6U-6

デジタル放送の字幕情報と発話傾向を考慮した発話者アノテーション*

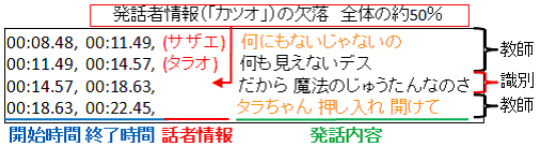
山室慶太 (法政大学大学院情報科学専攻), 伊藤克亘 (法政大学情報科学部)

1 まえがき

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 [1, 2, 3]．その一つとして話者情報のメタデータ化がある．本論文では従来手法でメタデータの自動抽出が行われていないドラマ，アニメーション，バラエティ番組を対象とする．これらの番組の全ての台詞に対して発話者情報を話者識別によってメタデータとして抽出する．そこで本論文では，デジタル放送に付与されているテレビ字幕の情報を活用することで識別モデルの高精度化を行い，同時に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．

2 テレビ字幕の活用

本論文で扱う字幕はデジタル放送に初めから付与されているものである．現在日本のテレビ放送の字幕が必要とされる番組に字幕が付与されている割合は総放送時間の約 69.9%である [5]．そのため，多くの番組で字幕情報を活用できる．テレビ字幕には字幕の表示時間，発話内容，発話者，字幕のフォントやカラーに関する情報が含まれている．特に発話者の情報は初めて登場した人物には必ず付加されている．しかし，それ以降の台詞の場合，映像で話者が特定できない場面でのみしか発話者情報が付加されない．そのためドラマとアニメの場合，発話情報が付加されている台詞は 1 人の話者あたり 1〜5 発話，またバラエティの場合，10〜40 発話程度である．例外として，主役級の人物は字幕のカラーが固有の色のため，これにより話者の識別が可能となる．これらの情報から発話者が分かる台詞は全体の約 50%ほどで，残りの 50%の台詞は話者識別によって判別する必要がある．本研究では字幕から発話者の情報が分かる台詞を学習データとして確率モデルを構築し，これを用いて残りの発話者不明な台詞を話者識別する．そのため，学習に使用するデータ量は従来研究より非常に少ないものとなる．



3 字幕情報を用いた話者識別処理

音声区間検出：本論文では音声データは 1 台詞単位で分析，識別される．1 台詞の抽出には字幕の時間情報を用いる．しかし，字幕の時間は実際の音声とずれていることがあるため，広めのフレームで台詞を抽出し，その後の STRAIGHT を用いた音声区間検出によってより正確な音声データを抽出する．

雑音除去：本論文で扱う音声はステレオ音源である．台詞の多くは中央に位置しており，また BGM など雑音は左右に位置することが多い．そのため，本研究では音声の定位を推定する ICC[7] を用いて発話と雑音を分離する．

音素アライメント：本論文では音素単位で確率モデルの学習を行う．学習には音声特徴を音素単位で分析する必要がある．そのため音素アライメントによって各台詞を音素単位で分割する．アライメントの前処理として字幕の発話内容から形態素解析によって読み仮名を抽出する．アライメントの音素をあらかじめ指定しておくことでより正確なアライメント結果を得ることが出来る．音素アライメントには音声認識エンジン Julius[6] を用いる．

音素単位モデル：本論文では音素アライメントの結果から話者ごとに音素単位の HMM モデルを構築した．モデルは 35 種類の音素モデルを話者ごとに用意し，学習データ量は 1 人当たり 1〜40 台詞，平均で 3 台詞となり，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．音声特徴量にはフレーム長 25ms，フレームシフト長 10ms で分析したメル周波数ケプストラム（MFCC）12 次+対数パワー 1 次+各 Δ の計 26 次を用いた．

モデル選択：本論文で扱う学習データの量は各話者によってばらつきがある．そのため，情報量基準の BIC を用いて識別に有効な音素モデルを選択する．本論文では X:データ，λ:モデル，d:モデルのパラメータ数とし，式（1）で BIC スコアを決定している．BIC スコアによって選択される音素モデルは 3〜5 音素程度となっている．

