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
MVE2011-12(2011-05)

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～

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
E-mail: †{skumagai,kdoman,ttakahashi,ddeguchi,ide,murase}@murase.m.is.nagoya-u.ac.jp

あらまし　ニュース映像中の人物の発言シーンはマルチメディア情報を豊富に含み，資料価値が高い．発言シーンの抽出には顔領域の位置や大きさを利用するアプローチが考えられる．しかし，ナレーションシーンのように被写体と話者が一致していないシーンも存在するため，それだけでは発言シーンを必ずしも抽出できない．そこで我々は，発生する音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して被写体と話者の一致・不一致を識別する手法を提案してきた．しかしながら，理想的な環境で撮影した映像に対する評価のみで，実際に放送されるニュース映像に対する評価にとどまっていた．本稿では，理想的な環境で撮影した映像を用いた実験とその結果，および実際に放送されたニュース映像を用いた実験とその結果について報告する．これら 2 つの実験から，提案手法の有効性および有用性を確認した．

キーワード　発言シーン抽出，視聴覚統合，ニュース映像，口唇動作特徴

Detection of Inconsistency between Face and Speaker Focusing on the Co-occurrence of Lip Motion and Audio

— An Application to News Video and its Evaluation —

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
E-mail: †{skumagai,kdoman,ttakahashi,ddeguchi,ide,murase}@murase.m.is.nagoya-u.ac.jp

Abstract　Speech scenes in news videos contain a wealth of multimedia information, and are valuable as archived material. In order to extract speech scenes from news videos, there is an approach that uses the position and size of a face region. However, it is difficult to extract them with only the approach, since news videos contain scenes where the speakers are not the subjects such as in narration scenes. To solve this problem, we have been proposing a method to detect the inconsistency between face and speaker focusing on the co-occurrence of the lip motion and the speech. However, the evaluations for the proposed method were performed in an ideal condition without much noise. In this paper, we report the investigation on the performance of the proposed method not only with videos captured in ideal conditions but also with actual broadcasted news videos. Their results showed the effectiveness and the usefulness of our method.

Key words　speech scene extraction, auditory-visual integration, news video, lip motion feature

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] や登場人物の人間関係に注目した研究 [2,3] 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発言シーンに注目している．このようなシーンは，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 [3,4]．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 [5] として取り上げられるなど需要は高い．そこで本研究では，ニュース映像から発言シーンを抽出する手法に注目する．

発言シーンでは，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アプローチが考えられる．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) のナレーションシーンのように被写体と話者が一致していないシーンも存在する．このシーンでは，被写体の発した音声は流れておらず，アナウンサなどの番組関係者の音声が流れている．よって，発言シーンのみを抽出するためには，まず顔領域の位置や大きさの情報を用いて候補シーンを抽出し，そこから被写体と話者が一致していないシーンを除去する必要がある．これを解決するために，本研究では，口唇動作と音声の共起性に基づき被写体と話者の一致・不一致を識別することで発言シーンを抽出することを考える．

堀井らは，映像中の話者検出のための手法を提案している [6]．しかし，映像中に必ず話者が存在する状況を想定しており，被写体と話者の一致・不一致を識別することは困難である．また，これまでにも口唇動作と音声の共起性に着目した発言シーンの抽出手法 [7] は提案されているが，それぞれ単一の口唇動作特徴と音声特徴のみを用いており，識別精度が不十分であった．そこで我々は，発生する音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し，被写体と話者の一致・不一致を識別する手法を提案した [8]．しかしながら，理想的な環境で撮影した映像に対する評価にとどまり，実際のニュース映像に対する有用性に関する評価が不十分であった．

そこで本稿では，これまで提案してきた被写体と話者の一致・不一致識別手法を理想的な環境で撮影した映像および実際のニュース映像に適用し，有効性および有用性を評価した結果について報告する．以降，2 節では提案手法について述べる．3 節では，提案手法の有効性およ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 節でまとめる．

