161C64
2016/3/2

WithYou: 音声認識を用いたインタラクティブシャドーイングコーチ

張 シンレイ[1,a)]　味八木崇[1,b)]　暦本純一[1,2,c)]

概要：
外国語習得において，お手本のモデル音声を聞きながら即座に復唱するシャドーイングはスピーキングの練習に効果的であることが知られている．一方，シャドーイングでは学習者の能力と関係なくモデル音声が再生されるため，非熟練者にとっては練習が難しい．言い間違いなどでモデル音声から遅れてしまった場合，学習者は音声の再生を手動で停止するか音声の進度に合わせて文章を飛ばしていく必要があり，練習の効率が低下してしまう．本論文では，シャドーイング練習時に学習者の音声をリアルタイムに音声認識することで発話進度を検出し，モデル音声の再生を学習者の進度に合わせる手法を提案する．提案手法を実現した学習システム WithYou によってユーザ評価実験を行った結果，提案手法では従来のシャドーイングよりも効率よく練習ができることを確認した．また，ユーザスタディから得られたスピーキング支援に関する知見についても議論する．

WithYou: An Interactive Shadowing Coach with Speech Recognition

Zhang Xinlei[1,a)]　Takashi Miyaki[1,b)]　Jun Rekimoto[1,2,c)]

Abstract: Speech shadowing is a proven-effective way of practicing speaking skills when learning foreign languages. Though effective, however, it is very difficult doing because the sound playback is irrelevant to learner ' s speech. When learner fails to catch up, he/she will have to reset the playback or jumps directed to where is being played, resulting in a waste of time and learner ' s resilience in doing it. In this paper, we propose a shadowing support system called WithYou. By comparing on-the-fly recognition result and the words currently being played, WithYou is able to detect choke in learner ' s speech and rewind the playback position to the corresponding place where the choke is. We implemented and tested two version of WithYou, and finds our approach is effective in helping users do better in shadowing. We will also discuss the implications we learned from our user study for speaking practice support.

1. はじめに

「私の言語の限界は，私の世界の限界を意味する」という言葉がある．外国語，特に世界共通言語を習得することは，国際的な場でコミュニケーションを図るために不可欠であると考えられる．また，言語の四つのスキル [1] の中で，スピーキングは円滑なコミュニケーションのために重要なスキルであると考えられている [2].

その一方で，スピーキングスキルの習得は初学者にとって容易ではない．先行研究では，第二言語が話される環境にいたり，その言語でコミュニケーションを取ったりすることがスピーキングスキルの習得に効果的であると報告されている [1] が，第二言語学習者が常にそのような環境にいることは難しく，多くの場合は独学でスピーキングを練習しなくてはならない.

スピーキングの練習においては，シャドーイング (shadowing) という，モデル音声を聞きながら聞いた単語を即座に復唱する手法の有効性が知られている [3]．また，シャ

[1] 東京大学大学院 学際情報学府
III, The University of Tokyo
[2] ソニーコンピュータサイエンス研究所
Sony Computer Science Laboratories, Inc. 3-14-13 Higashigotanda, Shinagawa-ku, Tokyo 1410022, Japan
[a)] xiaolei0114@gmail.com
[b)] miyaki@acm.org
[c)] rekimoto@acm.org

ドーイングを練習することで学習者のスピーキングスキルが様々な面で向上することも示されている [4]．しかし，シャドーイングの練習は大変難しく，たとえ有効性を知っていても，その難しさによってこの練習方法を諦めてしまう学習者がいる [5]．これは，シャドーイングをする時に，学習者は自分の声とモデル音声の両方に注意を払わなければならず，また聞いた音声をできるだけ遅延なく復唱しなければならないためであろう．言い間違えたり言いたい単語の発音が遅れたりすると，モデル音声の再生進度との差が大きく開いてしまい，学習者は練習を中断して音声の再生位置を自分が止まったところに調整し再生し直すか，追いかけられなかった部分を諦めて今再生されている単語に合わせて練習を続けることになる．どちらの方法でも失敗したところをすぐに練習し直すことができないため，多くの時間が無駄になってしまう．**図 1** はシャドーイングの典型的な失敗例を示している．

このような問題を解決するために，学習者の進度に合わせて音声の再生位置を調整するシャドーイング練習支援手法，およびそれにもとづいた学習システム WithYou を提案する．提案する手法では，学習者の音声をリアルタイムに音声認識し，モデル音声の再生進度と学習者の発話を単語単位で比べることで，学習者がその音声について正しくシャドーイングしているかどうかを検出する．学習者がついていけていない場合は，提示音によってシャドーイングが失敗したことを学習者に知らせ，モデル音声の再生位置を学習者が失敗したところに戻す．学習者は自分の失敗したところから直ちに練習し直すことができるようになり，手動で音声の再生進度を調整することなく，効率的に練習を行うことが可能になる．

