Vol.2018-SLP-123 No.3
2018/7/27

# シャドーイング音声自動評価における耐雑音化と回帰を用いた高精度化

梶島 優[1,a)]　齋藤 大輔[1,b)]　峯松 信明[1,c)]　山内 豊[2,d)]　伊藤 佳世子[3,e)]

**概要**：外国語音声教育の現場では，多人数で同時に学習活動を行なうことがしばしばある．音声処理技術を用いた教育支援を検討する場合，そのような環境で収録された音声にはバブルノイズが含まれることが多く，問題になることがある．本研究ではシャドーイング音声を対象とした自動評価において，耐雑音性能向上のために次の検討を行った．1）ネイティブ英語音声にバブルノイズを重畳して音響モデルを構築（マルチコンディション学習），2）ノイズを含んだシャドーイング音声を用いた音響モデル適応，である．その結果，音素事後確率に基づく GOP スコアを用いた手法では前者の場合で精度が向上し，提示音声とシャドーイング音声間の DTW を用いた手法では後者の場合で精度が向上した．また，各種特徴量の追加と回帰モデルの導入により手動評価と同等程度の精度を達成した．

## 1. はじめに

シャドーイングとは，聞こえてくる音声を即座に復唱する行為である．リスニングとスピーキングを同時に行い，かつ意味理解を伴うことから，第 2 言語学習において高い学習効果がある．例えば，多読・読み上げ・リスニングなどの学習法に比べ，シャドーイングがより効果的な学習法であると報告されている [1], [2]．様々な実験でシャドーイングが学習者のリスニング力や理解度を向上させることが報告されているが，いずれも一ヶ月程度の期間継続的に学習を行った場合に効果が確認されている [3]．また，シャドーイングはビジネス英会話スクールなどでも導入されており，毎日シャドーイングを続けることがカリキュラムとして組み込まれている [4]．

このように，より良い学習効果を得るには継続的な学習が重要であると考えられる．しかしながら，学習者は自身のシャドーイングが上達しているか判断することができず，モチベーションを保つことが難しい．シャドーイング音声を自動で評価し，学習のモチベーションを向上させることが求められる．

我々はこれまで，シャドーイング音声の手動評価スコアと相関の高い素性として GOP(Goodness of Pronunciation) を取上げ，HMM(Hidden Markov Model) を用いた推定 [5]，DNN(Deep Neural Network) を用いた推定を検討し [6]，TOEIC スコアや手動スコアとの相関分析によりその効果を示してきた．更には，シャドーイング音声中の単語脱落の検出 [7] などを行い，その結果と HMM-GOP を素性とした回帰問題として，手動スコアの自動予測を検討した．

[6] では DNN-GOP 以外にも，モデル話者音声と学習者のシャドーイング音声の比較を，両者を音素事後確率ベクトル系列化し，それらの DTW(Dynamic Time Warping) として捉え (DNN-DTW)，その効果を示した．しかし，DNN-GOP, DNN-DTW スコアと手動評価スコアとは話者単位では高い相関が出ていたものの，文単位では十分な相関は出ていなかった．実環境への応用を考えると，文単位での評価が必要である．

本研究では [6] の研究成果を踏まえ，その精度向上を試みた．まず，耐雑音性能向上のため，1) バブルノイズを重畳した音声を用いた音響モデルのマルチコンディション学習を検討した．次に，2）実際のシャドーイング音声を用いた音響モデル適応を検討した．更に，3)GOP・DTW 以外の素性を導入し，回帰問題として手動スコアを予測するタスクについても検討した．その結果，文単位でも比較的良好な精度を得ることができた．

[1] 東京大学大学院工学系研究科
Graduate School of Engeneering, The University of Tokyo
[2] 創価大学
Soka University
[3] 高野山大学
Koyasan University
[a)] kabashima@gavo.t.u-tokyo.ac.jp
[b)] dsk_saito@gavo.t.u-tokyo.ac.jp
[c)] mine@gavo.t.u-tokyo.ac.jp
[d)] yutaka@soka.ac.jp
[e)] itou@koyasan-u.ac.jp

