音声強調手法を用いた構音障害者の音声認識

宮本，千琴
滝口，哲也
有木，康雄

(Citation)
神戸大学都市安全研究センター研究報告,13:75-80

(Issue Date)
2009-03

(Resource Type)
departmental bulletin paper

(Version)
Version of Record

(JaLCDOI)
https://doi.org/10.24546/81001954

(URL)
https://hdl.handle.net/20.500.14094/81001954

神戸大学都市安全研究センター
研究報告，第13号，平成21年３月

音声強調手法を用いた構音障害者の音声認識

Dysarthric Speech Recognition Using Speech Enhancement

宮本　千琴[1)]
Miyamoto Chikoto
滝口　哲也[2)]
Tetsuya Takiguchi
有木　康雄[3)]
Yasuo Ariki

概要：音声認識技術は現在，様々な環境下や場面において利用される機会が増加している．しかし，言語障害者などの障害者を対象としているものは非常に少ない．本稿では，脳性麻痺による構音障害者の音声認識の検討を行う．構音障害者の発話スタイルは健常者と大きく異なり，認識精度の改善の一つとして構音障害者特有の特徴量抽出が必要であると考えられる．本稿では，健常者の音声に比べて構音障害者の音声は明瞭度が低く，スペクトル情報が埋もれてしまっている点に着目し，ローカルピーク強調による音声強調法を用いた特徴量抽出法を検討する．

キーワード：構音障害，言語障害，脳性麻痺

１．はじめに

近年，音声認識技術の発展に伴い[1)]，様々な環境下や場面での利用が期待されている．例えばカーナビゲーションの操作や会議音声の議事録化など様々な分野に応用されている．また，これまでは成人を対象とした音声認識が多くみられたが，最近では成人だけでなく高齢者や子供など成人と発話スタイルが異なる人も対象としており，様々な人が利用する機会が増えている[2)3)]．

現在，日本に言語障害者は４万2000人いるとされている[4)]．文献[5)]では，構音障害者音声を対象とした音響モデルの適応の検証を行っており，文献[6)]では，特徴量抽出や音響モデルの構築を行っているが，言語障害者などの障害者を対象としているものは非常に少ない．手足の不自由を患っている場合など音声に頼るほかない場合，構音障害者を対象とした音声認識の実現が期待され，障害者の就業機会の増加や講演時の補助等への活用などが望まれる．

構音障害とは言葉を正しく明瞭に発音できない症状である．構音障害の原因の一つとして，脳性麻痺が考えられる．脳性麻痺は乳児1000人につき2～4人の割合で起こる．脳性麻痺の定義として，1968年の厚生労働省脳性麻痺研究班は，「受胎から４週間以内の新生児までの間に生じた，脳の非進行性病変に基づく，永続的な，しかし変化しうる運動および姿勢の異常である．その症状は満２才までに発現する．進行性疾患や一過性運動障害，または将来正常化するであろうと思われる運動発達遅滞は除外する．」としている．

脳性麻痺とは，筋肉の動きをつかさどる脳の部分が受けた損傷が原因で筋肉の制御ができなくなり，けいれんや麻痺，そのほかの神経障害が起こる症状のことである．出生前，出生時，出生直後の脳への酸素供給，出生前の胎内感染，妊娠中毒症，分娩時の外傷，仮死状態，未熟出生，出生後の脳を覆う組織の炎症や外傷性損傷などが原因として考えられる．脳性麻痺は，脳の損傷部分によって主に痙直型（大脳皮質），アテトーゼ型（中脳もしくは脳基底核），失調型（小脳），混合型（脳の広範囲）に分類される．

本稿では，アテトーゼ型の脳性麻痺による構音障害者を対象としている．アテトーゼ型は脳性麻痺患者の約20％に発生し，筋肉が不随に動き，正常に制御できないアテトーゼと呼ばれる症状が見られる．とくに

緊張状態にあるときや，意図的動作を行うときに見られる．[7) 8)]

本稿では，構音障害者を対象とした音声認識の実現のために，まず従来用いられる不特定話者モデル，および構音障害者モデルでの音声認識を行う．そして，認識結果より考えられる構音障害者特有の問題に対する改善を検討する．そこで本稿では，健常者の音声に比べて構音障害者の音声は明瞭度が低いと考えられるため，ローカルピーク強調による音声強調手法 LPE（Local Peak Enhancement）を用いた特徴量抽出法を検討する．構音障害者音声に対してローカルピークを強調することによって，埋もれてしまっているスペクトル情報の抽出を行う．

