ユーザ訂正を活用したポッドキャスト音響ダイアライゼーションシステム*

◎佐々木 洋子，緒方 淳，後藤 真孝 (産総研)

1 はじめに

ポッドキャストやウェブラジオ，投稿動画など，Web上には多数の音コンテンツがあり日々更新され続けている．人がこうした大量の情報を活用するためには，テキスト情報のキーワード検索やランキングのように必要なコンテンツを整理，抽出する技術が不可欠である．実環境の制約のない音響信号に対し，何の音であるか，どのような内容であるかを人がわかる一般的な名称でコンピュータに理解させることが重要である．

これまでにも，人の声を対象とした音声認識や話者認識，音楽理解を目指した各種音楽情報処理など，特定の音を対象とした多くの認識技術が発展してきている．人の声，楽曲といった特定種類の音に限らず様々な音が混在する実環境の多様な音へ対応するためには，これらの前段処理として，まず「何の音か」を理解する技術 [1, 2] が不可欠である．

本稿では，入力された一連の音響信号に対し「どの部分が何の音か」を求める問題である，音響ダイアライゼーション (Audio Diarization)[3] について述べる．近年，音声認識分野では Rich transcription や Speaker diarization の研究が活発であり，対話コンテンツを対象として「誰がいつ話したか」を推定する技術が発展してきている [4]．会話シーンの分析システム [5] やポッドキャスト中の注目箇所として笑いや相づちを検出するシステム [6] も提案されている．これらを音声以外の一般的な音へ拡張した音響ダイアライゼーションについては，様々な音のモデル化方法や未知の音の扱いなど，まだ課題が多いのが現状である．

本研究では，Web上で日々更新される音メディアであるポッドキャストを対象として，音響ダイアライゼーション結果を可視化し，聴きながら誤りの訂正や情報付加が可能な視聴インタフェース「PodDiarizer」を提案する．提案システムは認識モデルの学習データとして固定的なデータセットのみを利用するのではなく，日々変化するWeb上のメディアも併用する．一般ユーザに自動認識の結果を訂正してもらうことで，継続的に更新可能なモデル構築の枠組みを作る，という実環境の音に対応するため一方法を示すものである．

2 PodDiarizer

本研究では，ポッドキャストを対象とした音響ダイアライゼーションシステムを扱う．ポッドキャストは，音声や音楽，物音など多くの種類の音を含み，Web上で多数のデータが日々更新されている．背景音楽や複数人の同時発話など条件も様々であるため，音響ダイアライゼーションのためのモデルを構築するデータとして利用可能である．ここでは，ポッドキャストを対象として音響ダイアライゼーションの結果を可視化し，ユーザがコンテンツを聴きながら間違いの訂正や情報付加が可能な視聴インタフェース PodDiarizer を提案する．

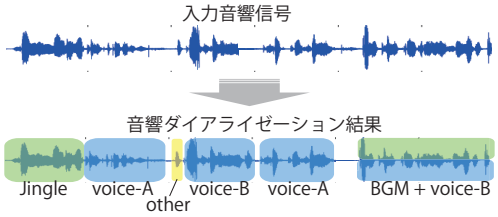

図 1 音響ダイアライゼーション

音響ダイアライゼーションとは，図 1 のように一連の音響信号に対し，類似した音のセグメントごとに切り分け，各セグメントが何の音であるかを判定する機能である．PodDiarizer では，ユーザとのインタラクションを通じて本機能の実現を目指す．まず入力コンテンツに対しコンピュータによる自動認識を行い，ダイアライゼーション結果を可視化インタフェース上に表示する．ユーザは可視化インタフェースを通してポッドキャストを視聴することで，自動認識に含まれる誤り箇所がわかる．提案インタフェースは，誤りの訂正や認識システムにない音ラベルの登録機能を備えており，ユーザに編集してもらうことでモデル構築のためのデータを蓄積する．蓄積したデータを基に認識システムのモデルを更新し，編集していない部分の認識結果も改善することでユーザの利便性も向上する．

