情報処理学会論文誌 Vol.61 No.10 1647–1657 (Oct. 2020)

音韻検査のための非語の音声認識に有効な特徴量の検討

多々納 俊治[1,a)] 縄手 雅彦[2,b)] 伊藤 史人[2,c)] 酒向 慎司[3,d)]

受付日 2019年12月24日, 採録日 2020年7月7日

概要：発達性ディスレクシアは学習障害の主要な症状の 1 つであり，早期発見は介入および読みの療育をするにあたって非常に重要である．PC を用いた簡易なスクリーニングテストが提案されており，テキストの読み上げの正確さ，単語の逆読み，文字の削除の能力についての評価データおよび回答潜時が自動的に記録される．しかし，正誤判定はテストの実施者によって行わなければならず，自動化が望まれている．正誤判定の部分を自動化するためには，検査の課題語にある意味を持たない言葉である非語に対応した音声認識技術が必要であるが，従来の音声認識では，非語に対する認識精度は低いのが現状である．そこで従来の音声認識の機能を補強しつつ，非語に対する正解率（accuracy）を音韻検査に実用できるレベルまで向上させなければならない．本研究では，ソースコードが無料で公開されており，自由に改造が可能な音声認識エンジン Julius に非語の正誤を判別する機構を組み込むことにより，非語に対する accuracy の向上を試みた．また，音声の特徴量に 7 つの候補をあげ，その組合せによる accuracy の動向を検討した．その結果，対象の非語によっては 75.0%から 95.0%，全体の平均値は 87.5%の accuracy を得た．

キーワード：音声認識，非語，機械学習，音韻検査

Study on Effective Combination of Features for Non-word Speech Recognition of Phonological Examination

Toshiharu Tadano[1,a)] Masahiko Nawate[2,b)] Fumihito Ito[2,c)] Shinji Sako[3,d)]

Received: December 24, 2019, Accepted: July 7, 2020

Abstract: Developmental dyslexia is a main element of learning disability and its early detection is very important for intervention and reading treatment. A convenient screening test using PC has been published and the answer times in text reading, reversed reading of word and mora skip of word are automatically recorded in the test. However, the correctness determination must be done by tester. In order to automate those test, a speech recognition technology corresponding to a non-word that are non-meaningful words used in an examination is necessary, but in conventional speech recognition, recognition precision for non-words is low. Therefore, while reinforcing the function of conventional speech recognition, the accuracy for non-words to be improved to a level that can be practically used for phoneme examination. In this study, we have tried to improve the accuracy for non-words by incorporating a mechanism to determine non-word correctness into Julius, which is in the public domain and can be modified freely. In addition, six candidates are given as feature quantities of speech, and the trend of the accuracy by the combination is examined. As a result, depending on the target non-word, the accuracy was 75.0% to 95.0%, and the overall average value was 87.5%.

Keywords: speech recognition, non-word, machine learning, phonological examination

1. はじめに

「読み書き困難児のための音読・音韻処理能力簡易スク

[1] 出雲医療看護専門学校
IZUMO COLLEGE OF MEDICAL NURSE, Izumo, Shimane 693–0001, Japan
[2] 島根大学
Shimane University, Matsue, Shimane 690–8504, Japan
[3] 名古屋工業大学
Nagoya Institute Technology, Nagoya, Aichi 466–8555, Japan

[a)] syungsan@road.ocn.ne.jp
[b)] nawate@ecs.shimane-u.ac.jp
[c)] fumi@ecs.shimane-u.ac.jp
[d)] s.sako@nitech.ac.jp

リーニング検査ソフト ELC」[4] は，ディスレクシアと呼ばれる読み書き困難児をスクリーニングする検査ソフトである．この検査ソフトは児童の読みに疑問を持った教師が，学校で簡便に使用できることを目的に開発されたものであり，PC を使用して検査全体を 10 分程度で行うことができる．検査内容は 3 部から構成されており，まず音韻特徴をとらえる課題として短文を読み上げさせ，その正確さを評価する「短文音読課題」，そして音韻意識を評価する課題として，呈示された言葉を逆から読ませたり，言葉を抜いて省略して読ませる「音韻操作課題（逆唱・削除）」，またデコーディングを評価する課題として「単語非語音読課題」を課すことになっている．

それぞれの検査は，検査時に児童の音声を録音したのち，検査者が，その録音内容から児童の音読・音韻処理能力について評価する．また，読みの能力や音韻操作の能力の評価は，流暢性が重要な指標となるため，回答潜時と呼ばれる課題提示から発話までかかった反応時間と，発話開始から発話終了まで時間を自動計測することが特徴となっている．しかし，現時点では 3 つの検査項目において，回答された音声の内容についての正誤判定は，検査者自身が行わなければならず，検査者の負担を減らし，正確な正誤判定を用いた信頼性の高い検査を効率良く実施するうえで，音声認識技術の活用による自動化が期待されている．

国外では，紙面による検査の代わりに非語の読み上げ音声評価の自動化を試みる研究はいくつか行われている [1], [2], [3]．いずれも，児童の読字能力の評価基準は ELC と同様に非語を対象とした読み上げの流暢性と発音の正確性の測定が主となっており，発話時間の測定と回答の正誤判定を音声認識により実現させる試みである．ただ，国外には読み書きに困難を示す児童の，状況に応じた発音のアノテーション付きのコーパスを収めたデータベースを保持する機関が存在しており，研究促進の要となっている場合があるが，日本国内にはなく，一般の音声認識が対象としない非語の音声認識を作成する場合は，学習用データをゼロから採取する必要がある．

ここで ELC の「音韻操作課題」で出題される 16 種類の言葉を**表 1** に示す．逆唱については左の列の言葉を逆さ読みした右の列の言葉を読み上げ，削除については左の列の言葉から文字を抜いた右の列の言葉を読み上げる内容である．回答として読み上げるべき言葉はすべて言語的に意味を持たない非語になっており，児童の音韻操作能力の特性をとらえやすい音素で構成されている．

ところで，最近のコンピュータを利用した音声認識の技術の進歩は目覚ましく丁寧に発声された音声の単語認識率は 95 %前後を達成している [5]．一般的な音声認識エンジンの構造は各単語とそれを発声したときの音声の特徴パラメータ系列との関係を表す音響モデルと単語の並びの制約を表す言語モデルからなり，観測された音響特徴パラメータに最も適合する単語列を膨大な検索空間上を検索することによって実現されている [5]．

表 1 ELC の音韻操作課題
Table 1 ELC phoneme manipulation task.

