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.

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

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
††† 国立情報学研究所　〒 101-8430 東京都千代区一ツ橋 2-1-2
E-mail: †{skumagai,kdoman,ttakahashi,ddeguchi,ide,murase}@murase.m.is.nagoya-u.ac.jp

あらまし　本報告では，ニュース映像中の人物の発言シーンの抽出を目的として，被写体と話者の不一致を検出する手法を提案する．被写体と話者が一致している発言シーンにおいては，被写体の口唇動作と話者の音声波形には高い共起性がみられる．そこで本研究では，口唇動作から得られる画像特徴と音声波形から得られる音声特徴の相関に基づいて作成された特徴ベクトルを用いて被写体と話者の一致・不一致を識別することを考える．実験により，最高で 78.3% の識別率が得られ，本手法の有効性を確認した．

キーワード　視聴覚統合，ニュース映像，発言シーン抽出，正規化相互相関

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

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

Abstract　We propose a method for detection of inconsistency between face and speaker to extract speech scenes in news videos. High co-occurrence of lip motion and audio features is observed in speech scenes where the face matches the speaker. Focusing on this, our method detects inconsistency between face and speaker with feature vectors based on correlations between image features from lip motions and audio features from speech waveform. We obtained up to 78.3% detection accuracy in our experiments, which showed the effectiveness of our method.

Key words　auditory-visual integration, news video, speech scene extraction, normalized cross correlation

1. はじめに

ニュース映像中の番組関係者以外の人物の発言シーンは，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 [1] や要約映像の生成などの支援に役立つ．発言シーンにおいては話者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，これらの情報のみの利用では，被写体以外の人物，主にアナウンサなど番組関係者の声が流れるシーンの誤抽出を防ぐことは困難である．これを解決するために，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手法が提案されている [2]．ただしこの手法では，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る．そこで本研究では，発生する音とそれに伴う口唇形状から得られる複数特徴の統合利用により，発言シーン抽出における被写体と話者の一致・不一致の高精度な識別を目指す．

2. 提案手法

まず，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）を入力として，そこから複数の口唇動作特徴と音声特徴を抽出する．その後，口唇動作特徴と音声特徴の相関に基づく特徴ベクトルをもとに SVM により被写体と話者の一致・不一致の識別を行う．

2.1 口唇動作特徴および音声特徴の抽出

提案手法では，フェイスショット中の n（$n = 1, \ldots, N$）フレーム目から，口唇動作特徴 $v_i(n)$（$i = 1, \ldots, 4$）および音声特徴 $a_j(n)$（$j = 1, \ldots, 22$）をそれぞれ以下のように抽出する．

2.1.1 口唇動作特徴の抽出

口唇動作特徴として口の形状を表す特徴と口の開閉の程度を表す特徴を利用する．口の形状を表す特徴は口唇領域の縦横比 $v_1(n)$ と，その前後フレーム間の変化量 $v_2(n)$ とする．口の開閉の程度を表す特徴は口内領域の面積 $v_3(n)$ と，その前後フレーム間の変化量 $v_4(n)$ とする．

2.1.2 音声特徴の抽出

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を利用する．声の大きさを表す特徴は音声信号の平均パワー $a_1(n)$ と，その前後フレーム間の変化量 $a_2(n)$ とする．音素の違いを表す特徴は 10 次の線形予測係数 $a_j(n)$（$j = 3, \ldots, 12$）と，その前後フレーム間の変化量 $a_j(n)$（$j = 13, \ldots, 22$）とする．

2.2 特徴ベクトルの作成

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 $\boldsymbol{a}_j$（$j = 1, \ldots, 22$）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，正規化相互相関 $c_{i,j}$ を算出する．これにより得られた $c_{i,j}$ を用いて，次のような 88（$= 4 \times 22$）次元の特徴ベクトル $\boldsymbol{c}$ を作成する．

$$\boldsymbol{c} = (c_{1,1}, c_{1,2}, \ldots, c_{4,22})^{\mathrm{T}} \tag{3}$$

提案手法では，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，上式で計算される $\boldsymbol{c}$ を利用する．

2.3 SVM による学習・識別

学習段階では，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 SVM の学習を行う．識別段階では，学習した SVM を用いて被写体と話者の一致・不一致を識別する．

3. 評価実験

従来手法 [2] をベースとした比較手法と提案手法の被写体と話者の一致・不一致識別精度を比較した．なお，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．

3.1 実験方法

10 名（人物 A〜J）にそれぞれ異なるニュース記事（2000 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 3,121 秒のフェイスショットを収集した．これらの映像と音声を用いて表 1 に示すような 5 つの学習セットを作成し，5-fold Cross Validation により提案手法の識別精度を評価した．なお，識別精度への影響を排除するため，口唇領域および口内領域は人手で切り出した．

表 1　学習セット（被写体／話者）

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

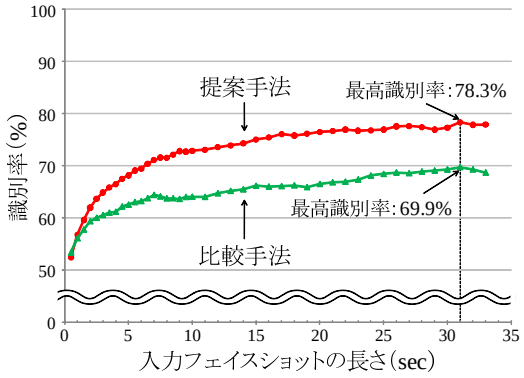

図 1　入力フェイスショットの長さの変化に対する識別率の比較

表 2　提案手法における各セットの識別率（最高識別率時）

	セット 1	セット 2	セット 3	セット 4	セット 5	平均
識別率	67.1%	56.8%	90.8%	92.1%	84.7%	78.3%

3.2 実験結果

図 1 に入力フェイスショットの長さの変化に対する識別率の比較を示す．全体的に比較手法より提案手法の方が高い識別率を示し，提案手法では最大で 78.3%，比較手法では最大で 69.9% の識別精度が得られた．また，提案手法において最高識別率を示したときの各セットの識別率を表 2 に示す．この中で，セット 2 に対する識別率は 56.8% であり，他のセットに対する識別率と比べて低かった．これに関して，セット 2 における人物 B は口の形状や開閉の程度の変化が他の人物と比べて少なく，このような個人差が識別率低下の要因の一つであると考えられる．

4. むすび

本報告では，発言シーン抽出のための被写体と話者の不一致検出手法を提案した．提案手法は，発声する音とそれに伴う口唇形状の共起性に着目し，口唇動作特徴と音声特徴における時系列特徴量の相関をもとに被写体と話者の一致・不一致を識別するものである．評価実験により最高で 78.3% の識別率が得られ，提案手法の有効性が確認できた．今後は，実際のニュース映像への適用および個人差に影響を受けにくい特徴の検討を行う．

謝辞　本研究の一部は文部科学省科学研究費補助金による．

文　献

[1] 關岡 他: “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”, PRMU2005-301, pp.277–282, Mar. 2006.
[2] 小林 他: “ニュース映像における話者と被写体の不一致検出”, FIT2007 講演論文集, pp.191–192, Sep. 2007.