テレビ字幕を利用した番組検索向け発話者アノテーション

山室慶太[†1]　伊藤克亘[†2]

近年放送される大量の映像コンテンツを管理するため，番組内情報をメタデータとして付加する研究が行われている．その一つとして話者情報のメタデータ化がある．しかし，多くの研究は一つのジャンルにのみ対応した手法である．本研究ではドラマ，アニメーション，バラエティ番組を対象とした字幕情報による発話者情報のアノテーション手法を提案する．具体的には，字幕によって音素単位で高精度化され識別モデルを用いて話者識別を行う．識別モデルはベイズ情報量基準（BIC）によって有効な音素モデルが選択される．また，学習データの少ない話者のモデルを改善するため，発話傾向を考慮した話者の出現頻度を用いた識別結果を重み付けする手法を提案する．提案手法を従来の GMM による話者識別と比較した結果，提案手法によって識別性能が 14.14%改善された．

Speaker annotation using closed caption for scene retrieval of television broadcast

Keita Yamamuro[†1] and Katunobu itou[†2]

There has recently been much research on annotation systems for television broadcast management. One approach is to manage the television broadcast by the metadata of speaker information. However, most of the methods developed have specialized in only one genre. Therefore, in this study we targeted three genres drama, animation, and variety and proposed a method of annotating indexical information through metadata obtained from television captions. Specifically, the information from the captions is used to create a phoneme HMM that is then used for speaker identification. The proposed system selects the most appropriate phonemic model from several candidate models based on the Bayesian information criterion (BIC) of likelihood and data. The identification results were weighted by the tendency of utterance. Characters in 30 television programs were identified with a recognition accuracy of 67.76%.

1. まえがき

近年放送のデジタル化とともに，多チャンネル化を向かえ，大量の映像コンテンツが数多く制作・放送されるようになりつつある．これに伴い，映像コンテンツへ番組内容に関する情報をメタデータとして付加する研究が盛んに行われている．付加されるメタデータは，各シーンのイベント情報や，人物情報，内容の要約など様々である．これらの情報は大量の映像コンテンツの中から視聴者がシーンの検索や編集，番組ハイライトなどの要求を支援することが可能となる．しかし，このような映像コンテンツへのメタデータ付与は手動で行われていることが多く，膨大な作業時間が必要であるという問題がある．従って，効率的なメタデータの作成には映像コンテンツの内容を自動で抽出する必要があり，これまでに画像処理，音声処理，自然言語処理など様々な技術を用いた研究が行われてきた[1)–3)]．

メタデータによるシーン内検索を行うのためには発話者情報の付加が重要であると考えられる．そのため，本研究では映像コンテンツに含まれる音声データを処理することで映像内の全ての台詞に対して発話者の情報をメタデータとして付加する．発話者情報は特定人物の登場シーン検索や人物相関情報の取得などに活用できる．しかし，これまでの映像コンテンツに対する音声処理を用いたメタデータ抽出手法はニュース，スポーツなど一部のジャンルの番組で提案されることが多く，ドラマ，アニメーション，バラエティなどのジャンルに対する研究はほとんど行われていない．これらの番組はニュース番組に比べ，環境音や雑音が多く含まれていることや，識別対象者によって学習データの量にばらつきがあること，登場人物の同時発話による誤認識など，自動で発話者情報を抽出することが難しい．

本研究では，音声データに加え映像コンテンツに付随している字幕情報を活用する手法を提案する．従来のニュース番組などの識別で提案された話者単位モデルを字幕情報に含まれる時間情報と発話内容を用いることで，より詳細な音素単位モデルを構築し，発話者の識別を行う．また，識別モデルの学習に用いる音素データ量のばらつきを考慮し，識別に有効なモデルを情報量基準の一つである BIC を用いてモデル選択する．識別結果に対しては字幕情報から登場人物の発話数の傾向を事前確率として用いることで話者識別の結果に重み付けを行い，登場回数が少ない話者の識別を補助する．

†1 法政大学
Hosei University
†2 法政大学
Hosei University

2. デジタル放送の発話者識別

テレビ放送への情報のアノテーションはいくつも研究されてきた．これらの目的は番組の検索，ダイジェスト映像の編集，音声の音声強調やモーフィングなどである．また，これらの提案手法は画像認識，音声認識，自然言語処理技術など様々である．

例えばニュース番組の放送音声に対する研究が行われてきた[6),9)]．ニュース番組は放送が多く，また多彩な情報が含まれている．そのため，その中から視聴者の要求する情報のみを得る場合に話者識別の利用は有効であるといえる．その他にも，スポーツ番組のシーン情報のアノテーションが研究されている．この場合，番組のハイライトが重要なシーン情報から編集できる．

しかし，これらの研究の多くはニュースかスポーツの番組を対象としている．ニュースの場合，識別に使われる音声は雑音が少なく，はっきりと発話されているため，識別音声の区間抽出や話者同定は比較的容易である．しかし，多くのジャンルの番組では様々な雑音が入るため，直接有声区間検出を行うには不向きである．また，各登場人物の発話回数が大きくばらつくことが多いため，全員分の学習データ量を十分に確保することが困難であることが多い．

スポーツの場合,「ゴール」や「シュート」などの特定のキーワードを処理するため，シーン情報が抽出しやすい[7)]．しかし，多ジャンルの番組では発話内容が多様になるため特定キーワードを用いたメタデータの生成は困難である．そのため，ニュースやスポーツ番組以外を対象とした手法を提案し，多くのジャンルの番組でアノテーションを行えるようにする必要がある．

そこで本研究ではドラマ，アニメーション，バラエティのテレビ番組から話者識別によって発話者情報をメタデータとして付与することを目標とする．そのため，すでに多くのテレビ放送に付与され，さらに今後も多くの番組に付与される予定である字幕の情報を本研究で活用することで，話者情報に関するメタデータを自動生成する．

3. 字幕情報の活用

3.1 テレビ字幕

本研究では台詞の発話者識別によって番組内インデックスを付加する．番組内インデックスとしてはシーンの時間情報，話者情報，発話内容を一つのメタデータとして扱う．これらのメタデータは映像コンテンツの音声を話者識別処理することで自動的に話者情報を抽出する．しかし，音声データのみの手がかりで情報を抽出した場合，メタデータの抽出精度が低くなってしまう．そのため，より精度の高い情報抽出を行うためデジタル放送に付加されている字幕情報を活用する．字幕情報は，従来研究[16)]でも映像の情報を抽出する際の手がかりとして活用されており，デジタル放送で話者情報を抽出することにも有効であると考えられる．本研究では，シーン検索などに活用できる話者情報を生成し，映像コンテンツへ自動付与する．

