

修士論文

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に関する研究



2020 年 1 月 30 日

指導教員 峯松 信明 教授

電気系工学専攻
37-176419 井上 雄介

本論文は東京大学大学院工学系研究科に修士号授与の要件として提出した修士論文である.

内容梗概

外国語発音学習の主目的は、母語話者に十分理解されやすい発音の獲得である。ところが自国で学ぶ学習者の多くは授業以外で母語話者と接する機会が少ないため、その獲得が困難である場合が多い。また一般の母語話者が面と向かって発音を厳しく指摘することはあまりなく、婉曲的あるいは上辺だけの指摘である場合が少なくない。学習者音声の可解性を評価する方法として、応用言語学の分野では母語話者にアンケートを課し、主観的に評価させる場合が多い。しかし、主観評価は評価基準の統一が難しく、評価者の恣意性が含まれ得るといった問題がある。そこで本研究では学習者音声の可解性をより客観的に計測する手法として、「母語話者シャドーイング」を提案する。シャドーイングは聴取音声を出来るだけ即座に復唱する行為で、一般的にはリスニング、及びスピーキング能力を向上させるために、学習者が母語話者音声に対して行うものである。一方、母語話者が学習者音声に対してシャドーイングを行う場合、その学習者音声の可解性が低ければ、シャドーイングが崩れると考えられる。すなわち、母語話者シャドーイングの円滑度を定量化することで、聴取音声の可解性が計測可能であると予想される。この方法論を検証するため、本研究ではベトナム人日本語学習者の読み上げ音声収録、及び収録した学習者音声を日本語母語話者にシャドーイングさせる実験を行った。実験の結果、1) 学習者音声の可解性、及びシャドーイングの円滑度それぞれに関する主観スコア間に強い相関が見られ、2) 学習者音声の可解性に関する主観スコアは、学習者音声の GOP よりも母語話者シャドーイング音声の GOP とより強い相関を示した。これらの結果から、「可解性の高い発音は即ちシャドーイングしやすい発音である」と考えることの妥当性が示された。以上により学習者音声の可解性評価に対して、本提案手法を用いることが有効であると主張する。

目次

第 1 章	序論	1
1.1	研究の背景	1
1.2	研究の目的	2
1.3	本論文の構成	2
第 2 章	先行研究	3
2.1	はじめに	3
2.2	学習者音声の「母語話者らしさ」に着目した研究	3
2.2.1	学習者シャドーイング音声の自動評価	3
2.2.2	音素事後確率に基づく GOP スコア (DNN-GOP [1])	5
2.2.3	事後確率ベクトルに基づく発話比較 (DNN-DTW [1])	6
2.3	学習者音声の「理解しやすさ」に着目した研究	8
2.3.1	学習者音声の了解性計測 [2, 3]	8
2.3.2	主観評価に基づく学習者音声の可解性計測 [4]	9
2.3.3	生体情報解析に基づく聴取時の負担感推定 [5]	10
第 3 章	提案手法	11
3.1	はじめに	11
3.2	母語話者シャドーイング	11
3.3	シャドーイングの円滑度を表す特徴量	11
3.3.1	調音に関する特徴量	12
3.3.2	シャドーイング遅れに関する特徴量	13
3.4	重回帰分析による予測精度の向上	13
3.5	学習者間相互シャドーイング	14
3.6	相互チュータリングアプリ開発	15
第 4 章	実験	16
4.1	はじめに	16
4.2	日本語音響モデルの構築	16
4.3	音声コーパス収録	17
4.4	母語話者シャドーイング実験の条件	19
4.4.1	被験者の構成	19
4.4.2	シャドーイングの主観評価, 及び客観評価	19
4.4.3	線形回帰モデルを用いた予測精度向上	19
4.5	母語話者シャドーイング実験の結果	20
4.5.1	被験者グループ毎の主観評価スコア	20

4.5.2	二つの主観評価スコア間の相関	21
4.5.3	DNN-GOP スコアと二つの主観評価スコア	21
4.5.4	各特微量と主観スコアの相関	22
4.5.5	線形回帰モデルによる予測結果	23
4.6	時系列アノテーションとしての母語話者シャドーイング	23
4.6.1	音素単位 GOP 時系列ラベリングのサンプル	23
4.6.2	CD 音声, および学習者音声との比較	25
4.6.3	CSJ の評価データを用いた検定	25
4.7	学習者間相互シャドーイングの妥当性検討	28
4.8	実用化に向けたモバイルアプリの開発	29
4.8.1	モバイルアプリの試験運用	30
第 5 章	実応用へ向けた関連研究	31
5.1	はじめに	31
5.2	生体情報解析に基づく学習者音声の可解性計測 [6]	31
5.3	母語話者シャドーイングは了解性・可解性のどちらを反映しているか [7]	32
5.3.1	文章難度を考慮した読み上げ音声の収録	33
5.3.2	シャドーイング実験	34
5.4	音声アノテーションへの応用に向けた検討	35
5.4.1	ベトナム人日本語学習者音声の大規模収録 [8]	35
5.4.2	読み上げ音声とシャドー音声の比較 [9]	37
第 6 章	結論	40
6.1	まとめ	40
6.2	今後の課題	40
	謝辞	41
	参考文献	42
	発表文献	46

目次

2.1	事後確率計算のための DNN モデル	5
2.2	DNN-GOP の計算フロー [10]	6
2.3	$m \times n$ 時系列間の DTW	7
2.4	DTW の局所パスの制約と重み	7
2.5	各学習者グループにおける単語単位の了解性 [3].	9
3.1	一般的なシャドーイング	12
3.2	母語話者シャドーイング	12
3.3	シャドーイング遅れ時間	13
3.4	学習者間相互シャドーイング	14
4.1	カラオケスタイルの読み上げ音声収録 web	17
4.2	母語話者シャドーイング実験の全体像	18
4.3	シャドワーの主観評価間相関係数のヒストグラム	20
4.4	シャドーイング音声の GOP と主観スコアの相関	21
4.5	学習者音声の GOP と主観スコアの相関	21
4.6	GOP 時系列ラベリングの計算方法	24
4.7	GOP 時系列ラベリングサンプル 1	24
4.8	GOP 時系列ラベリングサンプル 2	24
4.9	GOP 時系列ラベリングの比較 1	26
4.10	GOP 時系列ラベリングの比較 2	26
4.11	各音素毎の GOP 平均	27
4.12	「シャドーされたいか」の回答結果	28
4.13	「シャドーしたいか」の回答結果	28
4.14	開発したアプリの概念図	29
4.15	レッスン画面	30
5.1	表情センサー [6]	32
5.2	表情解析実験 [6]	32
5.3	表情アクションユニット [6]	33
5.4	nGOP の結果	36
5.5	sGOP の結果	36
5.6	シャドーイング遅れ時間の結果	36
5.7	読み上げ文の例（上段は初級者用，下段は中級者用）	37
5.8	シャドーイング音声と読み上げ音声の DTW[9]	38

表目次

2.1	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）	4
4.1	DNN 音響モデルのネットワーク構成	17
4.2	各特徴量の説明	18
4.3	ベトナム人日本語音声 (VJ) に対する主観スコアの平均値	20
4.4	日本人日本語音声 (NJ) に対する主観スコアの平均値	20
4.5	各特徴量と主観スコアの相関係数	22
4.6	重回帰分析の結果	23
4.7	One-way ANOVA の結果	25
5.1	プロのナレータに読み上げさせた文章セット	33
5.2	文章セットの特徴比較	34
5.3	被験者毎の S_C と S_S 間の相関係数 [8]	38

第1章

序論

1.1 研究の背景

第二言語獲得の為には、スピーキング、リスニング、ライティング、リーディングの4技能全てを習得する必要があるが、特にスピーキングとリスニングにおいては、他者との音声コミュニケーションが求められる。リスニングに関してはCD等の音声教材を用いても訓練可能であるが、スピーキングに関しては他者とのコミュニケーションを妨げるような発音誤りを把握する必要があるため、故に母語話者と接する機会をより多く持たなければならない。しかし、実際には自国で学ぶ学習者の多くは授業以外で母語話者と接する機会が少ないため、これを技術的に支援する対話形式のCALL (Computer-Aided Language Learning) システムが研究されてきた [11, 12, 13].

このシステムは発音誤りや文法誤りを自動的に検出し、その誤りをどのように修正すべきかといったフィードバックを返す。この時、母語話者音声で訓練した音響モデルを用いて学習者音声を評価する機会が多い。つまり母語話者によるモデル音声と学習者音声との比較に基づき評価を行うことになるが、これはすなわち学習者発話の native-likeness を計測していることに相当する。しかし、母語話者と全く同一の発音を目指す指導方法に疑念を抱く語学教師も多く、さらに、外国語訛りの程度によってはコミュニケーションが妨げられないことが知られてる [4, 14, 15].

例えば、英語は国際的に広く使用される言語であるため、多種多様な外国語訛りが受け入れられている。インドやシンガポール、フィリピンなど多くの国で英語が公用語とされているが、彼らは独自の訛りで英語を話す上、それをアイデンティティと考えている場合もある。世界諸英語 [16, 17] という言葉は、英語の現状をよく特徴付けている。

しかし、多様な外国語訛りが受け入れられている英語であっても、コミュニケーションが妨げられるケースがあるのは事実である [3]. 多くの学習者は、母語話者のような発音ではなく、母語話者に十分理解されやすい発音を獲得したいと願っている。しかし、一般に自国で学ぶ学習者が母語話者と会話する機会は少ないため、独りよがりな発音となってしまうことも多い。

学習者発音に対する評価として、応用言語学の分野では intelligibility と comprehensibility という指標がよく用いられる [4]. 本研究では、[4] に倣ってそれぞれを以下のように定義する。intelligibility は与えられた発話に対して、単語などの言語単位でどれだけ正確に聞き取られるかを示す指標である。intelligibility の度合いは母語話者に発話を書き起こさせることにより客観的な測定が可能である。一方 comprehensibility は、与えられた発話内容の理解に対する認知負荷を示す指標であり、母語話者にアンケートを課すことで主観的に評価することが多い。以上の定義から、本研究では intelligibility を了解性、comprehensibility を可解性と訳す。また聴取時の負担感という文脈では、listening effort, あるいは cognitive load という言葉もよく用いられる

[5, 18, 19, 20]. 発話内容を正しく理解するためには、単語を正しく同定する必要があるため、可解性は了解性を包含する概念であると考えられる。例えばある発話のすべての単語を正しく同定できた（了解性が高い）としても、発話内容の理解に努力を要した場合には、その発話の可解性は高いとは見なされない。

[4, 14, 15] では、外国語訛りの程度によっては、了解性、及び可解性を下げないことが示されている。学校での発音指導、及びCALLシステムにおいては、了解性或いは可解性を著しく低下させるような発音誤りを優先的に指摘するべきである。

では学習者自身が、可解性を下げる発音誤りを自覚するにはどうすればよいか。一般の母語話者は面と向かって学習者の発音を厳しく指摘することは少なく、婉曲的あるいは上辺だけの指摘をする場合が多い。学習者はより率直な指摘を求めらるだろう。

1.2 研究の目的

本研究の目的は、母語話者が感じる学習者音声の可解性をより率直かつ客観的な方法で推定することである。そこで、本研究では母語話者に学習者の音声をシャドーイングさせることを提案する。シャドーイングとは、音声を聴きながらその終了を待たずに復唱を開始する行為で、一般的には学習者がリスニングやスピーキング能力向上を目的として行う。一方、本研究では学習者音声を母語話者にシャドーイングさせる。シャドーイングは即座の反応を求められるため、その学習者音声の可解性が低ければ、シャドーイングが崩れると考えられる。つまり、シャドーイングの円滑度（smoothness）を定量化することで、母語話者が感じた学習者音声の可解性を客観的に測定できると予想される。

著者が知る限り、母語話者に学習者のシャドーイングを課すことはL2（第二言語教育）研究の中では初の試みであり、さらに可解性を客観的に測定する試みも今まであまり検討されてこなかった。これを実現する為、本研究では学習者読み上げ音声の収録、及び母語話者シャドーイング実験を行う。そしてシャドーイングの円滑度を表す様々な特徴量と被験者が付した主観評価スコアとの比較、及び学習者に対するアンケート調査の結果から、本提案手法の妥当性を検討する。

1.3 本論文の構成

本論文は全6章から構成される。第1章（本章）では、本研究の着想に至った背景や研究の目的、及び本論文の構成を述べ、第2章では、学習者音声評価に関する先行研究を紹介する。第3章では、本研究で提案する手法に関して述べ、第4章では、提案手法を検証するための実験について述べる。第5章では、応用に向けた更なる検討に関する関連研究に関して述べ、第6章では、本研究の達成点と残された課題についてまとめる。

第2章

先行研究

2.1 はじめに

本章では学習者音声の発音評価に関する先行研究について述べる。まず2.2節では、学習者音声の「母語話者らしさ」に基づいて発音評価を行った例として、学習者が実施したシャドーイング音声の分析によって、教師が付したスコアを予測した研究に関して述べる。特に、シャドーイング音声の分析は本研究でも行うため、その手法に関しても詳しく解説する。次に2.3節では、学習者音声の「理解しやすさ」、すなわち了解性や可解性に基づく発音評価として、様々なアプローチで取り組んだ例を紹介する。これらの研究例の紹介を通して、本研究の立ち位置を示す。

2.2 学習者音声の「母語話者らしさ」に着目した研究

学習者発音を矯正する場合、教師は自身が内在的に持つモデル発音との差異に基づいて指導を行う。また、技術的にこれを実現しようとする場合には、母語話者発音モデルとの差異を自動検出することになる。続く節では、学習者の発音能力は母語話者音声のシャドーイングに反映されると仮定し、シャドーイング音声を音響的に分析した研究を紹介する。

2.2.1 学習者シャドーイング音声の自動評価

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り、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と報告されている [21, 22]。しかし、聴取と同時に発音も行うため、シャドーイングが上手くできているかどうかを本人が判断するのは難しく、また教師が生徒一人ひとりの評価を行うことも非常に手間がかかるため、導入が難しいという現状がある。シャドーイング音声の自動評価が可能になれば、学習効果の高いシャドーイングをより手軽に教育現場に導入できる。

そこで、Luoらは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 [23]。2000年代の研究であり、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。

Yue, Kabashima らはシャドーイング音声から英語教師数名が付したスコアの推定を試みている [10, 25]. 本研究で用いる母語話者シャドーイング音声の円滑度に関する特徴量計算はこの研究を参考にしたため、以下に実験の詳細を示す.

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た [10]. 対象者は3つの異なる大学から参加した合計 124 名の日本人英語学習者である. シャドーイングはテキストを提示しながら行う場合もあるが, [10] ではより音声情報に集中させるためテキストを提示せずにシャドーイングさせた. シャドーイング音声は全部で 55 文から構成され, 4 つの異なるトピックから選出した. また, シャドーイング自体が比較的難しい行為であるため, それぞれの文音声について 4 回ずつシャドーイングを実施し, 最も学習者の実力が反映されているであろう 4 回目の音声を分析対象とした.

シャドーイング音声コーパスに対して英語教師による手動評価が実施された. シャドーイング収録時の 55 文のうち, 英語教師の助言のもと 10 文が手動評価の対象として選択された. 最終的に 124 人 × 10 文 = 1240 文の音声が手動採点された. ただし, 評価の際には各文をフレーズ単位に分けた為, 実際の採点対象はは 2 倍 ~ 3 倍ほどの数となる.

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. 以下 3 つの観点に基づき採点を実施した.

- 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 点満点の点数が付与された.

[25] ではシャドーイング音声の分析で得られた特徴量から, 回帰モデルを用いて手動スコアの予測を行っている. 各特徴量の詳細な計算方法は [25] を参照してもらうこととし, ここでは学習者単位での結果に関してのみ表 2.1 に示す.

表 2.1: 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 [1]	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

以上の結果から, シャドーイング音声を音響的に分析することで教師が付したスコアを比較的高精度で自動評価することが可能であることが示された.

本研究では, 分析する対象が「母語話者」のシャドーイング音声となるが, シャドーイングの円滑度を計測する目的では同じ特徴量を用いることができる. そこで具体的な計算方法に関して述べる.

