6P-8

デジタル放送の字幕情報を用いた発話者のアノテーション*

山室慶太 (法政大学大学院情報科学専攻), 伊藤克亘 (法政大学情報科学部)

1 背景

近年放送のディジタル化とともに，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放送されるようになりつつある．このような状況では視聴者が最も視聴したい映像をすぐに得ることが困難になる．そのため，放送番組や各種情報（メタデータ）を蓄積できるホームサーバーを利用したサーバ型放送サービスが検討されている [1, 2]．このようなサービスにおいてメタデータは，番組検索や番組のハイライトを選択して視聴できるダイジェスト視聴，視聴履歴からお勧め番組を提示するナビゲーション視聴などに利用できる．特にメタデータの中でも番組中のシーン内容に関するものは検索やハイライト編集などに欠かせない情報である．しかし，シーン内容に関するメタデータの作成を手作業で行う場合，膨大な時間と手間が必要となる．多くの番組が放送されている現状でそのような作業は非効率的であり，実施することは困難である．そのためメタデータを効率的に付与する支援システムとして，画像認識，音声認識，自然言語処理技術などを組み合わせた自動付与システムの考案が盛んになってきている [3, 4, 6]．そこで本研究では字幕データと話者識別による発話者情報のアノテーション手法について提案する．

2 字幕データに基づく話者識別

本研究では字幕データの持つ情報に基づき発話音声を話者識別することによって半自動的に番組内の発話者情報に関するアノテーション付けを行う．提案手法では字幕データが発話内容の情報を持つことを活用し，音素単位で話者を識別する．その際に従来のすべての音声に対して話者識別を行う方法ではなく，モデル選択を行い話者の識別に有効であると考えられる音素モデルのみ利用することで識別率の向上を図る．本研究では有効モデルの選択にモデルの分散，尤度，データ数などを利用する．

2.1 メタデータ

メタデータには番組のタイトル，カテゴリ，概要説明，出演者，制作者，放送時間など様々なものが当てはまる．メタデータはこれらの情報から目的の情報にたどり着くための効率的な検索手段を提供することが可能である．従来研究ではサッカー番組のアナウンスコメントを分類することで試合内容のメタデータを生成しているものや [2]，ニュース番組を映像・音声処理，自然言語処理などのメディア処理によって，シーンごとのニューストピックをメタデータとして生成しているものがある [4]．これらはメタデータを生成することでユーザが要求するシーンを検索可能にするための研究である．

本研究では番組の出演者が発話した内容を「誰が，いつ，何を話しているか」といった情報を分析し，メタデータとして情報の付加を目的としている．

2.2 字幕データ

現在，多くのテレビ局が放送番組に字幕データを付与しようとしている．例えば，ゴールデンタイムの時間帯ではどの放送局でも字幕データが付与されており，NHK では生放送番組にもある程度字幕放送に対応している．そのため，今後ほとんどの放送番組で字幕データを活用できると考えられる．

この字幕データには字幕の表示タイミングと表示内容，フォントの種類や色などの情報が含まれている．しかし，字幕データには一部の台詞にしか発話者情報が付与されておらず，多くの台詞は誰の発話かわからないのが現状である．発話者情報はシーン抽出や登場人物の検索に有効な情報なためメタデータの一種として重要な情報であると考えられる．そのため，字幕データの持つ発話内容などの情報を有効活用し，番組内の音声を話者識別によって発話者ごとに分類する．

本研究では字幕データの持つ情報を話者識別に活用している．まず，字幕の表示タイミングには発話の開始時間と終了時間の情報が書き込まれている．このデータを用いることで各話者の発話部分を切り出してくることが可能である．切り出した音声は，発話者がわかる場合は学習データとして，わからない場合はテストデータとして用いられる．このとき発話者の情報は字幕の内容とフォントの色から一部判別可能である．また，台詞内容からは，話者を識別するモデルの構築に利用できる．音声データの発話内容がわかっているため，音素単位での細かなモデルの構築が可能となる．

本研究ではこれらの情報を活用して，発話者のわからない部分の台詞について話者の特定を行っている．

2.3 話者識別モデル

本研究の扱うデータは，発話者は不明だが発話内容は字幕データによって明らかであるという特徴がある．これは発話内容のデータを用いて音素アライメントを取ることで音素単位での分析を可能としている．そのため，本研究では台詞の発話者を音素 HMM による話者識別によって行う．

音素 HMM の構築のためには音声データから音素情報を取り出してくる必要がある．本研究では字幕データと音素アライメントを利用して音素情報を抽出した．まず字幕データから音声データに対応する文章の情報を取り出す．この文章を形態素解析などによって，音素の並びに変換する．この変換した音素の並びを用いて音素アライメントをとることで，音声データから各音素に対応した区間を抽出する．音素 HMM は抽出した各音素の音声区間を用いて音素単位で構築される．今回話者ごとに状態数 3 の音素 HMM を用意し，音素の種類は 35 とした．

* Annotation of indexical information by caption of digital broadcast by Keita Yamamuro. (Graduate School of information science, Hosei University) et. al.