現在，日本のテレビ放送において総放送時間に占める字幕放送時間の割合は 40.1%である．また，字幕放送が不可能な生放送番組や，深夜帯の番組以外の字幕放送が必要と思われるテレビ放送を対象とした場合は，総放送時間の 69.9%に字幕が付与されている．現在も更なる字幕放送の普及へ取り組まれているため，今後はより多くのテレビ放送が字幕放送に対応すると予測される．そのため多くの番組で字幕情報による手法を活用できると考えられる．

3.2 テレビ字幕の持つ情報

字幕情報は字幕の表示時間に関する情報と発話内容に関する情報，字幕のフォントや色情報，そして発話者に関する情報が含まれている．しかし，字幕には全ての台詞の時間情報と発話内容が含まれているが，発話者に関する情報は一部の台詞にしか付加されていない（図 1）．そのため，本研究ではこの話者情報が欠落している台詞に対して話者識別による発話者ラベルの付与を行う．発話者情報が付加されている台詞の条件は，初めて登場した人物の最初の台詞と，画面に多数に人物が映っており，誰が発話しているか分からないような台詞に付加される．ドラマ，アニメーションの場合，発話者が分かりやすい場面が多いため，発話者情報が付加されている一人あたりの台詞は約 4〜20 秒（約 1〜6 発話程度）である．また，バラエティの場合，登場人物が混在している場面が多く，約 30〜120 秒（約 10〜40 発話程度）は発話者情報が付加されている．また，主役の人物の字幕には色づけがされており，この色情報を使うことでも主役級の発話者の台詞は特定可能である．しかし，それでも全体の 5 割程度の台詞しか話者情報を得られることが出来ない．そのため，この 5 割の発話を話者識別モデルの構築に用いる学習データとして活用する．本研究ではこのモデルを用いて残り 5 割の発話者不明な台詞に関して発話者情報を推定し，映像コンテンツへ付加することで，シーン検索などが可能になるメタデータの作成を行う． まず，テレビ字幕は音声の開始時間と終了時間の情報を持っている．このデータを用いることで各話者の発話部分を切り出せる．また，字幕に発話者情報が付加されている台詞は，教師データとしてモデルの学習に使われる．発話者情報が付加されていない台詞に関しては，話者識別によって発

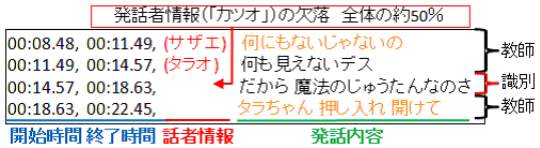

図 1　字幕情報の内容

話者情報が付加される．また，テレビ字幕の内容は，話者を識別するモデル構築に役立つ．音声の発話内容はテレビ字幕によってわかっているため，音声は音素単位で分析できる．本研究はテレビ字幕によって作られた音素モデルによって，テレビ放送の話者を識別する．

4. 字幕情報を用いた話者識別

本研究では字幕情報を用いて 4 段階の処理を行う (図 2)．最初に識別モデル構築のための前処理を行う．まず録画したテレビ番組から音声と字幕を抽出する．抽出された音声は雑音除去と有声区間検出処理によって 1 台詞単位に分けられる．また，字幕は発話内容の形態素解析を行い音素情報を抽出する．これらの前処理の結果を用いて識別モデルの構築を行う．本研究では音素単位のモデルを構築するため字幕の形態素解析結果を用いて音声の各音素区間をアライメント解析する．音素単位モデルはアライメント結果の音素区間を用いて音素ごとに学習を行う．次に構築された識別モデルを用いて話者識別を行う．すべての音素モデルが識別に有効とは限らないため本研究では BIC を用いて識別に有効なモデルのみを選択して識別を行う．最後に識別結果へ発話の傾向を考慮した事前確率による重みづけを行う．発話傾向を考慮することで学習データ量が少なくモデルの性能が不十分な話者の識別精度を補う．

5. 識別モデル構築のための前処理

5.1 雑 音 除 去

5.1.1 ICC(inter-channel cross-correlation) を用いた雑音除去

本研究の音声データは 1 台詞単位で分析される．ニュース番組のような雑音の少ない音声と違い，多ジャンルの番組では様々な雑音が入っている．これらの雑音によって有声区間の推定ミスや話者の誤識別を起こす可能性がある．そのため，前処理として音声データに雑音低減処理することで識別性能の向上を図る．今回識別に用いる音声がステレオであることを活用し，中央音声の抽出手法を検討する[18]．中央音声の推定にはチャンネル間の相互相

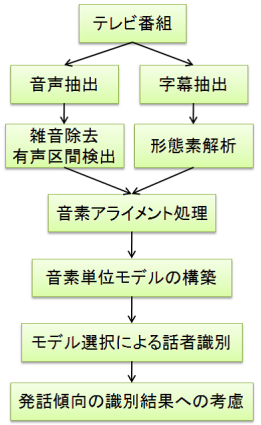

図 2　処理手順

関（ICC）を用いる．この手法は周波数帯域ごとに ICC を用いて，その帯域の角度を推定し，その角度にしたがって帯域の信号を中央 Sc(n) と左右の音に分配する．中央音声の推定計算は式（1）のようになる．

$$Sc(n) = -ICC(n) * f(1/d) * (Slt(n) + Srt(n)) \quad (1)$$

Slt(n) は元のステレオ音源の左チャンネルの n 番目の帯域の信号であり，Srt は右チャンネル，また f(1/d) は距離の逆数に比例するチャンネル分離のための関数である．ICC は式（2，3）によって計算される．

$$|\Gamma(n,k)|^2 = \frac{\Phi_{ij}(n,k)\Phi^*_{ji}(n,k)}{\Phi_{ii}(n,k)\Phi^*_{jj}(n,k)} \quad (2)$$

$$\Phi_{ij}(n,k) = E\{S_{i,n}(k)S^*_{j,n}(k)\} \quad (3)$$

登場人物の台詞は多くの場合，中央に位置している．それに対し，BGM や環境音などは左右の適当な場所に位置していることが多いため左右のチャネルにレベル差が生じる．このレベル差を用いて中央音を抽出することで BGM などの雑音を低減し，音声を強調する．

