八零、零四五九。
0310 － ：九一九九、五三一三です。
0320 通訳者: はい、分かりました。カードの期限はいつですか。
0330 申込者: ［えーと］来年の四月までですね。
0340 通訳者: かしこまりました。田中様、ご住所とお電話番号、お願いできますか。
0350 申込者: はい、［えー］ワシントンホテル、五〇七号室です。
0360 － ：電話番号が二六七、五二一、零三八七です。
0370 通訳者: 二六七の五二一の零三八七、ワシントンホテルの五百七号室でございますね。
0380 申込者: はい、そうです。
0390 通訳者: はい、田中様。では、わたくしメアリー・フィリップスが担当いたしました。お越しをお待ちしております。失礼いたします。

図 3: 対話例 (TAC22012)

3 会議音声の話者セグメンテーション

音声言語情報処理技術の一つの応用として、会議の議事録自動作成が挙げられる。これを実現するには、会議音声の認識、発話の要約に加えて、話者セグメンテーションが必要になる。話者セグメンテーションは、複数の話者が発話した会議や討論の収録音声に対して、話者交替を検出しそれぞれの話者区間を同定する処理である。

3.1 多数話者モデルに基づいた話者の特徴表現

これまでに行われてきた話者セグメンテーションの研究では、同一話者が発声していると判断された音声区間から話者モデルを構成し、次発話区間が同じ話者かどうかを照合する逐次的な話者照合に基づく手法が一般的であった [6]。このような処理では、次発話が同じ話者かどうかの判定を誤った場合に、誤りが伝搬しセグメンテーションの精度が劣化する恐れがある。

人間の声の話者性は「誰の声に似ているか」という観点から特徴付けることができる。すなわち、話者が違えば「A さんの声に似ている」「B さんの声には似ていない」といった傾向も変わってくる。近年、音声データベースの整備が進み、多数話者の音声が比較的容易に利用可能になってきた。そこで、ある発話の特徴を、あらかじめ用意された多数話者との類似度 (あるいは距離) によって表現し、発話を話者ごとに分類することを考える。多数話者の話者数を J、多数話者を $S_j(j = 1, \cdots, J)$ として、ある発話 U_i の音声と話者 S_j の音声の類似度が p_{ij} であるとすると、発話 U_i の特徴ベクトル V_i は

$$V_i = \{p_{i1}, p_{i2}, \cdots, p_{iJ}\} \tag{8}$$

で表される。会議や討論の収録音声では話者数が既知のことが多いため、収録音声を発話に分割した後で、式 (8) の特徴ベクトルを用いてクラスタ数既知のクラスタリングを行うことによって発話を分類し、一つのクラスタが一人の話者に対応すると考えて話者セグメンテーション結果を得る。発話 U_i と 話者 S_j の類似度としては、話者 S_j の音声のモデルを作成しておき、そのモデルによる発話 U_i の尤度を求めて、p_{ij} とする。

3.2 発話分類

多数話者との類似度による発話の特徴表現に基づく話者分類の基本的性能を調べるため、発話の話者分類を行う予備的実験を行った。多数話者の話者モデルとしては ergodic-HMM (以下、EHMM) と GMM の二種類を試みた。多数話者としては、JNAS コーパスの 306 名 (男性 153 名、女性 153 名) と ATR コーパスの B-set の 10 名 (男性 6 名、女性 4 名) の合計 316 名を用いた。以下、これらの話者を学習話者と呼ぶ。話者モデルの学習に用いた学習データは次のとおりである。

表 8: 話者クラスタリングにおける対象話者

話者番号	性別	発話音韻バランス文のセット ID
m036	男性	A
m005	男性	B
m044	男性	C
m010	男性	D
m131	男性	E
f091	女性	F
f087	女性	G
f144a	女性	H
f001	女性	I
f116	女性	J