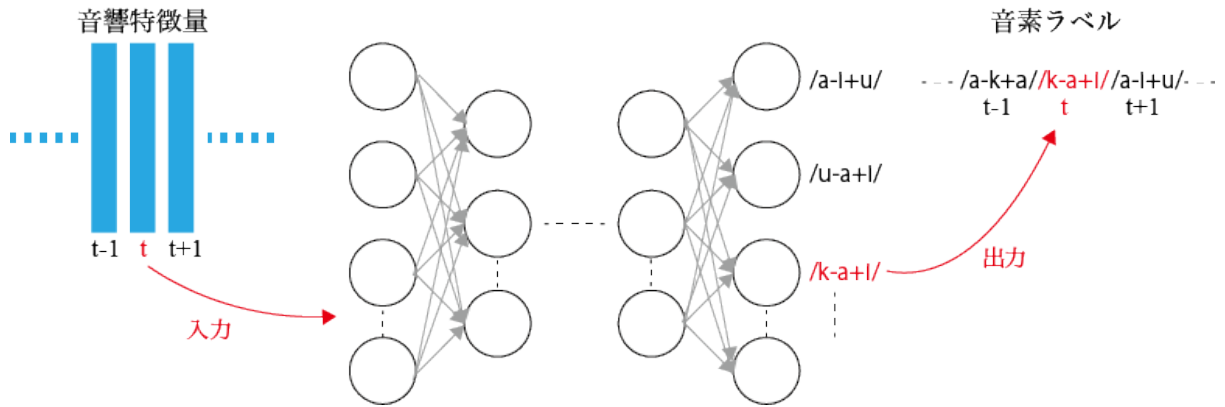


図 2.1: 事後確率計算のための DNN モデル

2.2.2 音素事後確率に基づく GOP スコア (DNN-GOP [1])

調音制御の正確さを示す尺度として GOP (Goodness Of Pronunciation) [26] という特徴量がよく用いられる。シャドーイングの調音評価においても同様のスコアを適用することができる [23, 27, 10]。GOP は音素事後確率 $P(c_i|o_t)$ として定義される。ただし、 o_t は各時刻 t で観測される音響スペクトル特徴量であり、 c_i は音素 i のクラスである。すなわち、ある時刻 t のスペクトル特徴量 o_t に対して、どの音素 c_i が意図されたのかを確率分布として表現していることになる。

さて、近年の音声認識では HMM (Hidden Markov Model) の出力確率計算に、GMM (Gaussian mixture model) ではなく、DNN (Deep Neural Network) を応用することでより高い認識精度を得ている。この DNN-HMM モデルの DNN の部分を取り出すことで、事後確率を直接求めることが可能になった。図 2.1 にその模式図を示す。DNN の学習には、ある入力に対する出力ラベルのペアが必要となるが、音声認識においては入力特徴量は MFCC (Mel Frequency Cepstral Coefficient) 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、音素ラベルとは音素そのものではなく、前後の音素までを含めるトライフォンが用いられる。図の例では、 $/k-a+i/$ は当該音素が $/a/$ で前の音素が $/k/$ 、後ろの音素は $/i/$ であることを表している。実際にはすべてのトライフォンを識別すると計算コストが大きいので、決定木によるクラスタリングによりトップダウン式に縮退され数千のクラスとなる。つまりこの DNN は、ある音響特徴量 o 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 c に相当するかを予測するモデルとなり、 $P(c|o)$ で表される GOP を直接計算できるようになる [28]。

続いて、DNN-GOP の具体的な計算方法について図 2.2 に示す。まず音声波形から入力特徴量となる MFCC を計算する。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と DNN の順伝播計算をそれぞれ適用する。強制アライメントの結果からは、各フレームに対応する音素の情報が得られ、DNN の順伝播計算の結果からは、各フレームにおける音素事後確率 (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 とも言う。) が得られる。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合計され、無音以外のフレーム長 D_x で除することで正規化される。

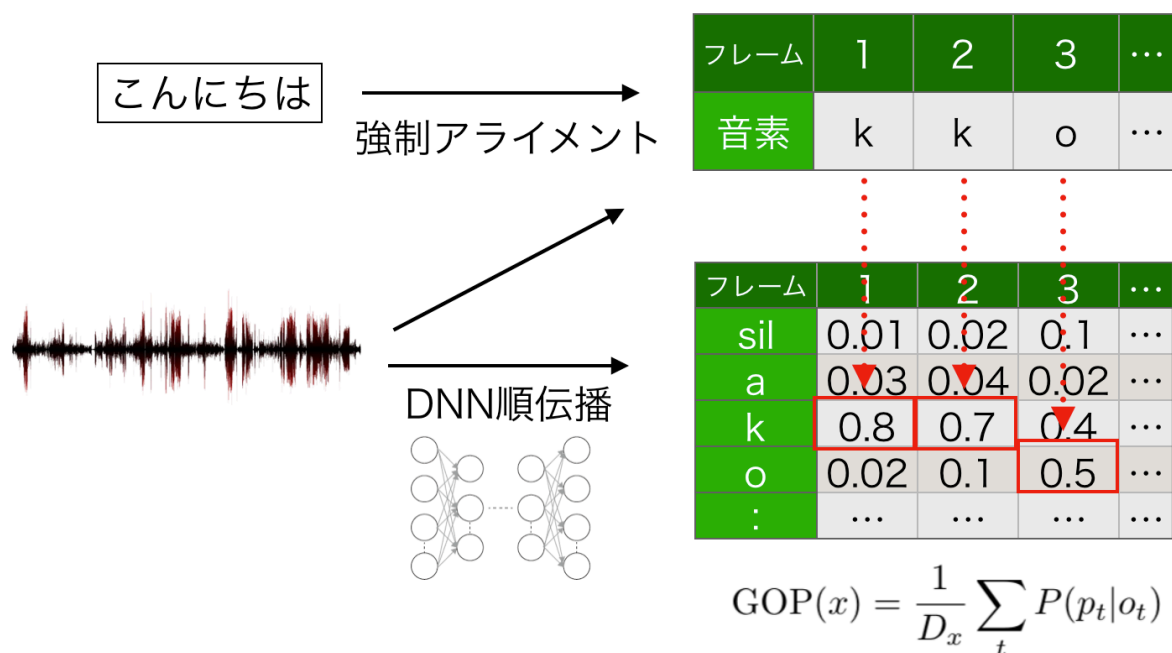


図 2.2: DNN-GOP の計算フロー [10]

2.2.3 事後確率ベクトルに基づく発話比較 (DNN-DTW [1])

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[1]。DTW (Dynamic Time Warping: 動的時間伸縮法)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。以下ではDTWに関して解説する。

図 2.3 に示すように m フレームの音響特徴量 $X = \{x_1, \dots, x_m\}$ と n フレームの音響特徴量 $Y = \{y_1, \dots, y_n\}$ 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 (2.1) のように表現できる。

$$\min_{\{(i_n, j_n)\}} \left[\sum_{k=1} d(i_k, j_k) \right] \quad (2.1)$$

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 2.4 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 (2.2) の漸化式が得られる。

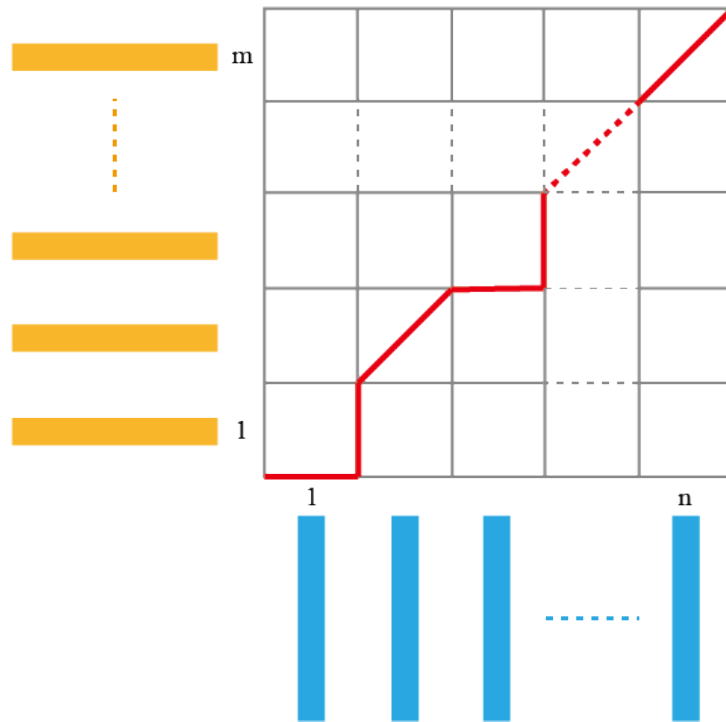


図 2.3: $m \times n$ 時系列間の DTW

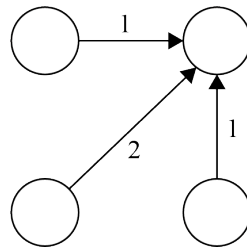


図 2.4: DTW の局所パスの制約と重み

$$D(i, j) = \min \begin{bmatrix} D(i, j - 1) + d(i, j) \\ D(i - 1, j - 1) + 2d(i, j) \\ D(i - 1, j) + d(i, j) \end{bmatrix} \quad (2.2)$$

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、MFCC などのスペクトルを表す特徴量には音素の情報以外にも、話者や体格の情報も多分に含まれている。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 [29, 30]。

[1] では、DNN-DTW をシャドーイング音声評価に適用している。シャドーイングを行う場合、母語話者が録音した、手本となるモデル音声が存在する。この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 を適用することで、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、[1] ではバタチャリヤ距離を用いている。離散確率ベクトル $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ 間のバタチャリヤ距離は式 (2.3) で計算できる。[1] では、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。本研究では、モデル音声で学習者の読み上げ音声に、学習者によるシャドーイング音声で母語話者によるシャドーイング音声となる。

$$D_{BD}(\mathbf{a}, \mathbf{b}) = -\log \left(\sum_i \sqrt{a_i b_i} \right) \quad (2.3)$$

さて、ここで DNN-GOP と DNN-DTW の関係性に関して言及しておく。GOP を計算するためには、シャドーイング音声と、聴取音声を読み上げる際に意図されたと思われる音素列、すなわち読み上げテキストが必要となる。この時モデル音声は計算に使用しないため、テキストとシャドーイング音声間の比較をしていると考えても良い。一方 DTW の計算にはシャドーイング音声とモデル音声の両方が必要である。言い方を変えれば、読み上げ時の音素列の情報（読み上げテキスト）は必要ない。DTW は音素事後確率ベクトル系列同士の比較であるが、GOP の場合は、音素列の情報があるという前提なので、DTW でモデル音声を用いる代わりに、当該音素のみが確率 1 をもつ、すなわち 1-hot のベクトル系列を用いて DTW を行っているとも考えることもできる。

2.3 学習者音声の「理解しやすさ」に着目した研究

2.2.1 節で述べた、学習者のシャドーイング音声評価では、母語話者音声コーパスを基に訓練した音響モデルを用いて音響分析をしている。これは母語話者発音に対する学習者発音の違いを定量化していることに該当し、つまり学習者発音の「母語話者らしさ」を評価しているといえる。しかし、1.1 節で述べたように、一定程度の外国語訛りであればコミュニケーションに支障をきたすことはない。学習目的にもよるが、実際は母語話者と全く同一の発音の獲得は非常にコストがかかるため、コミュニケーションに重きを置くのであれば、了解性や可解性をより重視すべきである。すなわち、母語話者の聴取能力上許容されうる発音逸脱を超えない程度の発音の獲得が望まれる。学習者音声の了解性や可解性を評価するためには、母語話者に聴取させるという工程が必ず必要となる。何故なら了解性は聴取者が実際に聞き取った単語等の文章の構成要素単位で計算されるものであり、可解性は聴取者が心理的に感じる尺度であるからである。そこで以下では学習者音声の了解性・可解性に着目した研究例をいくつか示す。

2.3.1 学習者音声の了解性計測 [2, 3]

[2, 3] では学習者音声の了解性を客観的に計測することを目的とした研究である。[2] は米国在住の移民 (L1 は様々)、[3] は日本人大学生の英語読み上げ音声を電話回線越しに米語母語話者に呈示した。母語話者には呈示した音声を書き起こすのではなく復唱するように指示した。そして復唱された音声は録音され、後日第三者によって書き起こされた。この書き起こし結果を元にして、発話毎に単語単位の同定率、すなわち了解性が計算された。

了解性は外国語訛りに対する聴取者の経験に依存すると考えられる。そこで [3] では、今まで日本人と接触経験のない米語母語話者を聴取者として採用した。また呈示音声は、ERJ(English Read by Japanese) コーパス [31] より日本人大学生 200 名に対し一人当たり平均 4 文音声を選択し、合計 800 音声とした。そして録音した英語読み上げ音声を米語母語話者 173 名に呈示した。この時、1 音声あたり平均 20 名の母語話者を割り当てた。また英語教師による主観評価の結果を元に、学習者を 7 つのグループにわけた。

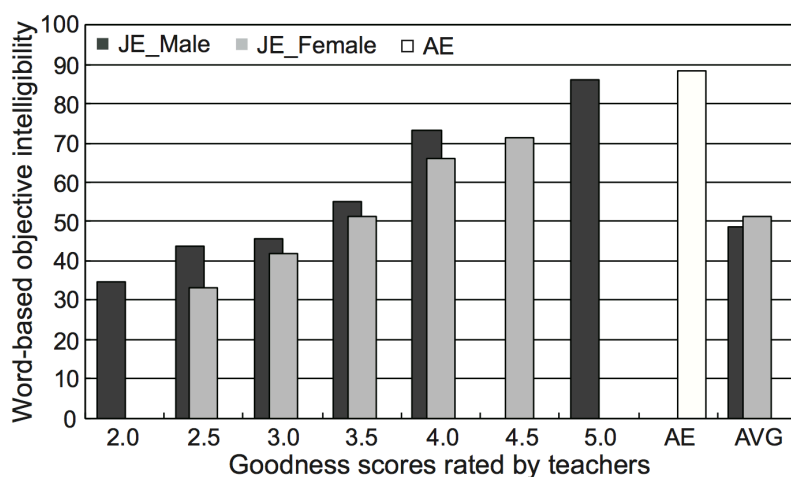


図 2.5: 各学習者グループにおける単語単位の了解性 [3]

図 2.5 は各グループに対する単語単位の了解性を示している。単語単位の了解性とは単語単位の平均同定率のことであるが、母語話者の音声に対してはおよそ 90%(連続する文音声は内容的に繋がりはなく、かつ話者も変わるという多少人工的な実験であったため、母語話者音声でも約 90%の正答率であった。)であったのに対し、日本人英語学習者全体の単語了解性は約 50%ととても低い値となった。

2.3.2 主観評価に基づく学習者音声の可解性計測 [4]

[4] では、アンケートを母語話者に課し、学習者音声の可解性を主観的なスコアで評価している。実験で用いた刺激音声は、カナダの大学に留学している 10 名の中国語母語話者と 2 名の英語母語話者による英語音声であり、あるストーリーをイラストで提示した後に、内容を英語で描写するというインストラクションの元に収録された自然発話音声である。この音声の開始 30 秒部分から、一度の聴取で書き起こし可能な程度の短いフレーズでそれぞれ 3 音声ずつ切り出し、合計 36 音声の刺激音声セットとした。

収録された学習者音声の評価実験は 2 回に分けて行われた。評価者は 18 名の英語母語話者である。1 回目はそれぞれの学習者音声を提示した後に、聴取内容の書き起こしをさせた。これは了解性の計測に用いられる。さらに書き起こし後には、9 段階のアンケートで可解性を評価させた。その際、1 には「とても簡単に理解できる」、9 には「聞き取りが不可能である」というラベルを付した。

続いて 4 日後に 2 回目の評価実験が行われた。2 回目の実験では、1 回目の実験と同様の刺激音声を提示し、今度は外国語アクセントに対して 9 段階の評価がなされた。ラベルは、1=「外国語アクセントは存在しない」、9=「かなり強い外国語アクセントが存在する」とした。

実験の結果、18 名中 17 名の評価者において、可解性と外国語アクセントのスコア間に有意的な正の相関が見られた。しかし相関係数は 0.41-0.82 と、評価者により大きく異なった。この結果から外国語アクセントは、聴取者によっては必ずしも可解性を大きく妨げるものではないということが言える。

2.3.3 生体情報解析に基づく聴取時の負担感推定 [5]

可解性は聴取者が心理的に感じる「聞き取りにくさ」の尺度として定義されることはすでに述べたとおりである。生理心理学の領域では、脳波や心拍数・瞳孔の開き等の人体の生体的特徴量を計測し、それらの特徴量から心理状態を予測するということが行われている。この技術を応用し、学習者音声聴取時の母語話者に生体センサーを取り付け、可解性等の心理的負担を予測する研究もある [5, 18, 19, 20]。

[5] では、脳波 (EEG:Electroencephalogram) の解析を通して、母語話者音声聴取時と学習者音声聴取時の listening effort の違いを分析している。実験方法の概略を述べる。

