社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

ショット内及びショット間の画像・音声特徴に着目したスピーチショット抽出

熊谷　章吾† 　道満　恵介† 　高橋　友和††

出口　大輔††† 　井手　一郎† 　村瀬　洋†

† 名古屋大学 大学院情報科学研究科　〒 464-8601 愛知県名古屋市千種区不老町
†† 岐阜聖徳学園大学 経済情報学部　〒 500-8288 岐阜県岐阜市中鶉 1-38
††† 名古屋大学 情報連携統括本部　〒 464-8601 愛知県名古屋市千種区不老町
E-mail: †{skumagai,kdoman,ttakahashi,ddeguchi,ide,murase}@murase.m.is.nagoya-u.ac.jp

あらまし　本報告では，ショット内及びショット間の特徴に基づく被写体と話者の異同判定によるニュース映像からのスピーチショット抽出手法を提案する．スピーチショットはマルチメディア情報を豊富に含み，資料的価値が高い．そこで我々はこれまで，被写体の口唇動作と話者の声から得られる複数の音声特徴と画像特徴の相関に基づく被写体と話者の異同判定手法を提案してきた．この手法は，音声ノイズの少ないショットに対しては高精度な異同判定が可能であるが，多量の音声ノイズを含むショットに対しての異同判定は困難であった．そこで本報告では，2 段階の処理による被写体と話者の異同判定手法を提案する．まず第 1 段階で，これまでに提案した手法により異同判定を行う．その後，第 2 段階で，ショット内及びその前後のショットとの間に表れる特徴的な画像・音声の性質に基づいて異同判定を行う．スピーチショット抽出実験の結果，提案手法の有効性を確認した．

キーワード　スピーチショット抽出，ニュース映像，映像検索，画像・音声特徴

Extraction of Speech Shots Focusing on Visual and Audio Features within and between Shots

Shogo KUMAGAI†, Keisuke DOMAN†, Tomokazu TAKAHASHI††,

Daisuke DEGUCHI†††, Ichiro IDE†, and Hiroshi MURASE†

† Graduate School of Information Science, Nagoya University, Japan
†† Faculty of Economics and Information, Gifu Shotoku Gakuen University, Japan
††† Information and Communications Headquarters, Nagoya University, Japan
E-mail: †{skumagai,kdoman,ttakahashi,ddeguchi,ide,murase}@murase.m.is.nagoya-u.ac.jp

Abstract　We propose a method to extract speech shots from news videos using detecting the inconsisteny between a subject and the speaker focusing on features within and between shots. Speech shots in news videos contain a wealth of multimedia information, and are valuable as archived material. To extract speech shots, we have previously proposed a method to detect the inconsistency between a subject and the speaker based on the co-occurrence between a subject's lip motion and the speaker's voice. This previous method could detect the inconsistency in a shot with little audio noises. However, it is difficult to detect the inconsistency in a shot with significant amount of audio noises. In order to deal with this problem, the proposed method detects the inconsisteny between a subject and the speaker in two steps. The first step detects the inconsistency by our previous method, and the second step detects the inconsistency based on the intra- and inter- shot features. Experimental results showed the effectiveness of the proposed method.

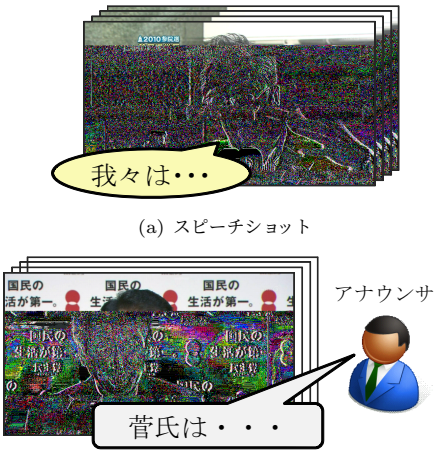
Key words　Speech shot extraction, news video, video retrieval, audio-visual features

1. はじめに

近年，大量にアーカイブされた映像の再利用や効率的な閲覧を支援する技術が必要とされている．さまざまな映像の中でもニュース映像は実世界の出来事に密接に関連しており，資料素材としての価値が高い．ニュース映像においては特に人物に関する情報が重要であり，人物名からの顔画像検索に関する研究 [1,2] や登場人物の人間関係に注目した研究 [3,4] など多くの研究がなされている．その中でも我々は，インタビューや記者会見，選挙演説など，番組関係者以外の人物のスピーチショットに注目している．このようなショットは，話者の表情や態度，声のトーンなど，テキストではわかりにくいマルチメディア情報を豊富に含み，発言集や要約映像の生成などの支援に役立つ [5,6]．また，その抽出に関しては，映像検索ワークショップ TRECVID のタスク [7] としても取り上げられていたこともある．そこで本研究では，ニュース映像からスピーチショットを抽出する技術に注目する．

