研究開発

オンライン話者クラスタリング技術と議事録作成支援への応用

越仲 孝文・長友 健太郎

要　旨

話者クラスタリングとは、複数の話者が交替で発話する状況において、個々の話者の性質や話者数などの事前知識を使わずに、発話を話者ごとに分類する技術であり、会議などでの音声認識精度向上、発言者名の入力支援などに有用です。本稿では、リアルタイム動作が可能で高精度な話者クラスタリング技術と、議事録作成支援への応用例について述べます。

キーワード

●会議音声認識 ●議事録作成 ●発言者推定 ●オンライン処理 ●隠れマルコフモデル

1. はじめに

クラスタリングとは、データ解析の用語で、多数のデータを似たもの同士まとめてグループ化（分類）する技術の総称です。クラスタリングで得られた一つひとつのグループのことをクラスタと呼びます。特に話者クラスタリングといった場合は、誰のものか分からない一連の発話を、声の類似性に基づいて話者ごとに分類する技術を指します。この場合、クラスタは話者と一対一対応することが期待されます（**図1**）。

音声認識において話者クラスタリングは、教師なし話者適応（入力音声とその認識結果を用いて、個々の話者にシステムを適応化して音声認識精度を向上させる技術）の前段に置かれ、話者ごとに発話を分類して話者適応化処理に送る役割を果たします（第4章参照）。発話が正しく分類できていないと、後段の話者適応化処理も適切に行われないので、認識精度の向上が得られなくなります。

また、大量の音声あるいは映像コンテンツから、ある特定の話者の発言を見つけたいというニーズがしばしばあります。話者クラスタリングにより話者ごとの分類がなされていれば、同一話者の発話をまとめて検索したり、任意の発話と類似した発話を検索したりできます。

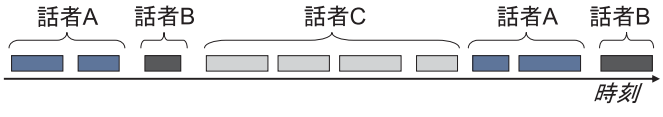

図1　話者クラスタリングの概念図。話者A, B, C, …の会話における一連の発話を、話者ごとにグループ化

このように話者クラスタリングは、音声情報処理の重要な要素技術の1つといえます。複数の話者が交替で話すようなシチュエーション（放送、会議など）への音声認識の適用が現実的に考えられるようになった1990年代後半から、この技術の研究が盛んに行われています。

従来提案されている話者クラスタリング方式は、バッチ処理として実行されるものがほとんどでした。すなわち、クラスタリング対象となるすべての発話が取得されてから処理を開始し、すべての発話を俯瞰して発話間の類似度を測り、分類を行っていました。

しかし、このようなバッチ処理タイプの方式は、即時性が要求される用途には使えません。例えば、議事録作成支援[1]のような用途では、会議中の発言や発言者をリアルタイムで書き起こして、会議終了後に短時間で議事録を発行したいというニーズがあり、会議中に取得される発話を高精度で逐次分類していく、オンライン処理が可能な方式が必要とされています。オンライン処理タイプの方式も過去に提案されていますが[2]、幾分アドホックなアプローチで、分類精度も高くありませんでした。本稿では、このような背景に基づいて開発した、高精度オンライン話者クラスタリング方式を紹介します。

2. オンライン話者クラスタリング

2.1 従来方式とその問題点

オンライン動作可能な話者クラスタリングの従来方式は、leader-followerクラスタリング（LFC）と呼ばれる一般的なクラスタリングアルゴリズムの考え方に基づくもので、次のように動作します。

1) 最初の発話 X_1 が入力されると、その発話のみからなるクラスタを作成する。

2) n 番目の発話 X_n が入力されると、 X_n と既存クラスタとの類似度を計算し、類似度最大のクラスタに X_n を分類する。ただし、類似度が所定のしきい値よりも小さい場合は、 X_n のみからなる新規クラスタを作成する。

