Vol.2020-MUS-127 No.38
Vol.2020-SLP-132 No.38
2020/6/7

学習者間相互シャドーイングの実現に向けた音声分析条件と発音教示生成に関する実験的検討

青谷　和真[1]　安藤　慎太郎[2]　井上　雄介[3]　齋藤　大輔[2]　峯松　信明[2]

概要：学習者の発話を母語話者がシャドーイングする逆シャドーイングにより，聴取時の聞き取りやすさ (可解性) を音響的に観測することができる．この逆シャドーイングを，多言語を対象として学習者同士でシャドーしシャドーされる相互シャドーイングへと発展させることができれば，逆シャドー音声の崩れを可解性アノテーションとした大規模データ収集や，学習者へのより直感的なフィードバックが実現できる．本研究では，学習者間相互シャドーイングの実現に向けた音声分析条件と発音教示生成に関する実験的検討を行う．逆シャドー音声の崩れの計測は，逆シャドー音声と (テキストを見ながら行う) スクリプトシャドー音声とを Posteriorgram に変換して比較していたが，様々な言語への応用を考えると，Posteriorgram への変換そのものが困難となることが予想される．同一話者内での発話比較であることから，スペクトルに基づく音響量による直接比較でも十分な精度が期待できる．本研究ではまず，声道長を変化させた場合の (同一話者内) 発話比較を行った．実験結果より，MFCC や PLPCC といった音響特徴量を使用し，距離尺度としてユークリッド距離を用いた場合には，体格の違いが発話間の距離に与える影響が小さいことが示された．また，日本語と中国語を対象として実際に逆シャドーイングによる発話評価を行った実験では，学習者自身による自己評価とは異なる客観評価を示せる可能性を確認できた．

キーワード：相互シャドーイング，可解性，発話比較，Posteriorgram，DTW，MFCC，PLPCC，教示生成

1. はじめに

近年のグローバル化の進行により外国語を運用する機会が増えており，外国語を習得する必要性は高まっている．こうした移民の増加は日本でも起こっており，特に移民の就職に関する支援が重要な課題である．移民の就職先などは外国語の能力に大きく影響を受けるため，コミュニケーションを取ることが求められる職種に就くためには，一定水準以上の語学力が必要となる．こうした状況で学習者に求められているのは，母語話者のように正しい発音で話すことよりも，むしろ，いかに理解される言葉を話すことが出来るかである．一般に学習者の話す言葉は本人には分かり易く，他者にとっての理解し易さ (可解性) が低い場合であっても，学習者本人がそれに気づかないことが少なくない．即ち，自身の発話に対する可解性フィードバックを学習の段階で十分に得ることは，一般に困難である．そこで，学習者発話の可解性を自動推定することができれば，非常に実用的な語学学習支援となる可能性がある．

これまでの研究では，聴取者の瞳孔や，表情の変化といった生理的現象から聞き手の認知負荷を計測する手法が試されることがあった [1, 2]．しかし，特殊な機器が必要となるため，大規模な学習者を対象とするにはコスト的に問題がある．より簡単に可解性を計測/推定する方法として，学習者の発話を聞き手 (多くは母語話者) にシャドーイングさせ，その音声の崩れに着目する手法が提案されている [3, 4]．シャドーイングは発話内容の理解が迅速にできなければ，遅れる，発話が乱れるなどの現象となって観測される [5]．また，計測コストも低く，非常に有望な手法と言える．

学習者発話とそれに対する母語話者の逆シャドー音声から算出した崩れの度合いは一種のデータとアノテーションに相当する．発話の可解性の自動推定を機械学習のタスクとする場合，データ量がモデル精度に直結するため，アノテーションを伴う大量データの収集が極めて重要である．母語話者による逆シャドーイングは計測のコストは低いが，どのように母語話者 (シャドワー) を集めるのか，という問題が残る．そこで有用となるのが学習者間相互シャドーイング (Inter-Learner Shadowing；ILS) である．

[1] 東京大学大学院情報理工学系研究科
[2] 東京大学大学院工学系研究科
[3] 株式会社 OneTerrace

図 1: 学習者間相互シャドーイングの概要図

全ての学習者はある言語の母語話者であり，その言語を学ぶ学習者は必ず存在する．ILS とは，異なる母国語を持つ学習者同士で互いの発話をシャドーする，というものであり，lang-8 の発話版とも言える [6]．各学習者は，自身の母語を学ぶ学習者の発話をシャドーするだけであり，語学教師のように専門的な知識は必要ないため，ILS へ参加するコストは低い．さらに ILS の枠組みに参加することで，学習者は自身の発話の可解性という普段の学習では得にくい情報を知ることができる．ILS の概要図を図 1に示す．