2. 提案手法

提案手法は，図 2 に示すように，学習と識別の 2 つの段階からなる．学習段階では，まず学習用のフェイスショット（顔領域を含む映像とそれに対応する音声区間）から特徴抽出し，特徴ベクトルを作成する．次に，全ての学習用フェイスショットから抽出された特徴ベクトルを学習し，識別器を構築する．一方，識別段階では，発言シーン候補であるフェイスショットから学習段階と同様に特徴ベクトルを作成し，構築した識別器を用いて被写体と話者の一致・不一致の識別を行う．以降，特徴抽出，識別器構築，一致・不一致識別の順に説明する．

(a) 被写体と話者が一致するシーン

(b) 被写体と話者が一致しないシーン

図 1　ニュース映像の例

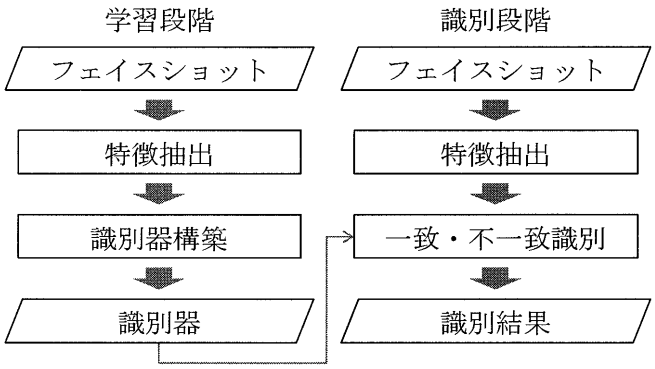


図 2　提案手法における処理の流れ

2.1 特徴抽出

特徴抽出の流れを図 3 に示す．提案手法ではまず，口唇動作特徴 $v_i(n)$ $(i = 1, \ldots, 4)$ と音声特徴 $a_j(n)$ $(j = 1, \ldots, 26)$ をフェイスショット中の n $(n = 1, \ldots, N)$ フレーム目とそれに対応する音声区間から抽出する．次に，抽出した口唇動作特徴と音声特徴の相関を，各組み合わせに対して算出し，それをもとに識別器の入力となる特徴ベクトルを作成する．以降，各処理について詳述する．

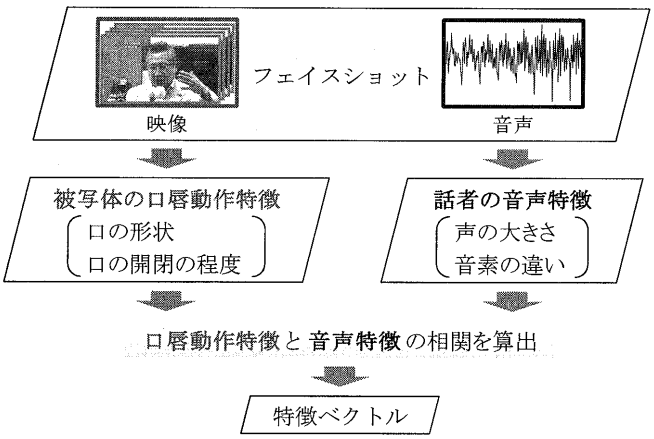
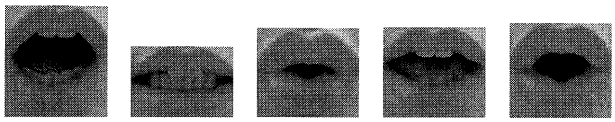


図 3　特徴抽出の流れ



(a)「あ」　(b)「い」　(c)「う」　(d)「え」　(e)「お」

図 4　母音発声時の口唇領域

2.1.1　口唇動作特徴の抽出

発声する音と密接に関連する口唇動作特徴として，口の形状と口の開閉の程度に注目する．具体的には，口の形状は口唇領域の縦横比で，口の開閉の程度は口唇領域の面積で表現する．これらの特徴は，図 4 に示す母音発声時の口唇領域の様子からわかるように，音素ごとに異なる．例えば「あ」は「い」よりも口の形状が縦長になり，口を大きく開いている．また，視覚情報をもとに発話内容の認識を行う読唇 [9] にも利用されており，発声する音声から得られる音声特徴との相関が高いと考えられる．

