スマートフォンで撮影された発話動画識別の検討

雫, 優太 / SHIZUKU, Yuta

(出版者 / Publisher)
法政大学大学院理工学研究科
(雑誌名 / Journal or Publication Title)
法政大学大学院紀要. 理工学・工学研究科編
(巻 / Volume)
61
(開始ページ / Start Page)
1
(終了ページ / End Page)
4
(発行年 / Year)
2020-03-24
(URL)
https://doi.org/10.15002/00022888

スマートフォンで撮影された発話動画識別の検討

STUDY ON IDENTIFICATION OF SPOKEN VIDEO TAKEN BY SMARTPHONE

雫　優太
Yuta SHIZUKU
指導教員　赤松茂

法政大学大学院理工学研究科応用情報工学専攻修士課程

In this paper, we examined the identification of spoken movie containing camera shakes captured by smartphones. The motion at each pixel of the time-series image was obtained by extracting the optical flow. From the optical flow around the mouth, the optical flow obtained in the region where relatively little movement is expected during utterance was subtracted, and identification was performed using Support Vector Machine. As a basic performance assessment of the speech word recognition, using a 6 word of "a-ri-ga-to", "o-me-de-to", "ko-n-ni-chi-wa", "ko-n-ba-n-wa" ,"sa-yo-na-ra" "su-mi-ma-se-n".
KeyWords:　*Spoken Word*，*Lip Reading*，*Optical Flow*，*Support Vector Machine*

１．はじめに

近年，音声認識技術の発達により，携帯電話での音声によるインターフェイスや音声認識による家電の操作などさまざまな音声認識技術が実用化されている．しかし，現在の音声認識技術には，実環境など雑音環境下で頑健に音声認識を行うことが難しいため，その解決手法の一つとして，唇動画像を用いた画像認識が注目され，研究が進められている[1]．

従来の画像読唇研究[1]では，室内での固定カメラで撮影されたデータを対象に解析等が行われていた．しかしながら，実際日常生活で撮影される動画は，スマートフォンなどで撮影されることが多く[2]，手振れなどのノイズが入ったものが主となる．そのため，本研究では読唇技術を実用的に利用することを目的として，手振れ等のノイズが入った動画像に対する識別率向上の検討を行う．

２．実験方法

（１）実験概要

本実験では，スマートデバイスで撮影された発話動画の識別精度向上を目的とした．初めに，スマートデバイスで発話動画を撮影した場合，識別にどれほど影響を与えるのかを示した．その後，手振れの軽減を行ったオプティカルフローを特徴量として，K-Nearest-Neighbor 法と SupportVectorMachine を用いて識別を行った．識別実験の手順概要を図 1，図 2 に示す．

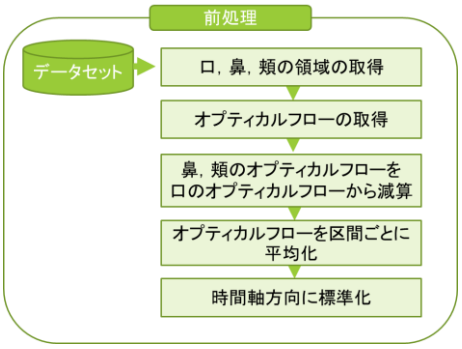

図 1　前処理概要

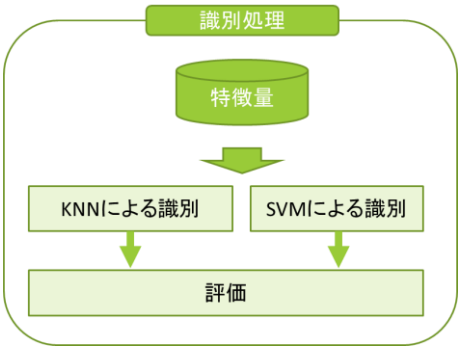

図 2　識別処理概要

（２）手振れが発話動画の識別に与える影響の検証

手振れが含まれる発話動画が，オプティカルフローに対してどれほどの影響があるのか検証するために予備実験を行った．

撮影は，スマートフォンである iphone10 を用い，以下の(A)，(B)の条件で撮影を行った．撮影時のフレームレートは 30fps とした．

(A) 安定した土台にカメラをセットし，発話動画の撮影を行う．
(B) カメラを持った状態で，何もせずに顔の撮影を行う．

(A), (B)で発生するオプティカルフローを比較することで，手振れによる影響を検証した．また，(A)の発話時には”ありがとう”，”おめでとう”，”こんにちは“，”こんばんは“，”さようなら“，”すみません“の 6 単語で行い，被験者 5 名の平均を取った．

