Institute of Electronics, Information, and Communication Engineers

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

信学技報
IEICE Technical Report
IE2011-147,MVE2011-109(2012-3)

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

This article is a technical report without peer review, and its polished and/or extended version may be published elsewhere.
Copyright ©2012 by IEICE
NII-Electronic Library Service

Institute of Electronics, Information, and Communication Engineers

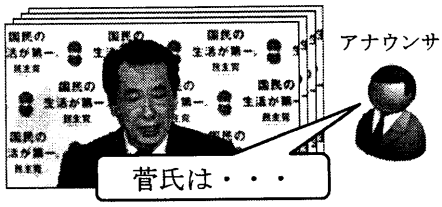
1. はじめに

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

NII-Electronic Library Service

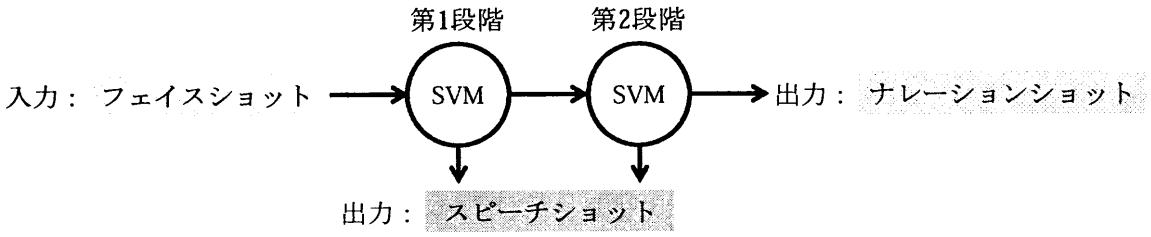


図 2　提案手法における処理の流れ

音声特徴には相関が表れないことが予想される．そこで第 1 段階では，口唇動作と音声から抽出される複数の画像・音声特徴の相関を基に作成した以下の特徴ベクトル $\boldsymbol{c}$ を SVM で識別することにより被写体と話者の異同判定を行う．

$$\boldsymbol{c} = (c_{1,1}, c_{1,2}, \ldots, c_{4,26})^{\mathrm{T}} \tag{1}$$

ここで，$c_{i,j}$ は，画像特徴ベクトル $\boldsymbol{v}_i$（$i = 1, \ldots, 4$）と音声特徴ベクトル $\boldsymbol{a}_j$（$j = 1, \ldots, 26$）の正規化相互相関である．$\boldsymbol{v}_i$, $\boldsymbol{a}_j$ は，フェイスショット（N フレーム）の各フレームから抽出した特徴を時系列順に並べたものであり，以下で表される．

$$\boldsymbol{v}_i = (v_i(1), \ldots, v_i(N))^{\mathrm{T}} \tag{2}$$

$$\boldsymbol{a}_j = (a_j(1), \ldots, a_j(N))^{\mathrm{T}} \tag{3}$$

なお，利用する画像・音声特徴は以下の通りである．

画像特徴

- 口唇領域の縦横比 $v_1(n)$ 及びその動的特徴量 $v_2(n)$
- 口唇領域の面積 $v_3(n)$ 及びその動的特徴量 $v_4(n)$

音声特徴

- 音声信号の平均パワー $a_1(n)$ 及びその動的特徴量 $a_2(n)$
- MFCC（12 次）$a_j(n)$（$j = 3, \ldots, 14$）
  及びその動的特徴量 $a_j(n)$（$j = 15, \ldots, 26$）

2.2　第 2 段階：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

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．そのため，スピーチショットにおいては，ショット内やその前後のショットとの間に，ある程度特徴的な画像・音声の傾向が存在すると考えられる．そこで第 2 段階では，ショット内から抽出した画像・音声特徴 f_{w_1}, f_{w_2}，及びショット間から抽出した画像・音声特徴 f_{b_i}（$i = 1, \ldots, 6$）を基に次式で表される特徴ベクトルを作成し，それを SVM で識別することにより被写体と話者の異同判定を行う．

$$\boldsymbol{f} = (f_{w_1}, f_{w_2}, f_{b_1}, \ldots, f_{b_6})^{\mathrm{T}} \tag{4}$$