提案システムは，ユーザとインタラクション可能な視聴インタフェースを通じて継続的にデータ収集することで，様々な音を扱う音響ダイアライゼーションにおいて，大規模なデータ収集，変化に合わせたモデル更新が可能となる．さらにユーザが新たに音ラベルを登録することでシステムのモデルにない未知の音への対応を実現する．

2.1 PodDiarizer の基本機能

図 2 に PodDiarizer の概要を図示する．提案する PodDiarizer には，1) 音響ダイアライゼーション結果の可視化機能，2) ユーザによる誤り訂正/情報付加機能，の大きく 2 つの機能がある．提案システムは，ユーザが聴きながら気軽に編集できるよう，マウス操作を中心とする簡便なインタフェース上に，より信頼度の高い順に修正候補を提示する．そのために，音響ダイアライゼーション処理においては，認識結果に対する信頼性の評価が可能な手法を採用し，一つの認識結果だけでなく，複数の認識結果を修正候補として出力する機能を実現する．

*A Podcast Audio Diarization System Using User Correction. by Yoko Sasaki, Jun Ogata, and Masataka Goto (AIST)

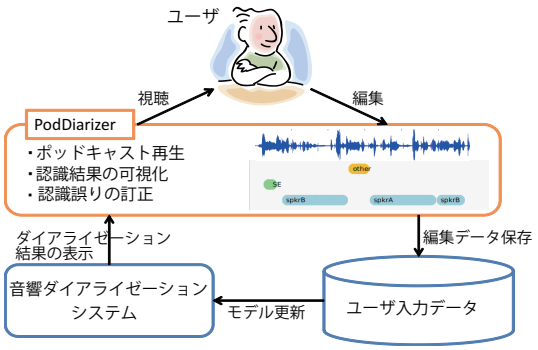

図 2　ユーザ協力型音響ダイアライゼーションシステム PodDiarizer

2.2　インタフェースの構成

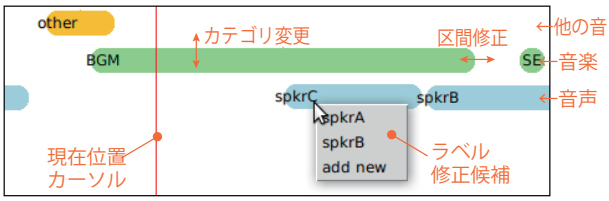

図 3　編集画面の例

ダイアライゼーションの可視化・編集インタフェースの例を図 3 に示す．表示画面は，音声，音楽，その他の音の 3 つのカテゴリに対応して 3 段に分かれており，それぞれのカテゴリに対し，区間を示すバーと，BGM や話者 A といった音のラベル名が表示されている．また横軸の時間軸に対し，現在再生中の位置を赤線で示している．

ユーザによる入力機能は以下の 6 つである．

1. バー両端の左右ドラッグで区間修正
2. 左ダブルクリックでバーの分割
3. 画面下へドラッグでバーの削除
4. 右クリックでラベル修正候補表示 (信頼性の高い順)
5. バーの上下ドラッグでカテゴリ変更
6. 新規ラベル名登録

6. の機能により話者クラスタへの個人名付与，およびモデルにない音の登録が可能となっている．

3　音響ダイアライゼーションの実現方法

本節ではシステムによる自動認識部分の実現方法について述べる．様々な音へ対応するため，大分類から詳細認識へ続く階層的なシステム構成を採用した．

3.1　システム構成

ポッドキャストの音響信号から，背景音楽 (BGM), Jingle(番組の節目に挿入される短い音楽などの総称), 効果音 (Sound Effect, SE), 話者ごとに分類した音声，その他の音，それぞれの混合区間，無音区間を推定する．システムの概略を図 4 に示す．大きく分けて，セグメンテーション，大分類，詳細認識の 3 つの部分からなる．以下，各モジュールの信号処理手法について説明する．

