修士論文

シャドーイング音声自動評価の高精度化と実用化
に関する検討

2019 年 1 月 31 日

指導教員　峯松 信明 教授

電気系工学専攻
37-176430　椛島 優

内容梗概

スマートフォンの普及により，音声の収集が容易になり大規模な音声コーパスを構築することが可能になった。それにより，音声認識や音声合成の品質が大幅に向上し音声の技術を知らない一般的な人でも簡単に音声認識システムが利用できるようになった。また，単なる音声認識だけでなく音声認識技術を活用した発音教育ソフトなどの開発も進められてきており，スマートフォン上で利用できることからその利用者も増えてきている。

現在利用されている発音教育支援システムの多くはユーザーに単語を発声させ，正しい発音との差異を検出して何らかの点数をユーザーにフィードバックするというものがほとんどである。しかしながら，単語の発音が良いだけでは連続した文を発声することは難しく，リズムやアクセントを習得することができない。そのため，より実践的な語学力を身につけるためには文単位あるいはそれ以上の長さを単位とした練習が必要となる。特に，シャドーイングは近年の語学学習で広く用いられている練習法であり，ある文の音読音声をなるべく遅れずに復唱することで行われる。ネイティブが発声した音声を真似ることでリズムや単語の連結などを習得することができる。シャドーイングは学習者自身で行えることが利点ではあるものの，その良し悪しを評価するということはあまり行われていない。

我々の研究室では以前からシャドーイングに着目しその自動評価を行う研究を進めてきた。近年の深層学習技術の発展にともない自動評価の精度が実用化できるレベルに近づいてきた。

そこで本研究では，シャドーイングの自動評価の高精度化および実用化に向けた検討を行う。まず，先行研究で挙げれられていたいくつかの問題点について調査・実験を行った。また，精度向上のため複数の特徴量を追加したうえで，回帰モデルの導入を行った。

次に，実用化における大きな課題である雑音に頑健な自動評価手法の検討を行った。語学教育は一般に教室などの多人数の話者が存在する環境で行われることが多く，学習者の音声以外に他人の音声がノイズとして含まれることになる。特にシャドーイングの場合，提示された音声に合わせて発声を行うため他の話者と内容が同一かつ時間的にもほぼ同一な音声がノイズとなってくる。これは，技術的に除去するには極めて困難な種類のノイズとなる。本研究では，このようなノイズについて実際の教室で音声収録を行いその影響を調査した。また，近年の深層学習技術を用いてどの程度ノイズの影響が抑制できるかについても実験を行った。結果として，人と比べても十分な精度を達成し，雑音に対する耐性も向上し，雑音に頑健な自動評価に近づくことができた。

目次

図目次

表目次

第1章

序論

1.1 研究の背景

スマートフォンの普及にともない膨大な音声コーパスの収集が可能になった。それと同時に深層学習の技術を用いた音声認識技術の研究が進み，高い精度の音声認識が実現可能となり音声認識技術を搭載した機器が生活のあらゆる場面で見られるようになった。音声認識に使用されている技術は単なる音声認識にとどまらず，発音教育ソフトなどのも開発にも応用されてきている。例えば,「ELSA [1]」は学習者の発声を録音し，ネイティブの発音との違いを検出し，その良し悪しを段階をわけてフィードバックしてくれる。その他にも数多くの語学教育アプリケーションが開発されており，現代人にとって語学学習に音声認識技術を利用することは当たり前になってきている。

ここで挙げたようなものは提示された文を読み上げてその発音を評価するという形をとっている。一方で，文ではなく音声を提示してその音声を復唱するという練習法もある。このときに復唱ではなく，提示された音声になるべく遅れずに復唱を行うことをシャドーイングという。

シャドーイングは語学教育の現場で広く導入されている練習法で，聞きながら復唱するという認知的に負荷の高い活動であり初学者から上級者まで利用されている練習法である。シャドーイングは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ことが報告されている [8, 9]。第二言語の獲得において最終的に目指すところは，1）相手の発話を聞く，2）内容を理解する，3）自分の発話内容をプラニングする，4）それを発声として生成する，などを同時に行えるようになることである。そのため，シャドーイングのようにこのような処理をパラレルに走らせるような訓練が必要であると言える。また，シャドーイングはビジネス英会話スクールなどの学習プランにも取り入れられており，継続的にシャドーイングを行うことで受講者も自身の英会話スキルが向上していることを実感できているようだ [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 [10]。

[1] ELSA https://www.elsanow.io/home
[2] PROGRIT https://www.progrit.co.jp

1.2　研究の目的

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難しく，モチベーションを保つことができなくなってしまう。シャドーイング音声を自動で評価し，学習者のモチベーションを向上させることが求められる。シャドーイングの評価も基本的には発音の評価と同等であるが，細かい発音の前にそもそも聞き取れていなかったり，単語そのものが上手く発声できていなかったりすることが多々ある。したがって，発音評価システムのようにそれぞれの音素を細かく比較してその結果を返すというより，お手本の音声に比べてどの程度の達成度であったかを点数化するような採点が望ましい。また，学習者自身にシャドーイング音声の評価がフィードバックされることで学習を継続するためのモチベーション向上につながる。シャドーイングが上手くできるようになった後は，より難しい教材を選んだり，単語単位で細かい発音を練習するような方向に切り替えれば良い。

そこで本研究では，細かい誤りを検出するシステムではなくある発声に対して点数付けを行うシステム，いわゆる「Overall assessment」を行う自動評価システムの構築を目指す。また，実際の教育現場での実用化を踏まえて学校の教室など多人数で同時に発声するような環境で動作することを目標とする。

1.3　本論文の構成

本論文は全 5 章から構成される。この第 1 章では，本研究に関連する近年の研究の流れを簡単に述べ，本研究の有用性と本論文の流れについて説明する。第 2 章では，これまでの音声認識技術に基づく発音評価の基礎や関連研究・先行研究について紹介する。第 3 章では，先行研究における自動評価手法の改善を行い，高精度化に向けた手法および実験とその結果を示す。第 4 章では，実用化をふまえ，音声に含まれる雑音に頑健な自動評価に向けた検討を行う。さらに，雑音収録実験とその結果および耐雑音手法の効果について報告する。第 5 章では，本論文によって達成された課題と解決すべき問題点についてまとめる。

第2章

音声認識技術による外国語学習支援

2.1 はじめに

国際化の進む現代社会において外国語の習得は誰もが直面する課題となっている。移民が増えていることもあり，日本国内では英語を習得したい日本人だけでなく日本語を習得したい外国人も増えていることだろう。政治的な側面からみても，小学校教育に英語が正式な科目として採用されるなど外国語教育の需要はますます高まっている。教育の需要が増えると当然，語学教師が必要となってくるわけだが，教師を必要としている学習者の数を考えると十分な数の語学教師を確保することは難しい。そこで，CALL（Computer Assisted Language Learning）やMALL（Mobile Assisted Language Learning）といったコンピュータやスマートフォンなどの機器を用いた学習支援や効率的な教育が注目されている。これらには音声教材や画像教材を提示するといった使用用途も含まれる [14] が，より高機能なものには学習者の発音を自動で評価するようなシステムが含まれる。

[5] はCALL中のでも特にCAPT（Computer Assisted Pronunciation Training）と呼ばれるものの一例である。図 2.1 はそのシステムをスマートフォン上で動かしたときの画面の例である。詳細を説明することは行わないが，単音単位で練習するステップや発音の似た単語を2つ提示しているステップや，音声合成システムを用いて学習者に正しい発音を示しているステップがある。このようにCALLの分野では音声認識技術に限らず音声合成の技術やマルチメディア処理の技術が統合的にシステムとして導入されることが多い。本研究では，音声合成やマルチメディア処理の部分には触れずに音声認識技術をベースにした外国語学習支援に焦点を置く。

2.2 発音誤り検出技術

ここでは，音声認識技術を利用した外国語学習支援の代表的な例である発音誤り検出技術の概観について触れる。

2.2.1 発音誤りの種類

発音誤りとは何を意味するのか。ネイティブのように発音することを正しいとするか，聞き手が識別可能な発音であれば正しいとするか，などの問題もある。また，音というものは連続的なものであって，正しい発音と誤った発音に明確な境界を定義することは難しい。発音誤りの検出方法について述べる前に，発音における誤りについて大まかな種類を挙げておく。 [33] によると

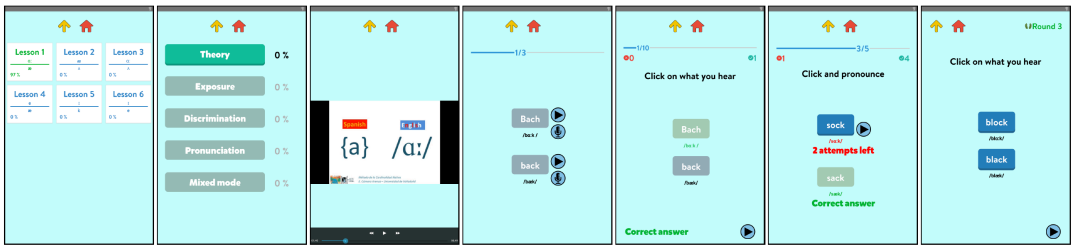


図 2.1: スマートフォン上で動作する CAPT の画面例 [5]

表 2.1: 発音誤り検出に用いられる音響的特徴量の一例 [33]

発音誤りの種類	音響特徴量
音素的	対数尤度比,GOP
	母音継続長
	フォルマント周波数
韻律的	F0
	パワー（単位時間当たり）
	無音時間（単位時間当たり）
	話速 (WPM)

発音誤りは，図 2.2 のように大きく 2 つに分けることができる。一つは，音素的な (Phonemic) 誤りである。音素的な誤りは挿入誤り (Insertion)，置換誤り (Substitution)，脱落誤り（Deletion）歪み（Distortion error）などにわけられる。日本人が英語を発音するとき不必要な母音を挿入してしまうことは挿入誤りの典型的な例である。歪みとは，どの音素にも分類できないような曖昧な音のことで，実際にはこのような誤りがほとんどだと言われている [27]。

もう一つは，韻律的な誤りである。韻律的な誤りには強勢（Stress），音節単位でのリズム (Rhythm)，抑揚 (Intonation) の誤りがある。韻律的な間違いの 3 つについては，言語や文献によって解釈が異なるが，強勢は主に読み上げ速度の遅速，強勢は声の強弱，抑揚は高低によるものと表現できる。

2.2.2　発音誤り検出に用いられる主な特徴量

ここでは，音声処理技術を用いて発音誤りを自動的に検出する際にどのような特徴量が使われるのか，代表的な例を取り上げて紹介する。表 2.1 に発音誤りがどのような特徴に現れるかいくつかの例を示す。GOP（Goodness Of Pronunciation）は発音評価や誤り検出における基本的な評価指標の一つである。GOP の詳細やその計算方法については 2.2.3 節で述べる。フォルマント周波数は声道の形状により決まる共振周波数のことで，第一・第二フォルマント周波数を用いると日本語五母音をおおよそ分類することができる。F0 は声帯振動の基本周波数のことである。音の高さに対する心理量をピッチと呼ぶが，これに対応する物理量が F0 である。つまり，発音の高低の変化は F0 の変化としてとらえることができる。韻律的な発音の誤りを検出する場合このような特徴を利用することがある。

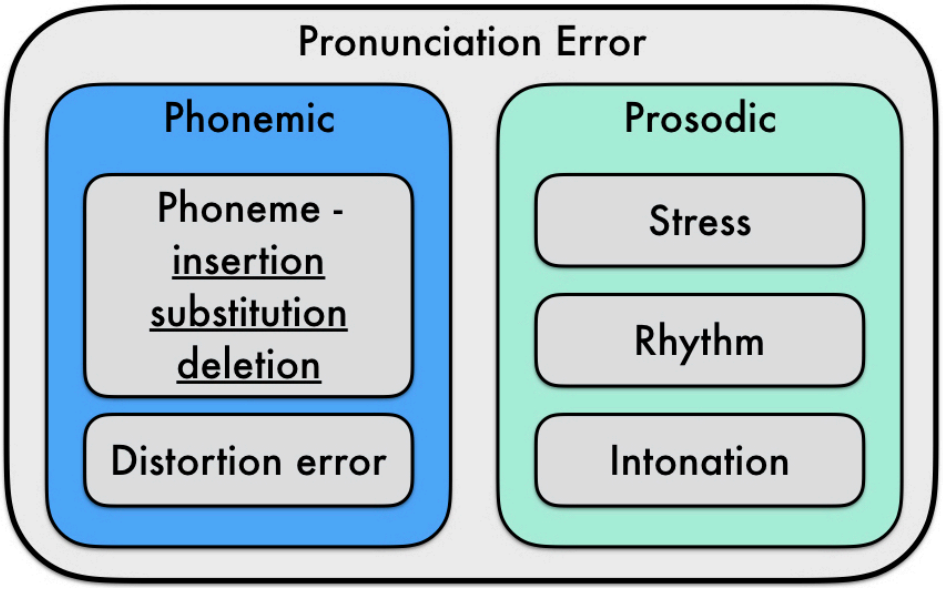


図 2.2: 発音誤りの種類 [33]

2.2.3　音響特徴量の抽出方法

2.2.2 節で紹介したような音響特徴量を計算するにはどのような手順が必要になるか，音素的なものをいくつか取り上げて紹介する。また，フォルマント周波数や F0 については [39] を参照していただきたい。

人の声は声帯の振動による音源波形の生成と声道形状の変形による共振周波数の変化からなる。音声の分野では声帯の振動（ソース）と声道の形状（フィルタ）によるモデルのことをソースフィルタモデルと呼ぶ。声帯の振動は F0 に関係してくるもので，音素的な特徴量にはなりえない。音素の違いは声道形状の変化によるものである。では，どのようにして人の音声から声道の形状の情報，すなわち音素的な情報を取り出すのか。その手順はを大まかに示すと図 2.3 のようになる。上から順に説明していく。

1. 音声波形のサンプリング・量子化
   音声は空気中を伝わる波である。これを機械的に処理するためにデジタル信号処理では，アナログ波形を離散的な値にサンプリング・量子化を行う。一般に人の音声を分析する場合は，サンプリング周波数が 16kHz 量子化ビット数が 16bit 程度である。
2. 短時間フーリエ解析
   入力音声から音素的な情報を取り出すためには短い時間で，波形の変化を捉える必要がある。そのため，得られた音声波形にハミング窓などの窓掛けを行い短時間の波形を切り出す。その後，離散フーリエ変換を行い周波数領域のスペクトルに変換する。
3. パワースペクトル
   フーリエ変換の結果からパワースペクトルを計算すると楕形のスペクトルが得られる。このうち微細な変動は声帯の振動を，スペクトルの概形（包絡）は声道の形状を表している。音声認識や音素的な違いのみを区別したい場合は，声道の形状の情報を効率的に得られることが望ましい。そこで，音声の分野では次に述べる MFCC（Mel-Frequency Cepstum Coefficients）が一般的な音響特徴量として広く使われている。

