音声言語情報処理 42－9
（２００２．７．１３）

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

秋田 祐哉[†‡]　河原 達也[†‡]　奥乃 博[†]

[†] 京都大学 情報学研究科 知能情報学専攻

[‡] 科学技術振興事業団 さきがけ研究 21

〒 606-8501 京都市左京区吉田本町

e-mail: akita@kuis.kyoto-u.ac.jp

あらまし　討論などの長時間音声の教師なし話者インデキシングのために、多数話者との類似度を用いたオフラインのインデキシング手法を提案する。音声データベースから構築した多数話者モデルによる話者識別スコアを成分として発話ごとに話者ベクトルを構成し、これをクラスタリングすることにより話者インデキシングを実現する。 また討論においては司会が特別な役割を持ち、その発話が非常に多いという特徴から、司会のみ固有の話者モデルを構築し、クラスタリングに先立って話者照合を行うことでさらなる精度の向上を図る。実際の討論音声を用いた実験の結果、88.2%のインデキシング精度を得た。

Unsupervised Speaker Indexing for Discussion Speech Using Large Scale Speech Database

Yuya Akita[†‡]　Tatsuya Kawahara[†‡]　Hiroshi G. Okuno[†]

[†]Graduate School of Informatics

Kyoto University, Kyoto 606-8501, Japan

[‡]Japan Science and Technology Corporation PRESTO

e-mail: akita@kuis.kyoto-u.ac.jp

Abstract　We address an unsupervised speaker indexing method using similarity measures to the speakers in large scale speech database. At first, speaker characterization vectors are generated by speaker identification with a large number of speakers of speech database. Then, the dimension of vectors is reduced by KL-transformation and these vectors are clustered into participant speakers of discussion. To enhance the indexing, we also introduce a model of chairperson who speaks more often than others, and perform speaker verification before clustering. Indexing accuracy of 88.2% is achieved using real discussion speech.

1 はじめに

近年、計算機の能力向上とマルチメディア技術の発展により、長時間の音声をディジタル・アーカイブ化して活用することが可能となっている。音声メディアの場合、テキストのようなブラウジングや全文検索のできるメディアと異なり、所望の部分を検索するためには多くの時間と手間がかかる。そこで適切なインデックス (索引) を構成することがアーカイブを構築する上で非常に重要である。

本研究では、討論音声を対象に話者情報を自動インデキシングする手法について検討する。討論音声は複数の話者からなり、話者情報は討論の展開を追う上で重要な情報である。

従来、このような話者インデキシングは話者識別の問題として種々の研究がなされてきたが、そのほとんどは、事前に対象話者の音声を別途用意して話者の音声特徴をモデル化 (教師付き学習) し、未知入力音声のモデルとの類似度を利用して話者を識別する手法である。このような手法は、十分な学習データがあれば高い精度で話者を識別できることが報告されている。しかし、事前にモデル学習用音声を準備する必要があるため、討論音声のような毎回参加者が異なる環境では不便である。さらに、参加者が共通であるとしても、各話者の音声の短期的あるいは長期的な時間変動にどのように対処するかという問題もある。

そこで本研究では、当該の討論音声のみを用い、学習データを必要としない教師なしの手法を提案する。

2 教師なし話者インデキシングの検討

2.1 オフラインのインデキシング

教師なし話者インデキシングを行う手法はこれまでにいくつか研究されているが、オンライン処理で逐次的に話者モデルを構築して話者の検出・分類を行う手法が大半である [1, 2, 3, 4]。基本的な枠組みは、それまでの発話の話者モデルを利用して照合を行い、照合のスコアが閾値を上回った場合はその話者として、下回った場合は新たな話者としてインデキシングを行うものである。新たな話者とされた場合は、その発話を用いて話者モデルの学習も同時に行っている。話者特徴を表現するモデルとしては話者部分空間 [1]、混合ガウス分布モデル (GMM)[2]、ベクトル量子化に基づくコードブック [3, 4] などがある。また、話者の個人性そのものを表現するのではなく、音声認識に用いられる音響モデルに話者適応を施して話者の識別に利用する手法もある [5]。いずれの研究でも、インデキシングはあらかじめポーズ長等により区切られた発話単位で行われている。