2．LPE を用いた特徴量抽出

(1) 提案手法の概要

現在，音声認識システムにおける音声特徴量として MFCC（Mel Frequency CepstralCoefficient）が広く用いられている．この特徴量はメル尺度フィルタバンクの出力を対数変換し，離散コサイン変換（Discrete Cosine Transformation: DCT）を適用したケプストラムである．

本稿では，健常者の音声に比べて構音障害者の音声は明瞭度が低いと考えられるため，ローカルピーク強調による音声強調手法 LPE（Local Peak Enhancement）を用いる．文献[9)]では，雑音環境下で LPE を行うことによって，認識精度の改善を可能としている．本研究では，構音障害者音声に対して LPE を行いローカルピークを強調することによって，埋もれてしまっているスペクトル情報の抽出を行う．

(2) LPE (Local Peak Enhancement)

LPE（Local Peak Enhancement）[9)]とは，処理フレーム毎に調波構造部分を強調するように設計されたフィルタを用いる音声強調手法である．有声音の区間と無声音の区間の区別をすることなくフィルタの設計を行うことができ，また，基本周波数を推定する必要もないので動作が安定である．

この手法の手順と各手順の処理結果を図 1 に示す．

観測音声のスペクトルを $y_t(j)$ とする．ここで，t はフレーム番号を表す．

まず，観測音声の対数スペクトル $Y_t(j)$ を得る．

$$Y_t(j) = \log(y_t(j))$$

次に，対数スペクトルに対して DCT を行い，ケプストラム $C_t(i)$ を得る．

$$C_t(i) = \sum_j D(i,j) \cdot Y_t(j)$$

ここで，$D(i,j)$ は離散コサイン変換行列を表す．

人間の音声の調波構造の間隔より狭い変化と広い変化の部分を除去するために，ケプストラムの低次項と高次項を 0 に設定する．

$$\hat{C}_t(i) = \begin{cases} \varepsilon \cdot C_t(i), & if \quad i < lower\_cep \quad or \quad i > upper\_cep \\ C_t(i), & else \end{cases}$$

本稿では，ε を 0 の近似値として 10^{-3} に定めた．また，FFT のサンプル数が 512 点であるとき，$lower\_cep$ ＝1　$upper\_cep$ ＝120 とした．これは，本稿で使用した構音障害者の音声の第一フォルマント周波数が取りうる範囲に相当する．