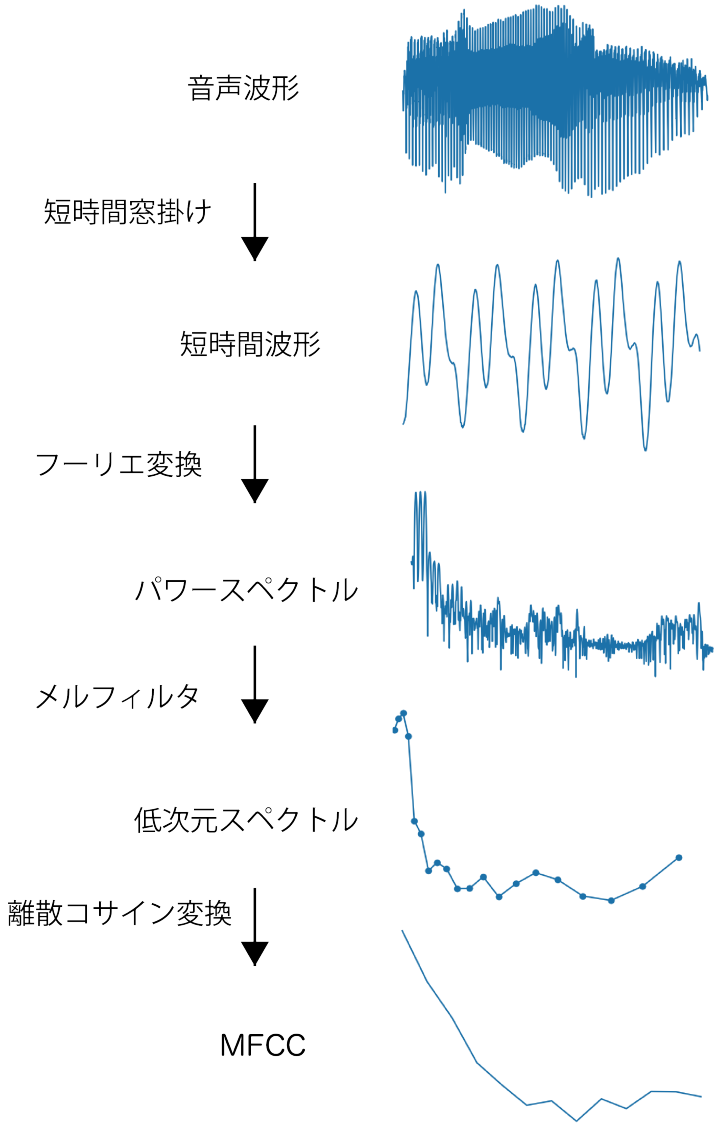


図 2.3: 音響特徴量の抽出手順

4. MFCC（Mel-Frequency Cepstum Coefficients）の抽出
   フーリエ変換の結果から得られるパワースペクトルはそのまま扱うには次元数が多い。そのため，パワースペクトルにメルフィルタをかけて低次元のスペクトル特徴量に変換する。メルフィルタとは人間の聴覚特性に合わせて，低い周波数では分解能が高く，高い周波数では分解能が低くなるようなフィルタである。数式では式 (2.1) のように書ける。図 2.4 にメルフィルタの例を示す。この処理を行うことのよりパワースペクトルをより低次元で表すことができる。最後にこの低次元スペクトルに対して離散コサイン変換を行う。このうち低次の 13 次元程度を取り出したもを MFCC と呼ぶ。

$$Mel(f) = 1127 \log_{10}(1 + \frac{f}{700}) \tag{2.1}$$

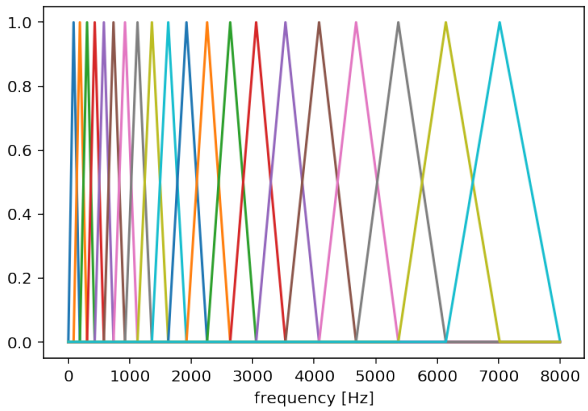


図 2.4: メルフィルタの一例

さて，ここで紹介した MFCC は音韻情報をよく表している特徴量ではあるが，人の声のスペクトルは話者によって異なったり，同じ話者でも複数回発声すれば微妙に異なってくる。そのため，そのゆらぎを上手くモデル化する必要がある。音声認識の分野ではこのような音のモデルのことを音響モデルと呼び，GMM（Gaussian Mixture Model）や DNN（Deep Neural Network）を用いてモデル化することが一般的である。単純な音素認識器を構成したければ解くべき問題は式 (2.2) のように定式化できる。右辺はベイズの定理を適用した結果である。この音響モデルの役割はある入力音響特徴量 o（MFCC など）が与えられたときにそれがどの音素が意図されたもの p（単語）であったかを確率的に推測することである。$P(p)$ はある音素の出現頻度を示す事前確率である。$P(o|p)$ はある音素が決まったときのどのような音響特徴量が出力されるかをモデル化するものであり，これを音響モデルと呼ぶ。また，近年の DNN の枠組みでは $P(p|o)$ をそのまま推定することが可能となっている [34]。

$$\hat{p} = \underset{p}{\operatorname{argmax}}\, P(p|o) = \underset{p}{\operatorname{argmax}}\, P(o|p)P(p) \tag{2.2}$$

発音誤り検出などの大部分はこの音響モデルに依るところが大きい。次節ではこの音響モデルを利用した発音誤り検出手法について紹介する。

2.2.4　発音誤り検出手法

ここでは，音響モデルを利用した音素的な発音誤り検出手法について紹介する。音響モデルはある音響特徴量を与えたときにそれがどの音素であったかを表す尤度を計算することができる。そこで尤度の比を利用した発音誤り検出手法が提案されている [12]。この手法ではまず，異なる 2 つの音響モデルを用意する。一つはネイティブの正しいとされる音声によって訓練されたもの（λ_C），もう一つは，非ネイティブの誤りを含んだ音声によって訓練されたもの（λ_M）である。これら 2 つの音響モデルに対してそれぞれ尤度計算を行う。対数尤度比 LLR（Log Likelihood Ratio）は式 (2.3) によって計算される。

$$LLR(p) = \log(\frac{P(O|p,\lambda_M)}{P(O|p,\lambda_C)}) \tag{2.3}$$

適切なしきい値を設定することで誤り検出を行うことができる。例えば日本人英語の誤り検出をしたければネイティブ英語（正しい英語）で構築した音響モデルと日本人英語で構築された音響モ

デルの 2 つを用意する必要がある。音響モデルを構築するには大量の音声コーパスが必要となるため母語以外の音声コーパスを入手しなければいけないという点が課題となる。

2.2.5　スコアリング手法

前節で述べた手法はしきい値を定めて正しい発音か誤った発音か，の 2 値分類を行うようなものであった。しかしながら，一般に誤りとそうでないものの境界というのははっきりしておらず，正誤のラベル付けを行うには音声学に長けた語学教師の協力が必要不可欠である。そこで，明確に正誤をきめるのではなく発声に対して点数付け（スコアリング）を行う場合がある。この分野では GOP が用いられることが多い。

GOP とは音素事後確率のことで $P(p|o)$ のことである [25]。これは調音制御の適切さを表す指標とされる。ここで p は意図とされた音素であるが，GOP の計算の際にはこの意図された音素を決定する必要がある。これは HMM（Hidden Markov Model）による強制アライメントによって実現できる。ある発声 x とその書き起こしが与えられたときその発声に対する GOP は式 (2.4) で計算できる。t は時刻を表す。D_x はフレーム数である。

$$\mathrm{GOP}(x) = \frac{1}{D_x} \sum_t P(p_t|o_t) \tag{2.4}$$

GOP は確率であるため 0〜1 までの値を取る。これをスコアとすれば自動スコアリングが可能となる。人間がつけた手動の点数を目的変数とした回帰モデルを用いることで GOP と手動スコアとの関係を導くことができる。また，GOP に対してしきい値を定めることで誤り検出を行う場合もある [26]。

2.3　シャドーイング音声の自動評価

ここでは，本研究の直接の先行研究となるシャドーイング音声を対象とした自動評価手法について紹介する。

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 [9, 13]。効果的な学習には，学習者に対するフィードバックを必要とするが，知識豊富な教師を十分用意することは難しい。シャドーイング音声を自動で評価し，教示を返すシステムが求められている。

Luo らはシャドーイングに学習者の総合的な英語能力を表していると考え，シャドーイング音声から学習者の TOEIC スコアを推定することを試みている [18]。2000 年代の研究であり，DNN ではなく GMM-HMM の音響モデルを使用しているため，あまり高い精度は実現できていないが，シャドーイング音声から学習者の英語能力が推定可能であることが示されている。

Shi らはシャドーイング音声を分析し，どのような誤りが存在するかを解析している [24]。最も多いのは単語の脱落であったため，強制アライメントの際に用いるネットワークにおいて各単語の発音が無音相当の音素と置換されることを許容するようなネットワークを用いることで単語脱落の検出を行っていた。

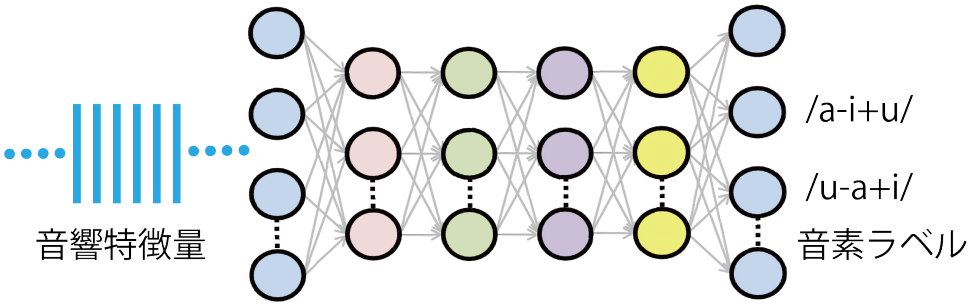


図 2.5: 音響モデルにおける DNN のイメージ図

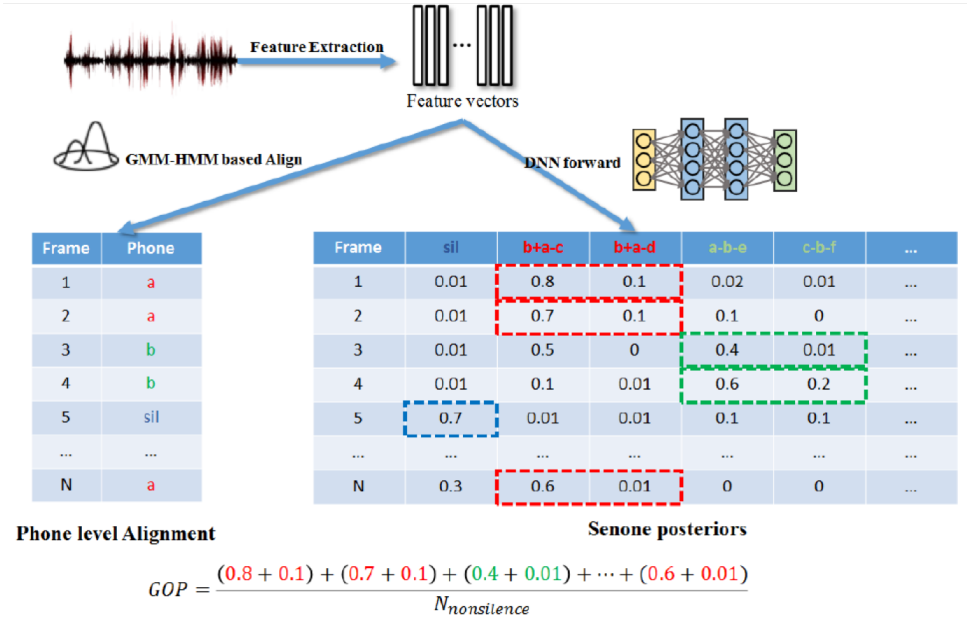
2.3.1　DNN-GOP に基づくシャドーイング音声の自動評価 [36]

近年の深層学習の発展により音声認識の分野においても大きな変化があった。すでに何度か述べている音響モデルもそのうちの一つである。GMM を用いる場合 GOP の計算のためには，事後確率を直接求めることができないため，近似式を用いて計算を行っていたが DNN の登場により，事後確率を直接求めることが可能になった。図 2.5 にその模式図を示す。DNN とは中間層を深くしたニューラルネットワークである。ニューラルネットワークとはある入力と出力の間の非線形な変換を表現するためのものである。したがってニューラルネットワークの学習には入力と出力ラベルのペアが存在することが前提となる。音声認識のおいては入力特徴量は 2.2.2 節で説明したような MFCC などの音響特徴量が，出力には音素ラベルが用いられる。ここでいう音素ラベルとは音素そのものではなく，トライフォンと呼ばれる前後の音素を考慮したものが用いられる。図の例では，/a-i+u/は当該音素が/i/で前の音素が/a/後ろの音素は/u/であることを表す。さらにこのトライフォンは決定木によるクラスタリングによって縮退され，実際には数千のクラスとなる。つまり，ある音響特徴量が与えられたときに数千クラスのうちのどの音素クラスに相当するかを予測するモデルを構築することなる。これは，まさに $P(p|o)$ のモデル化そのものであり，GOP を直接計算することができるようになる [32]。

Yue らはこの DNN-GOP をシャドーイングの評価に適用している [36]。DNN-GOP の計算方法について図 2.6 に示す。図上部の特徴量抽出についてはすでに述べた。実際には MFCC を数フレーム分結合するなど幾つかの前処理を行っているが，これについては後の実験の部分で説明する。入力特徴量はフレームごとに与えられるため，これらに強制アライメント（図左）と DNN の順伝播計算（図右）をそれぞれ行う。強制アライメントの結果から各フレームと音素の対応が得られる。同時に，DNN の順伝播計算の結果から各フレームにおける音素事後確率[1]が計算される。DNN-GOP の計算では，強制アライメントの結果から得られた音素に対応する複数の音素クラスの事後確率を合計し，そのフレームにおける GOP スコアとする。ただし，強制アライメントの結果が無音に相当する音素であった場合は計算から除外する。最終的にある発話の GOP スコアは全フレームで計算，合計され無音以外のフレーム数 $N_{nonsilence}$ で除することで正規化される。

Yue らは DNN-GOP によるシャドーイング評価の実験のため，日本人大学生の英語学習者のシャドーイング音声の収録実験を行った。その音声に対して英語母語話者による手動評価を行っている。結果として，DNN-GOP と学習者のシャドーイング音声には高い相関があることが示されている。

[1] 正確には状態共有を行ったトライフォンの HMM 状態の事後確率，専門用語で Senone とも書く。



Frame	Phone
1	a
2	a
3	b
4	b
5	sil
...	...
N	a

Frame	sil	b+a-c	b+a-d	a-b-e	c-b-f	...
1	0.01	0.8	0.1	0.02	0.01	...
2	0.01	0.7	0.1	0.1	0	...
3	0.01	0.5	0	0.4	0.01	...
4	0.01	0.1	0.01	0.6	0.2	...
5	0.7	0.01	0.01	0.1	0.1	...
...	...	...	...	...	...	...
N	0.3	0.6	0.01	0	0	...