以降，ショット自体の性質を表す画像・音声特徴，隣接ショットとの関係性を表す画像・音声特徴について順に述べる．

2.2.1　ショット自体の性質を表す画像・音声特徴

ショット自体の性質を表す画像・音声特徴として，ショット内における画像的変化，ショットに含まれる音声ノイズ量を利用する．以降，各特徴について述べる．

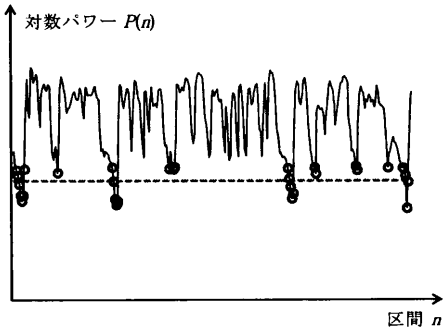


図 3　音声ノイズ量の算出（円：$P_{\mathrm{low}}(m)$（$m = 1, \ldots, \lceil N/10 \rceil$），点線：$f_{w_2}$）

(1) ショット内における画像的変化 f_{w_1}

インタビューや記者会見において，一般に被写体は移動せず一定の場所に留まり発話する．また，被写体の様子を捉えようとするため激しいカメラモーションも少ない．そのため，ショット全体を通して画像的変化は少なくなる．この点に着目し，提案手法ではショットの最初のフレームと最後のフレームの相違度を利用する．

相違度には，画像間の類似性評価指標として一般的に用いられる RGB ヒストグラム間の Bhattacharyya 距離を利用する．具体的には，次式によりショット内における画像的変化 f_{w_1} を求める．

$$f_{w_1} = \sqrt{1 - \sum_i \sqrt{H_f(i) H_l(i)}} \tag{5}$$

ここで，H_f または H_l は，ショットの最初または最後のフレームにおける正規化 RGB ヒストグラムである．

(2) ショットに含まれる音声ノイズ量 f_{w_2}

スピーチショットは静かな環境で撮影されるとは限らず，周囲の人の声や騒音を含むことがある．そのため，ある程度の音声ノイズが含まれることが予想される．一方，被写体と話者が異なる映像においては，主に静かな環境で発話する番組関係者の音声が流れる．そのため，音声ノイズが含まれることは少ない．このことから，提案手法ではショット内の音声ノイズ量を利用する．

具体的にはまず，ショットを N 個の区間に分割し，各区間から音声信号の平均パワーを求め，その対数を取ったものを $P(n)$（$n = 1, \ldots, N$）とする．次に，$P(n)$ の値を降順にソートしてその下位一割を取り出し，これを $P_{\mathrm{low}}(m)$（$m = 1, \ldots, \lceil N/10 \rceil$）（図 3 中の円）とする．この $P_{\mathrm{low}}(m)$ の平均（図 3 中の点線）

NII-Electronic Library Service

をショットに含まれる音声ノイズ量 f_{w_2} として利用する．

$$f_{w_2} = \frac{1}{\lceil N/10 \rceil} \sum_{m=1}^{\lceil N/10 \rceil} P_{\mathrm{low}}(m) \tag{6}$$

2.2.2 隣接ショットとの関係性を表す画像・音声特徴

隣接ショットとの関係性を表す画像・音声特徴として，ショットの切り替わりにおける画像的変化，ショットの切り替わりにおける音量，ショット間での音声ノイズ量の差を利用する．以降，各特徴について述べる．

(1) ショットの切り替わりにおける画像的変化 f_{b_1}, f_{b_2}

ある人物が発話する映像は，複数のスピーチショットの組み合わせによって構成されることが多い．そのため，画像的に似たスピーチショットは連続することが多い．この点に着目し，提案手法では判定対象であるショットの代表フレームとそれに隣接するショットの代表フレームの相違度を利用する．