ILS を教育インフラとして提供できれば，各言語に対して様々な可解性のラベルがついた学習者の音声が集まることになる．そこで，ある学習者に対して，「自身の外国語発話が（その言語の）母語話者にとってどのくらい理解しやすいのか」を，同程度の可解性ラベルがつけられた自身の母国語を学ぶ学習者音声を聴取させることで提示すれば，より直感的な教示となる可能性がある．学習者に自身の現状をより直感的に，かつ客観的に把握させることが可能となり，(特に上級者に対しては) 学習のモチベーションを向上させることや，自身の語学力が当該言語のコミュニティーにどのように受け入れられるのか (例えばどういう仕事ができるのか) を，想像することができるだろう．

2. 本研究の目的

本研究では，様々な言語の学習者 (及び様々な言語の母語話者) を相互シャドーイングに参加させ，可解性アノテーションを自動生成しようとする場合に生じうる 2 種類の問題について検討する．一つ目はシャドー音声の崩れの定量化に関する問題，もう一つは教示生成に関する問題である．

聞き手による (逆) シャドー音声から，その崩れを計測する必要があるが，これは，逆シャドー後に，(学習者が読み上げた) テキストを参照しながらシャドーさせ (スクリプトシャドーイング．以降，S シャドーと表記)，両音声と時系列として比較して行う．S シャドーはテキストを参照したシャドーであり，一番上手なシャドー音声が取得できる．

この 2 発話比較であるが，シャドワー A によるシャドー崩れと，シャドワー B によるシャドー崩れを比較する場合，定量的に計測されたシャドー崩れが年齢，性別などの非言語的要因に依存しないことが前提条件である．例えば大人の声は太く，子供の声は細いため，大人のシャドー崩れの方が，より小さく計測される傾向があれば，異なるシャドワー間でのシャドー崩れは直接比較できなくなる．従来この問題に対し，[7] では，比較対象の 2 発話を Posteriorgram に変換した上で，両者を DTW により比較していた．

しかし，Posteriorgram を求めるためには，当該言語の DNN 音声認識用の音響モデルが必要となる．学習者間相互シャドーイングにおいて任意の言語でシャドー崩れを求める場合，全言語（全方言）の DNN モデルが必要となるが，これは現実的ではない．年齢，性別，体格による音声の音響的相違は，音声認識においても大きな問題となる（人によって声は異なる）．本タスクの場合，同一話者内での発話比較（シャドー音声と S シャドー音声）であるため，Spectrogram に基づく音響量であっても，発話差（崩れ）の大きさ推定において，年齢，性別，体格の影響は小さいと期待される．聴取者が感じる心理的差異を，Spectrogram に基づく音響量による 2 発話比較を通して推測する場合，その音響量が人間の聴覚特性を反映しているものが都合がよい．本研究では，聴覚特性を考慮した周波数（メル周波数）軸上でのケプストラムである MFCC と，更に音の強さ（ラウドネス）に対する聴覚特性を考慮した PLPCC を検討する．

二つ目の問題は，教示生成に関する問題である．学習者にとって本人の外国語音声は（母語話者の音声よりも）理解し易い場合もあるだろう．少なくとも自分が理解し難いように外国語を話す学習者はいない．つまり学習者が自身の発声に対して持つ可解性と，聞き手が感じる可解性とが乖離していれば，本手法の意義は高まる．逆にそれが小さければ，相互にシャドーする必要などなくなる．本研究ではこの点について実際にデータを収集して検討する．

3. 先行研究

3.1 GOP による評価と Posteriorgram

学習者の発話の評価として，Goodness Of Pronunciation（GOP）を用いる方法がある [8]．GOP とは対象音声が，意図した発音（音素列）としてどの程度妥当かを表す値である．発話 O=(o_1, o_2, . , o_T) に対する HMM 音響モデルを用いた GOP は，以下となる．

$$GOP(O,\ p_1,\ p_2, .\ ,\ p_N) = \frac{1}{D_O}\sum_t \log\left(P(q_t|o_t)\right) \quad (1)$$

$$= \frac{1}{D_O}\sum_t \log\left(\frac{P(o_t|q_t)P(q_t)}{\sum_{q\in Q}P(o_t|q)P(q)}\right) \quad (2)$$

$$\approx \frac{1}{D_O}\sum_t \log\left(\frac{P(o_t|q_t)}{\max_{q\in Q}P(o_t|q)}\right) \quad (3)$$

ただし，$p_1, p_2, . , p_N$ は意図された音素列，q_t は時刻 t において意図された音素であり，Q は全音素の集合，D_O は発話 O の継続長である．HMM は生成モデルであるため，音素事後確率は近似的に求めざるをえない．しかし，DNN