まず刺激音声の収録には、イギリス人英語母語話者と韓国人英語学習者がそれぞれ1名ずつ参加した。2人はそれぞれ720文の英語の文を読み上げた。読み上げた720文は、文末の単語の予測可能性 (predictability) に応じて、それぞれ3種類に分類される。一つ目は簡単に予測できるもの (例: Beef and milk come from cows)。二つ目は予測が難しいと考えられるもの (例: The man draws pictures of cows)。三つ目は予測が不可能と考えられる (意味的に誤りがあるもの) (例: Beef and milk come from bays) である。これらの音声を、イギリス人英語母語話者23名と韓国人英語学習者21名に提示する。

提示の際には、片耳には母語話者音声、もう片方の耳には学習者音声を流し、どちらか一方に集中するよう指示する。そして、聞こえた音声がつつ目の意味的に誤りがあるもの文音声であると判断した場合にはスイッチを押す。このようにすることで聞き取りに集中させ、生体反応を出やすくする。

実験の結果、韓国人被験者の EEG 信号の位相と聴取音声の同期度合いは、イギリス人被験者と比較して高かった。これは第二言語を聴取する上で母語話者よりも集中を要する為、より音声信号の周期と同期した反応が出た為であると考えられる。さらに母語話者が学習者音声に着目して聴取している時には、母語話者音声に着目して聴取している際と比較して、EEG と聴取音声の位相同期の度合いが高かった。

また文末の単語に対して、意味内容の処理中に発火すると考えられている N400 という電位を調べたところ、この場合も学習者の方がより大きな反応があった。

以上の実験のように、生体センサーを用いた聴取者の負担感から、聴取音声の可解性を予測することは可能である。([20] では、学習者訛りに起因する聞き取りの負担感ではなく、ノイズ音声に対する負担感を推定している。)しかし応用を考えた場合、常に母語話者にセンサーの装着を求めることは現実的ではない上、センサー機器を一定数用意するためのコストも必要である。

第3章

提案手法

3.1 はじめに

第2章では、第二言語発音学習の一つの目標として了解性や可解性に注目することの有用性について述べた。しかしそれらは母語話者の主観的な評価を要するため、自動的に計測することは一般的には難しい。また可解性の測定に生体計測機器を持ち込むことは現実的ではない。そこで本研究では可解性をより手軽かつ率直に計測可能な、母語話者シャドーイングという手法を提案する。続く節ではその詳細に関して述べる。

3.2 母語話者シャドーイング

[3]では聴取者に復唱という一定のタスクを課すことで客観的に了解性を測定した。しかし復唱方法の統制を取らなかった為、復唱までの遅れ時間や知識による補完などは聴取者に依存していたことが予想される。すなわち、得られた結果に、聴取者の聴取態度の違いに起因するノイズがかなり混入していたのではないかと想像される。

ここで復唱するまでの遅れ時間を最小化していくと復唱はシャドーイングとなり、知識による補完や推測がほぼ排除されると考えられる。すなわち単なる復唱の場合は、聴取後の（オフラインな）評価となってしまうが、シャドーイングであれば、音声を聴取している瞬間に感じるオンラインな聞き取り難さを反映できると考えられる。図 3.2 に母語話者シャドーイングの概念図を示す。

瞬間的に感じる聞き取り難さがシャドーイングの円滑度に反映されるということは、シャドーイングの円滑度は了解性というよりも可解性をより表していると考えられる。なお、この点に関しての実験的な検討は、5.3節で行う。

母語話者シャドーイングにおいては学習者の訛った発音を真似るのではなく、聞き取った単語を母語話者としての発音で再生するように指示する。このようにすることで、学習者音声の可解性の高い部分は母語話者らしい発音に変換され、可解性の低い部分はシャドーイングが崩れるような状況が生み出される。

3.3 シャドーイングの円滑度を表す特徴量

本研究では、シャドーイングの円滑度は、調音的な崩れと、遅れ時間で表現可能であると仮定した。シャドーイングの遅れ時間を大きく取れば当然シャドーの調音的な崩れは少なくなると予

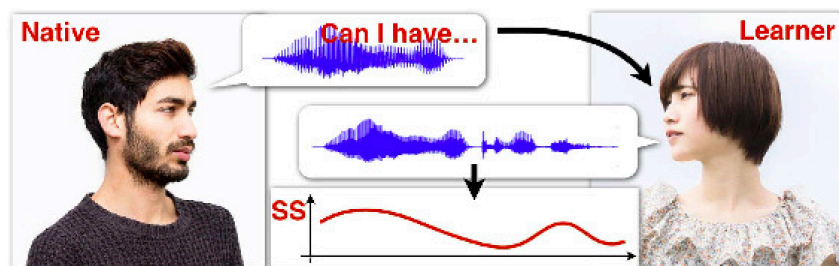


図 3.1: 一般的なシャドーイング

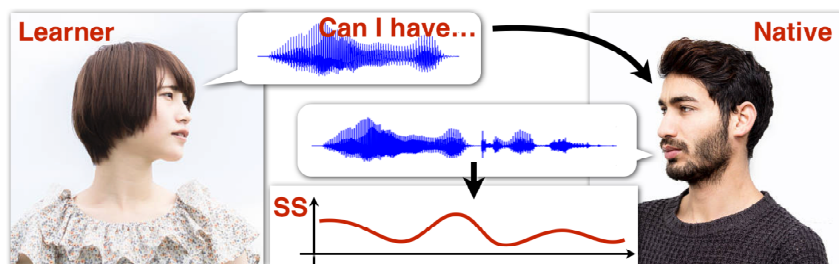


図 3.2: 母語話者シャドーイング

想される。よってこれら二つの特徴量は、遅れ時間を大きく取ってなるべく完璧に復唱する、或いは調音的な崩れを気にせずシャドーイングがなるべく遅れないように復唱する、という二つの異なるシャドーイング戦略をカバーできると考えられる。続く節では、それぞれどのような特徴量を分析に用いるかを述べる。

3.3.1 調音に関する特徴量

シャドーイング音声の調音の崩れに対しては、以下の特徴量を計算した。

- Phoneme-based GOP
 GOP の計算方法に関しては、2.2.2 節にてすでに述べた。一般に母音は子音に比べ時間長が長く、発話全体で事後確率を平均する場合、母音の影響が強調される。このバイアスを防ぐため、各音素区間ごとにフレーム平均 GOP を計算し、それを全出現音素数で平均する方法をとる。これを音素平均 GOP とする。
- Posteriorgram-based DTW
 2.2.3 節で述べた方法と同様の方法で計算した。
- WRR (Word Recognition Rate)
 WRR は自動音声認識の精度を評価する際に広く用いられる指標で、以下のように計算される。

$$WRR = \frac{N - D - S}{N} \quad (3.1)$$

ただし、N は正解テキストの単語数、S は置換単語数 (Substitution)、D は消失単語数 (Deletion) を表す。

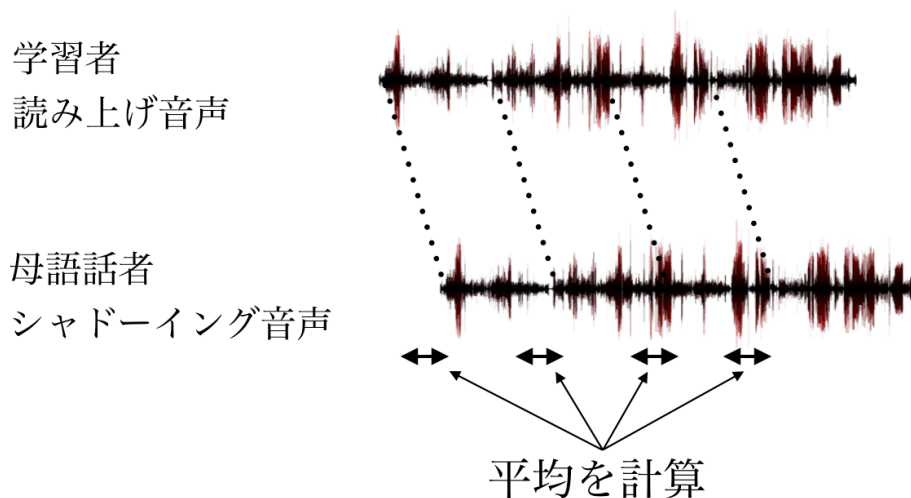


図 3.3: シャドーイング遅れ時間

3.3.2 シャドーイング遅れに関する特徴量

シャドーイング遅れ時間は、強制アライメントにより学習者音声とそれに対応する母語話者シャドーイング音声それぞれの音素境界時間を取得し、対応する音素境界対の比較により、その遅れを計算した。図 3.3 に示すように、二つの音声間の音素単位の遅れ時間の平均をシャドーイング音声の遅れ時間と定義する。

3.4 重回帰分析による予測精度の向上

各特徴量を回帰モデルに適用することで、より高精度にスコアを予測することが可能である。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。

$$Y = w_0 + w_1x_1 + \dots + w_nx_n \quad (3.2)$$

$$(3.3)$$

ここで、 $\mathbf{X} = (1, x_1, \dots, x_D)^\top$ とすると

$$Y = \mathbf{w}^\top \mathbf{X} \quad (3.4)$$

このときの重み \mathbf{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。 $\Phi(x) = \exp\{-\frac{(x-\mu)^2}{2s^2}\}$

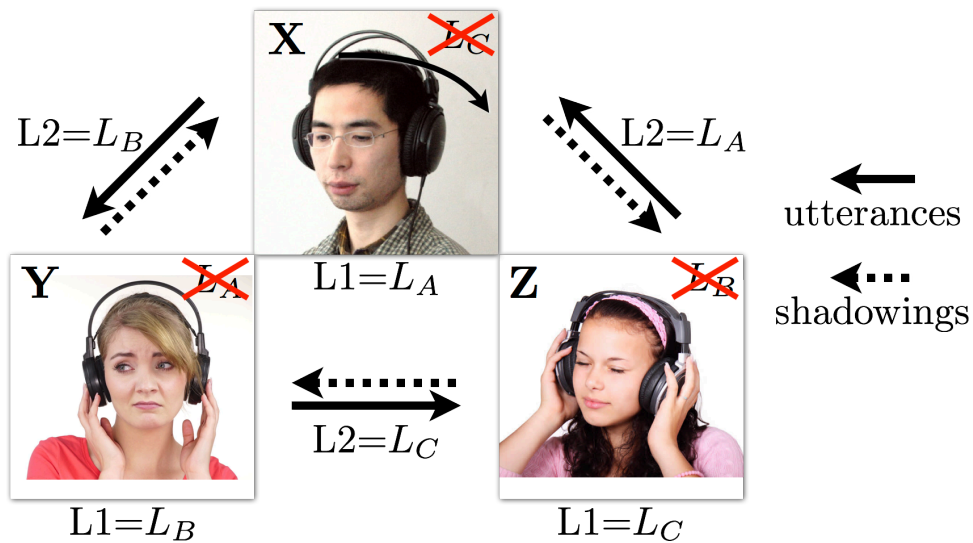


図 3.4: 学習者間相互シャドーイング

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充実しており、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きる。本研究では、いくつかの回帰モデルを構築し、その中から最も適したモデルを選定する。

3.5 学習者間相互シャドーイング

本研究では、学習者音声に対する可解性自動計測のフレームワークとして、母語話者シャドーイングという新たな方法論を提案している。ところがこの方法論においては、一つの重大な問題が存在する。それは学習者の数に対して十分な数の母語話者をどのように集めるかということである。

著者は、この問題は学習者間相互シャドーイングというフレームワークによって解決できると考える。すべての言語学習者は学習者であると同時に、少なくとも一つの言語の母語話者であり、その言語の学習者音声を母語話者シャドーすることが可能である。図 3.4 は学習者間相互シャドーイングの構図を示している。

言語 L_A を母語とし、言語 L_B を学習しているが、言語 L_C は学習したことがない学習者 X が言語 L_B で文章を読み上げる。次に言語 L_B を母語とし、言語 L_C を学習しているが、言語 L_A は学習したことがない学習者 Y が、学習者 X の音声をシャドーイングする。同様に言語 L_C を母語とし、言語 L_A を学習しているが、言語 L_B は学習したことがない学習者 Z が、学習者 Y の音声をシャドーイングする。最後に学習者 Z の音声を学習者 X がシャドーイングする。

この枠組みには更なる利点がある。どの学習者にとっても、自身の発音の可解性が低い人はいない。最高ではないかもしれないが、十分に高い可解性を有する発音であることは自明である。つまりその発音が母語話者にとってどれほど聞き取り難いかを学習者本人は自覚することが難しい。もし様々な言語に対して母語話者シャドーイングの円滑度をスコア化出来れば、学習者は自身の発音の可解性を、自身の母語を学ぶ学習者音声の聴取によって体感することができる。

本研究では、この枠組みの有効性を実験的に検討した訳ではないが、ベトナム・ハノイにて日

本語を学ぶ外国語大学の学生に対して、母語話者にシャドーしてもらうこと、また、ベトナム語を学ぶ外国人の音声をシャドーすること（即ち、学習者間相互シャドーイング）の妥当性をアンケート調査した。その結果も報告する。

3.6 相互チュータリングアプリ開発

外国語訛りが強くなる原因として考えられる要因は多数存在するが、その訛りが矯正されないのは圧倒的な母語話者との接触不足に原因があるだろう。しかし途上国では教師の数が圧倒的に足りないため、十分な音声教育を受けることは容易ではない。学習者間相互シャドーイングは母語話者数の不足を解決するための一つの方法であると言えよう。

ただし母語話者シャドーイングに参加する可能性のある母語話者は、やはり何かの言語の学習者であり、つまり「語学学習者」というコミュニティで閉じてしまっている。

昨今 Uber（タクシー配車アプリ）や、Uber eats（フードデリバリーアプリ）など、プロではないが最低限サービスを提供するには十分な能力を持ったユーザー（例えば運転ができる等）を、そのサービスの享受者と直接つなげる CtoC（Customer to Customer）のサービスが多く登場している。これを語学学習に当てはめれば、「会話」という最低限のサービスを提供可能なユーザーとして、すべての母語話者がなりうる。すなわち、教師・語学学習者以外の「一般の」母語話者であっても、中上級者相手であれば、母語を学習する学習者の会話相手になることは十分可能であると考えられる。

そこで、この方法論を検証するため、実際にスマートフォンのアプリケーション開発を行い、その運用実績を報告する。

第4章

実験

4.1 はじめに

本章では本研究の提案手法、母語話者シャドーイングの妥当性を検討するための実験に関して述べる。最初に実験の分析に用いる日本語音響モデルの構築に関して述べた後に、実験で用いる音声コーパスの収録方法と、シャドーイング実験に関して述べる。そしてシャドーイング実験の解析結果と、学習者に対して実施したアンケート結果を元に、本手法の妥当性を主張し、精度向上の為に更なる検討に関して述べる。その後時系列アノテーションとしての母語話者シャドーイングの応用に関して述べ、最後に、本手法を実社会に応用する為のモバイルアプリ開発に関して述べる。

4.2 日本語音響モデルの構築

ここでは、音響分析に用いる日本語音響モデルの構築方法に関して述べる。音響モデルの構築に必要なのは、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、日本語話し言葉コーパス (CSJ, Corpus of Spontaneous Japanese) を用いた。CSJとは、学会公演と模擬公演音声を含む、合計 600 時間ほどの音声コーパスである。

DNN 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 [32]。DNN 音響モデルの学習は、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 音響モデルの学習には、フレーム単位でラベル付けされたデータセットが必要となるが、これを人手で用意することは難しい。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した結果を用いる。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 (Cepstrum 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。



図 4.1: カラオケスタイルの読み上げ音声収録 web

次に、話者適応 GMM 音響モデルの学習を行う。MFCC は、同一音素であっても話者に依存してパラメータに違いがある。従来は、GMM の確率モデルで表現することで話者依存性を吸収していたが、学習データに存在しない未知のデータに対しては精度が低くなってしまふ。そこで、入力特徴量に対して事前に変換をかけ、話者の正規化を行う。

KALDI のデフォルトレシピでは、MFCC を数フレーム連結し、LDA (Linear Discriminant 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 Linear 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。詳細は、[33] を参照。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と fMLLR 特徴量を用いた DNN の学習を行う。DNN の学習時には、まず RBM (Restricted Boltzmann Machine) を用いて事前学習を行い、さらに誤差逆伝播法によってファインチューニングを行う。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。一方、音声認識の目標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と考えられる。KALDI の nnet レシピでは、sMBR (sequential Minimum Bayes Risk) 基準による学習 [34] 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 10% 程度になる。

表 4.1 に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 4.1: 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
日本語音響モデル	1400	6 層	2000

