ラジオ放送話者ダイアライゼーションシステム
Speaker Diarization System For Radio broadcast

安田 沙弥香
Sayaka Yasuda
法政大学情報科学部ディジタルメディア学科
E-mail: sayaka.yasuda.zz@stu.hosei.ac.jp

Abstract

This paper proposes a method for constructing a speaker diarization system for radio broadcast to estimate a speaker for making it easier to understand a conversational content in radio broadcast.This system automatically processes segmentation of a radio program,classification of speech and music,suppression of BGM and speaker identification.As the result of identification,even a name of a speaker is estimated and visualized.Radio broadcast includes not only speech but music and BGM,so it is hard to set a range of application of speaker identification.So,to classify speech and music,Bayesian information criterion based segmentation is performed.When classified speech signals convolute music and speech,BGM is removed using the extraction of the repeating musical structure.Speaker identification is performed for each speech signal.Although casts of radio are many,the amount of data of each speaker becomes small by the effect of BGM.Therefore it is appropriate to construct speaker models that adapt Gaussian Mixture Models-Universal Background Models so that it makes possible to identify a speaker with high-accuracy.The speaker diarization system for radio broadcast that had 39.5% accuracy is constructed using a proposal method.Because the speaker identification rate of clear data that is extracted manually from a radio program was 96.7%,it is concluded that the multi-speaker identification using actual radio broadcast data is realized.

1 まえがき

ラジオ放送はテレビと違い映像や字幕情報がない．現在ラジオでは，受信機での聴取以外にも IP サイマルラジオ放送によるインターネット経由での聴取が定着しつつある．しかし音声以外の字幕情報などは配信されておらず，未だラジオの情報量は少ない．複数の話者がいると誰が話しているのか，といった話者了解度が低く話の内容が理解しにくくラジオ離れが進む原因のひとつになっている．そこで本研究では，話者識別の技術を用いてラジオでの発話者が誰であるかを推定し，発話者の名前を PC の画面上に表示する話者ダイアライゼーションシステムを構築し，ラジオ離れの問題を解決する．

話者識別とは発話者が誰であるかを判定するシステムである．特に「いつ，誰が話しているか」という情報を推定する技術を話者ダイアライゼーションと言う．従来研究では会議音声に多く用いられ [1]，議事録の自動作成などに応用されている．またポッドキャストや TV 放送への適応例はあるが，実際に配信されている音声データの特徴のみを用い，名前まで同定するシステムは少ない [2][3]．話者識別をラジオへ適応する問題は，識別対象が多人数であること，背景雑音として音楽や複数人の声の重なりが存在し，事前に集められる話者データやその量が少なくなることである．

本研究では IP サイマルラジオを対象として，セグメンテーションから音声・音楽分類，BGM 処理，話者の識別まで全て自動で行うシステムを提案する．結果は発話者の名前を PC の画面上に表示する．またシステムのユーザが混乱を招かないものにするため，識別される話者のモデルが存在しない場合は未知話者であることの表示や，識別結果に確実性が持てない場合は複数人の候補者名を表示する．

2 ラジオ放送での話者ダイアライゼーション

複数話者による会話の分析を行う研究では発声の内容だけでなく誰がどの部分を話しているのかという情報も重要である．そのような情報を抽出する処理を話者ダイアライゼーションと呼ぶ．一連の音響信号を入力としセグメンテーションやクラスタリングを行い，各信号が誰の発話であるかを推定する．現在，この技術はテレビや会議，電話音声に多く応用されている．名前まで同定するダイアライゼーションシステムでは，話者識別や音源方向推定などの手法が用いられている．本研究ではラジオ放送を対象として話者識別を用いたシステムを実現する．会議とは違い，ラジオ全体の総出演人数を考慮すると 1,000 人を越え，その中から数人に絞るため識別対象が多人数となる．また，マイクや人物の位置もユーザーが指定できないため音源方向の情報や，テレビ放送の字幕情報などはラジオ放送では活用できない．しかしラジオの番組表には出演者の名前が記載されているためこの情報は活用できる．

