機械読唇を用いた母音認識および子音認識による発話語推定

Abe, Kiyoshiro / 阿部, 聖志朗

(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. 理工学研究科編
(巻 / Volume)
63
(開始ページ / Start Page)
1
(終了ページ / End Page)
4
(発行年 / Year)
2022-03-24
(URL)
https://doi.org/10.15002/00025370

機械読唇を用いた母音認識 および子音認識による発話語推定

Estimating Spoken Words by Vowel and Consonant Identification Using Machine Lip Reading

阿部聖志朗
Kiyoshiro Abe
指導教員　平原誠

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 recent years, with the development of speech recognition technology, communication between people has become smoother. However, it is not as smooth in noisy environments and for people with speech disabilities. In this study, we propose a machine lipreading method that estimates the content of speech based on video information alone. Machine lipreading of Japanese generally involves vowel identification. Therefore, we attempt to perform consonant recognition in addition to vowel recognition, and compare these results with a word corpus to estimate the speech content.

Key Words *: lipreading, vowel, consonant*

１．はじめに

機械読唇とは，動画像情報のみで発話内容を推定する技術である．これは，雑音化の音声認識や聴覚障がい者のコミュニケーション支援として研究が進められている．

機械学習においては口唇周辺部の特徴量を抽出，学習することにより日本語の母音認識を行う手法が主流である[1]．予め決められた単語を予測する単語認識は高い精度が出ている[2][3]．しかし「イヌ」と「ニス」のように同じ母音の並びでも異なる単語が存在する．そのような単語は母音のみからでは予測できないため認識対象となる単語を限定することで対応しているのが一般的である．そこで本研究は，母音認識の結果に子音認識の結果を組み合わせることで発話内容を予測する．母音および子音の認識結果と単語コーパスとを比較して発話内容に似ている単語を列挙するシステムを提案する．

２．提案手法

本システムは特徴抽出部と母音認識部，子音認識部，単語列挙部の４つから成る．

特徴抽出部には唇，鼻，輪郭，眉などの座標点を取得できるソフトウェアライブラリを使用した[4]．得られた座標点から「口唇中央の縦幅」と「口唇中央の横幅」に加え「鼻先から顎の中央までの縦幅」の３つを特徴量とした．

母音認識部と子音認識部はそれぞれ RNN (Recurrent Neural Network)を用いて学習により構築した．

母音認識部は５つの母音に閉口状態の「ん」を加

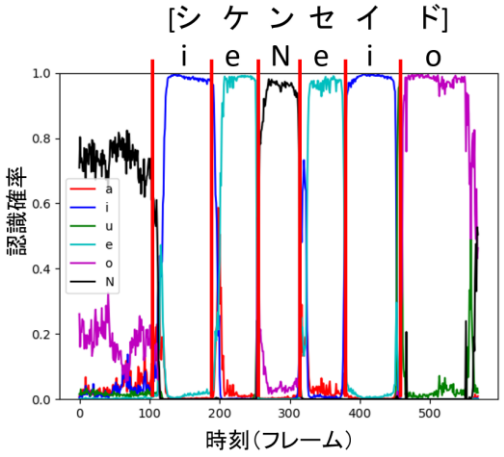

図１　母音認識部の認識確率と時間遷移

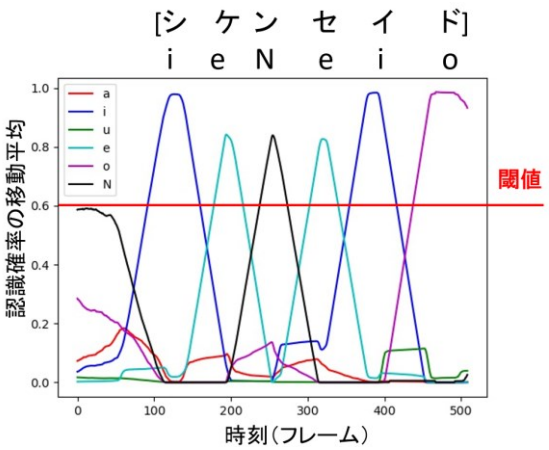

図２　母音認識部の認識確率の移動平均

えた[a,i,u,e,o,N]の6種類を分類する. 母音認識部は(N=50)フレームを1つの入力時系列とし, [a,i,u,e,o,N]それぞれに対応する認識確率を出力する. 図1は「シケンセイド」に対して入力時系列Nフレームを1フレームずつずらして得られた6分類の認識確率の時間遷移である. 図2は図1の認識確率の移動平均であり, 予め設定した閾値θ_1(=0.6)以上の山を取ることによって母音列(予測母音列)に変換した. 図2の場合は[ieNeio]となる.