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

発話傾向による事前確率：登場回数の少ない人物は学習データ量が少なく識別されにくい．そこで，登場回数の少ない人物は偏ったシーンのみに登場する傾向があることを考慮し，一度発話した人物が次に発話するまでの間隔を事前確率として識別結果に重み付けする．発話傾向から求められた発話間隔に基づく分布確率は式（2）によってモデル尤度へ重み付けされる．P(X | S):各話者の識別結果の尤度，f:発話傾向の分布関数，k:次の発話までの間隔，α:各項の重み係数である．今回 α は 0.7 と設定している．

$$\text{尤度} = \alpha P(X|S) + (1-\alpha)f(k) \quad (2)$$

4 評価実験

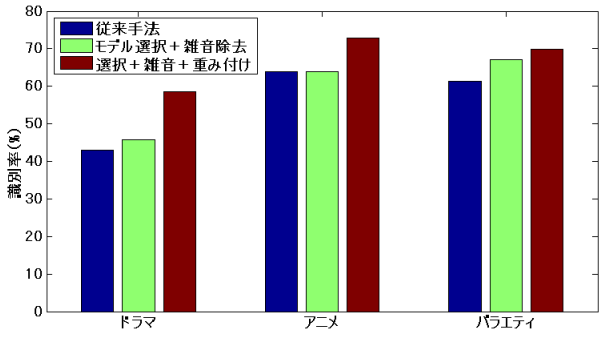
提案手法の有効性を検証するため，性能を従来手法と比較した．比較には日本のドラマ，アニメーション，バラエティ番組のテレビ放送を録画し，評価データとして用いた．評価データにはドラマ 10 番組，アニメー

* Annotation of speaker using closed caption and speech tendency of television broadcasting. by Keita Yamamuro. (Graduate School of information science, Hosei University) et al.

ション 10 番組，バラエティ10 番組の合計 30 番組を用意し，全ての台詞を対象に話者識別を行った．実験条件は以下の表に示す．この評価では従来の GMM による手法 [2] を用いた結果と，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．

評価番組数	30 番組
識別対象人数	12 (1 番組あたりの平均)
発話数	400-600(30 min), 900-1100(1 hr)
識別モデル	3 状態の HMM
音素モデルの種類	35 種類の音素
サンプリング周波数	16 kHz
フレーム長/シフト長	25 ms / 10 ms
音声特徴量	MFCC (1-12) + 対数パワー (1) + Δ (計 26 次元)

実験結果を以下に示す．話者識別率は全ての手法を組み合わせたものが最も良く，ドラマのとき 58.61%，アニメーションのとき 72.89%，バラエティのとき 69.80%となり，全体で 67.10%となった．



5 考察

3 つのジャンルの中でドラマが極端に識別率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 20 人以上の多くの人物が登場し，また一人当たりの学習データを十分に得られなかったためと考えられる．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ドラマよりも性能が良い．こちら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，学習データを十分に確保出来たと考えられる．識別性能は学習データ量に対応しており，今回の実験の場合 70%程度の識別率を得るには 20 前後の台詞を学習データとして用いる必要があった．

このことから，ドラマの場合は学習データの追加による識別性能の改善が考えられる．そこで，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．

学習に使用した番組数	1	2	3	4
識別率（%）	64.31	66.31	66.52	69.01

1 話単位で学習データを追加した結果，64.31%から 69.01%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が 11.6 個から 29.5 個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．このことから，登場人物が同じ俳優あるいは声優ならば，他番組からの音声も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．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，学習データを聴いたところ，話者の音声の状態が悪いものが多かった．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．

6 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．その結果，アニメのモデル選択+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 67.10%となった．しかし，提案手法の結果を従来のニュース番組の話者識別結果と比較した場合，ドラマ番組に関しては提案手法の性能は不十分であった．

今回の識別結果から現在の学習データ量では高精度の識別を行うには不十分である．そのため，今後は識別結果を再利用し，モデルを繰り返し学習させることを検討する．一度識別を行った後，特に尤度の高い識別結果を学習データとして新たに追加することで学習データ量の増加を図る．

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, 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”, 信学技報, 37-42, 2005.

[2] A. Messina et al., “A Complete System for Automatic News Programme Annotation Based on Multimodal Analysis”, Image Analysis for Multimedia Interactive Services, pp. 219-222, 2008.

[3] M. H. Kolekar et al., “A Novel Framework for Semantic Annotation of Soccer Sports Video Sequences”, IET 5th European Conference on Visual Media Production, pp. 1-9, 2008.

[4] 池田, “HMM の構造探索による音素モデルの生成”, 信学論, 10-18, 1995.

[5] “字幕情報の実績” 総務省, 2009.

[6] T. Kawahara et al., “Recent Progress of Open-Source LVCSR Engine Julius and Japanese Model Repository”, Proc. ICSLP, Vol. 4, pp. 3069-3072, 2004.

[7] D. Jang et al., “Center channel separation based on spatial analysis”, In Proc. 11th Int. conf. Digital Audio Effects, pp. DAFX-08, 2008.