## 2. 先行研究

### 2.1 DNN-GOP

GOP とは観測された音響特徴量 $o$ とクラス $i$ の音素 $p_i$ に対して定義される音素事後確率 $P(p_i|o)$ のことであり，ある発声に対してどのくらい音素同士が区別して発声されているかを表す指標と解釈される．DNN-GOP はこれを DNN を用いて求めたものである．入力特徴量ベクトルに対して，それがどの音素状態であるかを学習した DNN により計算できる．シャドーイングの場合，学習者が聴取した母語話者音声（多くの場合読み上げ音声）が利用できるため，学習者が発声すべき音素列の情報を利用できる．この音素列を使って，学習者音声の徴量ベクトル系列に対して強制アライメントを行う[*1]．これにより，学習者音声の各フレームが対応している音素が得られる[*2]．一方，各フレームは DNN により音素状態事後確率ベクトルへと変換されるため，ここから各フレームに対し，当該音素の音素事後確率を取得する[*3]．

### 2.2 DNN-DTW

DTW とは，2 つの時系列に対して系列同士の累積距離が最も小さくなる対応付けを求める技術である．ここで系列同士の累積距離とは，系列を構成する要素間の局所距離の総和である．DNN の出力である音素状態（クラスタリングにより状態共有されたトライフォンの状態）事後確率を用いて，音声特徴ベクトルは音素事後確率ベクトルへと変換できる．DNN-DTW とは，2 つの音素事後確率ベクトル系列に対して行なう DTW である．モデル音声とシャドーイング音声に対して DNN-DTW を適用する．適切なシャドーイングができていれば累積距離は小さくなり，評価の指標とすることができる．ここで，局所距離としてはバタチャリヤ距離を用いる．また，クラス数（状態数）が数千ともなれば（事後確率化に使われる）言語への依存性は低くなることが予想され，[6] では，DNN-DTW の言語非依存性が実験的に示されている．

## 3. 雑音環境下での音声認識

一般に音声認識システムでは，静寂な環境下で収録された音声を用いて音響モデルの学習を行なう．学習時と似た環境下の音声に対して高い精度で認識するが，そうでない場合に認識精度の低下が問題になる．

本研究で対象としているシャドーイングの音声は集団で学習を行なう教室などで収録されることが想定される．そのため，収録音声に他人のシャドーイング音声などのバブルノイズが含まれることが多い．実環境での応用を考えるとこのような雑音に頑健な自動評価システムが望ましい．雑音に頑健な音響モデルの構築方法として，1）マルチコンディション学習，2）モデル適応，3）特徴量強調などが挙げられる．

### 3.1 マルチコンディション学習

使用時に含まれる雑音が予想できる場合は，雑音を重畳した音声を用いて音響モデルの学習を行なうことで耐雑音性を上げることが可能である． [8] では，クリーンな音声に様々な雑音を数種類の SN 比で重畳した音声を用いて学習（マルチコンディション学習）を行なうことで，ノイズ環境下音声認識タスクにおいて良好な結果が得られたことが報告されている．残響下音声認識のタスクにおいてもその効果が確認されている [9]．

### 3.2 音響モデルの適応

音響モデルの学習には大量のデータが必要となり，一般に多くの時間と計算資源を要する．そのため，使用環境に合わせて学習をやりなおすことが難しい．そこで，少量の音声データを用いた適応処理が行なわれる．耐雑音性を向上させたい場合，雑音が含まれる音声あるいは適応させたい環境の音声を用いて DNN の再学習を行う． [10] では，話者適応のタスクにおいて従来の GMM における適応技術より，DNN の再学習による適応の方が性能が良くなることが報告されている．また，L2 正則化を行なうことでモデルの汎化性能が向上することも報告されている．再学習を行なう場合，特定の層のみに限定して DNN のパラメータを更新することがしばしば行われる．

### 3.3 特徴量強調による雑音除去

上に述べた手法はいずれも音響モデル側で雑音の影響を少なくするアプローチであった．一方で，入力する音声について何らかの前処理を行なうことで雑音の影響を少なくするという方法も考えられる．代表的な手法としてスペクトルサブトラクション [11] がある．音声区間から非音声区間のスペクトルを差し引くことで，雑音の影響を抑えることができる．雑音が加算性かつ定常である場合に有効である．

近年では DNN を応用した雑音除去の手法が提案されている．DAE(Denoising AutoEncoder) もその一つである．DAE は欠損のあるデータから頑健な特徴（中間層出力）を取り出すための手法として提案された [12]．AutoEncoder では出力が入力そのものになるように学習を行なうが，DAE では入力に雑音を付与したものを与え，出力にクリーンな特徴量が得られるように学習を行なう．音声の分野

