Vol.2011-SLP-87 No.16
2011/7/23

PodDiarizer: ポッドキャスト音声認識・理解のためのユーザ訂正活用型音響ダイアライゼーションシステム

佐々木　洋子[†1]　緒方　淳[†1]　後藤　真孝[†1]

本稿では，ポッドキャスト等の Web 上の音コンテンツ中の典型的な音響イベント（背景音楽，Jingle，効果音，話者別発話，笑い声等）の区間やそれらの混合区間を自動的に検出する音響ダイアライゼーションシステム「PodDiarizer」を提案する．こうした音コンテンツに対する音響ダイアライゼーションは，実用的な音声認識・理解のために不可欠である．PodDiarizer では，音響信号の再生に同期して，音響イベントの検出結果をスクロール表示し，その誤りをユーザが容易に訂正できるユーザインタフェースを提供する．その誤り訂正結果を用いて音響イベントのモデルを改善することで，音響ダイアライゼーションシステムの性能を向上させることができる．ポッドキャストを対象とした実験により，音響ダイアライゼーションの性能を評価し，音声認識システムに適用した際の性能向上を確認した．

PodDiarizer: An Audio Diarization System Based on User Corrections for Speech Recognition and Understanding of Podcasts

Yoko Sasaki,[†1] Jun Ogata[†1] and Masataka Goto[†1]

The paper proposes an audio diarization system, *PodDiarizer*, that can automatically detect typical audio events such as background music, jingle, sound effect, spoken voice of each speaker, laugh, and a combination of these, in *podcast* audio files on the web. Audio diarization of such audio content is indispensable for practical speech recognition and understanding. PodDiarizer provides a user interface in which users can easily correct diarization errors by editing on the scrolling visualization in synchronization with the audio playback. The results of the error correction can then be used to improve the performance of our diarization system by updating models for audio events. We evaluated the performance of audio diarization for podcasts and confirmed that diarization results improved speech recognition performances.

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．また，いつ何の音がしたかがわかると，ライフログデータの自動分析や自律型ロボットの環境認識など，音環境理解としての様々な応用も期待できる．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響信号を扱うためには，各種認識技術の前段処理として，まず何の音なのかを理解する技術[1)–3)] が有用である．

本稿では，図 1 のように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)[4)] について述べる．近年，

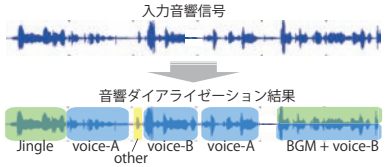


図 1　音響ダイアライゼーション
Fig. 1　Audio diarization

音声認識分野では Rich transcription や話者ダイアライゼーションの研究が活発[5),6)] であり，対話コンテンツを中心に「誰がいつ話したか」を推定する技術が発展してきている[7)]．会話シーンの分析システム[8)] 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[9)] も提案されている．しかし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，課題が多く困難であった．

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を備えた音響ダイアライゼーションシステム「PodDiarizer」を提案する．音響ダイアライゼーションにおける各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

†1 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)

利用するのではなく，日々変化する Web 上の音コンテンツも併用する．実環境中の音に対応するためのひとつのアプローチとして，一般ユーザにダイアライゼーション結果を訂正してもらうことで，未知の音モデルの獲得や，継続的に更新可能なモデル構築が可能となる．こうした音声システムでユーザの誤り訂正という貢献を活用する枠組みとその重要性は，音声情報検索システム PodCastle[10] で初めて提唱された．

以下の章では，まず提案する音響ダイアライゼーションシステムの概要とユーザインタフェースについて述べる．次に音響ダイアライゼーション機能として，音響イベントごとの信号処理手法，ユーザ訂正を利用した学習手法について説明する．最後に構築したシステムの初期性能とユーザ入力データを追加学習の有効性を，実際のポッドキャストを利用した音響ダイアライゼーション実験により評価する．また提案する音響ダイアライゼーションシステムを前処理として利用した際の音声認識性能についても比較検証する．

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を備えた音響ダイアライゼーションシステム，PodDiarizer を提案する．

2.1 PodDiarizer の基本機能

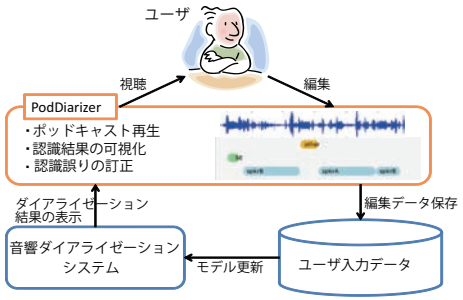
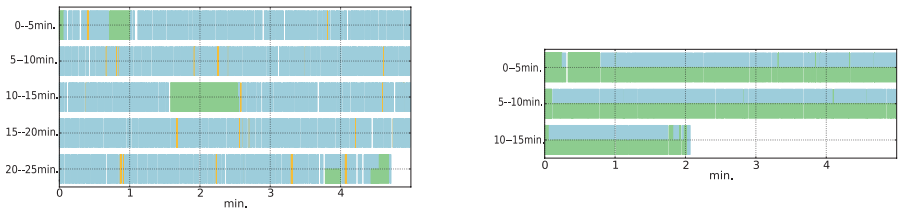