相違度には，f_{w_1} と同様に，RGB ヒストグラム間の Bhattacharyya 距離を用いる．まず，判定対象であるショットがその直前のショットから切り替わる点における画像的変化 f_{b_1} は次式で計算される．

$$f_{b_1} = \sqrt{1 - \sum_i \sqrt{H_f(i) H_l'(i)}} \tag{7}$$

ここで，H_f または H_l' は，判定対象であるショットの最初のフレームまたはその直前のショットの最後のフレームにおける正規化 RGB ヒストグラムである．次に，判定対象であるショットからその直後のショットへ切り替わる点における画像的変化 f_{b_2} は次式で計算される．

$$f_{b_2} = \sqrt{1 - \sum_i \sqrt{H_l(i) H_f''(i)}} \tag{8}$$

ここで，H_l と H_f'' は，それぞれ判定対象であるショットの最後のフレームとその直後のショットの最初のフレームにおける正規化ヒストグラムである．

(2) ショットの切り替わりにおける音量 f_{b_3}, f_{b_4}

スピーチショットの始端及び終端においては，音量が下がることが多い．これは，ショットを結合して一つの映像を作成した場合に視聴者に違和感を与えないためであると考えられる．一方，被写体と話者が異なる映像においては，画像的な切り替わりに関係なく番組関係者がナレーションを行う場面が多く見られる．そのため，画像的な切り替わりにおいて音量が下がらないことがある．このことから，提案手法ではショットの切り替わりにおける音量を利用する．

具体的には，ショットの切り替わりの前後 1/30 秒，合計 1/15 秒における音声信号の平均パワーを利用する．まず，判定対象であるショットがその直前のショットから切り替わる点における音量 f_{b_3} は次式で計算される．

$$f_{b_3} = \frac{1}{T'} \sum_{t_1' \leq t \leq t_2'} x'^2(t) \tag{9}$$

ここで，t_1' または t_2' はショットが切り替わる時刻の 1/30 秒前または後，$x'(t)$ は時刻 t（$t_1' \leq t \leq t_2'$）の音声出力のサンプリング値，T' は $t_1' \leq t \leq t_2'$ での音声出力のサンプル数である．次に，判定対象であるショットからその直後のショットへ切り替わる点における音量 f_{b_4} は次式で計算される．

$$f_{b_4} = \frac{1}{T''} \sum_{t_1'' \leq t \leq t_2''} x''^2(t) \tag{10}$$

ここで，t_1'' または t_2'' はショットが切り替わる時刻の 1/30 秒前または後，$x''(t)$ は時刻 t（$t_1'' \leq t \leq t_2''$）の音声出力のサンプリング値，T'' は $t_1'' \leq t \leq t_2''$ での音声出力のサンプル数である．

(3) ショット間での音声ノイズ量の差 f_{b_5}, f_{b_6}

スピーチショットには音声ノイズが含まれることが多い．ただし，その量は撮影環境によって様々であるため，絶対量のみで判断するのは困難である．しかしながら，隣接するショットと比較して音声ノイズの量が多い場合には，スピーチショットである可能性が高くなると考えられる．このことから，提案手法では判定対象であるショットと隣接ショットにおける音声ノイズ量の差を利用する．

具体的にはまず，式 6 と同様の方法で，判定対象であるショットにおける音声ノイズ量 f_{w_2} とその直前及び直後のショットにおける音声ノイズ量 f_{w_2}'，f_{w_2}'' を求める．これらを用いて，判定対象であるショットとその直前及び直後のショットにおける音声ノイズ量の差 f_{b_5}，f_{b_6} をそれぞれ次式で求める．

$$f_{b_5} = f_{w_2} - f_{w_2}' \tag{11}$$

$$f_{b_6} = f_{w_2} - f_{w_2}'' \tag{12}$$

表 1 作成したデータセットとその内訳

セット	被写体＝話者	被写体≠話者	合計
1	48	13	61
2	37	25	62
3	49	13	62
4	55	19	74
5	48	18	66
6	41	31	72
7	43	19	62
合計	321	138	459