図2 はその概念図です。LFCの考え方は大変シンプルで、実装も容易ですが、次のような問題点を含みます。

・ 分類が決定論的で、いずれのクラスタに分類すべきか決め難い曖昧なケースでも、必ずいずれか1つのクラスタに分類しなければならない。

・ 過去の履歴を考慮しておらず、その時点までのクラスタ構成を正しいとして発話を分類する。つまり、過去に分類誤りはないという非現実的な仮定が入っている。

人間の発話という現実世界のデータを扱う場合、分類が非常に難しいケース、誤った分類をしてしまうケースは必ず起こるものであり、これらのケースを考慮した方式が必要です。

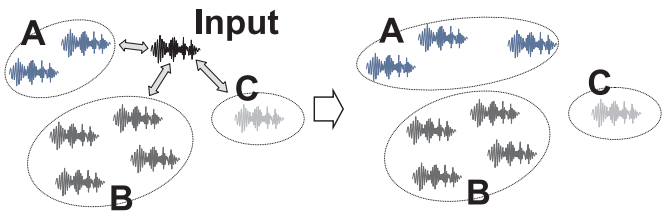

図2　オンライン話者クラスタリングの従来方式

2.2 提案方式

前述の問題点を解決するために、提案方式では隠れマルコフモデル（HMM）と呼ばれる確率モデルを用います。HMMは、音声認識システムにおいて音韻のモデル化（音響モデリング）に広く用いられる道具で、音声のような可変長データを扱うのに適しています。

ここでは、**図3** に示すような N 個の状態（図の中央に縦に並んだノード）を有するHMMを用います。HMMは初期状態 I から動作を開始し、確率 π_i で状態 i（$i=1,\cdots,N$）に遷移します。状態 i は確率1- a_i で自己遷移をくり返すので、しばらくの間、状態 i に滞留することになります。状態 i に滞留する間、状態 i に固有の確率分布 $f(x|\lambda_i)$ に従い、短時間の音声に相当する特徴ベクトルが出力されます。確率 a_i で終状態 F に到達したときには、特徴ベクトルの系列 $x_1, x_2,\cdots, x_T$ が出力されています。その後、また初期状態 I に戻り、1から N のいずれかの状態に遷移し…というような動作をくり返します。

このような動作をするHMMは、複数の話者が交替で発話するような状況における発話の生成過程そのものです。すなわち、初期状態 I は発話の始まりに対応し、状態 i への遷移は i 番目の話者が発話することに対応します。その発話は特徴ベクトル系列 $X=(x_1, x_2,\cdots, x_T)$ として観測されます。終状態 F で話者 i の発話が終わり、また次の話者が発話を開始します。

話者クラスタリングは、上述した発話の生成過程の逆問題、つまり与えられた発話を用いたHMMの学習問題として定式化できます。提案方式では、順次入力される発話 $X_1, X_2,\cdots$ を用いて、HMMのパラメータ π_i, a_i, λ_i（$i=1,\cdots,N$）を逐次推定します。推定されたパラメータを用いると、ある発話 $X=(x_1, x_2,\cdots, x_T)$ がどの状態（話者）から発生したか

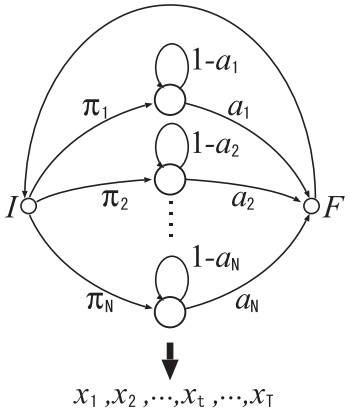

図3　提案方式の基礎となる隠れマルコフモデル

を次式により確率的に計算することができます。

$$P(i|X) \propto \pi_i a_i (1-a_i)^{T-1} \prod_{t=1}^{T} f(x_t|\lambda_i).$$

もちろん、外部に出力する分類結果は、「$P(i|X)$が最大となるi」のように決定論的な形式になりますが、内部ではこのように確率論的に結果を保持して処理を進めることにより、分類誤りの可能性を考慮した話者クラスタリングが可能となります。