図 2.6: DNN-GOP の計算フロー図 [35]

2.3.2　DNN-DTW に基づくシャドーイング音声の自動評価 [36]

GOP と同様に音素事後確率を用いたシャドーイング音声の自動評価手法として DNN-DTW が提案されている [36]。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。DTW について簡単に説明しておく。図 2.7 に示すように m フレームの音響特徴量 $X = \{x_1, \cdots, x_m\}$ と n フレームの音響特徴量 $Y = \{y_1, \cdots, y_n\}$ が与えられたとする。このとき，i 番目のフレームと j 番目のフレームの間に局所距離 $d(i, j)$（ユークリッド距離など）を定義する。k 番目の対応点を (i_k, j_k) とすると，X と Y の対応付けは対応点列（パス $\{(i_n, j_n)\}$ として表現できる。このとき解くべき問題は式 (2.5) のように表現できる。

$$\min_{\{(i_n,j_n)\}} \left[\sum_{k=1} d(i_k, j_k) \right] \tag{2.5}$$

この問題は次のいくつかの制約を課すことによって動的計画法を用いて解くことができる。

1. X と Y の始点同士，終点同士が対応する。
2. 対応点列はそれぞれの時間軸に対して順序が逆転しない。
3. 直前の対応点から次の対応点に対して図 2.8 のような制約を定める。

対応点 (i, j) の局所距離を $d(i, j)$ とし，(i, j) に至るまでの累積距離の最小値を $D(i, j)$ と書くと式 (2.6) の漸化式が得られる。

$$D(i,j) = min \begin{bmatrix} D(i, j-1) + d(i,j) \\ D(i-1, j-1) + 2d(i,j) \\ D(i-1, j) + d(i,j) \end{bmatrix} \tag{2.6}$$

さて，この DTW を音響特徴量間で行うと発話間の距離（類似度）を測ることができる。しかしながら，前にも説明したように MFCC などのスペクトルを表す特徴量をそのまま使用した場合，音素

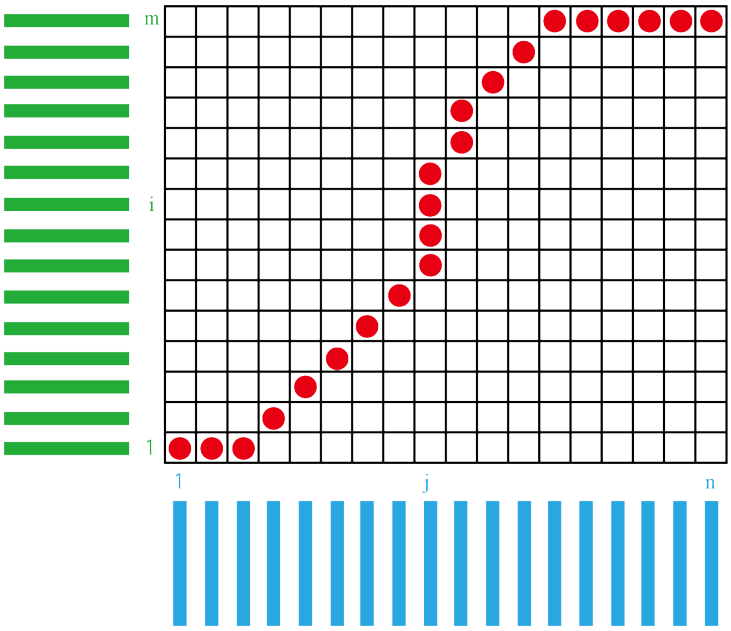


図 2.7: $m \times n$ 時系列間の DTW

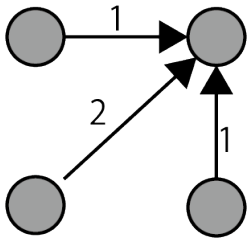


図 2.8: DTW の局所パスの制約と重み

の違いだけでなく話者の違いによっても距離が変化してきてしまう。そこで，話者の違いに頑健な DNN の出力である音素事後確率ベクトルを利用した DTW（Posteriorgram-DTW,DNN-DTW）による発話比較などが検討されている [21, 3]。

Yue らは DNN-DTW をシャドーイング音声の評価に適用している [36]。シャドーイングを行う場合，お手本となるモデル音声が存在する。モデル音声と学習者音声をそれぞれ音素事後確率ベクトルに変換し，両者に DTW を適用すればモデル音声と学習者音声との距離を計算することができる。この場合，出力ベクトルは確率分布であるため，分布間の距離を用いることが考えられるが [36] ではバタチャリヤ距離を用いている。離散確率ベクトル $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$ 間のバタチャリヤ距離は式 (2.7) で計算できる。DNN-GOP の場合と同様に，学習者シャドーイング音声に対して DTW 計算を行いその累積距離を学習者のスコアとした場合に手動スコアとの間に高い相関があったことが示されている。

$$D_{BD}(\boldsymbol{a}, \boldsymbol{b}) = -\log\left(\sum_i \sqrt{a_i b_i}\right) \tag{2.7}$$

また，音素事後確率は 2.3.1 節でも説明したようにそのクラス数は数千におよぶ。英語音声を評価する場合は英語の音響モデルを用いることが適切であるが，クラス数が数千にもなると異なる言語同士でも音として共通な音素クラスが存在することが予想される。実験では英語音声のシャ

ドーイング評価を日本語の音響モデルを用いて事後確率化した場合においても，DTW の結果と手動スコアの間に一定の相関があることが示されていた。実験的にではあるが DNN-DTW には言語非依存性があると言える。

第3章

音素事後確率に基づくシャドーイング音声の高精度化

3.1 はじめに

2章では外国語学習支援に関する研究について紹介した。本研究では [36] を先行研究に位置づけ，シャドーイング音声の自動評価について取り上げる。

[36] では，DNN-GOP および DNN-DTW など音素事後確率に基づくシャドーイング音声の自動評価手法が提案されていた。しかしながら，これらの手法には次のような幾つかの課題が残っている。

1. GOP の計算において単純にフレーム数で平均をとっている。
2. DTW における音素事後確率のクラス数が最適化されていない。
3. 話者単位での相関は高いが，文単位での相関が低い。

本章ではこれらの課題についての検討および実験について報告する。

3.2 手動評価付きシャドーイング音声コーパス

実験の説明に入る前に，先行研究 [36] および本研究において主に用いられたシャドーイング音声コーパスについて説明しておく。

3.2.1 シャドーイング音声の収録

Yue らは，自動評価手法の有効性を検証するためにシャドーイング音声コーパスの収集を行っている [36]。対象者は 124 名の日本人大学生で 3 つの異なる大学で収録を行った。シャドーイングはテキストを見ながら行う場合もあるが，テキストを見せずにシャドーイングを行った。シャドーイングは全部で 55 文あり，4 つの異なるトピックから選ばれた。また，シャドーイング自体が比較的難しいタスクであるため，初めて聞く音声をシャドーイングする場合上手くシャドーイングできない可能性がある。そのため，それぞれの文音声について 4 回づつシャドーイングを行っている。表 3.1 にシャドーイングに用いた文の一例を示す。

収録環境だが，3 つの大学のうち A 大学（およそ 6 割）の音声は学生個人の自宅で収録された音声で，できるだけ静かな環境での収録を依頼している。マイクはデスクトップ型のマイクを用いている。残りの B・C 大学の学生については，学校の教室内で収録を行っている。そのため，周辺の学生のシャドーイング音声が収録されてしまうことが予想される。それをできるだけ防ぐた

表 3.1: シャドーイング音声収録に用いた文の一例 [35]

トピック	文番号	書き起こし
My name is Akira	1	Hi, my name is Akira.
My name is Akira	4	I'm studying photography too. Shall we exchange some photos we've taken, and discuss them on the Internet.
MacDonald	1	The MacDonald's house has been broken into.
MacDonald	10	It was you who kicked the door, wasn't it?
Valentine	1	February 14th is a day for people who have fallen in love.
Valentine	12	People wrote their own words on the cards. Usually a kind or funny message.
Fugu	1	In 1996, three men in California were taken into a hospital with strange symptoms.
Fugu	4	The hospital doctors thought the men had been poisoned but couldn't work out what's wrong with them.

めに，イヤーフックマイクを用いることで収録の対象者の口元へできるだけ近づけて，それ以外の学習者からは遠ざけるようにしている。

3.2.2　シャーイング音声の手動評価

[36] では，前節のシャドーイング音声コーパスに対して語学教師による手動評価を行っている。シャドーイングの収録そのものは 55 文行われたが，手動採点は専門家の知識と経験が必要でありその全てを採点するには多大な時間を要する。そこで，英語教師の助言のもと 55 文のうち 10 文が手動評価の対象として選択された。また，4 回のシャドーイング音声のうち 4 回目の音声を使用した。最終的に 124 人× 10 文=1240 文 [1]の音声が手動採点されることとなった。また，シャドーイングの特性上，途中まではシャドーできているが途中からできなくなってしまうということが起こりうる。そのため，一文を 2〜3 のフレーズに区切りフレーズごとに採点を行った。

次に，採点方法について説明する。採点者は米語を母語とする英語教師 3 名でうち二人はアメリカと日本のハーフである。3 名とも日本人英語学習者に慣れているため，教育的な観点から評価ができることが期待される。また，採点基準については次の 3 つを用意した。

- 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。1〜5 点
- 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。1〜5 点
- 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。1〜5 点（より厳密には，そのように聞こえるか）。

合計で最低 3 点，最大 15 点満点の点数が各フレーズごとに付与された。

自動評価とは人間の採点をシミュレーションするわけだが，そもそもの人間の採点に大きな個

[1] 一部の音声は収録できていなかったため全部で 1206 文

表 3.2: 3 人の評価者の各スコアにおける平均値と標準偏差 [35]

	Phoneme (P)	**Prosody (S)**	**Correctness(C)**
SA-平均	1.9	4.2	4.1
SA-標準偏差	0.61	0.56	0.54
SB-平均	2.9	2.8	3.7
SB-標準偏差	0.64	0.82	0.68
SC-平均	2.7	2.9	3.7
SC-標準偏差	0.71	0.78	0.69

人差があっては一般的な自動評価システムを構築することは難しい。そこで，参考として 3 人の評価者の手動スコアの統計情報を示しておく。まず，3 人の評価者（SA，SB および SC）のスコアごとの平均値と標準偏差を表 3.2 に示す。評価者 SA は P スコアについてかなり厳しい採点をしていることがわかる。一方で SB と SC は 3 つのスコアで似た採点の傾向があることがわかる。また，C スコアは全体的に高い点数がつけられている。これは，評価対象の音声が 4 回目のシャドーイング音声であり，単語の脱落が少なかったことに起因すると思われる。

次に，評価者間のスコアの関係性を見るために，評価者間のスコアの相関を示す。相関はフレーズ単位，文単位，話者単位で計算された。それぞれ，表 3.3, 3.4, 3.5 に示す。

表 3.3: フレーズ単位で計算した評価者間相関 [35]

	P	**S**	**C**	**P+S+C**
SA_SB	0.47	0.42	0.67	0.71
SA_SC	0.43	0.47	0.65	0.69
SB_SC	0.56	0.54	0.70	0.74

表 3.4: 文単位で計算した評価者間相関 [35]

	P	**S**	**C**	**P+S+C**
SA_SB	0.57	0.46	0.72	0.73
SA_SC	0.52	0.54	0.72	0.73
SB_SC	0.66	0.62	0.77	0.80

表 3.5: 話者単位で計算した評価者間相関 [35]

	P	**S**	**C**	**P+S+C**
SA_SB	0.74	0.63	0.85	0.87
SA_SC	0.72	0.73	0.87	0.86
SB_SC	0.84	0.72	0.86	0.87

P+S+C とは各スコアの合計である。全体的な傾向として文スコアはフレーズスコアを平均したもの，話者スコアは文スコアを平均したものであるため，相関も話者単位が最も高くなる。フレーズ単位，文単位ともに P,S スコアにおいて相関が低い傾向があるが C スコアおよび合計スコ

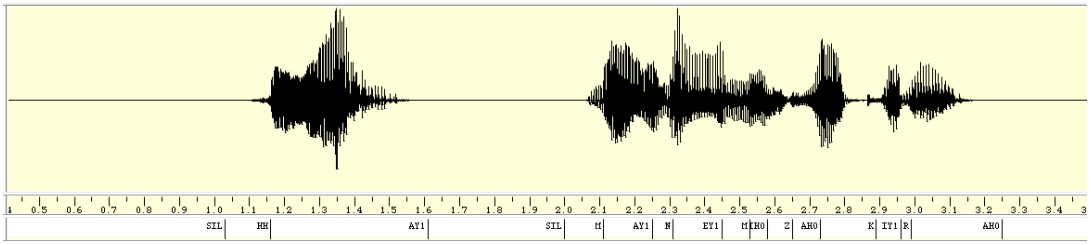


図 3.1: 英語音声に対するアライメント結果

アではある程度高い相関がある。また評価者 SA と SB は平均と標準偏差に似た傾向があったように，相関も他の組み合わせに比べ高くなっている。いずれの場合も合計スコアに関しては一定の相関が見られる。評価者が 3 名と少ないこともあるが，手動スコアを機械的に予測するタスクは現実的な問題だと言えるだろう。以下の実験では，この手動スコア付き音声コーパスを主に扱う。

3.3　DNN-GOP における計算方法の変更

2.3.1 節では DNN-GOP を用いたシャドーイング音声の自動評価す手法について説明した。先行研究において DNN-GOP の計算は式 (2.4) で計算されていた。ある発話 x の GOP を計算する場合，各フレームの GOP の総和を計算し，その発話の継続長すなわち全フレーム数で割って平均をとることになる。しかしながら，通常音素ごとにその継続長は異なる。特に母音と子音を比べた場合，母音の継続長のほうが長くなることは想像できる。図 3.1 に英語音声のアライメント結果を示す [2]。冒頭から見ていくと初めに/SIL/の文字があるこれは，無音（Silence）を表す。その後,/HH/,/AY1/と続いている。数字の 1 は第一強勢母音であることを表している。この/HH AY/の部分は英語の「Hi」に相当する。ここでそれぞれの継続長をみてみると，/HH/に比べ/AY1/のほうが 3 倍ほど長いことがわかる。他の母音と子音を比べてみてもそのような傾向があることは明らかである。

さて，このように母音と子音では継続長がかなり違ってくる。そのため，GOP を計算する場合に発話の全フレーム数で平均化した場合に母音の違いを大きく評価していることになる。このような評価方針が人の評価方針と一致しているならば問題ないが，そうでない場合は，自動評価の精度を下げる要因になりうる。