	逆唱		削除	
単語	あたま	またあ	ねくたい	ねたい
	かめら	らめか	たまねぎ	たねぎ
	たまご	ごまた	あさがお	あさが
	つくえ	えくつ	ひまわり	ひまり
非語	みしけ	けしみ	いそれす	いそす
	たぐめ	めぐた	なゆかた	なゆか
	たちの	のちた	ぶとみご	ぶとご
	せとく	くとせ	よですち	よすち

ところが，一般の音声認識で非語を対象にした場合，その単語認識率は 70%～80%にとどまり，単語に比べ低くなっている [5]．

これは一般の音声認識のほとんどが周辺の単語の出現頻度をみて回答にふさわしい言葉を探し出す大語彙連続音声認識であり，非語を対象とした認識手段は備わっていないことや，辞書には非語が最初から登録されていないことが起因している．仮に一般の音声認識エンジンに非語を新規登録したとしても，発音の類似したコーパスのなかの膨大な語のなかから影響を受けるため，登録した非語に近い言葉を誤認識する確率が高い．

また，モーラタイプライターを用いて，1 モーラずつを認識する場合，モーラの組合せで，あらゆる非語の認識に対応可能だが，モーラタイプライタは 1 モーラに対する単語認識率が低く，認識させたい言葉のモーラ数が増えるほど各モーラが同時に正解する同事象確率は低くなる．さらにまた，辞書に登録した孤立した単語のみを分類することを考えた場合においては，正解として認識させたくない間違い語を登録する必要がある．しかし，辞書にすべての間違い語を誤りデータとして列挙し登録するのことは困難である．

本研究においてまず，ELC の検査項目の回答の正誤を正確に判定するシステムを実現する場合，回答として読み上げるべき非語に対し，正しくいえたかどうかの「二値分類の認識率（正解率）」（accuracy）を高めなければならない．そのためには，一般の音声認識にはない回答として読み上げるべき非語のみに特化した音声認識の開発が必要である．

ところで，一般的に従来の音声認識は，音響モデルに DP マッチングを拡張した隠れマルコフモデル（HMM）と HMM の各状態の音響特徴量のパターンを連続分布でモデル化する混合ガウス分布（GMM）を導入した GMM-HMM と，言語モデルには，単語の連接規則（文法）をオートマトンで記述したものから，その遷移を確率的なものにし，その確率をコーパスから最尤推定する N-gram モデルを導

入したシステムが使われてきた [5], [6]．最近では，音響モデルについては，GMM による確率計算をディープニューラルネットワーク（DNN）に置き換えた DNN-HMM が，言語モデルについては，リカレントニューラルネットワーク（RNN）を N-gram と併用するモデルを使用することで，全体として単語認識率が向上している [6]．

このような理由から，本研究では音声認識器から得られる特徴量を用いて，各非語の二値分類器を構築した．

我々はこれまでフリーの音声認識エンジンである Julius [7] に機械学習のサポートベクタマシーン（SVM）を組み込む実験により，非語に対する accuracy の検証を行った [8], [9]．SVM による試行では音声認識器である Julius を特徴量生成器として使用し，Julius が内部的に作成する 25 次元の MFCC とパワー項，および尤度を SVM に特徴量として学習させ，カーネルトリックなしで単一境界面による線形分離を行った結果，全非語において平均 86%の accuracy を得た．

ELC の検査項目すべてにおいて平均 86%程度では実用性に欠ける．今回は別法として特徴量抽出から分類までの過程を近年深層学習の分野で高い識別能力を示しているニューラルネットワークで構築し実験を行った．同時に特徴量においては Julius の出力のみに頼らず，SVM のときに比べ，新たにフレーム数，DTW 距離，フォルマント，音素セグメンテーションの結果を追加し特徴量の要素を増やした．またニューラルネットワークの学習アルゴリズムを最適化することで，全般的な accuracy の向上を試みた．

ニューラルネットワークの基本構造は人間の脳内のニューロンを模したユニットを並列・多層に連結させたものであり，近年では大量の学習データが用意できる画像認識等の分野において，畳み込みニューラルネットワーク（CNN）を利用する等して高い accuracy を得た事例が報告されている [12]．

ニューラルネットワークを使用した音声認識では，まず波形そのものを学習させ，ネットワークに特徴量をすべて抽出させる End-to-End 学習による手法が考えられるが，学習に十分な量のデータを収集することが困難であることから，実現が難しい．また，時間軸方向のデータ量が多く高次元数のため計算コストが高くなり，現時点の一般のコンピュータでは処理を施すことに問題があると考えられる．そこで，計算負荷を軽減することを考えるとともに，少量の学習データでも弁別に有効な特徴量の候補をいくつかあげ，それらを組み合わせたものをコンピュータにあらかじめ与え，そのなかで accuracy の最も高かったものを選抜することにする．

特徴量の候補については，従来の一般的な音声認識に使われていた MFCC が非語の認識に有効であると考えられるが，それ以外の特徴量についても十分検討されていたとはいえないため，本研究では非語の音声認識に有効な特徴量について MFCC を含め，6 つの特徴量の候補について検討する．

具体的に，本研究では新たに，Julius の出力する MFCC の時間系列について，見本のものと学習データのものとを伸縮同期させ波形の類似度を算出する手法である動的時間伸縮法（DTW：Dynamic Time Wraping）[13] を行った結果と，Julius のセグメンテーションキットによる音素セグメンテーションの結果，およびフォルマント [14]．をそれぞれ算出し特徴量として利用することを提案する．

2. 目的

本研究は ELC の検査項目の正誤判定を自動化するために，音声認識の非語に対する精度を向上させることを目的とする．そのために既存の音声認識エンジン Julius が出力した音声特徴量を利用し，外部に機械学習を組み込むことで，結果を 2 値分類させる．また，提案した特徴量である DTW の結果と音素セグメンテーションの結果，およびフォルマント，そのほか，合わせて 6 つの特徴量について，同じ構造のニューラルネットワーク内で，それぞれの組み合わせを交互に学習させた場合に accuracy がどのように変化するかを検証し，非語の音声認識に有効な特徴量を検討する．