パラメータの推定には、HMMの学習アルゴリズムとしてよく知られたBaum-Welch法などを適用することが可能です。一般的なBaum-Welch法はバッチ処理であり、オンライン処理はできませんが、提案方式では、一部のデータのみ参照して学習を行うインクリメンタル学習の考え方[4]に基づいて、直近のb個の発話のみを用いてパラメータを更新するオンライン処理（ブロックインクリメンタル学習）としています。bをここではバッファサイズと呼ぶことにします。バッファを持つことによって、過去のb個の発話については、分類結果を補正して動作することが可能となります。つまり、過去の分類結果を顧みて、自身の犯した間違いを改めることで、その後の分類をより正しく行うという機構が備わっています。

更に提案方式では、Baum-Welch法に代えて、学習データの多寡に応じてパラメータ推定結果の信頼性を見積もるベイズ学習の枠組み（変分ベイズ法）を取り入れ、発話の時間長が短い場合など、学習データ量が乏しい状況でも頑健に動作するように配慮しています。話者数Nの推定についても、ベイズ学習の枠組みにのっとり、事後確率$P(N|X)$を計算することによりNを決定するモデル選択を逐次行っています。

以上は、提案方式についてできるだけ平易な説明を試みたものですが、定式化などのより厳密な説明については、既発表の文献[5]などを参照してください。

3. 評価実験

提案方式の有効性を示すために、会議音声を用いたいくつかの話者クラスタリング実験を行っています。ここではその中でも基本的な結果を紹介します。

実験に使用したのは、約2時間（20分×6件）の日本語会議音声データ（22 kHz, 16 bit PCM）で、1件当たりには、6〜10人の話者から発せられた500発話が含まれます。1発話当たりの継続時間長は、平均1.90秒（最小0.24秒、最大10.86秒）で、この種の実験では比較的難易度の高いタスクといえます。

この評価データを用いて、話者クラスタリングの単体性能、すなわち発話をどのくらい精度よく話者ごとに分類できたかを調べます。クラスタリングの良し悪しをどう測るかはさほど自明ではなく、いくつかの指標がありますが、ここではRand Indexと呼ばれる指標を用いました。Rand Indexは、話者クラスタリングに限らずデータ分割問題の汎用的な誤り率の指標で、ここでは「無作為に選んだ2つの発話が、話者が同じなのに異なるクラスタに分類されるか、話者が異なるのに同じクラスタに分類される確率」と定義されます。

クラスタ数を段階的に変えて、クラスタ数と誤り率（Rand Index）の関係を求めたのが**図4**です。従来方式（Conventional）は前述のLFC。提案方式（Proposed）については、3種類のバッファサイズb=1,2,4を試しています。すべてのクラスタ数において、提案方式が従来方式よりも低い誤り率を達成していることが分かります。また、バッファサイズbの増加により、誤り率を更に低減できていることが分かります。過去の発話をなるべく多く保持して、これらの分類結果を補正することが、その後の分類によい影響を与えているといえます。このように、過去に犯した間違いに気づき、正しながら動作するというのが、提案方式の大きな特徴です。

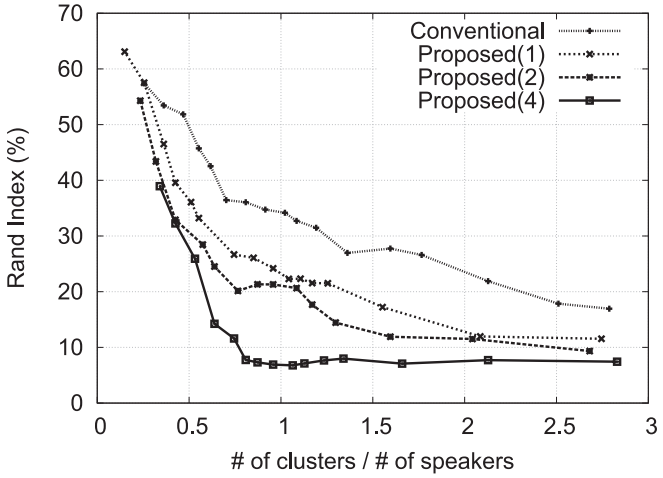

図4　従来法（Conventional）と提案法（Proposed, b =1,2, 4）の精度比較

4. 議事録作成支援への応用

本稿で示したオンライン話者クラスタリング技術に関して、議事録作成支援[1]での応用例を紹介します。

音声認識を用いた議事録作成支援が有効に機能するためには、不特定話者に対して高い認識精度を維持することが重要です。音声認識の分野では、入力音声とその認識結果を用いて、認識に使用する音韻のモデル（音響モデル、以下AMと略す）を入力音声に適応化させる「教師なし話者適応」という手法が提案されています。教師なし話者適応[6]を話者クラスタリングと組み合わせることにより、話者ごとにAMを更新し、認識精度を逐次的に向上させることが可能となります。

