

映像/音響ジングルの検出に基づくトピック分割

Topic Segmentation by Detecting Video/Audio Jingles

岩元 浩太†
Kota Iwamoto

大網 亮磨†
Ryoma Oami

1. はじめに

放送番組などの映像を意味的な単位であるトピックに分割することは、映像の閲覧・視聴にとって効果的である。例えば、バラエティ番組や情報番組をコーナー単位に分割すると、見たいコーナーへの頭出しが可能となる。

映像のトピック分割として、音声認識された発話内容に対して、テキストのトピック分割技術を適用する方法がある[1][2]。この方法は、音声認識精度が高く、トピックごと単語分布の差が顕著であるニュース番組など以外には適用が困難である。

また個別の番組に特化して、ルールや機械学習を適用する方法が提案されている。ルールを適用する方法では、トピック分割するためのその番組特有のルールを構築する[3]。機械学習を適用する方法では、番組ごとにトピック分割点の正解データを識別器に学習させる[4][5]。これらルールや機械学習を適用する方法は、個別の番組ごとにルール構築や識別器の学習を行う必要があるため、バラエティ番組・情報番組・教育番組・ドキュメンタリーなどの様々な番組に対して適用するのは現実的ではない。

そこで本稿では、様々な番組に対して汎用的に適用でき、個別の番組ごとの調整が不要な方法として、映像/音響のジングルを検出することによりトピック分割を行う方法を提案する。ここでジングルとは、トピック分割点(トピックの開始・終了・境界)を明示的に示すために挿入される特殊な映像/音響信号のことである。例えば映像ジングルは図1に示すようなコーナーのタイトル画像など、音響ジングルはコーナーの冒頭を告知する効果音・短い音楽などである。ジングルは映像の制作者がトピック分割点を告知するために意図的に挿入したものであるため、正確にジングルを検出することで、ユーザにとって納得性の高いトピック分割ができる。



図1: 映像ジングルの例

2. 提案するジングル検出手法

提案するジングル検出手法は、トピック分割点となるジングルが同じ番組で繰返し使用されることに着目し(例えば、同一シリーズ番組では同じコーナーの冒頭に同じジングルが毎回挿入される)、まず、繰返し発生する映像/音響信号(繰返し信号)を検出する。繰返し信号検出だけでは、ジングル以外の信号を過剰に検出するため(例えば同一カメラで撮影された映像や拍手・笑い声などの音)、次に、真のジングル(トピック分割点)が汎用的に持つ特性(表1)を用い

て判別する。汎用的な特性を用いるため、個別の番組に特化した学習が不要である点が特徴である。

ジングルの検出手順を示す。まず、(1)映像/音響の繰返し信号をジングルの候補点として検出する。次に(2)各候補点からジングルの汎用特性を反映する特徴量を抽出する。最後に(3)SVMに基づいてジングルを判別する。以下それぞれの詳細を示す。

2.1 繰返し信号の検出

映像/音響の繰返し信号は、同一シリーズの番組を対象に、複数の放送回に共通に表れる類似信号をクラスタリングして検出する。映像信号の繰返し検出ではショットを単位とし、ショットの先頭画像の視覚的特徴量[6]をクラスタリングして検出する。音響信号の繰返し検出では、音響フレームのパワースペクトルのサブバンド別の平均パワーを特徴量とし、[7]の照合方式により類似区間をクラスタリングし、繰返し信号を検出する。検出された繰返し信号(の時間位置)がジングルの候補点となる。

2.2 ジングル判別のための汎用特徴量抽出

真のジングルを判別するために、各候補点に対して、ジングルの汎用特性を反映する3つの特徴量(汎用特徴量)を抽出する。この3つの特徴量、およびそれぞれ対応するジングルの汎用特性を表1にまとめる。

表1: ジングル判別のための汎用特徴量

抽出する汎用特徴量	ジングルが持つ汎用特性
シーン・話者の変化度	ジングルの前後で、シーンや話者が変化する
繰返し信号の時間的密集度	繰返し信号が時間的に密集して(パースト的に)発生しない
映像/音響繰返し信号の共起度	映像/音響の繰返し信号が共起することが多い