（３）識別に用いたデータセット

識別を行うために，データセットとして，SSSD[3][4]を用いた．SSSD はスマートデバイスで撮影された男女 36 名の発話動画が含まれる読唇研究向けの公開データベースである．発話単語は”ゼロ”，“イチ”などの数字や，“ありがとう”，“おめでとう”などの日常会話単語が含まれ，それぞれ 30 サンプルが含まれている．本研究では日常会話でよく使われることと，発話時間の影響を最小にすることを考え”ありがとう”，”おめでとう”，”こんにちは“，”こんばんは“，”さようなら“，”すみません“の 6 単語を選択した．顔画像の例を図 3 に示す．

図 3　顔画像例

（４）領域の取得

発話時のオプティカルフローを検出するために口領域の取得を行った．加えて，手振れによって発生するオプティカルフローを検出するために，発話時に影響が少ないと考えられる領域の取得を行った．領域を図 4 に示す．

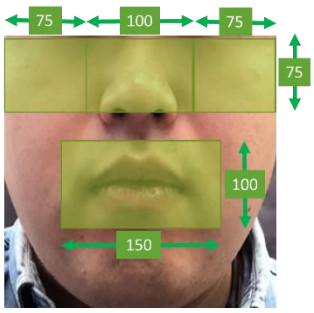

図 4　取得した領域

領域の取得には SSSD に含まれる顔の特徴点を用いた．特徴点の座標に基づき，口領域を 150pixel×100pixel，鼻領域を 100pixel×75pixel，頬領域を 75pixel×75pixel で取得した．提供される特徴点は OpenPose によって取得されている．

（５）オプティカルフローの取得

a）オプティカルフローとは

オプティカルフローとは，デジタル画像内の物体の動きをベクトルで表したもので，動画像にいくつかの仮定を置き画素の移動値を検出する方法である．

b）取得方法

(4)で得られた領域に対して，フレーム間のオプティカルフローを計算した．オプティカルフローの計算には空間的局所最適化法の 1 つである Lucas-Kanade 法を用いた．得られたオプティカルフローの例を図 5 に示す．

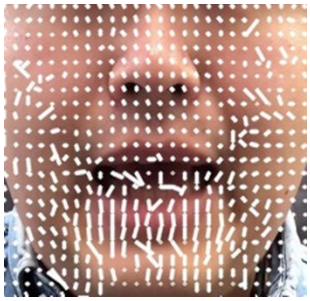

図 5　発話時のオプティカルフロー

（６）手振れの減算

手振れを検出するために，鼻，頬領域で得られたオプティカルフローを用いた．それぞれの領域で得られたオプティカルフローの最頻値を手振れによるオプティカルフローとして口領域のオプティカルフローから減算した．

（７）区間ごとの平均化

オプティカルフローを抽出する際，場所によっては正確に検出できない場合もある．そういった，誤検出に対しロバストにするため，本研究では口画像を 8×8 の領域に分け，それぞれの領域の平均速度ベクトルを特徴量として使用した．口画像の領域分けを図 6 に示す．

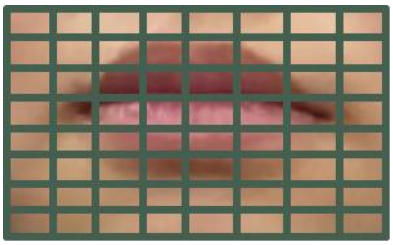

図 6　平均ベクトル算出のための領域分け

（８）時間軸の標準化

（７）で得られた特徴ベクトルは，各フレームで計算され，発話動画全体のフレームを n フレームとしたときに n-1 個の特徴ベクトルが得られる．発話の時間は発

話者によって異なるため，得られる特徴ベクトルの数も異なる．そのため，その差異に対してロバストな特徴量とするために，時間軸の標準化を行い特徴ベクトルを 10 区間に分けた．時間軸標準化のイメージを図 7 に示す．