オンラインの手法では、インデキシングの時点で得られている音声のみを利用して話者のモデルを構築する。途中の段階ではモデルの学習量が十分に確保できないため、音響的変動により同一話者が複数のモデルに分散したり、あるいは雑音等の重畳した不適当な音声区間でモデルを構築してしまうなどの問題が生じうる。それに加えて、話者数をどのように制御するかが難しい問題である。時間に伴って話者クラスタ数が増え続けるので、特に討論音声のように長時間の音声ではその数が膨大になる。事前に話者数を与えたとしても、途中の段階で新たな話者かどうかの判定は難しい。

これに対して本研究では、すべての音声データからクラスタリングを行うオフラインのインデキシングを考える。入力音声すべてをインデキシング処理の最初から利用可能であり、入力音声の特徴の分布を把握した上で話者の分類を行うことができる。また指定された話者数へのクラスタリングの制御が容易である。本研究ではアーカイブ構築のための非リアルタイムのインデキシングを想定しており、話者数を指定することも含めて、このような前提は妥当である。

2.2 話者ベクトルによるインデキシング

オフラインのインデキシングの例として、Ergodic HMM において状態を話者、状態遷移を話者の遷移と見なして、信号 (音声特徴量) レベルでクラスタリングを行う研究もある [6]。ただし、モデルを全く用いないで、特徴量そのものでクラスタリングを行う手法では、特徴量自身の揺らぎにより話者の分類精度が大きく変動するおそれがある。[6] では評価用の音声として音声データベースを利用しているが、4 名の話者で 68-79%の精度しか得られてい

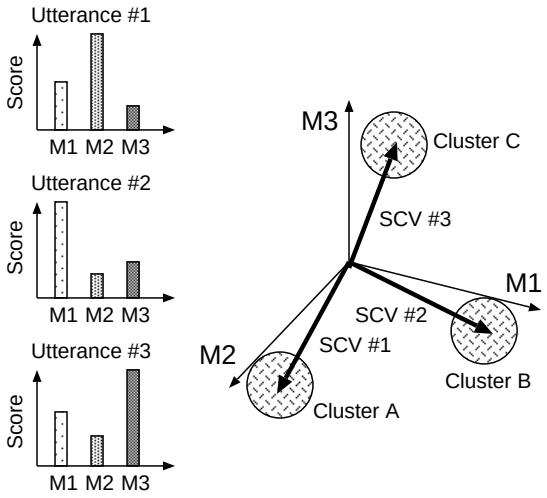

図 1: 話者ベクトルの概念図

ない。さらに、自然な話し言葉の音声には種々の雑音や歪みが含まれるため、特徴量に大きな変動が生じる。

そこで本研究では、音声の特徴を別の話者との類似性で表現することを考える [7]。未知入力発話を、多数の話者との類似度を成分とした話者ベクトル (Speaker Characterization Vector) に写像することで、特徴量の揺らぎの影響を軽減することを図る。類似度の計算対象となる話者数が少ない場合は話者特徴を十分に表現できないが、多数の話者を含む音声データベースを用いることで解決を図る。モデルには GMM を利用し、類似度として多数話者モデルを用いた話者識別のスコアを採用する。この例として [7] では、話者識別の話者候補を削減し計算量を抑えるために話者ベクトルによる類似度比較が用いられているが、本研究ではこれを話者クラスタリングそのものに適用する。

図 1 に話者ベクトルの概念図を示す。入力発話ごとに、音声データベースから構築した話者モデルのそれぞれに対して話者識別スコアを計算する。これを成分とするベクトルを構成し、これをクラスタリングすることで話者インデキシングを実現する。