を用いれば入力を音響特徴量，出力を音素ラベル (前後の音素も考慮に入れたトライフォン) としてモデル学習し，順伝播計算することで音素事後確率を直接計算することができる [9]．DNN 音響モデルの学習は，まず音声データから MFCC のような特徴量を抽出し，特徴量ベクトルを得る．この特徴量ベクトルに GMM-HMM で強制アラインメントを行い，各フレームと HMM 状態の対応を取る．この HMM 状態を出力ラベルとして用い，入力フレームから HMM 状態を同定する DNN を構築する．この DNN を使えば，入力ベクトルに順伝播計算を行い，HMM の状態に対する事後確率を計算できる．これを音素事後確率にする場合は，ある音素に該当する状態に対する事後確率の和を計算する．DNN を用いた GOP は，以下となる．

$$GOP(O, p_1, p_2, . , p_N) = \frac{1}{D_O} \sum_t P(q_t|o_t) \qquad (4)$$

以上は，与えられた学習者発話に対して，意図された音素列が既知である場合の議論であり，その各音素に対して事後確率を計算している．意図された音素列が不明であっても，各時刻の音声特徴に対して全音素（全状態）の事後確率を計算できる．その確率分布をベクトルとみなし，音声を事後確率ベクトル時系列として捉えたのが Posteriorgram である．こうすると年齢，性別，体格などの違いが抑制されるため，話者の個人性が凡そ捨象された音声表現となる．

3.2 シャドーイングによる発話の評価

前節の GOP は発音が母語話者発音に近いかを評価する尺度であり，発話の可解性を直接評価する指標ではない．そこで，学習者の発話が母語話者にいかに容易に理解されるかを測定するために，逆シャドーイングが提案されている．シャドーイングとは一般的には，母語話者による発話を学習者が聞きながら出来るだけ遅れないように繰り返す手法である．この関係を逆転し，学習者の音声を母語話者がシャドーし，その際の発話の崩れ具合から学習者の発話の可解性を評価する．学習者の発話が聞き取り易ければシャドーはスムーズに行なえるが，聞き取りにくい場合には発話が崩れる [5]．[3] では逆シャドー音声に対して GOP スコアを求めて，崩れとしている．[4] では逆シャドー音声とテキスト読み上げ音声との間で DNN-DTW(DNN Posteriorgram に基づく DTW) を行い，崩れを計算している．本研究では [10] で提案された，1) 学習者音声を逆シャドーする，2) 次に学習者が読み上げたテキストを見ながらシャドーする（S シャドー）タスクを通して得られた 2 発話を DTW で比較する．また，Posteriorgram のみならず，MFCC，PLPCC についても検討する．

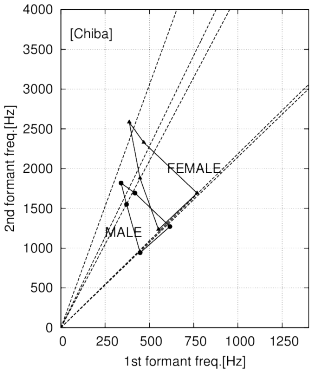

図 2: 声道長の違いと母音フォルマント周波数 [11]
声道長が長くなるにつれてフォルマント周波数は小さくなる．

4. 話者性の低減についての検討

4.1 使用する特徴量

音響特徴量を直接用いて 2 発話間の DTW を行った場合に，話者の年齢や体格の違いが DTW 差異の大きさにどのように影響するのかを実験的に検討する．声道長 (体格) が伸びると，フォルマント周波数は下がる．第 n フォルマント周波数を F_n とした時，喉の形状を断面積一定の閉管であるとモデル化すれば式 5のように表せる．

$$F_n = \frac{c}{4l}(2n+1) \qquad (5)$$

ただし，c は音速，l は声道長である．これより声道長が伸びると，フォルマント周波数が下がることがわかる．図 2に声道長の違いによるフォルマント周波数の変化を表した図を示す．理論的には声道長が無限長になれば，全母音のフォルマント周波数は 0[Hz] に接近し，区別できなくなる．同一話者の 2 発話を比較する場合，片方の発声では/a/の部分がもう片方では/i/となった場合を考えれば，発話間比較は同一話者内の母音差となる．音声学の分野では母音差異を（F_1, F_2）のユークリッド距離で表現することがあるが，これを使うと，話者内の発話比較（発話差異）の計量（測定量）は，体格の影響を直接的に受けることになる．これは学習者間相互シャドーイングには適さない．

そこで音響特徴量として MFCC，PLPCC を使用することを検討する．[12] では FFT ケプストラム空間において声道長の違いは一次変換で近似され，かつその一次変換は回転性の高い行列となることが報告されている．2 つの母音のケプストラムを $\boldsymbol{c}$, $\boldsymbol{d}$ とすると体格の違いは回転行列 A をかけることに近似でき，2 母音間の距離は

$$|A\boldsymbol{c} - A\boldsymbol{d}| = |A(\boldsymbol{c} - \boldsymbol{d})| \fallingdotseq |\boldsymbol{c} - \boldsymbol{d}| \qquad (6)$$