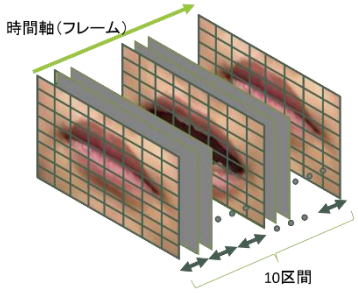

図 7　時間軸の標準化

（９）K-Nearest-Neighbor 法による識別

K-Nearest-Neighbor 法（KNN）とは，教師あり学習に含まれる分類アルゴリズムの一つで，距離情報に基づき探索する手法である．本実験ではベースライン手法として KNN を用いた．

（１０）SupportVectorMachine による識別

SupporVectorMachine（SVM）とは，学習データを用いて複数のクラスを分類する線を生成し，その線に基づき新規データがどのクラスに分類されるかを識別する手法である．SVM は従来の発話検出でも利用されており，高い精度が確認されている[5]．本実験では 6 単語の識別を行うため，複数の SVM を作成し識別を行うことで，最も尤度が高いものを結果として出力した．

（１１）評価

評価は，Leave-One-Out 法を用いた．Leave-One-Out 法とは，いくつかのモデルを生成したときに，どのモデルが最適であるのかを評価する解析方法の一つである．全サンプル数が n 個ある場合，特定のサンプルを 1 つ取り出しテストデータとし，残りの n-1 個を学習データとして検証を行う．これを繰り返し，すべてのサンプルをテストデータにした合計の精度を結果として出力する．

３．実験結果

（１）手振れがオプティカルフローへ及ぼす影響

速度ベクトル（オプティカルフロー）の大きさの出現頻度を図 8 に示す．

図中の Mobilie が２．（１）における（B）の状態を撮影したものを表し，それ以外は各発話時に（A）の状態を撮影したものを表す．図 8 を見ると，話時に発生するオプティカルフローの大きさは，主に 1.5～3.5 の大きさで出現していることがわかる．それに対し，発話をせずに手で持っていただけの場合も同様に 1.5~3.5 の範囲内で多く発生していることが確認された．

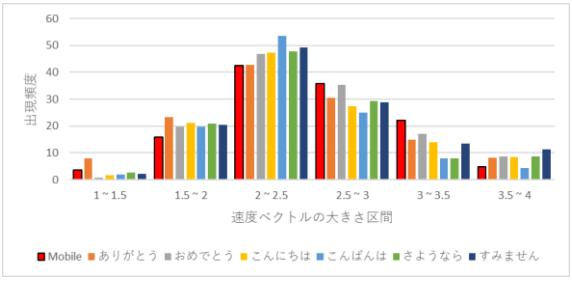

図 8　速度ベクトルの出現頻度

（２）KNN を用いた識別結果

KNN を用いた際の識別結果を図 9 に示す．各単語の混同行列を表１～４に示す．手振れとして得られた各領域のオプティカルフローを引き識別を行った場合，より高い識別率を得ることができた．

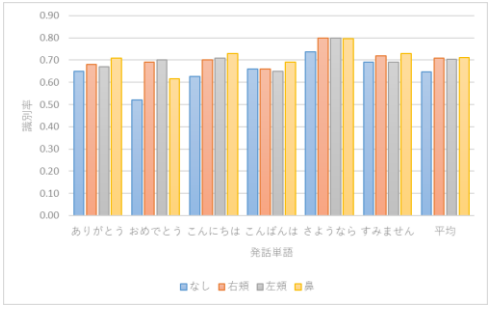

図 9　KNN を用いた際の識別率

表 1　口領域のみの混同行列(KNN)