5.1.2 ICC の性能評価

signal-noise ratio（SN 比）を用いた ICC の性能評価を行った．評価には ICC 処理前と

後の音声データの SN 比を計測し，雑音除去の性能を比較した．評価データにはドラマ，アニメーション，バラエティをそれぞれ 10 番組の計 30 番組の音声データを対象とした．1 番組あたりのデータ量は平均 426 発話である．評価結果は図 3 に示す．

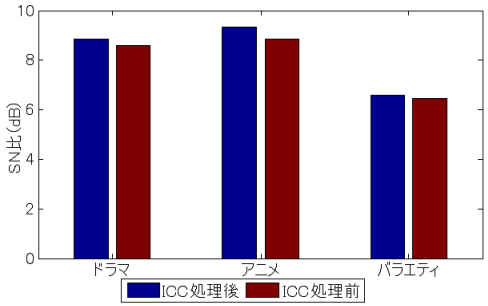

図 3　ICC の性能評価結果

SN 比は ICC 処理によってドラマが 0.27，アニメーションが 0.48，バラエティが 0.15 改善された．また各ジャンルの SN 比はドラマの 8.85，アニメーションの 9.33 と比較してバラエティは 6.61 となっており，雑音が多く含まれていることが分かる．そのため，バラエティ番組は話者識別を行う際に雑音の影響が出てしまう可能性がある．

5.2　有声区間検出

5.2.1　字幕情報を用いた有声区間検出

本研究では雑音除去を前処理として行っているが，それでも音声データにはニュース番組よりも多くの雑音が含まれている．従来手法では識別データを有声区間検出によって推定している．しかし，本研究で扱う番組は雑音が多く含まれているため推定ミスが発生しやすい．そこで有声区間検出の精度を向上させるために 1 台詞部分の抽出には字幕の時間情報を活用する．時間情報は字幕表示の開始時間と終了時間の情報である．しかし，この字幕の時間情報は実際の映像の音声の時間とずれていることや，次の台詞の音声が含まれていることがある．そのため，本研究では一度字幕の時間情報を元に前後 1 秒冗長な区間を残して音声を抽出し，その後有声区間検出を行うことで正確な有声区間を特定している．有声区間検出には STRAIGHT を用いる．番組の音源をそのまま STRAIGHT を用いて有声区間検出を行うと図 4 のような検出結果となる．緑の枠内は検出したい有声区間であり，赤い枠内は雑音区間である．この雑音部分は走っている音がなっているがこのような雑音部分を有声区間と判定してしまうことがある．ニュース番組に比べこのような雑音が多くなる多ジャンルの番組では本来の有性区間を正確に検出することが困難である．そのため，事前に字幕情報から有声区間を抜き出すことでその後の有声区間検出処理の精度を向上させることが出来る．

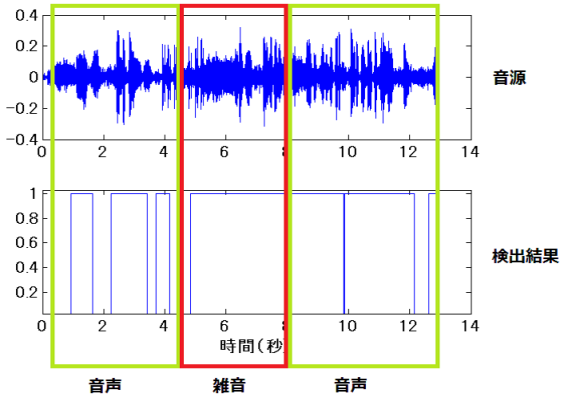

図 4　有声区間検出のみを用いた識別音声の抽出

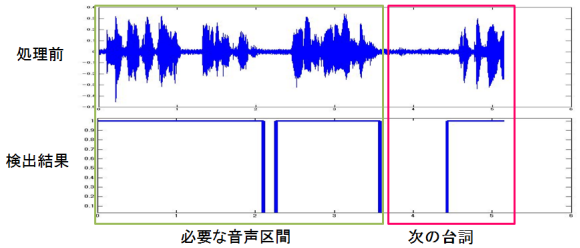

図 5　有声区間検出による識別音声の抽出

5.2.2　有声区間検出の性能評価

有声区間検出の性能を評価するため有声区間検出処理を行ったアニメーションの音声をランダムに 50 個選択し，字幕の発話内容が含まれているか比較することで検出性能評価を行った．検出精度を評価したところ 48 個が正しく検出できており，失敗した 2 個の音声は，

台詞が長かったため字幕の情報が途中で 2 つに分けられていたことが原因でのミスであった．しかし，識別に用いる音声は全て含まれているため大きな問題とはならない．

6. 話者識別モデル

6.1 音素アライメント

本研究では従来の話者単位モデルとは別に音素単位でモデルを構築する．そのため，各話者の音声特徴を音素単位で分析する必要がある．音声データの分析には音素アライメントを用いた．音素アライメントは各台詞の音声区間で使われている音素区間を推定することができる．これにより台詞を音素単位で分割することが可能となる．本研究で使われる識別モデルはこの分析結果の音素情報を用いて構築される．

音素アライメントには音声認識エンジンの Julius と字幕の発話内容を用いて分析する[14]．Julius は，音声認識システムの開発・研究のためのオープンソースの高性能な汎用大語彙連続音声認識エンジンである．Julius は音声認識結果を用いて，入力に対する音素単位のアライメントを実行することができる．その結果，音素ごとの境界フレームと平均音響尤度が出力される．この音素ごとの境界フレームを用いることで音素単位の情報を抜き出している．そして抜き出された各音素データは音素ごとに識別モデルへの学習データとして用いられる．音素アライメントは事前に言語情報を与えることで分析精度が向上するため，字幕の発話内容を活用することでより正確な分析結果を得ることが出来る．

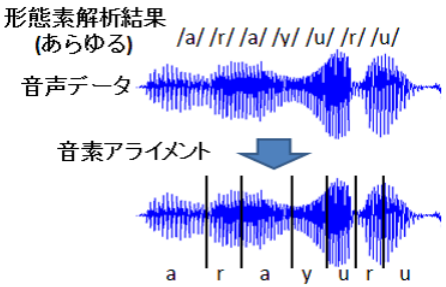

図 6 音素アライメントの例

6.2 音素単位モデル

従来研究では隠れマルコフモデル（HMM）を用いた話者単位の識別モデルが提案されている．話者単位モデルは学習データに各話者の平均的な声の特徴を用いることで構築されている．ニュース番組では識別対象の人数が少ないため，この平均的な声の特徴を用いてもある程度は識別可能である．しかし，ドラマ，アニメーション，バラエティの場合識別対象となる人数が多くなるためより高い識別精度を必要とする．しかし，音素モデルでは 1 音素当たりの学習データ量は少なくなってしまう．本研究で扱うことのできる学習データの量は 1 人当たり 1〜40 台詞（平均 3〜120 秒），平均で 5 台詞（約 20 秒）となり，その 1 台詞内に含まれる音素数は 3〜55 音素，平均で 10 音素となっている．