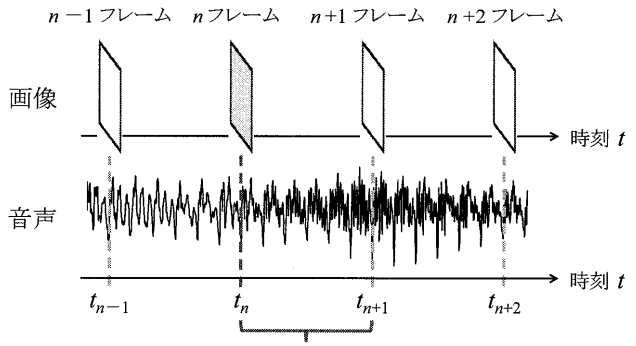
以上から，口唇動作特徴を次のように定義する．

- 口の形状：口唇領域の縦横比 $v_1(n)$ およびその前後フレーム間の変化量 $v_2(n)$
- 口の開閉の程度：口唇領域の面積 $v_3(n)$ およびその前後フレーム間の変化量 $v_4(n)$

なお，口唇領域の抽出手法はこれまでにも数多く提案されている．例えば，Active Appearance Model [10] により高精度な抽出が可能であると報告されており，本研究でもこれらが適用可能であると考えられる．

2.1.2　音声特徴の抽出

発声時の口の動きと密接に関連する音声特徴として，声の大きさを表す特徴と音素の違いを表す特徴に注目する．具体的には，声の大きさは音声信号の平均パワーで，音素の違いはメル周波数ケプストラム係数（MFCC）で表現する．音声信号の平均パワーは，発話の有無を調べる際に有効であるため発話検出 [6] などに用いられる特徴であり，口唇動作の有無を表す画像特徴との相関が高いと考えられる．一方，MFCC は，音声のスペクトルのうち，音素の違いに対応するスペクトル包絡構造を表す代表的な特徴であり，音声認識など音声処理の分野で広く利用される [11]．音素の違いは口唇や声道の形状の違いにより生み出されるため，音素の違いを表すスペクトル包絡から得られる特徴は画像特徴との相関が高いと考えられる．



n フレーム目の画像に対応する音声の分析区間

図 5　音声の分析区間

以上から，音声特徴を次のように定義する．

- 声の大きさ：音声信号の平均パワー $a_1(n)$ およびその前後フレーム間の変化量 $a_2(n)$
- 音素の違い：MFCC（12 次）$a_j(n)$（$j = 3, \ldots, 14$）およびその前後フレーム間の変化量 $a_j(n)$（$j = 15, \ldots, 26$）

なお，映像のフレームレートと音声のサンプリングレートは一般に異なるため，図 5 に示すように，各フレーム間の音声信号を用いて音声特徴を抽出する．

2.1.3　口唇動作特徴と音声特徴の相関の算出

N フレームの映像から抽出された特徴を時系列順に並べ，次のような口唇動作特徴ベクトル $\boldsymbol{v}_i$（$i = 1, \ldots, 4$）と音声特徴ベクトル $\boldsymbol{a}_j$（$j = 1, \ldots, 26$）を作成する．

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

$$\boldsymbol{a}_j = (a_j(1), \ldots, a_j(N))^{\mathrm{T}} \tag{2}$$

次に，$\boldsymbol{v}_i$ と $\boldsymbol{a}_j$ の各組み合わせに対し，次式で正規化相互相関 $c_{i,j}$ を算出する．

$$c_{i,j} = \frac{\sum_{n=1}^{N} \left(v_i(n) - \overline{v_i}\right)\left(a_j(n) - \overline{a_j}\right)}{\sqrt{\sum_{n=1}^{N} \left(v_i(n) - \overline{v_i}\right)^2} \sqrt{\sum_{n=1}^{N} \left(a_j(n) - \overline{a_j}\right)^2}} \tag{3}$$

ここで $\overline{v_i} = \frac{1}{N}\sum_{n=1}^{N} v_i(n)$，$\overline{a_j} = \frac{1}{N}\sum_{n=1}^{N} a_j(n)$ である．これにより得られた $c_{i,j}$ を用いて，次のような 104（$= 4 \times 26$）次元の特徴ベクトル $\boldsymbol{c}$ を作成する．