本研究では、再生制御ユーザーインターフェイスのユーザビリティを検討するために二つのバージョンの WithYou を実装した。手動版は、学習者自身が失敗したと感じたときにボタンを押すと、システムがミスのあった部分までモデル音声の再生進度を戻す仕組みになっている。自動版では、システムが学習者のミスを検出すると、ミスのあったところまで自動的に戻って音声が再生される。

2. WithYou を使ったシャドーイング

WithYou を用いたシャドーイング練習の流れを説明する。まず、学習者はシャドーイングで練習したい原稿とそれに対応した音声データを用意する。ここでの音声データは、ナレーター音声の録音やボイスシンセサイザーによって自動合成された音声を想定している。原稿と音声データから、システムはモデル音声に対して音声認識を行い、データに含まれる単語と各単語が出てくるタイミングを自動的に分析する。以後、この分析で得られたデータをアライメントデータと呼ぶ。アライメントデータが得られたら、システム側の準備は完了となる。

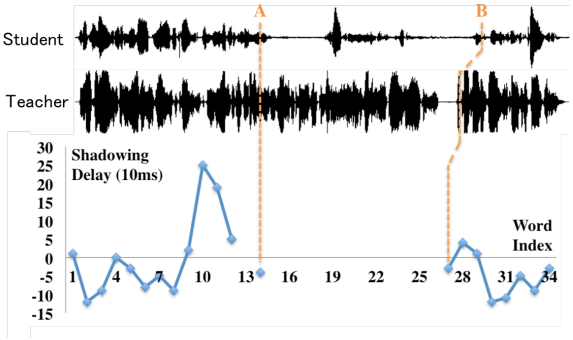

図 1 シャドーイングの典型的な失敗例：学習者はラベル A の時刻まではモデル音声についていけていた．ラベル A の時刻の少し手前から言いよどみが生じ，ラベル B の時刻まではモデル音声についていくことができず，復唱することができなかった．

Fig. 1 An Example of Shadowing failure: Language learner succeeded in following the model sound until the timing marked by label A, but then fail to catch up until the timing labeled by B

システムの準備が整うと、画面にウィンドウが表示され、システムは学習者からの操作を待機する状態になる。学習者がヘッドセットを装着しモデル音声の再生を始めるとシャドーイング練習が始まる。練習中に学習者が言いよどんだりモデル音声に追いつけなくなったりすると、システムはシャドーイングのミスがあったと判断する。手動版では、学習者はミスをしたと思ったときに押したらシステムは提示音を鳴らし、学習者がミスをした部分からモデル音声の再生を始める。自動版では、学習者のミスが検出されるとシステムが自動的に提示音を鳴らし、ミスのあった部分から再生する。この二つのバージョンにおいて、モデル音声が学習者の失敗したところから再生できることにより効率的な練習ができるようになり、また自分の気づかない失敗の練習もできるようになる。自動版の WithYou を使ったシャドーイング練習を**図 2** で説明する

3. システム構成

3.1 音声認識

本システムでは，モデル音声の再生位置制御を実現するために，音声認識技術によって学習者の発話進度をリアルタイムで検出する（**図 3** を参照）．この機能は，オープンソースの汎用大語彙連続音声認識エンジン Julius を使って実現している [14]．Julius は高度にモジュール化されており，ソフトウェアを修正することなく，言語モデルと音響モデルを簡単に組み替えることができる．このため，小語彙の音声対話システムからディクテーションまで様々な幅広い用途の音声認識に対応でき，さらに特定の言語に依存しないインタラクションシステムをつくることが可能になる．これが本システムで Julius を利用した最大の理由である．

本システムでは，Julius を独立した一つのスレッドで常

図 2 自動版の WithYou を使ったシャドーイング流れ：1. 練習のお手本となるモデル音声が再生されている． 2. 学習者がモデル音声を聞いて復唱するが，「mine」の部分でミスしてモデル音声に遅れてしまう． 3. システムがミスを検出し，モデル音声の再生位置を学習者がミスしたところまで戻す．学習者はそこから練習を再開できる．

Fig. 2 Shadowing using WithYou: 1. Sound file for doing speaking practice is being played. 2. User trying to catch up, and failed at the word "mine". 3. System detects that miss, and rewinds the playback to where user made mistake

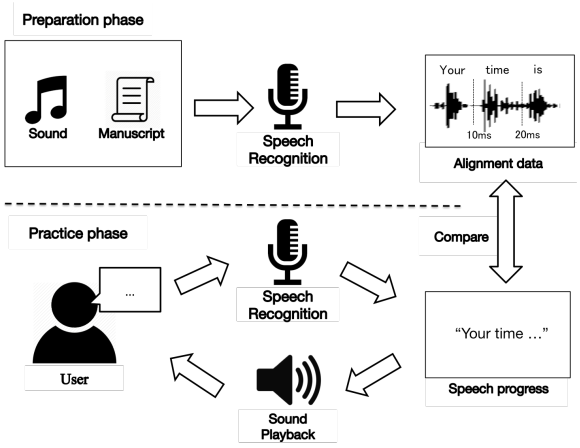

図 3 システム構成

Fig. 3 System Architecture

に動かしており，学習者の音声がマイクから入力されると，その認識の結果がリアルタイムに出力される．Julius を含めシステム全体は C 言語をベースにしているため可搬性に優れており，様々なデバイスに簡単に移植することができる．