本研究では字幕情報と音素アライメントによって分析された音素データを学習データとして用いて音素単位で話者ごとに HMM のモデルを構築した．音素単位モデルは音声認識に用いられる音素数とほぼ同程度の 35 種類を話者一人ごとに用意している．また，HMM の状態数は 3 に設定している．音素ごとに話者の識別を行うことでより詳細な話者ごとの特徴の違いを識別可能にする．モデルの構築には隠れマルコフツールキット（HTK）を使用した[13]．HTK は，隠れマルコフモデルを使った音声認識システムを作るためのソフトウェアツールキットである．本研究では HTK を用いて音素単位モデルを構築する．

6.3 音声特徴量

話者モデル学習のため特徴量にはメル周波数ケプストラム（MFCC）12 次+対数パワー 1 次とそれぞれの Δ の計 26 次を用いる．MFCC は人間の聴覚尺度に近い，対数スケールのメル周波数軸上で分析を行う．そのため，人間の声道特徴を表現しており，話者ごとに異なる値を示す．分析単位はフレーム長 25ms，フレームシフト長 10ms となっている．

7. ベイズ情報量規準（BIC）によるモデル選択

7.1 モデル選択

提案した音素単位モデルは 35 種類すべての音素に関して学習を行っている．しかし，学習用の音声データは字幕情報に話者情報が与えられているもののみを用いているため，すべての音素データを均等に得ることが出来ない．例えば，母音/a/，/i/のような音素ならば大量に含まれているため過学習を起こしてしまう可能性がある．また子音/ch/，/hy/のような音素の場合，一つも含まれておらず学習が不十分になってしまうことがある．そのため，学習に使われた音素データ量は各音素によって違っており，学習された話者モデルの中には十分な学習をできていない音素モデルが存在する可能性がある．これらの音素モデルを識別に用いてしまうと識別性能の低下が考えられる．

本研究では，提案手法の 35 種類の音素モデルと従来手法の話者単位モデル（GMM）の

中から識別に有効と思われる識別モデルのみを選択し，話者の識別に用いる．認識モデルの選択手法は，発話継続時間による認識モデル選択[5] や雑音特徴による雑音モデル選択[8] など認識性能の向上のために多くの手法が提案されている[11),15)]．本研究では，BIC 基準に基づいて話者識別に有効なモデルを選択する．

7.2 BIC

BIC は情報量を基準とした確率モデル選択の評価基準のひとつである．n 個のデータ $X = x_1, x_2, ..., x_N$ に関する r 個のモデル候補を $\lambda = \lambda_1, \lambda_2, ..., \lambda_i$ とする．このときモデル λ のパラメータ数を d とすると BIC は次式のように表される．

$$BIC_i = \log P(X|\lambda_i) - \frac{1}{2}(d + \frac{1}{2}d(d+1))\log(N) \quad (4)$$

式（4）の第 2 項はモデルのパラメータ数が増えた場合に増加するペナルティを示している．ペナルティによって過学習を考慮したモデル評価を行っている．これにより BIC の値が大きいほど最適なモデルであると考えられる．本研究ではこの BIC のスコア差の比較によって学習の十分なモデルを選択している．今回の実験データでは平均で 4 音素のモデルが選択されている．

8. 発話傾向による事前確率の追加

8.1 発話の傾向

テレビ放送の話者識別において問題となるのは，登場回数の少ない人物の識別である．ドラマ，アニメーション，バラエティでは，番組内での発話数が 10 回未満となる人物が登場人物の約 5 割を占めることが多く，また話者は発話数が少ないため学習データを多く確保することが出来ない（図 7）．そのため，発話数の少ない話者の識別モデルでは学習が不十分になることが多く，識別結果において一度も識別されないことが多くなってしまう．そのため，本研究では発話の傾向を字幕から事前確率として学習し，発話数の少ない話者に対し重み付けを行うことで，学習データ量の違いに対応させる．

字幕から得られる発話の傾向として，発話数の少ない話者は短い時間の中に発話が集中している．図 8 はアニメーション 1 番組内で登場した人物が発話した回数と発話した時間帯のグラフである．この場合，話者 4 と話者 5 は 1 部のシーンにのみしか出演していないことが分かる．これは，登場回数の少ない人物は 1 シーンにのみ出演していることが多く，そのシーン以外にはほとんど登場しないためである．

また，字幕には初めて登場する人物の台詞には必ず発話者名が付与されている．そのため，発話数の少ない人物はその初めて登場したシーンにのみ出演している可能性が高いと考えられる．実際にアニメーション 20 番組，ドラマ 15 番組，バラエティ 15 番組の合計 50 番組を用いて台詞を発話した人物が次に台詞を発話するまでの回数を調査したところ図 9 の結果が得られた．図 9 では次に発話するまでの間隔が広がるにつれて同じ話者が登場する確率が少なくなっていることが分かる．本研究では，このデータをポアソン分布に近似して用いることで，識別結果に重み付けを行う．この処理によって識別されにくい話者が，初めて発話した付近の台詞では高いスコアを示すことが出来るように調整される．