ラジオ放送話者ダイアライゼーションの問題点はラジオ番組の識別対象者が多人数であることに加え，背景雑音として BGM や声の重なりが多く存在することである．よって事前に収集できる学習データが少なく，また識別率が低下する．さらに一発話の時間も短いものしか収集できない．よって BGM 付き音声かつ少量データからの大規模人数話者識別の性能を向上させる手法を検討する．音楽は非定常な雑音であるため，音楽の構造に着目した除去方法を検討する．モデル化は少量のデータからも学習できるものでなくてはならない．また事前にラジオに出演する全ての話者モデルを構築できない．よって事前にモデルを持たない話者である場合は，未知話者と判断するなど誤識別を最小限に抑えシステムユーザの混乱を防がなければならない．これらの問題点を踏まえ，ダイアライゼーションシステムの構築をする．

3 システム構成

ラジオ話者ダイアライゼーションシステムの実現方法について述べる．システム全体の概略を図 1 に示す．話者の名前まで同定するため事前に学習した話者モデルを用いる．IP サイマルラジオの一番組を入力信号とし，セグメンテーションを行い音声，BGM 付き音声，音楽，無音部分に分類を行う．その後，音楽と音声が重畳された BGM 付き音声は，BGM 除去の処理をし音声区間とする．音声区間では信号内に複数人が発話している場合があるためさらにセグメンテーションをし単独発話に

Supervisor: Prof. Katunobu Itou

対して話者識別を行う．

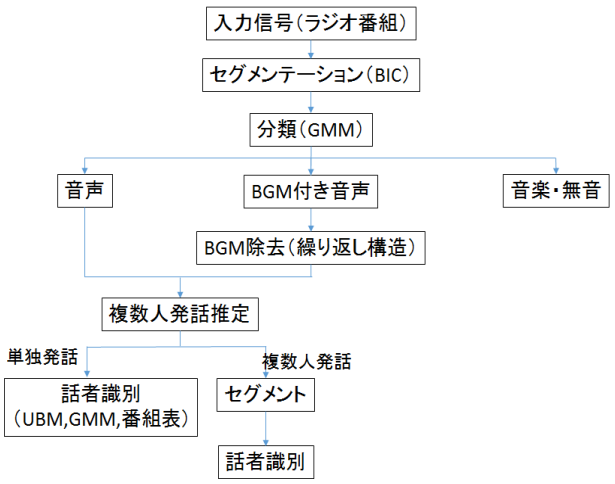

図 1. ラジオ話者ダイアライゼーションの処理の流れ

3.1 ベイズ情報量基準によるセグメンテーション

ベイズ情報量基準 (BIC) を用いて類似した音が含まれるセグメンテーション手法が提案されている [4]．この手法を用い一連の入力信号に切り分ける．データの集合を $X = \{x_i \in R : i = 1, ..., N\}$, モデル候補を $M = \{M_i : i = 1, ..., K\}$ とする．このとき $L(X, M)$ をモデル M の最大尤度とし，$\#(M)$ をモデル M のパラメータ数とすると BIC は式 (1) のように定義される．

$$\mathrm{BIC}(M) = \log L(X, M) - \lambda \frac{1}{2} \#(M) \log(N) \qquad (1)$$

BIC を利用したセグメンテーションでは，データが N 点のある区間に対して 3 つのモデルを想定する．1 つは区間全体 (0〜N) のモデル $M_0 = \mathrm{N}(\mu_0, \Sigma_0)$，他の 2 つは時刻 0〜$j$ と時刻 $j+1$〜N までのモデル $M_{12} = \{M_1, M_2\} = \{\mathrm{N}(\mu_1, \Sigma_1), \mathrm{N}(\mu_2, \Sigma_2)\}$ とする．式 (2) が極大値となる時刻 j がセグメントの境界となる．

$$\begin{aligned} \Delta\mathrm{BIC}(j) &= \mathrm{BIC}(M_0) - \mathrm{BIC}(M_{12}) \\ &= \frac{1}{2}(N \log|\Sigma_0| - j \log|\Sigma_1| - (N-j) \log|\Sigma_2|) \\ &\quad - \frac{1}{2}\lambda \left(d + \frac{1}{2} d(d+1) \right) \log N \end{aligned} \qquad (2)$$

ここで，d は特徴量の次元数である．本研究ではこの手法を用い音声，音楽などの変化の検出を行うため可変長窓を用いた分割を行う [5]．手順は以下のとおりである．

(1) フレームを最小幅に設定する（$[a, b] = [1, 2]$）．
(2) BIC を用いて a, b 間に変化点があるか入力の最初の点から探索する．
(3) a, b 間に変化点がない場合 ($\Delta BIC < 0$)，最小フレーム幅を足し，処理を続ける．
(4) a, b 間に変化点があった場合 ($\Delta BIC > 0$)，その点を始点として処理を続ける．
(5) 入力の最後まで (3)(4) を繰り返す．