そこで，本研究では GOP を計算する際に一度音素ごとにスコアを平均化し，そのうえで音素数で平均をとる手法を用いることを提案する。このような計算方法自体は，GOP を使用した研究で過去に検討された例があり [17]，今回はそれを DNN-GOP に適用する。本研究では，フーレム数で単純な平均をとったものを fGOP とよび，音素ごとに一度平均をとって再度音素数で平均をとったものを pGOP と呼ぶ。pGOP は次の式で計算される。ただし，N_p は発話中の出現音素数で $\mathrm{fGOP_i}$ は i 番目の音素に対する fGOP で式 (2.4) で計算される。

$$\mathrm{pGOP}(x) = \frac{1}{N_p} \sum_{i=1}^{N_p} \mathrm{fGOP_i} \tag{3.1}$$

[2] praat を使用:http://www.fon.hum.uva.nl/praat/

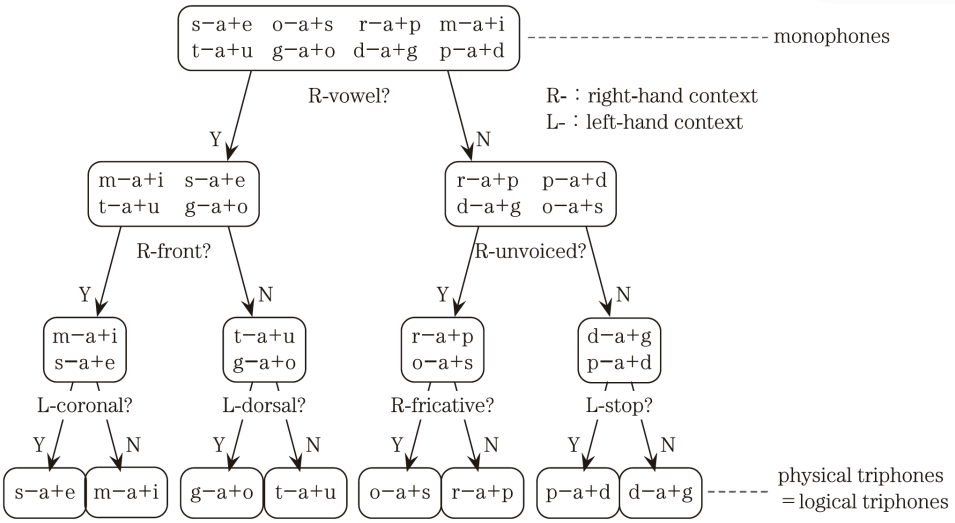


図 3.2: 音素環境に基づくトップダウンクラスタリング [39]

3.4　DNN-DTW における事後確率化のクラス数最適化

[36] では，音素事後確率ベクトル同士の DTW による自動評価手法を提案していた。この事後確率ベクトルの次元が数千におよぶことはすでに説明した。先行研究では，クラス数約 3000 において実験を行っていたが果たしてこの 3000 という数は最適なものなのだろうか。特に，先行研究において日本語の音響モデルを用いて事後確率化を行った場合にクラス数が約 3000 のものと約 9000 のものでは手動スコアとの相関に大きな違いが見られていた。このような結果から，どこかに最適なクラス数が存在することが予想される。クラス数が 3000 ということは音の違いを 3000 個に分類した上で音声同士の比較を行っていることになる。この数が大きくなればより細かい違いを考慮して比較を行い，小さければ大雑把な比較を行うということになる。また，DTW は局所距離の計算に内積計算を含むため次元数が多いと計算量が膨大になり，リアルタイムにスコアをフィードバックすることが難しくなる。本研究では，シャドーイング評価において最適なクラス数を調査することを目的とするため，計算時間等の議論は行わないが，事後確率ベクトルのスパース性を利用した高速化について検討されたものがある [40]。

では，どのようにして事後確率化のクラス数は決定されているのか。クラスのラベルにはトライフォンを適度に共有したものが用いられることは説明した。仮に，音素数が 40 だった場合 40^3 のトライフォンが存在し，HMM は通常 3 状態で構成されるためさらにその 3 倍の数の HMM 状態が存在することになる。この膨大なクラス数を減らすために，音素決定木によるトップダウンクラスタリングが用いられる。図 3.2 に音素決定木の一例を示す。この例では，まず「後ろ（時間的に次）の音素が母音であるか」という質問を用意しあてはまるものとそうでないもので 2 つに分割している。さらに,「無声音であるか」など弁別素性に基づく質問を用意し，最終的には全てのトライフォンは別々のクラスタとして分割される。仮にトライフォンを共有せずにそのまま扱った場合，訓練サンプルが不足し過学習が起きてしまう。音声認識ではおおよそ数千程度の状態数になるようにクラスタリングが行われる。このクラス数は GMM の音響モデルを学習する際に事前に決定される。そのため，このクラス数を適当に変更して音響モデルの学習を行えば様々なクラス数の DNN 音響モデルを構築することができる。本研究ではそのような試行を繰り返して最適なクラス数を調査することを目的とする。

3.5　回帰モデルによる自動評価の高精度化

さて，先行研究では DNN-GOP や DNN-DTW と手動スコアとの相関が話者単位で高いことが主張されていた。確かに，学校のクラス分けテストで数文の音声をシャドーイングしてもらいそのスコアを自動的に算出するといった用途であれば，話者単位での精度が十分あれば問題ないかもしれない。しかしながら，語学学習の現場で実際に使うことを考えると，一文一文に対して高い精度で自動評価を行えたほうが学習者のモチベーションにもつながる上，リアルタイムなフィードバックも可能となる。本研究では，自動評価の高精度化を目指して GOP，DTW 以外の特徴量の追加とそれらを用いた回帰モデルの構築を行った。

3.5.1　各種特徴量の追加

自動評価の精度が低い原因として，予測するための変数が少ないことが挙げられる。機械学習を用いた予測タスクでは複数の特徴量を用意することでその精度を上げることがしばしば行われる。本研究でもそれに倣い，GOP・DTW 以外の特徴量を追加する。今回追加で用意した特徴量は次の 3 種類である。

- 音素クラス別 GOP
- 単語認識率（WRR, Word Recognition Rate）
- 無音率（シャドーイング音声中の無音時間を音声時間で割った比率）

まず，音素クラス別 GOP について説明する。pGOP の計算では音素を単位として計算していたが、音素はさまざまな観点からさらに細かく分類することができる。例えば調音の観点から分類すれば、有声音・無声音・母音・子音などが挙げられる。また、英語の場合、強勢の位置によって第一強勢・第二強勢といった分け方もできる。本研究では、このような種類別 GOP を計算することで、これまで大雑把に見てきた発音の部分についてより細かい分析を行う。今回は、通常の GOP に加え、母音・子音・強勢なし・第一強勢・第二強勢の 5 つの種類別 GOP を計算する。また，母音の強勢位置については CMU pronunciation dictionary [1] を参考にした。特徴量の多様性を増やす目的もあるが，音声学的な観点から分類した特徴量を追加することによってどの部分が手動評価において重視されているかを調査する目的もある。

次に単語認識率について説明する。音響モデルはもともと言語モデルと組み合わせて用いる音声認識のためのモジュールであった。GOP などを計算している音響モデルも音声認識システムを構築する過程でできたものであり，同様の環境で音声認識を行うことが可能である。通常音声認識システムの評価においては単語誤り率（WER, Word Error Rate）が用いられる。今回はスコアの予測を行うため，正の相関を持つように WRR を用いる。 3.3 に WRR の説明図を示す。

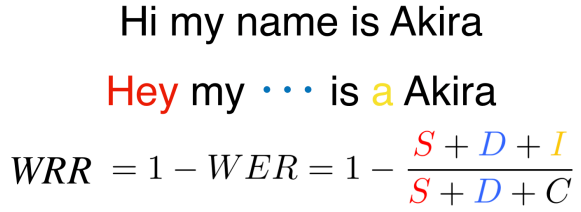


図 3.3: WRR の計算式

図中上の「Hi my name is Akira」は正解となる単語列でその下の「Hey my … is a Akira」となっているのが音声認識の結果である。単語の誤りは置換（S, Subsutitute），削除（D, Deletion），挿入（I, Insertion）の 3 つに分類される。置換とは単語そのものが別の単語に置き換わってしまうこと，削除はあるべき単語が削除（省略）されてしまうこと，挿入は不必要な単語が挿入されてしまうことを示す。C は正解（Correct）を表している。GOP では，文を既知とした上で処理を行うが WRR では音響モデルと言語モデルのスコアをもとに単語仮説を生成するため，微妙に異なる観点の特徴量 [3]となりうる。

最後に無音率について説明する。 2.3 章で説明したように，シャドーイングにおける誤りの多くは単語の脱落であった。また，単語の途中までは言えたが途中から言えなくなる場合もある。つまり，シャドーイングが上手くできないということは，黙ってしまう時間が長くなってしまうということになる。そのため，収録音声中の無音時間が長ければ長いほど手動スコアは低くなるはずである。しかし，単純に無音時間をそのまま扱ってしまうと長い音声ほど無音時間は長くなってしまい，音声ごとに平等な評価ができない。そのため，音声中の無音時間を全音声の時間で割った比率を特徴量として用いる。

音声中の無音時間を自動的に検出するためには，VAD（Voice Activity Detection）が有効であると考えられる。VAD とは音声の発声区間を検出するための技術である [37] が言い換えれば発声していない区間を検出することになる。リアルタイム音声認識などではしばしばこのような技術が用いられる。しかし，VAD は録音環境ごとにしきい値を定める必要があり今回のシャドーイング音声のように収録環境が多様な場合にはあまり適していない。そこで，本研究では全ての単語の発音辞書に無音を追加することで単語そのものが無音に置換されることを許すネットワークで強制アライメントを行う。強制アライメントの結果から無音区間の時間情報を得ることができる。

3.5.2　重回帰分析によるスコア予測

回帰分析とは，ある目標となる連続的な値 Y と入力変数 X の間の関係 $Y = f(X)$ を求めることである。X が 1 次元であれば単回帰，n 次元であれば重回帰と呼ぶ。最も単純な回帰モデルは線形回帰モデルである。X が n 次元のベクトルであるとき，重み w を用いて Y は次の式で表現できると仮定する。

$$Y = w_o + w_1 x_1 + \cdots + w_n x_n \tag{3.2}$$

$$\tag{3.3}$$

ここで，$\boldsymbol{X} = (1, x_1, \cdots, x_D)^\top$ とすると

$$Y = \boldsymbol{w}\top\boldsymbol{X} \tag{3.4}$$

このときの重み $\boldsymbol{w}$ は最小二乗法による最適化で求めることができる。線形回帰では，Y と X の間に線形な関係を仮定するため，表現力はあまり高くない。そこで X に対して非線形の基底関数 $\Phi(X)$ をかけることで非線形な関係をモデル化することができる。$\Phi(x) = x^n$ とすると多項式カーネルとなり n を上げることでより複雑なモデルを構築することができる。$\Phi(x) = \exp\{-\frac{(x-\mu)^2}{2s^2}\}$ とすると，ガウスカーネルとなる。これは，実質的に無限の次元をもつ多項式カーネルとなるため，基底関数としては最も表現力が高い。

線形回帰は最も単純なモデルであるが，そのほかにも様々なモデルが存在する。最近では機械学習向けのライブラリが発達しており比較的容易に様々な種類の回帰モデルを構築することがで

[3]ネイティブに正解を見せず書き取らせたらどのくらい書きとれるか，のシミュレーションとなる。

きる。本研究では，いくつかの回帰モデルを構築することで手動スコアを予測するのに適しているモデルを選定する。

3.6　実験

ここでは，シャドーイング音声自動評価をの高精度化を目的として行った実験について報告する。

3.6.1　実験条件

第 3 章で行う実験について音響モデルなど共通な実験条件について説明する。

- DNN 英語音響モデル
  WSJ コーパスを用いて構築した英語音響モデル。GOP・DTW の計算に用いた。
- DNN 日本語音響モデル
  CSJ コーパスを用いて構築した日本語音響モデル。DTW の計算に用いた。

まず，DNN 英語音響モデルの構築ついて説明する。音響モデルの構築ために必要なのは大規模な音声コーパスとその書き起こしである。本研究では，大規模な英語音声コーパスとして代表的な WSJ（Wall Street Journal）コーパスを用いた [7]。WSJ コーパスは音声認識システムの構築を目的として収録されたコーパスで，英語ニュース記事の読み上げ音声で構成されている。WSJ は WSJ0 と WSJ1 の 2 つにわかれており，合計で約 83 時間の音声で構成されている。

DNN 英語音響モデルの構築には，オープンソースの音声認識ツールキットである KALDI を用いた [20]。DNN 英語音響モデルは大まかに次のような手順で行われる。

1. GMM 音響モデルの構築
2. 話者適応 GMM 音響モデルの構築
3. DNN 音響モデルの構築

まず，GMM 音響モデルの構築について簡単に説明する。DNN 音響モデルを作るために，フレーム単位でラベル付けされたデータセットが必要になる。これを人手で容易することは難しい。そこで通常，GMM で音響モデルを構築し，GMM 音響モデルでアライメントをとり（ラベル付けを行い）その結果を用いて DNN の音響モデルを構築する。GMM 音響モデルでは入力特徴量に MFCC とその動的特徴量である Δ と $\Delta\Delta$ を用いる。また，前処理として CMVN（Cepstrurm Mean Variance Normalization）を行う。これは，MFCC の平均・分散ベクトルを正規化する処理であり，録音環境などの違いによる変化を抑制する効果がある。初めに，モノフォンで構成された音響モデルを学習しアライメントを生成する。その結果を用いてトライフォンの音響モデルを学習する。

次に，話者適応 GMM 音響モデルの学習を行う。前にも説明した通り，MFCC は同一音素でも話者などによって微妙に異なってくる。GMM でモデル化することで，その多様性を許容していたが，学習データにないものに対する精度は低くなってしまう。そこで，特徴量に対して変換をかけることで話者の正規化を行うことがある。KALDI のデフォルトレシピでは，MFCC を数フレーム連結し，LDA（Linear Discreminant Analysis）により次元圧縮を行い，MLLT（Maximum Likelihood Linear Transform）によりベクトル同士の相関を削減し，fMLLR（feature-space Maximum Likelihood Liner Regression）を適用することで話者の特性を正規化した特徴量を作成

している。詳細については，[22] を参考にしてもらいたい。この話者正規化特徴量（以下，fMLLR 特徴量）を用いて GMM を学習することで話者適応 GMM を構築することができる。

最後に，話者適応 GMM 音響モデルのアライメント結果＋ fMLLR 特徴量を用いた DNN の学習を行う。DNN の学習は RBM（Restricted Boltzman Machine）による事前学習と誤差逆伝播法によるファインチューニングから成る。また，音響モデルとしての DNN のタスクはクラス分類であるため，損失関数にはクロスエントロピー誤差が用いられる。クロスエントロピー基準の学習では，フレーム単位（各時刻ごと）で誤差を最小化する。しかしながら，音声認識で目的とするところは単語誤り率を最小化することである。そのため，時系列を考慮した損失関数を定義したほうが精度がよくなることが考えられる。KALDI の nnet レシピ [4]では，sMBR（sequential Minimum Bayes Risk）基準による学習 [29] を最後に行うことで音声認識の精度を上げている。最終的に，評価データに対する単語誤り率は 3% 程度になる。

日本語音響モデルについても同様の方法で音響モデルの構築を行った。用いたコーパスは日本語話し言葉コーパス（CSJ, Corpus of Spontaneous Japanese）で，学会公演と模擬公演音声の合計 600 時間ほどの音声コーパスである。表 3.6 に DNN 英語・日本語音響モデルのネットワーク構成を示す。ここに示すモデルは，先行研究で使用されていた音響モデルと同一のものである。

