構音障がい者による発話内容の認識と
変換を目的とした音響モデルの検討

生体計測工学研究室　2181210　井野竣介
（主査：京相 雅樹 教授　　副査：桐生 昭吾 教授，　小林 千尋 講師）

1. はじめに

“構音障がい”とは、言葉を理解しているが、音を作る器官やその動きに問題があって発音がうまくできない状態のことである。手足が不自由な重度障がい者において音声認識は非常に重要である。しかし、従来のものでは発話に障がいがない人(成人男女)の音声を対象としているため、構音障がい者などの発話スタイルが異なった方では従来の音響モデルでは認識が困難である。音声認識技術は現在、車内でのカーナビの操作、会議における書き起こしなど様々な環境下や場面において使用される機会が増加している。音声認識技術を応用して、音声で家電を操作することなどは非常に便利だが、私は半身麻痺などによって構音障がいを併発した方にこそ必要なシステムであると考えている。しかし、一般的な音声認識システムでは構音障がい者の音声内容を高い精度で捉えることはできない。以上の経緯により、構音障がい者の方に特化した音声認識を研究している。

その先行研究[1]として、構音障がい者の音声を利用して特別に調整した音響モデルを作成したものがある。しかし、音響モデルを作成するのに大量の学習データが必要であったり、認識率が 70%以下であったりするなどの問題点があった。

そこで、私は音素分析をするための学習モデルに、読唇術システムで得た特徴量を組み込むことを考えた。しかし、読唇術には同口形異音語などの問題がある。これは、「煙草」「ナマコ」「卵」のように、同じ口の動きをする単語のことである。

構音障がい者の音素認識誤りの傾向を調べた先行研究[2]では、構音障がい者 3 名を対象とした音素認識実験により正解率が低下している音素が母音、子音ともに類似しており、特に母音/a/,/i/,/u/,/e/,/o/の正解率が低いことが分かった。そのため、読唇術システムによって母音の識別ができれば、不完全な音声信号をフォローすることができるのではないかと構想した。

本研究では、画像処理による読唇術のデータを音声認識システムに組み込むことで識別精度の向上の検討をする。本提案システムは、動画の音声から特徴量を取得する従来の音声認識に加えて、動画フレームから話者の口唇の動きを分析することで特徴量を取得する部分で構成されている。また、これらの特徴量から話者の発話内容を予測して出力する機能を持つ。最終的には自然発話の字幕化・テキスト化を目標としているが、本研究では母音 5 種の識別を目標としている。評価方法としては、識別モデルの性能評価を行った。また、動画フレーム毎の予想音素を出力することで、最終的な自然発話の字幕化・テキスト化という目標に向けたシステムとした。

2. 理論

2.1 音韻と音素

音声には、さまざまな情報が含まれるが、音声認識にとって重要な情報は音韻と音素についての情報である。音韻とは、ある言語における識別のために必要な最小な単位の集合として定義される。音韻は言語によって異なる。

一方、音素とは、音韻と同様に音声を構成する単位だが、音声の物理的な特徴で分類されたものである。国際音声学協会(International Phonetic Association)では世界中のあらゆる言語を表現するために必要な音素、国際音声字母(International Phonetic Alphabet)を定めている。音素と音韻の対応は一般に言語により異なる。

音素は母音と子音に分けられる。母音は声道や舌の形を変え、その結果フォルマント周波数を変えることで生成される。Fig.2-1 に日本語の母音の第 1 フォルマント(F1)と第 2 フォルマント(F2)の分布を示す。F1 と F2 によりこれらの 5 母音がきれいに分類されることがわかる。子音は唇や歯茎、鼻腔などさまざまな器官を利用して生成される。

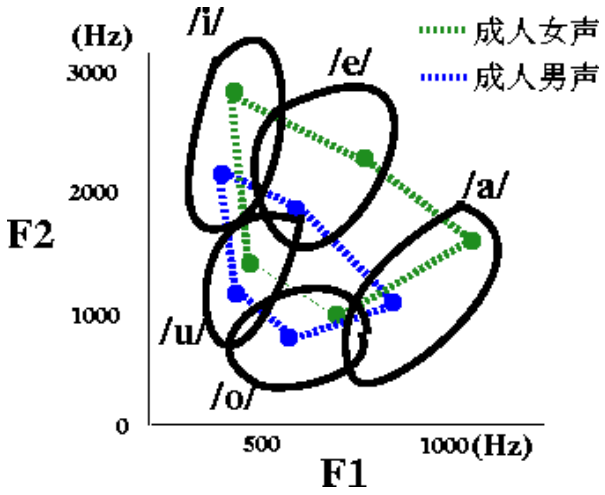

Fig.2-1 日本人発声の第 1 フォルマント(F1)と第 2 フォルマント(F2)の分布 [3]