3. 評価実験

ニュース映像からのスピーチショット抽出に関する提案手法の有用性を評価するための実験とその結果について述べ，考察を加える．

3.1 実験方法

実際に放送されたニュース映像（NHK ニュース 7）7 日分からフェイスショット 459 本を人手で抽出した．その内訳を表 1 に示す．ここで，各セットは 1 日分のニュース映像に対応している．これらのフェイスショットに対して，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手法（提案手法における第 1 段階）を適用し，被写体と話者が同一であると判定されたショットをスピーチショットとして抽出した．なお，口唇領域は，輝度及び色情報を利用したシンプルかつ高速な手法により

NII-Electronic Library Service

表 2　スピーチショットの抽出精度の比較

	適合率	再現率	F 値
第 1 段階のみ	0.949 (56/ 59)	0.174 (56/321)	0.294
第 2 段階のみ	0.879 (270/307)	0.841 (270/321)	0.860
提案手法	0.882 (276/313)	0.860 (276/321)	**0.871**

自動で抽出した．その後，提案手法における第 1 段階で被写体と話者が異なると判定されたショットに対して，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手法（提案手法における第 2 段階）を適用し，被写体と話者が同一であると判定されたショットをスピーチショットとして抽出した．

提案手法における第 2 段階の識別器の学習には，学習セットに対して提案手法における第 1 段階の処理で被写体と話者が異なると判定されたサンプルのみを利用した．このとき，1 つのセットを評価用セット，残りの 6 つのセットを学習用セットとし，各セットにおける抽出精度の平均を全体の抽出精度とした．

評価基準としては，スピーチショットの抽出に関する適合率，再現率，F 値を利用した．適合率が高いほど誤抽出が少ないことを表し，再現率が高いほど抽出漏れが少ないことを表す．適合率と再現率のそれぞれが高ければ抽出性能が優れていることを意味するが，両者はトレードオフの関係にある．そのため，総合的な性能評価には，適合率と再現率の調和平均である F 値を用いた．

3.2 実験結果

実験結果を表 2 に示す．F 値に関して，提案手法における第 1 段階のみでは 0.294，第 2 段階のみでは 0.860，提案手法では 0.871 となった．提案手法における第 1 段階のみ，第 2 段階のみの場合と比べて提案手法のほうが高い F 値が得られたことから，2 段階から構成される提案手法の有効性を確認した．また，適合率に関しては，提案手法における第 1 段階のみでは 0.949，第 2 段階のみでは 0.879，提案手法では 0.882 となり，第 1 段階のみの場合に最も高くなった．

3.3 考　　察

提案手法における 2 段階の識別処理の有効性，第 2 段階で利用する各特徴の有効性について考察を述べる．

2 段階の識別処理の有効性：提案手法における第 1 段階と第 2 段階を組み合わせることでスピーチショットの抽出精度が向上した．これは，異なる性質を用いて異同判定を行ったためである．提案手法における第 1 段階は口唇動作と音声の共起性を基に判定を行う手法であり，画像・音声特徴を正確に抽出できれば高精度なスピーチショット抽出が可能である．しかしながら，画像・音声ノイズが含まれる場合には画像・音声特徴を正確に抽出することができないため，スピーチショットの抽出漏れが生じる．一方，提案手法における第 2 段階は，第 1 段階の判定精度の低下の原因となる音声ノイズなどに表れる特異な画像・音声特徴を利用して判定を行う．しかしながら，あくまで確率的な判定であり，特徴的な画像・音声の傾向が表れない場合には正しく判定できない．提案手法はこのような長所と短所を持つ 2 つの手法を組み合わせて判定を行うものである．これらの手法が双方の短所を長所で補い合ったことにより，性能が向上したと考えられる．