また，一度発話した人物が次に発話する確率は累積分布にしたがって増加していく．本研

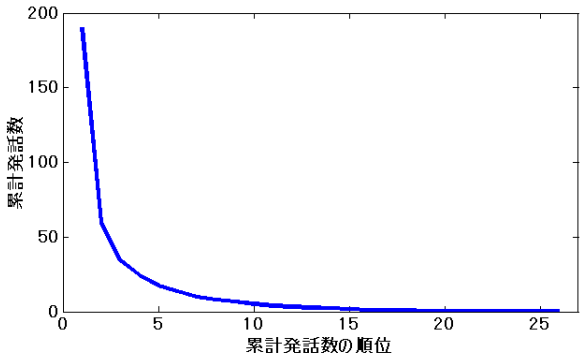

図 7　1 番組内での話者ごとの発話回数

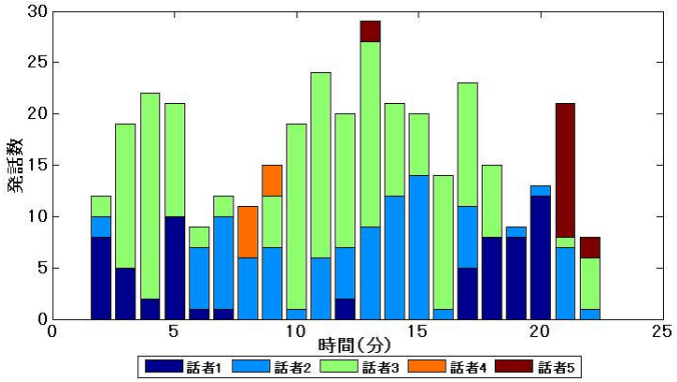

図 8　1 番組内の話者の発話数の分布

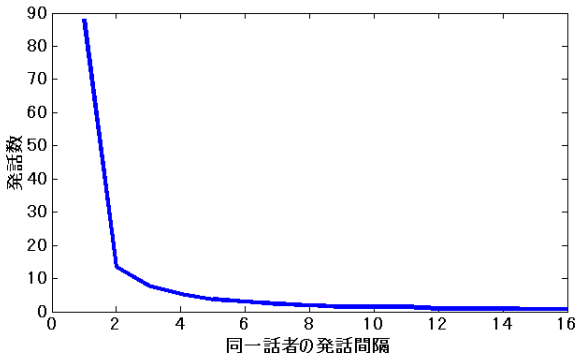

図 9　同一人物が次に発話するまでの回数

究では二つの確率分布を組み合わせて発話傾向を考慮した重み付けを行う（図 10）.

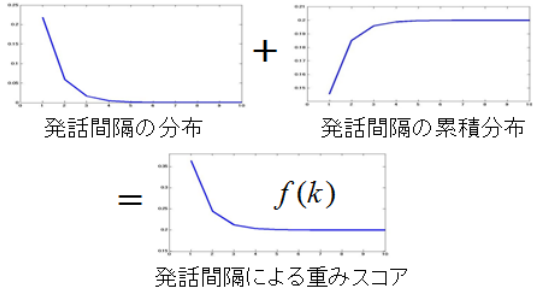

図 10　発話間隔に基づく重み付け関数

8.2　発話内容に基づくの登場人物の傾向

本研究では発話内容に含まれる登場人物の名前の情報から取得できる傾向も話者識別へ考慮する．例として図 1 の場合 4 つ目の台詞に「タラちゃん」と登場人物の名前が含まれている．このような台詞が合った場合この付近でこの名前の人物が登場している可能性が高いと考えられる．そのため，本研究では発話内容に登場人物の名前が含まれていた場合，その人物の尤度を高くなるように重み付けする．台詞に話者の名前が含まれる割合に関して，ドラマ，アニメーション，バラエティの計 30 番組に対して調査した結果，全台詞数の約 6.08%に登場人物の名前が含まれていた．

同時に発話内容を形態素解析して話者の発話特徴を識別結果に反映させている．具体的には学習データの発話の語尾に「です」,「ます」,「まし（た）」の三種類の助動詞が用いられている場合，その人物は丁寧語で発話する可能性があると考えられる．そのため評価データの発話内容が丁寧語である場合，これらの人物の尤度が高くなるように重み付けする．

8.3　発話傾向に基づいた識別尤度の重み付け

発話傾向から求められた発話間隔に基づく分布確率と発話内容を考慮した分布確率は次式のようにして，モデル尤度へ重み付けされる．

$$loglikelihood = (1-\alpha+\beta)\log P(X|S) + \alpha \log f(k) + \beta \log f(l) + \gamma \log f(r) \qquad (5)$$

このとき，P(X | S) は各話者の識別結果の尤度，f は発話傾向の分布関数となる．また，パラメータ k は次の発話までの間隔，パラメータ l は名前が含まれていた台詞からの間隔，パラメータ r は丁寧語の有無，α，β，γ は各項の重み係数である．ドラマ，アニメーション，バラエティの各 10 番組を用いて試験的に評価を行った結果，今回 α は 0.3，β は 0.1 と設定した．

9.　話者識別性能の評価実験

9.1　実 験 概 要

BIC によるモデル選択の有効性を検証するため，モデル選択の有無による性能を比較した．比較には日本のドラマ，アニメーション，バラエティ番組のテレビ放送を録画し，評価データとして用いた．評価データには，ドラマ 10 番組，アニメーション 10 番組，バラエティ10 番組で合計 30 番組を用意した．これらの放送時間はドラマ，バラエティの場合 1 時間，アニメーションの場合 30 分間である．実験ではこの 30 番組の全ての台詞を対象に話者識別を行った．一番組内の台詞は 30 分番組で約 400-600，1 時間番組で約 900-1100 であり，この内字幕情報から推定できる約 5 割の台詞は識別モデルの学習データとして用いる．また残りの約 5 割の台詞は話者識別の識別対象のデータとなる．そのため，識別対象のデータ数は平均ドラマ 254 発話，アニメーション 188 発話，バラエティ229 発話となる．詳細な実験条件は表 1 に示す．

今回の評価では従来の GMM による手法[6] を用いた結果，雑音除去+有効モデル選択によって選ばれたモデルのみを用いて話者識別を行った結果，モデル選択+発話傾向による重み付け処理を用いた話者識別の結果の 3 つの手法に関して比較を行った．また，提案しているそれぞれの手法の単体性能評価も同時に行った．単体性能評価のモデル選択には音素モデ

ルを使い，その他の手法には従来手法と同様の GMM を用いている．

表 1　実験条件

評価番組数	30 番組
平均識別対象人数	ドラマ:16，アニメ:7，バラエティ:13
平均識別発話数	ドラマ:254，アニメ:188，バラエティ:229
識別モデル	3 状態の HMM
音素モデルの種類	/a/,/b/,/by/,/ch/,/d/,/dy/,/e/,/f/,/g/ /gy/,/h/,/hy/,/i/,/j/,/k/,/ky/,/m/,/my/ /n/,/NN/,/ny/,/o/,/p/,/py/,/q/,/r/ /ry/,/s/,/sh/,/t/,/ts/,/u/,/w/,/y/,/z/
サンプリング周波数	16 kHz
フレームシフト長	10 ms
フレーム長	25 ms
音声特徴量	MFCC (1-12) + 対数パワー (1) + Δ (計 26 次元)