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

上式で計算される $\boldsymbol{c}$ を，顔領域を含む N フレームの映像を表現する特徴ベクトルとして利用する．

2.2　識別器構築

被写体と話者が一致するフェイスショットと一致しないフェイスショットから作成した特徴ベクトルをもとにサポートベクターマシン（SVM）を構築する．

2.3　一致・不一致識別

入力されたフェイスショットから特徴ベクトルを作成し，学

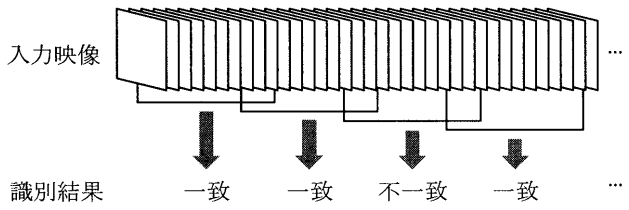


図 6　実 験 概 要

表 1　提案手法と比較手法で用いる口唇動作特徴と音声特徴

	口唇動作特徴		音声特徴	
	口唇領域の縦横比 (v_1,v_2)	口唇領域の面積 (v_3,v_4)	平均パワー (a_1,a_2)	MFCC $(a_3 \sim a_{26})$
提案手法	○	○	○	○
比較手法	○	×	○	×

習段階で構築した SVM により被写体と話者の一致・不一致を識別する．これにより，発言シーンの抽出を行う．

3. 評 価 実 験

提案手法の有効性および有用性を以下の 2 つの実験により調査した．まず，理想的な環境で撮影した映像を用いて，提案手法の有効性を調べた．次に，実際のニュース映像を用いて，提案手法の有用性を調べた．それぞれの実験では，図 6 に示すように，入力映像から一定の長さの区間を抽出し，それぞれの区間に対し提案手法を適用した．なお，本実験では，入力映像中で区間同士の部分的な重なりを許した場合に得られる全ての区間を利用した．評価基準としては，次式に示す識別率を利用した．

$$\text{識別率} = \frac{\text{正識別区間数}}{\text{総区間数}} \tag{5}$$

以降，それぞれの実験について順に述べる．

3.1 実験 1：撮影映像への適用

提案手法で利用する口唇動作特徴と音声特徴，およびそれらの統合方法の有効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．

3.1.1 実 験 方 法

人物 A～人物 J の 10 名にそれぞれ異なるニュース記事（2,000 文字程度）を朗読してもらい，その様子を撮影することで合計 3,481 秒のフェイスショット（1,440×810 pixel）を収集した．なお，識別精度への影響を排除するため，撮影は騒音のない静かな室内で行った．また，口唇領域はテンプレートマッチングなどにより自動抽出した後，人手で修正した．

収集した映像から表 2 に示すような 5 つのデータセットを作成し，5-fold Cross Validation により識別精度を評価した．その際，学習と識別に用いる入力区間の長さを 0.5 秒（$N = 15$）から 10 秒（$N = 300$）まで 0.5 秒ずつ変化させたときの識別率の変化を調べた．

評価に際して，従来手法 [7] に基づく比較手法との比較を行った．提案手法と比較手法の違いは，表 1 に示すように識別に利用する特徴のみである．比較手法は，口唇動作特徴として口唇領域の縦横比（v_1, v_2），音声特徴として音声信号の平均パワー（a_1, a_2）のみを用いる手法である．

表 2　撮影映像から作成したデータセット（被写体／話者）

セット	1	2	3	4	5
被写体＝話者	A / A	C / C	E / E	G / G	I / I
	B / B	D / D	F / F	H / H	J / J
被写体≠話者	A / B	C / D	E / F	G / H	I / J
	B / A	D / C	F / E	H / G	J / I