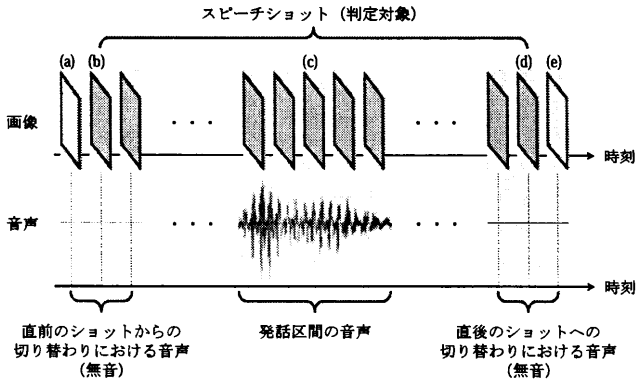


図 4　ショットの切り替わりにおける音量を利用した場合の識別成功例

(a) 直前のショットの最後のフレーム　(b) スピーチショット（判定対象）の最初のフレーム　(c) スピーチショット（判定対象）中のフレーム

(d) スピーチショット（判定対象）の最後のフレーム　(e) 直後のショットの最初のフレーム

図 5　図 4 の (a)〜(e) に対応するフレーム

第 2 段階で利用する各特徴の有効性：第 2 段階で利用する各特徴の有効性を評価するために，被写体と話者の異同判定実験において比較手法との比較を行った．その結果を表 3 に示す．利用する特徴を減らしたいずれの比較手法よりも提案手法の方が高い識別率が得られた．さらに，いずれの特徴を追加した場合でも識別率が向上したことから，個々の特徴には効果があるといえる．実際，各特徴を追加することで正しく識別できるようになったショットが存在し，そこには各特徴の傾向が顕著に表れていた．その中で最も識別率の向上に寄与した特徴は，ショットの切り替わりにおける音量 f_{b_3}，f_{b_4} であった．スピーチショットの切り替わりにおける音量を利用した場合に識別に成功した例を図 4，図 5 に示す．直前のショットからの切り替わりにおける音声波形及び直後のショットへの切り替わりにおける音声波形は，ショット中の発話区間における音声波形と比べて振幅が小さく，音量が小さいことがわかる．提案手法ではこのようなスピーチショットの傾向を表す特徴を利用することで，正しく判定することができたと考えられる．

一方で，提案手法では識別率 0.808，比較手法のうち最も高いものでは識別率 0.797 となり，その差は 0.011 と小さかった．提案手法で算出した特徴によって着目した性質を適切に表現できてはいるものの，より良い算出方法があると考えられる．例

NII-Electronic Library Service

表 3　第 2 段階で利用する各特徴の有効性評価のための比較

手法	ショット自体の性質を表す特徴		隣接ショットとの関係性を表す特徴			識別率
	画像的変化 f_{w_1}	音声ノイズ量 f_{w_2}	画像的変化 f_{b_1}, f_{b_2}	音量 f_{b_3}, f_{b_4}	音声ノイズ量の差 f_{b_5}, f_{b_6}	
比較 A		✓	✓	✓	✓	0.760
比較 B	✓		✓	✓	✓	0.797
比較 C	✓	✓		✓	✓	0.771
比較 D	✓	✓	✓		✓	0.758
比較 E	✓	✓	✓	✓		0.780
提案	✓	✓	✓	✓	✓	0.808

えば，音声ノイズ量は話者が発話していないと思われる区間の音量から算出している．これは，人間の発話においては一定の「間」が存在するという仮定の上に成り立っている．そのため，「間」をおかずに発声し続けている場合には音声ノイズ量を正しく算出することができない．識別精度向上のためには，着目した性質をより正確に捉える特徴の算出方法を検討する必要がある．また，提案手法では，ショット間の性質を表す特徴を算出する際に，前後のショットの内容については考慮していない．顔検出処理により，各ショットがフェイスショットであるかどうかは判定できる．また，第 1 段階における処理により，一部のフェイスショットについてはスピーチショットであることが特定できる．これらの処理結果を第 2 段階で利用することで，更に高精度な判定ができると考えられる．

4. むすび

本稿では，被写体と話者の異同判定を利用したニュース映像からのスピーチショット抽出手法を提案した．提案手法は 2 段階の処理から構成される．第 1 段階は，口唇動作と音声の共起性に着目し，複数の画像特徴と音声特徴の相関を利用して被写体と話者の異同判定を行う手法である．第 2 段階は，ショット自体やそれに隣接するショットとの間に表れる特異的な性質に着目し，ショット内及びショット間から抽出した画像・音声特徴を利用して被写体と話者の異同判定を行う手法である．画像・音声特徴を正確に抽出できれば高精度な判定が可能な第 1 段階と，ショット内及びショット間の画像・音声特徴に基づいて確率的に判定する第 2 段階を組み合わせることで，様々なスピーチショットに対しても高精度な抽出を可能とする．