3. 方法

本研究の実験には**図 1** のような入力層と中間層そして出力層からなる単純 3 層パーセプトロン（MLP）を機械学習のネットワークとして非語ごとに学習して使用する．ユニット数は，入力層を選択した特徴量に応じて 1 から全特徴量の次元の総和である 110 までの可変長とし，中間層を 128，出力数を正誤 2 値分類のために 1 とし，すべてのユニットを全結合したものに対し，活性化関数に ReLU 関数を適用した．ただし最終出力はシグモイド関数を使用した．また各層間に Dropout を破棄率 0.15 で挟み込み，損失関数に Binary Cross Entropy，最適化関数に adadelta を使用した．また計算の高速化を図るため計算は GPU で行った．

まず，非語の音声を，量子化ビット数を 16 bit，サンプリングレートを 16 kHz の WAVE ファイルとして，大学生 21 歳前後の男女 12 人から，正しい発音のデータを 10 サンプル，および間違いの発音データを 40 サンプル程度採取し，データクレンジングを施した．このデータクレンジングと，のちの音声データの追加等により各非語のデータ数にばらつきが生じたが，すべてのサンプルに正誤の教師用のラベルを振って用意した．そのうち各非語において 10 個ずつの正解と不正解のデータを計 20 個，各話者のものを均等にテストデータに回した．差し引いた各非語の訓練データ数の内訳は表 4 に示す様である．その際，間違いデータの採取については，間違い文字列生成プログラムを

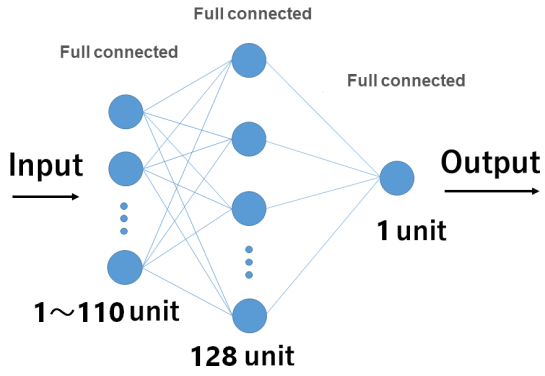

図 1 作成したネットワーク
Fig. 1 Created network.

別途作成し，その出力を参考に発音してもらい，音声データを採取した．

間違い文字列生成プログラムは，正解の 3 文字からなる非語の文字列に対して，先頭 2 文字が一致する 3 文字列と，先頭 1 文字が一致する 3 文字列と，または，そのほかのまったく異なる並びの 3 文字列の順番で優先順位を付け，優先順位の高い文字列をより多く生成するように，重みつき乱択アルゴリズムで実現している．

次に同じ学習データを使用し，MFCC とパワー項，音響モデルと言語モデルの尤度を学習させた後，新たに提案した特徴量の DTW の結果と音素セグメンテーションによる結果，およびフォルマントを含めて計 6 つの特徴量について，学習時の accuracy の推移グラフを参照し，最適な組み合わせを Forward Stepwise Selection（FSS）手法により選抜する．

FSS 手法とは，最初に 1 特徴量のみを使って評価を行い，最も accuracy の高いものを選ぶ．次に，残りの特徴量から 1 つずつを選び，最初に選択した特徴量と組み合わせた 2 特徴量による最適な組合せを探す．そしてその次は選択した 2 特徴量＋他の 1 特徴量の組合せを探すという流れで，有効な特徴量の組合せを調べる手法である [15]．

3.1 特徴量について

音声の特徴量の候補は以下の 6 つをあげる．

(1) 39 次元の MFCC とパワー項の各次元についての総フレーム長の平均値．
(2) 文法モード Julius に非語を登録したときの，それに対する尤度．
(3) 1/100 秒の解像度に設定した場合の音声の総フレーム数．
(4) 正解の見本音声の MFCC に対して次元ごとに計算した DTW 距離．
(5) 非語の各音素についての第 1 と第 2 フォルマント．
(6) Julius のセグメンテーションキットによる各音素の発話開始フレームと終了フレーム数と各音素についての尤度を並べたもの．

以降，各特徴量について詳しく説明する．

3.1.1 MFCC

MFCC はメル周波数ケプストラム係数（Mel-Frequency Cepstrum Coefficients）と呼ばれ，人間の声道特性および人間の音声知覚を表す特徴量であり，音声認識で使用する特徴量としては，一般的なものである．実験には MFCC を 12 次元および Delta 量の 12 次元，DeltaDelta 量の 12 次元にそれぞれの Power 項を含めた計 39 次元の各次元について発話全体（総フレーム）の平均値を 1 つずつ使用する．

3.1.2 尤度（スコア）

文法モードの Julius の辞書に対象の非語を登録した場合に Julius が出力する学習データの正解文字列に対しての音響モデルと言語モデルの尤度（スコア）を 1 次元として使用する．

3.1.3 フレーム数

時間軸方向の尺度としてタイムシフト幅（解像度）を 1/100 秒を単位として設定した場合の学習データの音声フレーム数を 1 次元として使用する．

3.1.4 DTW 距離

DTW は 2 つの時系列データの各点の距離を比較したうえで，2 つの時系列データがなるべく重なり合うように伸縮して，2 つの時系列データの最小距離を求めるアルゴリズムであり，2 つの時系列の長さが異なっていても距離を求められるのが特徴である．距離が近いほど 2 つの時系列データは一致度が高いことになるので，この距離は類似度を表す指針となる．ここで時系列データ x と時系列データ y の伸縮調整後の時刻 k におけるアライメントを $\omega_k = (\omega_k^x, \omega_k^y)$ と置いた場合，DTW による時系列データどうしの距離は距離尺度に L1 ノルム（マンハッタン距離）を利用する場合に式 (1) で定義される．

$$\mathrm{DTW}(x, y) = \min \sum_{k}^{K} |x_{\omega_k^x} - y_{\omega_k^y}| \tag{1}$$

たとえば，式 (1) にて計算した「つくえ」の逆唱である「えくつ」についての正解データと不正解データの 2 次元目の MFCC に対して，非語の正しい見本の時系列データとの DTW の結果をグラフで見ると，**図 2**，**図 3** のようになり，正解のデータに対する DTW 距離で一致度の高いほうの最隣接距離は約 0.09686 となり，不正解のデータに対する DTW 距離で一致度の低いほうの最隣接距離は約 0.16794 となる．グラフからも，正解の包絡と不正解の包絡とでは形状に一致度の差があることが確認できる．このようにして，すべての学習データの MFCC の 39 次元に対し，非語の正しい見本の時系列データと伸縮同期させた場合の最隣接距離を次元ごとに計算したものを特徴量とする．したがって次元数は MFCC と同じく 39 次元となる．なお，見本の音声は，音声合成ソフトにより生成したもの