日本語には、母音、子音以外に、小さい「ゃ」に対応する拗音(子音と半母音/y/の結合)、「ん」に対応する撥音、小さい「っ」に対応する促音がある。また、二重母音(/ei/、/ou/など)や長母音(/e:/など)を音素とすることもある。

ここで、調音位置、調音様式、有声と無声の別など、音韻ごとの特徴を表すものを、特に弁別素性と呼ぶ。音韻は弁別素性の束として表現可能である。例えば母音/o/の弁別素性は、{母音、舌の高さが中央、舌の前後位置が奥}となる。

音声は、必ずしも語学辞書の発音表記に忠実に発音されるとは限らず、人間の調音器官は、筋肉が連続的に動くことで別々の音素に対応する声道の形を作っている。
また、音素の音響的特徴は、その音素だけでなく、その前後にどんな音素が続いているかによって変化する。音声認識において問題となるこの現象を調音結合と呼ぶ。

3. 研究方法

3.1 実験 1：読唇術システムの評価実験

読唇術システムを作成し、そのデータを分析することで構音障がい者の発話時における口唇の動き方の特徴の分析や、母音 5 種の分類を行った。計測方法としては Fig.4-1、読唇術システム全体の流れは Fig.4-2 に示す。このように被験者 2 名(男性構音障がい者 1 名・男性健常者 1 名)が 1 秒ごとに閉口と発話(/a,i,u,e,o/)を 5 回ずつ繰り返した動画を撮影し、これを 2 回行い、安定部分を切り取った。安定部分とは、発話中口唇の動きが安定したことを目視で確認した部分である。被験者 1 名につき、学習用データ 25 個とテスト用データ 25 個を取得した。計測データから、特徴量を算出した。特徴量を 2 種の学習モデル(SVM･NN)によってクラス識別をし、その正解率を算出した。