図 2 ユーザ協力型音響ダイアライゼーションシステム PodDiarizer
Fig. 2 User collaborative audio diarization system: PodDiarizer

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 つのインタフェース機能がある．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．

2.2 インタフェースの構成



a) 対話番組　　b) 経済解説番組

図 3 ダイアライゼーション結果の全体表示
Fig. 3 Overview window

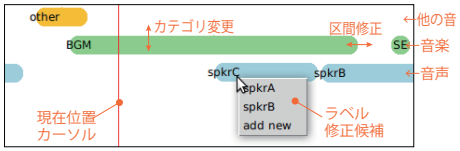


図 4 編集画面の例
Fig. 4 Snapshot of the PodDiarizer interface

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 3,4 に示す．図 3 はコンテンツ全体のダイアライゼーション結果を音の種類ごとに色分け表示したものである．音声区間を青，音楽区間を緑，その他の音区間を黄色で表示している．この画面により，聞く前におおよそどんな種類のコンテンツなのかを把握可能である．たとえば図 3 はどちらも人の発話がメインの番組であるが，a) では音声を示す青領域が細かく区切れているのに対し，b) では比較的長く話している様子がわかる．また b) では発話中に音楽が流れている．

図 4 の編集画面は，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，大分類から詳細識別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．また音声区間については発話と区別するために笑い声の検出も行う．システムの概略を図 5 に示す．大きく分けて，セグメンテーション，大分類，詳細識別の 3 つの部分からなる．以下，図 5 の赤枠で示した各モジュールの信号処理手法について説明する．

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)[11] 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．

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

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．

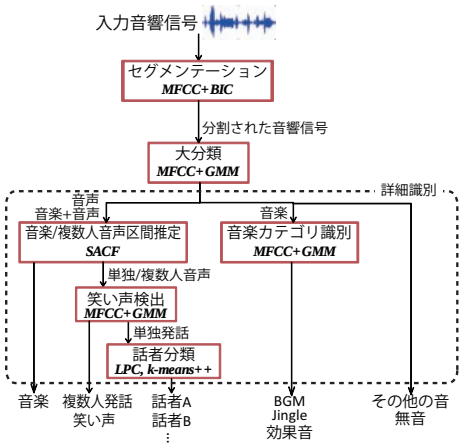


図 5 音響ダイアライゼーション処理の流れ
Fig. 5 Calculation flow of the audio diarization system

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，識別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．また音楽/複数人音声区間検出の結果，音声と判断された区間に対し笑い声の検出を行う．ここでは笑い声を含む区間を全て笑い声区間とし，単独発話，複数人同時発話，笑い声の識別を行う．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・複数人音声区間推定

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)[12] 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[13),14)]，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 法を基本に，クラスタ中心の初期化法を改良した k-means++[15] 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

システムのダイアライゼーション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，システムの学習に利用することでダイアライゼーション性能を向上させる．GMM による識別を行うモジュールについては，訂正データを正解データとして追加し GMM を再学習する．再学習では初期モデルに対し追加データを用いて MLLR による適応を行う．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

提案する音響ダイアライゼーションの基本性能，ユーザ訂正を利用した学習の効果を確認するため，実際のポッドキャストを用いて評価実験を行った．ただし現時点では実際のユーザによる訂正データが収集されていないため，ここではシミュレーション実験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，データの追加によって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検証する．また音声認識システムの前処理として本音響ダイアライゼーション手法を適用し，音声認識率にどのように影響するかを検証する．

4.1 実験条件

実験には実際に Web 上で配信されているポッドキャスト 56 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 28 番組 (grpA，10 時間 22 分)，ユーザ訂正データと仮定した追加学習用の 12 番組 (grpB，4 時間 9 分)，性能評価用の 16 番組 (grpC，5 時間 54 分) 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，市販の放送用音楽素材集[16] も合わせて使用した．

4.2 初期性能と訂正データの効果

システムの基本性能となる初期状態でのダイアライゼーション性能と訂正データ追加の効果を評価する．ユーザ入力データは完全とは限らず，追加学習用のデータにも誤りが含まれることが想定される．そこで grpB の正解データに擬似的に誤りを含め，どの程度影響するかを検証した．