除去された後のケプストラム $\hat{C}_t(i)$ に対して逆離散コサイン変換(IDCT) を行い，対数スペクトル $W_t(j)$ を得る．

$$W_t(j) = \sum_i D^{-1}(j,i) \cdot \hat{C}_t(i)$$

そして，指数関数によってスペクトル表現 $w_t(j)$ に変換し，平均が 1 になるように正規化を行う．

$$w_t(j) = \exp(W_t(j))$$

$$\overline{w}_t(j) = w_t(j) \cdot \frac{Num\_bin}{\sum_k^{Num\_bin} w_t(k)}$$

ここで，Num_bin は，FFT のサンプル数を表す．

フィルタとして得られた $\overline{w}_t(j)$ によって，強調音声のスペクトル $z_t(j)$ を得る．

$$z_t(j) = \overline{w}_t(j) \cdot y_t(j)$$

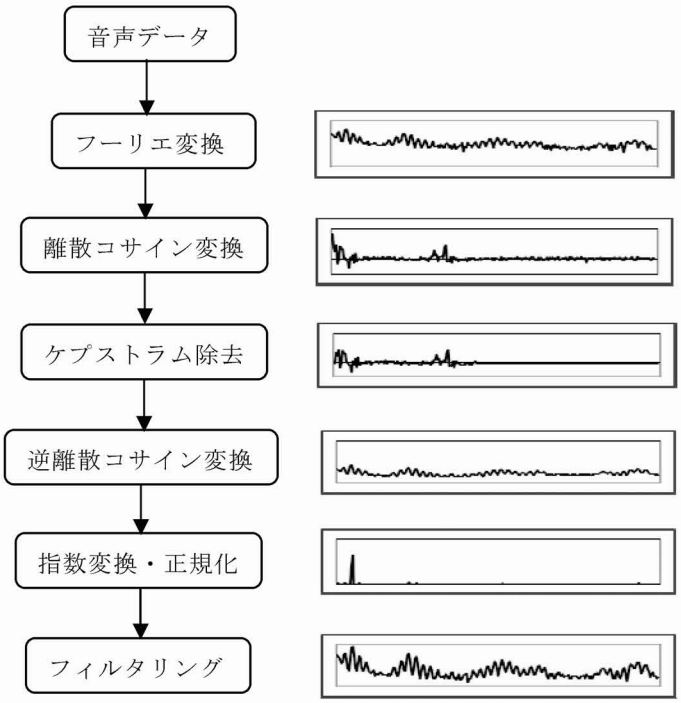

図 1.　LPE の手順

３．評価実験

(1) 実験条件

実験用データとして構音障害者，健常者それぞれ 1 名のデータを収録した．発話内容として，ATR 音素バランス単語（216 単語）から 210 単語を無作為に選択し，図 2 のように各単語を 5 回連続発話した音声を収録した．図 4，図 5 に構音障害者，健常者の音声波形の例を示す．実験には HTK[10] を用いる．

(2) 不特定話者モデルでの認識実験

初めに，不特定話者モデルでの認識実験を行った．音声データのサンプリング周波数は 16kHz，ハミング窓長は 25msec，フレーム周期は 10msec であり，音響モデルには文献 [11] に含まれている不特定話者モデルを用いた．評価用データは，1,050(210 単語× 5 回）個を用いた．認識結果を図 3 に示す．健常者においては 91.6%の単語正解精度が得られるが，構音障害者においては 3.1%しか認識できず，健常者と発話スタイルが異なるため，不特定話者モデルでの認識は困難であることがわかる．

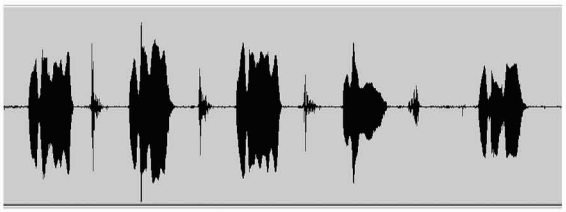

図 2.　収録音声データ

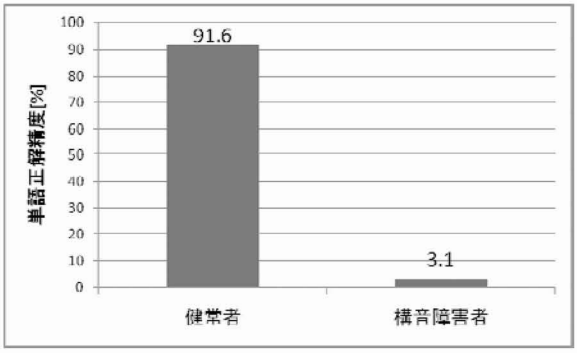

図 3.　不特定話者モデルでの認識実験結果

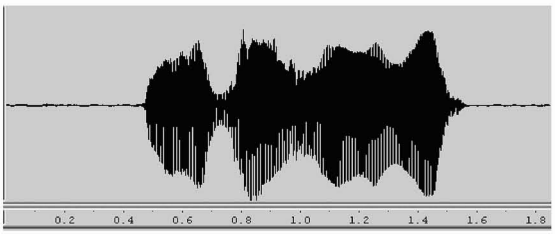

図 4.　構音障害者の音声例//i k i g a i

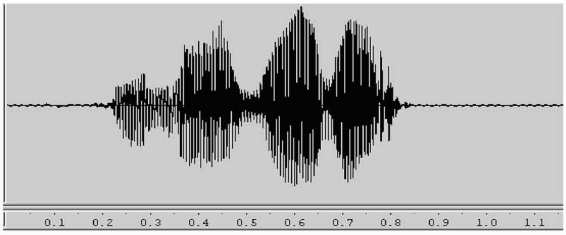

図 5.　健常者の音声例//i k i g a i

(3) 構音障害者モデルでの認識実験

不特定話者モデルでの認識が困難であることから，構音障害者の音響モデルを作成し認識実験を行った．1 回目の発話の評価を行う場合は，2〜5 回目の発話を音響モデルの学習に用いた．これを各発話に対して行う．5 回発話全体の平均の認識結果を図 6 に示す．