4.3 音声コーパス収録

本研究では対象言語を日本語、学習者をベトナム人とした。呈示音声としてベトナム人学習者の日本語音声を収録するとともに、比較のため日本語母語話者の音声も収録した。

この時、話速が遅すぎると可解性は常に高くなり、発音の影響がシャドーに反映されにくいと考えられる。そこで音声収録の際には話速の統制を行った。

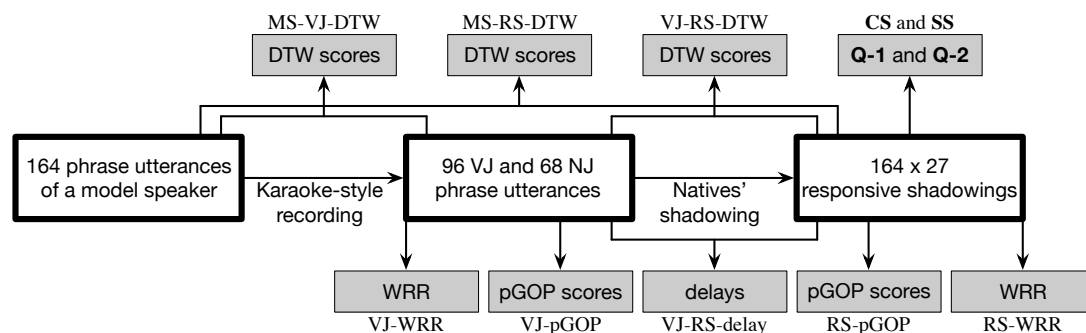


図 4.2: 母語話者シャドーイング実験の全体像

表 4.2: 各特徴量の説明

特徴量	特徴量の説明	特徴量	特徴量の説明
RS-pGOP	シャドー音声の GOP	VJ-pGOP	学習者音声の GOP
VJ-RS-delay	シャドー音声の遅れ時間	MS-RS-DTW	CD, シャドー音声間の DTW
VJ-RS-DTW	学習者, シャドー音声間の DTW	MS-VJ-DTW	CD, 学習者音声間の DTW
RS-WRR	シャドー音声の単語認識率	VJ-WRR	学習者音声の単語認識率
CS	可解性に関する主観スコア	SS	シャドーの出来に関する主観スコア

ベトナム人を選んだ理由は 2 つある。一つ目の理由は、昨今超高齢化社会を迎えた日本が、新しい労働力の担い手として、建設業等でベトナム人を多く受け入れるような政策を実施しており [35]、今後来日するベトナム人が増加すると考えられるからである。

二つ目の理由は、ベトナム語の特性に起因するかどうかは不明であるが、日本語教師の中で特にベトナム語訛りの日本語は聞き取りが難しい、という意見が多かったからである。

続いて、テキストは中級レベルの日本語の教科書を採用した [36]。この教科書には音声 CD が付属しており、日本人ナレーターによるモデル音声 that 収録されている。この教科書から 10 文章を選出した。1 文章あたり平均約 16 フレーズ (文より短い文節群)、合計 164 個の異なりフレーズである。この時、固有名詞を含む文章は除外した。また日本語の読みやすさを計算するツール、Jreadability [37] を用いて、これら 10 文章が同一レベルに属することを確認した。

10 文章中のそれぞれのフレーズを、6 名のベトナム人 (男性 3 名, 女性 3 名) と 6 名の母語話者 (男性 3 名, 女性 3 名) に読み上げさせた。6 名のベトナム人学習者は、3 名が学習歴 3 年未満 (平均 2.7 年) の中級レベル、残り 3 名が学習歴 3 年以上 (平均 5.8 年) の上級レベルである。さらに読み上げ時の話速統制のため、図 4.1 に示すカラオケスタイルの録音アプリケーションを用いた。

このアプリケーションでは、CD モデル音声に対する強制アライメントにより得られた時間情報に合わせて、各フレーズ中の文字色が変化する。読み上げの際に吃ったり、言い間違えたりした場合には何度でもやり直しを許した。

最終的にベトナム人学習者 1 人辺り約 100 音声、母語話者 1 人あたり 164 音声 that 得られた。ベトナム人学習者の場合習熟度によって収録所要時間に差があり、得られた音声の数に差がある。そのうちベトナム人日本語音声 (VJ) 96 音声と、日本人日本語音声 (NJ) 68 音声を提示音声として選択した。VJ と NJ には重複するフレーズは存在せず、164 個の異なりフレーズとなっている。

4.4 母語話者シャドーイング実験の条件

4.4.1 被験者の構成

シャドーイングを円滑に行えるかどうかは、ベトナム人日本語に対する経験に依存することは容易に予想される。

本研究では以下3グループの母語話者被験者を用意した。

NS-1 ベトナム人留学生との会話経験が全くない大学生

NS-2 研究室でベトナム人留学生と会話経験がある大学生

NS-3 ベトナム人留学生に日本語を日常的に教える教師

上から17名、5名、5名の日本人(20歳以上)が被験者として実験に参加した。実験のインストラクションとして、呈示された音声をシャドーイングする際に、決して訛りを真似しないように指示した。また呈示音声を単語単位で同定し、標準的な日本語でシャドーするように指示した。

4.4.2 シャドーイングの主観評価、及び客観評価

実験には合計27名の被験者が参加した。被験者全員がVJ 96音声、NJ 68音声、及び解析に用いないダミー音声36音声を合計200音声をシャドーイングした。ダミー音声は、非日本語母語話者による日本語読み上げ音声コーパス Japanese Read by Foreigners (JRF) [38] から、ベトナム人による読み上げ音声を選択した。音声呈示はヘッドホンを通してランダムな順序で行われ、音声収録には単一指向性のイヤーフックマイクを用いた。また1音声のシャドーイングが終わる度に、下記2つの主観評価を課した。

Q-1 呈示音声がどれくらい理解し易かったか

Q-2 シャドーイングがどれくらいスムーズであったか

Q-1は可解性に関する質問であり、Q-2はシャドーイングの円滑度に関する質問である。どちらの質問も7段階評価とした。常識的に考えると、これら2つの値は高い相関となることが予想される。

母語話者シャドーイング実験の全体図を図4.2に示す。ただしシャドーイング音声をRS(Responsive Shadowing)とし、各特徴量の略称も示した。RSに関係する特徴量は、提示した164フレーズに対して被験者間の平均値とした。また、各特徴量の説明を表4.2に示す。

4.4.3 線形回帰モデルを用いた予測精度向上

これまで述べた特徴量を説明変数とし、線形回帰モデルを構築することで主観スコアを予測する。これにより素の特徴量一つ一つを使う場合と比較して、高精度な予測が可能となる。

線形回帰モデルにはLassoを用いた。Lasso回帰は線形回帰モデルにL1正則化を行なったものであり、変数選択の性質を持っている。これによりスコア予測における各説明変数の寄与の大きさを知ることができる。

表 4.3: ベトナム人日本語音声 (VJ) に対する主観スコアの平均値

	NS-1	NS-2	NS-3
comprehensibility (CS)	4.13	4.20	4.24
smoothness (SS)	4.58	4.45	4.78

表 4.4: 日本人日本語音声 (NJ) に対する主観スコアの平均値

	NS-1	NS-2	NS-3
comprehensibility (CS)	6.53	6.75	6.40
smoothness (SS)	6.17	6.08	5.87

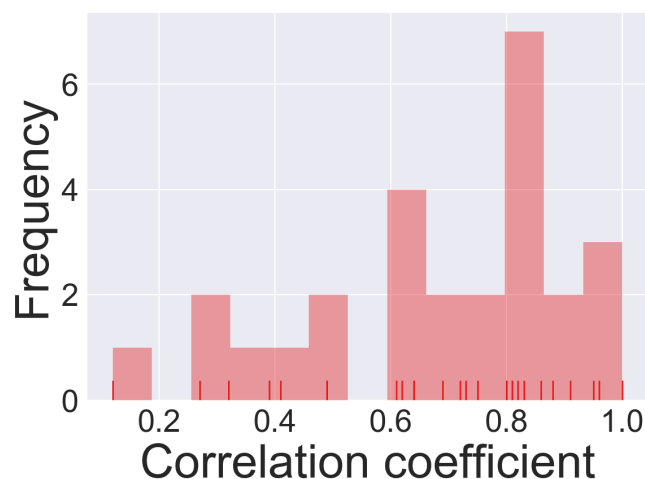


図 4.3: シャドワーの主観評価間相関係数のヒストグラム

4.5 母語話者シャドーイング実験の結果

4.5.1 被験者グループ毎の主観評価スコア

同じ学習者音声に対し、4.4.1 節に示した 3 グループ間で異なる可解性スコア (CS), 及び円滑度スコア (SS) が得られると予想される。表 4.3 に、VJ に対する 2 つのスコアの平均値を被験者グループ別に示す。一元配置分散分析の結果、SS に関して **NS-1-NS-3** 間、及び **NS-2-NS-3** 間にのみ有意差 ($p < 0.05$) が見られた。CS においては有意差の見られる組み合わせは存在しなかったが、**NS-1** から **NS-3** にかけて上昇する傾向があることがわかった。可解性評価において、学習者発音に対する慣れの影響が有意的な差を生じないということは、[39] でも報告されている。ただし、今回の教師グループ **NS-3** は、ベトナム語の学習経験はなかったが、仮に学習者の母語を学んだ経験がある教師であれば、結果は変わる可能性がある。今後も検討すべき事項である。表 4.4 は NJ に関する平均値である。

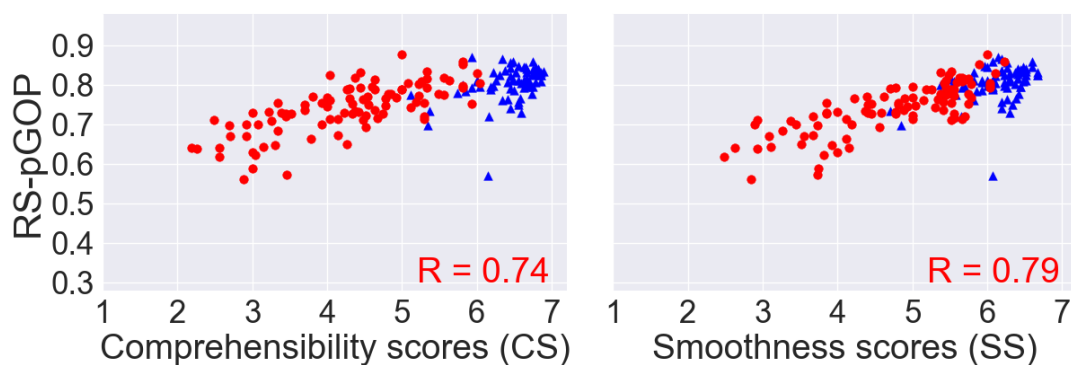


図 4.4: シャドーイング音声の GOP と主観スコアの相関

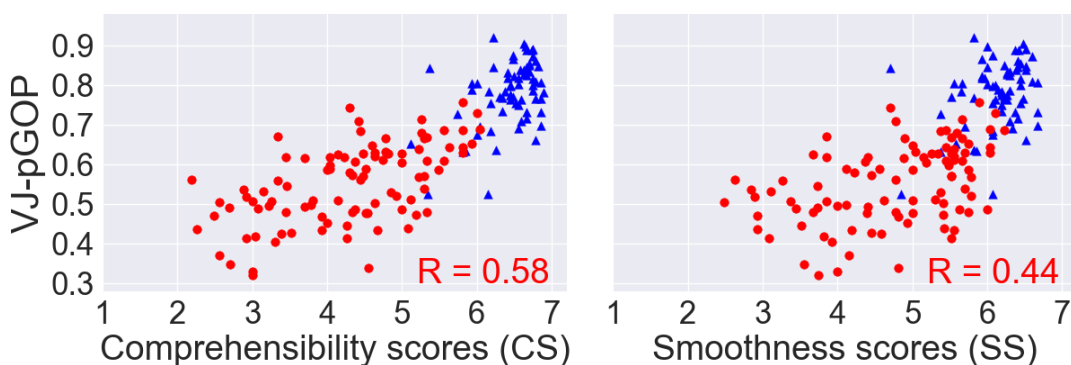


図 4.5: 学習者音声の GOP と主観スコアの相関

4.5.2 二つの主観評価スコア間の相関

二つの主観評価スコア CS, SS 間の相関を被験者毎に計算した。その結果、平均値は 0.68 とそれほど高くない値となった。4.3 は被験者毎の CS, SS 間の相関係数のヒストグラムを示している。27 名の内、7 名の被験者は CS, SS 間の相関が低かった（平均 0.36）。この 7 名の被験者は、学習者音声の可解性、及びシャドーの円滑度に対する評価基準が異なっていたと考えられる。評価基準の統制のため、可解性とシャドーの円滑度の各スコアの基準となる音声を示し、事前にコンセンサスを取るべきであったと考えられる。しかし十分にコンセンサスを取れていたとしても、多少の評価基準の不一致は避けられないと考えられる。また残りの 20 名の相関係数の平均値は 0.79 と高い値となった。

4.5.3 DNN-GOP スコアと二つの主観評価スコア

VJ 96 音声と NJ 68 音声のそれぞれに対して、27 名のシャドワーによるシャドーイング音声、CS、及び SS が被験者実験で得られた。そして全てのシャドーイング音声に対して RS-pGOP を計算した。さらに 164 個の呈示音声毎に、すべての被験者間の RS-pGOP、CS、SS の平均値を計算した。図 4.4 は CS と RS-pGOP の平均値間、及び SS と RS-pGOP の平均値間の相関を表している。赤点は VJ、青点は NJ の音声を表している。各図の red **R** は VJ のみに関して計算された相関係数である。図 4.4 の RS-pGOP と SS のどちらもシャドーイング音声に対して直接得られたスコアであり、高い相関が得られるのは自然であると考えられる。興味深いことに、CS はシャドーイング音声に対するスコアではなく学習者音声に対するスコアであるにも関わらず、RS-pGOP

表 4.5: 各特微量と主観スコアの相関係数

特微量	CS	SS	特微量	CS	SS
RS-pGOP	0.74	0.79	VJ-pGOP	0.58	0.44
VJ-RS-delay	-0.59	-0.69	MS-RS-DTW	-0.68	-0.71
VJ-RS-DTW	-0.60	-0.61	MS-VJ-DTW	-0.60	-0.51
RS-WRR	0.53	0.57	VJ-WRR	0.47	0.43

と高い相関を示した。

音素平均 GOP をシャドーイング音声ではなく、学習者音声 (VJ, 及び NJ) に対しても計算した。VJ-pGOP と、2つの主観評価スコアの相関を図 4.5 に示す。VJ-pGOP 及び CS は学習者音声から直接得られた値であるが、相関係数は図 4.4 の値よりも低い値となった。以上の結果から、GOP に基づいた学習者音声の可解性評価においては、学習者音声そのものよりも母語話者シャドーイング音声を分析した方がより適切であると考えられる。これには二つの理由があると考えられる。一つは学習者音声に対する可解性スコア CS は通常聴取者に依存するが、学習者音声から計算された VJ-pGOP は聴取者には全く依存しない点である。もう一つの理由は、母語話者シャドーイング音声は母語話者による発話なので、非母語話者音声に対して GOP を計算する場合と比較して、技術的により安定していると考えられる点である。

4.5.4 各特微量と主観スコアの相関

表 4.5 に、シャドーの円滑度を表す各特微量と、主観スコア CS 及び SS 間の相関係数を示す。ただし VJ96 フレーズについてのみ計算した。

VJ-RS-delay が負の相関なのは、可解性の低い音声ほどシャドーイングが遅れやすい、すなわちシャドーイング遅れ時間が大きくなるためと考えられる。また DTW スコアも負の相関となっているが、これは可解性の低い音声ほど二つの音声の音素事後確率ベクトル系列の類似度が下がり、累積距離が大きくなるためと考えられる。

特に注目すべきは、母語話者シャドーイングに関する特微量が、それと対応する VJ に関する特微量と比較してより高い相関を示している点である。特に RS-pGOP, VJ-RS-delay, MS-RS-DTW は CS と SS に対して高い相関を示しており、回帰モデル構築の際に有効な特微量であると考えられる。表 4.5 より、学習者音声の可解性が観測対象である場合、学習者音声そのものを解析するよりも母語話者シャドー音声を解析する方が有益であるということが言える。

また GOP, 及び DTW に関する特微量は、WRR に関する特微量よりも高い相関を示した。(RS-pGOP & MS-RS-DTW > RS-WRR, VJ-pGOP & MS-VJ-DTW > VJ-WRR)。ASR (Automatic Speech Recognition) モデルは通常、母語話者音声コーパスで学習を行い、母語話者音声を正しく認識するように最適化されている。つまり学習者の発音誤りに対する聞き手 (母語話者) の許容度を測定するために ASR 技術を用いることは適当であるとは言えない。ASR モデルを学習者音声コーパスで学習を行い学習者音声の認識率を高めたとしても、それは聞き手の許容度を表す指標にはならず、つまり可解性の測定には適当でない。母語話者シャドーイングが利用可能である場合には、それを用いる方がより有効である。