特徴量は 12 次元 MFCC，ΔMFCC，正規化対数パワーの計 25 次元とする．

3.2 音声と音楽の GMM 分類

BIC でセグメントされた信号を音声，音楽，BGM 付き音声，無音区間にカテゴリ分類する．本研究での音楽は歌声入りの音楽，ジングル，効果音とする．事前学習として各分類それぞれのデータを IP サイマルラジオより収集し，GMM でモデル化する．セグメントされた信号と分類モデルとの尤度を比較してカテゴリ分類する．また著しくパワーが低い信号は無音区間とする．

3.3 複数人発話区間のセグメンテーション

話者ダイアライゼーションにおいて話者を識別する音声は一人の話者のみが含まれるようにしなければならない．しかし BIC によりセグメンテーションされた音声には話者交代がある可能性が高い．よってさらにセグメントを行い単独発話にする．本研究では 2 つの手法を検討する．

3.3.1 パワー変動

対数パワーに基づくセグメンテーションを行う．話者交代や息継ぎの際，音が途切れるようにパワーの変動が生じる．この変動を検出することで一単語ごとに区切れ話者の交代がある音声でも単独発話にできる．

3.3.2 F0 変動

基本周波数 (F0) の変動に基づき単独発話へのセグメントを行う．呼気によって区切られる区間内では F0 が上昇し，のち下降するという性質に注目する．

3.4 BGM 除去

ラジオの背景雑音には定常な雑音だけでなく BGM の非定常な雑音が存在する．SNR ごとに識別をし，SNR と話者識別の精度の関係を観察する．実際に話者 8 人のクリーンデータに対し，SNR7 段階ごとに BGM を付与した各話者 7，計 56 サンプルによって実験した結果を図 2 に示す．

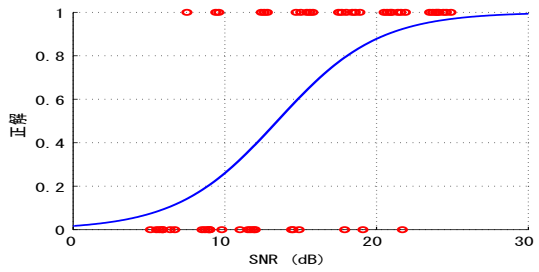

図 2. SNR と話者識別の正解数との関係

図より，SNR が大きくなると正解確率が上がるという関係性が見られる．よって BGM は話者識別に影響を与えるため，除去を行い識別率の向上を図る．本研究では 2 つの手法を検討する．

3.4.1 スペクトル減算法

一つ目は比較的単純で短時間で処理を行えるスペクトル減算法（以下 SS 法）[6] である．SS 法は雑音のパワースペクトルの平均値を推定し，観測信号のパワースペクトルから引くことで雑音の低減を行う．

3.4.2 繰り返し構造の抽出

二つ目は音楽の基本的な特徴である繰り返しに着目し，音楽と声に分離する手法 [7] である．音楽部分はある程度の間隔の繰り返し構造を持ち，音声部分は繰り返し構造を持たないことを活用する．

3.4.3 ラジオ音声における BGM 処理の比較

図 3, 図 4, 図 5 は，IP サイマルラジオ放送から録音した BGM 付きの 6 秒間の音声信号と SS 法, 繰り返し構造の抽出で雑音を除去したスペクトログラムである．背景雑音が抑制され，音声部分が強調されていることがわかる．

前述より SNR が大きくなるにつれ識別率が向上すると評価した．よって SNR が大きくなるよう BGM 除去をし，性能評価を行った．識別率は 22.5% から SS 法は 42.5%，繰り返し構造抽出後は 52.5% に向上した．よって BGM 処理は有効である．本システムでは精度の良かった繰り返し構造抽出法を用いる．