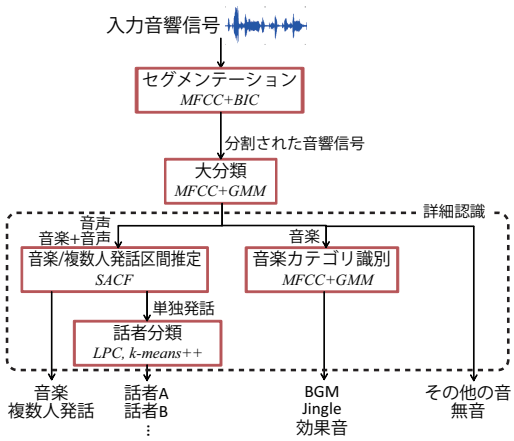

図 4　音響ダイアライゼーション処理の流れ

3.1.1　BIC によるセグメンテーション

まず一連の音響信号を，類似した音が含まれるセグメントごとに切り分ける．本システムではセグメンテーション手法としてベイズ情報量基準 (Bayesian Information Criterion, BIC)[7] を用いる．n 個のデータ $X = x_1, x_2, ..., x_n$ に関する m 個のモデル候補 $M = M_1, M_2, ..., M_m$ が与えられたとすると，BIC 値は次式で定義される．

$$BIC(M_i) = \log P(X|M_i) - \frac{1}{2}\lambda d_i \log n \qquad (1)$$

ただし d_i はモデル M_i の自由パラメータ数，P はデータ X に対するモデル M_i の尤度を表す．

BIC に基づく音響信号の分割では，データ数 n のある区間に対し，境界点 $j(1 < j < n)$ で 2 つの区間に分けた場合 (M_{12}) および，区間全体のモデル (M_0) を想定し，それぞれの BIC 値を比較する．j を連続的に変化させ，$\Delta BIC(j) = BIC(M_0) - BIC(M_{12})$ が極大値をとる点を分割境界とする．特徴量には 12 次元 MFCC，正規化した対数パワー，Δ12 次元 MFCC，Δ 対数パワーの 26 次元の特徴量を用いる．

3.1.2　GMM による大分類および音楽カテゴリ識別

切り分けた各セグメントを，音楽，音声，音楽+音声，その他，無音の 5 種類に大分類する．ここでの音楽とは背景音楽の他に Jingle や効果音といった放送用の人工音を含めるものとする．事前学習として，各カテゴリでフレームごとに求めた音響特徴量を GMM(Gaussian Mixture Model) でモデル化し，認識時にそれらのモデルに対する尤度を比較することで音の種類を判定する．特徴量は前節で述べた 26 次元ベクトルとする．

大分類の結果が音楽区間の場合には，同様に GMM の尤度比較で，BGM，Jingle, SE の識別を行う．ユーザ入力インタフェースでは，尤度の高い順に修正候補として表示する．

3.1.3　SACF による音楽・複数人発話区間推定

音楽区間の推定および音声区間における複数人発話の検出には，フレーム単位の特徴に比べ有用性の高い時系列データの特徴量として，各時刻で推定した基

本周波数 (F0) を時間方向に追跡した F0 軌跡を用いる．F0 推定には高雑音環境で有効であると言われている SACF(Summary Auto Correlation Function)[8] を用いる．SACF は，音響信号を内耳フィルターバンク (Cochlear Filterbank) に通し，各チャネル出力の自己相関関数 (Auto Correlation Function, ACF) を求め，全チャネルの ACF の和として求められる．