2.4 有効モデルの選択

従来の話者識別ではテスト用データの音声全体を識別している [5]．しかし，発話内容がわかっているのであれば音素アライメントを取ることで音素単位の識別が可能である．そのため，本研究では話者識別に有効であると判断できる音素 HMM とそれに対応する音声データの一部を用いて識別を行う．このとき識別に有効な HMM は音素 HMM の分散や尤度，学習データ数などから，判断しその音素の特徴をうまく学習できていると判断できるモデルのみ，話者の識別に用いる．本研究では 35 種類の音素モデルの中から話者の識別に有効と判断できるモデルを選択して識別に利用している．従来から発話継続時間による認識モデル選択 [7] 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モデル選択手法が提案されている．そこで，本研究では音素モデルの分散，尤度，データ数などに注目して有効モデルの選択を行った．有効度が上位と判断されたもののみを識別に用いることで，すべての音素データを必要としないため従来よりも学習データが少なくてすむ．

3 性能評価

3.1 実験内容

今回有効音素モデルの選択による話者識別性能を評価するため，モデル選択によって選ばれた音素モデルによる話者識別とモデル選択を行わず使えるすべての音素モデルによる話者識別を行い，それぞれの識別性能を比較した．また，有効なモデルを選択する際に上位 1〜6 までのモデルを用いて識別に有効な音素モデルとして用いることが出来るか比較をした．

3.2 実験条件

本研究では，スポーツやニュース番組よりもある程度の登場人物が固定されているためドラマ番組を対象とした．今回評価用データとしてドラマ「獣医ドリトル」の 1 話分（約 50 分）を用いた．この作品には主な登場人物が 9 名登場しており，中心人物の 3 名は字幕データから台詞を判断できるため，今回残りの 6 名の登場人物を識別した．識別対象とする音声は全部で 180 文となる．
今回モデルの学習データとして，話者ごとに 5 発話分の音声データを用いており，分析条件は表 1 に示す．

表 1. 分析条件

対象人数	男性 4 名，女性 2 名（計 6 名）
標本化	16 kHz
量子化	16 bit
フレーム周期	10 ms
フレーム長	25 ms
特徴量	MFCC（1-12），対数パワー（1）＋ Δ（計 26 次元）

3.3 実験結果

表 2 はモデル構築時の分散，尤度とデータ数から判断した上位 6 つの音素モデルによる識別結果と，全音素モデルを利用して識別を行った結果である．女性 1 以外のすべての識別対象者は選択された音素モデルのいずれかで識別率が全音素モデルによる識別率を上回る結果となった．

表 2. モデル選択の有無による識別率（%）

	男性 1	男性 2	男性 3	男性 4	女性 1	女性 2
a	35.2	26.1	4.2	47.8	12.5	22.8
d	0.0	0.0	4.2	36.4	42.9	72.5
e	2.7	0.0	1.1	72.1	52.6	10.5
i	14.3	0.0	12.3	5.6	0.0	86.7
n	21.4	89.5	0.0	8.3	0.0	66.7
s	33.3	0.0	73.2	25.8	0.0	1.9
all	12.0	57.1	31.8	23.1	80.0	28.8

3.4 考察内容

今回の識別結果では最高でも約 90 %の識別率であり，この識別率のままメタデータを番組に付加すると放送の約 1 割は間違った情報を送信してしまうことになる．そのため現在の識別率では発話者の情報をメタデータとして付加するには不十分であると考えられる．しかし，男性 2 の n や，男性 3 の s，女性 2 の i などのいくつかの認識率は全音素モデルによる識別結果よりも大きく上回っていることがわかる．このことから，各発話者に対してそれぞれ識別率の高い音素モデルが存在していると考えることが出来る．このことから例えば男性 2 の情報を知りたい場合，有効モデルの選択によって n の音素モデルを選択することが出来れば男性 2 に関しての識別は高い性能を得られる可能性がある．

4 あとがき

本研究では字幕情報とモデル選択によって放送音声の話者識別を行った．比較評価の結果，モデル選択によって有効モデルを判断することで全モデルを採用するよりも高い性能を得られる可能性がある．また，評価に用いたデータはドラマ 1 話分の音声データであったが，このような連続ものの放送番組では放送を重ねていくことで，学習データは増加していく．そのため，今後はこのような追加データもモデル選択などに利用することで性能を上げていく．

参考文献

[1] 藤澤俊之, 他“ メタデータの規格とサービス例 ”映情学誌 133-157, 2007-02

[2] 馬場 秋継, 他“ サーバー方放送におけるメタデータ利用技術の位置検討 ”IEICE 104(279), 11-16, 2004-08-27

[3] 山田 一郎, 他“ 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 ”IEICE 37-42, 2005-02-18

[4] 桑野秀豪, 他“ 映像・音声認識，自然言語処理の適応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 Vol.61 No.6 pp.842-852 2007

[5] 小坂哲夫，他“ 音素モデルを用いた話者ベクトルに基づく話者識別 ”IEICE J90-D(12), 3201-3209, 2007-12-01

[6] Sylvain Meignier,“ Step-bystep and integrated approaches in broadcast news speaker diarization ”Computer Speech and Language 20(2006) 303-330

[7] 西田 昌史，他“ 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 ”IEICE 504-512, 2004-02-01

[8] 張 志鵬，他“ 頑健な区間検出とモデル適応に基づく雑音下音声認識 ”IEICE 2004-12-14