3.5 話者識別

本研究ではラジオ番組から話者の名前を識別し表示する．よって各話者モデルを事前に構築するための話者データを IP

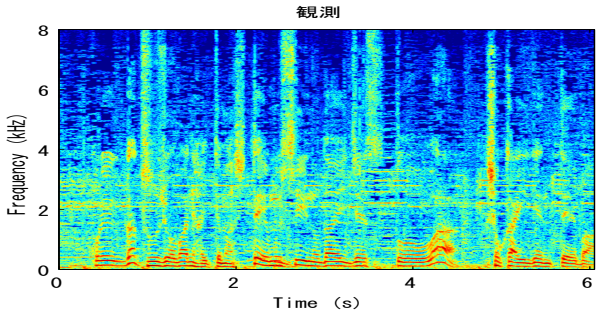

図 3. 元音声のスペクトログラム

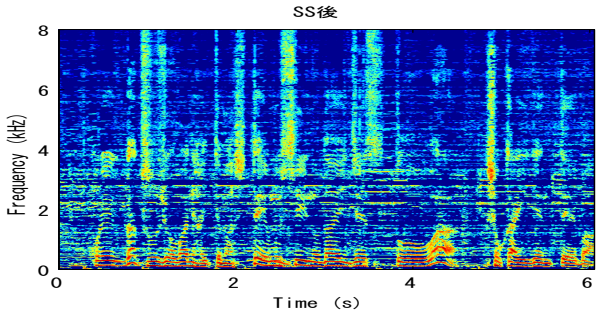

図 4. SS 後のスペクトログラム

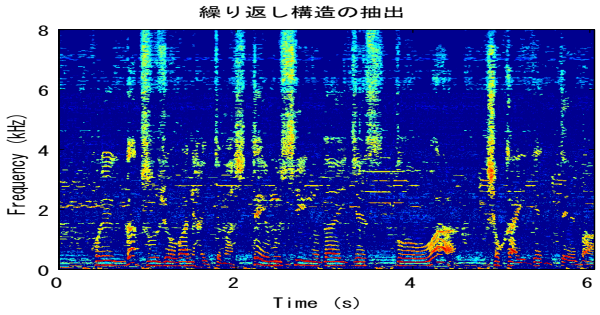

図 5. 繰り返し構造抽出後のスペクトログラム

サイマルラジオより収集する．48 番組，約 50 時間分よりデータを収集した．収集したデータの詳細は表 1 に示す．

表 1. 収集データ

話者数	153 名 (男性:82 名，女性:71 名)
収集期間	2013/2/5〜2013/11/27
ファイル数	1779
全時間長	164.4 分 (平均 : 5.5 秒/1 ファイル)

これらのデータから 153 名分の話者モデルを構築する．大規模人数かつ事前学習が少量のデータでの話者識別となるため，背景モデルを導入し MAP により話者モデルを構成する GMM-UBM（Gaussian Mixture Models-Universal Background Model）を用いる [8].

3.5.1 GMM-UBM

UBM とは話者の特徴分布を表すために学習された一つの大規模 GMM で，雑音などを除く人間の声のみで構成される背景モデルである．また各話者のモデルはこの UBM を適応させ学習する．これは識別において学習とテスト時では発声するテキストが毎回違うことや，BGM により事前に収集できるデータが少量であるため，その話者が発声しうるすべての範囲のモデルを作ることができない．よって音声の一般的なモデルとなる UBM を適応させることでその話者の発声範囲をカバーする．話者適応の方法は，最初に EM アルゴリズムの期待値ステップによって訓練話者のパラメータの期待値を求める．次に式 (3) によって新しい統計量に更新する．

$$
\begin{aligned}
\hat{w} &= [\alpha_i^w \frac{n_i}{T} + (1-\alpha_i^w) w_i]\lambda \\
\hat{\boldsymbol{\mu}} &= \alpha_i^m E_i(\boldsymbol{x}) + (1-\alpha_i^m)\boldsymbol{\mu}_i \\
\hat{\boldsymbol{\sigma}} &= \alpha_i^v E_i(\boldsymbol{x}^2) + (1-\alpha_i^v)(\boldsymbol{\sigma}_i^2 + \boldsymbol{\mu}_i^2) - \hat{\boldsymbol{\mu}_i^2} \\
\alpha_i^\rho &= \frac{n_i}{n_i + r^\rho}
\end{aligned}
\tag{3}
$$

ここで，w は重み係数，$\boldsymbol{\mu}$ は平均，$\boldsymbol{\sigma}$ は分散である．また $\alpha_i^\rho, \rho \in \{w, m, v\}$ で適応の強度を表している．従来，この手法を利用した単一話者検出の実験では等価エラー率 10% という高い精度が示されている [8].