		入力単語					
		ありがとう	おめでとう	こんにちは	こんばんは	さようなら	すみません
正解単語	ありがとう	0.65	0.07	0.10	0.08	0.08	0.02
	おめでとう	0.05	0.52	0.10	0.18	0.10	0.05
	こんにちは	0.01	0.04	0.63	0.20	0.07	0.05
	こんばんは	0.01	0.02	0.15	0.66	0.11	0.05
	さようなら	0.02	0.03	0.06	0.12	0.74	0.03
	すみません	0.00	0.01	0.06	0.18	0.06	0.69

表 2　鼻領域を用いた際の混同行列(KNN)

		入力単語					
		ありがとう	おめでとう	こんにちは	こんばんは	さようなら	すみません
正解単語	ありがとう	0.71	0.06	0.08	0.06	0.08	0.02
	おめでとう	0.04	0.62	0.06	0.13	0.09	0.06
	こんにちは	0.02	0.02	0.73	0.14	0.06	0.03
	こんばんは	0.01	0.02	0.13	0.69	0.09	0.06
	さようなら	0.02	0.00	0.05	0.07	0.80	0.06
	すみません	0.01	0.02	0.02	0.14	0.08	0.73

表 3　右頬領域を用いた際の混同行列(KNN)

		入力単語					
		ありがとう	おめでとう	こんにちは	こんばんは	さようなら	すみません
正解単語	ありがとう	0.68	0.07	0.06	0.08	0.10	0.01
	おめでとう	0.03	0.69	0.06	0.06	0.12	0.04
	こんにちは	0.05	0.01	0.70	0.15	0.08	0.01
	こんばんは	0.03	0.04	0.13	0.66	0.11	0.03
	さようなら	0.05	0.00	0.07	0.08	0.80	0.00
	すみません	0.03	0.01	0.03	0.16	0.05	0.72

表 4　左頬領域を用いた際の混同行列(KNN)

		入力単語					
		ありがとう	おめでとう	こんにちは	こんばんは	さようなら	すみません
正解単語	ありがとう	0.67	0.07	0.05	0.07	0.13	0.01
	おめでとう	0.02	0.70	0.04	0.04	0.16	0.04
	こんにちは	0.03	0.00	0.71	0.16	0.07	0.03
	こんばんは	0.01	0.05	0.16	0.65	0.09	0.04
	さようなら	0.04	0.00	0.06	0.08	0.80	0.02
	すみません	0.02	0.05	0.03	0.14	0.07	0.69

（３）SVM を用いた際の識別結果

SVM を用いた際の識別結果を図 10 に示す．各単語の混同行列を表５～８に示す．KNN 同様，手振れとして得られた各領域のオプティカルフローを引き識別を行った場合，より高い識別率を得ることができた．また，KNN で識別を行った場合より全体的に精度が向上した．

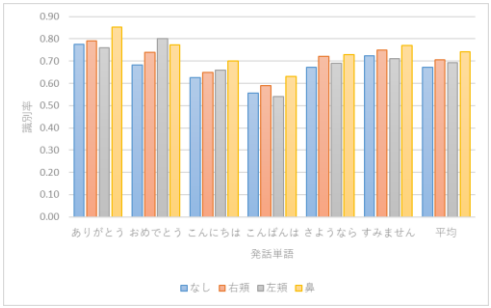

図 10　SVM を用いた際の識別率

表 5　口領域のみの混同行列(SVM)

		入力単語					
		ありがとう	おめでとう	こんにちは	こんばんは	さようなら	すみません
正解単語	ありがとう	0.78	0.07	0.05	0.02	0.06	0.02
	おめでとう	0.10	0.68	0.05	0.07	0.05	0.04
	こんにちは	0.01	0.09	0.63	0.12	0.08	0.06
	こんばんは	0.01	0.09	0.12	0.56	0.10	0.12
	さようなら	0.05	0.08	0.09	0.03	0.67	0.08
	すみません	0.02	0.03	0.07	0.06	0.10	0.72

表 6　鼻領域を用いた際の混同行列(SVM)