話者ベクトルへの写像が十分な話者分離性を持っているかを、実際の討論音声 (表 1) を用いて予備調査した。話者モデルのマッチング傾向を図 2 に示す。図中の横軸は話者 ID、縦軸は話者識別スコアの平均であり、討論の参加者ごとにスコアの高いもののみをプロットしている。上に位置するほどそのモデルが討論話者の音声によくマッチングしている。この図から、討論の参加者によってよくマッチングするモデルが異なっていることがわかる。

表 1: 評価用の討論音声

番組名	NHK『日曜討論』2001 年 9 月 30 日
参加者	8 名 (男性 7 名、女性 1 名)
時間	48 分
発話総数	127

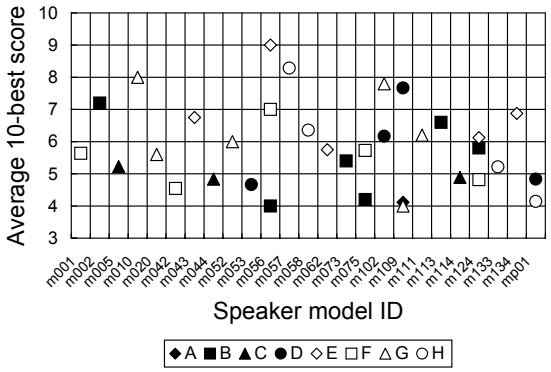

図 2: 話者モデルのマッチング傾向

3 話者インデキシング手法

提案する話者インデキシングの処理の流れを図 3 に示す。まず、未知の入力発話について、録音系等の影響を抑えるためにケプストラム平均正規化 (CMN) を適用する。次に、音声データベースによる多数話者 GMM に対する識別スコアが計算される。各識別スコアを成分とした話者ベクトルが構成されるが、話者分離性を高めるために KL 変換による次元の圧縮を行う。最後に LBG アルゴリズムを用いてベクトルを分類する。クラスタ数、すなわち話者数は事前に与えるものとする。

評価用の討論音声は表 1 に示すものである。この討論は、政治・経済などの分野の時事問題について、数名の政治家や学者などにより行われている。また討論には司会が存在する。発言機会は司会によって与えられるため複数話者による発言の重複は少ないが、相づちや笑い声などの重複は存在する。本研究ではこのような重複区間は除去していない。また、話者交代ごとに区切った各区間を発話と定義し、入力音声は (相づちを除いて) あらかじめ発話

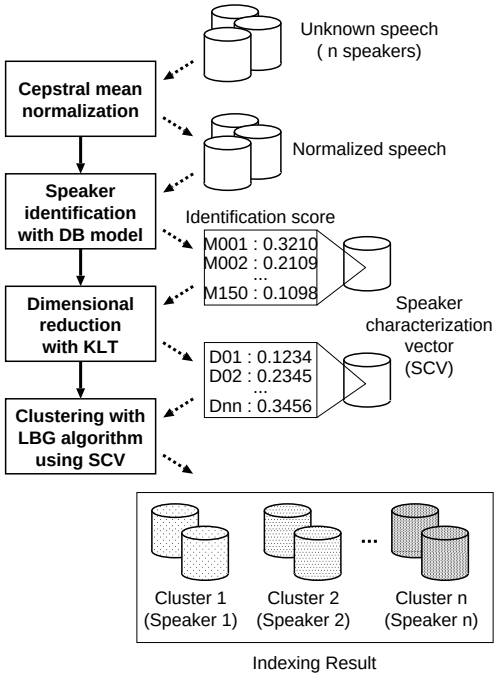

図 3: インデキシング処理の概要

表 2: 多数話者モデルの仕様

データベース	日本音響学会 JNAS データベース
話者数	304 名 (男性 153 名、女性 151 名)
発話文	音素バランス文 45 文
特徴量	MFCC(12), ΔMFCC(12), ΔEnergy(1) 計 25 次元
混合数	32 混合

ごとに区切られているものとする。

インデキシングの精度を高めるためには、話者モデルに個人性以外の点でマッチングしないようにすることと、入力音声の個人性を強調することが重要である。この観点からインデキシング手法に加えた検討について、以下に詳しく述べる。