9.2 実験結果

実験結果を図 11 に示す．話者識別率は雑音除去+モデル選択+重み付けを行ったものが最も良く，ドラマのとき 60.01%，アニメーションのとき 72.89%，バラエティのとき 70.39%となり，全体で 67.76%となった．この結果は従来手法よりも約 14.14%改善していたが，ドラマに関しての識別は他のものと比較して約 10%ほど低い結果となった．

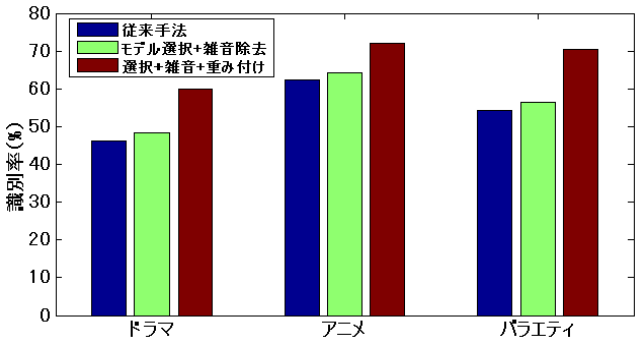

図 11　提案手法の評価結果

また，各手法の単体評価性能を図 12 に示す．従来手法に対して雑音除去と発話傾向の重み付け処理を行うことで性能が向上した．特に発話傾向は識別モデルのみでは性能が低いドラマ，バラエティ番組に対して効果的で，従来のモデルより識別性能が 8.58%向上している．モデル選択に関してはアニメーションでは従来のモデルよりも性能が 2.37%向上しているがドラマ，バラエティ番組に関しては性能が 6.79%低くなってしまっている．また，発話内容の傾向としてはバラエティ番組のみ識別性能が向上した．

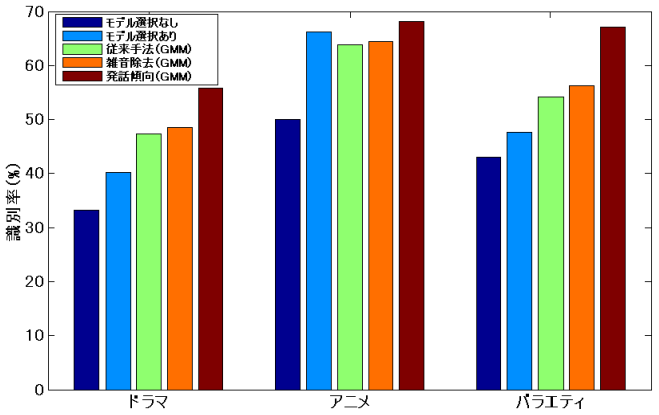

図 12　各手法の比較結果

また，提案手法の識別結果を字幕情報の話者情報と組み合わせた場合の結果を図 13 に示す．字幕情報との組み合わせの結果，全ての台詞に対する正しい話者情報の付加率はドラマ 75.43%，アニメーション 87.48%，バラエティ87.63%となった．この結果から現在 1 番組あたり約 83%の台詞に対して話者情報の付加が可能であると考えられる．

9.3 考　　察

提案手法によって全ての番組において識別率は改善された．3 つのジャンルの中でドラマが極端に識別率の低い結果となったが，これはほとんどのドラマ作品において登場人物が 20 人以上と非常に多くの人物が登場し，また一人当たりの台詞が少なく学習データ量を十分に確保することが出来ないためであると考えられる．同様にバラエティ作品も多くの登場人物が出演しているが，こちらは識別率がドラマに比べ高い結果となっている．これは字幕の話者情報がドラマよりも多く付加されているため，識別モデルが十分な学習を出来ていた

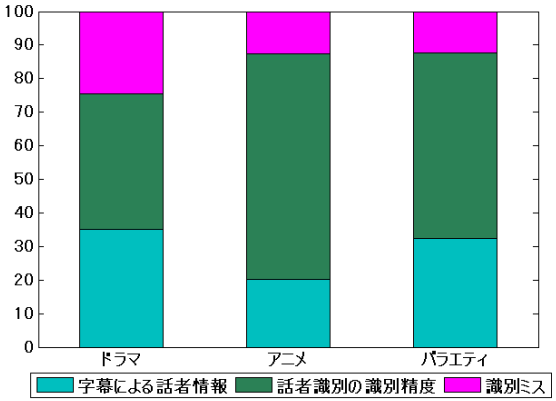

図 13　字幕情報と話者識別を組み合わせた結果

ためと考えられる．学習データ量と識別性能の関係は図 15 のようになっており，70%程度の識別率に達するためには 100 秒前後の台詞を学習データとして用いることが必要であると推測できる．

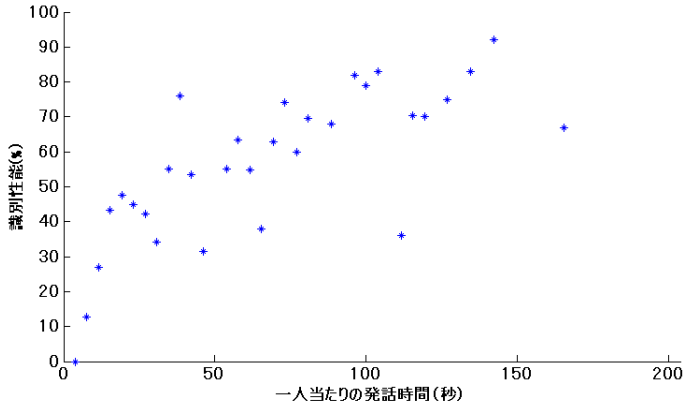

図 14　学習データ量と識別率の関係（発話傾向の考慮なし）

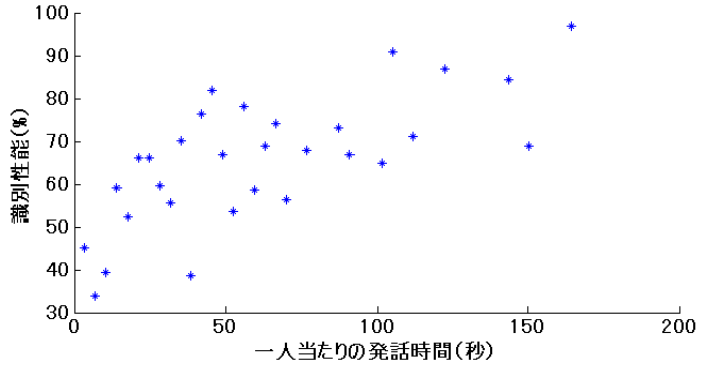

図 15　学習データ量と識別率の関係（発話傾向の考慮あり）

このことから，ドラマの場合は学習データを追加することで識別性能の改善が考えられる．そこで，5 番組を対象に同じドラマ内で放送話数の違う映像から同一人物の音声を学習データとして利用する実験を行った．