2.2.1 シーン・話者の変化度

真のジングル(トピック分割点)の前後では、トピックの変化に伴ってシーンが変化したり(スタジオ→ロケーション)、話者が変化したり(司会者→ナレーター)する人が多い。シーン・話者の変化度は、この変化の度合いを表す指標であり、Normalized Cut 値(以後、NCut)[8]を用いる。具体的には、映像の時間軸上の区間をノードとし、ノード間の重み値(類似度)を定義したグラフを構成し、候補点の時間位置でグラフを2分割した場合のNCutを算出して特徴量とする。NCutが小さいほど変化度が大きく、ジングルである確からしが高くなる。

シーンの変化度を表すNCutの算出には、映像のショットをノードとし、ショットの視覚的特徴量の類似度に基づいたノード間の重み値 W を用いる。重み値 W は、

$$W = \exp(-\Delta I^2 / \sigma) \times S \quad (1)$$

で算出する。ここで Δt はノード(ショット)間の時間間隔、 S はショット間の視覚的特徴量[6]の類似度、 σ はスケール係数である。

† NEC 共通基盤ソフトウェア研究所

話者の変化度を表す NCut の算出には、発話セグメントをノードとし、話者クラスタリングで求まる話者クラスタの一致/不一致に基づいたノード間の重み値 W を用いる。具体的には、話者クラスタが一致する場合は $S=1$ 、不一致の場合は $S=0$ とし、式(1)によって重み値 W を算出する。

2.2.2 繰返し信号の時間的密集度

同一カメラで撮影された映像や拍手・笑い声などの音に起因する繰返し信号とは異なり、真のジングルの場合、同一のジングル(同一クラスタの繰返し信号)が時間軸上の短い時間区間に密集することはない(短い時間間隔でトピックが頻繁に変わるにことはない)。繰返し信号の時間的密集度は、繰返し信号が密集して発生する度合いを表す指標であり、候補点に対して、同一クラスタの繰返し信号の時間的な近接度合いで定まるペナルティ値を加算して算出する。あるクラスタに含まれる繰返し信号の数を N 、それぞれの時刻を $\{t_1, \dots, t_i, \dots, t_N\}$ と表すと、 i 番目の繰返し信号(候補点)の時間的密集度 D_i は、次式で算出される。ただし α は定数である。

$$D_i = \sum_{n=1, n \neq i}^N \exp\left\{-\frac{(t_i - t_n)^2}{\alpha}\right\} \quad (2)$$

2.2.3 映像/音響繰返し信号の共起度

真のジングルでは、映像の繰返し信号と音響の繰返し信号が共起する場合が多い。映像/音響繰返し信号の共起度は、この共起の度合いを表す指標である。共起度 C は、候補点(繰返し信号)の前後 s 秒以内に候補点とは別種の繰返し信号が(候補点が映像の繰返し信号の場合は音響の繰返し信号が存在する場合は1、存在しない場合は0とする。

2.3 SVMによるジングル判別

各々の候補点(繰返し信号)に対して、前節で述べた特徴量を用いて SVM によるジングル判別を行う。各候補点の特徴量は、前節で述べた特徴量に(NCUT に関しては2つの σ を使用)に加えて、候補点の繰返し信号が属するクラスタの平均値も含めた計12次元を用いる。SVMの識別関数の出力値が規定の閾値以上(>0)となる場合に、その候補点をジングル、すなわちトピック分割点と判定する。この際、ある時間幅以内にジングルと判定される候補点が複数存在する場合は、識別関数の出力値が最大となる点をジングル(トピック分割点)として採用する。なお、汎用な特徴量を用いているため、SVMの学習は、個別の番組ごとではなく、バラエティ番組・情報番組・教育番組・ドキュメンタリーなどの様々なジャンルの映像をまとめて行う。