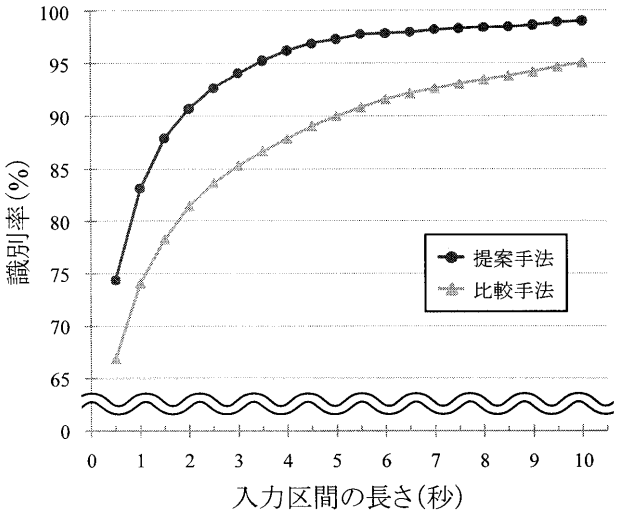


図 7　撮影映像に対する一致・不一致の識別率

3.1.2 実 験 結 果

入力区間の長さの変化に対する識別率の比較を図 7 に示す．入力区間の長さによらず，比較手法より提案手法の方が高い識別率が得られた．これにより，提案手法の有効性を確認した．また，両手法とも入力区間の長さが長くなるにつれて識別率が向上した．両手法とも識別率が最高となったのは入力区間の長さが 10 秒の時で，提案手法では 99.1%，比較手法では 95.1%の識別率が得られた．

3.1.3 考　　察

複数の口唇動作特徴と音声特徴の統合利用の有効性と入力区間の長さと識別率の関係について考察を述べる．

複数の口唇動作特徴と音声特徴の統合利用の有効性：比較手法より提案手法の方が全体を通して高い識別精度が得られた．比較手法と提案手法の違いは，口唇領域の面積および MFCC を利用するかどうかだけである．口唇領域の縦横比では口の形状は表現できるが，口の開閉の程度は表現できない．また，音声の平均パワーでは声の大きさは表現できるが，音素の違いは表現できない．口唇領域の縦横比および平均パワーに加え，口唇領域の面積および MFCC を統合利用することで，口唇動作と音声の共起を捉えやすくなり，識別率が向上したと考えられる．以上のことから，提案手法で利用する口唇動作特徴と音声特徴，およびその統合方法の有効性を確認した．

入力区間の長さと識別率の関係：提案手法と比較手法の両手法において，入力区間の長さが長くなるほど識別率が向上した．これは，映像が長いほど多くの情報を利用できるためであると考えられ，人間の感覚とも一致している．なお，ニュース映像

表 3　ニュース映像中のフェイスショットの撮影環境と種類の内訳

	記者会見	インタビュー	選挙演説
屋内	6	4	0
屋外	0	7	3

表 4　ニュース映像に対する一致・不一致の識別率

	被写体＝話者	被写体≠話者	平均
識別率（%）	53.3	99.7	76.5

への適用を考えた場合，3〜5 秒程度の短い映像に対しても高い精度で識別できることが望ましい．そこで，入力区間の長さが 0.5 秒〜5 秒の場合の識別率に注目すると，比較手法より提案手法の方が 7%〜10%高く，短い映像に対しても高精度に識別できた．このことからも提案手法の有効性を確認した．

3.2　実験 2：ニュース映像への適用

ニュース映像に対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．

3.2.1　実験方法

実際に放送されたニュース映像（NHK ニュース 7）における被写体と話者の一致するフェイスショット（1,440×810 pixel）を 20 本用意した．各映像は長さが 8 秒〜12 秒あり，表 3 に示すように複数の異なる撮影環境と種類を含む．また，抽出した 20 本それぞれの音声を，別のシーンにおけるアナウンサの音声に置き換えることで，被写体と話者の一致しないフェイスショットを 20 本作成した．なお，識別精度への影響を排除するため，口唇領域は人手で切り出した．これらのフェイスショットに対して表 2 のデータセット全てを用いて構築した識別器により被写体と話者の一致・不一致の識別を行った．なお，学習と識別に用いる入力区間の長さは，実験 1 の結果および実際のニュース映像におけるフェイスショットの長さを考慮し，5 秒（$N = 150$）とした．

3.2.2　実験結果

提案手法をニュース映像中のフェイスショットへ適用した結果を表 4 に示す．被写体と話者が一致しているフェイスショットに対しては 53.3%，一致していないフェイスショットに対しては 99.7%，平均で 76.5% の識別率が得られた．