3.1 多数話者モデルの学習

多数話者モデルは表 2 にある仕様で学習した GMM である。これが話者個人性を十分に表現していることが重要である。JNAS データベースは研究用の不特定話者モデルの構築に最も広く用いられており、十分な話者を含むと考えられる。学習には音素バランス文のみを使用した。また音声から個人性に寄与しないと考えられる文頭・文末・文中の無音区間を除去した。そのために、「日本語ディクテーション基本ソフトウェア」[8] に含まれる monophone HMM(43 音素、4 混合、性別非依存) を利用し、非音素モデルには無音系の 3 種の HMM(silB, silE, sp) の、音素モデルにはその他の HMM のガウス分布を用いて GMM を作成した。無音検出の結果、フレーム単位で全体の 29.2% が除去された。この音声をもとに構築した多数話者モデルで話者ベクトルを算出しインデキシングを行ったところ、無音を除去しないすべての音声区間で構築したモデルの場合と比較して、精度は 2.3%(46.5% → 48.8%) 改善した。同一モデルに複数の話者がマッチングする傾向が大きく減少した。

3.2 入力音声のケプストラム平均正規化

多数話者モデルの学習用音声と入力される討論音声とでは録音系が異なるため、特徴量 (メル周波数ケプストラム) の正規化によりその影響を除去する。特徴量からケプストラムの平均を減じて正規化を行うが、平均を発話単位で算出し正規化すると、発話に含まれる話者個人性まで正規化されて失われる。したがってケプストラム平均はすべての発話を用いて求める。このように正規化した場合、発話単位で行った場合に比べ 6.3%(48.8%→ 55.1%) 精度が改善した。

3.3 KL 変換による話者ベクトルの次元圧縮

ある発話に対して話者ベクトルを算出すると、大多数の話者モデルについて識別スコアが 0 に近い値をとる。このような情報のない成分は、クラスタリングにおいてはむしろ類似していると見なされ、話者分離性が低下する。そこで、これらのインデキシングに寄与しない成分を除去するために、KL 変換を用いて次元の圧縮を図った。KL 変換による次元の圧縮とクラスタリング精度の関係を図 4 に示す。図中破線の累積寄与率が 90%前後で識別精度が最大となった。ベクトルの元の次元数 304 に対し 15 分の 1 程度で十分であることがわかる。また、これは有意な値を持つ成分の数におおむね一致している (図 2 の横軸)。KL 変換により、55.1%から 67.7% へと精度を 12.6%高めることができた。

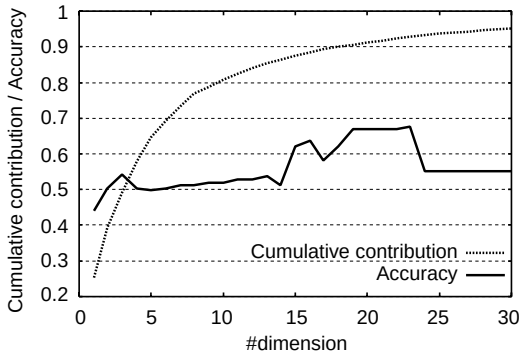

図 4: KL 変換によるクラスタリング精度の改善

表 3: クラスタリング結果

ID	C1	C2	C3	C4	C5	C6	C7	C8	total
A	32	7	3	1	1	2	11	3	60
B		5							5
C			16	1				1	18
D				6					6
E					8				8
F	1	2	1		7				11
G							5		5
H								14	14
total	33	14	20	8	16	2	16	18	127

4 司会モデルの導入

前章のインデキシング結果 (表 3) は十分な精度とはいえない。インデキシングの誤りを分析したところ、参加者中もっとも発話数の多い司会 (表 3 の話者 A) の発話が複数のクラスタに分散してクラスタが正しく話者に対応できないことが大きな原因となっている。参考のため、司会の発話 (127 発話中 60 発話) をすべて除外してインデキシングを行ったところ、精度は 86.6%に達した。