となり，異なる声道長の話者による 2 つの発話間のユークリッド距離は不変であると近似できる．この事実を実験的に検証するが，具体的なケプストラム係数としては周波数軸を聴覚特性を参照して非線形に伸縮した MFCC 及び，

音の強度に対する聴覚特性も考慮した PLPCC を用いる．

4.2 音声資料とケプストラム距離

日本語，中国語に共通する母音/a/，/i/，/u/間の距離を体格，言語を変えて比較することでそれぞれの要因が距離に与える影響について検討した．使用した音声は中国語上級話者である日本人から日本語，中国語それぞれの/a/，/i/，/u/を録音した．その音声に対し，声道長変換を施して擬似的に声道長の長い話者 (1.2 倍) による発声，そのままの話者による発声，声道長の短い話者 (1/1.2 倍) による発声を作成した (以下，これらの発声を巨人声，通常声，小人声と呼ぶ)．図 3に日本人話者による日本語の「あ」の発話の 3 つの Spectrogram を示す．声道長の違いによってフォルマント周波数が異なっていることが確認できる．

各音声サンプルに対して MFCC12 次元，PLPCC12 次元の時系列を抽出し，各母音のフレーム平均を算出し，母音間距離を計算した．距離尺度としてはユークリッド距離 (式 7)，コサイン距離 (式 8) を用いた．

$$d(\boldsymbol{x},\ \boldsymbol{y}) = |\boldsymbol{x} - \boldsymbol{y}| = \sqrt{\sum_i (x_i - y_i)^2} \qquad (7)$$

$$1 - \cos(\boldsymbol{x},\ \boldsymbol{y}) = 1 - \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{|\boldsymbol{x}||\boldsymbol{y}|} \qquad (8)$$

なお，$\boldsymbol{x}$, $\boldsymbol{y}$ を単位ベクトル化したものを $\boldsymbol{x'}$, $\boldsymbol{y'}$ とすると，$d(\boldsymbol{x'},\ \boldsymbol{y'})^2$ は式 9のようにコサイン距離を用いて表せるため，2 つの距離尺度の違いは正規化の有無になる．

$$\begin{aligned} d(\boldsymbol{x'}, \boldsymbol{y'})^2 &= |\boldsymbol{x'}|^2 - 2\boldsymbol{x'} \cdot \boldsymbol{y'} + |\boldsymbol{y'}|^2 \\ &= 2(1 - \cos(\boldsymbol{x'}, \boldsymbol{y'})) \end{aligned} \qquad (9)$$

各条件におけるユークリッド距離，コサイン距離での母音間距離をそれぞれ表 1に示す．MFCC でユークリッド距離を計算した場合，声道長の変化により生じる差は，巨人声と小人声での中国語の/a/と/u/で最大で 9.7%の差が生じたが，その他の場合では約 5%以内の差となっている．一方でコサイン距離の場合は通常声と巨人声で中国語の/a/と/u/で 35%程の差となっており，全体的に，コサイン距離の方が体格の影響が大きい．日本語と中国語で同一母音間の距離を比較した場合においてもユークリッド距離での優位性が見られる．一方で PLPCC を用いた場合は小人声と巨人声を比較すると 10%程度の差があり，MFCC と比較して体格の影響はやや，大きくなっている．この結果から特徴量として MFCC，距離尺度としてユークリッド距離を用いた場合には声道長や言語の違いによる影響はそれほど大きくならないことが確認できた．

5. シャドーイングによる可解性評価

5.1 実験概要

日本語と中国語を学習対象とする異なる学習者から，学習言語の読み上げ音声を収集し，それを母語話者にシャドー及び S シャドーさせ，両者を DTW で比較する．この場合，Posteriorgram，MFCC，PLPCC の 3 種類の DTW を行う．シャドーと S シャドーの違い (前者の崩れ) は，被験者実験により主観的に定量化することができる．ここでは，主観的な崩れと DTW による崩れの相関分析を行う．

次に，学習者自身が自分の外国語音声に対して想定する可解性（母語話がどの程度楽々と理解してくれるのか）と，実際の可解性（母語話者が楽々理解できたかどうか）に開きがあるのかどうかについても，実験的に検討した．

5.2 学習者による音声収録

シャドワーに提示する学習者音声は，同一内容を（例え話者が違っていたとしても）繰り返して示すべきではない．即ち様々な学習者から，異なる内容の読み上げ音声を提供してもらう必要がある．そこで，日本語学習者（中国人），中国語学習者（国籍は多様）それぞれの学習者に自身で自由に作文をしてもらい，その文章を読み上げた音声を収録した．読み上げでは録音のやり直しを許し，言い直し等を含まない音声としている．中国語学習者（国籍多様）8 名，日本語学習者 16 名（国籍中国）から，独自の作文に対する読み上げ音声を収録した．