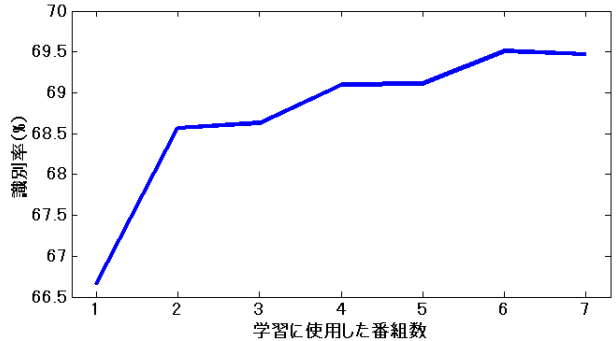

図 16　追加データによる識別結果

1 話単位で学習データを追加した結果，66.66%から 69.47%まで識別性能が改善された．これは一人当たりの学習データ量の平均時間が 40.6 秒から 156.3 秒に増加したことによっ

てモデルの性能が向上したためと考えられる．学習データの追加前と後でモデル選択された音素に違いがあるか比較した結果，データ追加前よりもデータ追加後のほうが子音の音素モデルが多く選択されていた．これは，データ量が少ないとき，子音の音素モデルは十分に話者の特徴を学習できなかったためと考えられる．また，学習データ量が平均 150 秒を超えても識別率が 70%を上回ることはなかった．これは登場する話者すべての学習データ量が増えているのではなく，レギュラーで毎回登場するような人物のみの学習データ量が増えているためであると考えられる．識別性能をより改善させるにはすべての人物の学習データを増加させる必要がある．このことから，すべての人物の学習データ量を確保するため，登場人物が同じ俳優あるいは声優の音声を他番組から学習データとして用いることで識別性能の向上が期待できる．

また，図 14 と図 15 は発話傾向を適応前と適応後のデータ量ごとの識別結果率の関係である．これらの図から発話内容の適応によってデータ量が少ない話者でも識別性能が改善されていることがわかる．

誤識別をしている音声は，BGM に歌声が混ざっていることや発話傾向の事前確率が悪さしている可能性がある．発話傾向は発話数の極端に少ない話者を考慮した手法のため，すべての話者の学習データが十分に確保できる場合，識別誤りの原因となる可能性がある．

また，単体性能評価のモデル選択においてドラマ，バラエティ番組の識別率が悪かった理由としては，学習データの音声状態が悪いものが多かったことが原因と考えられる．特にバラエティで識別率が悪かった話者の学習データは雑音が多く，SN 比は-6 となった．識別率が高い話者の音声は状態が良く，SN 比は 30 ほどとなることから，このような SN 比の悪い音声がモデルの学習に悪影響を与えていると考えられる．また，刑事ドラマでは電話や無線によって加工されているものが存在し，これらのデータも学習に悪影響を与えていた可能性がある．これらの雑音の影響でモデルが音素単位でうまく学習できなかったと考えられる．

発話傾向による重み付けについては，登場人物が 2 人の番組に関しては約 7%識別率が低下していた．この番組ではすべての人物の発話数が多かったため，発話数の少ない人物を考慮した今回の確率分布では補うことが出来ないと思われる．そのため，発話者が一定以上になるまで重み付けを考慮せずに話者識別を行うことで解決できる可能性がある．

発話内容の特徴による重み付けでは，ドラマ，アニメーション番組において識別性能が低下した．この理由としてドラマとアニメーションでは場面と話し相手が頻繁に変わることが原因であると考えられる．バラエティ番組に関しては同じスタジオ内で変化しない出演者と会話することが多いため，話し方の特徴があまり変化しなかったことが識別性能を向上させたと考えられる．

図 13 では字幕から得られる話者情報がアニメーションと比較してドラマが非常に多くなっている．これはドラマの一人当たりの学習データ量がアニメーションよりも多いのではなく，識別対象の人数がドラマのほうが約 2 倍多いためこのような結果となった．

今回の話者識別結果は 67.76%であったが，字幕情報からすでに分かっている話者情報と組み合わせることで，約 83%の台詞に話者情報の付加が可能であった．今回提案したような学習データの少ない例での話者識別結果としては文献[19] がある．この手法では一人当たり約 5 分の会話音声を学習データとして用いており，話者識別を行った識別性能は 86.6%となっている．これに対し提案した手法で用いた学習データ量はドラマ，アニメーションの場合で一人当たり約 20 秒，学習データ量が多く取得できるバラエティ番組でも約 1 分であった．しかし，一人当たりの学習データ量を約 140 秒分確保することができた場合，従来手法と同程度の識別性能を得ることができた．

10. 識別結果の再利用

10.1 識別尤度の高い音声の再利用

話者識別の評価結果から学習データ量を十分に確保する必要がある．そのため，本研究では識別結果を再利用する手法について検討を行った．この手法では評価データを話者識別結果に基づいて学習データとして再利用する．しかし，すべての識別結果が正しい結果を求められているわけではないため，再利用に用いるデータは厳選される必要がある．本研究では話者識別結果は各話者のモデルから得られる尤度を比較して行っているため，最も尤度が高い識別結果を出したモデルと 2 番目の識別結果の尤度差を比較することによってモデルの再学習に用いるデータを選択している．今回，識別結果の第一候補と第二候補の尤度に 0.088 以上の差があった場合，正しい識別結果と判断しモデルの再学習用のデータとして用いる．

10.2 再利用手法の評価

識別結果の再利用手法の性能評価を行った．評価対象のデータにはドラマ，アニメーション，バラエティの各 5 番組，計 15 番組を用いた．評価結果を図 17 に示す．

識別結果の再利用を行った結果，アニメーションは 0.86%，バラエティは 0.36%識別結果が改善され，ドラマに関しては 1.27%低下した．識別性能がほとんど変化しなかった原因は，学習データ量の少ない話者の結果があまり再利用されなかったためだと考えられる．全体の識別性能を向上させるためにはすべての話者の学習データを十分に確保する必要があるが，データ量の少ない話者は識別結果の尤度も低くなってしまうため，再利用の対象に