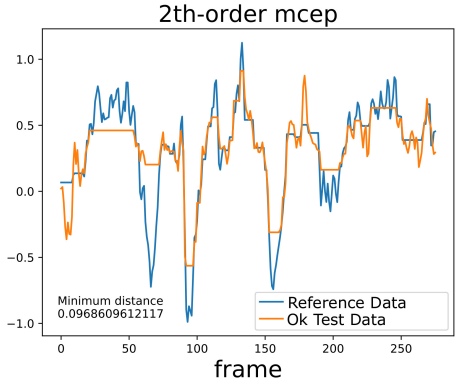

図 2 「えくつ」についての正解のデータに対する DTW の結果
Fig. 2 DTW results of correct data about “Ekutsu”.

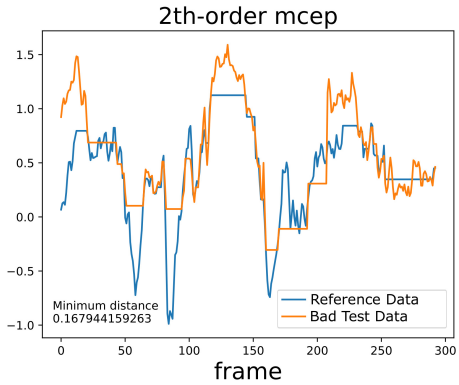

図 3 「えくつ」についての不正解のデータに対する DTW の結果
Fig. 3 DTW results of incorrect data about “Ekutsu”.

を使用した．

3.1.5 音素セグメンテーション

音素セグメンテーションの取得については Julius のセグメンテーションキットを利用する．セグメンテーションキットは，用意した音声データの発話内容に相当する文字列情報をテキスト形式で与えた場合，その音声を構成する音素それぞれの音声波形の振幅と零交差の回数に基づき算出された発話開始時間と発話終了時間の音声有効区間の分割フレーム，ならびに，それぞれの音素に対しての音響モデル，言語モデル（スコア）を抽出する機能を備えている．たとえば，「つくえ」の逆唱である非語の「えくつ」の音素列を正解のテキストの見本として与えた場合について，正解の「えくつ」をセグメンテーションキットに入力した場合の出力結果は**表 2** のようになっている．また「えくつ」を正解とした場合に不正解である「またあ」をセグメンテーションキットに入力した場合の出力結果は**表 3** のようになる．各音素に関してのスコアは正解の場合が大きく，不正解の場合は小さくなる．

「えくつ」は母音と子音の音素に分割され，それぞれの音素について発音開始時間と発話終了時間，および音響モデルと言語モデルのスコアが情報として付与されている．

表 2 「えくつ」についての正解音素セグメンテーションの結果
Table 2 Results of correct phoneme segmentation about “Ekutsu”.

開始時間	終了時間	スコア	発話音素
0.0000000	0.8925000	−19.671556	silB
0.8925000	1.0225000	−27.144936	E
1.0225000	1.1025000	−26.987701	k
1.1025000	1.2425000	−26.617083	u
1.2425000	1.3525000	−26.231667	ts
1.3525000	1.5225000	−23.172693	u
1.5225000	2.5925000	−19.618814	silE

表 3 「えくつ」についての不正解音素セグメンテーションの結果
Table 3 Results of incorrect phoneme segmentation about “Ekutsu”.

開始時間	終了時間	スコア	発話音素
0.0000000	0.3425000	−23.469467	silB
0.3425000	0.3725000	−27.347067	E
0.3725000	0.4025000	−30.164429	k
0.4025000	0.6625000	−24.368580	u
0.6625000	0.7225000	−27.699484	ts
0.7225000	1.2825000	−23.481785	u
1.2825000	2.2925000	−23.492867	silE

これらの非語の正解テキストを見本とした場合の採取音声データの音声区間とスコアをあわせたセグメンテーションキットの出力値を，横 1 列に並べたものをすべての学習データから特徴量として抽出した．また，各音素についてのセグメンテーションをとるので，対象の非語の音素数は 5 か 6 であり，次元数は少なくて (音素数 5 × 開始及と終了フレーム数の 2)+ 各音素のスコアの 5 = 15 次元となる．また多くて (音素数 6 × 開始及と終了フレーム数の 2) + 各音素のスコアの 6 = 18 次元となる．

3.1.6 フォルマント

フォルマントについては，まず各音素区間の $y_0, \cdots, y_n$ からなるスペクトル包絡 y_n を直前の y_i $(i = 0, \cdots, n-1)$ を使って，前向き線形予測の式 (2) で近似した後，Linear Prediction Coding (LPC) 係数 a_i をレビンソン・ダービン再帰法のアルゴリズムによって算出したものを利用する．実際には式 (2) の伝達関数である式 (3) の $E(z)$ を人間の声道の共鳴する特性をモデル化したものとして定義し，その周波数特性をフォルマントとした．

$$y_n \approx -\sum_{i=1}^{k} a_i y_{n-i} \tag{2}$$

$$\frac{Y(z)}{E(z)} = \frac{1}{1 + \sum_{i=1}^{k} a_i \exp(-ij\omega)} \tag{3}$$

$(z = \exp(j\omega)$，$(\omega = 2\pi f$：角周波数$))$

フォルマントを採取する各音素区間の特定についてはセグメンテーションキットの出力結果である音素ごとの分

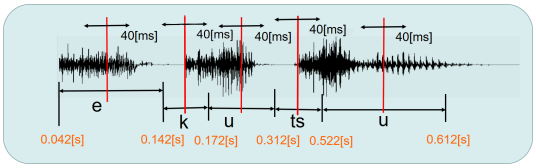

図 4 フォルマントの算出方法
Fig. 4 Formant calculation method.