5.3 シャドーイングと S シャドーイング音声の収録

学習者音声を母語話者にシャドー，及びその直後に S シャドーさせた．中国語学習者の中国語音声は，前節の日本語作文読み上げ音声収録に参加した中国人 (日本語学習者) が行った．より具体的には，中国語学習者音声から連続する 20 秒程度を切り出したものに加え，中国語母語話者音声も中国語教科書から 5 つ選択して追加した．シャドワー 1 人がシャドーする音声は合計 28 個である．日本語学習者の日本語音声も，同様に 20 秒程度の音声となるように切り出し，日本語母語話者の音声も加えた．なお，シャドワーは本学の学部 3 年生 3 名である．また，シャドワー 1 人がシャドーする音声は合計 32 個である．そもそもシャドーイングに慣れていない場合もあるため，母語話者の母国語音声を使ってシャドーイングの練習を行ない，本番のシャドーイング収録を行なった．シャドーイングは録音を一回しか行えないようにしているが，S シャドーイングは繰り返しの録音を許可した．

5.4 シャドー/S シャドー音声の文節自動分割

日本語，中国語の S シャドーイング音声に対し HTK を用いて強制アラインメントを行った．シャドーイング音声に対しては，発音の崩れや読み飛ばしが原因でアラインメントをとることが難しいため，S シャドーイング音声とシャドーイング音声の間で DTW を行い，DTW パスの情

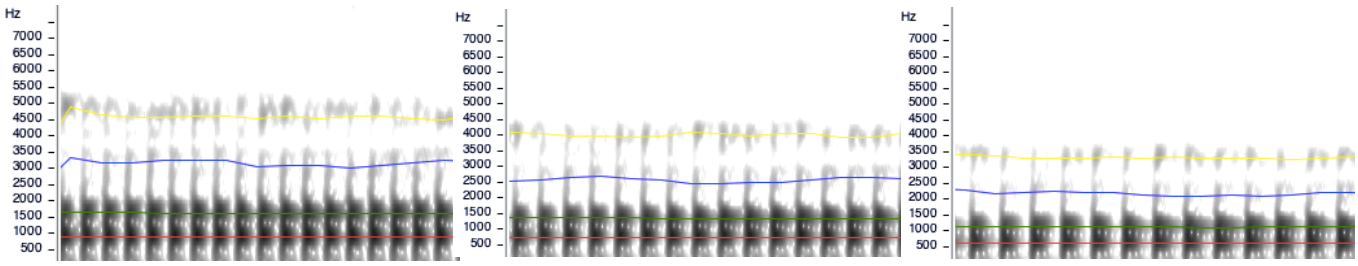

図 3: 日本人話者による「あ」の発話に声道長変換をかけた音声．左から順に小人声，通常声，巨人声である．赤/緑/青/黄の線は F_1，F_2，F_3，F_4 を表している．

表 1: 特徴量，距離尺度，言語ごとの各母音間距離

/a/と/i/ (MFCC)

体格	距離尺度	距離 (日/中)
小人	Euclid	53.54/52.23
通常	Euclid	54.52/52.98
巨人	Euclid	54.93/54.29
小人	cos	0.87/0.76
通常	cos	0.95/0.87
巨人	cos	0.79/0.70

/a/と/u/ (MFCC)

体格	距離尺度	距離 (日/中)
小人	Euclid	36.28/34.46
通常	Euclid	35.66/33.61
巨人	Euclid	35.79/31.41
小人	cos	0.67/0.58
通常	cos	0.70/0.61
巨人	cos	0.62/0.45

/i/と/u/ (MFCC)

体格	距離尺度	距離 (日/中)
小人	Euclid	42.26/38.67
通常	Euclid	43.69/40.20
巨人	Euclid	40.97/39.38
小人	cos	0.45/0.38
通常	cos	0.50/0.40
巨人	cos	0.45/0.33

/a/と/i/ (PLPCC)

体格	距離尺度	距離 (日/中)
小人	Euclid	6.83/6.51
通常	Euclid	6.16/6.11
巨人	Euclid	6.15/6.35
小人	cos	1.35/1.33
通常	cos	1.33/1.24
巨人	cos	1.39/1.35

/a/と/u/ (PLPCC)

体格	距離尺度	距離 (日/中)
小人	Euclid	7.95/7.00
通常	Euclid	7.30/7.32
巨人	Euclid	6.44/6.26
小人	cos	1.37/1.33
通常	cos	1.33/1.24
巨人	cos	1.18/0.88