3.5.2 番組表の活用

ラジオには番組表が存在する．その中にはラジオの出演者情報が載っており，事前に番組内に出る人物を特定できる．しかし番組表に載っていない人物も出演する可能性はある．よってあらかじめ番組表から得られる人物の事前確率を高くすることで，識別率の向上を図る．

番組表に載っている話者は，番組の半分以上に出演している．よって本研究では事前実験により事前確率として 0.5 を実際に番組表から得られた話者で振り分け，残りの 0.5 をモデルを持つ話者と未知話者に振り分けた．

3.6 話者決定法

事前にラジオに出演する全ての話者モデルを構築できない．よって事前にモデルを持たない話者が番組に出演した際に，システムユーザの混乱を防ぐため未知話者である判断をしその表示を行う．この決定には未知話者とモデル有りの尤度比を用いる．ここでは識別対象の話者モデルを λ_{hyp}，GMM-UBM を $\lambda_{\overline{(hyp)}}$ とする．対数尤度比は式 4 で評価できる．

$$
\lambda(x) = \log p(X|\lambda_{hyp}) - \log p(X|\lambda_{\overline{(hyp)}}) \tag{4}
$$

未知話者とモデル有りの尤度では，平均的にモデル有りの方が尤度比が高い．よって話者の決定法として以下のルールを定める．事前に調べた結果この二つでは平均的にモデル有りの方が尤度比が高い．よって話者の決定法として次の通りルールを定める．尤度比がモデル有りの平均より大きい場合は，識別結果の話者一名を正解とし表示する．尤度比が未知話者の平均より小さい場合は，モデルを持たない未知話者と表示する．尤度比がモデル有と未知話者の平均の間である場合は，尤度の高かった 5 名の名前表示を行う．

4 評価

話者識別の性能，ダイアライゼーションの性能の評価をするため実際のラジオ放送を用いて実験を行った．

4.1 話者識別の評価

153 名のモデルからテスト話者 13 名（男性 : 7 名，女性 : 6 名）を識別する実験を行った．テスト話者 13 名の学習モデルは 20 秒以上とし，テストファイルはモデル構築に含まないファイルを用いた．テストファイルは全 241 個用意した．特徴パラメータは，標本化周波数 16kHz，フレーム長 64ms，シフト長 32ms の分析条件で，12 次元の MFCC とその ΔMFCC の計 24 次元とした．UBM は前述の IP サイマルラジオからの収集データ全てを用い，32 混合で構築し，平均と分散を適応させ話者モデルを作成した．識別結果は 96.7% となった．13 名のみのモデルから実験したところ 98.8% となり，多数のモデルを使った場合でも，あまり性能は変わらない．誤識別した音声は相槌などの 1 秒以下の音声であった．よって識別する音声は 1 秒以上の音声であることが望ましい．

4.2 ダイアライゼーションシステムの評価

実際の IP サイマルラジオ 2 番組から評価を行った．番組の詳細は表 2 に示す．

表 2. テスト番組の詳細

番組	話者数	データ長 (s)	音声 (s)	BGM 音声 (s)
1(2013/11/27)	5	714	337	351
2(2013/7/23)	6	600	272	143

最初に，この番組から GMM 分類した性能を評価する．評価尺度は誤受理 (false acceptance rate: FAR)，誤棄却 (false rejection rate: FRR) を用いた [1].

$$\mathrm{FAR} = \frac{N_{FA}}{N_{ns}} \times 100$$
$$\mathrm{FRR} = \frac{N_{FR}}{N_{s}} \times 100 \qquad (5)$$

ここで $N_{ns}, N_s, N_{FA}, N_{FR}$ はそれぞれ，非音声フレーム数，音声フレーム数，非音声の誤検出フレーム数，音声の誤検出フレーム数である．表 3 に GMM 分類の結果を示す.

表 3. GMM 分類の結果 (%)

	FAR	FRR
1	10.0	0.0
2	0	6.7
平均	5.0	3.4

表 3 より 2 番組とも，どちらのエラー率も小さく抑えられている．本研究で用いた音楽・音声の分類法は有用であることを示している.

次にダイアライゼーションにおける評価を行った．番組は表 2 を用いる．ここでは複数人発話区間のセグメンテーション手法のパワー変動，F0 変動それぞれでの性能を比較する．評価指標には式 (6) で表される DER を用いた [9].

$$\mathrm{DER} = \frac{\text{誤受理} + \text{誤棄却} + \text{話者誤り}}{\text{全データ時間長}} \times 100 \qquad (6)$$

誤受理 (false alarm speaker: FA) は話者区間の誤検出，誤棄却 (missed speaker: MS) は話者区間の未検出，話者誤り (speaker error: SE) は誤った話者への誤識別を指す．最初にパワーセグメンテーションでのエラー率を表 4 に示す.