割フレームを利用した．波形データ中の音素 1 つ 1 つは，**図 4** のようにセグメンテーションキットの出力結果によって発話区間で分割することができる．そのうち各音素の発話区間の中心に各音素の発音強度のピークがあると仮定し，そこから左右両端へ幅 40 ms の区間で前向き線形予測分析し，第 1 フォルマント（F1 値）と第 2 フォルマント（F2 値）を算出する過程をすべての学習データについて施した．次元数は各非語の音素数に依存し，対象の非語の音素数は 5 か 6 であるので次元数は少なくて音素数の 5 × 第 1 フォルマントおよび第 2 フォルマントの数の 2 = 10 次元である．また多くて音素数の 6 × 第 1 フォルマントおよび第 2 フォルマントの数の 2 = 12 次元である．

3.2 accuracy について

本研究で使用する accuracy とは，以下のような式で定義する．

$$accuracy = \frac{\text{正しく分類できた数}}{\text{総学習データ数}} \tag{4}$$

3.3 学習の流れ

採取した各非語の音声データからテストデータとして正解と間違いをそれぞれ 10 個ずつで 20 個，各話者別に均等に選ぶ．そして残りの音声データの 10%を検証データに，残りの 90%を訓練データとして使用する．そして候補にあげた 6 つの特徴量を学習データから抽出する．その後，すべての学習データごとの特徴量に Z-score 関数，式 (5) による標準化を施したものを学習させ，FFS 手法による検証を行う．

$$z_i = \frac{x_i - \mu}{\sigma} \qquad (\mu : \text{平均}, \sigma : \text{標準偏差}) \tag{5}$$

1 回の学習ごとに検証データについての accuracy を評価し，その推移を記録していく．またシステムの学習方式は訓練データ 1 つの学習を 1 回行うたびにパラメータを更新する逐次学習法（オンライン学習法）ではなく，すべての学習データをまとめて学習した後でパラメータを更新する試行を 1 epoch とし，それを 300 epoch まで行った．以上のような試行を 10 分割クロスバリデーション検証を用いて行った．そして最後にテストデータによる評価を行い，それぞれの特徴量の組合せので accuracy の推移に，どのような差異がでるか結果の数値とグラフから比較検証し，考察を行った．

3.4 追加実験

以下の追加実験を行った．
(1) 間違いデータの割合による認識率の違いの検証
(2) 話者クローズの実験
(3) 全非語のデータ数を揃えた場合の実験
これらの実験の手法，結果および考察は 6 章に詳しく述べる．

4. 結果

4.1 FSS 手法による結果

FSS 手法を用いて各非語の最適な特徴量の組合せを導いた．「えくつ」について各試行で得た最適な特徴量の組合せを順番で並べていくと次のようになった．ここで各試行における評価基準は，テストデータによる accuracy の高さとし，accuracy が同率の場合は loss を用いて比較し，その値が低いほうを選択した．

(1) 第 1 試行での特徴量の最適な組合せ：Segmentation，78.95%
(2) 第 2 試行での特徴量の最適な組合せ：Segmentation+MFCC，89.47%
(3) 第 3 試行での特徴量の最適な組合せ：Segmentation+MFCC+Frame，89.47%
(4) 第 4 試行での特徴量の最適な組合せ：Segmentation+MFCC+Frame+Score，95.00%
(5) 第 5 試行での特徴量の最適な組合せ：Segmentation+MFCC+Frame+Score+DTW，78.95%
(6) 第 6 試行での特徴量の最適な組合せ：Segmentation+MFCC+Frame+Score+DTW+Formane，68.42%

ここで「えくつ」においては，途中で最も accuracy の高かった [Segmentation+MFCC+Frame+Score] を最適な特徴量の組み合わせとして採択した．

4.2 各非語の accuracy の推移

学習対象の 8 つの非語のうちの 1 つである「つくえ」の逆唱の「えくつ」についての学習結果を**図 5**，**図 6**，**図 7**，**図 8** に示す．図には FSS 手法により採択した上位 3 つの特徴量がプロットされている．「えくつ」は音素セグメンテーションと MFCC に付け加えフレーム数とスコアの組合せで最高の accuracy となった．

次に「あたま」の逆唱の「またあ」についての学習結果を**図 9**，**図 10**，**図 11**，**図 12** に示す．「またあ」は MFCC とフレームのみの組合せで最高の accuracy となった．

そのほかの残り合わせ計 6 つの非語の学習結果については，**表 4** に示すように非語の種類によって結果に差異が出たが，すべての非語それぞれについての最適な特徴量の組合せで 75.0%から 95.0%，全体の平均値は 87.5%の accuracy を得た．

またすべての非語についての学習結果の評価は F1 Score

表 4 全非語の学習の結果

Table 4 Results of learning all non-words.

学習対象	訓練データ比率（正答：誤答）	Epoch of MLP	Accuracy by MLP	Loss by MLP	MLP における特徴量の組合せ	Epoch of SVM	Accuracy by SVM
えくつ	189:290	300	95.00%	0.22	Segmentation+MFCC+Frame+Score	48	88.00%
ごまた	29:68	300	95.00%	0.19	MFCC+Frame+Segmentation	43	98.67%
けしみ	31:33	300	75.00%	0.69	Frame+Score	43	96.00%
くとせ	28:48	300	90.00%	0.42	MFCC+Frame+Score	43	85.33%
またあ	79:91	300	90.00%	0.41	MFCC+Frame	43	56.00%
めぐた	38:47	300	75.00%	1.05	MFCC+Segmentation+Frame+Score	43	97.33%
のちた	23:25	300	95.00%	0.24	Segmentation+Frame+Score	43	88.00%
らめか	27:36	300	85.00%	0.37	Segmentation+Frame	43	78.68%

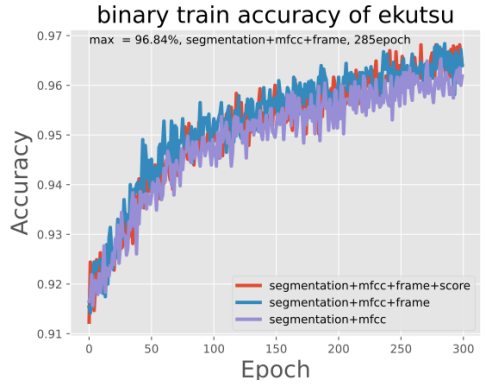

図 5 「えくつ」についての訓練 accuracy の結果

Fig. 5 Results of training accuracy about “Ekutsu”.