スピーチショットにおいては，図 1(a) のように人物の顔領域が中央付近に大きく映ることが多いため，抽出の際には顔領域の位置や大きさを利用する方法が考えられる．しかし，顔領域が中央付近に大きく映る映像の中には，図 1(b) のナレーションショットのように被写体と話者が異なるショットも存在する．このショットでは，被写体の発した音声は重畳されておらず，アナウンサなどの番組関係者の発した音声が重畳されている．このように，ニュース映像中には被写体と話者が同一人物であるショットと異なる人物であるショットが存在する．そのためスピーチショットを抽出するためには，被写体と話者の異同判定が必要となる．

関連研究として，堀井らは，口唇動作と音声のタイミング構造に基づき話者を検出する手法を提案した [8]．しかし，この手法は，映像中の複数の人物のうち，発話している人物を特定するためのものであり，映像中に含まれていない人物の発話を想定していない．従って，本研究で扱う問題とは性質が異なる．また，小林らは，口唇動作と音声の共起性に着目した手法を提案している [9]．しかしながら，口唇動作と音声を表す特徴としてそれぞれ単一の画像・音声特徴のみを用いており，判定精度が不十分であった．これに対して本研究では，発生する音声とそれに伴う口唇動作から得られる複数の音声特徴と画像特徴の相関を利用する手法を提案した [10,11]．また，これらの中で，音声ノイズの少ないショットに対する有効性を確認した．しかしながら，記者会見におけるシャッタ音や屋外における騒音などの音声ノイズが含まれるショットに対する異同判定は困難であった．よって，音声ノイズが含まれるショットに対しては，口唇動作と音声の共起性以外に着目した被写体と話者の異同判定が必要である．これに関して，ニュース映像は，情報をより明確に伝えることを目的としており，スピーチショットは，映像中の人物の言動に対して視聴者の注目が集まるように撮影及び編集される．スピーチショットとナレーションショットでは，そのような「見せ方」に関して異なる傾向が存在すると考えられる．そのため，それを基にした確率的な判定が可能であると考えられる．



(a) スピーチショット

(b) ナレーションショット

図 1　ニュース映像の例

そこで本報告では，2 段階の処理による被写体と話者の異同判定手法を提案する．第 1 段階では，これまでに提案した被写体の口唇動作と話者の声との共起に基づく手法 [10,11] により判定を行う．第 2 段階では，ニュース映像中に表れる特異的な性質を利用して確率的な判定を行う．

以降，2. では提案手法について述べる．3. では，提案手法の有効性及び有用性を評価するための実験について述べ，考察する．最後に 4. でまとめる．

2. 提案手法

提案手法における処理の流れを図 2 に示す．提案手法は 2 段階の処理によりフェイスショット（人物の顔領域が大きく映るショット）における被写体と話者の異同判定を行う．まず第 1 段階では，口唇動作と音声の共起に基づき被写体と話者の異同判定を行う．ここでは，発声する音とそれに伴う口唇動作から得られる複数の画像・音声特徴（口の形状や開閉の程度，声の大きさや音素の違い）の相関を基に特徴ベクトルを作成し，それを SVM で識別することで被写体と話者の異同判定を行う．続く第 2 段階では，ショット内及びショット間の特徴に基づき被写体と話者の異同判定を行う．ここでは，スピーチショット内及びその前後のショットとの間に表れる特徴的な画像・音声（ショット内及びショット間の画像的変化や音声ノイズ量など）の傾向を基に特徴ベクトルを作成し，それを SVM で識別することで被写体と話者の異同判定を行う．以上の 2 段階の処理により被写体と話者の異同判定を行う．以降，各段階について順に説明する．

2.1 第 1 段階：口唇動作と音声の共起に基づく被写体と話者の異同判定

被写体と話者が同一であるショットにおいては，被写体の口唇動作と話者の声に高い共起性が見られ，そこから得られる画像特徴と音声特徴に強い相関が表れることが予想される．一方，被写体と話者が異なるショットにおいては，被写体の口唇動作と話者の声は無関係であるため，そこから得られる画像特徴と