表 4. ダイアライゼーションの結果 (%)

	DER	FA	MS	SE
1	46.0	0.9	36.5	8.6
2	43.1	1.2	26.8	15.2
平均	44.6	1.1	31.7	11.9

表より FA が小さく押さえられる．これは前述の通り GMM 分類が効いているためである．一方 MS が多かった．これは識別音声に複数人の話者が含まれたことが原因である．理由として，複数人発話区間の分離法にパワーを採用したためである．話者交代や発話のオーバーラップの際，パワーの変動が少なく推定できなかった.

次に，F0 セグメンテーションでのエラー率を表 5 に示す.

表 5. ダイアライゼーションの結果 (F0 変動)(%)

	DER	FA	MS	SE
1	38.8	1.4	10.0	27.4
2	40.1	3.3	17.4	19.4
平均	39.5	2.4	13.7	23.4

表よりパワー変動と比べ，MS が押さえられた．しかし依然高いエラー率となった．またパワー変動に比べ SE が多くなった．これは F0 でセグメントを行う際に，1 秒程度と短くセグメントされ 4.1 で示したように誤識別が多くなったためである．よって話者交代に頑健かつひとつのセグメントが長くなるよう，発話のオーバーラップのみを検出しセグメンテーションする手法を適用することで改善が見込める.

また，未知話者の判断でのエラー率が高くなった．よって未知話者の尤度比を用いた判断は有効性が見られなかった．改善のためにはユーザーが聞きながら話者名の修正をし，モデル更新を可能にすることが挙げられる．そのほかには BGM 付き音声の識別が悪くなった．BGM を除去した音声は，クリーンデータと比べると音声の劣化が見られる．除去音声とクリーンデータでのモデルとの尤度の比較では対応がとれていないために識別率が低下している．よって除去音声でのモデル構築を行い，除去音声はこのモデルと比較することで識別率の改善が見込める.

5 あとがき

本研究では，ラジオ放送で発話者が誰であるかを推定するための，セグメンテーションから話者識別まで自動で行う話者ダイアライゼーションシステムを構築した．話者識別部では UBM を適応させることで，識別率 96.7% を実現した．ラジオ放送話者ダイアライゼーションシステムとしては DER39.5% と精度が低くなった．しかし FA のエラー率は 2% 程度と低くなり，BIC と GMM 分類が有効であった．話者交代でのセグメンテーションはパワーと F0 の変動では，F0 の変動の方が DER が 5 ポイントよくなり，MS を押さえられた．発話のオーバーラップへ対応するセグメンテーション手法の検討や BGM 除去音声でのモデルを構築し識別することで DER の精度向上が見込める．また固定的なデータセットのみを利用するのではなくユーザが修正し更新可能なモデル構築をできるようにすることが今後の課題である..

参考文献

[1] 荒木, 他, “音声区間検出と方向情報を用いた会議音声話者識別システムとその評価”, 音響論集, pp.1-2，2008.

[2] Poignant J. et.al. “Unsupervised Speaker Identification using Overlaid Texts in TV Broadcast”, the 13rd Annual Conference of the International Speech Communication Association, INTERSPEECH, 2012.

[3] 佐々木, 他, “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”, 音響論集, pp.63-66, 2011.

[4] S. S. Chen, et.al., “Speaker, environment and channel change detection and clustering via the bayesian information criterion”, In InProc. of the DARPA Broadcast News Transcription and Understanding Workshop, pp. 127-132, 1998.

[5] 河原, 他, “音声会話コンテンツにおける聴衆の反応に基づく音響イベントとホットスポットの検出”, 情処学論誌, Vol.52, No.12, pp.3363-3373, 2011.

[6] Steven F. Boll, “Suppression of Acoustic Noise in Speech Using Spectral Subtraction” ,IEEE Trans.Acoustics Speech, and Sigbal Processing, VOL.ASSP-27, No.2, pp.113-120, 1979 .

[7] Rafii Z, et.al., “A simple music/voice separation method based on the extraction of the repeating musical structure”, In Proc. ICASSP, pp.221-224, 2011.

[8] Douglas A., et.al., “Speaker Verification Using Adapted Gaussian Mixture Models”, Digit. Signal Process. 10, pp.10-41,2000.

[9] The Rich Transcription Spring 2003 (RT-03S) Evaluation Plan, `http://www.itl.nist.gov/iaui/894.01/tests/rt/2003-spring/index.html`.