/i/と/u/ (PLPCC)

体格	距離尺度	距離 (日/中)
小人	Euclid	3.57/6.44
通常	Euclid	3.66/5.95
巨人	Euclid	3.22/5.57
小人	cos	0.79/0.68
通常	cos	0.89/0.70
巨人	cos	0.73/0.83

報から S シャドー音声の文節境界を得た．このようにして，シャドー，S シャドーの全発話対に対して，文節単位での DTW 正規化距離を求めた．シャドーイングの崩れは，発話全体で崩れることもあるが，多くは特定の文節で崩れる．1 発話中に日本語の場合 25.2 個，中国語の場合 13.0 個文節があるため，シャドーの崩れを主観的に評価する場合は，文節単位で行なうのが適切だと判断した．

5.5 シャドー崩れの主観スコアと客観スコアの相関分析

日本語学習者音声に対するシャドー音声，S シャドー音声を文節に自動分割し，文節単位で両者の差異 (崩れ) を主観評価した．評価者は成人男性 1 人であり，7 段階での評価とした (7 に近いほどシャドーイングの崩れは小さい)．

シャドー音声と S シャドー音声の DTW 計算の際には，局所距離としてユークリッド距離，特徴量として MFCC12 次元，MFCC と Δ特徴量，ΔΔ特徴量を合わせた 36 次元（MFCC+Δ＋ΔΔ），PLP+Δ＋ΔΔ，比較対象として日本語話言葉コーパス（CSJ）を用いた構築した DNN 音響モデルによる Posteriorgram を用いた．結果を表 2に示す．MFCC よりも Posteriorgram を使用した方が相関は高くなったがその差はシャドワー B の場合は 2 ポイント，C の場合は 4 ポイント程である．Posteriorgram の算出には DNN 音響モデルが必要となる．学習者間相互シャドーイングは任意の言語の音響モデルが必要としておりこれは実現が難しい．本実験で示された音響量だけを用いた精度は，Posteriorgram との比較という意味において十分な精度が出ていると考察できる．更に，周波数軸のみならず強度についても聴覚特性（等ラウドネス曲線）を考慮した PLPCC の方が僅かではあるが，何れの場合も相関が高かった．なおシャドワー A は B，C と比べて相関が低くなっているが，これはこのシャドワーの滑舌が影響している．この話者は S シャドー時も滑舌があまりよくない話者であったため，シャドーが崩れた場合でもシャドーイング音声と S シャドー音声の DTW の値が小さくなり，相関が悪くなっていると考えられる．シャドーイングは認知負荷が高いタスクである．学習者相互シャドーイングにシャドワーとしての参加者を募る場合，母語音声のシャドーイングの円滑さや，音読の円滑さなど，その話者の母語話者としての特性を考慮して選抜する必要についても検討すべきであろう。

全シャドワーの MFCC+Δ＋ΔΔ を用いた際の文節単位での崩れの評価と DTW の値の散布図を図 4に示す．シャドワーごとに差はあるが主観スコアと DTW による値の間に相関があることが読み取れる．

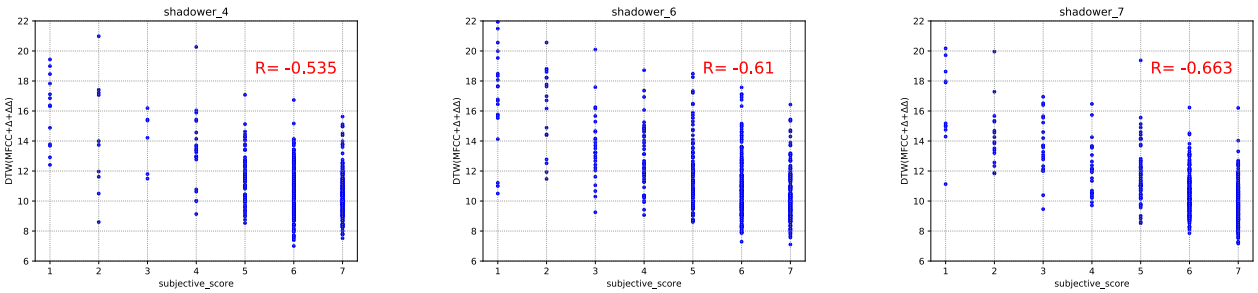

図 4: 文節単位での崩れの主観スコアと DTW 値の散布図 (左からシャドワー A/B/C)