[*1] 学習者音声を，母語話者音声中の音素列が発声された結果として解釈している．
[*2] 各フレームが対応する音素状態（クラス）が得られるが，ここでは音素情報を取得する．
[*3] 状態単位（クラス）単位での事後確率が使えるが，同一音素だが異なる状態についての事後確率も考慮し，当該音素に対する状態事後確率の和を計算することで，音素事後確率としている．

では MFCC（Mel-Frequency Cepstral Coefficients) などの周波数領域の特徴量を用いて DAE を構成する． [13] では MFCC より lmfb(対数メルフィルタバンクの出力) を用いた方が性能が良いことが報告されている．

## 4. 予備実験

先行研究では，DNN-GOP や DNN-DTW が手動評価と高い相関を持つことが示されていた．我々は，シャドーイング音声に含まれる雑音がその精度に悪影響を与えているのではないかと推測した．我々が行なうシャドーイング音声収録は，冒頭に約 1 秒間の無音区間（無発声区間）が含まれる．この区間におけるパワー（RMS,Root Mean Square）を計算し GOP，DTW との相関を計算した．また，RMS が中程度のものはスコアに対する影響が小さいと考え，RMS が大きい上位 25%と下位 25%のみを取り出して相関を計算した．その結果 GOP は-0.23，DTW が 0.28 とわずかに相関見られ，5%で有意であった．[*4]

このことから，収録された音声に含まれる雑音が自動評価の精度に悪影響を与えていることが確認できた．本研究ではこのような雑音の影響を抑えるための手法について検討した．

## 5. 実験

本研究では，3 つの実験を行った．1 つ目はクリーン音声にバブルノイズを重畳した音声を用いた音響モデルのマルチコンディション学習とその効果．2 つ目はシャドーイング音声を用いた音響モデルの適応とその効果．3 つ目はそれらのモデルを用いて計算した DNN-GOP/DNN-DTW などを説明変数とした回帰である．

### 5.1 利用したコーパス

本研究で用いられた音声コーパスは 3 つである．1 つは DNN 学習用音声コーパス，残り 2 つは日本人による英語シャドーイング音声コーパスである．

**WSJ** Wall Street Journal の読み上げ音声コーパス．約 80 時間・4 万発話からなる．全ての音響モデルの学習に使用された．

**手動スコア付きシャドーイング音声コーパス** 124 名の大学生のシャドーイング音声．一話者につき 10 文の音声がある．実際の収録では，シャドーイングは各文に対して 4 回ずつ行われ，本実験で使用するのは，4 回目のシャドーイング音声である．自宅で収録されたものと教室で収録されたものがある．この音声には，日本人英語に対する知識を有する米国人・カナダ人英語教師 3 名による，下記 3 つの観点からの手動スコアが付与されている．

- Phoneme(P) 各文の個々の音素が，どの程度適切に生成できているか．
- Suprasegmental(S) 韻律・超分節的な側面が，どの程度適切に生成できているか．
- Correctness(C) 母語話者音声の各単語を同定して，シャドーできているか（より厳密には，そのように聞こえるか）．

各尺度に対して 5 点満点で評価され，15 点満点となる．本稿では手動スコアとして 3 人の平均値を用いる．

**モデル適応用シャドーイング音声**

手動評価が行われていないシャドーイング音声．手動スコア付き音声とは別人物の発声で，別時期に収録された．

- A 大学音声
  自宅など静寂な環境で収録されており比較的雑音が少ない．24 文× 53 名=1,272 文音声
- B 大学音声
  一般教室などで収録されており，他人のシャドーイング音声などのバブルノイズが含まれる場合が多い．
  24 文× 99 名=2,376 文音声

合計 3,648 文音声を音響モデルの適応に用いた．

### 5.2 バブルノイズを用いたマルチコンディション学習

バブルノイズを用いたマルチコンディション学習を行った．はじめに，実験を行なうにあたって WSJ コーパス中の音声データに対してノイズを重畳した．重畳するバブルノイズは， [14] で公開されているデータセットで用いられているネイティブ英語バブルノイズ音声を使用した．ノイズを重畳する際の SN 比については，ノイズレベルの強弱を考え，A:0,5,10,15dB と B:5,10,15,20,∞dB，の 2 通りを用意した．また，A,B 各々において比率は等しくなるようにした．