- JNAS コーパスの話者 : コーパス中の全発話を用いた。文数は、新聞記事の読み上げおよび音韻バランス文を合わせて約 150 文であり、話者によって若干異なる。
- ATR コーパスの話者 : B-set の全音韻バランス文を用いた。文数は 503 文である。

発話分類の対象とした話者は、表 8に示す 10 名で、全員 JNAS コーパスの話者で学習話者に含まれている話者である。JNAS コーパスでは、一人の話者が音韻バランス文 (ATR 503 文) の 10 セットのうちの一つのセットだけを発声している。異なる話者で同じ文セットが重ならないように話者を選択した。(実際には、JNAS コーパス中で与えられる発声リストにおいて、それぞれの文セットの先頭の話者を男女比を考慮して選択した。)

音声の特徴量は 12 次の MFCC と Δ MFCC および Δ パワーの 25 次元で表現し、分析フレーム 25ms、シフト 10ms で分析した。

(3.2.1) ergodic-HMM での結果

EHMM では、状態数を 5, 10, 20, 30 と変えて実験を行った。出力分布の混合数は 1 で、HTK を用いてモデルを学習した。クラスタリングでは得られる発話クラスタと話者の関係は決まらないため、話者同定率が最も高くなるようにクラスタ番号と話者の対応付けを行い、話者同定率を求めた。結果を表 9に示す。多数話者との類似度による発話の特徴表現では、比較すべき話者の違いによる発話尤度 (式 (8) における $p_{i1}, p_{i2}, \cdots, p_{iJ}$) 間の相対的な差が重要であり、尤度自体の大きさにはあまり意味がない。発話の長さ、発声内容の違いなどによる発話間の尤度の違いを軽減するために発話の尤度の正規化を行った。すなわち、発話の特徴ベクト

表 9: ergodic-HMM による話者同定率 [%]

状態数	発話の尤度の正規化		
	なし	全話者	平均
5	48.9	92.2	99.0
10	50.7	99.2	98.8
20	57.3	99.0	98.8
30	59.2	98.4	98.8

ルを式 (8) に代わり、

$$V_i = \{p_{i1} - \bar{p}_i, p_{i2} - \bar{p}_i, \cdots, p_{iJ} - \bar{p}_i\} \tag{9}$$

とする。$\bar{p}_i$ を求める方法として、

- 全話者の音声のモデルを作成しておき、そのモデルによる発話 U_i に対する尤度を $\bar{p}$ とする。この結果を表 9の「全話者」として示す。
- 複数話者のモデルによる結果の平均値を $\bar{p}_i$ とする。すなわち、$\bar{p}_i = \sum_{j=1}^{J} p_{ij}$ とする。この結果を表 9の「平均」として示す。

の二種類を試みた。表 9から正規化を行うことで話者同定率が大きく向上することがわかる。全話者モデルによる方法と平均尤度を用いる手法では大きな差はなく、簡便さから考えると平均尤度を利用する手法で十分であると言える。また、状態数を変化させても結果にそれほどの差はなかった。

(3.2.2) GMM での結果

GMM では、混合数を 1, 2, 4, 8, 16, 32, 64, 128 と変えて発話の分類を行った。また、EHMM の場合と同様に尤度の正規化を行った。結果を表 10に示す。EHMM の場合と同様に尤度正規化の効果は大きいことがわかる。正規化手法の比較では、平均を使う方が若干話者同定率が良く、混合数の変化に対しても安定した結果となっている。

3.3 クラスタリングによる学習話者の削減

本手法では、あらかじめ用意しておく多数話者 (学習話者) 中に声質の類似した話者が多く含まれていてもあまり意味がなく、代表的な声質の話者がある程度含まれていれば良いと考えられる。そこで、学習話者をクラスタリングして代表話者を選択することによって、話者数 J

表 10: GMM による話者同定率 [%]

混合数	発話の尤度の正規化		
	なし	全話者	平均
1	28.6	74.2	60.0
2	39.8	78.9	90.3
4	41.4	98.8	99.2
8	48.1	92.2	99.4
16	50.9	99.4	99.4
32	56.9	99.2	99.6
64	59.0	99.0	99.4
128	60.4	98.8	99.2