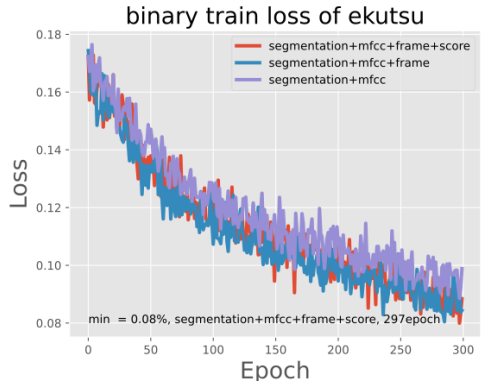

図 6 「えくつ」についての訓練 loss の結果

Fig. 6 Results of training loss about “Ekutsu”.

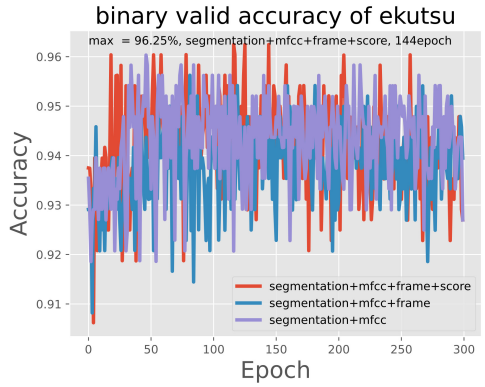

図 7 「えくつ」についての検証 accuracy の結果

Fig. 7 Results of verification accuracy about “Ekutsu”.

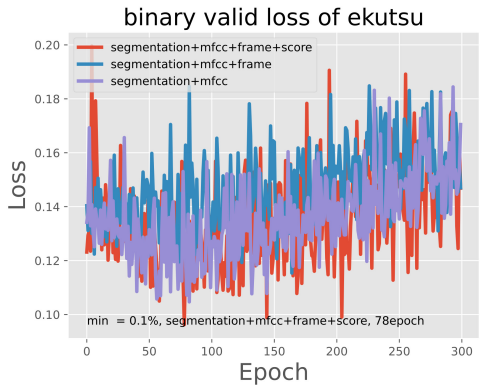

図 8 「えくつ」についての検証 loss の結果

Fig. 8 Results of verification loss about “Ekutsu”.

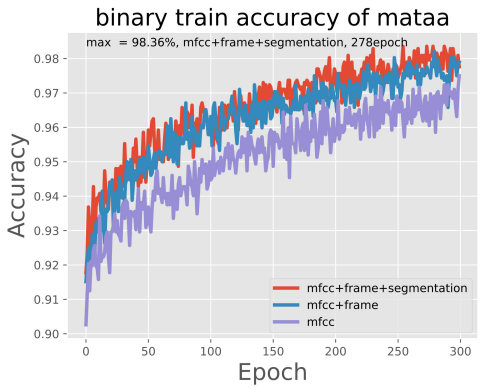

図 9 「またあ」についての訓練 accuracy の結果

Fig. 9 Results of training accuracy about “Mataa”.

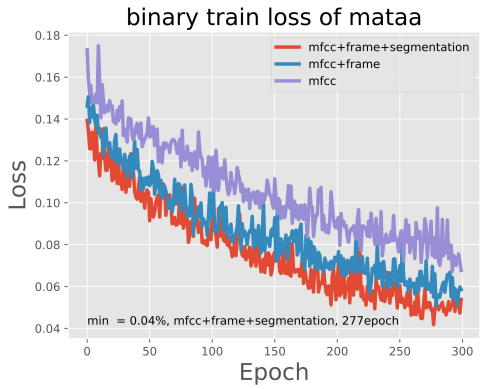

図 10 「またあ」についての訓練 loss の結果

Fig. 10 Results of training loss about “Mataa”.

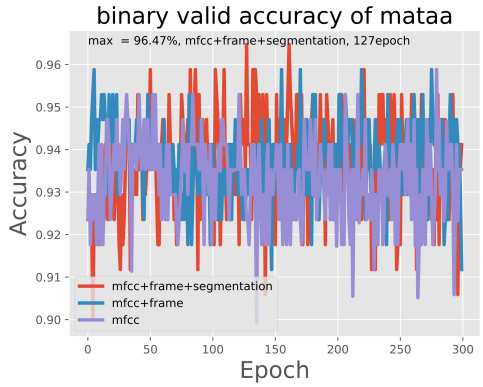

図 11 「またあ」についての検証 accuracy の結果
Fig. 11 Results of verification accuracy about “Mataa”.

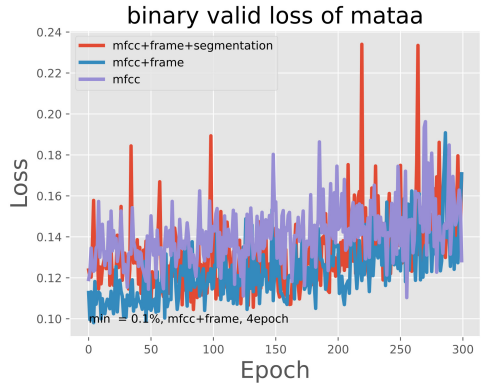

図 12 「またあ」についての検証 loss の結果
Fig. 12 Results of verification loss about “Mataa”.

表 5 F1 Score 等による比較
Table 5 Comparison by F1 score etc.

学習対象	Precision	Recall	F1 Score
えくつ	1.00	0.90	0.95
ごまた	1.00	0.90	0.95
けしみ	0.73	0.80	0.76
くとせ	0.90	0.90	0.90
またあ	1.00	0.80	0.89
めぐた	0.86	0.60	0.71
のちた	1.00	0.90	0.95
らめか	0.89	0.80	0.84

等でも行った．**表 5** にそれを示す．

5. 考察

SVM による実験結果と比較すると，今回の実験結果は 8 つの非語について accuracy の最大値は低いが，平均な accuracy は SVM の 86%より 87.5%と高い値が得られた．

5.1 各特徴量の寄与について

各非語によって最適な特徴量の組合せに違いがでたが，いずれの場合にも MFCC もしくは音素セグメンテーションが含まれている．

MFCC は一般的に音声認識に使用するうえで有効とされる既知の特徴量であることから，相応の結果が出たと考えられる．

音素セグメンテーションが accuracy の向上に寄与した理由としては，音素セグメンテーションは個々の音素に対する発話時間と言語および音響モデルの尤度情報が分割して含まれており，各音素を特徴付ける量となっている．またそれらは，時間軸方向に区分された音素有効区間情報であり，時系列の情報を付け加えた指針として accuracy の向上に効果的な特徴量になったと考えられる．