子音認識部は, 研究当初は子音の行である[k,s,t,n,h,m,y,r,w]の9種類を分類していたが, 口を閉じる動作を含む「ま行」と「ぱ行」, 「ば行」のみ認識できることが分かった. そのため子音認識部を「口を閉じる子音」であるかどうかの2種類を分類するものとして再構築した. 子音認識部も母音認識部と同様にNフレームを1つの入力時系列とし「口を閉じる子音」の認識確率を出力する. 予測子音列への変換は次のとおりである. まず母音に変換した区間で子音の認識確率の移動平均が予め設定した閾値θ_2 (=0.7) 以上であれば, その区間を「口を閉じる子音」とする. ただし, 「ん」の区間で閾値θ_2以上となった場合は「口を閉じる子音」として変換しないようする. これは「ん」の口に対しての子音が存在しないためである.

単語列挙部は, まず予測母音列とコーパス内の単語の母音列とを比較する. コーパスは, 名詞と形容詞が合計約60万単語含まれているものである. 比較には, レーベンシュタイン距離という指標を使用した. これは2つの文字列がどの程度異なっているかを示す指標であり, 部分一致や長さが違う場合にも対応できる. この指標に沿ってコーパス内の母音列が似ている単語を列挙した. このとき, 予測子音列も利用した. 例えば予測子音列の2文字目が「口が閉じる子音」である場合, コーパス内の2文字目が「ま行」と「ぱ行」, 「ば行」を含む単語のみを対象として比較し列挙した.

3. 実験

3. 1 実験データ

母音認識部の学習データ用の動画は, 各母音につき50個撮影した. 子音認識部の学習データ用の動画は各行の音ができるだけ均等になるように合計50個撮影した. 単語列挙部のテストデータは, コーパス内からランダムに選んだ30個の単語を撮影した. 動画は60fpsのカメラを用いて撮影した.

3. 2 母音認識部

母音認識部の入力時系列長Nは50フレーム(入力時系列)として1フレームずつずらしながら入力した. 中間素子数を64とし重みの初期値を変えて5回実験を行った. 母音の認識率は89.9 (±0.5) %であった. テストデータ「ガメンミギオク」を発話した時の認識結果を図3に示す.