を削減することを考える。いくつかの発話に対する話者モデルの尤度を話者の特徴として k-means 法でクラスタリングを行った。全学習話者のうち、音素バランス文の A セットを発声している話者 42 名 (ATR コーパスの 10 名、JNAS コーパスの 32 名) から、A セットの発話を話者ごとに異なる発話を一発話ずつ順に選択し対象発話とした。従って、42 発話に対するモデルの尤度が話者クラスタリングにおける話者の特徴ベクトルになる。クラスタ数は 2, 4, 8, 16, 32, 64, 128 を試みた。各クラスタのクラスタ中心に最も近い話者を代表として取り出し、削減された話者数で 3.2節の発話分類を行った。平均値を用いた尤度の正規化を行った場合の結果を表 11に示す。EHMM と GMM のどちらの話者モデルにおいても 32 名程度の話者で十分な性能が得られている。

3.4 発話クラスタリングに基づいた会議音声の話者セグメンテーション

RWC の会議音声コーパスを用いて、会議音声の話者セグメンテーションを行った。用いた会議音声データの話者数、発話数、長さ、内容を表 12にまとめておく。会議音声を 200ms 以上の無音区間で発話単位に自動分割し、コーパスに付与されている転記データを参照して、一人の話者だけが発声していると判断された発話だけを発話クラスタリングの対象とした。表 12中のデータごとの「発話数」において、「合計」は自動分割された全発話数、「単一話者」は一人の話者だけが発声していると判断された発話数を示している。

平均を用いて尤度の正規化を行った場合の結果を表 13に示す。EHMM, GMM とも、状態数や混合数、学習話者数によらず、発話の話者ごとの分類がほとんどできていない結果となってしまった。

表 11: 削減された学習話者による話者同定率 [%] (平均値を用いた尤度の正規化)

(a) EHMM

状態数	学習話者数							
	2	4	8	16	32	64	128	all
5	43.5	65.4	82.3	95.0	98.4	99.0	98.8	99.0
10	41.9	84.1	91.1	95.6	99.0	98.6	98.6	98.8
20	40.8	63.8	95.0	95.6	98.4	97.8	97.8	98.8
30	46.9	73.4	93.4	93.8	97.2	98.0	98.8	98.8

(b) GMM

混合数	学習話者数							
	2	4	8	16	32	64	128	all
1	27.2	41.4	53.5	49.9	59.4	60.4	60.8	60.0
2	44.7	68.4	82.7	90.9	86.1	88.3	88.3	90.3
4	41.7	70.8	80.7	96.6	99.4	99.0	99.0	99.2
8	48.9	79.3	93.8	98.6	99.0	99.2	99.4	99.4
16	47.1	84.7	81.5	96.6	98.6	99.2	99.2	99.4
32	50.5	68.0	95.4	96.8	98.4	98.8	98.4	99.6
64	49.9	74.4	91.8	98.0	98.6	99.4	99.4	99.4
128	43.5	56.7	86.5	93.6	98.0	98.8	99.8	99.2

表 12: RWC 会議音声コーパス

会議 ID	話者数			発話数		長さ [分]	内容
	男性	女性	合計	単一話者	合計		
M01	3	1	4	306	387	21.50	ツアー企画
M02	2	2	4	386	434	19.14	ツアー企画
M03	1	4	5	311	430	23.12	ホームページ活性化
M04	2	2	4	213	315	20.58	ホームページ活性化
M05	3	2	5	458	578	35.45	ツアー企画
M06	2	3	5	345	447	20.52	メールマガジン
M07	2	3	5	483	626	33.15	ホームページでの企画

表 13: 会議音声の発話分類率 [%] (平均値を用いた尤度の正規化)