次に，雑音を重畳した音声を用いて音響モデルの構築を行った．入力特徴量には，MFCC13 次元の特徴量を用いた．前処理として CMN（Cepstral Mean Normalization），LDA（Linear Discriminant Analysis），MLLT（Maximum Likelihood Linear Transform）および話者正規化の fMLLR（feature-space Maximum Likelihood Linear Regression）を適応した．音響モデルの構築については Kaldi の WSJ レシピ [15] に基づいて行った．

2 つの音響モデル（A，B）を用いて音声認識実験を行った．認識対象は WSJ の評価データセット 333 文と手動評価付きシャドーイング音声からランダムに選択した 500 文である．また，WSJ のデータセットについては SN 比別に WER（Word Error Rate，単語誤り率）を計算した．表 1 に結果を示す．

モデル A は B に比べ SN 比が低い部分で WER が低くなっていることがわかる．しかし，SN 比が高い音声やクリーンな音声やシャドーイング音声についてはモデル B の

[*4] 雑音レベルが高くなると，（音響モデルはクリーン音声で構築されているため）GOP スコアは低くなる．一方，DTW 距離は（モデル音声はクリーン音声であるため）大きくなる．

表 1　単語誤り率 [%]

| モデル | WSJ0dB | WSJ5dB | WSJ10dB | WSJ15dB | WSJ クリーン | シャドー音声 |
|---|---|---|---|---|---|---|
| クリーン | 88.5 | 46.5 | 17.5 | 7.69 | 3.33 | 72.4 |
| A | 16.5 | 7.60 | 5.39 | 4.50 | 4.52 | 73.9 |
| B | 22.4 | 8.84 | 4.96 | 4.09 | 3.93 | 71.2 |

表 2　単語誤り率（sMBR 最小化基準,20 エポック）[%]

| 更新するパラメータ | WSJ 評価データ | シャドー音声 |
|---|---|---|
| 入力層のみ | 4.06 | 68.1 |
| 出力層のみ | 3.86 | 68.9 |
| 全て | 4.45 | 60.0 |
| なし | 3.33 | 72.4 |

方が WER が低くなった．もともとのクリーンなモデルに比べると，モデル B においてシャドーイング音声に対する WER が低くなった．シャドーイング音声には比較的静かな環境で収録されたものも存在しているので全体的に SN 比が小さいモデル B において WER が下がったと考えられる．以下の実験ではモデル B をマルチコンディション音響モデルとして使用する．

### 5.3　シャドーイング音声を用いたモデル適応

シャドーイング音声を用いて音響モデルの適応を行なう．適応処理では 5.1 節に示したモデル適応用シャドーイング音声を用いた．

#### 5.3.1　sMBR 最小化基準の再学習

音声認識では時系列を扱うため，ある系列に対して予測精度が高くなるように学習を行なうことがある．sMBR（state-level Minimum Bayes Risk）, 状態系列を対象にしたベイズリスク最小化基準では，ある正解単語 $W_u$ に対して予測された単語 $W$ の HMM 状態系列の正解精度を目的関数に含む．そのため，系列全体としての認識誤りを最小化することができる．[16] では sMBR などの系列識別学習が，単純に各時刻における予測誤差を最小化するように学習した場合より精度がよくなることが報告されている．Kaldi における DNN のレシピでも最終的に sMBR 最小化基準での学習が行われている．本研究でもそれにならい，sMBR 最小化基準でモデル適応を行った．

はじめに，予備実験としてパラメータを更新する層を限定した場合のモデル適応について調査を行った．クリーンな音響モデルをもとに，1）入力層のみ．2）最終層のみ．3）すべての層のパラメータを更新する場合にわけて再学習，認識実験を行った．また，隠れ層の数は 6 層，2048 ノードとした．表 2 に各場合における音声認識結果を示す（20 エポック）．

結果として，WSJ の評価データに対する WER は大きな差がでなかったがシャドーイング音声に対しては，全ての層のパラメータを更新する場合が最も WER が低くなった．以下，sMBR 基準の実験では全ての層のパラメータを