また，FSS 手法による特徴量選択の過程の観察から，スコアとフレーム数は音素セグメンテーションまたは MFCC による accuracy に少量のポイントを加える役割を果たしていると考えられる

一方，DTW 距離，およびフォルマントはいずれの特徴量の組合せにも含まれなかった．

DTW 距離は学習データにおける MFCC の時系列波形の見本との最隣接距離であるが，それは 1 カ所の限定的な位置情報であり，時間変化する音声全体の特徴をとらえる量ではないと考えられる．また，今回見本の音声は音声合成ソフトで生成したものであり，その品質の良しあしで，結果が相対的に変化する可能性あり，特徴量としては適切ではないと考えられる．

フォルマントは，非語における各音素の推定値であり，DTW と同様に音声全体の特徴をとらえる量ではないと考えられる．またセグメンテーションキットによる音素区間の特定を利用したフォルマント検出範囲が 40 ms では不十分な可能性もある．

5.2 各非語の学習データ数と結果の相関

非語の種類により accuracy の結果に差異が生じた原因は，学習データ数の偏りが原因である可能性が大きい．収集数に不均衡なデータがある場合，認識の結果はデータの多いクラスのほうに引っ張られる傾向があるためである．

そこでまず，Precision（適合率），Recall（再現率），および F1 Score（F 尺度）を計算したところ，「めぐた」，「けしみ」においては 3 つの値とも 0.8 をきっているが，均等な分布であり，その他の非語についても 3 つの値とも 0.9 程度でバランスのよい分布を維持していることが分かった．このことから，今回の実験結果は学習データの数の正誤の数の内訳の違いにより，正誤どちらかに傾いた認識をするものではないということが考えられる．一方，学習データ数と accuracy の相関係数を計算したところ 0.38 と，小さな正の相関があった．このことから，各非語の accuracy が学習データ数に依存していると考えられる．不均衡データについての検証は，6 章の追加実験にて詳しく行った．

表 7 「えくつ」における話者クローズの実験結果
Table 7 Experiment results of speaker close in “Ekutsu”.

学習対象	訓練データ比率（正答：誤答）	Epoch	Accuracy	Loss	Precision	Recall	F1 Score
えくつ	931:599	300	94.00	0.92	0.52	0.94	0.94

表 8 全非語のデータ数を揃えた学習の結果
Table 8 Results of learning with same number of non-word data.

学習対象	訓練データ比率（正答：誤答）	Epoch	Accuracy	Loss	特徴量の組合せ	Precision	Recall	F1 Score
えくつ	23:25	300	60.00%	1.27	Segmentation+MFCC+Frame+Score	0.57	0.80	0.67
ごまた	23:25	300	70.00%	11.33	Segmentation+MFCC+Frame+Score	0.64	0.90	0.75
けしみ	23:25	300	65.00%	1.31	Segmentation+MFCC+Frame+Score	0.64	0.70	0.67
くとせ	23:25	300	80.00%	0.55	Segmentation+MFCC+Frame+Score	0.80	0.80	0.80
またあ	23:25	300	70.00%	3.24	Segmentation+MFCC+Frame+Score	0.67	0.80	0.73
めぐた	23:25	300	65.00%	2.94	Segmentation+MFCC+Frame+Score	0.71	0.50	0.59
のちた	23:25	300	85.00%	0.31	Segmentation+MFCC+Frame+Score	0.89	0.80	0.84
らめか	23:25	300	75.00%	1.08	Segmentation+MFCC+Frame+Score	0.78	0.70	0.74

表 6 「えくつ」における間違いデータの割合と精度
Table 6 Percentage and accuracy of incorrect data in “Ekutsu”.

条件	Accuracy	Loss	Precision	Recall	F1 Score
すべて既知	80.00	0.92	0.75	0.90	0.82
すべて未知	95.00	0.19	1.00	0.90	0.95

6. 追加実験

6.1 間違いデータの割合による認識率の違い

実験で使用したテストデータのうち，10 個は間違いの発音データであるが，それが訓練データに存在する既知のものか，まったく新しい未知のものかで，全体の認識率に違いが生じるはずである．そこで「えくつ」に関して間違いのテストデータの数を確認したところ，そのうち 8 個が訓練データの発音内容と一致するものであった．間違いのテストデータの 8 割が訓練データに存在する既知のものである場合の「えくつ」の accuracy は 95.00%である．ここで間違いのテストデータのすべてが訓練データに存在する既知のものである場合と，すべてがまったく違う未知のものである場合の実験を「えくつ」について行い，結果を比較した．学習モデルは FSS 手法により得られた最良の特徴量の組み合わせ（Segmentation+MFCC+Frame+Score）で構築されたものを使用した．結果は**表 6** のようである．本来，テストデータが既知である割合が大きいほうが精度が高く，テストデータが未知である割合が大きいほうが精度が低くなると予測されるが，実験結果はそれとは逆となった．未知のデータに対し精度が高いことは，ロバスト性が高いと評価できる．

6.2 話者クローズの実験

データ数の一番多い「えくつ」について新たに同一人物の発話を採取して，データ数を増やし，話者クローズの実験を行った．また，テストデータも同じ話者のもののみ正解と不正解を 100 個に増やした場合の実験結果を**表 7** に示す．

話者オープンのときよりも accuracy が低下し，また F1 Score，Precision，Recall も低下しアンバランスになった．

データ数とテスト条件が異なるため単純な比較はできないが，本来ならば，汎用性が低くなるが，訓練データとテストデータの特徴の一致率がより高い話者クローズの精度が大きく，汎用性は高いが，訓練データとテストデータの特徴の一致率がより低い話者オープンの精度が小さくなることが予測されるが，実験では逆の結果となった．

これらは，データ数の少なさと実験条件の違いが影響したと考えられる．

6.3 全非語のデータ数を揃えた実験

クラスのなかでもデータ数が最も小さいものに全クラスのデータ数を揃え（正解データ数：23 個，不正解データ数：25 個），全クラスのデータ数が同じ場合の実験を行った．結果は**表 8** のようである．accuracy はすべての非語において低下し，Presicion，Recall，F1 Score ともに低く，アンバランスなものになった．全体的な accurcy の低下と各非語による差は，前の実験に比べ学習データ数を減少させたことが原因と考えられる．また，本来，全クラスでデータ数を揃えた場合 Presition，Recall，F1 Score がバランス良く向上することが予測できるが，データ数を全クラスで揃えたのにかかわらず，結果がアンバランスに低下した理由も，データ数が少ないためだと考えられる．追加実験においては学習データの少なさが影響し，評価結果が予測と異なった．さらに学習データの採取が必要であると考えられる．