表 4.6: 重回帰分析の結果

モデル	CS	SS
Lasso	0.81	0.86
inter-rater	0.66	0.59

4.5.5 線形回帰モデルによる予測結果

CS, SS に対してそれぞれ Lasso 回帰モデルを構築した。なおデータの数は VJ96 フレーズであり, RS に関する特徴量は母語話者被験者間の平均値を用いてる。データを 3:1 の割合で訓練データおよびテストデータに分け, 訓練データの中で 3-fold のクロスバリデーションを行いパラメータを決定した。テストデータの取り方によって予測精度も変わるため, データ分割・パラメータ調整・テストデータ予測を 1 セットとし, 合計 50 セットの精度評価値の平均値を計算した。なお予測値の精度評価指標は, 正解データとの相関係数とした。

表 4.6 に Lasso モデルの結果と, 比較のため被験者間の相関係数を示す。被験者間の相関係数の計算方法は次のように行った。まず被験者 27 名のうち 1 名と 26 名のグループに分ける。そして 26 名の CS, SS の平均値を計算し, 残りの 1 名の CS, SS と相関を計算する。これを 27 名の被験者がそれぞれ 1 名の被験者となるよう繰り返し計算する。最後に, 計算された 27 名の相関係数の平均値が被験者間の相関係数である。Lasso モデルで予測する値 CS, SS が, 被験者 27 名の平均値であることから, 以上のような計算方法を採用した。

回帰モデルの予測結果は被験者間相関と比較して, かなり高い相関を示した。この結果から, 学習者音声の可解性自動計測に本モデルを用いることの有効性が示された。

4.6 時系列アノテーションとしての母語話者シャドーイング

これまで, GOP スコアは一つの音声に対してまず分析フレーム単位で計算し, 続いて音素単位で平均をとり, 最終的に全音素の平均を計算して該当音声に対して一つの GOP スコアを計算していた。分析フレーム単位の GOP スコア, 或いは音素単位の GOP スコアを用いれば, 該当音声に対して GOP スコアの時系列ラベルが得られる。学習者音声を評価する場合の GOP スコアは, 計算の精度を高めるために音素全体の平均を取っていた。しかし, 母語話者シャドーイングの枠組みでは, 一つの学習者音声に対して, 複数の母語話者がシャドーイングを行っており, 母語話者シャドー音声間で平均を計算することが可能である。

具体的な計算方法を図 4.6 に示す。GOP スコアを時系列ラベリングすることが出来れば, 学習者音声の可解性がどこで下がったのかを知ることができる可能性がある。また, 発音指導においてもどの部分を重点的に指導すべきか, といった指標として使える可能性がある。

4.6.1 音素単位 GOP 時系列ラベリングのサンプル

実際に, 先述した方法によって音素単位で GOP の時系列を計算した。結果を, 図 4.7, 及び図 4.8 に示す。得られたラベルを見ると, 単語頭の音素の GOP が低くなる傾向があるように思われる。仮説としては, シャドーイングにおいて文頭或いは単語頭の予測は困難であるため, 崩れやすい傾向があると考えられる。

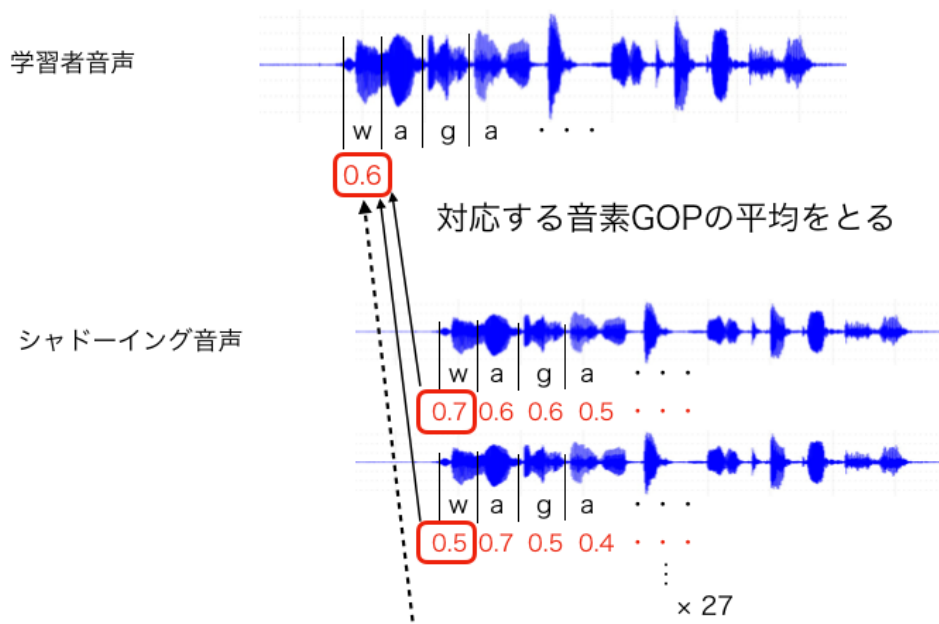


図 4.6: GOP 時系列ラベリングの計算方法

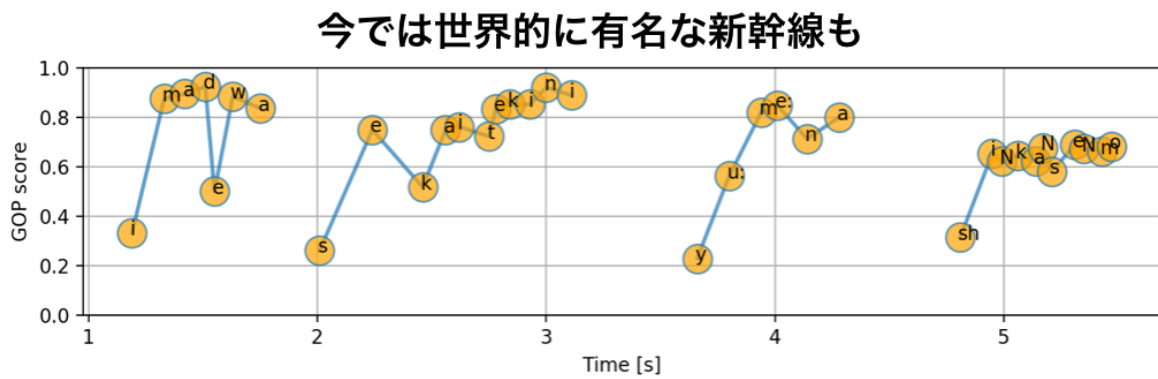


図 4.7: GOP 時系列ラベリングサンプル 1

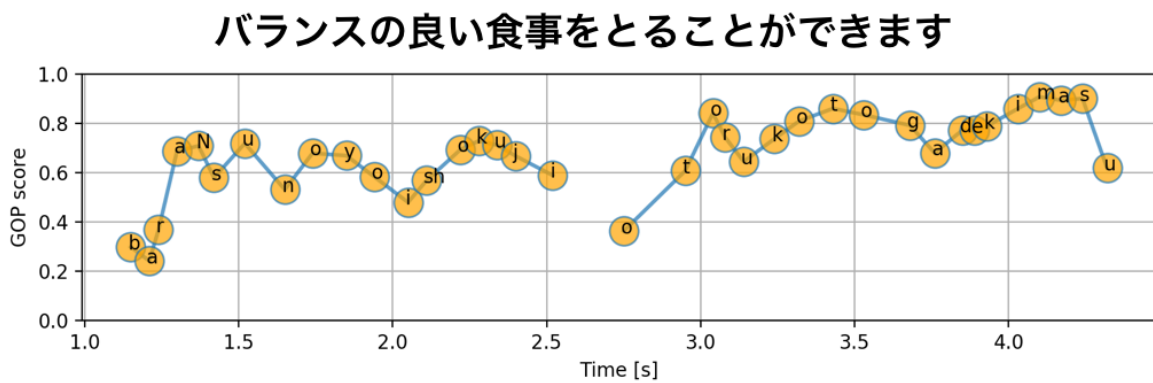


図 4.8: GOP 時系列ラベリングサンプル 2

表 4.7: One-way ANOVA の結果

音声データ	Begining	Internal	Ending	Singleton
CSJ 検証データ	0.81	0.86	0.85	0.70
教科書 CD 音声	0.67	0.82	0.78	0.61
学習者音声	0.59	0.68	0.68	0.54
シャドーイング音声	0.64	0.79	0.79	0.65

4.6.2 CD 音声, および学習者音声との比較

この仮説を検討するため, CD 音声, また学習者音声に対しても GOP 時系列ラベリングを付した. ここで CD 音声とは, 図 4.2 に示した教科書に付属している CD の, 母語話者による読み上げ音声である. ここで CD 音声の無音部分は振幅が 0 となるようにノイズ除去が施されており, 強制アライメントがうまく行えなかったため, S/N 比 40dB となるようホワイトノイズを加算した後に分析を行った. それぞれのサンプルに対する CD 音声, 学習者音声, シャドー音声の GOP ラベリングの結果を, 図 4.9, 及び図 4.10 に示す. ここで, 各音素の後ろに表記してある, B, I, E, S は, それぞれ Beginning (単語頭), Internal (単語頭, 単語尾以外), Ending (単語尾), Singleton (孤立音) を表している. また比較のため, 横軸は全て学習者音声の強制アライメント情報に合わせて表示している.

図 4.9, 及び図 4.10 から, 読み上げ音声であっても単語頭, すなわち Beginning の音素が低くなる傾向があるように思われる.

4.6.3 CSJ の評価データを用いた検定

単語頭の音素の GOP が低くなる傾向があるのか, kaldi で CSJ コーパスを使って音響モデルを学習する際の, 評価データを用いて検定を行った. ここで評価データには 5 時間ほどの音声が含まれている. 評価データすべてに付与した音素単位 GOP を, B, I, E, S 毎にグループ分けし, One-way ANOVA を用いて, それぞれの平均値に統計的有意差が存在するか検定を行った. 検定結果を, 表 4.7 に示す. データに対して有意差検定を実施したところ, B および S は I と E に対して全ての組み合わせで有意差があった.

この結果から, Singleton も単語頭であると考えれば, すべての音声データにおいて単語頭と Internal の値の間に有意差があった. つまりここから言えるのは, 読み上げ音声であっても, シャドーイング音声であっても, 単語頭の音素は低くなる傾向があるということである. ただし, CSJ の評価データと比較して, 他の音声は Beginning と Internal の平均値差が大きいことがわかる. 今回教科書から採択したフレーズに特異な現象である可能性もある.

また CSJ の評価データ中の各音素の GOP 平均を, 図 4.11 に示す. 図 4.11 から分かることは, 音素によって GOP にバラツキがあることである. 全体の平均値と, 各音素に対して One-way ANOVA を行うと, 平均値と有意差がある音素が複数存在した. これらの音素に対して平均値補正を行うことで, 時系列 GOP ラベリングの信頼性がより上がる可能性がある.

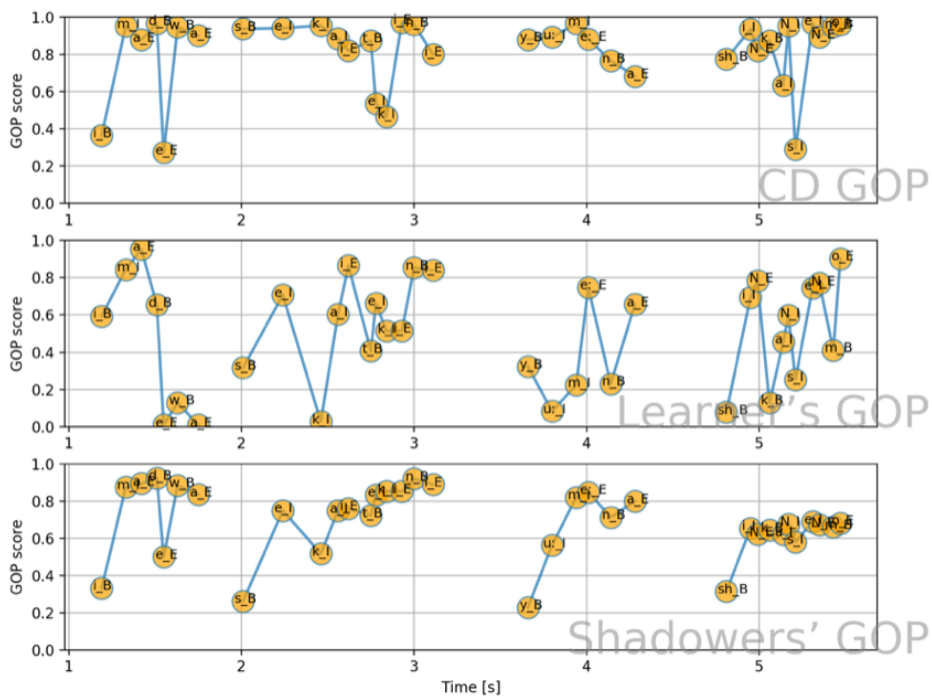


図 4.9: GOP 時系列ラベリングの比較 1

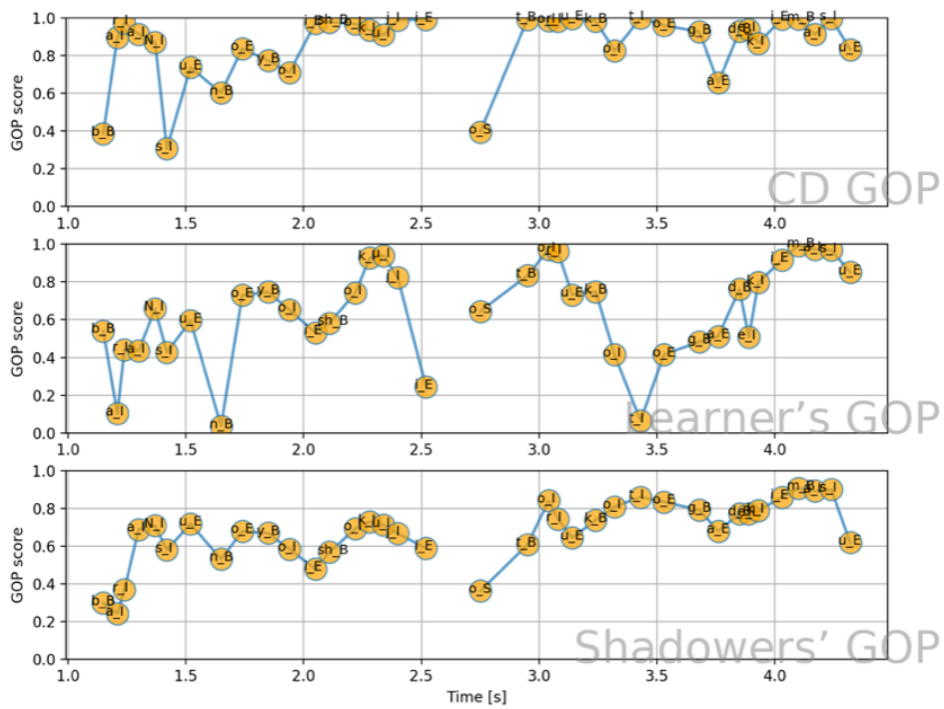


図 4.10: GOP 時系列ラベリングの比較 2

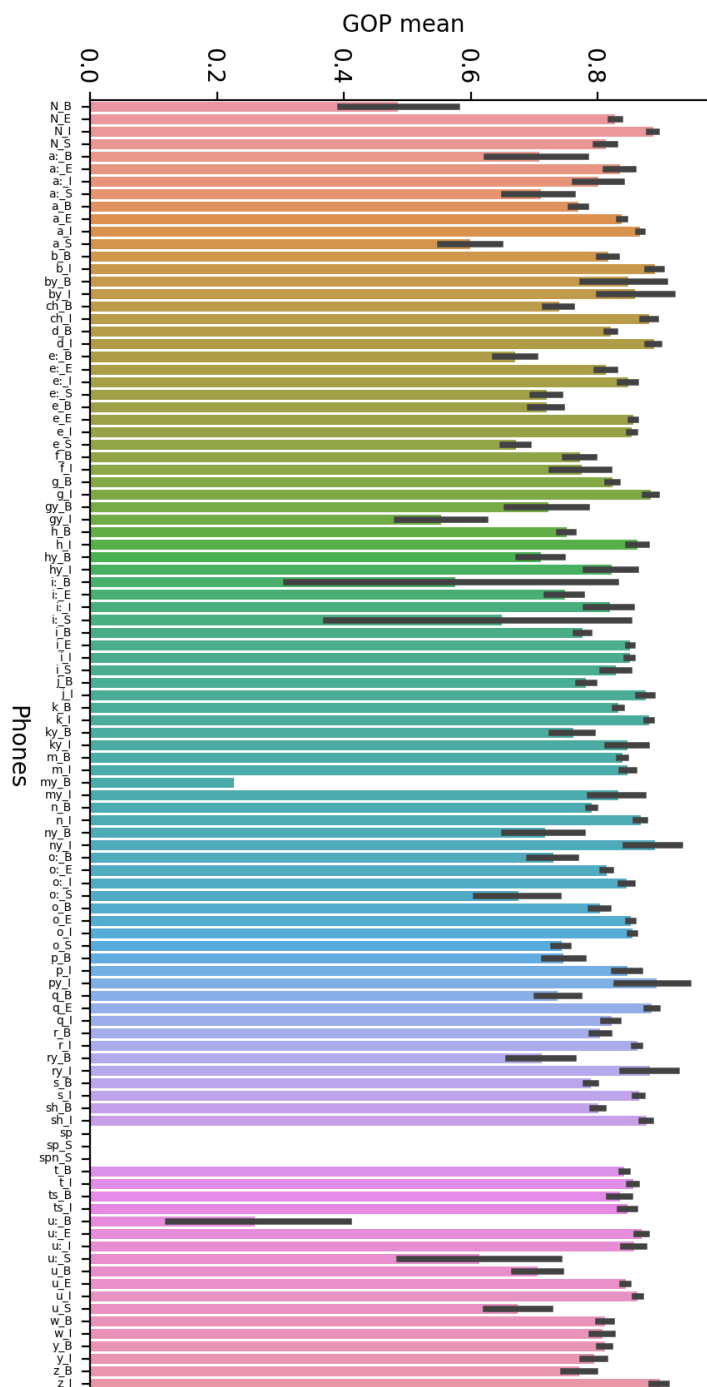


図 4.11: 各音素毎の GOP 平均

相互シャドーイングについて、母語話者として、どう思いますか？ / What do you think of inter-learner shadowing as native speaker?