表 3.6: DNN 音響モデルのネットワーク構成

モデル	入力次元数	中間層	出力クラス数
英語音響モデル	440	6 層	3458
日本語音響モデル	1400	6 層	2856,9429

3.6.2　音素平均 GOP と手動スコアの相関検証実験

3.3 節では，音素単位で平均をとる音素平均 GOP を取ることを提案した。実際に検証するために，既存手法である fGOP と pGOP を全シャドーイング音声について計算し，手動スコアとの相関を計算した。表 3.7 に話者単位で計算した各 GOP と手動スコアとの相関を示す。

表 3.7: 各種 GOP と手動スコアとの相関（話者単位）

GOP	**P**	**S**	**C**	**P+S+C**
fGOP	0.74	0.83	0.71	0.83
pGOP	0.79	0.84	0.78	0.88

結果として全てのスコアにおいて相関が上昇した。このことから，音素継続長を考慮して正規化することの有効性が確かめられた。また，超分節的な観点である S スコアとの相関はあまり上昇していない一方で，P や C など，音素的な部分に着目しているスコアのほうが上昇度が高かったことから，音素レベルでの細かい評価での精度が上昇したことが伺える。これ以降の実験では，fGOP ではなく，pGOP を積極的に使用していく。

[4]KALDI は Github で管理されているため，最新版のレシピではデフォルトが別のレシピに変更されている場合がある。

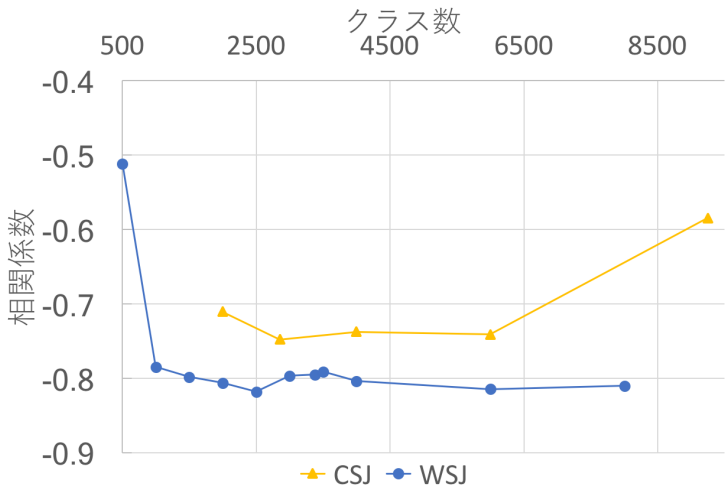


図 3.4: DTW と手動スコアとの相関

3.6.3　DNN-DTW のクラス数最適化と手動スコアとの相関検証実験

次に，DNN-DTW におけるクラス数変更実験を行った。先行研究では，表 3.6 に示した英語・日本語音響モデル 3 つを用いて音声を事後確率化していた。本研究では，両者ともに様々なクラス数で音響モデルを構築することで手動スコアとの相関の変化を調査する。それぞれ，先行研究で用いられていたクラス数を基準に増加・減少させ，手動スコアとの相関が最大になるようなクラス数を実験的に調査した。一定期間続けての相関の減少が見られた場合はそこで終了とする。図 3.4 に話者単位でのスコア相関の変化を示す。

距離が大きいほど手動スコアは下がるため、負の相関となる。WSJ においては、クラス数が 2000〜8000 の間で高い相関がでており、わずかではあるが 2500 でピークがでていた。500 で極端な落ち込みが見られたが，クラス数が少ないと、DNN 学習時の精度そのものが悪くなる場合があり、その影響も考えられる。CSJ においては、2000 での相関は若干悪くなっていたが，それ以降は大きな変動はなく，クラス数を極端に上げると大きく相関は下落した。クラス数増加による相関の下落は WSJ では観測されなかった。WSJ と CSJ を比較すると，クラス数は 2500 以上あれば相関値は安定し，差は 0.05 程である。日本語音素数は米語音素数よりも少ない。本研究では，日本人の英語学習を対象としているが，外国人の日本語学習者を対象とし，これを米語 DNN で評価した場合に同様の差が観測されるか否かは興味のあるところである。なお，クラス数増加による CSJ の精度劣化は，言語差によるものと解釈される。

3.6.4　回帰モデル構築と相関検証実験

3.5.2 で導入した特徴量を用いて，回帰モデルの構築を行う。初めに，特徴量別に手動スコアとの相関を計算した。表 3.8 にその結果を示す。表中の vGOP,cGOP はそれぞれ母音（Vowel）のみ，子音（Consonant）のみについて計算した pGOP である。v0GOP,v1GOP,v2GOP はそれぞれ第 1 強勢，第 2 強勢，強勢なしの母音について計算したものである。母音・子音 GOP において P(Phoneme) に対する相関に大きな差があることがわかった。全体的に見ても，子音 GOP は音素平均 GOP と同程度の相関を示した。日本語と英語を比較すると、子音体系よりも母音体系の方が差が大きく、その意味では母音 GOP の方が相関が高くなることが予想されたが，結果は逆であった。不適切な子音生成の方が評価者にとって「耳障り」であったことが示唆されるが，発

表 3.8: 各種特徴量と手動スコアとの相関（話者単位）

特徴量	**P**	**S**	**C**	**P+S+C**
fGOP	0.74	0.83	0.71	0.83
pGOP	0.79	0.84	0.78	0.88
vGOP	0.70	0.83	0.70	0.81
cGOP	0.79	0.82	0.78	0.87
v1GOP	0.63	0.78	0.64	0.75
v2GOP	0.42	0.41	0.43	0.46
v0GOP	0.71	0.75	0.78	0.78
DNN-DTW	-0.66	-0.84	-0.69	-0.80
無音率	-0.34	-0.21	-0.29	-0.30
WRR	0.79	0.81	0.71	0.84

音指導に際して母音と子音のどちらに力点を置くべきかなど，検討事項である。第二強勢の相関が低い理由として、一つは文によっては第二強勢音素が出現しないものがあること、もう一つは GOP スコアがほぼ 0 である音声が半数を占めており分布が極端化していることが挙げられる。無音率はわずかに相関が見られたが強い相関ではなかった。相関が低い理由として，4 回目のシャドーイング音声を採点対象としたため，ある程度黙ることなくシャドーできていたことが考えられる。また，WRR も一定の相関があったが，pGOP などに比べると多少低くなっている。

次に，これらの特徴量を説明変数として回帰モデルの構築を行った。GOP については従来法より相関の高かった pGOP を用いた。v2GOP については文によってはスコアが 0 になることがあるため除いた。表 3.9 に使用したモデルと説明変数および目的変数を示す。

表 3.9: 回帰モデルの実験設定

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

Lasso（Least absolute shrinkage and selection operator）とは，通常の線形回帰に正則化項を加えたもので，過学習を防ぐことができる。また，L1 正則化を行うと，推定したパラメータの一部が 0 となりスパースな解が得られる。つまり，重回帰のように説明変数が多く存在するときにそれらの一部の重みを 0 にすることで特徴選択をしてくれることになる。SVR（Support Vector Regression）は，SVM（Support Vector Machine）を回帰に応用したもので，ガウスカーネルを用いることで高い表現力を持つ。RadomForest は通常，分類のタスクに用いられるが，値を細かく離散化することで回帰を分類問題として解くことができる。回帰モデルの構築には scikit-learn [2] を用いた。モデルの学習および評価には 4 分割交差検定を使用した。また，モデルにハイパーパラメータが存在する場合は適宜グリッドサーチを用いて最適なパラメータを設定した。

表 3.10 に回帰モデルの予測スコアと手動スコアとの話者単位での相関を示す。

fGOP はベースラインとして示している。モデル別に結果を比較すると，SVR が全体的に高い精度があることがわかる。Lasso についても C を除いて SVR と同等程度の精度があり，線形な

表 3.10: 回帰モデルの予測スコアと手動スコアとの相関（話者単位）

モデル	**P**	**S**	**C**	**P+S+C**
fGOP [36]	0.74	0.83	0.71	0.83
Lasso	0.84	0.89	0.76	0.90
SVR	0.85	0.89	0.83	0.89
Random Forest	0.77	0.84	0.79	0.86
評価者間相関	0.77	0.69	0.86	0.87

モデルでも十分な精度が出る事がわかる。Random Forest はベースラインに比べると高い精度であったが，他のモデルに比べると多少精度が劣る。合計スコアでみると Lasso モデルが 0.90 と最も高い相関があった。また，表の一番下に評価者間相関を示している。これは表 3.5 の評価者同士の相関の平均をとったものである。ベースラインでは，評価者間相関を超える精度はでていなかったが回帰モデルを用いることで，C スコアを除くスコアで評価者間相関を超える相関がでている。

同様にして計算した文単位での相関を表 3.11 に示す。

表 3.11: 回帰モデルの予測スコアと手動スコアとの相関（文単位）

モデル	**P**	**S**	**C**	**P+S+C**
fGOP [36]	0.64	0.64	0.53	0.68
Lasso	0.68	0.73	0.65	0.77
SVR	0.70	0.73	0.68	0.78
Random Forest	0.67	0.68	0.61	0.74
評価者間相関	0.58	0.54	0.74	0.75

話者単位で計算した場合と同様に，ベースラインに比べて相関が上昇していることがわかる。話者単位で相関を計算する場合平均化によって回帰モデルを使わない場合でも高い相関がでていたが，文単位では回帰モデルを使う場合とそうでない場合で顕著な差がでている。話者単位の場合と同様に，SVR・Lasso で高い精度がでており，合計スコアに対する相関では，SVR が最も高く 0.78 であった。これは評価者間相関の 0.75 よりも高く，文単位で十分な精度の評価ができているといえる。また，スコア C についてのみ相関が低い結果となっている。これについては， [16] で別な枠組みの実験を提案し，考察を述べている。

以上の結果から回帰モデルを用いることで高精度に自動評価ができることがわかった。しかしながら，様々な特徴量を用いたためにどの特徴量がスコアに影響を与えているのかが不明瞭になってしまった。教育的な観点からみても，どのような特徴量がスコアに影響しているかを知ることは有意義である。先程述べたように，Lasso 回帰は特徴選択の特性があるため，不要な特徴量の回帰係数は 0 になる。したがって，Lasso 回帰における特徴量ごとの重みから，その特徴量の重要度を知ることができる。表 3.12, 3.13 に Lasso 回帰において重みが大きかった特徴量の上位 3 つを示す。ただし，入力特徴量は全て正規化・標準化の処理を行っている。

まず，我々が主な評価指標として使用している pGOP や DTW についてみると，WRR などの特徴量に比べ重要度が高い事がわかる。特に，DTW は合計で 7 回表にリストされている。一方で，WRR はスコア P においてのみリストされている。WRR は言語モデルに依存するため，シャ

表 3.12: 重要度の高い特徴量上位 3 つ（話者単位）

	P	S	C	P+S+C
1	pGOP	DTW	pGOP	pGOP
2	WRR	vGOP	DTW	v1GOP
3	v1GOP	無音率	cGOP	DTW

表 3.13: 重要度の高い特徴量上位 3 つ（文単位）

	P	S	C	P+S+C
1	pGOP	DTW	DTW	DTW
2	DTW	pGOP	cGOP	cGOP
3	WRR	cGOP	無音率	無音率

ドーイングするコンテンツによって変化してしまう。発音評価のタスクでは，文章（発話内容）を既知として GOP や DTW などの手法が適しているといえるだろう。本研究では，GOP については様々な亜種を導入したが，DTW については特になにも行っていない。今後，DTW についてもいくつかの亜種を導入することを検討するべきだろう。

3.7　まとめ

本章では，DNN-GOP において音素ごとに平均をとる pGOP を導入し精度の向上を図った。さらに，DNN-DTW においてそのクラス数を変化させて，自動評価に最適なクラス数を調査した。最後に，複数の特徴量を用いた回帰モデルを用いることで自動評価の精度を向上させ，評価者間相関を上回る精度を達成した。

第4章

シャドーイング音声自動評価の耐雑音性向上

4.1 はじめに

前節では，シャドーイング音声の高精度化に向けて回帰モデルの構築などを行った。次に目指すべき目標は実際の教育現場で，我々の自動評価手法を導入することである。ここで問題になることが想定されるのが，教室などで音声の録音を行った場合に他人の音声が収録されてしまうことである。実際，これまでの実験で使用してきたシャドーイング音声コーパスのうち 4 割ほどの音声は教室で収録されたものであり，そのような雑音が含まれるものが存在する。DNN-GOP や DNN-DTW において音声を事後確率に変換する際に，学習データは雑音のない音声であるのに対して，雑音のある音声を入力すると事後確率の推定の精度が落ちてしまう。その場合，当然自動評価の精度も下がってくることになる。本研究では，雑音環境下でも頑健に動作するシャドーイング自動評価システムの構築を目指す。本章では，はじめに音声における雑音について簡単に述べ，音声認識システムに及ぼす影響について述べる。その後，音声認識において雑音除去・抑圧に用いられている代表的な手法をいくつか取り上げ，我々の提案する自動評価手法に取り入れる。

4.2 音声における雑音とその抑圧

雑音とはどのようなものを示すのか。ここでは，音声における雑音について簡単に紹介する。その上でそれらの雑音を抑制するための手法について述べる。まず，雑音の種類は次ように大きく 4 つに分類できる [38]。

- 加算性雑音: 周囲雑音。例）エアコン・換気扇の音。話し声など。
- 乗算性雑音: 回線歪みなど音声の伝達中に受けるもの。例）電話・無線音声。反響。
- 定常雑音: 時間的な変化が少ない雑音。例）エアコン・換気扇の音。電源ノイズ。
- 非定常雑音: 時間的な変化が大きい擦音。例）話し声。

加算性雑音とは音声信号にそのまま重畳される雑音のことである。所望の音声信号以外の周囲音声は全て周囲雑音となる。音声信号を $x(t)$ 加算性雑音を $n(t)$ とすると，雑音を含む音声 $y(t)$ は次のように表される。

$$y(t) = x(t) + n(t) \tag{4.1}$$

実際に音声と雑音を個別に収録し波形同士を足し合わせると，雑音の混ざった音声を作ることができる。一方で，乗算性雑音とは音声信号が伝達されるときにその伝達経路の特性などにより生

じる雑音である。例えば，電話や無線においてダウンサンプリングされたり，伝送路の損失により微妙に異なる音声信号に変化したりするものが当てはまる。$h(t)$ を伝搬による歪みを表すフィルタとすると, 乗算性雑音による音声信号は次のように表される。

$$y(t) = x(t) * h(t) \tag{4.2}$$

$*$ は畳み込みを表す[1]。これら加算性雑音と乗算性雑音はそれぞれさらに定常雑音と・非定常雑音の 2 つに分けることができる。定常雑音とは時間的な変化 [2]が少ない雑音のことで，エアコンの駆動音など機械的な音が当てはまる。非定常雑音は時間的な変化が大きい雑音で，人の話し声などが当てはまる。特に人の話し声でも，大勢が同時に話していて内容が聞き取れないようなものをバブルノイズという。機械的な音でも工事の音など，音の大きさが頻繁に変化したりする場合は非定常雑音となる。