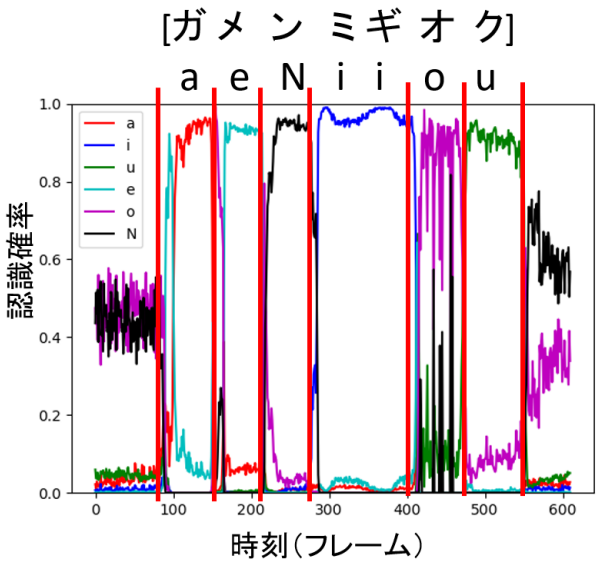

図3 「ガメンミギオク」に対する母音認識部の認識確率の時間遷移

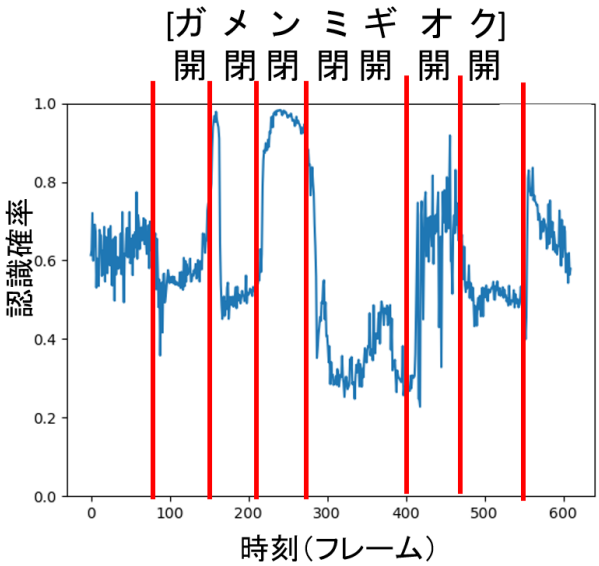

図4 「ガメンミギオク」に対する子音認識部の認識確率の時間遷移

3. 3 子音認識部

子音認識部の入力時系列長Nは50フレーム(入力時系列)として1フレームずつずらしながら入力した. 中間素子数を32とし重みの初期値を変えて5回実験を行った. 子音の認識率は68.8 (±0.8) %であった. テストデータ「ガメンミギオク」を発話した時の出力結果を図4に示す.

3. 4 単語列挙部

単語列挙部は, テストデータに対する母音認識部および子音認識部の結果を元に単語を列挙した. 予測母音列と正解単語の母音列との平均レーベンシュタイン距離は0.295 (±0.012) であった. 図5にテストデータ30個それぞれに対するレーベンシュタイン距離を示す. 表1はテストデータ「ガメンミギオク」に対する予測母音列とのレ

ーベンシュタイン距離が小さい単語を列挙した結果である．なお予測子音列を単語列挙に反映させた結果，全てのテストデータに対して精度が下がってしまった．表 2 は予測子音列を反映させた場合の「ガメンミギオク」に対する結果である．

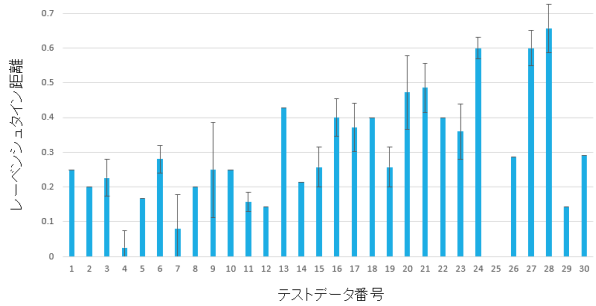

図 5 それぞれのテストデータのレーベンシュタイン距離

表 1 「ガメンミギオク」の予測母音列に基づく予測結果

距離	母音列	単語
0.0000	aeNiiou	ガメンミギオク
0.0000	aeNiiou	カゲンリキジョウ
0.1250	aneNiiou	サンテンイチコウ
0.1250	aeniNiou	ハッセンニンイジョウ
0.1250	aieNiiou	ダイセンニシショウ
0.1250	aaeNiiou	ナナテンイチコウ
0.1250	aeNeiiou	タメンテキキノウ
0.1250	aeNiaiou	シャケンヒタイオウ
0.1250	aeNiinou	アレンジウィンドウ

表 2. 「ガメンミギオク」の予測母音列および予測子音列に基づく予測結果

距離	子音列	母音列	単語
0.1429	s–jrp-	eNiiou	センジリッポウ
0.2500	kws–bjgy-	aaeiiiou	カワセイビジギョウ
0.2500	s–n–m–k-	aNeNeiou	サンネンメイコウ
0.2500	hnbshm–chzk	aaeiaiiou	ハナベシマイチゾク
0.2500	--z–m---	aieNeiou	アイゼンメイオウ
0.2500	h–m–p–jy-	eieNiiou	ヘイメンピージョウ
0.2500	k–s–b-	aNaiou	カンサイボウ
0.2500	g–s–b-	eNuiou	ゲンスイボウ
0.2500	n–sb-	aiiou	ナイシボウ