こうして求めた F0 軌跡に基づいて，音楽区間および複数人発話区間を推定する．音楽/音声の分類では，音楽信号は音声に比べ周波数変化が小さいという知見から，F0 軌跡が水平な直線のときに音楽，それ以外を音声と判定し [9, 10]，音声と判定された F0 軌跡が複数重なる区間を複数人発話区間とする．F0 軌跡の音楽判定法は以下の通りである．まずある固定区間長のデータに含まれる F0 を時間方向に加算し，周波数方向の F0 ヒストグラムを作成する．周波数変動の少ない音楽の F0 軌跡はヒストグラムのピークとして現れるため，ヒストグラムのピーク周波数を求め，その周波数帯に完全に含まれる F0 軌跡を音楽と判定する．

3.1.4 話者クラスタリング

音声区間に対する話者の分類では，計算コストが低い非階層的クラスタリングである k-means 法を基本とし，クラスタ中心の初期化法を改良した k-means++[11] を用いる．k-means++はランダムに初期値を与える k-means に比べ，初期化の計算量が多いものの収束が速く，また外れ値による悪影響を低減可能である．特徴量にはセグメントごとの正規化平均スペクトル包絡を用いる．スペクトル包絡は線形予測分析 (LPC) により求めた．音声ラベルがついたそれぞれのセグメントごとに平均スペクトルを算出し，入力音響信号内での話者クラスタリングを行う．ユーザインタフェースにおけるラベル修正候補の表示順は，クラスタ中心からのユークリッド距離で決定し，距離の近い順に全クラスタを選択候補として表示する．

3.2 ユーザ訂正の利用

システムの自動認識結果に対しユーザが区間やラベル名を修正したデータを蓄積し，ダイアライゼーションシステムの性能を向上させる．GMM 認識については，訂正データを正解データとして追加し GMM を再学習する．音楽・複数人発話推定については，訂正データを基に音楽の F0 軌跡の平均持続長と周波数変動幅を求め，直線 (音楽) 判定のパラメータである，ヒストグラムを作成するデータ区間長を更新する．話者クラスタリングについては，正解クラスタの分散を求めクラスタ数を更新する．

なおユーザが話者ラベルに登録した個人名については，本稿では認識モデルの更新には使用しない．ただしユーザの視点では具体名をつけることで編集しやすくなるという効果がある．今後多くのデータを集めることで，コンテンツをまたがる話者分類や話者名による検索などの応用が考えられる．

4 評価実験

音響ダイアライゼーションの性能とユーザインタフェースの使いやすさを評価するため実際のポッドキャストを用いて実験を行った．ユーザ訂正のデータはまだ得られていないが，ここでは予備調査として正解付きのデータの一部をユーザ訂正データと仮定して，少量ではあるが，データを追加することでどの程度音響ダイアライゼーションの性能が向上するかをモデルの更新前後で比較し，現在のシステムの有効性と課題を検証する．

4.1 実験条件

実験には Web 上で配信されているポッドキャスト 58 番組を使用した．全て 16bit,16kHz サンプリングで，合計 20 時間 25 分のデータである．人の発話が約 14 時間含まれ，このうち 15.6% は複数人の同時発話であった．正解ラベル付きの本データセットを，初期モデル作成用の 31 番組 (grpA), ユーザ訂正データと仮定した追加学習用の 15 番組 (grpB)，性能評価用の 10 番組 (grpC) の 3 グループに分けて評価を行った．さらに 3 種類の音楽識別用 GMM 作成には，音楽素材集 Palule[12] も合わせて使用した．

4.2 初期性能と訂正データの効果

まずシステムの基本性能となる初期状態での認識性能と訂正データ追加の効果を評価する．grpA のデータから作成した初期 GMM と，ユーザによる訂正データと仮定した grpB のデータを加え再構築した GMM による認識正解率を比較した．対象となるポッドキャストでは複数音の混合区間も含まれるが，ここではラベルごとに独立にフレームごとの認識正解率を求めた．結果を表 1 にまとめる．GMM は表にある 8 種類を作成した．上 5 行が大分類，下 3 行が音楽識別の結果である．左列の各ラベルに対し，2 列目が初期モデルの正解率，4 列目が正解データ追加後の正解率となっている．また 3,5 列目はそれぞれの GMM 作成に用いたデータ長である．大分類の正解率は平均で初期状態の 68.0%から追加学習後は 69.9%に向上した．