		入力単語					
		ありがとう	おめでとう	こんにちは	こんばんは	さようなら	すみません
正解単語	ありがとう	0.85	0.06	0.03	0.00	0.05	0.00
	おめでとう	0.06	0.77	0.06	0.03	0.04	0.04
	こんにちは	0.02	0.05	0.70	0.12	0.08	0.04
	こんばんは	0.01	0.08	0.06	0.63	0.09	0.13
	さようなら	0.02	0.07	0.06	0.05	0.73	0.07
	すみません	0.01	0.01	0.06	0.07	0.07	0.77

表 7　右頬領域を用いた際の混同行列(SVM)

		入力単語					
		ありがとう	おめでとう	こんにちは	こんばんは	さようなら	すみません
正解単語	ありがとう	0.79	0.09	0.06	0.02	0.04	0.00
	おめでとう	0.05	0.74	0.05	0.06	0.02	0.08
	こんにちは	0.01	0.06	0.65	0.17	0.09	0.02
	こんばんは	0.01	0.06	0.12	0.59	0.07	0.15
	さようなら	0.00	0.05	0.09	0.11	0.72	0.03
	すみません	0.01	0.07	0.05	0.10	0.02	0.75

表 8　左頬領域を用いた際の混同行列(SVM)

		入力単語					
		ありがとう	おめでとう	こんにちは	こんばんは	さようなら	すみません
正解単語	ありがとう	0.76	0.10	0.09	0.01	0.04	0.00
	おめでとう	0.05	0.80	0.04	0.04	0.03	0.04
	こんにちは	0.02	0.03	0.66	0.19	0.10	0.00
	こんばんは	0.00	0.08	0.14	0.54	0.10	0.14
	さようなら	0.02	0.05	0.08	0.12	0.69	0.04
	すみません	0.01	0.05	0.07	0.10	0.06	0.71

４．考察

手振れによるオプティカルフローが識別に影響を与えるという結果から，手振れの影響を軽減すれば識別精度が一様に上がると考えられる．しかし，図 10 を見ると，右頬領域を用いた際，単語によっては精度が落ちてしまう領域があった．同様に，鼻領域を利用した際，ありがとうのみ精度が他の領域より精度が低い場合があった．これは，発話時に対象領域に動きがあり，それによって発生したオプティカルフローがノイズになってしまったと考えられる．図 9，図 10 を見るとこんにちは，こんばんはの識別率が低かった．混同行列を確認すると，こんにちはとこんばんはではお互いに誤識別が起こっていることが分かる．これは，発話時の筋肉の動かし方が似ているため，発生するオプティカルフローも近い値になってしまうためだと考えられる．

５．今後の展望

スマートフォンで撮影された動画の識別を検討し，識別率を向上させることができた．しかし，差分領域として選んだ個所によって識別率の低下が起きる個所があった．そのため，今回使用したデータセットには含まれていないが，より発話時に動きが少ないと考えられる目の付近などその他の領域を手振れ検出する領域として設定し検証することが今後の課題として挙げられる．また，発話時の筋肉の動きが似ている場合，互いを誤識別してしまうことがあった．そのため，それらの識別を正確に行うため，利用する特徴量の検討も今後の課題として挙げられる．

謝辞

本研究の一部には,科学研究費補助金(基盤(C)19K12188)の助成を得た．

参考文献

[1] S. Agrawal, V. R. Omprakash and Ranvijay, "Lip reading techniques: A survey," 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), Bangalore, 2016, pp. 753-757.

[2] 総務省，“インターネット利用の広がり，” 総務省, 2019.Available:https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h30/html/nd142110.html. [アクセス日: 15 1 2020].

[3] 齊藤 剛史, 窪川 美智子: SSSD：スマートデバイスを用いた読唇技術向け日本語データベース, 電子情報通信学会技術研究報告, vol.117, no.513, pp.163-168, 2018.

[4] T. Saitoh , M. Kubokawa, “SSSD: Speech Scene Database by Smart Device for Visual Speech Recognition,” Proc. of ICPR2018, pp.3228-3232, 2018.

[5] A.A.Shaikh, D.K.Kumar, W.C.Yau, M.Z.C.Azemin and J. Gubbi, "Lip reading using optical flow and support vector machines," 2010 3rd International Congress on Image and Signal Processing, Yantai, 2010, pp. 327-330.