Fig.3-1 計測方法

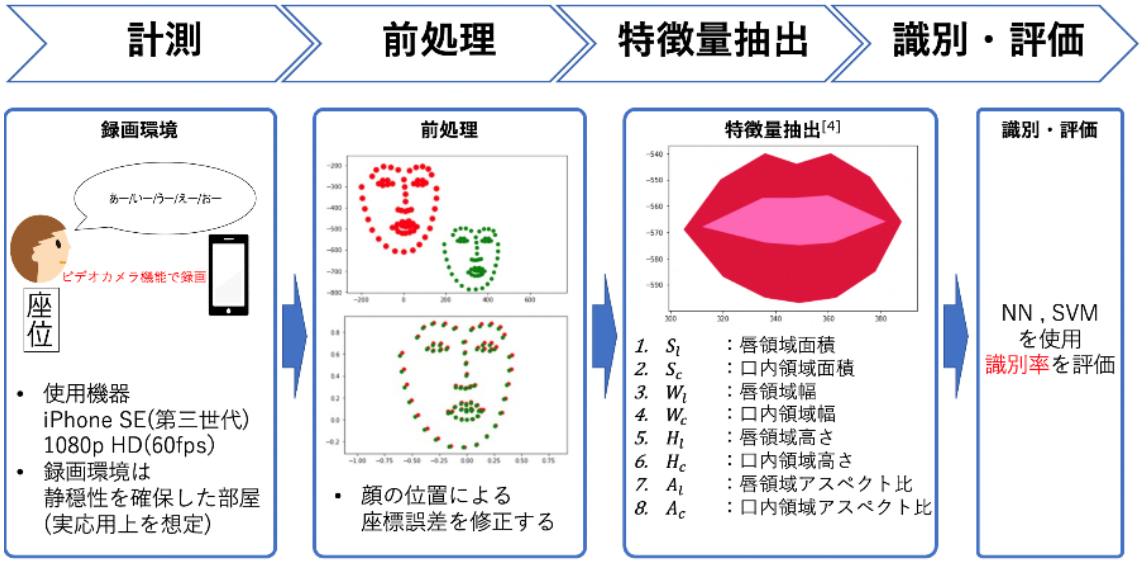

Fig.3-2 読唇術システム

3.2 実験 2：学習モデルの精度評価実験

Fig.3-5 に示した構想システムにより実験 2 で用いた動画を解析した。安定部分を切り取るのではなく自動で無音部分をカットし、発話フレーム時の特徴量を取得した。母音 5 種の動画それぞれについて教師ラベルを付与し、ランダムで学習データ 70%、テストデータ 30%にスプリットして学習モデルの精度を評価した。

3.3 実験 3：構想システムの性能評価実験

本構想システムは、動画フレームから話者の口唇の動きを分析することで特徴量を取得する読唇術ブロックと動画の音声から特徴量を取得する音声特徴量抽出ブロックで構成されている。また、これらの特徴量から話者の発話内容を予測して出力する音素識別ブロックを持つ。

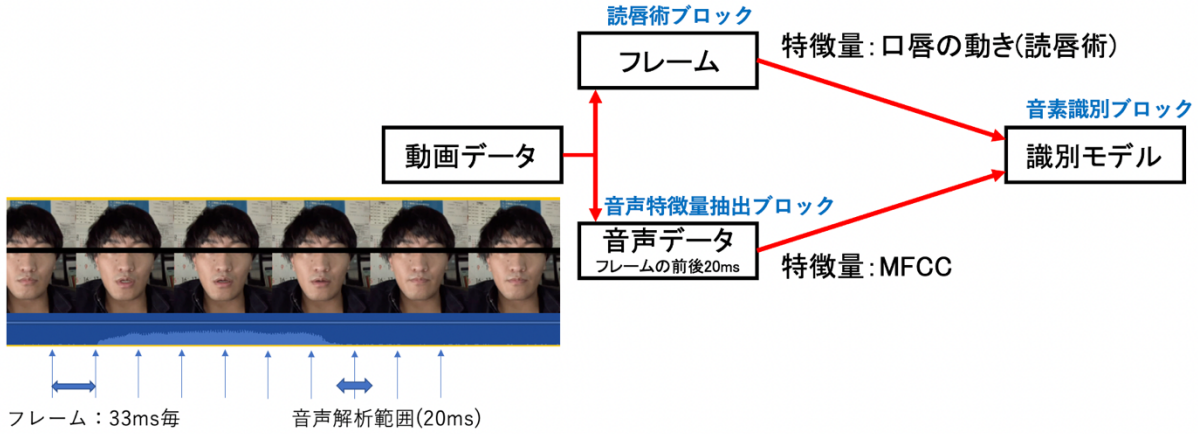

Fig.3-3 構想システム概要

実験 2 で取得した全ての特徴量で学習したモデルにより、未知の動画データの発話内容をフレーム毎に識別した。今回は母音のみで「あ、え、い、お、う」と事前発話した動画を入力した。各フレームに対する音素分析の結果を出力した。実験環境は実験 1 と同様である。

4. 結果・考察

4.1 実験１：読唇術システムの評価実験

被験者 2 名(男性構音障がい者 1 名・男性健常者 1 名)についてサポートベクターマシンで識別した結果を Fig.4-1,2 に示す。構音障がい者については正答率 72%で、母音のみの識別にも関わらず目標の 90%よりも低いという結果になった。健常者についても 84%で従来の音響モデルによる識別率よりも低いという結果になった。原因として学習データとテストデータの少なさにあると考えられる。本実験では、動画データから画像データの切り出し、解析までの作業時間に時間が掛かってしまった。そのため、１つの外れ値に大きな影響を受けてしまったと考えられる。

構音障がい者		識別結果(個)					テストデータ数	正答率
		a	i	u	e	o		
正解値	a	5	0	0	0	0	5	100%
	i	0	5	0	0	0	5	100%
	u	0	1	1	0	3	5	20%
	e	0	2	0	3	0	5	60%
	o	0	0	1	0	4	5	80%
回答数		5	8	2	3	7	25	72%