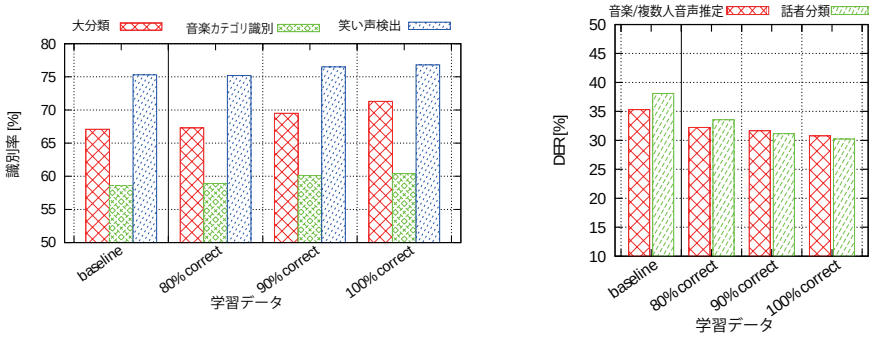
まず図 5 中，MFCC+GMM を採用した 3 つのモジュールについて，grpA のデータから作成した初期 GMM と，ユーザによる訂正データと仮定した grpB のデータを加え追加学習した GMM による識別率を比較した．

モジュールごとの平均識別率を図 6a) に示す．図左端，grpA で学習した初期状態での識別率は，大分類が 67.1%，音楽カテゴリ識別が 58.6%，笑い声検出が 75.3%となった．大分類では混合音に対する誤りが多く,「音楽+音声」と「音声」の誤りが目立った．BGM/Jingle/

Vol.2011-SLP-87 No.16
2011/7/23

効果音を識別する音楽カテゴリ識別は性能が低く現状の GMM によるモデル化では高精度な識別は困難であった．笑い声検出では，複数人の同時発話と笑い声の混同がみられた．

これをベースラインとして，grpB のデータで再学習させると完全な正解データでは平均 7.7%識別率が向上した．20%誤りを含む学習データでは，正解率が下がった音響イベントもあったが，平均正解率は初期状態とほぼ同値となり，学習データに含まれる誤りが 10%以下ではいずれの音響イベントも性能が向上する結果となった．再学習時のモデルパラメータ推定手法として MLLR を用いているため少ない追加データに対し有効に働いていると言える．



a) GMM モジュールの識別率　　b) 音楽区間推定，話者分類の DER

図 6 訂正データを利用した音響ダイアライゼーションの評価
Fig. 6 Evaluation of audio diarization using user correction data

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った．評価指標には DER(Diarization Error Rate)[17] 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．ここでは音響イベントごとに独立に 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 と同様に擬似的に誤りを含めた追加学習データで学習前後の DER を求めた結果を図 6b) に示す．初期状態での DER は，音楽/複数人発話推定が 35.3%，話者分類が 38.1%となった．この結果をベースラインとして，grpB のデータで学習した結果を比較すると，正解率 80%のデータであっても，学習後は平均 3.8%DER が減少しており，誤りを含むデータであっても追加データでパラメータの更新が有効に働いていることが確認できる．

表 1 音響ダイアライゼーションを利用した音声認識結果
Table 1 Speech recognition results using audio diarization

	ベースライン	音楽	他の音	笑い声	複数人音声	**全機能**
単語誤り率 [%]	61.40	60.32	61.06	61.26	60.67	**58.97**
認識対象時間長	5.90 時間	5.77 時間	5.81 時間	5.86 時間	5.68 時間	**5.43 時間**

4.3 音声認識システムへの適用

提案する音響ダイアライゼーション手法を音声認識システムに適用した例を示す．提案法の有効性を音声認識性能 (ポッドキャストに対する書き起こし精度) で評価した．本実験では，音響ダイアライゼーションによりいかに後段の音声認識で対象とすべき音声発話を特定できたか，音声認識率にどのように影響するかを調査する．

PodDiarizer で得られたダイアライゼーション結果を用いて，音声認識すべき区間を抽出する．ここでは音声が含まれない部分および笑い声，複数人の同時発話区間を音声認識対象から除外する．音声認識には PodCastle 音声認識システム[18] を用いており，語彙サイズ 24K の 3-gram による大語彙連続音声認識を行った．grpC のテストデータ 16 番組に対し，コンテンツ全体を書き起こした場合と音響ダイアライゼーション結果を用いて音声以外を除外した場合の認識性能を比較した．音響ダイアライゼーションには 100%正解のデータで学習後のモデルを用いた．テストデータは全体で 5 時間 54 分あり，このうち 4 時間 47 分が発話区間である．評価指標には単語誤り率を用いた．なお本テストデータは，複数人での対談や芸能人のラジオ番組，発話中の音楽重畳など，音声認識が困難なタスクとなっている．