表 2: DTW 値と主観的なシャドー崩れとの相関

シャドワー	特徴量	相関係数
A	MFCC	-0.526
A	MFCC+Δ + ΔΔ	-0.535
A	PLP+Δ + ΔΔ	-0.573
A	posteiorgram	-0.631
B	MFCC	-0.591
B	MFCC+Δ + ΔΔ	-0.610
B	PLP+Δ + ΔΔ	-0.614
B	posteiorgram	-0.633
C	MFCC	-0.651
C	MFCC+Δ + ΔΔ	-0.663
C	PLP+Δ + ΔΔ	-0.670
C	posteiorgram	-0.703

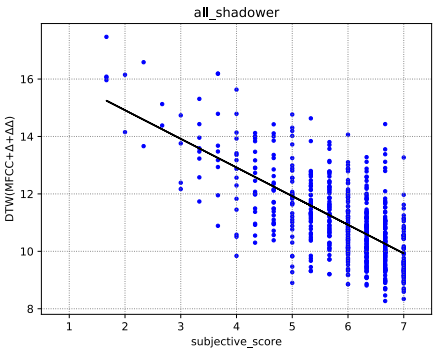

図 5: 3 人のシャドワーの平均

6. 自己評価と他己評価との乖離

図 4の 3 名分のシャドワーによる文節単位の DTW 値の平均値と，3 名のシャドー/S シャドー対に対する崩れ主観スコアの平均をプロットした（図 5）．図中の点は，学習者作文中の各文節（音声）に相当する．この回帰直線から，学習者の発話の上手さを（シャドー・S シャドー間の）DTW 値に基づいて 7 段階で定義した（以下この評価値を客観スコア o と呼ぶ）を，式 10で定めた．これは DTW 値が大きいほどシャドー時の崩れが大きいことを意味している．

$$o = \begin{cases} 1 & (15.5 < DTW) \\ 2 & (14.5 < DTW \leq 15.5) \\ 3 & (13.5 < DTW \leq 14.5) \\ 4 & (12.5 < DTW \leq 13.5) \\ 5 & (11.5 < DTW \leq 12.5) \\ 6 & (10.5 < DTW \leq 11.5) \\ 7 & (DTW \leq 10.5) \end{cases} \quad (10)$$

上記は中国人の日本語学習者の日本語音声を日本人母語話者がシャドー，S シャドーした結果から導出された式である．この式を，様々な中国語学習者の中国語音声を，中国人（即ち日本語学習者）がシャドーした場合の崩れに対して適用し，外国語訛りの中国語音声を同様に 1〜7 と分類した．なお，日本語学習者は複数いるが，個々の学習者の（母語話者としての）シャドー音声を使い，各学習者毎にレベル 1 から 7 に相当する外国人中国語音声を定義した．

その後，中国人日本語学習者 16 名からシャドー及び S シャドー収録を早期に終えた 3 名に対して，自身の作文音声を文節単位で日本人がシャドーした場合にどのくらい崩れると思うのか，を以下のように自己診断させた．まず，崩れ 1〜7 の様々な外国語訛りの中国語音声を聴かせ，自分の日本語音声が文節単位で，どのレベルの外国語訛りの中国語に相当するのかを答えさせた．自身の日本語に自信があれば崩れが小さい (7) 中国語を選び，自信の日本語音声に自信がなければ逆 (1) を選ぶことになる．

中国人の日本語学習者 3 名が主観的に選んだスコアと，日本人によるシャドー崩れから計算された客観的な崩れの対応を表 3に示す．評価の単位は文節であり，数値は該当する (主観スコア，客観スコア) の文節の頻度である．両スコアが類似している場合を黒で，主観スコアの方が客観スコアより 2 以上大きい場合を赤で，逆に主観スコアが客観スコアより 2 以上小さい場合を青で示した．黒の領域に占める割合は左から，61%，55%，70%と過半数を超えているが，黒以外の領域の様子が 3 者では大きく異なった。赤:青を個数で表現すると左から，9:7，2:16，15:0 とバランスよく分布している学習者，過度に自信がない学習者，過度に自信を持つ学習者となった．自信がない学習者に対し

	客観スコア						
	1	2	3	4	5	6	7
1	0	0	0	0	0	0	0
主 2	1	0	0	0	0	0	0
観 3	1	0	0	0	1	0	2
ス 4	0	0	1	0	1	1	2
コ 5	0	0	0	0	2	2	1
ア 6	1	0	0	4	1	4	3
7	0	0	0	0	3	4	6

	客観スコア						
	1	2	3	4	5	6	7
1	0	0	0	0	1	3	1
主 2	0	0	1	0	1	2	0
観 3	0	0	0	0	1	0	1
ス 4	0	0	0	0	1	3	2
コ 5	0	0	1	1	1	4	1
ア 6	0	0	0	1	2	5	4
7	0	0	0	0	0	1	2

	客観スコア						
	1	2	3	4	5	6	7
1	0	0	0	0	0	0	0
主 2	0	0	0	0	0	0	0
観 3	0	0	0	0	0	0	0
ス 4	0	0	0	0	0	0	0
コ 5	0	0	1	0	0	2	0
ア 6	0	0	0	0	0	0	0
7	1	3	1	3	6	11	22

表 3: 3 人の中国人日本語学習者による自身の発話への文節単位での自己評価 (主観スコア) と他己評価 (客観スコア)
表中の値はそれぞれのスコアが選択された回数を示す.