4.2.1　信号処理による雑音抑圧

さて，これらの雑音が音声に含まれている場合に雑音のみを取り除くことを考える。ここでは簡単のため定常雑音である場合のみ取り上げる。雑音が加算性である場合，理想的には加算された信号をそのものを差し引くことができれば元の音声のみを取り出すことができる。時間領域の波形からそのまま減算する場合と，周波数領域に変換してスペクトルを差し引く場合がある。後者はスペクトルサブトラクションと呼ばれ広く知られている [6]。一方で乗算性雑音は時間領域での処理が難しい。先程示したように，乗算性雑音は時間領域では畳み込み演算になる。そこで，この信号に対してフーリエ変換をかけて周波数領域に変換すると

$$Y(\omega) = X(\omega)H(\omega) \tag{4.3}$$

となり，両辺の log をとると

$$\log Y(\omega) = \log X(\omega) + \log H(\omega) \tag{4.4}$$

と和の形になり，この領域で減算を行えば良いことになる。しかし，乗算性雑音は音声区間・非音声区間に関係なく存在するので，これを求めるのは自明ではない。そこで，対数スペクトルを離散フーリエ変換し，ケプストラム領域において長い区間に渡るケプストラムを計算し，それを乗算性雑音とみなし差し引くことで雑音を除去をする CMN（Cepstral mean normalization）[4] という方法が良く用いられる。非定常雑音についても信号処理的に抑圧する手法が存在するが，複数のマイクを使用することが前提であるため，今回のように単一マイクを使用した収録では考えにくいためここでは割愛する。

さて，ここで紹介した手法は信号処理的に雑音を抑圧するための手法である。しかしながら，音声認識において達成したいことは雑音の有無に関わらず正しい認識結果を出すことであり，必ずしも信号レベルで雑音を除去する必要はない。本研究では DNN 音響モデルを用いた雑音抑圧手法について次のようなものを取り上げる。

- 音響モデルのマルチコンディション学習
- 雑音音声を用いた音響モデルの適応
- DAE を用いた特徴量強調

[1]周波数領域では掛け算となるため乗算性と呼ばれる
[2]ここでは数秒間観測したときの変化のこと

この章の残りではこれらの技術についての説明を行い，ノイズを含むシャドーイング音声の自動評価に適応しその効果を検証する。

4.2.2　音響モデルのマルチコンディション学習

雑音によって音声認識の精度が下がるということは，言い換えると音声認識は学習データにない音声に弱いということである。これは音声に限ったことではないが，機械学習にしても深層学習にしても，学習データを用いてモデルのパラメータを決定し，未知のデータに対して推論を行う。したがってクリーンな音声を使って学習した音響モデルを用いて雑音の混ざった英語音声を認識しようとするとその精度は下がってしまう。一方で，雑音の混ざった音声で学習を行うとクリーンな音声を認識する際の精度が下がってしまう。あるいは雑音音声と言ってもエアコンなどの定常雑音を含む音声を学習データとし，認識対象がバブルノイズを含む音声であった場合，認識率の改善はあまり望めないだろう。そこで，学習データに様々なドメインの音声を用いて学習を行うことがしばしば行われる。これを音響モデルのマルチコンディション学習という。

[23] 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。また，乗算性雑音であ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 [19]。

4.2.3　音響モデルの適応

音響モデルの学習には大量の音声コーパスが必要となり，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しばしば行なわれる。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。 [11] 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。

4.2.4　DAE を用いた特徴量強調

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 [6] 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 [30]。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。DAE の学習は事前学習と誤差逆伝播によるファインチューニングで構成され，事前学習では RBM や DAE を 1 層ずつ学習し積み上げたものを用いる場合もある。また，出力をそのまま利用する DAE の場合中間表現を取り出す必要は必ずしもないため，通常の DNN のように各層の

ノード数を揃える場合もある [31]。音声の分野では MFCC などの周波数領域の特徴量を用いて DAE を構成する。 [15] 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。

4.3　実験

ここでは，シャドーイング音声に対して次の 3 つの耐雑音音声認識手法を適用し，手動評価精度の改善を試みる。

4.3.1　バブルノイズを用いたマルチコンディション学習と相関検証実験

本研究で用いている手動評価付き音声コーパスに含まれている雑音の多くは教室内で収録したことによる人の話し声（シャドーイング）などのバブルノイズであった。したがって，音響モデルを構築する際にバブルノイズを重畳した音声を用いれば，そのような音声についても精度を落とさずに自動評価できることが考えられる。本研究で，英語音響モデルの構築に利用している WSJ コーパスはそれ自体は，通常のクリーンな音声である。そこで，別途用意したバブルノイズのみの音声を WSJ の音声に重畳する。今回は， [28] で公開されている雑音コーパスのうちの英語音声のバブルノイズ音声を使用する。またマルチコンディション学習というと様々な環境での音声を用いて学習することを示すことが多いが，本実験ではバブルノイズ単一でも SN 比を複数用意することでマルチコンディション学習とする。雑音を重畳する際の SN 比については，ノイズレベルの強弱を考え，

- A: 0,5,10,15 dB
- 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。使用する音声コーパスが雑音を含む WSJ 音声コーパスであること以外は 3.6.1 節で説明したものと同様の手順で学習を行った。

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 比別に単語誤り率を計算した。表 4.1 にその結果を示す。モデル「クリーン」は通常の WSJ 音声を用いたベースラインモデルである。

表 4.1: マルチコンディションモデルを用いた音声認識における単語誤り率 [%]

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

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。

次に，マルチコンディションモデルを用いて GOP・DTW を計算した。ただし，マルチコンディションモデルを用いる場合は，ノイズのあるシャドーイング音声についてはモデルとのミスマッチが小さくなることが期待されるが，モデル（お手本）となる音声はクリーンな環境で収録された音声であるため，そこでミスマッチが生じてしまう。そのため，DTW においてはシャドーイング音声の事後確率化についてはマルチコンディションモデルを使用し，モデル音声の事後確率化にはクリーンな音響モデルを用いた。

表 4.2 にモデルごとの GOP・DTW と手動スコアの合計点（P+S+C）との相関を示す。ただし，括弧内の数値はモデル音声の事後確率化にクリーンなモデルを用いた場合の結果である。

表 4.2: GOP・DTW と手動スコアの相関（文単位）

特徴量	クリーン	マルチ
GOP	0.73	0.75
DTW	-0.71	-0.69（-0.74）

GOP に関しては若干ではあるが相関が上昇した，雑音の混ざった音声にに対する精度が上昇したことが予想される。また，DTW については両音声の事後確率化にマルチコンディションモデルにおいて相関が下がっていた。予想通り，モデル音声とのミスマッチによる影響だと考えられる。一方で，モデル音声のみクリーンなモデルで事後確率化を行った場合，相関が 0.74 まで上昇していた。2 つのモデルを用いることで，両音声に対して正しく事後確率化が行えていると考えられる。

4.3.2　シャドーイング音声を用いたモデル適応と相関検証実験

シャドーイング音声を用いてモデル適応を行う。ここでは，手動スコア付きシャドーイング音声とは別に次のような適応用シャドーイング音声を用意した。この場合雑音が含まれた音声であるとともに，日本人の英語音声となるため，複数の側面に対する適応がかかることになる。GOP は滑舌の良さを表す特徴量であるが，それはネイティブ音声で学習した場合であって，日本人英語で学習すると GOP は日本人英語らしさを評価するような特徴量になってしまう。そのため，発音評価の観点からは悪影響を及ぼす可能性があるが，適度に適応がかかったほうが良い可能性もあるため実験的に確かめておく。本実験では次の音声を適応用音声コーパスとして用いた。

- K 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=1272 文音声
- S 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=2376 文音声

合計 3648 文音声を音響モデルの適応に用いた。

音響モデルの適応は基本的に適応用のデータを使って追加学習を行うことであるが，KALDI のデフォルトレシピではクロスエントロピー基準で学習した後に，sMBR 基準の系列学習を行っている。そのため，追加学習についてどちらの基準で学習を行うべきかは理論的には決定しがたい。

表 4.3: 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

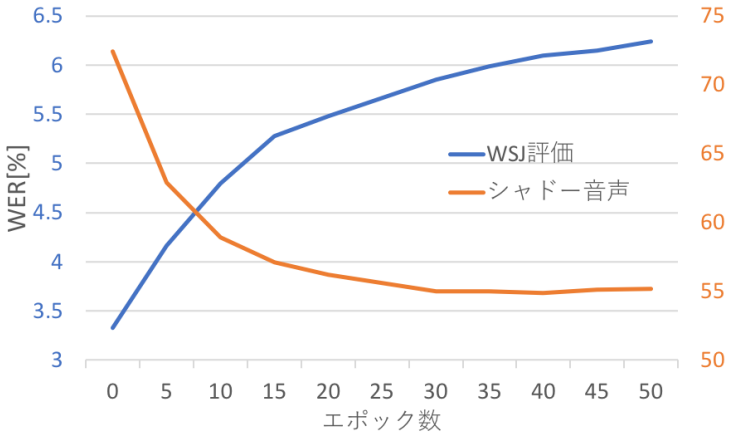


図 4.1: 5 エポックごとの各データセットに対する WER

そこで本実験では，通常のクロスエントロピー基準で再学習した場合と，sMBR 基準で再学習した場合について実験を行う。

sMBR 基準の再学習

はじめに，予備実験として DNN の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 4.3 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更新するものとする。

予備実験の結果を踏まえ，音響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 を評価用シャドーイングデータに対する WER が下がらなくなるまで繰り返す

4.1 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 4.1 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。

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，フレーム正解率向上が 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を行った。表 4.4 に各場合における音声認識の結果を示す。

表 4.4: 単語誤り率（クロスエントロピー基準，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

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 4.5 に各場合における音声認識の結果を示す。

表 4.5: 単語誤り率（クロスエントロピー基準，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

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 が悪くなっているため，必ずしも良いとは言えない。 [11] 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。

相関分析による精度評価

最後に，適応を行った音響モデルを用いて GOP・DTW の計算を行い，手動スコアとの相関を計算した。表 4.6 にモデルごとの GOP・DTW と手動スコアの合計点との相関を示す。ただし，括弧内の数値はモデル音声の事後確率化にクリーンなモデルを用いた場合である。
（C:クリーン, CX:クリーン+クロスエントロピー, CS:クリーン+sMBR，MX:マルチコンディション+クロスエントロピー，MS:マルチコンディション+sMBR）

表 4.6: GOP・DTW と手動スコアの相関（文単位）

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

GOP については，全てのモデルであまり変化がないか相関が減少する傾向があった。予想していた通り，シャドーイング音声が日本人英語であったため，雑音環境だけでなく日本人英語の発音にも適応してしまったと考えられる。また，クロスエントロピー基準よりも sMBR 基準のほうが相関は高い傾向にあり，系列学習の効果は大きいと言えるだろう。

DTW について見てみると，CS モデルにおいて相関が上がっていることがわかる。クリーンなモデルに適応をかけたことでモデル音声およびシャドーイング音声に対する事後確率化の精度が上がったことが予想される。また，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。

4.3.3　DAE の学習と相関検証実験

音声の分野でも，DAE を用いた研究例がある。 [19] では残響化音声認識のタスクに DAE を適応している。本研究ではこれに倣い DAE の構築を行う。

DAE では，ノイズを含むデータからクリーンなデータへの変換を学習する。本実験では WSJ の音声コーパスにバブルノイズを人工的に重畳した。SN 比は $5, 10, 15, 20, \infty$dB とした。（これは 4.3.1 節のマルチコンディション学習で用いたデータセットと同一である。） [19] では MFCC より LMFB（対数メルフィルタバンク）の方がより性能が良くなることが示されている。しかし，現状の DNN では fMLLR 適応を行っており LMFB との組み合わせが難しい。したがって本実験では入力特徴量に MFCC を用いる。DAE の構築には，RBM の事前学習を行い, 続いてバックプロパゲーションによるファインチューニングを行った。表 4.7 にネットワークの構造を示す。音響モデルの入力に与える場合は 429 次元（$13 \times 3 \times 11$）のうち 13 次元を取り出して当該時刻の MFCC として扱う。

次に，DAE の効果を確認するために音声認識実験を行った。音響モデルには 1. 通常のクリーンな音声で学習したモデル（クリーンモデル）2. 以前の実験で用いたマルチコンディションモデルを用いた。また，それぞれの場合においてノイズの混ざった学習用データを DAE に与え，その出力を用いて音響モデルの再学習を行った場合（sMBR 基準 5 エポック）も実験した。認識対象の音声には WSJ の音声に 1. 学習時と同じバブルノイズ 2. 車内（Car）ノイズ 5dB,3. 学習データには含まれないバブルノイズ（Meeting）5dB の 3 つのノイズをそれぞれ重畳した WSJ の評価セットを用いた。表 4.8 に音声認識の単語誤り率を示す。

表 4.7: DAE のネットワーク構造

入力特徴量	MFCC+Δ + ΔΔ 11 フレーム
隠れ層	5 層 1024 ノード
活性化関数	Sigmoid
損失関数	平均二乗誤差
学習率（初期値）	10^{-5}

表 4.8: WSJ 評価セットに対する単語誤り率 [%]

モデル	バブル 0dB	5dB	10dB	5dB	雑音なし	車 5dB	会議室 5dB
クリーン	88.5	46.5	17.5	7.69	**3.33**	3.86	28.5
クリーン+DAE	20.2	8.63	5.12	4.32	3.44	3.99	26.8
クリーン+DAE 再	18.6	8.10	5.02	3.99	3.58	3.76	26.5
マルチ	23.7	8.84	**4.77**	**3.93**	3.40	3.76	**14.0**
マルチ+DAE	**16.1**	**7.46**	4.91	4.06	3.40	**3.63**	19.7
マルチ+DAE 再	18.8	8.05	5.00	3.99	3.58	3.72	26.0

まず，クリーンモデルの結果について見てみると DAE の効果によってクローズドなバブルノイズに対する WER が大幅に改善されていることがわかる。また，DAE 特徴量で再学習を行うことで多少の改善がみられる。一方でノイズなし音声に対する WER は上がっており，トレードオフの関係にあることがわかる。車内ノイズについてはどのモデルの場合もあまり大きな差がなく周波数特性の違うノイズはあまり悪影響がないことがわかる。会議室ノイズは DAE による改善があまり見られず，学習データにないノイズへの効果が弱くなる傾向がみられた。

次にマルチコンディションモデルの結果について見てみると，クローズドなバブルノイズ 0,5dB において DAE 特徴量と組み合わせた場合に WER が最も低くなっている。ノイズが大きいものについてはマルチコンディションモデルと DAE の組み合わせにより WER がより改善されることがわかる。その他の部分については，DAE 特徴量を使う場合とそうでない場合であまり差が生じなかったが，会議室ノイズについてはマルチコンディションモデルでそのままの特徴量を使った場合でも最も WER が低くなっており，マルチコンディションモデルが学習データにないノイズにも対応できていることがわかる。また，DAE 特徴量を用いた再学習による改善はみられなかった。

次に，DAE による特徴量強調を行った上で GOP・DTW を計算した。表 4.9 にモデルごとの GOP・DTW と手動スコアの合計点との相関を示す。

表 4.9: GOP・DTW と手動スコアの相関（文単位）

特徴量	クリーン	マルチ	クリーン+DAE	マルチ+DAE
GOP	0.73	0.75	0.72	0.74
DTW	-0.71	-0.69(0.74)	-0.73	-0.69