特定話者モデルを用いることで，構音障害者において 89.5%の精度が得られた．図 7 に健常者における発話回数ごとの認識結果，図 8 に構音障害者における発話回数ごとの認識結果を示す．

ここで，構音障害者（図 4）と健常者（図 5）の音声波形を比較することにより，構音障害者の音声は健常者の音声より明瞭度が低いことが考えられ，この問題を考慮すべきである．

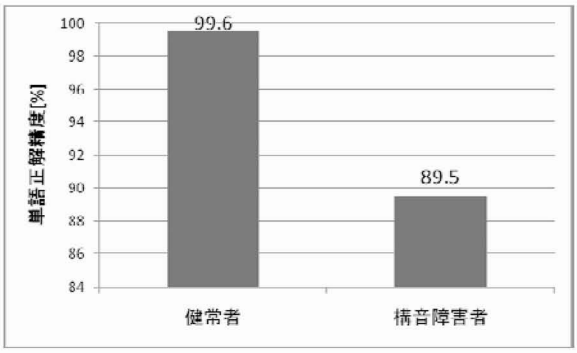

図 6.　特定話者モデルでの認識実験結果

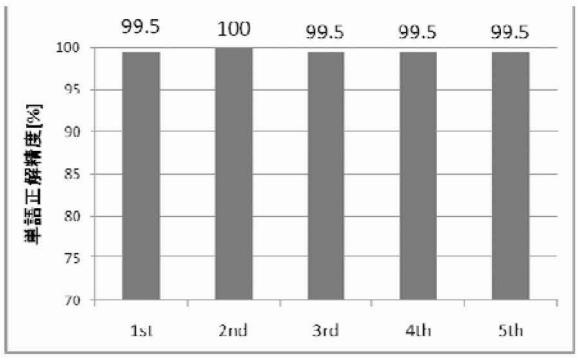

図 7.　健常者における発話回数ごとの認識率

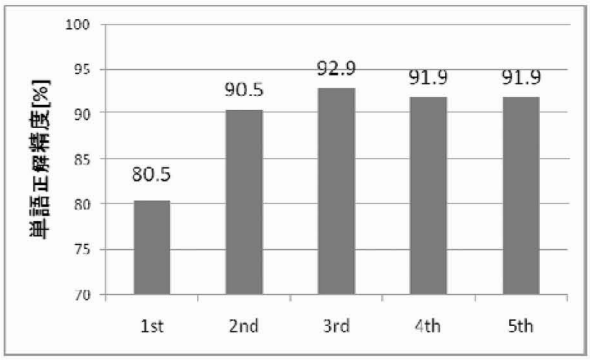

図 8.　構音障害者における発話回数ごとの認識率

(4) LPE を用いた認識実験

MFCC の特徴量と，LPE を適応した MFCC の特徴量（LPE）を用いて実験を行った．5 回発話全体の平均の認識結果を図 9 に，発話毎の認識結果を図 10 に示す．

5回発話全体の平均の認識精度において，ある程度の認識率が得られたものの，MFCCに及ばない結果に留まった．発話毎の認識精度において，5回目発話で0.5%の認識精度の改善を得た．しかし、ほとんどの発話でMFCCと同程度の認識精度あるいは下回る結果となった．この原因として，LPE を行うときのケプストラムを除去する範囲を一定にし,適応したことが考えられる.フォルマントは音素ごとに現れる周波数が異なり，また母音ではほぼ同一のフォルマントが見られるが，子音はフォルマントのはっきりしたものとはっきりしないものがあり，これらを考慮する必要がある．フレーム毎にフォルマント情報を抽出しフィルタを構築することが認識精度の改善に有効であると考えられる．