結果を表 1 にまとめる．「ベースライン」とはコンテンツ全体を一定の無音を基準に区分化し，それらを全て音声認識システムに入力したときの結果である．これを基準に，音響ダイアライゼーション結果から，音楽 (BGM, Jingle, 効果音)，他の音，笑い声，複数人音声を音声認識対象から除外した場合の効果をそれぞれ検証した．「全機能」とは上記 4 種類を全てを音声認識対象から除外した結果である．

ベースラインとなるコンテンツ全体を音声認識システムへ入力した場合と比較して，提案システムで抽出された発話区間のみを入力とすることで挿入誤りが低減され，全体として 2.4% 単語誤り率が減少した．非発話区間とした音の種類ごとに個別に検証すると，いずれもベースラインより低い単語誤り率となっており，音声認識前のフィルタとして音響ダイアライゼーションシステムが有効に機能していることが確認できる．

本節では音響ダイアライゼーション結果利用の一例として，発話以外の音による認識誤りを減少させる効果を示した．音響ダイアライゼーション結果には様々な情報が含まれており，このほかにも，話者ごとの音響モデル適応，BGM を考慮した音声認識，複数人同時発話のための音声認識等，状況に応じた適切な認識手法を選択する手段として様々な利用が考えられる．

4.4 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「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について述べた．ポッドキャストを対象とし，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を備えた音響ダイアライゼーションシステム「PodDiarizer」を提案した．本システムは，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．

本稿では，ユーザの協力を得ることで継続的にモデルを更新可能な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．また映画やテレビの音，日常環境での収録音など，ポッドキャスト以外の音コンテンツに対しても本システムを利用し，異なる条件下における提案システムの有効性や課題を検証していきたい．

参 考 文 献

1) Keansub Lee. *Analysis of Environmental Sounds*. PhD thesis, Columbia University, 2009.
2) Lie Lu, Rui Cai, and Alan Hanjalic. Audio elements based auditory scene segmentation. In *Proc. of International Conference on Acoustics, Speech and Signal Processing*, volume 5, pages 17–20, 2006.
3) 林田 亘平, 溝口 遊, 森勢 将雅, 西浦 敬信. 擬音語 HMM に基づく音場ディクテーションの検討. In 電子情報通信学会技術研究報告, SIP, volume 110, pages 55–66, 2010.
4) 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*, volume 5, pages 953–956, 2005.
5) Kentaro Ishizuka, Shoko Araki, Kazuhiro Otsuka, Tomohiro Nakatani, and Masakiyo Fujimoto. A speaker diarization method based on the probabilistic fusion of audio-visual location information. In *Proc. of the 2009 International Conference on Multimodal Interfaces*, 2009.
6) 藤本 雅清. 音声区間検出の基礎と最新の研究動向. In 電子情報通信学会技術報告, SP, volume 110, pages 7–12, 2010.
7) Margarita Kotti, Vassiliki Moschou, and Constantine Kotropoulos. Review: Speaker segmentation and clustering. *Signal Processing*, 88(5):1091–1124, 2008.
8) 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〜. In 日本音響学会 2010 年秋期研究発表会講演論文集, 2010.
9) Kouhei Sumi, Tatsuya Kawahara, Jun Ogata, and Masataka Goto. Acoustic event detection for spotting hot spots in podcasts. In *Proc. of 10th Annual Conference of the International Speech Communication Association*, pages 1143–1146, 2009.
10) 後藤 真孝, 緒方 淳, 江渡 浩一郎. PodCastle: ユーザ貢献により性能が向上する音声情報検索システム. 人工知能学会論文誌, 25(1):104–113, 2010.
11) Scott S. Chen and P. S. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the bayesian information criterion. In *Proc. of the DARPA Broadcast News Transcription and Understanding Workshop*, pages 127–132, 1998.
12) DeLiang Wang and Guy J. Brown. *Computational Auditory Scene Analysis*. IEEE Press, 2006.
13) Michael Jerome Hawley. *Structure out of Sound*. PhD thesis, Massachusetts Institute of Technology, 1993.
14) Klaus Seyerlehner, Tim Pohle, Markus Schedl, and Gerhard Widmer. Automatic music detection in television productions. In *Proc. of the 10th International Conference on Digital Audio Effects*, 2007.
15) David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In *Proc. of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, 2007.
16) Sound library: Palule. http://tobiuo.sytes.net/palule/palule_intro.htm.
17) Diarization Error Rate. http://www.xavieranguera.com/phdthesis/node108.html.
18) Jun Ogata and Masataka Goto. PodCastle: Collaborative training of acoustic models on the basis of wisdom of crowds for podcast transcription. In *Proc. of 10th Annual Conference of the International Speech Communication Association*, pages 1491–1494, 2009.