表 1　GMM 認識の正解率

ラベル名	初期モデル		追加学習後	
	正解率 [%]	データ長 [hour]	正解率 [%]	データ長 [hour]
音声	75.53	7.73	76.24	11.75
音楽	65.24	0.38	67.99	0.57
音楽+音声	69.38	1.64	73.76	2.35
その他の音	69.65	0.02	68.53	0.03
無音	60.42	1.00	63.04	1.47
BGM	70.04	1.42	70.49	1.53
Jingle	58.76	0.15	58.47	0.19
効果音	47.55	0.04	47.55	0.04

次に SACF による音楽/複数人発話推定，および話者クラスタリングの評価を行う． 評価指標には DER(Diarization Error Rate)[13] を用いた．DER は NIST で提案されている評価尺度で，次式で表される．

$$DER = \frac{\text{誤受理，誤棄却した時間長}}{\text{全データ時間長}} \times 100[\%] \quad (2)$$

誤受理とは対象音がない区間で誤って検出することを指し，逆に後棄却とは対象音を検出できなかった区間を指す．ここではラベルごとに独立に DER を求める．評価では NIST の測定基準に従い，正解ラベルに対し前後 250ms までのずれを許容した．GMM の追加学習と同じ grpB のデータから音楽判定に用いるデータ区間長および話者のクラスタ数を更新し, grpC

に対する DER をパラメータの更新前後で比較した．音楽判定を行うデータ長は，初期状態で設定した 3 秒が 1.75 秒となり，話者クラスタ数は最大 8 と設定したが，更新後は平均 2.8 となった．表 2 の結果は学習後に性能が向上しており，より適切なパラメータに近づいたといえる．パラメータの更新により主に音楽区間の誤受理 (音楽区間以外を誤って音楽と判定) と複数人発話の誤棄却 (複数人発話区間を検出できなかった部分) が減少した．

表 2　音楽/複数人発話推定，話者分類の DER

	初期状態 [%]	学習後 [%]
音楽区間推定	35.27	31.06
複数人発話推定	41.09	36.50
話者分類	38.41	30.44

4.3　ユーザ訂正機能の評価

インタフェースの機能と使いやすさを評価する予備実験として，訂正のしやすさの指標となる，話者ラベルの訂正候補に含まれる正解ラベルの割合 (セグメント単位の正解率) を求めた．平均 3.2 人の発話を含むポッドキャストに対する話者ラベルの表示結果は，第 1 候補のみに含まれる正解率が 64.7%，同様に第 2 候補までが 82.4%，第 3 候補までが 91.1%となった．

さらに著者の一人が実際にポッドキャストを視聴しながら，再生の一時停止を一切しないという条件でどれだけ訂正可能かを調査した．3 番組 (計 1 時間 36 分) について一回の視聴で訂正可能なデータ量 (訂正データのフレーム単位の正解率) を評価したところ，初期状態で正解率が平均 60.2% のデータに対し，訂正後は 88.7%となった．聴きながら少ない労力で大部分を編集可能なインタフェースであるといえる．区間修正のミスは比較的少なく，頻繁に話者が交替する会話に対し一回の視聴では訂正しきれない部分が残った．

5　おわりに

本稿では，ポッドキャストを対象として「どの部分が何の音か」を推定する音響ダイアライゼーションシステムを構築し，ダイアライゼーション結果を可視化することで聴きながらユーザがシステムの認識誤りを訂正可能な視聴インタフェース PodDiarizer を提案した．本システムは，Web 上に大量のデータがあるポッドキャストを対象とし，コンピュータによる認識誤りをユーザに訂正してもらうことでデータを蓄積し，未知の条件が多い実環境の音へ対応できる柔軟なモデルを構築可能であることが特徴である．今後の展開としては，ユーザ入力データの蓄積により更新される認識モデルを定量的に評価し，後続の音声認識やシーン分析の前処理として本システムによる音響ダイアライゼーション技術で可能になることを示していく予定である．このような後続のアプリケーションの性能が向上することを示していくことでユーザにとっても利便性が向上し，さらなるユーザ協力を期待できる．