Fig.4-1 構音障がい者 SVM 結果

健常者		識別結果(個)					テストデータ数	正答率
		a	i	u	e	o		
正解値	a	5	0	0	0	0	5	100%
	i	0	4	0	1	0	5	80%
	u	0	0	5	0	0	5	100%
	e	0	0	0	5	0	5	100%
	o	0	0	1	2	2	5	40%
回答数		5	4	6	8	2	25	84%

Fig.4-2 健常者 SVM 結果

被験者 2 名(男性構音障がい者 1 名・男性健常者 1 名)についてニューラルネットワークで識別した結果を Fig.4-1,2 に示す。構音障がい者については正答率 88%で、母音のみの識別ではあるが、目標の 90%に近い精度が出た結果になった。健常者については 84%で従来の音響モデルによる識別率よりも低いという結果になった。識別音素を増やしても高い識別率を維持することができれば、本研究システムにおいて音声認識をフォローできるのではないかと考えられる。

構音障がい者		識別結果(個)					テストデータ数	正答率
		a	i	u	e	o		
正解値	a	5	0	0	0	0	5	100%
	i	0	5	0	0	0	5	100%
	u	0	0	3	1	1	5	60%
	e	0	0	0	5	0	5	100%
	o	0	0	1	0	4	5	80%
回答数		4	8	4	5	4	25	88%

Fig.4-3 構音障がい者 NN 結果

健常者		識別結果(個)					テストデータ数	正答率
		a	i	u	e	o		
正解値	a	5	0	0	0	0	5	100%
	i	0	4	0	1	0	5	80%
	u	0	0	5	0	0	5	100%
	e	1	0	0	4	0	5	80%
	o	2	0	1	0	2	5	40%
回答数		8	4	4	5	2	25	84%

Fig.4-4 健常者 NN 結果

4.2 実験２：読唇術システムの評価実験

実験 2 で得た結果を Table4-1 に示す。

Table 4-1 学習モデルの精度評価

被験者	学習モデル	音声のみ	読唇術のみ	音声+読唇術
健常者	SVM	98.9%	91.2%	100%
	NN	98.9%	91.2%	100%
構音障がい者	SVM	90.1%	93.4%	97.8%
	NN	90.1%	93.4%	97.8%

学習モデルについて、サポートベクターマシンとニューラルネットワークの結果は同一だった。健常者の場合は音声のみの精度が読唇術の精度よりも高かったが、構音障がい者の場合は読唇術の精度の方が音声のみによる精度よりも高かった。したがって、音声が不完全な構音障がい者について 2 つの特徴量を組み合わせた学習モデルの精度が最も高かったことからも、研究目的を達成していると考えられる。

4.3 実験３：構想システムの性能評価実験

実験 3 で得た結果を Fig.5-6,7 に示す。また、学習モデルについてサポートベクターマシンとニューラルネットワークの結果は同一だったため省略する。本システムでは、各フレームに対する音素分析の結果を出力している。

健常者結果　　自然発音発音：あえいおう
音声+読唇術
あ,あ,あ,あ,あ,あ,あ,あ,あ,あ,あ,あ,あ,あ,あ,あ,あ,あ,え,え,え,あ,え,あ,え,え,え,え,え,え,え,え,え,え,え,え,い,い,い,い,い,い,い,い,い,い,い,い,い,お,お,お,お,お,お,お,お,お,お,お,お,お,お,お,お,い,い,う,う,う,う,う,う,う,う,う,う,う,う,う,う,う,う,う,
音声のみ
あ,あ,あ,あ,あ,あ,あ,あ,あ,あ,あ,あ,あ,あ,あ,あ,あ,え,え,え,い,え,お,え,え,え,え,え,え,え,え,え,え,え,え,い,い,い,い,い,い,い,い,い,い,い,い,い,お,お,お,お,お,お,お,お,お,お,お,お,お,お,お,お,い,い,う,う,う,う,う,う,う,う,う,う,う,う,う,う,う,う,う,
読唇術のみ
い,え,あ,あ,あ,あ,あ,あ,あ,あ,あ,あ,あ,あ,あ,え,あ,え,え,え,あ,あ,あ,あ,あ,あ,あ,あ,あ,あ,あ,あ,あ,え,あ,あ,あ,あ,え,い,い,い,い,い,い,い,い,い,い,い,い,い,お,お,お,う,お,お,お,お,お,お,お,お,お,う,お,お,お,お,う,う,う,う,う,う,う,い,う,い,う,う,う,う,う,