結果としてマルチコンディションモデルを上回るものはなかった。音声認識の結果から SN 比が低い部分では DAE との組み合わせにより改善が見られたが，現在使っている音声はそこまで大きな雑音は含まれないため，マルチコンディションモデルで十分対応できているものと思われる。

4.3.4　シャドーイング雑音収録実験

これまでの実験では，手動評価付きシャドーイング音声コーパスを用いて議論を進めてきた。このコーパスのうちに教室内で収録された音声が含まれるため，それらの影響を少なくするために耐雑音音声認識に用いられる技術を導入した。しかしながら，手動評価のために収集したコーパスであったため収録されている雑音について細かい条件設定などはされていない。

そこで，より現実的な場面を想定した教室内雑音収録を行った。収録実験において変化させる条件は主に次の 4 つである。

1. 人数
2. 発話内容
3. 発話タイミング
4. シャドーイング時に参照するテキストの有無

各項目に対して説明する。まず人数ついてだが，周りにいる人の数が変われば収録時の雑音の特性も変化してくる。例えば，少人数であれば発話内容が比較的明瞭な雑音となる。一方，大人数の場合，多くの発声が混在するため発話内容が不明瞭となりガヤガヤとしたノイズ（バブルノイズ）となる。

次に，発話内容についてだが，通常の会話などを考えるとある特定の人の音声を収録したい場合にその周りに同じ発話内容を発声している人がいるという状況はあまり多くない。しかし語学教育の現場では比較的頻繁に生じる。例えば，教師が提示した文や音声を学習者全員で読み上げるといったことがある。シャドーイングもこれに当てはまる。当然，発話内容が同一の場合は雑音を除去することが難しくなってくる。

3 つ目の発話タイミングについてだが，これは発話内容が同一である場合のみ考慮するべき条件となる。発話内容が同一であってもそのタイミング（時間的な位置）がずれれば全体としては不明瞭な音声になる。しかし，シャドーイングは通常モデル音声のペースに合わせて復唱を行う。そのタイミングに微妙な個人差はあれど，ずれは 1 秒程度に収まるだろう。特に，開始するタイミングを完全に揃えれば全員が同じ発話内容を同じタイミングで話すことになり，雑音としては除去することがさらに難しくなってくる。

最後のシャドーイング時に参照するテキストの有無についてだが，これは単純にテキストがあると大きな声になり，ないと小さな声になるだろうという予想の元設定した条件である。通常シャドーイングはテキストを見ずに音のみを聞いて行われるが，あまり慣れていないとシャドーできずに黙ってしまうことが考えられる。テキストを見せることによって聞き取れないことで黙ってしまうことをできるだけなくすようにできる。こうすることで学習者の声量をある程度コントロールでき，雑音が大きくなる場合と小さくなる場合をコントロールすることができる。

以上のことを踏まえ収録実験の条件は表 4.10 のように設定した。人数については，当初予定していた人数よりも収録実験当日の参加者が少なかったので，減らし方に一貫性がなくなってしまっているが，おおよそ半分づつ減らすようにしている。また，4,8 人などの少人数の場合は参加者のシャドーイングが極端に小さいとほとんど音声が収録されない可能性があるので，実験前あるいは実験中にある程度大きな声でシャドーできている人を選定し，その人らを残すようにした。

表 4.10: シャドーイング雑音収録実験における諸条件

人数	発話内容	発話タイミング	テキスト
35,14,8,4	揃える・揃えない	揃える・揃えない	あり・なし

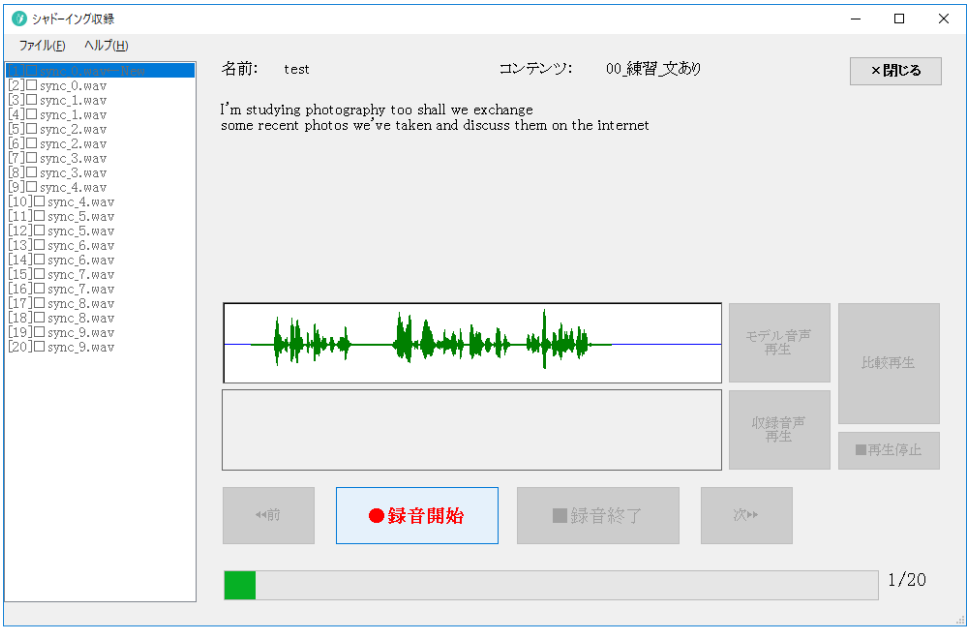


図 4.2: 収録ソフトの再生・録音画面

収録環境と設備

上記で説明した条件以外に収録環境について説明しておく。まず，収録場所についてだが，より現実的な環境での収録を行うため，都内にある大学の CALL 教室で行った。CALL 教室は最大で 60 名収容でき，各机には個人用 PC およびヘッドセットが取り付けられている。また，教員が示したスライドなどを見るための参照用のモニターが設置されている。参加者は PC の前に座ってヘッドセットを装着してシャドーイングすることになる。また音声の収録には我々の研究室の OB が作成したシャドーイング収録ソフトを用いた。図 4.2 に収録ソフトのスクリーンショットを示す。本来，音声の録音のために作られたものであるが，今回は主に再生用途で使用した。録音開始ボタンを押すとシャドーイング用の音声が再生される。参加者にテキストを見せたい場合は，画面上部にテキストを表示することができるようになっている。また，音声を連続で再生することができるため，開始タイミングを揃えたい場合は最初のタイミングさえ揃えれば続けて再生される音声のタイミングも自然と揃うようになる。肝心の録音についてたが，今回の収録では時間的な同期をとる必要がある。そこで，同様のソフトを録音者にも使用してもらい，他の参加者と同じタイミングで録音開始・終了ボタンを押してもらうようにした。ただし，録音者はシャドーイングをしない。図 4.3 に実際の録音の様子を示す。

録音用のマイクはイヤーフック型ものを用いた。マイクの中では比較的他人の声が混入しづらいものである。シャドーイングする文章は発話内容を揃えない場合については，20 秒程度の音声を用意し，発話内容を揃えるものについては，手動評価を行った 10 文の音声と同じものを用い

図 4.3: 収録の様子

た。これにより，収録した雑音を手動評価付き音声に重畳したときの影響を調査することができる。発話タイミングを揃える場合については，参照用モニターに時刻 [3]を表示し「時刻が XX 秒になったら開始」とすることでタイミングを揃えた。また，人数を減らす場合はマイクの位置からできるだけ離れた座席に座ってもらうようにした。

実際の収録は大まかに次の様な手順で行った。

1. 発話内容を揃えずにシャドー
2. 発話内容を揃えて，自然なタイミングでシャドー
3. 発話内を揃えて，タイミングも揃えてシャドー

上記手順を 1. 文章を見ながら，2. 文章を見ないで行う。これを 1 セットとして 1 セット終了ごとに人数を減らして再度同じ手順を繰り返した。したがって参加者は最大で 4 セット繰り返すことになる。

4.3.5　シャドーイング雑音と耐雑音手法による効果

収録した雑音ををもともと雑音の少ない手動評価付き音声（約 700 発声）に重畳した。初めに，人数の違いおよび発話内容の一致などの条件がどのくらい影響を及ぼすかを見るために，SN 比は 5dB で固定とした。まず，何も対策をしていないクリーン音響モデルで GOP・DTW の計算を行い，手動スコアとの文単位の相関を計算した。結果を図 4.4a， 4.4b に示す。

まず，GOP についてはノイズなしの元音声では相関が 0.74 ほどあったものが全体的に 0.5 前後に下がってしまっている。意図的に混ぜた雑音ではあるがクリーンなモデルだとかなり精度が落ちてしまうことがわかる。また，発話内容がバラバラなものを “ランダム”，発話内容が同一であるが開始タイミングがずれているものを “非同期”，発話内容が同一でかつ開始タイミングもそろっているものを “ 同期” として示している。特に，雑音の種類による傾向は見られないが，ノイズの種類によって若干の差が生じていることがわかる。また，人数の違いによる傾向も特には

[3]Time.is `https://time.is`

見られなかった。

DTW についてみると，GOP に比べて全体的に相関が低く，0.4 以下に減少してしまっていることがわかる。DTW では，音素事後確率の全ての次元を用いて計算を行っているため，雑音による影響が出やすいと考えられる。GOP の場合と同様に，人数や雑音の種類による傾向は見られなかった。

続いて，本研究で検討した耐雑音手法をこれらの音声に対して適用した。今回実験したのは次の 3 つである。

1. マルチコンディションモデル
2. 適応モデル（クリーン+sMBR 基準）
3. DAE+マルチコンディションモデル

これらを用いて GOP・DTW を計算し，手動スコアとの相関を計算した結果を図 4.4 に示す。

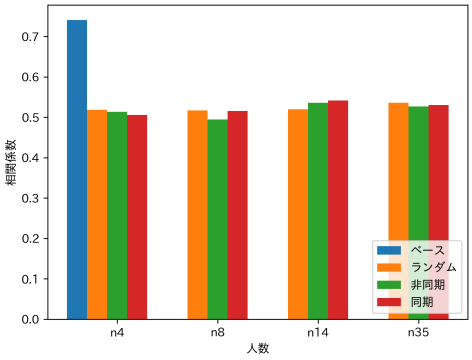
順番にみていくと，マルチコンディションモデルでは GOP については相関が約 0.1 ほど減少している。クリーンなモデルでは 0.5 程度まで下がっていたので，それと比較すると雑音の影響が軽減できている。DTW については，全体的に相関が 0.5 程度に下がっているが，クリーンなものと比べると 0.1 ほど高い。また，このときの事後確率化にはモデル音声だけクリーンな音響モデルを用いている。また，ランダムな雑音より同期・非同期雑音の場合に相関が高くなる傾向が見られた。ランダムな雑音では，事後確率ベクトルの分布が平坦になってしまうため，同期・非同期雑音に比べ局所距離が大きくなりやすいことが考えられる。

適応モデルでは GOP・DTW ともに相関が 0.5 前後となっている。GOP については，雑音の影響は抑制できているが日本人英語に適応してしまっていることが原因だろう。

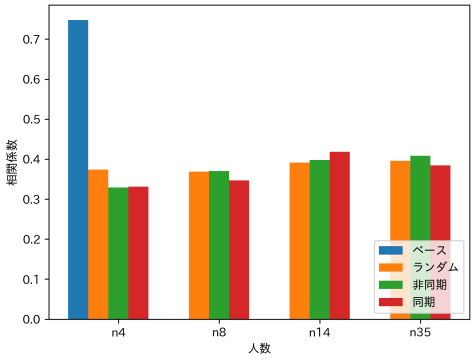
DAE とマルチコンディションモデルを組み合わせた場合では，GOP・DTW ともにマルチコンディションモデルのみの場合に比べて相関が減少していた。DAE との組み合わせにより，精度が上がることを期待していたが結果は逆であった。また，DTW においてランダムなシャドーイング雑音の場合に相関がより低くなる傾向がみられた。ランダムな雑音は学習データに用いたバブルノイズに特性としては近くなるので，他の雑音に比べ影響を少なくできると考えていたが逆の結果となった。本研究では，雑音除去後の音声の再合成が行っていいないため，聞いて判断することはできないので，事後確率ベクトルの分布の変化などを調査してみる必要がある。

4.3.6　まとめ

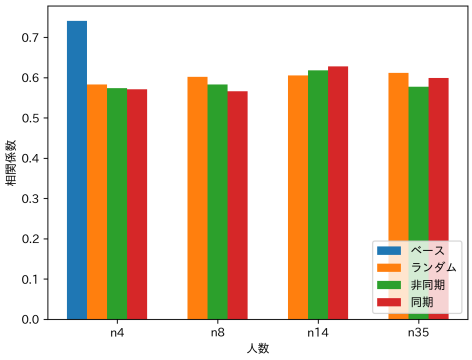
本章ではシャドーイング音声自動評価において，実用化を踏まえ，耐雑音性の向上を検討した。音声認識で用いられている耐雑音手法である 1）マルチコンディション学習，2）モデル適応，3）DAE による特徴量強調を検討し，雑音を含むシャドーイング音声評価に対して効果を検証した。その結果，いくつかの場合で精度が上がることがわかった。最後に，より現実的な場面を想定してシャドーイング雑音の収録実験を行った。また，その雑音を手動評価付き音声コーパスの一部に重畳し，その影響を調査した。さらに，本実験で検討した耐雑音手法を適用し一定の効果があることを検証した。