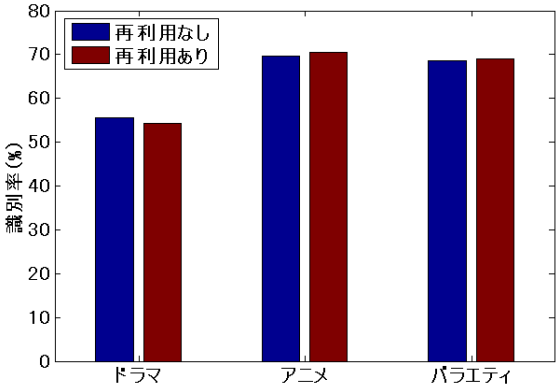

図 17　識別結果の再利用による識別結果

選択されなかった．そのため，識別性能があまり改善しなかったと思われる．また，再利用した音声データの中には間違った話者の音声が含まれていることがあった．これは，発話傾向を考慮した結果，間違った話者でも尤度が高くなってしまい誤推定してしまったことが原因の一つとして考えられる．また，ドラマの場合は話者識別の性能がほかに比べ低いため間違った話者の音声を持ってきてしまい性能が低下したと考えられる．

11. あ と が き

本研究では字幕情報を活用し，BIC によるモデル選択と発話傾向を考慮した重み付けによって映像コンテンツ 30 番組分の話者識別を行った．識別を行った結果，アニメーションのモデル選択+発話傾向による重み付けをした手法の 72.89%の識別率が最も良い結果となり，全体では識別率が 67.76%となった．さらに字幕から取得できる話者情報と組み合わせることで約 83%の話者情報が取得でき，これにより従来手法の 1/5 の学習データで同程度の話者情報がアノテーション可能であった．

今回用いたモデル選択の手法についても再検討する必要があると考えられる．今回の実験ではモデル選択によって一部のモデルのみを識別に利用していたが，その際に選択された音素が全ての話者の識別に関して有効であるとは限らない．そのため，モデル選択によるスコアから全ての音素モデルに重み付けを行い，識別には全ての音素モデルを用いることを検討している．これにより，話者ごとに識別に有効な音素モデルが違っていても全ての音素モデルを用いて検討することが可能になると考えられる．

本研究では識別した話者情報を実際に映像コンテンツへ付加し，利用する方法についても検討していく．映像コンテンツから抽出された話者情報は現在 MPEG-7 形式にしてメタデータとして映像へ付加することを検討している．また，この話者情報を用いて実際にシーン検索をする際にどの程度有効であるかシーン検索システムを構築して利用することを考えている．

参 考 文 献

1) 山田 一郎, 佐野 雅規, “アナウンスコメントを利用したサッカー番組メタデータ自動生成”, 信学会技報, pp. 37-42, 2005.
2) 桑野 秀豪, 松尾 義博, 川添 雄彦, “映像・音声認識, 自然言語処理の適用によるメタデータ生成の作業コスト削減効果に関する考察”, 映情学誌, pp. 842-852, 2007.
3) S. Meignier, D. Moraru, C. Fredouille, J. Bonastre, and L. Besacier, “Step-bystep and integrated approaches in broadcast news speaker diarization”, Computer Speech and Language, Vol. 20, pp. 303-330, 2006.
4) 小坂 哲夫, 赤津 達也, 加藤 正治, 好田 正紀, “音素モデルを用いた話者ベクトルに基づく話者識別”, 信学論, pp. 3201-3209, 2007.
5) 西田 昌史, 河原 達也, “BIC に基づく統計的話者モデル選択による教師なし話者インデキシング”, 信学論, pp. 504-512, 2004.
6) A. Messina, R. Borgotallo, G. Dimino, D. Airola Gnota, and L. Boch, “A Complete System for Automatic News Programme Annotation Based on Multimodal Analysis”, Image Analysis for Multimedia Interactive Services, pp. 219-222, 2008.
7) M. H. Kolekar, K. Palaniappan, and S. Sengupta, “A Novel Framework for Semantic Annotation of Soccer Sports Video Sequences”, IET 5th European Conference on Visual Media Production, pp. 1-9, 2008.
8) 張 志鵬, 古井 貞熙, “頑健な区間検出とモデル適応に基づく雑音下音声認識”, 情報処理, 2004.
9) Chuck Wooters, “The ICSI RT07s Speaker Diarization System”, Proceedings of the Second International Workshop on Classification of Events, Activities, and Relationships, Vol. 46, pp. 509-519, 2007.
10) 関口 一樹, 小杉 信, 向井 信彦, “映像と音情報を用いた野球中継の自動インデクシング”, 映情学誌 , pp. 41-46, 2006.
11) 池田 思朗, “HMM の構造探索による音素モデルの生成”, 信学論, pp. 10-18, 1995.
12) D. A. Reynolds and P. Torres-Carrasquillo, “Approaches and Applications of Audio Diarization”, Proc. ICASSP, Vol. 5, pp. 18-23, 2005.

13) P. C. Woodland, J. J. Odell, V. Valtchev, and S. J. Young, “Large vocabulary continuous speech recognition using HTK”, Proc. ICASSP, Vol. 2, pp. 125-128, 1994.
14) T. Kawahara, A. Lee, K. Takeda, K. Itou, and K. Shikano, “Recent Progress of Open-Source LVCSR Engine Julius and Japanese Model Repository”, Proc. ICSLP, Vol. 4, pp. 3069-3072, 2004.
15) M. Park and Jin-Young Ha, “Model Selection Criterion using Confusion Models for HMM Topology Optimization”, SICE-ICASE, 1004, 2006.
16) R. Turetsky and N. Dimitrova, “Screenplay alignment for closed-system speaker identification and analysis of feature films”, Proc. ICME, Vol. 3, pp. 1659-1662, 2004.
17) E. C. W. Koh, H. Sun and T. L. Nwe, “Speaker Diarization Using Direction of Arrival Estimate and Acoustic Feature Information: The I2R-NTU Submission for the NIST RT 2007 Evaluation”, Lecture Notes in Computer Science, Volume 4625, pp. 484-496, 2008.
18) D. Jang, J. Hong, K. K. Jung, and K. Kang, “Center channel separation based on spatial analysis”, In Proc. 11th Int. conf. Digital Audio Effects, pp. DAFX-08, 2008.
19) W. M. Campbell, J. P. Campbell, D. A. Reynolds, D. A. Jones, and T. R. Leek, “Phonetic Speaker Recognition with Support Vector Machines”, Advances in Neural Information Processing Systems, vol. 16, Morgan Kaufmann, 2004.