提案手法によるスピーチショット抽出の精度評価に関する実験では，第 1 段階のみでは F 値 0.294，第 2 段階のみでは F 値 0.860 であったのに対し，提案手法では F 値 0.871 となり，2 段階から構成される提案手法の有効性を確認した．

今後は，第 2 段階におけるショット内及びショット間の特徴に関して，顔検出処理の結果及び第 1 段階における処理の結果を利用した算出方法について検討していく．

謝辞　本研究の一部は科学研究費補助金及び国立情報学研究所との共同研究による．

文　献

[1] S. Satoh, Y. Nakamura, and T. Kanade, "Name-It: Naming and detecting faces in news videos," IEEE Multimedia, vol.6, no.1, pp.22–35, Jan.–Mar. 1999.

[2] D. Ozkan and P. Duygulu, "Finding people frequently appearing in news," Image and Video Retrieval, eds. by H. Sundaram, M. Naphade, J.R. Smith, and Y. Rui, vol.4071, pp.173–182, Lecture Notes in Computer Science, Springer, July 2006.

[3] 小笠原崇，高橋友和，井手一郎，村瀬　洋，"放送映像からの人物相関グラフの構築，" 第 19 回人工知能学会全国大会，no.1F4-02, June 2005.

[4] I. Ide, T. Kinoshita, H. Mo, N. Katayama, and S. Satoh, "trackThem: Exploring a large-scale news video archive by tracking human relations," Information Retrieval Technology, eds. by G.G. Lee, A. Yamada, H. Meng, and S.-H. Myaeng, vol.3689, pp.510–515, Lecture Notes in Computer Science, Springer, Oct. 2005.

[5] 井手一郎，關岡直城，小笠原崇，木下智義，孟　洋，片山紀生，佐藤真一，高橋友和，村瀬　洋，"NewsWho'sWho：ニュース映像アーカイブからの人物情報ポータルの構築，" 第 2 回デジタルコンテンツシンポジウム，no.1-3, June 2006.

[6] 關岡直城，高橋友和，井手一郎，村瀬　洋，"ニュース映像中のモノローグシーン検出による発言集の自動作成，" 電子情報通信学会技術研究報告（PRMU），PRMU2005-301, vol.105, no.674, pp.277–282, March 2006.

[7] A.F. Smeaton, P. Over, and W. Kraaij, "High-level feature detection from video in TRECVid: A 5-year retrospective of achievements," Multimedia Content Analysis, Theory and Applications, ed. by A. Divakaran, pp.151–174, Signals and Communication Technology, Springer, Dec. 2008.

[8] 堀井　悠，川嶋宏彰，松山隆司，"口唇動作と音声のタイミング構造に基づく話者検出，" 第 11 回画像の認識・理解シンポジウム（MIRU）講演論文集，no.OS8-1, pp.193–200, July 2008.

[9] 小林尊志，高橋友和，井手一郎，村瀬　洋，"ニュース映像における話者と被写体の不一致検出，" 第 6 回情報科学技術フォーラム（FIT2007）講演論文集，no.H-081, pp.191–192, Sept. 2007.

[10] 熊谷章吾，道満恵介，高橋友和，出口大輔，井手一郎，村瀬　洋，"口唇動作と音声の共起に着目した被写体と話者の不一致検出～ニュース映像への適用と評価～，" 電子情報通信学会マルチメディア・仮想環境基礎研究会（MVE）技術研究報告，vol.111, no.38, pp.75–80, May 2011.

[11] S. Kumagai, K. Doman, T. Takahashi, D. Deguchi, I. Ide, and H. Murase, "Detection of inconsistency between subject and speaker based on the co-occurrence of lip motion and voice towards speech scene extraction from news videos," Proc. of 2011 IEEE Intl. Symp. on Multimedia (ISM), pp.311–318, Dec. 2011.

NII-Electronic Library Service