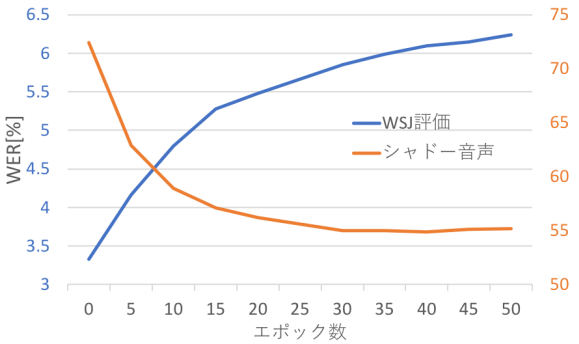


図 1　5 エポックごとの各データセットに対する WER

更新するものとする．

予備実験の結果を踏まえ，モデルの再学習は次のような手順で行った．また，学習率は $10^{-5}$ で固定とした．

(1) クリーン音響モデルを用いて適応用データのアライメントを取る
(2) sMBR 基準で DNN の再学習を 5 エポック行なう
(3) 2 でできたモデルを用いて再度適応用データのアライメントを取る
(4) 2,3 を WER が下がらなくなるまで繰り返す

図 1 に 5 エポックごとの音声認識結果を示す．

シャドーイング音声に対しての WER は 40 エポックで 54.9%と最小になりそれ以上は下がらなかった．また，その時の WSJ の評価セットに対する WER は 6.10%であった．

同様の実験をマルチコンディション音響モデルを初期モデルとした場合についても行った．ほぼ図 1 と同じような WER の変化が見られた．結果として，シャドーイング音声に対しての WER は 30 エポックで 55.4%と最小となった．また，その時の WSJ の評価セットに対する WER は 6.73%であった．いずれも，クリーンなモデルを初期モデルとした場合に比べてわずかに高かった．

#### 5.3.2　クロスエントロピー最小化基準の再学習

ここでは一般に DNN の学習において用いられる目的関数であるクロスエントロピー最小化に基づく音響モデルの適応を行なう．sMBR 同様，更新するパラメータを限定して再学習を行った．また，学習については適応データの 1 割をランダムに取り出し検証データとして交差検定を行い，フレーム認識率向上が 0.1%未満で学習を終了した（約 10 エポック）．学習率ははじめ 0.008 とし，学習が進むにつれて徐々に減少させた．初期アライメントの生成にはクリーン+sMBR の 40 エポック後のモデルを用いた．

まず，クリーンな音響モデルを初期モデルとして再学習

表 3　単語誤り率（初期モデル＝クリーン）[%]

| 更新するパラメータ | WSJ 評価データ | シャドー音声 |
|---|---|---|
| 入力層のみ | 27.4 | 91.1 |
| 出力層のみ | 7.35 | 50.4 |
| 全て | 10.5 | 50.7 |

表 4　単語誤り率（初期モデル＝マルチコンディション）[%]

| 更新するパラメータ | WSJ 評価データ | シャドー音声 |
|---|---|---|
| 入力層のみ | 32.0 | 91.5 |
| 出力層のみ | 8.54 | 49.6 |
| 全て | 11.1 | 50.6 |

表 5　構築したモデル（Xent=クロスエントロピー基準）

| モデル名 | 初期モデル | 適応 |
|---|---|---|
| C（ベース） | クリーン | なし |
| M | マルチコンディション | なし |
| CS | クリーン | sMBR, 全層 |
| MS | マルチコンディション | sMBR, 全層 |
| CX | クリーン | Xent, 最終層 |
| MX | マルチコンディション | Xent, 最終層 |

を行った．表 3 に各場合における音声認識の結果を示す．

出力層のみパラメータを更新した場合が最も WER が低くなる結果となった．また，入力層のみを更新した場合著しく精度が悪くなった．一部のパラメータを固定したことにより誤差の伝播が上手く行われなかったことが考えられる．

同様に，マルチコンディション音響モデルを初期モデルとした場合についても実験を行った．表 4 に各場合における音声認識の結果を示す．

こちらも出力層のパラメータのみを更新した場合に一番精度が良くなる結果となった．また，シャドーイング音声に対する WER は全ての実験の中で最も低くなったが，WSJ の評価データに対する WER が sMBR の 6.10%より高くなっているため，必ずしも良いとは言えない．[10] にならい L2 正則化による適応も検討したが，初期モデルの重みが十分小さいせいか，精度はあまり変わらなかったため割愛する．

### 5.4　相関分析による精度評価

ここまでで，合計 6 つの音響モデルが構築された．表 5 にまとめたものを示す．これらのモデルを用いて DNN-GOP[*5]・DNN-DTW の計算を行い，手動スコアとの相関係数を計算する．表 6 に手動スコアの合計点との相関を示す．

GOP については，モデル M の場合において相関が最も高くなっていた．シャドーイング音声を使用してモデル適応を行なう場合，雑音などの環境だけでなく，学習者音声すなわち日本人英語にも適応してしまう．その結果，不適切な発音であっても高い GOP スコアとなる．本研究での評価者は英語母語話者であるため，適応モデルでは相関が低くなり，適応なしのバブルノイズ耐性のあるモデル M が高くなったと考えられる．

[*5] 音素単位で GOP スコアを計算し，発声中のスコア合計を音素数で正規化する [17].