48 responses

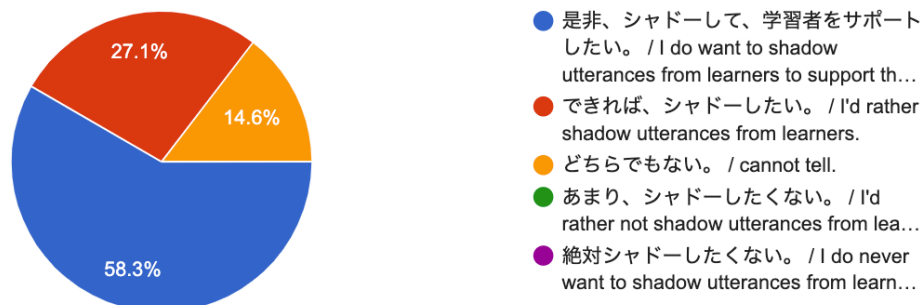


図 4.12: 「シャドーされたいか」の回答結果

相互シャドーイングについて、学習者として、どう思いますか？ / What do you think of inter-learner shadowing as learner?

48 responses

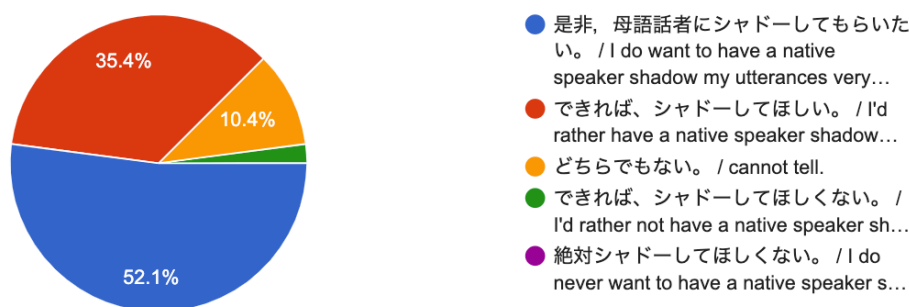


図 4.13: 「シャドーしたいか」の回答結果

4.7 学習者間相互シャドーイングの妥当性検討

学習者間相互シャドーイングの妥当性を検討するため、5.4.1で収録に参加した来日経験のないベトナム人学習者（大学にて日本語を学ぶベトナム人）52名に対して本手法を説明し、「日本人にシャドーされたいか」というアンケートを実施した。結果の分析を通して実応用の可能性を検討する。

図 4.12 にアンケート結果を示す。図 4.12 では、「是非、母語話者にシャドーしてもらいたい」、及び「できれば、シャドーしてもらいたい」と、全体の 87.5% を肯定的な意見が占めた。被験者は母語話者として、ベトナム語学習者音声のシャドーも体験させたため、本手法の有用性を実感出来たのだと考えられる。一方「シャドーされたくない」という意見であるが、母語話者シャドー音声に崩れる様子にショックを受けそうという理由であった。既に就職が決まり、日本語を使って母語話者と会話することが必須な学習者に対しては、本手法の実用価値は高いだろうが、

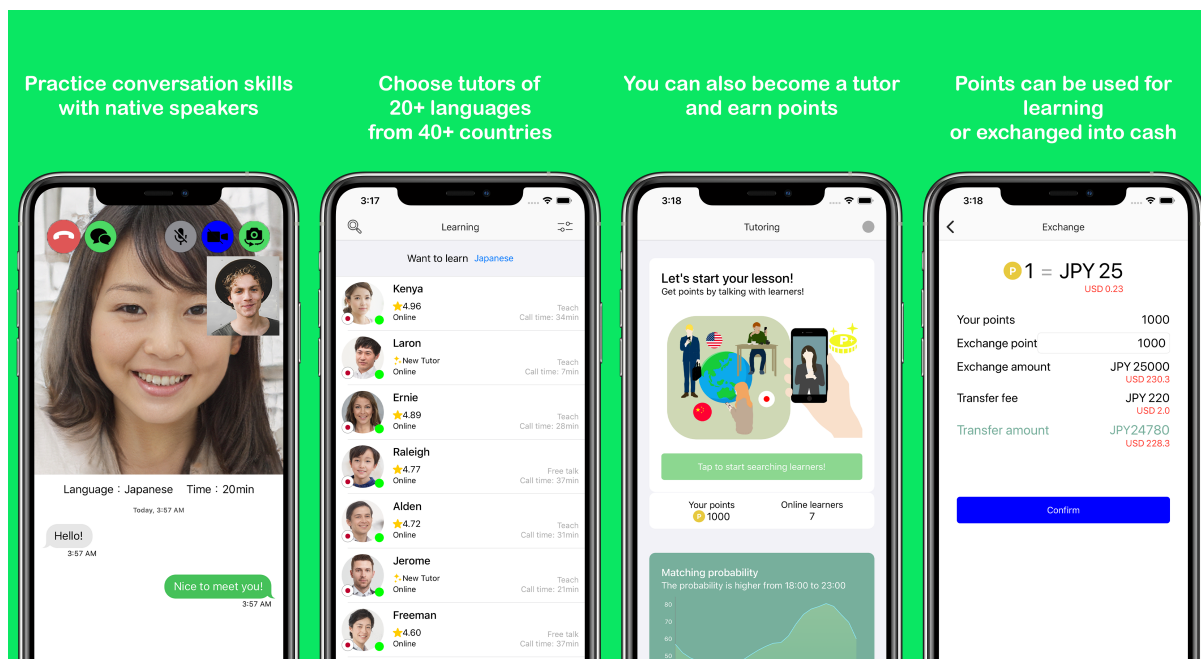


図 4.14: 開発したアプリの概念図

初級者に対して導入すべきかどうかは、教師の判断を待ちたい。

続いて、「母語話者としてベトナム語学習者音声のシャドーイングを通して、学習者のサポートをしたいか」というアンケートも実施した。結果を図 4.13 に示す。図 4.13 においても「サポートしたい」という肯定的な意見が 85.4%と、多く得られた。この結果から、学習者同士で相互にシャドーする枠組みであれば、より多くの学習者が参加する可能性がある。今後、アプリケーションの開発も検討したい。

4.8 実用化に向けたモバイルアプリの開発

3.5 節で述べたように、母語話者を必要な数集めることは非常にコストがかかるが、学習者同士が自発的にサポートし合うことが可能なプラットフォームを提供することが出来れば、安価にデータを集めることが可能であると考えられる。例えば学習者 **X** が自身の発音の可解性を評価されたい場合には、まず自分の母語の学習者である学習者 **Y** の音声をシャドーイングし、その後自身の音声を登録してその言語の母語話者にシャドーイングしてもらう。

このプラットフォームをそのまま実現した場合には、外国語学習者のみが参加することになるが、その場合問題になることとして、学習者人口の不均衡性が挙げられる。例えば、世界中で英語を第二言語として学ぶ学習者の数は英語母語話者の数よりも多いため、母語話者の中でも特に外国語学習者に限定してしまうと、英語をシャドーイングしてもらう母語話者を見つけるのが難しくなる。このような現象は、学習者数が多い他の言語にも起こりうる。問題を解決するためには、外国語学習者以外のユーザーも積極的に参加可能な構造を提供する必要がある。3.6 節で述べたようなアプリの開発が望まれる。

そこで、すべてのユーザーが自由にチューター、或いは生徒になれるモバイルアプリを開発した。アプリの概念を図 4.14 に示す。このモバイルアプリを学習者として使う場合には、アプリ内通貨を消費することで母語話者であるチューターと一定時間通話することができる。また反対に

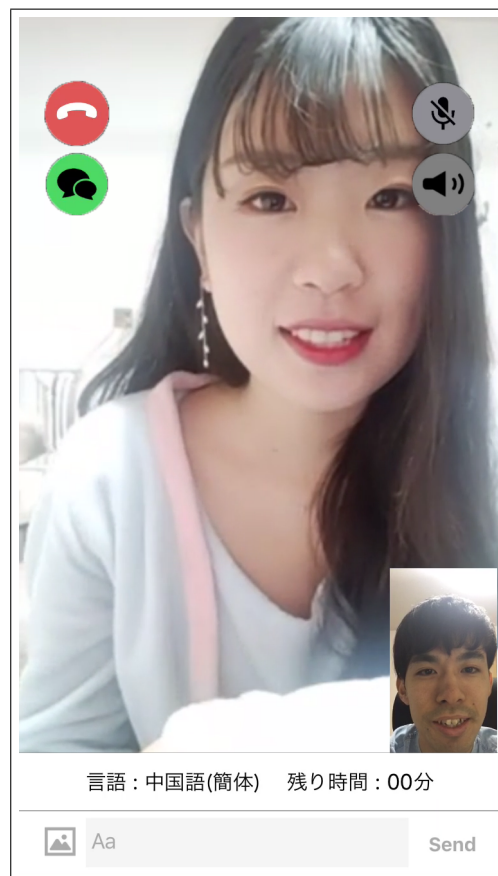


図 4.15: レッスン画面

チューターとして使う場合には、学習者との通話によってアプリ内通貨を獲得することができる。この獲得したアプリ内通貨は、自身の学習言語のチューターとの通話に充てることも可能であり、または現金化することも可能である。このフレームワークにより、外国語学習が目的ではない母語話者、即ち報酬のみを目的とする母語話者を取り込むことができると考えられる。

4.8.1 モバイルアプリの試験運用

ここまで述べた概念を実現するために、上述の機能を実装したモバイルアプリを試験的に運用した。運用開始から約2週間で合計464名のユーザがアプリをダウンロードし、295名がユーザー登録した。また14名の中国語母語話者、1名の広東語母語話者、9名の英語母語話者、34名の日本語母語話者、4名のベトナム語母語話者がチューターとして登録した。中国語のチューターと日本人学習者の実際のレッスン画面を図4.15に示す。2020年1月時点で約20回のレッスンが実施され、学習者からの評価は概ね良好である。今後ユーザー数が増えた場合には、学習者間相互シャドーイングのフレームワークを一つの機能として導入し、学習者音声・母語話者シャドーイング音声の平行データを大量に集め、コーパスを構築することを検討したい。

第5章

実応用へ向けた関連研究

5.1 はじめに

第4章では、本研究の提案手法である母語話者シャドーイングの有効性を実験的に検証した。本章では、実応用に向けて更なる検討を行った関連研究に関して述べる。

5.2 生体情報解析に基づく学習者音声の可解性計測 [6]

2.3.3では、EEG及び瞳孔を見ることで母語話者の聴取の様子をセンシングする研究事例を紹介した。このように生理心理学の実験では、被験者に何か一定のタスクを与えることで実験に専念させ、生体情報の変化を見ることが多い。[6]では、母語話者シャドーイングを一つのタスクとし、シャドー時の表情の変化を、筋電センサー、及び動画像を用いて解析し、刺激として提示した学習者音声の可解性を予測している。以下に実験の詳細を述べる。

実験で母語話者に提示した刺激音声は、約60名のベトナム人日本語学習者による日本語読み上げ音声3203音声で、時間的なバランスを考慮しつつ、それぞれ1時間半ほどの4つのデータセットに分けられた。それぞれのデータセットに対して各1名の日本人被験者がシャドーイングを行い、シャドーイング時には図5.1に示すような筋電センサーを取り付けた。実験の様子は図5.2に示す。また、動画像からの表情解析には図5.3に示す特徴量を用いた。

実験の結果、表情センサー及び動画像から抽出した特徴量と可解性の主観スコアとの間に0.6-0.9程度の比較的高い相関係数が得られた。

このように、聞き取りが難しい際には表情に一定の変化が現れることがわかった。しかし、筋電センサーの電位が被験者によって正負が反対に出るケースもあり、被験者によって表情の作り方には差があることがわかった。

さらに、表情に関しては文化差などもある為、やはり実応用するのであれば音響分析の結果を用いる方が汎用性が高いと言える。

ただし、音響分析の際に用いる音響モデルは日本語の音声コーパスを用いて構築したものであり、GOP計算などに必要は強制アライメントは日本語以外には適用できない。

[1]では、英語音声のDTW計算に日本語音響モデルを試しに適用した例があり一定の精度があることがわかっている。また[21]では、5言語の音素を国際音声記号(IPA: International Phonetic Alphabet)を用いて表記したコーパスを用いて音響モデルを構築し、英語と日本語の評価に適用している。

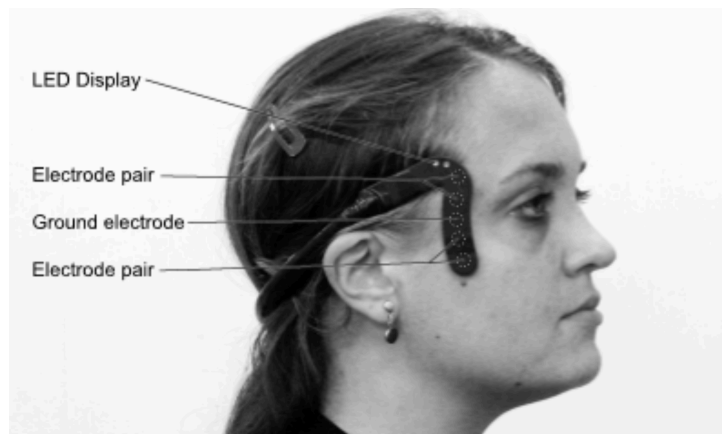


図 5.1: 表情センサー [6]



図 5.2: 表情解析実験 [6]

学習者音声評価技術の応用を考えた場合には、「言語非依存性」のあるシステムの開発が特に重要となるだろう。

5.3 母語話者シャドーイングは了解性・可解性のどちらを反映しているか [7]

本提案手法である母語話者シャドーイングであるが、聴取音声に含まれる単語を同定しながら復唱しているという点では了解性にも近い概念のように思われる。ここで了解性と可解性の違いについて再度確認すると、了解性とは実際に聞き取ることが出来た単語数を元に計測されるものであり、意味内容の理解は問わない。一方可解性とは意味内容の理解しやすさに対する尺度であり、単語と単語のつながり、すなわち統語的な要素の影響も受ける。そこで、シャドーイングの際に意味内容を理解しながら復唱しているのか、あるいは聞こえた単語を「音」として捉え復唱しているのかを実験的に分析することで、シャドーイングの円滑度は了解性と可解性のどちらに近い尺度であるかを検討可能である。

4節の実験では、日本語中級レベルの教科書から文章を選択した。また Jreadability というツールを用いて、それぞれの文章がおおよそ同レベルの文章難易度に属していることを確認した。よって学習者の発話内容の文法的な難易度は同一であるという仮定の元に解析を行った。