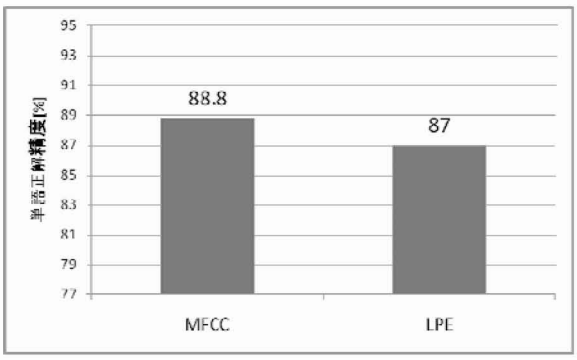

図 9.　提案手法による認識実験結果

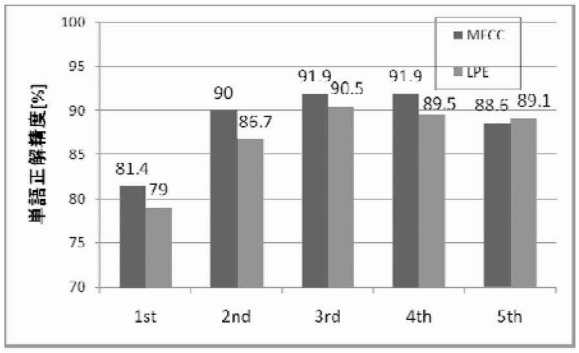

図 10.　提案手法による発話回数ごとの認識率

4．おわりに

本稿では，構音障害者の音声認識の検討として，従来用いられる不特定話者モデル，および構音障害者モデルでの音声認識実験を行い，健常者の音声に比べて構音障害者の音声は明瞭度が低く，スペクトル情報が埋もれてしまっている点に着目し,ローカルピーク強調による音声強調法を用いた特徴量抽出法を検討した．

今後の課題として，本研究では対象者が一名であったが，構音障害者特有の問題や音素情報などを確証してくために対象者を増やしていく予定である．また，構音障害者は健常者に比べてデータ収録の際の負担が大きいため，大量のデータが収録できない可能性がある．そこで，少量のデータでの音響モデルの構築は困難であるため，少量のデータでの音声認識手法の検討が重要となる．

参考文献

1）中川聖一，“音声認識研究の動向,” 電子情報通信学会論文誌，Vol.J83-D-Ⅱ，No.2,433-457，2000.

2）馬場朗，芳澤伸一，山田実一，李晃伸，鹿野清宏，“高齢者音響モデルによる大語彙連続音声認識,”電子情報通信学会論文誌，Vol.J85-D-Ⅱ，No.3，390-397，2002.

3）鮫島充，ランディゴメス，李晃伸，猿渡洋，鹿野清宏，“実環境における子供音声認識のための音韻モデルおよび教師なし話者適応の評価,” 情報処理学会論文誌，Vol.47，No.7，2295-2304，2006.

4）内閣府，“平成20 年版障害者白書.”

5）中村圭吾，田村直良，鹿野清宏，“発話障害者音声を対象にした健常者音響モデルの適応と検証,” 日本音響学会講演論文集，3-7-4，109-110，2005.

6）Hironori Matsumasa，Tetsuya Takiguchi，Yasuo Ariki，Ichao LI，Toshitaka Nakabayashi，“Integration of Metamodel and Acoustic Model for Speech Recognition,”Interspeech2008，2234-2237，2008.

7）Mark H. Beers，福島雅典，“メルクマニュアル医学百科最新家庭版,” 日経BP 社，2004.

8）S. Terry Canale，落合直之，藤井克之，“キャンベル整形外科手術書第4 巻小児の神経障害/小児の骨折・脱臼,” エルゼビア・ジャパン，2004.

9）O. Ichikawa，T. Fukuda，M. Nishimura，“Local Peak Enhancement Combined with Noise Reduction Algorithms for Robust Automatic Speech Recognition in Automobiles,” ICASSP2008，4869-4872，2008.

１０）“HTK (Hidden Markov Model Toolkit) ,” http://htk.eng.cam.ac.uk/.

１１）鹿野清宏，伊藤克亘，河原達也，武田一哉，山本幹雄，“音声認識システム,” オーム社，2001.

筆者：　1）宮本千琴，工学研究科情報知能学専攻，学生；2）滝口哲也，都市安全研究センター，講師；3）有木康雄，都市安全研究センター，教授

Dysarthric Speech Recognition Using Speech Enhancement

Chikoto Miyamoto
Tetsuya Takiguchi
Yasuo Ariki

Abstract

Recently, the accuracy of speaker-independent speech recognition has been remarkably improved by use of stochastic modeling of speech. However there has been very little research on orally-challenged people, such as those with speech impediments. In this paper, we discuss our efforts to build an acoustic model for a person with articulation disorders. The speaking style of a person with an articulation disorder differs considerably from that of a physically unimpaired person, and we need a feature extraction characteristic of a person with an articulation disorder. The sound spoken by a dysarthric speaker is less clear than a physically unimpaired person, and spectral information is buried in the sound. Therefore, we investigate a feature extraction method based on LPE (Local Peak Enhancement).