本研究で提案したユーザとのインタラクションを利用した音響ダイアライゼーションシステムは，Computer Vision 分野で活発に議論されている一般物体認識 (制約のない画像中に含まれる物体を一般的な名称で認識する問題)[14] を，音響信号で実現するためのひとつのアプローチであり，将来コンピュータが実環境の様々な音を扱うための第一歩としてその問題点や可能性を明確にする，という学術的な意義がある．また一般音響認識のモデル構築という目的の他に，使ってもらうことで専門家以外に技術を知ってもらうという社会的意義がある．今後、自律型ロボットの聴覚機能としての環境音理解や，ライフログデータへの音響情報の付加などの応用が期待できる．

本稿では，ユーザ協力を活用した音響ダイアライゼーションというコンセプトを提案し，システムの一構成法を示したが，基本となる信号処理手法をはじめまだ完全ではない．より多くのユーザ協力を得るためにも基本性能をあげることが重要であり，PodDiarizer を運用しながらその特長や課題を検証し，システムを更新していくことが今後の課題である．

参考文献

[1] K. Lee. *Analysis of Environmental Sounds*. PhD thesis, Columbia University, 2009.

[2] L. Lu, R. Cai, and Alan Hanjalic. Audio elements based auditory scene segmentation. In *Proc. of Acoustics, Speech and Signal Processing*, pp. V–V, Toulouse France, 2006.

[3] D. A. Reynolds and P. Torres-Carrasquillo. Approaches and applications of audio diarization. In *Proc. of International Conference on Acoustics, Speech, and Signal Processing 2005*, pp. V–953–956, Philadelphia, PA, USA, 2005.

[4] M. Kotti, V. Moschou, and C. Kotropoulos. Review: Speaker segmentation and clustering. *Signal Processing*, 88(5):1091–1124, 2008.

[5] 堀 貴明 他. いつ誰が何を話したか即座に認識するオンライン会話分析システム～(1) コンセプトとデザイン～. 日本音響学会 2010 年秋期研究発表会講演論文集, pp. 2-9-6, 2010.

[6] K. Sumi, T. Kawahara, J. Ogata, and M. Goto. Acoustic event detection for spotting hot spots in podcasts. In *Proc. of 10th Annual Conference of the International Speech Communication Association*, pp. 1143–1146, 2009.

[7] S. S. Chen and P. S. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the bayesian information criterion. In *In Proc. of the DARPA Broadcast News Transcription and Understanding Workshop*, pp. 127–132, 1998.

[8] D. Wang and G. J. Brown. *Computational Auditory Scene Analysis*. IEEE Press, 2006.

[9] M. Jerome Hawley. *Structure out of Sound*. PhD thesis, Massachusetts Institute of Technology, 1993.

[10] K. Seyerlehner, T. Pohle, M. Schedl, and G. Widmer. Automatic music detection in television productions. In *Proceedinfs of the 10th International Conference on Digital Audio Effects*, Bordeaux, France, 2007.

[11] D. Arthur and S. Vassilvitskii. k-means++: the advantages of careful seeding. In *Proc. of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 2007.

[12] Sound library:Palule. http://tobiuo.sytes.net/palule/palule_intro.htm.

[13] Diarization Error Rate. http://www.xavierangu era.com/phdthesis/node108.html.

[14] 柳井 啓司. 一般物体認識の現状と今後. 情報処理学会論文誌 コンピュータビジョンとイメージメディア, Vol.48, CVIM19, pp.1–24, 2007.