そこで、あらかじめ司会の発話を除外することを考える。司会のみ固有の話者モデルを構築し、クラスタリングに先立って話者照合を行い、検出された司会の発話はクラスタリングを行わないでインデキシングする。話者モデルの構築には学習データのラベリングが必要であるが、司会は特別な立場にあり、討論の最初に話すことが多い。このため実用的にはあまり問題にならないと考えられる。本稿では、司会の最初の発話の冒頭 1 分を抽出して司会モデルを学習する。

司会モデルによる照合を組み込んだ手法を図 5 に示す。司会の照合には、司会のモデルとその他の話者を表現する討論者モデルを用いる。入力音声に対する両モデルの尤度比が閾値以上である場合に司会と判別する。討論者モデルの構築には司会以外の話者の音声が必要であるが、司会モデルの初期学習用として与える最初の発話以外は入力音声のどの区間が司会のものであるか不明である。そのため、討論者モデルを未知入力音声全体で構築して一度話者照合を行い、入力発話を一度分類した上であらためて両方のモデルを学習する。

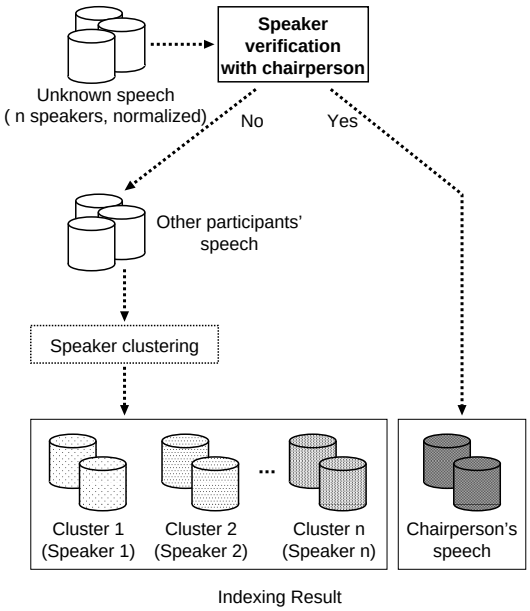

図 5: インデキシング処理の概要 (司会モデル導入時)

司会モデルによる照合と、討論者とされた発話に対するクラスタリングの結果を総合したインデキシング精度を図 6 に示す。図中、司会の照合結果については誤棄却率と誤受理率を、司会以外の発話に対してはクラスタリング精度を破線で表している。最終的なインデキシング精度は実線で示す。照合の閾値が 0.3 の場合にもっとも高い精度 88.2%を得ている。これは、おおむね照合で最もよい閾値に一致する。クラスタリングから司会の発話数をほぼ除外できたため、クラスタリング精度は 81.8%と向上している。

提案手法に対して、別の討論音声により評価を行った。利用した討論音声の仕様は表 4 の通りである。これら 2 つの討論音声のインデキシング精度を図 7、図 8 に示す。

6 月 24 日の討論音声では、クラスタリング精度

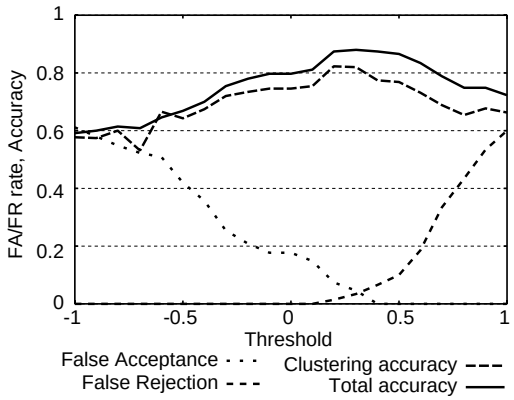

図 6: 司会モデル導入時のインデキシング精度

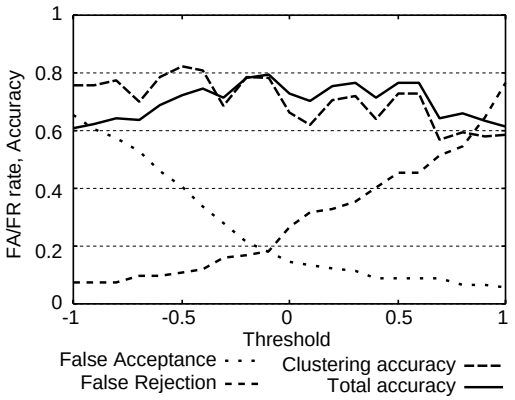

図 7: インデキシング精度 (2001 年 6 月 24 日)