表 6　6 種のモデルで計算した GOP・DTW と手動スコアの相関

| | C（ベース） | M | CS | MS | CX | MX |
|---|---|---|---|---|---|---|
| GOP | 0.73 | 0.75 | 0.72 | 0.74 | 0.65 | 0.65 |
| DTW | -0.71 | -0.69 | -0.75 | -0.71 | -0.68 | -0.63 |

表 7　回帰における実験設定

| 説明変数 | 目的変数 | 回帰モデル |
|---|---|---|
| 各種 GOP(5 種), DTW 距離, WRR, 無音率 | 手動スコア (P, S, C, P+S+C) | Lasso, SVR, RandomForest |

DTW については，モデル CS の場合に最も高くなっていた．DTW では，HMM 状態に対応するラベル（Senone）に対する事後確率ベクトル系列を用いている．そのため，HMM 状態誤りを最小化する sMBR において精度が良くなったと考えられる．また，モデル M はモデル C より相関が低くなっている．DNN-DTW では，学習者音声とモデル音声に対して音素事後確率の計算を行なう．そのため，学習者音声（雑音を含む音声）に頑健なモデル用いても，お手本音声（クリーン音声）とはミスマッチが生じてしまい精度が悪くなったと考えている．

### 5.5　各種特徴量と回帰を用いたスコア予測

最後に，回帰モデルを用いたスコア予測を行なう．音響モデルには，DNN-GOP,DNN-DTW においてそれぞれ相関が高くなっていたモデル M とモデル CS を使う．また，ベースラインとしてクリーンなモデル C を用いる．説明変数・目的変数・回帰モデルはそれぞれ表. 7 の通りである．説明変数の各種 GOP5 種とは，通常の GOP に加え，母音音素のみについて GOP を計算した母音 GOP，子音のみについて計算した子音 GOP．強勢の位置別に計算した，第 1 強勢 GOP，強勢なし GOP の 5 種類である．強勢位置については CMU Pronunciation Dictionary [18] を参考にした．また，第 2 強勢音素は文によって出現しないことがあるため除外した．WRR（Word Recognition Rate）は音声認識の単語正解率 [*6]で，無音率は収録音声中の無音時間を収録時間で割った比率である．

学習者音声 1206 文（一部収録が正常にできていないものは除いた）を 10 分割し交差検定により回帰モデルの構築・評価を行った．表 8 にモデルの予測スコアと手動スコアとの相関を示す．また，回帰モデルはおおよそ精度が最も高かった SVR の例を示す．

結果として，提案法のモデルはベースラインより高いか同程度の精度であった．また，モデル Mix_sel は手動スコ

[*6] Word Accuracy，$1 - WER$ で計算される．

表 8　SVR の予測スコアと手動スコアの相関（文単位）

| 音響モデル | P | S | C | P+S+C |
|---|---|---|---|---|
| C（ベース） | 0.70 | 0.73 | 0.68 | 0.78 |
| M | 0.73 | 0.73 | 0.68 | 0.79 |
| CS | 0.67 | 0.75 | 0.69 | 0.78 |
| Mix_sel | 0.73 | 0.76 | 0.71 | 0.81 |
| Mix_all | 0.75 | 0.77 | 0.72 | 0.81 |
| 評価者間相関 | 0.58 | 0.54 | 0.74 | 0.75 |

アと特徴間の相関を調べ CS と M において相関が高くなる方を選択して回帰モデルを構築した場合の結果である．（GOP に関する特徴量はモデル M それ以外はモデル CS から得られる特徴量を用いた）Mix_all は特徴量選択を人手で行わずに，全ての特徴量を入力した場合の結果である．

結果として，Mix_all が全てのスコアに対する相関で最も高くなっている．Mix_sel でも精度は向上したが，明示的に特徴選択を行わなくてもモデルが自動的に良い方を選択していると考えられる．複数の音響モデルを用いることでさらに精度を向上させられることが実験的に示された．