3. 評価実験

3.1 実験データ

バラエティ番組・音楽番組・情報番組・教育番組を含む12のシリーズ番組を各3回分、合計36の映像(総時間長1520分)を評価映像とした。トピック分割点となる映像/音響ジングルは合計387箇所であった。なお、意味的にトピックが変化した箇所でも、ジングルにより明示的に示されるものでなければ、正解データに入れていない。

評価は、シリーズ番組を単位とした leave-one-out 交差検定を行った。すなわち、あるシリーズ番組の映像(3回分)に対する精度を求める際には、それ以外のシリーズ番組の映像(33回分)を学習データとした(汎用性を評価するため)。

3.1 実験結果

提案するジングル検出の精度を、適合率=TP/(TP+FP)、と再現率=TP/(TP+FN)で評価した。TP(True Positive)は正解データのジングルのうち検出できた数、FN(False Negative)は検出漏れの数、FP(False Positive)は誤った検出(過剰検出)の数である。なお正解データのジングルの時間位置に対して、前後2秒以内に検出された場合に正解とした。

実験結果を図2に示す。適合率70.7%に対して再現率91.5%を達成している。9割程度の高い再現率を保ちつつ、スタジオシーンなどにおける撮影カメラの切り替えや、拍手・笑い声などにより過剰に検出される繰返し信号の多くを除外できている。ただし、拍手や笑い声が非常に多い一部のバラエティ番組などでは、他の番組と比較して適合率が低い結果となった。

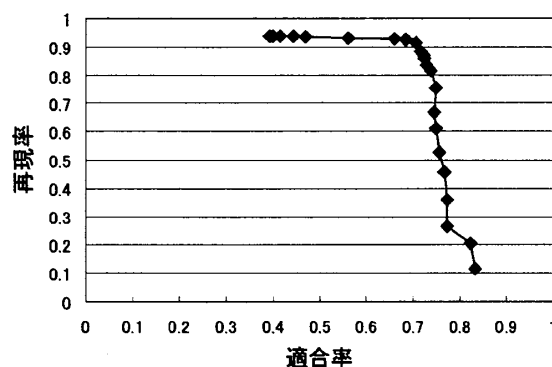


図2: 提案手法によるジングルの検出精度

4. まとめ

トピックの境界を明示的に示す映像/音響ジングルの検出に基づく映像のトピック分割手法を提案した。映像/音響の繰返し信号をジングルの候補として検出し、ジングルが汎用的に持つ特性に基づいて判別する。様々なジャンルの放送番組を用いて提案手法を評価した結果、適合率70.7%に対して再現率91.5%を達成した。

参考文献

- [1] Wu *et al.*, "Fudan University at TRECVID 2003", in Proc. of TRECVID 2003 Workshop, 2003/11.
- [2] M. Sugano *et al.*, "Shot Boundary Determination on MPEG Compressed Domain and Story Segmentation Experiments for TRECVID 2003", in Proc. of TRECVID 2003 Workshop, 2003/11.
- [3] G. M. Quenot *et al.*, "CLIPS-LIS-LSR-LABRI experiments at TRECVID 2004", in Proc. of the TRECVID 2004 Workshop, 2004/11.
- [4] 帆足ほか, "汎用特徴量に基づく動画画像話題分割手法", 信学論D, Vol. J89-D, No. 10, pp.2305-2314, 2006/10.
- [5] A. Amir *et al.*, "IBM Research TRECVID-2004 Video Retrieval System", in Proc. of the TRECVID 2004 Workshop, 2004/11.
- [6] 岩元ほか, "多次元特徴空間解析に基づく映像のカット検出手法", FIT2005 予稿集, 2005/09.
- [7] E. Kasutani *et al.*, "Video Material Archive System for Efficient Video Editing based on Media Identification", IEEE Proc. of ICME2004, Vol. 1, pp.727-730, 2004/06.
- [8] J. Shi *et al.*, "Normalized Cuts and Image Segmentation", IEEE Trans. on PAMI, Vol. 22, No. 8, 2000/08.