表 4: 討論音声

放送日	2001 年 6 月 24 日	2001 年 10 月 7 日
参加者	5 名 (すべて男性)	8 名 (男性 7 女性 1)
時間	49 分	47 分
発話総数	171	113

そのものは他の討論音声と同程度であるが、司会の照合精度が低いために全体のインデキシング精度も 79.5%にとどまっている。一方 10 月 7 日の討論音声では、クラスタリング精度・司会照合精度ともに高く、93.8%のインデキシング精度が得られた。

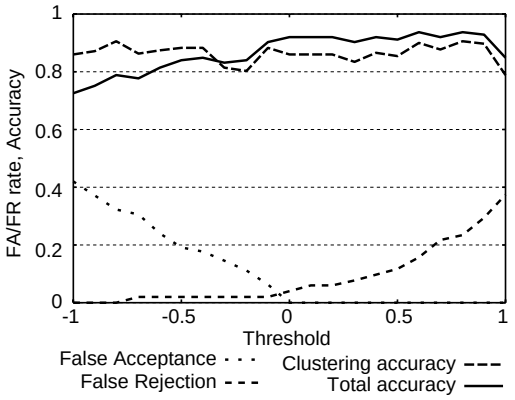

図 8: インデキシング精度 (2001 年 10 月 7 日)

5 結論

討論音声を対象とした教師なし話者インデキシングについて検討した。多数の話者からなる音声データベースを用いて話者モデルを構築し、それぞれの話者モデルとの類似度をベクトル化してクラスタリングすることでインデキシングを行う。また、司会の発話が他より顕著に多いという討論音声特有の特徴をもとに、司会のみ話者照合に基づく検出を行うことでさらなる精度の向上を図り、最終的なインデキシング精度は 88.2%となった。

今後はインデキシングの精度を高めるとともに、現在事後的に定めている司会照合の閾値や KL 変換の打ち切り次元数などのパラメータを適切に決定する手法についても検討する。また、インデキシングの結果を活用した討論音声の認識も行う予定である。

参考文献

[1] 緒方淳, 西田昌史, 有木康雄. 自動抽出されたアナウンサー発話に対するニュースディクテーションと記事分類. 情報処理学会研究報告, 98-SLP-21-5, 1998.

[2] 張志鵬, 古井貞熙. 話者交代検出を含むオンライン話者適応の検討. 日本音響学会秋季研究発表会講演論文集, 1-1-23, 1999.

[3] 森一将, 山本一公, 中川聖一. 発話間の VQ 歪みを用いたオンライン話者交替識別と話者クラスタリング. 電子情報通信学会技術研究報告, SP2000-18, 2000.

[4] 張巍, 中川聖一. 自動話者クラスタリングに基づく逐次話者適応化手法を用いた連続音声認識. 日本音響学会春季研究発表会講演論文集, 3-2-2, 2002.

[5] 田熊竜太, 岩野公司, 古井貞熙. 逐次話者適応を用いた並列処理型会議音声認識システムの検討. 日本音響学会春季研究発表会講演論文集, 2-5-16, 2002.

[6] 村上仁一, 杉山雅英, 渡辺秀行. Ergodic HMM を用いた未知・複数信号源クラスタリング問題の検討. 電子情報通信学会論文誌, Vol. J78-DII, No. 2, pp. 197–204, 1995.

[7] D. Sturim, D. Reynolds, E. Singer, and J. Campbell. Speaker indexing in large audio databases using anchor models. In *Proc. ICASSP*, Vol. 1, pp. 429–432, 2001.

[8] 鹿野清宏, 伊藤克亘, 河原達也, 武田一哉, 山本幹雄. 音声認識システム. オーム社, 2001.