図5に模式的に示すように、話者A、B、Cが交替で発言するような状況では、初出の話者、例えば話者Aの最初の発話に対しては、標準的な不特定話者のモデル（AM-0）で音声認識を行います。それと同時に教師なし話者適応が動作し、話者Aに適応化したモデル（AM-A）が作成され、2回目の発話の認識ではAM-Aが使用されます。このような動作を反復すると、AM-A’、AM-A”、…というように、音響モデルと話者Aの適合度が増し、逐次的に認識精度が向上します。

話者クラスタリング技術はまた、議事録における発言者名の入力支援機能を実現します。すなわち、ユーザ（議事録作成者）がある発話に対して発言者名を入力すると、話者が同一と推定されるほかの発話にも、同じ発言者名を自動的に一括入力することができます（**図6**）。この機能により、個々の発話に対してくり返し発言者名を入力する手間が省け、議事録作成の効率化を図ることができます。このような機能は、会議出席者の声を事前に登録しておくといった手間を必要としないので、運用上の利便性が高いといえます。

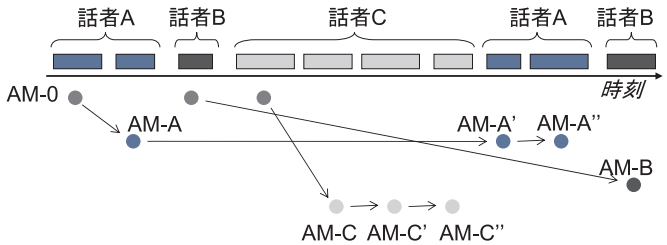

図5　話者クラスタリングと教師なし話者適応の組合せによる音響モデルの逐次更新

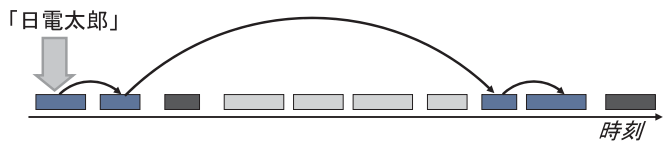

図6　発言者を推定して一括自動入力

5. おわりに

議事録作成支援のようにリアルタイム性が要求される用途を想定して開発した、オンライン話者クラスタリング技術を紹介し、従来法と比較して高い精度で、発話を話者ごとに分類できることを実験的に示しました。また、会議の議事録作成支援などの応用において、本技術が重要な役割を果たすことを述べました。議事録作成に限らず、音声認識技術は今後、多人数のコミュニケーションを助ける方向で応用の範囲をますます広げていくでしょう。我々の技術がその一助となれば幸いです。

参考文献

1) 千代章ほか、「議事録作成支援ソフトウェア VoiceGraphy」、NEC技報、 Vol.63、No.1、(2010年2月）pp.59-61.
2) D. Liu et al., “Online speaker clustering,” Proc. of IEEE Int’l Conf. on Acoustics, Speech, and Signal Processing (ICASSP), 2004.
3) R. Duda et al., “Pattern classification (Second Edition),” John Wiley & Sons Inc., 2001.
4) R. M. Neal et al., “A view of the EM algorithm that justifies incremental, sparse, and other variants,” Learning in Graphical Models, The MIT Press, 1998.
5) T. Koshinaka et al., “Online speaker clustering using incremental learning of an ergodic hidden Markov model,“ Proc. of IEEE Int’l Conf. on Acoustics, Speech, and Signal Processing (ICASSP), 2009.
6) K. Shinoda et al., “A structural Bayes approach to speaker adaptation,” IEEE Trans. On Speech Audio Processing, Vol.9, No.3, 2001.

執筆者プロフィール

越仲 孝文
共通基盤ソフトウェア研究所
主任研究員
電子情報通信学会
日本音響学会各会員

長友 健太郎
共通基盤ソフトウェア研究所
主任