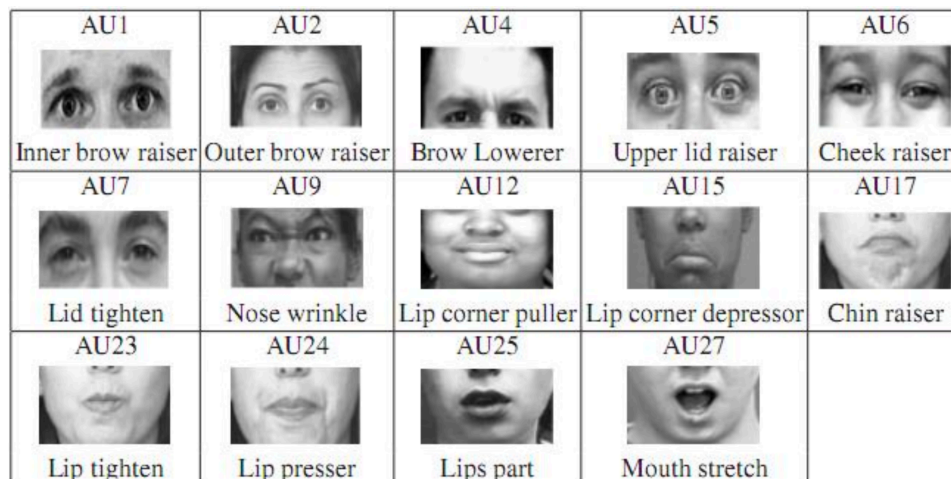


図 5.3: 表情アクションユニット [6]

表 5.1: プロのナレータに読み上げさせた文章セット

set	source
A	桃太郎
B	日本語学習者向けに平易な文で書かれた NHK News Web Easy の記事
C	NHK News Web Easy から抽出した単語をランダムに並べた文章
D	日本人向けに書かれた NHK News Web の記事
E	日経サイエンスの記事
F	ランダムな平仮名列

しかし実際には、諺や昔話などのような何度も耳にしたことのある文章と、学術的に書かれた文章とでは、母語話者が聴取する場合でも理解しやすさに差があることは容易に想像できる。

そこで文章的な難易度の異なる文章セット群を 6 セット用意し、それぞれを日本語母語話者である同一のプロのナレータに読み上げさせ、その音声を日本語母語話者にシャドーイングさせた。日本人のナレータによる明瞭な発音を、日本人がシャドーイングする為、読み上げ話者の発音による影響を取り除き、文章の難易度による影響のみを観測することができる。以下では実験の詳細を述べる。

5.3.1 文章難度を考慮した読み上げ音声の収録

読み上げに用いた文章セットを、表 5.1 に示す。まず最も聴き慣れた文章として、桃太郎の冒頭の数フレーズをセット A とした。桃太郎の冒頭部分は一般的な日本人であれば既知である為、シャドーイングが最も円滑に行われると予想される。続いてセット B は日本語学習者向けに平易な単語や文法を用いて書かれた NHK News Web Easy, というニュースサイトから引用したフレーズである。こちらも日本語母語話者であれば容易にシャドーイングが可能であると言える。セット C は NHK News Web Easy から機能語を省いて抽出した単語をランダムに並べたフレーズである。ここで 5.3 節での議論に立ち返ると、セット B は了解性・可解性ともに高いフレーズ、セット C は了解性は高いが、ランダムな単語間に意味的なつながりはない為、意味内容を理解で

表 5.2: 文章セットの特徴比較

set	単語頻度	単語間の予測性	フレーズ間の予測性	文法的複雑性
A	○	◎	◎	◎
B	◎	◎	○	◎
C	◎	×	×	×
D	○	○	○	○
E	△	○	○	○
F	×	×	×	×

きない、すなわち可解性が低いフレーズということになる。この二つのセット間のシャドーイング円滑度を比較することで、母語話者シャドーイングの円滑度は了解性、あるいは可解性のどちらをより反映しているのかが検討可能である。

セット D は日本人の読者を想定した NHK News Web の記事である。こちらもセット B と比較すると文章の難易度が高いと言える。さらに文章難易度の高い例として、日経サイエンスの記事から選択したフレーズをセット E とした。

またシャドーイングの際に聞こえた音を単純に真似することは可能であるのか、ということを検証する為、ランダムひらがな列フレーズをセット F とした。フレーズの作成の際には、セット A の桃太郎のフレーズのそれぞれのひらがなを、濁音や半濁音を除くランダムなひらがなに置き換え、読み上げの際は元の単語のアクセントに従うよう指示した。

それぞれの文章セットには 20 個のフレーズが含まれている。最終的に用意した文章セットの特徴を一覧にしたものを表 5.2 に示す。

以上の文章を日本語のプロのナレーターに読ませた。読み上げ音声は、話速の目安として 4 節で用いた日本語教科書音声に付属していた CD 音声を適宜提示しながら防音室で収録されたため、ノイズが含まれない明瞭な音声である。

5.3.2 シャドーイング実験

被験者として、7名の成人の日本語母語話者、男性5名、女性2名が実験に参加した。男性被験者は工学系の大学院生のため、文章セット E に対しては比較的慣れていると考えられる。一方女性被験者は秘書であり、文章セット E に対する慣れは少ないと言える。シャドーイング実験の手順は、本研究の実験手法と同様である。

続いて、シャドーイング音声の分析結果を図 5.4, 図 5.5, 図 5.6 に示す。なお、nGOP とはナレーター音声の GOP であり、sGOP とはシャドーイング音声の GOP である。

まず注目したいのが、図 5.4 を見ると分かるように、ナレーターであっても発話内容によっては GOP が変化する。すなわち、文章の難易度が発声のしやすさに影響を与えていることが分かる。T テストの結果、様々な組み合わせで有意差が見られたが、特に注目すべきは BC 間である。何故なら B と C に含まれる単語レベルは同等であるが、B は了解性・可解性が高く、C は了解性は B と同様で、可解性が低くなるはずだからである。

今回 nGOP, sGOP, シャドーイング遅れ時間のすべてで、BC 間に有意差が見られた。以上の結果から、文法的な可解性の違いがシャドーイングの円滑度に影響を与えることがわかったため、シャドーイングの円滑度は可解性をより反映していると言える。

応用言語学の分野では、了解性 (intelligibility) や可解性 (comprehensibility) の他にも、listening effort, cognitive load, interpreterability 等、学習者発音評価のために様々な用語が用いられているが、実際は研究者によってその定義はまちまちである。教育現場の立場から考えた場合には、そのような用語の細かな違いは寧ろそこまで重要ではなく、教師や学習者にとってわかりやすいものが求められている。

「ある学習者音声は母語話者にとってどれほどシャドーイングしやすいか」ということがわかりさえすれば十分だ、という語学教師の声もあり、筆者としては、細かな定義に関しては一度置いておき、実応用に向けてより重要な部分に注力すべきであると考えている。

学習者間相互シャドーイングのような枠組みを実現すれば、自身の外国語音声を母語とした聴取者のシャドー崩れが定量的に得られる。自分の母語を学ぶ学習者音声をシャドーすることによって、シャドーイングが崩れるような音声を聞いた時の聞き取りの難解さ、聞き取りの嫌悪感を学習者自身も実体験できる。つまり、自身の音声は母語話者のシャドーを崩すことの、「コミュニケーション上の意味」を実体験させることができる。

5.4 音声アノテーションへの応用に向けた検討

本研究では、母語話者シャドーイングの円滑度は学習者音声の可解性を反映していることを明らかにした。一般的に教師が行う発音評価は、学習者に対して一つのスコア、あるいは一つの発音全体に対して一つのスコアをつけるのが一般的で、より細かい単位で評価する場合もあるかもしれないが、とてもコストや時間がかかるため、大規模にコーパス化することは難しい。しかし、シャドーイング円滑度の定量化に用いた特徴量は各フレームごとに計算可能であり、つまり時系列データとして学習者音声の可解性をスコア化できる。特徴量計算の精度がより向上すれば、学習者音声に対する音声のアノテーションとして使える可能性がある。そこで続く節では、精度向上のために行った研究を紹介する。

5.4.1 ベトナム人日本語学習者音声の大規模収録 [8]

4章で収録に参加したベトナム人日本語学習者は日本の大学に通う大学生であり、日本の滞在歴が2年以上であったため、習熟度が比較的高い学習者であった可能性がある。そこで日本人との接触がより少ないと予想される、来日前の学習者を対象により大規模な収録を行った。

収録に参加した学習者は、ハノイ大学の日本語専攻の学習者30名、及びハノイ法科大学で日本語を第二専攻とする学習者30名の合計60名であり、27名の学習者は1年間のカリキュラムを修了したばかりで、残りの33名は2年または3年のカリキュラムを修了済みである。前者を初級者グループ、後者を中級者グループと定義する。

読み上げテキストは、[36, 40, 41, 42, 43, 44] から選択し、初級者グループに対しては50個、中級者グループに対しては55個のパラグラフを割り当てた。それぞれのテキストの例を図5.7に示す。また教科書の文章は文法的には誤りがないはずであるが、学習者が話せるようになりたい内容とは解離している場合もある。そこで、今回は学習者自身が作文した文章も読み上げさせた。作文した文章にはもちろん文法的な誤りも含まれる。前者の教科書から選択した文章の読み上げ時には、4.3節で述べたようなカラオケスタイルの録音方法で話速を統制し、後者の作文の読み上げ時には話速の統制は特に行わなかった。

以上の収録により、初級者グループからは914フレーズ、中級者グループからは1963フレーズの教科書読み上げ音声を得られた。音声の総時間長は12656秒(約3.5時間)である。また作

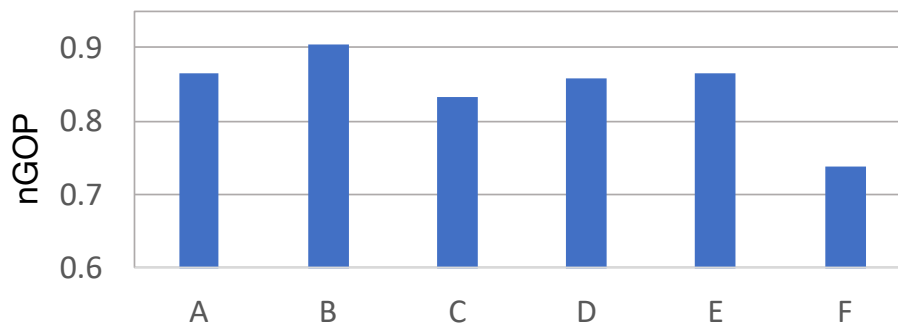


図 5.4: nGOP の結果

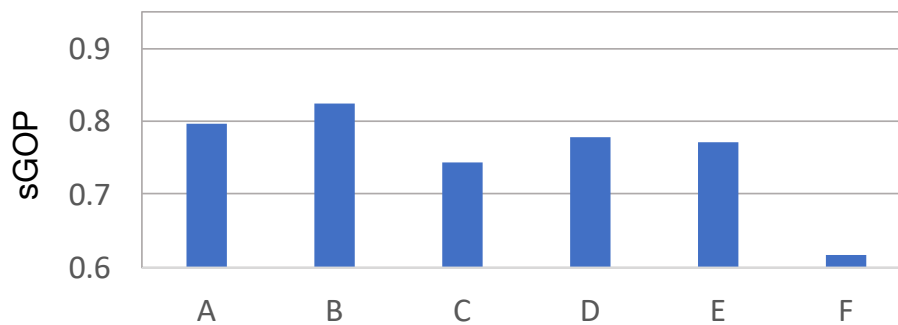


図 5.5: sGOP の結果

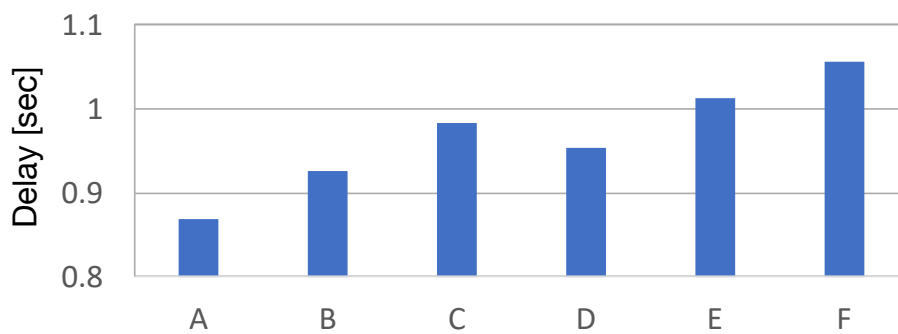


図 5.6: シャドーイング遅れ時間の結果

私の町には
ロシア人のかんこうきゃくがたくさん来ます。
私はロシア語がすこしできますから
ときどきバスの乗りかたをおしえます。
私もいつか
ロシアに旅行に行きたいです。

どこの文化にも昔話があります。
昔話は
昔の人の知恵や遠い昔に起きた事件などが
もとになっていると言われていました。
昔話の不思議な点は
遠く離れた場所に似た話があることです。
たとえばアジアの広い地域に
…

図 5.7: 読み上げ文の例（上段は初級者用，下段は中級者用）

文の読み上げ音声は 478 個得られたが、読み上げの単位に関しては学習者自信が決めたため、それぞれの文単位は異なる。例えば数文節からなるフレーズ毎に読み上げた場合もあれば、一つのパッセージをすべて一度に読み上げた場合もある。最終的に、総時間長 4808 秒（約 1.3 時間）の作文読み上げ音声を得られた。

以上の学習者読み上げ音声コーパスを、4 人の母語話者にシャドーイングさせた。一人当たり、1.2 時間分の学習者音声を割り当てた。またシャドーイングの手順は今回の実験と同様である。

シャドーイングの結果得られた、可解性に関する主観スコア S_C と、シャドーイングのしやすさに関する主観スコア S_S の間の相関係数を表 5.3 に示す。4 章で用いた教科書の読み上げ音声に対する相関係数の平均は 0.71 と、4.5.2 の結果 0.68 と比較するとほぼ変わらなかったが、特別低い被験者もいなかった。この要因は、実験の際に「スムーズなシャドーイングとは何か」という教示をより明示的に与えたことにあると推測される。

また、学習者自身の作文の読み上げ音声に対する相関係数は、教科書の読み上げ音声に対する相関係数よりも高かった。この要因として、教科書のレベルと学習者の習熟度がマッチしていなかった可能性が考えられる。すなわち、一部の学習者は未学習の単語を含む教科書の文章を読む必要があり、その単語を明瞭に発音する傾向がある。そのような音声をシャドーすることは容易であるが、聞いたときの自然性は低いため、可解性スコアが低くなったと考えられる。いずれにせよ、より効果的な学習のためには、学習者自身に自主的に作文を行わせる方が良いことは明らかである。

5.4.2 読み上げ音声とシャドー音声の比較 [9]

本研究では一つの学習者音声に対して、27 名のスコアを平均して音声単位の相関を計算した。何故なら今回用意した被験者において、学習者発音に対する慣れの度合いは影響しなかったため、母語話者が感じる可解性は凡そ等しいという仮定をおいたからである。しかし、応用を考えた場

表 5.3: 被験者毎の S_C と S_S 間の相関係数 [8]

読み上げ内容	S1	S2	S3	S4	avg.
教科書	0.83	0.85	0.73	0.69	0.78
4章で用いた教科書	0.80	0.82	0.64	0.56	0.71
学習者自身の作文	0.89	0.90	0.85	0.73	0.84

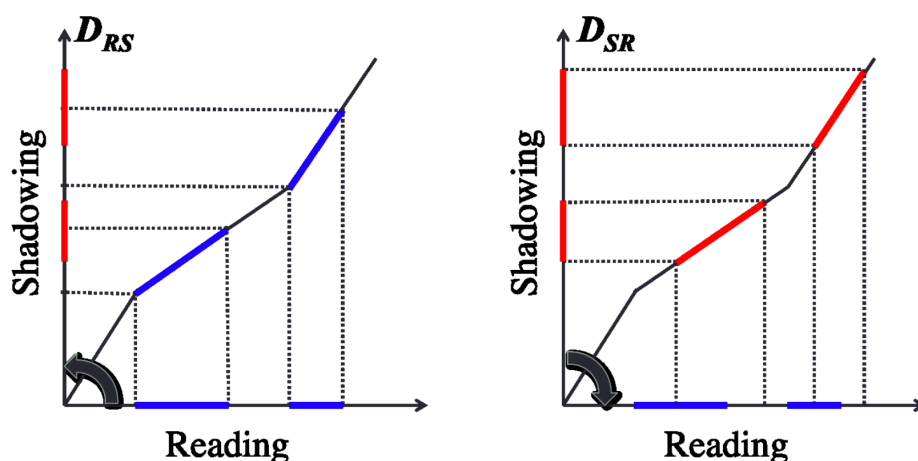


図 5.8: シャドーイング音声と読み上げ音声の DTW [9]

合に一つの音声に対して様々な属性の人がシャドーイングを行い、その平均値で可解性を代表するやり方はコストがかかる上、学習者としてはもう少し特定の属性の母語話者に対して自分の発音がどれほど通じるのかを知りたいはずである。（例えば、高齢者ばかりの職場へ海外就職するのであれば、高齢の母語話者にどれほど聞き取ってもらえるのかが重要である）また、シャドーイングを行う母語話者の発音は、母語であっても訛りや発音の癖がある場合もある。（例えば米語の例を考えれば、地域によってアクセントには差がある。）

そこで、母語話者個人により着目した解析が必要となるため、シャドーイング音声単位でより高精度な予測が出来るような解析を行いたい。

[9]では、母語話者に学習者のシャドーイングさせた後に、学習者が意図したテキストを提示し、そのテキストを読み上げさせた。この場合、同一話者・同一録音環境となるため、DTWの比較においてノイズがかなり少なくなる。別の考え方をすると、母語話者本人から正解音声を得ているとも言える。

さて、シャドーイング音声と読み上げ音声を比較する場合には、もう一点考慮すべきことがある。それは無音の位置である。シャドーイング音声であれば、聴取した学習者音声に元から含まれるポーズに応じた無音部分と、聞き取れずに何も言えなかった場合の無音が含まれるが、読み上げ音声の場合には、母語話者本人が自由に置いたポーズの部分が無音となる。

そこで [9]では、読み上げ音声の無音以外のフレームに該当する DTW の平均距離 D_{RS} と、シャドーイング音声の無音以外のフレームに該当する DTW の平均距離 D_{SR} を計算した。図 5.8 にその概念を示す。また、どちらの平均距離も可解性を一定程度反映していると考え、それぞれを一次結合した、 $\alpha D_{RS} + (1 - \alpha) D_{SR}$ の α の値を変えながら、可解性スコアに対する相関を計算した。

計算の結果、 α の取り方によっては 0.64 という相関係数が得られたが、この値は従来のシャ

ドーイング音声単位の相関係数と比較してかなり改善されている [8]. 今後はテキストを見せながらシャドーイングする「スクリプトシャドーイング」の音声と、テキストを見ずに実施したシャドーイング音声とを比較することで、より条件の揃った分析を行う予定である.