４． 考察

図 3 のようにテストデータに対して母音の認識が出来ているものが多かった．しかし，誤認識を起こす箇所もあり特に「あ」の部分を「え」と認識してしまうことがあった．これは「あ」と「え」の口が似ていることが原因だと考えられる．「う」と「お」の口は似ているが入力に「鼻先から顎の中央までの縦幅」を追加したことによって精度が向上した．同じように「あ」と「え」を差別化する特徴量が必要だと考えられる．

子音認識部は，「口を閉じる子音」の認識を行ったが誤認識が多かった．これは，学習の際「口を閉じる子音」の学習データを区別してないことが原因だと考えられる．例えば口を閉じる子音である「ま行」でも「ま，み，む，め，も」の５パターンがある．本研究では，それらを全て同じ「口を閉じる子音」として扱っているため，誤認識が生じたと考えられる．

図 5 から，それぞれのテストデータのレーベンシュタイン距離が大きく異なることが分かる．レーベンシュタイン距離が大きいテストデータを見ると，母音認識部の認識確率にノイズが多いことが分かった．このため母音認識部の認識確率を母音列に変換する際に多々誤変換が起こり，レーベンシュタイン距離が大きくなってしまったと考えられる．また表１のように予測母音列と似ている単語を見ると，レーベンシュタイン距離が同じ値になってしまうものが沢山あることが分かった．これは短い単語ほど同じ値になるものが多く，長い単語ほど少ないことが分かった．

表２を見ると予測子音列の出力を単語列挙に反映させた結果，精度が悪くなってしまった．テストデータ「ガメンミギオク」に対して予測母音列と予測子音列はそれぞれ[aeniiou]と[----閉--]であった．予測子音列[----閉--]の「閉」は「口を閉じる子音」である「ま行」および「ば行」，「ぱ行」のどれかであることを示している．「ガメンミギオク」に対して予測子音列は 5 文字目を「口を閉じる子音」であると予測しているが，実際には 4 文字目であるため精度が下がってしまった．このように子音列への変換で誤変換を起こしていることが多かった．また予測母音列と予測子音列から単語列挙を行なうとき，2 文字目が「口が閉じる子音」と予測した場合はコーパス内の 2 文字目が「ま行」と「ぱ行」，「ば行」を含む単語のみを対象にして比較し列挙する手法を取っている．そのため，予測子音列が正解単語の子音列と異なった場合に著しく精度が落ちたものと考えられる．

５． 今後の展望

子音認識を追加することによって精度向上を目指したが達成できなかった．これは子音認識部における認識精度および予測子音列への変換精度の低さに問題があると考えられる．子音認識部の改良には，唇の開閉をより細かく認識するための特徴の発見が必要であると考えられる．また予測子音列と正解単語の子音列との不一致により，単語列挙に大きな影響を及ぼした．単語列挙の精度向上のために，予測子音列の少しの誤認識を許容できる手法を取る必要があると考えられる．また予測母音列が正しくてもレーベンシュタイン距離が同じ値になる単語が複数あるため正解単語を当てることは難しい．そのため，単語のみではなく文脈を考慮したシステムが必要であると考えられる．

謝辞

本研究を進めるにあたり，終始適切な助言を賜り，また丁寧に指導してくださった，修士論文指導教員の平原誠准教授に心より感謝いたします.

参考文献

1) 間瀬健二， アレックス ペンドランド,”オプティカルフローを用いた読唇,”テレビジョン学会技術報告会 ITEJ Technical Report, Vol.12, No.44, pp.7-12, 1989.

2) 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，Vol.J90-D, No.4, pp.1105-1114, 2007.

3) 齊藤剛史，森下和敏，小西亮介,”発話シーンからのキーフレーム検出とキーフレームに基づく単語読唇,”電気学会論文誌，Vol.131, No.2, pp.418-424, 2011.

4) D. E. King,” Dlib-ml : A machine learning toolkit ,” Journal of Machine Learning Research, Vol.10, pp.1755-1758, 2009.