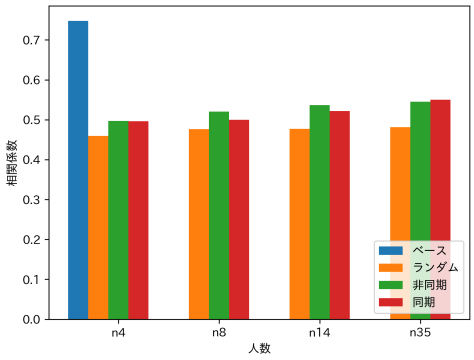
(a) GOP，クリーン



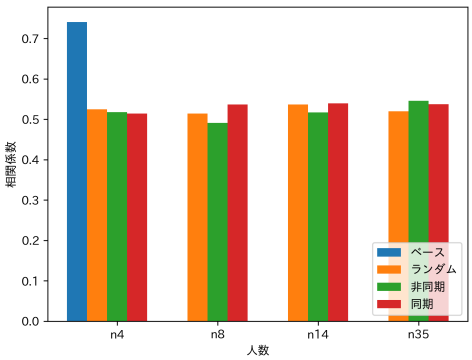
(b) DTW，クリーン



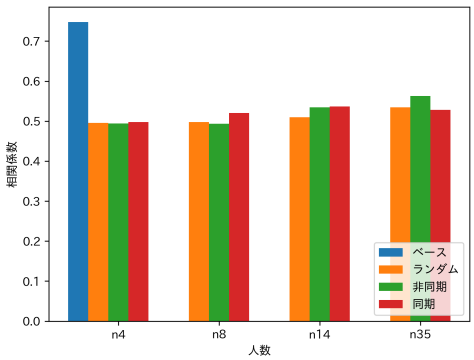
(c) GOP, マルチコンディションモデル



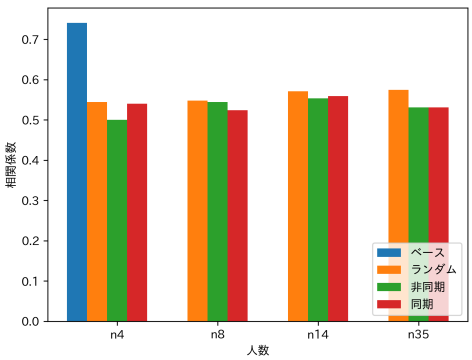
(d) DTW, マルチコンディションモデル



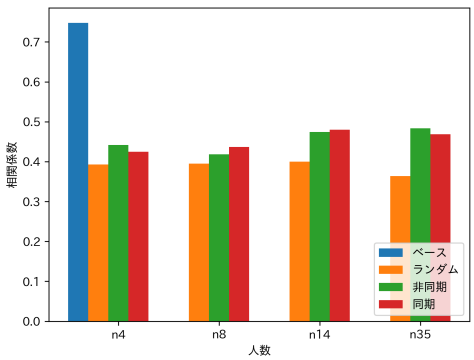
(e) GOP, 適応モデル



(f) DTW, 適応モデル



(g) GOP, DAE+マルチ



(h) DTW, DAE+マルチ

図 4.4: シャドーイング雑音を重畳した音声の相関の変化

第5章

結論

5.1 まとめ

本論文では、先行研究で提案されていた音素事後確率をもとにしたシャドーイング音声自動評価手法の高精度化および実用化に向けた検討を行った。前半の実験では，GOP の計算方法の変更や DTW の事後確率化における最適なクラス数の調査を行った。さらに，GOP・DTW 以外の特徴量を導入し，回帰モデルを用いることで評価者間の相関を超える精度を達成した。

後半の実験では，手動評価付き音声コーパス中に教室で収録された音声があることに注目し，音声に含まれる雑音を抑圧するための検討を行った。本実験では，1）マルチコンディション学習，2）モデル適応，3）DAE による特徴量強調を検討した。その結果，ある一定の精度の改善が見られた。さらに，実際の語学教育でに起こりうる雑音として，発話内容や発話タイミングなどの諸条件をコントロールした上で，教室内雑音の収録実験を行った。また，その雑音を手動評価付き音声コーパスに重畳することで，その雑音の影響及び，本研究で使用した耐雑音手法の効果を検証した。結果として，完全に雑音の影響をなくすことはできなかったが，ある一定の改善効果が見られた。雑音環境下で完全に実用化できるほどの精度は達成できなかったが，実用化に向けて一歩近づくことができたと考えている。

5.2 今後の課題

前半の高精度化の部分では，多くの特徴量を用いることで高精度化を実現したが，その結果として自動評価にかかる計算時間が増加してしまっている。実用化にあたって精度が重要であることは事実であるがリアルタイム性を追求する場合もっと計算時間を減らす工夫が必要となる。今後，精度と速度の両方の側面で現実的な自動評価に向けた検討を行う必要がある。

また，後半の実験ではいくつかの条件を設定して雑音の収録を行ったが，雑音対策についてはそれらの条件を考慮できていない。例えば，同一の発話内容である場合は音源分離の技術などを応用して特定の話者の音声だけを取り出すということが考えられる。今後はよりシャドーイング雑音に特価した対策を検討していきたい。

謝辞

本研究ならびに本論文の執筆にあたり，指導教員である峯松信明教授には多大なるご指導ご鞭撻を賜りましたこと，深く感謝いたします。研究生活では，外国語教育支援に関わる多くのプロジェクトに関わらせていただき，技術的な側面だけでなく教育的な側面での知識を得ることができました。また，不自由のない研究生活のために様々なサポートをしてくださった高橋登技官，秘書の池上恵さんに深く感謝いたします。

峯松・齋藤研究室の皆様には，修士 2 年間の研究生活において大変お世話になりました。外部から修士入学した私に対して内部の学生と別け隔てなく接していただきました。

齋藤大輔講師には，ミーティングなどで適切なご指摘をいただき，本研究を進めることができました。研究の進捗や計算機環境についても日頃からお気遣いいただき，感謝の念に堪えません。また，研究以外の部分でも雑談などをさせていただきました。

先輩方には，研究に必要なツールキットやプログラムの説明を丁寧にしていただき，円滑に研究を進めることができました。後輩の皆様には，様々な方向性の研究についてミーティングで報告していただき，モチベーションをいただきました。快く充実した大学院生活を送ることができたのは皆様のお陰です。本当にありがとうございました。

最後に，学生生活を支えてくださった家族と乃木坂 46 に心より感謝の意を表します。

2019 年 1 月 31 日

椛島　優

参考文献

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

[2] Scikit-learn. http://scikit-learn.org/stable/index.html.

[3] Lee Ann and Glass James. A comparison-based approach to mispronunciation detection. In *SLT*, pp. 382–387, 2012.

[4] B.S.Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *The Journal of the Acoustical Society of America*, Vol. 55, No. 6, pp. 1304–1312, 1974.

[5] Tejedor-Garcia Cristian, Escudero-Mancebo David, Gonzalez-Ferreras Cesar, Camara-Arenas Enrique, and Cardenoso-Payo Valentn. Evaluating the efficiency of synthetic voice for providing corrective feedback in a pronunciation training tool based on minimal pairs. In *7th ISCA Workshop on Speech and Language Technology in Education*, pp. 25–29. ISCA, 2017.

[6] Steven F.Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustic, Speech, and Signal Procesing*, Vol. 27, No. 2, pp. 113–120, 1979.

[7] Doug Paul John Garofalo, Graff David, and Pallet David. CSR-I (WSJ0) Complete. https://catalog.ldc.upenn.edu/LDC93S6A, 2007.

[8] Yo Hamada. The effectiveness of pre-and post-shadowing in improving listening comprehension skills. *The Language Teacher*, Vol. 38, No. 1, pp. 3–10, 2014.

[9] Yo Hamada. Shadowing: Who benefits and how? uncovering a booming efl teaching technique for listening comprehension. *Language Teaching Research*, Vol. 20, No. 1, pp. 35–52, 2016.

[10] Yo Hamada. Shadowing: What is it? how to use it. where will it go? *RELC Journal*, Vol. 0, No. 0, pp. 1–8, 2018.

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

[12] Franco Horacio, Neumeyer Lenonardo, Ramos Maria, and Bratt Harry. Automatic detection of phone-level mispronunciation for language learning. In *Eurospeech*, pp. 851–854, 1999.

[13] 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*, Vol. 11, No. 2, pp. 43–65, 2013.

[14] Masaaki Ishii, Takayuki Kunieda, Jun Murata, Ken Takehara, Shinsuke Harada, Yuto Karei, and Fuuko Nakano. タブレット型 CALL システムの開発・運用. 情報教育シンポジウム, pp. 163–167, 2017.

[15] 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*, pp. 616–620, 2014.

[16] Suguru Kabashima, Yusuke Inoue, Saito Daisuke, and Nobuaki Minematsu. DNN-BASED SCORING OF LANGUAGE LEARNERS' PROFICIENCY USING LEARNERS' SHADOWINGS AND NATIVE LISTENERS ' RESPONSIVE SHADOWINGS. In *SLT*, pp. 971–978, 2018.

[17] Neumeyer Leonardo, Franco Horacio, Digalakis Vassilios, and Weintraub Mitchel. Automatic scoring of pronunciation quality. *Speech Communication*, Vol. 30, pp. 83–93, 2000.

[18] 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*, pp. 1–4. IEEE, 2008.

[19] 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*, Vol. 2015, No. 1, p. 62, 2015.

[20] 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. 2011. IEEE Catalog No.: CFP11SRW-USB.

[21] Ramya Rasipuram, Milos Cernak, Alexandre Nanchen, and Mathew Magimai.-Doss. Automatic accentedness evaluation of non-native speech using phonetic and sub-phonetic posterior probabilities. In *Interspeech*, 2015.

[22] Shakti P Rath, Daniel Povey, Karel Veselỳ, and Jan Cernockỳ. Improved feature processing for deep neural networks. In *INTERSPEECH*, pp. 109–113, 2013.

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

[24] 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*, pp. 3142–3146, 2016.

[25] Witt Sike and Young Stece. Language learning based on non-native speech recognition. In *Eurospeech*, pp. 633–636, 1997.

[26] Witt Sike and Young Stece. Phone-level pronunciation scoring and assessment for interactive language learning. In *speech communication*, pp. 95–108, 2000.

[27] Yoon Su-Youn, Hasegawa-Johnson Mark, and Sproat Richard. Landmark based automated pronunciation error detection. In *InterSpeech*, 2010.

[28] 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.

[29] Karel Vesely, Arnab Ghoshal, Luks Burget, and Daniel Povey. Sequence-discriminative training of deep neural networks. In *INTERSPEECH*, pp. 2345–2349, 2013.

[30] 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, pp. 1096–1103, 2008.

[31] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.*, Vol. 11, pp. 3371–3408, 2010.

[32] Hu Wenping, Qian Yao, and K.Soong Frank. An improved DNN-based approach to mispronunciation detection and diagnosis of l2 learners' speech. In *SlaTE*, pp. 71–76, 2015.

[33] Silke M. Witt. Automatic error detection in pronunciation training: Where we are and where we need to go. pp. 1–8, 2013.

[34] Dong Yu, Li Deng, and George E. Dahl. Roles of pre-training and fine-tuning in context-dependent DBN-HMMs for real-world speech recognition. December 2010.

[35] Junwei Yue. DNN-based automatic assessment of shadowing speech. Master's thesis, The University of Tokyo, 2017.

[36] 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*, pp. 1422–1426, 2017.

[37] 石塚健太郎, 藤本雅清, 中谷智広. 音声区間検出技術の最近の研究動向. 日本音響学会誌, Vol. 65, No. 10, pp. 537–543, 2009.

[38] 篠田浩一. 音声認識. 講談社, 2017.

[39] 中川聖一, 小林聡, 峯松信明ほか. 音声言語処理と自然言語処理（増補）. コロナ社, 2018.

[40] 田中公啓, 椛島優, 齋藤大輔, 峯松信明. スパース性に着眼した posteriorgram のコンパクト化と DTW 発話比較における効果. 情報処理学会研究報告, Vol. 2018-SLP-125, No. 15, pp. 1–4, 2018.

発表文献

国際会議

[1] Suguru Kabashima, Yusuke Inoue, Daisuke Saito, Nobuaki Minematsu, “DNN-BASED SCORING OF LANGUAGE LEARNERS’ PROFICIENCY USING LEARNERS’ SHADOWINGS AND NATIVE LISTENERS’ RESPONSIVE SHADOWINGS”, The Proceedings of IEEE Spoken Language Technology (SLT) conference, 2018, pp.971-978,Athena, Greece

[2] Yusuke Inoue, Suguru Kabashima, Daisuke Saito, Nobuaki Minematsu, Kumi Kanamura, Yutaka Yamauchi, “A Study of Objective Measurement of Comprehensibility through Native Speakers’ Shadowing of Learners’ Utterances”, The Proceedings of INTERSPEECH 2018, 2018, pp.1651-1655, Hyderabad, India

[3] Nobuaki Minematsu, Yusuke Inoue, Suguru Kabashima, Daisuke Saito, Yutaka Yamauchi, Kumi Kanamura “Natives’ shadowability as objectively measured comprehensibility of non-native speech”, The Proceedings of 2nd International Symposium on Applied Phonetics (ISAPh2018), Fukushima, Japan

[4] Ryo Masumura, Suguru Kabashima, Takafumi Moriya, Satoshi Kobashikawa, Yoshikazu Yamaguchi, Yushi Aono, “Relevant Phonetic-aware Neural Acoustic Models using Native English and Japanese Speech for Japanese-English Automatic Speech Recognition”,Asia-Pacific Signal and Information Processing Association Annual Summit and Conference 2018（APSIPA2018）, 2018, Hawaii, USA

国内研究会・全国大会

[5] 椛島優, 塩澤文野, 齋藤大輔, 峯松信明, 山内豊, 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” 日本音響学会春季講演論文集, 2018, pp.1363-1366

[6] 椛島優, 張昊宇, 齋藤大輔, 峯松信明,　小橋川哲, 増村亮 “日本人英語に観測される発音多様性に関するコーパスに基づく定量的分析” 電子情報通信学会音声研究会, 2018,pp.69-74

[7] 椛島優, 齋藤大輔, 峯松信明, 山内豊, 伊藤佳世子, “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”, 情報処理学会音声言語情報処理研究会資料, 2018, p.1-6

[8] 椛島優, 齋藤大輔, 峯松信明, 山内豊, 伊藤佳世子 “シャドーイング音声自動評価における耐雑音性向上に関する検討” 日本音響学会秋季講演論文集, 2018, pp.829-832

[9] 椛島優, 齋藤大輔, 峯松信明, 山内豊, 伊藤佳世子, “教室内雑音が学習者シャドーイング音声の自動評価へ及ぼす影響に関する実験的検討”, 電子情報通信学会音声研究会, 2019, （発表予定）

[10] 井上雄介, 椛島優, 齋藤大輔, 峯松信明, 金村久美, 山内豊 “母語話者シャドーイングに基づく非母語話者音声の了解性計測に関する予備的検討” 日本音響学会春季講演論文集, 2018

[11] 井上雄介, 椛島優, 齋藤大輔, 峯松信明, 金村久美, 山内豊 “母語話者シャドーイングに基づく非母語話者音声の可解性自動計測”, 情報処理学会音声言語情報処理研究会資料, 2018

[12] 井上雄介, 椛島優, 齋藤大輔, 峯松信明 “母語話者シャドーイングに基づく可解性自動計測と回帰分析による高精度化” 日本音響学会秋季講演論文集, 2018

[13] 増村亮, 椛島優, 森谷崇史, 小橋川哲, 山口義和, 青野裕司 “ネイティブ日本語とネイティブ英語の音声データを活用した日本人英語向けニューラル音響モデルの検討” 日本音響学会秋季講演論文集, 2018

[14] 峯松信明, 井上雄介, 椛島優, 齋藤大輔, 金村久美, 山内豊 “母語話者シャドーイングとそれに基づく「聞き取り易さ」の客観的計測” 日本音声学会全国大会予稿集, 2018

[15] 井上雄介, 椛島優, 齋藤大輔, 峯松信明, “母語話者シャドーイングに基づく学習者音声の可解性自動計測と回帰分析による高精度化”, 情報処理学会音声言語情報処理研究会資料, 2018

[16] 田中公啓, 椛島優, 齋藤大輔, 峯松信明 “スパース性に着眼した Posteriorgram のコンパクト化と DTW 発話比較における効果”, 情報処理学会音声言語情報処理研究会資料, 2018

[17] 田中公啓, 椛島優, 齋藤大輔, 峯松信明 “スパース性に基づく Posteriorgram のコンパクト化と学習者・教師発話比較における効果” 日本音響学会春季講演論文集, 2019（発表予定）

学位論文

[18] 椛島優, “iPhone を活用した言語教育支援ソフトウェア開発（ネットワーク部分）” 久留米工業高等専門学校専攻科研究論文, 2017