Fig.4-5 健常者構想システム結果

構音障がい者結果	自然発音発音：あえいおう
音声+読唇術	
あ,あ,あ,あ,あ,あ,あ,あ,あ,あ,あ,え,え,え,え,え,え,え,え,え,え,え,え,え,え,え,い,い,い,い,い,い,い,い,い,い,い,い,い,い,え,え,お,う,う,お,お,お,お,お,お,お,お,お,お,お,お,う,お,う,う,う,う,う,う,う,う,う,う,う,	
音声のみ	
あ,あ,え,あ,あ,あ,あ,あ,あ,あ,え,え,え,お,お,え,お,お,お,お,え,お,お,え,え,い,い,い,い,い,い,い,い,い,い,い,い,い,い,い,え,う,う,お,お,お,お,お,お,お,え,お,お,お,お,お,お,い,お,う,お,う,う,う,う,う,う,う,う,う,う,う,	
読唇術のみ	
あ,あ,あ,あ,あ,あ,あ,あ,あ,あ,あ,え,あ,え,え,え,え,え,え,い,え,え,え,え,え,え,え,え,い,い,い,い,い,い,い,い,い,い,い,い,い,い,え,う,う,う,う,お,お,お,お,お,お,お,お,お,お,お,お,う,う,う,う,う,う,い,う,う,う,う,う,う,う,	

Fig.4-6 構音障がい者構想システム結果

実験 3 と同様に、健常者については音声のみであっても適切なテキスト化がされているようだが、構音障がい者は音声だけの学習モデルでは適切なテキスト化ができなかった。しかし、読唇術と組み合わせたものでは高い精度でテキスト化することができたのではないかと考えられる。

5. 終わりに

音声信号処理技術は、これまで健常者を対象としたものがその多くを占めてきた。しかし我が国において 2016 年に「障害者差別解消法」が施行され、ハードウェア・ソフトウェアの両面から各分野における情報技術支援の重要性は非常に高まっている。そこで、本研究では脳梗塞による半身麻痺による構音障がい者の音声認識の検討を行った。半身麻痺による構音障害者の発話スタイルは、神経麻痺による筋肉の弛緩が原因で健常者と大きく異なり不安定であるため、特定話者モデルでの音声認識には限界がある。構音障がい者の音素認識誤りの傾向を調べた先行研究では、構音障がい者 3 名を対象とした音素認識実験により正解率が低下している音素が母音、子音ともに類似しており、特に母音/a/,/i/,/u/,/e/,/o/の正解率が低いことが分かった。

これに対し本研究では、構音障がい者の音声認識精度の改善を目的とし、音声の特徴量に加えて動画から口唇の動きを特徴量として用いる手法を提案した。

結果として、構音障がい者については音声のみで行った音声認識によるテキスト化と比較して、読唇術の特徴量と組み合わせることで高い性能を出すことに成功した。しかしながら、母音のみの識別であり、完全な自然発話のテキスト化システムとはならなかった。そのため、今後は音素全ての識別を可能とするためにフレーム前後の動きまで考慮したり、△MFCC 特徴量の活用をしたりすることが重要であると考えられる。

参考文献

[1] 宮本千琴, 松政宏典, 滝口哲也, 有木康雄, 李義昭, 中林稔堯,
「構音障害者の連続音声認識の検討」, http://www.me.cs.scitec.kobe-u.ac.jp/~takigu/pdf/2009/miyamoto-asj09s.pdf, 最終閲覧日 2023/01/31.

[2] 吉岡利也, 高島遼一, 滝口哲也, 有木康雄, 李義昭,
「構音障害者の音素認識誤りの傾向」, 日本音響学会講演論文集, 2012/9, p137-140.

[3] 和歌山大学大学院システム工学研究科聴覚メディア研究室, 「音声 1」,
https://media.sys.wakayama-u.ac.jp/kawahara-lab/LOCAL/diss/diss7/S3_6.htm,
最終閲覧日 2023/01/31.

[4] 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」,
信学論 D, Vol.J90-D,No.4,pp.1105-1114, 2007/4.