ては，もっと自信を持って学習に臨むように，また自信過剰な学習者には，現実をデータに基づいて示すことができる．このように学習者の個性に応じた適切な教示が可能であることが示された．日本における英語教育がそうであるように，中国本土における日本語教育では，教師以外で日本人と会話する機会はほとんど無いのが現状である．このような状況で，母語話者が学習者の音声をどのように感じながら聞いているのかは，学習者には想像することが難しい．本枠組みの検討を継続することで，その枠組みは実現することが可能となるだろう．

7. まとめ

本研究では学習者間相互シャドーイングの実現に向け，シャドー音声の崩れの音響量による定量化と教示生成についての 2 点について検討を行った．

音響量による崩れの定量化について，まず声道長を変化させた 3 種類の音声の母音間距離を，音響量，距離尺度を変えて計算した．その結果，音響量として MFCC，距離尺度としてユークリッド距離を用いた場合は体格の差はによる影響は小さい事が確認出来た．次に日本語と中国語を対象として学習者間シャドーイングを行い，発話の評価を行った．音響量を用いた場合，主観評価の値と DTW の値の相関は Posterior を用いた場合よりは低かったが，評価の指標として有用である事が確認出来た．また，MFCC と PLPCC の比較ではわずかではあるが PLPCC の方が相関の値は良くなった．

シャドーイングによる教示生成の可能性については，学習者による発話の自己評価の結果と他己評価の乖離から，学習者自身の個性に合わせたより効果的な指導の可能性を示した．

今後の課題としては，本研究では 3 人の学習者による自己評価を 3 人でしか行えていないため，より多くの学習者を対象にしてデータを集める事が挙げられる．

参考文献

[1] T.Trisitichoke, S.Ando, D.Saito and N.Minematsu, “Analysis of Native Listeners’ Facial Microexpressions while Shadowing Non-native Speech – Potential of Shadowers’ Facial Expressions for Comprehensibility Prediction –,” in *Proc.INTERSPEECH*, 1861–1865, 2019.

[2] A.Govender and S.King, “Using pupillometry to measure the cognitive load of synthetic speech,” in *Proc.INTERSPEECH*, 2838–2842, 2018.

[3] Y. Inoue, S. Kabashima, D. Saito, N. Minematsu, K. Kanamura, and Y. Yamauchi, “A study of objective measurement of comprehensibility through native speakers shadowing of learners’ utterances,” in *Proc. INTERSPEECH*, 1651–1655, 2018.

[4] Z.Lin, Y.Inoue, T.Trisitichoke, S.Ando, D.Saito and N.Minematsu, “Native Listeners’ Shadowing of Non-native Utterances as Spoken Annotation Representing Comprehensibility of the Utterances,” in *Proc.SLaTE*, 2019.

[5] T.Trisitichoke, S.Ando, D.Saito and N.Minematsu, “Influence of content variations on smoothness of native speakers’ reverse shadowing,” in *Proc.ICPhS*, 2019.

[6] Lang-8, http://lang-8.com

[7] J.Yue, F.Shiozawa, S.Toyama, Y.Yamauchi, K.Ito, D.Saito and N.Minematsu, “Automatic scoring of shadowing speech based on DNN posteriors and their DTW,” in *Proc.INTERSPEECH*, 1422-1426, 2017.

[8] D.Luo, N.Minematsu, Y.Yamauchi, and K.Hirose, “Automatic assessment of language proficiency through shadowing,” in *Proc.ISCLP*, 1-4, 2008.

[9] H.Wenping, Q.Yao, and K.Soong Frank, “An improved DNN-based approach to mispronunciation detection and diagnosis of l2 learners’ speech,” in *Proc.SLaTE*, 71-76, 2015.

[10] 高島諒，“日本語母語話者の英語音声を対象とした逆シャドーイング法の精緻化と様々な L1 話者を用いた分析，” 東京大学工学部電気電子工学科卒業論文（2020）.

[11] 坂田聡，小林真理子，上田裕市．“日本語音声の地域性による母音ホルマント分布の相違と男女間平均声道長比の類似性，”日本音響学会講演論文集，257-258，2015.

[12] D.Saito, N.Minematsu and K.Hirose, “Decomposition of Rotational Distortion Caused by VTL Difference Using Eigenvalues of Its Transformation Matrix,” in *Proc.INTERSPEECH*, 1361-1364, 2008.