第6章

結論

6.1 まとめ

本研究では、学習者音声の可解性自動計測の新たな手法として母語話者シャドーイングを提案し、学習者読み上げ音声収録、及び母語話者シャドーイング実験を行なった。そしてシャドーの円滑度に関する特徴量を計算し、被験者が付した主観評価スコアとの比較によってこの手法の妥当性を示した。また得られた特徴量から主観評価スコアを予測する回帰モデルを構築し、予測した値と正解スコアに高い相関が得られた。以上より、本モデルを学習者音声の可解自動計測に用いることの有効性が示された。

6.2 今後の課題

シャドーイングを実施した母語話者に学習者が意図したテキストを読み上げさせ、読み上げ音声とシャドーイング音声を比較することが有効であることが示されている。しかし今後の課題として、母語話者シャドーイングを音声アノテーションとして使うためには、より精度の高い分析が必要である。現在考えている方法としては、テキストを提示しながらシャドーイングする「スクリプトシャドーイング」の導入である。スクリプトシャドーイングであれば、ポーズの位置や話速まで条件を揃えることが出来るため、よりノイズの少ない分析が可能である。また、将来的には母語話者シャドーイング音声そのものを生成可能な、「バーチャルシャドワー」の構築を目標としており、そのためにはより大規模なコーパスの収集が必須である。そこで、学習者間相互シャドーイングに向けて開発したモバイルアプリケーションをより活発化させ、より多くのユーザーに対してプラットフォームを提供することで、自然に多くのデータが集まるような枠組みを構築する必要がある。

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。専門外からやってきた知識の浅い私に対し，研究の心構えや実験作法等に関して丁寧にご指導をいただき，国内外の学会で発表出来るほどに成長することができました。また研究以外では，留学や起業という個人的な事柄に対しても快く背中を押していただき，私の人生を大きく変えるきっかけをいただくことが出来ました。

研究を進める上で様々な指摘やアドバイスを下さった齋藤大輔講師にも深く感謝いたします。齋藤講師からは，多角的な視点で問題を解釈することが大事であるということ学びました。また研究生活の様々な面においてサポートして下さった高橋登技官，秘書の池上恵さんにも深く感謝いたします。

研究室の先輩・同期・後輩の皆様にも感謝いたします。修士課程から東京大学に入学した私に対しても，分け隔てなく接していただき，研究のモチベーションを維持することができました。

また，ベトナム人日本語学習者の収録に関してご協力くださった名古屋経済大学の金村久美教授にも感謝いたします。金村教授のご協力がなければ本研究で成果を出すことは不可能でした。

最後に，今まで私を経済的・精神的な面で支えてくださった家族・友人に深く感謝いたします。本当にありがとうございました。

2020年1月30日

井上 雄介

参考文献

- [1] 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.
- [2] Jared Bernstein. Objective measurement of intelligibility. In *Proc. ICPHS*, pp. 1581–1584, 2003.
- [3] Nobuaki Minematsu, Koji Okabe, Keisuke Ogaki, and Keikichi Hirose. Measurement of objective intelligibility of japanese accented english using erj (english read by japanese) database. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.
- [4] Murray J. Munro and Tracey M. Derwing. Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, Vol. 45, No. 1, pp. 73–97, 1995.
- [5] Jieun Song and Paul Iverson. Listening effort during speech perception enhances auditory and lexical processing for non-native listeners and accents. *Cognition*, Vol. 179, pp. 163–170, 2018.
- [6] Tasavat Trisitichoke, Shintaro Ando, Daisuke Saito, and Nobuaki Minematsu. Analysis of Native Listeners’ Facial Microexpressions While Shadowing Non-Native Speech — Potential of Shadowers’ Facial Expressions for Comprehensibility Prediction. In *Proc. Interspeech 2019*, pp. 1861–1865, 2019.
- [7] Tasavat Trisitichoke, Shintaro Ando, Yusuke Inoue, Daisuke Saito, and Nobuaki Minematsu. Influence of content variations on smoothness of native speakers’ reverse shadowing. *Proc. ICPHS (to appear)*, 2019.
- [8] Ando Shintaro, Lin Zhenchao, Trisitichoke Tasavat, Inoue Yusuke, Yoshizawa Fuki, Saito Daisuke, and Minematsu Nobuaki. A large collection of sentences read aloud by vietnamese learners of japanese and native speaker’s reverse shadowings. In *Proc. Spring Meeting of Acoustic Society of Japan*, pp. 111–114, 2019.
- [9] Zhenchao Lin, Yusuke Inoue, Tasavat Trisitichoke, Shintaro Ando, Daisuke Saito, and Nobuaki Minematsu. Native listeners’ shadowing of non-native utterances as spoken annotation representing comprehensibility of the utterances. In *Proc. SLaTE 2019: 8th ISCA Workshop on Speech and Language Technology in Education*, pp. 43–47, 2019.

- [10] Junwei Yue. DNN-based automatic assessment of shadowing speech. Master's thesis, The University of Tokyo, 2017.
- [11] Reima Karhila, Sari Ylinen, Seppo Enarvi, Kalle Palomäki, Aleksander Nikulin, Olli Rantula, Vertti Viitanen, Krupakar Dhinakaran, Anna-Riikka Smolander, Heini Kallio, Katja Junntila, Maria Uther, Perttu Hämäläinen, Mikko Kurimo. Siak — a game for foreign language pronunciation learning. In *Proc. Interspeech 2017*, pp. 3429–3430, 2017.
- [12] W. Li, S. M. Siniscalchi, N. F. Chen, and C. Lee. Improving non-native mispronunciation detection and enriching diagnostic feedback with dnn-based speech attribute modeling. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6135–6139, March 2016.
- [13] Wei Li, Kehuang Li, Marco Siniscalchi, Nancy Chen, and Chin-Hui Lee. Detecting mispronunciations of l2 learners and providing corrective feedback using knowledge-guided and data-driven decision trees. pp. 3127–3131, 09 2016.
- [14] Murray J. Munro and Tracey M. Derwing. The functional load principle in esl pronunciation instruction: An exploratory study. *System*, Vol. 34, No. 4, pp. 520 – 531, 2006.
- [15] Tracey M Derwing and Murray J Munro. *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research*, Vol. 42. John Benjamins Publishing Company, 2015.
- [16] Braj B Kachru, Yamuna Kachru, and Cecil L Nelson. *The handbook of world Englishes*. Wiley Online Library, 2006.
- [17] Jennifer Jenkins. *World Englishes: A resource book for students*. Psychology Press, 2003.
- [18] Jeremy Goslin, Hester Duffy, and Caroline Floccia. An erp investigation of regional and foreign accent processing. *Brain and language*, Vol. 122, No. 2, pp. 92–102, 2012.
- [19] Anja Hahne. What's different in second-language processing? evidence from event-related brain potentials. *Journal of psycholinguistic research*, Vol. 30, No. 3, pp. 251–266, 2001.
- [20] Avashna Govender and Simon King. Using pupillometry to measure the cognitive load of synthetic speech. *System*, Vol. 50, p. 100, 2018.
- [21] 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.
- [22] 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.
- [23] 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.

- [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] 梶島優. 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. Master’s thesis, The University of Tokyo, 2018.
- [26] Silke M Witt and Steve J Young. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech communication*, Vol. 30, No. 2-3, pp. 95–108, 2000.
- [27] Dean Luo, Nobuaki Minematsu, Yutaka Yamauchi, and Keikichi Hirose. Analysis and comparison of automatic language proficiency assessment between shadowed sentences and read sentences. In *International Workshop on Speech and Language Technology in Education*, 2009.
- [28] 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.
- [29] 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.
- [30] Lee Ann and Glass James. A comparison-based approach to mispronunciation detection. In *SLT*, pp. 382–387, 2012.
- [31] Nobuaki Minematsu, Koji Okabe, Keisuke Ogaki, and Keikichi Hirose. Measurement of objective intelligibility of japanese accented english using erj (english read by japanese) database. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.
- [32] 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.
- [33] Shakti P Rath, Daniel Povey, Karel Veselý, and Jan Cernocký. Improved feature processing for deep neural networks. In *INTERSPEECH*, pp. 109–113, 2013.
- [34] Karel Vesely, Arnab Ghoshal, LukásBurget, Daniel Povey. Sequence-discriminative training of deep neural networks. In *INTERSPEECH*, pp. 2345–2349, 2013.
- [35] 惠羅さとみ. 高齢化する転換期の労働社会と移民労働者. *労働社会学研究*, Vol. 19, pp. 1–19, 2018.
- [36] 河野麻衣子 吉田佳世松浦真理子. 日本語音読トレーニング. アスク出版, 2014.
- [37] Jreadability. <https://jreadability.net>.

- [38] Nobuaki Minematsu, Kikuko Nishina, and Seiichi Nakagawa. Read speech database for foreign language learning. *THE JOURNAL OF THE ACOUSTICAL SOCIETY OF JAPAN*, Vol. 59, No. 6, pp. 345–350, 2003.
- [39] Sara Kennedy and Pavel Trofimovich. Intelligibility, comprehensibility, and accentedness of L2 speech: The role of listener experience and semantic context. *Canadian Modern Language Review*, Vol. 64, No. 3, pp. 459–489, 2008.
- [40] 赤木浩文 篠原亜紀中川千恵子. 伝わる発音が身につく! にほんご話し方トレーニング. アスク出版, 2015.
- [41] 許舜貞中川千恵子. さらに進んだプレゼンのための日本語発音練習帳. ひつじ書房, 2009.
- [42] 国際交流基金. まるごと: 日本のことばと文化 (入門 A1 りかい). 三修社, 2013.
- [43] 国際交流基金. まるごと: 日本のことばと文化 (初級 1 A2 りかい). 三修社, 2014.
- [44] 国際交流基金. まるごと: 日本のことばと文化 (初級 2 A2 りかい). 三修社, 2014.

発表文献

国際会議

- [1] 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, pp.1651-1655, Hyderabad, India, 2018
- [2] 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 (ISAPh), Fukushima, Japan, 2018
- [3] 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, pp.971-978, Athena, Greece, 2018
- [4] Tasavat Trisitichoke, Shintaro Ando, Yusuke Inoue, Daisuke Saito, Nobuaki Minematsu, “INFLUENCE OF CONTENT VARIATIONS ON SMOOTHNESS OF NATIVE SPEAKERS’ REVERSE SHADOWING”, International Congress of Phonetic Sciences (ICPhS), pp.2553-2557, Melbourne, Australia, 2019
- [5] Haoyu Zhang, Yusuke Inoue, Daisuke Saito, Nobuaki Minematsu, Yutaka Yamauchi, “Computer-aided high variability phonetic training to improve robustness of learners’ listening comprehension”, International Congress of Phonetic Sciences (ICPhS), pp.924-928, Melbourne, Australia, 2019
- [6] Zhenchao Lin, Yusuke Inoue, Tasavat Trisitichoke, Shintaro Ando, Daisuke Saito, Nobuaki Minematsu, “Native listeners’ shadowing of non-native utterances as spoken annotation representing comprehensibility of the utterances”, The Proceedings of Speech and Language Technologies in Education (SLate) , pp.43-47, Graz, Austria, 2019
- [7] Shintaro Ando, Zhenchao Lin, Tasavat Trisitichoke, Yusuke Inoue, Fuki Yoshizawa, Daisuke Saito, Nobuaki Minematsu, “A Large Collection of Sentences Read Aloud by Vietnamese Learners of Japanese and Native Speaker’s Reverse Shadowings”, The Proceedings of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA) conference, Cebu, Philippines, 2019

国内研究会・全国大会

- [8] 張昊宇, 井上雄介, 齋藤大輔, 峯松信明, 増田斐那子, 山内豊 “音声変形技術を用いた外国語聴取能力の頑健化に向けた実験的検討” 日本音響学会秋季講演論文集, pp.371-372, 2017
- [9] 張昊宇, 井上雄介, 齋藤大輔, 峯松信明, 山内豊, 増田斐那子 “音声変形技術を用いた外国語聴解教材の声質操作とそれを用いた日本人英語学習者の聴解能力の頑健性に関する実験的検討” 電子情報通信学会音声研究会資料, pp.31-34, 2018
- [10] 張昊宇, 井上雄介, 齋藤大輔, 峯松信明, 山内豊 “音響変形を施した音声の聴取に基づく外国語理解能力の頑健性に関する実験的検討” 日本音響学会春季講演論文集, pp.1367-1370, 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
- [12] 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
- [13] 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
- [14] Tasavat Trisitichoke, Shintaro Ando, Yusuke Inoue, Daisuke Saito, Nobuaki Minematsu “Influence of content variations on native speakers’ performance of shadowing” The Journal of The Acoustical Society of Japan, 2018
- [15] 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
- [16] 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
- [17] Tasavat Trisitichoke, Shintaro Ando, Yusuke Inoue, Daisuke Saito, Nobuaki Minematsu “Influence of content variations on native speakers’ fluency of shadowing” IPSJ SIG Technical Report, 2018
- [18] Shintaro Ando, Tasavat Trisitichoke, Yusuke Inoue, Fuki Yoshizawa, Daisuke Saito, Nobuaki Minematsu “A large collection of Japanese sentences read aloud by Vietnamese learners and native speakers’ responsive shadowings” The Journal of The Acoustical Society of Japan, 2019
- [19] 安藤慎太郎, トリシティシヨーク・タサバット, 井上雄介, 吉澤風希, 齋藤大輔, 峯松信明 “ベトナム人学習者による日本語読み上げ音声とそれに対する母語話者のシャドーイング音声の収録” 電子情報通信学会音声研究会資料, 2019
- [20] Zhenchao Lin, Yusuke Inoue, Tasavat Trisitichoke, Shintaro Ando, Daisuke Saito, Nobuaki Minematsu “Native Listeners Shadowing of Non-native Utterances as Spoken Annotation Representing Comprehensibility of Non-native Utterances” The Journal of The Acoustical Society of Japan, 2019

- [21] 安藤慎太郎, 井上雄介, 齋藤大輔, 峯松信明, “Posteriorgram-DTW に基づく発話比較における言語依存性の低減に関する検討” 日本音響学会秋季講演論文集, 2019
- [22] Zhenchao Lin, Yusuke Inoue, Shintaro Ando, Daisuke Saito, Nobuaki Minematsu “Native Listeners’ Shadowing of Non-native Utterances as Spoken Annotation for Comprehensibility of the Utterances” IPSJ SIG Technical Report, 2019
- [23] 井上雄介, 峯松信明, 金村久美 “ネット環境を利用した母語話者との音声インタラクションの拡充 – 相互シャドーイングと相互チュータリングを例にとって –” 日本語教育方法研究会, 2020 (提出済み)