3.2.3　考察

識別失敗の原因，口唇領域の抽出精度と識別精度の関係，提案手法の有用性について考察を述べる．

識別失敗の原因：被写体と話者が一致している映像を不一致と誤識別した映像の中には，大きな騒音が含まれる映像が多かった．このような映像からは話者の声だけでなく周囲の騒音からも音声特徴を抽出してしまうため，口唇動作との共起性を的確に捉えることが難しくなる．これに対する改善策としては，ノイズ除去や音源分離などにより話者の音声のみから音声特徴を抽出することが考えられる．また，音声ノイズにロバストな特徴を利用する方法や音声ノイズを含む映像を学習サンプルとして利用する方法なども有効であると考えられる．

口唇領域の抽出精度と識別精度の関係：本実験では口唇領域を手動で抽出したが，発言シーン抽出の自動化を行うためには自動で抽出する必要がある．しかし，ニュース映像中の発言シーンにおいては，カメラのフラッシュや屋外における影など照明条件が変動しやすく，口唇領域の高精度な抽出は難しい．口唇動作領域を正確に抽出できなければ，音声波形との共起性を上手く捉えることが難しくなる．そのため，高精度な口唇領域の抽出とともに，抽出精度の影響を受けにくい口唇動作特徴の利用や抽出誤差を含む学習サンプルを用いた識別器の構築などによる対応が必要である．

提案手法の有用性：提案手法を例えば発言集の生成に適用する際には，抽出した発言シーンの中に非発言シーンが含まれることは望ましくない．そのため，特に，非発言シーンを不一致と正しく識別することが重要である．この点，提案手法では，被写体と話者が一致している映像に対する識別精度には改善の余地があるものの，被写体と話者が一致していない映像に対しては 99.7%と非常に高い精度で識別できた．このことから，提案手法の有用性を確認した．

4. むすび

本稿では，これまで提案してきた被写体と話者の一致・不一致識別手法を，理想的な環境で撮影した映像を用いた実験に加えて，ニュース映像を用いた実験により評価した結果について報告した．理想的な環境で撮影した映像を用いた実験の結果，被写体と話者の一致・不一致を最大で 99.1%の精度で識別できたことから，提案手法の有効性を確認した．また，ニュース映像を用いた実験の結果，被写体と話者の一致していない映像に対して 99.7%の識別率が得られたことから，提案手法の有用性を確認した．今後の課題は，ノイズ除去や音源分離などによる話者の音声のみからの特徴抽出，高精度な口唇領域の抽出等を検討していく．また，ニュース映像から発言シーンを抽出するため，ニュース番組の構造やクローズドキャプションの利用も検討していく．

謝辞　本研究の一部は科学研究費補助金による．

文　献

[1] D. Ozkan and P. Duygulu: “Finding People Frequently Appearing in News”, Proc. 5th Intl. Conf on Image and Video Retrieval, Lecture Notes in Computer Science, 4071, pp. 173–182, July 2006.

[2] 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, pp. 1–4, June 2005.

[3] 井手一郎, 闕岡直城, 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.

[4] 闕岡直城, 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）, vol. 105, no. 674, pp. 277–282, Mar. 2006.

[5] A. F. Smeaton, P. Over, and W. Kraaij: “High level feature detection from video in TRECVid: A 5-year retrospective of achievements”, in Multimedia Content Analysis, Theory and Applications, (A. Divakaran, ed.), Springer, pp.151–174, 2008.

[6] 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, pp.193–200, July 2008.

[7] 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（FIT）講演論文集, pp.191–192, Sept. 2007.

[8] 熊谷章吾, 道満恵介, 高橋友和, 出口大輔, 井手一郎, 村瀬 洋: “口唇動作特徴と音声特徴の共起性に基づく被写体と話者の不一致検出”, 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.110, no.35, pp.51–52, May 2010.

[9] 齊藤剛史, 小西亮介: “トラジェクトリ特徴量に基づく単語読唇”, 電子情報通信学会論文誌（D）, vol.J90-D, no.4, pp.1105–1114, Apr. 2007.

[10] T.F. Cootes, G.J. Edwards, and C.J. Taylor: “Active Appearance Models”, Proc. the 5th European Conference on Computer Vision（ECCV）, vol. 2, pp.484–498, June 1998.

[11] G. Potamianos and C. Neti: “Audio-Visual Speech Recognition in Challenging Environments”, Proc. the 8th European Conference on Speech Communication and Technology（EUROSPEECH）, pp.1293–1296, Sept. 2003.