(a) EHMM

状態数	学習話者数							
	2	4	8	16	32	64	128	all
5	34.3	33.0	33.4	32.8	32.6	34.7	35.1	35.4
10	32.9	34.6	36.6	33.2	33.9	33.3	32.7	35.3
20	33.4	32.9	34.1	32.9	33.6	33.4	33.5	33.3
30	32.2	34.8	35.1	34.6	34.3	34.6	34.1	33.7

(b) GMM

混合数	学習話者数							
	2	4	8	16	32	64	128	all
1	34.1	33.6	34.2	34.6	32.7	33.7	33.4	33.7
2	32.2	34.1	33.0	33.5	32.3	34.3	34.2	33.7
4	33.5	32.7	33.0	34.5	33.8	33.0	32.5	33.0
8	33.7	32.6	34.2	34.8	32.9	34.1	34.2	34.3
16	31.3	35.4	33.6	35.4	33.7	34.1	32.9	33.4
32	32.0	35.7	33.1	32.8	35.0	33.4	33.3	33.3
64	32.9	35.7	35.2	33.2	35.0	33.7	34.0	34.0
128	32.7	33.7	36.5	35.6	32.8	33.8	34.5	33.5

3.5 考察

音韻バランス文を対象とした予備的実験における話者ごとの発話分類では、3.2節で述べたように、非常にうまく発話が分類された。この実験では、対象した 10 名の話者が学習話者に含まれているクローズドな実験であるが、3.3節で述べた学習話者の削減を行った場合には、例えばクラスタ数 (学習話者数) が 32 と時には、対象の 10 名はほとんど学習話者に含まれておらず、オープンな話者に対しても機能している。しかし、3.4節で述べたように、RWC コーパスの会議音声に対する実験では、まったく発話の分類ができなかった。秋田らは本手法とほとんど同じアイディアに基づいた手法で会議音声の発話分類がうまくいくことを報告している [7]。秋田らは、CMN を利用することで性能が向上することを指摘しており、本研究で用いた会議音声においても同様の効果は期待できるものの、他の要因による性能の劣化も考えられ、今後さらに結果の分析が必要である。

4　おわりに

今後も大量データの蓄積が進み、大量データを対象としたデータマイニングや質問応答などの発見／検索型の情報処理技術やコーパスに基づいた音声認識、翻訳などの変換型の情報処理技術がますます重要になると思われる。それに伴い、蓄積されたデータを使いやすくするためのタグ付与や構造化などの技術の必要性も高まってくる。本報告では、音声データを自動的に構造化するためにクラスタリングを用いる手法について述べた。クラスタリングで得られる結果は、基本的には、データのセグメントへの分割である。データを自動的に構造化するという観点からは、今後、セグメント間の関係の分類や認識が興味ある研究テーマとなるであろう。

参考文献

[1] 山下, 小磯, 堀内 : “音声対話に対する談話セグメントのタグ方式の検討”, 人工知能学会誌, 14, 2, pp.282-289 (1999).

[2] 井上, 山下 : “要約のための重要文検出における F0 モデルの利用”, 電子情報通信学会技術報告, SP2001-131, pp.47-54 (2002).

[3] 北, 津田, 獅々堀 : “情報検索アルゴリズム”, 共立出版 (2002).

[4] D. Cutting, D.R. Karger, and J.O. Pedersen : “Constant Interaction-Time Scatter/Gather Browsing of Very Large Document Collections”, Proc. of SIGIR '93, pp.126-135 (1993).

[5] 清水, 大野, 樋口 : “文のクラスタリングに基づく統計的言語モデル”, 日本音響学会春季講演論文集, 1-6-14, pp.31-32 (1998-3).

[6] 西田, 有木 : “自動学習による話者セグメンテーション”, 電子情報通信学会技術報告, SP97-57, pp.1-6 (1997).

[7] 秋田, 河原, 奥乃 : “多数話者音声データベースを用いた討論音声の教師なし話者インデキシング”, 情報処理学会研究報告, SLP-42-9 (2002).