7. まとめ

ELC の検査項目を自動化するために音声認識の非語に

対する accuracy を向上させるべく Julius を特徴量生成器として利用し，ニューラルネットワークによって 2 値分類する実験を行った．

特徴量として一般的な MFCC に加え，DTW を利用した見本音声との DTW 距離と，セグメンテーションキットを利用した音声有効区間情報と音素ごとのスコア，およびフォルマントを含め 6 つの音声特徴量から学習に有効な組み合わせを FSS 手法により検証した．

その結果，特徴量の組合せは音素セグメンテーションと MFCC が有効であり，それらを含めた組合せで，非語によっては 75.0%から 95.0%，全体の平均値は 87.5%の accuracy を得た．

また DTW 距離とフォルマントは本実験では特徴量として適切ではないことが分かった．

今回の実験データやモデル構造では，音声特徴量を与えず，機械にすべての学習を行わせる場合よりも，手動で音声の弁別に優れた特徴量を与えた場合に accuracy は向上することが判明した．

また，本研究で使用したニューラルネットワークは全結合 3 層構成の単純なものであった．今後は時系列データを取り扱うためのリカレントニューラルネットワークを使用し，accuracy が向上するかどうかを検討する．また，学習データ数を増やす必要もある．

謝辞 音声データの収集に協力していただいた「山陰発達障害当事者会スモステの会」の方々，ならび島根大学教育学部のボランティアの方々に御礼申し上げます．

参考文献

[1] Duchateau, J., Cleuren, L., Van Hamme, H. and Ghes, P.: Automatic Assessment of Children's Reading Level, *Proc. Interspeech 2007* (2007).

[2] Black, M.P., Membe, S., Tepperman, J. and Narayanan, S.S.: Automatic Prediction of Children's Reading Ability for High-Level Literacy AssessmentAutomatic evaluation of reading aloud performance in children, *IEEE Trans. ASLP*, Vol.19, No.4 (2011).

[3] Proença, J., Celorico, D., Candeias, S., Lopes, C. and Perdigão, F.: Children's Reading Aloud Performance: A Database and Automatic Detection of Disfluencies, *Proc. INTERSPEECH*, pp.1655–1659 (2015).

[4] 加藤醇子，安藤壽子，原 恵子，縄手雅彦：読み書き困難児のための音読・音韻処理能力簡易スクリーニング検査 ELC Easy Literacy Check，図書文化 (2016).

[5] 中川聖一：ここまでできる音声ドキュメント処理技術，平成 24 年度電気関係学会東海支部連合大会，S3-1 (2012).

[6] 河原達也：[特別講演] 音声認識技術の展開，IEICE Technical Report：信学技報，Vol.115，No.388，pp.111–116 (2015).

[7] 名古屋工業大学：Julius now on GitHub，入手先 ⟨http://julius.osdn.jp/⟩.

[8] 多々納俊治，縄手雅彦，伊藤史人，酒向慎司：外部学習ライブラリによる Julius の精度向上の研究，ヒューマンインタフェースシンポジウム 2016 (2016).

[9] 多々納俊治，縄手雅彦，伊藤史人，酒向慎司：SVM による非語の正誤判定を用いた音韻検査の自動化の検討，HCG シンポジウム 2016 (2016).

[10] National Taiwan University: LIBLINEAR - A Library for Large Linear Classification, available from ⟨https://www.csie.ntu.edu.tw/~cjlin/liblinear/⟩

[11] 斎藤康毅：ゼロから作る Deep LearningDeep - python で学ぶディープラニングの理論と実装，オーム社 (2016).

[12] 藤吉弘亘，戸田智基：音・画像情報処理における特徴表現-まとめ-，入手先 ⟨http://vision.cs.chubu.ac.jp/~hf/MIRU_KIKU2016_Fujiyoshi_Toda.pdf⟩.

[13] haripo：Dynamic Time Warping（動的時間伸縮法），入手先 ⟨https://haripo.com/articles/2017/dynamic-time-warping/⟩.

[14] 加藤重弘，安藤智子：音声学講義，研究社 (2016).

[15] Tang Ba Nhat, 目良和也, 黒澤義明, 竹澤寿幸：音声に含まれる感情を考慮した自然言語対話システム，Human-Agent Interaction Symposium 2014，P-11, p.11 (2014).

[16] 荒木雅弘：フリーソフトでつくる音声認識システム パターン認識・機械学習の初歩から対話システムまで，森北出版 (2007).

[17] aidiary：人工知能に関する断創録，入手先 ⟨http://aidiary.hatenablog.com/entry/20150310/1425983455⟩.

多々納 俊治

2013 年放送大学大学院自然環境科学プログラム修了．2016 年出雲医療看護専門学校非常勤講師．2017 年島根大学大学院総合理工学研究科博士後期課程単位取得退学後現在に至る．

縄手 雅彦 （正会員）

1987 年広島大学大学院工学研究科修了．1987 年名古屋大学助手．1990 年広島大学助手．1996 年島根大学助教授．2007 年同教授，現在に至る．2003 年より福祉情報工学の研究を始め，近年は発達障がい児の学習支援ソフトの開発を中心に障がい児・者支援への ICT 活用のための研究に従事．電子情報処理学会発達障害支援研究会委員長．工学博士．

伊藤 史人 （正会員）

2011 年岩手県立大学ソフトウェア情報学研究科修了．一橋大学情報基盤センターを経て現在に至る．重度障害者のコミュニケーション支援技術に携わる．近年は，視線入力訓練ソフトウェアやバリアフリーマップアプリの開発に注力．博士（ソフトウェア情報学）．

酒向 慎司 （正会員）

2004 年名古屋工業大学大学院電気情報工学専攻博士後期課程修了．同年東京大学大学院情報理工学研究科特任助手．2007 年名古屋工業大学大学院情報工学専攻助教．2017 年同准教授．博士（工学）．音楽情報処理，音声情報処理，手話認識の研究に従事．2009・2016 年度日本情報処理学会山下記念研究賞，2011 年度電子情報通信学会ヒューマンコミュニケーション賞受賞．電子情報通信学会，日本音響学会，人工知能学会，IEEE 各会員．