4 章の予備実験同様に RMS と新しく計算した GOP・DTW との相関を計算してみたが，ほぼ変化が見られなかった．回帰における精度は向上したが雑音の影響は以前として残っていると考えられる．

## 6. まとめ

本研究では先行研究を踏まえ，耐雑音性能向上について検討を行った．結果としてマルチコンディション音響モデルを用いた場合に DNN-GOP の精度が向上し，クリーンなモデルを sMBR 基準で適応した場合に DNN-DTW の精度が向上した．適応処理を行なうと，環境だけでなく日本人の英語にも適応がかかってしまう．[19] では，適応時にもとのモデルと適応後のモデルの KL-divergence を用いた正則化が提案されている．今後，検討していきたい．また，今回行わなかった特徴量強調による雑音除去の手法についても検討していきたい．

**謝辞**　本研究は科研費 JP16H03084，JP16H03447，JP26240022 の助成を受けました．ここに感謝の意を表します．

### 参考文献

[1] Yo Hamada. Shadowing: Who benefits and how? uncovering a booming efl teaching technique for listening comprehension. *Language Teaching Research*, 20(1):35–52, 2016.

[2] Kun Ting Hsieh, Da Hui Dong, and Li Yi Wang. A preliminary study of applying shadowing technique to english intonation instruction. *Taiwan Journal of Linguistics*, 11(2):43–65, 2013.

[3] Yo Hamada. Shadowing: What is it? how to use it. where will it go? *RELC Journal*, 0(0):1–8, 2018.

[4] PROGRIT. https://www.progrit.co.jp.

[5] Dean Luo, Nobuaki Minematsu, Yutaka Yamauchi, and Keikichi Hirose. Automatic assessment of language proficiency through shadowing. In *Chinese Spoken Language Processing, 2008. ISCSLP'08. 6th International Symposium on*, pages 1–4. IEEE, 2008.

[6] Junwei Yue, Fumiya Shiozawa, Shohei Toyama, Yutaka Yamauchi, Kayoko Ito, Daisuke Saito, and Nobuaki Minematsu. Automatic scoring of shadowing speech based on dnn posteriors and their dtw. In *INTERSPEECH 2017*, pages 1422–1426, 2017.

[7] Shuju Shi, Yosuke Kashiwagi, Shohei Toyama, Junwei Yue, Yutaka Yamauchi, Daisuke Saito, and Nobuaki Minematsu. Automatic assessment and error detection of shadowing speech: Case of english spoken by japanese learners. In *INTERSPEECH*, pages 3142–3146, 2016.

[8] Michael L. Seltzer, Dong Yu, and Yongqiang Wang. An investigation of deep neural networks for noise robust speech recognition. In *ICASSP*, pages 7398–7402, 2013.

[9] Masato Mimura, Shinsuke Sakai, and Tatsuya Kawahara. Reverberant speech recognition combining deep neural networks and deep autoencoders augmented with a phone-class feature. *EURASIP Journal on Advances in Signal Processing*, 2015(1):62, 2015.

[10] Liao Hank. Speaker adaptation of context dependent deep neural networks. In *ICASSP*, pages 7947–7950, 2013.

[11] S.F.BBoll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustic, Speech, and Signal Procesing*, 27(2):113–120, 1979.

[12] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 1096–1103, 2008.

[13] Du Jun, Wang Qing, Gao Tian, Xu Yong, Dai Lirong, and Lee Chin-Hui. Robust speech recognition with speech enhanced deep neural networks. In *INTERSPEECH*, pages 616–620, 2014.

[14] Valentini-Botinhao and Cassia.
noisy speech database for training speech enhancement algorithms and tts models
, 2017. https://datashare.is.ed.ac.uk/handle/10283/2791.

[15] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011.

[16] Karel Veselý, Arnab Ghoshal, Lukás Burget, and Daniel Povey. Sequence-discriminative training of deep neural networks. In *INTERSPEECH*, pages 2345–2349, 2013.

[17] 椛島優, 塩澤文野, 齋藤大輔, 峯松信明, 山内豊, and 伊藤佳世子. DNN-GOP と DNN-DTW に 基づくシャドーイング音声自動評価の高精度化. Technical report, 2018.

[18] The CMU pronouncing dictionary
. http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[19] Dong Yu, Kaisheng Yao, Hang Su, Gang Li, and Frank Seide. KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition. *ICASSP*, pages 7893–7897, 2013.