3.2 アライメント分析

声データの再生を単語単位で制御する場合、音声データの中で各単語が出てくる順序とタイミングを把握する必要がある。この分析は Forced Alignment という処理によって行う。Forced Alignment は、音声データと正解のスクリプトを用意し、音声データからスクリプトの各単語が現れるタイミングを抽出する処理である。本システムでは Julius をベースに Forced Alignment を実装し、学習者が用意した原稿と音声データでアライメント分析を行う。

3.3 記述文法による構文制限

学習者の発話進度を把握するには，リアルタイムで音声を認識しなければならない．しかし，一つの言語には何万個もの単語があり，それらすべてをリアルタイムで認識しようとすると，計算量が膨大になり認識精度も低下する．一方，シャドーイングの練習では学習者は原稿の内容通りに発話していくため，発すべき単語とその順番は決まっている．このため，システム側では全ての単語を候補として音声認識を行う必要がなく，原稿に含まれている単語だけを候補として計算することで，よりロバストで高速な音声認識を実現できる．

このような構文制限をつけて音声認識を行うには，発話されうる文のパターンを定義する「記述文法」を認識エンジンに与える必要がある．記述文法は.grammar ファイルと.voca ファイルによって構成される．.grammar ファイルは認識できる単語間の繋ぎ方を定義し，.voca ファイルは各単語の表記とその読み方の音素列を認識エンジンに登録する．

記述文法によって音声認識に制限をつけることで，認識処理の出力は.grammar ファイルに書かれた文に限定され，認識には.voca ファイルに登録した音素列の音響モデルが利用される．

3.4 音響モデルと言語モデル

音響モデルと言語モデルは音声認識を機能させるために必要なオブジェクトである．音響モデルは音響的特徴と言語的特徴を結びつけるもので，言語モデルはある単語列が文としてその言語に出てくる確率を決めるものである．

WithYou では，音響モデルにオンラインで提供されている無料の英語音響モデルを使用し，言語モデルには先ほど述べた記述文法を利用している．このために，学習者が用意する原稿から自動的に記述文法を生成するプログラムを実装した．

3.5 無音区間の対応

シャドーイングをしている時に言いよどみが生じ，学習者がモデル音声から遅れると，音声データの中では比較的長い「無音区間」が生じる．音声認識にとって無音区間は「利用者が話していない状態」に等しく，一般的な音声認識エンジンでは無音区間があると認識が止まってしまう．しかし，シャドーイングでは学習者が発話に詰まってから言い直すことはよくあり，音声認識がそこで止まることは望ましくない．

そこで本システムでは，音響モデルに無音区間を代表する「無音単語」用の音響モデルを導入し，言語モデルにも無音単語を一つの単語として全ての文章の単語の後ろに追加した．これによって無音区間を単語として認識することができるようになり，言い詰まった後に言い直しても継続的に認識処理を実行できるようになった．

3.6 シャドーイングの失敗判定

シャドーイングの支援をおこなう際，どのような基準でシャドーイングの失敗を判定するかは重要な問題である．学習者はモデル音声を即座に復唱するとはいえ，モデル音声の再生と完全に同じタイミングで発話することは非現実的である．逆に，音声を聞いてから復唱するまでの反応時間の許容範囲を広げすぎると，練習としての効果が薄れてしまう．このことから，シャドーイングを行うとき，モデル音声からの時間的遅延をどの程度許容するかを明確に定義する必要がある．

本研究では，この遅延を「モデル音声の再生進度と学習者の発話進度の差分」と定義する．実際のシステムではこの閾値 N を 5 に設定している．つまり，モデル音声の再生進度と学習者の発話進度の差分が 5 単語を超えると，シャドーイングが失敗したと判定される．この閾値は予備実験によって決定した．

予備実験では，被験者 9 名に手動版システムを使ったシャドーイングを体験してもらい，被験者が「R」を押したときの「モデル音声の進捗と学習者の進捗の差分」を計測した．その結果，$N = 5$ が妥当であることが分かった $(6/9, Mean 4.88, SD 0.6)$.

4. 評価実験

WithYou を用いたシャドーイング練習の効率とシステムの有効性を検証するために，ユーザーテストを実施した．

4.1 実験の条件と手順

本実験ではアメリカの Educational Testing Service(ETS) による英語リスニングテストのスクリプトを 3 つ利用し，システムなし・手動版システム使用・自動版システム使用の 3 条件でシャドーイングの練習をするタスクを設定した．

被験者のシャドーイングに対する認識を統一するため，実験前にシャドーイングの定義を説明し，一般的なシャドーイングのデモビデオを見せた．練習方法に対して不明な点がないことを確認してから，被験者に実験で使用する文章の印刷原稿を渡し，単語や文章の理解に問題がないかどうかを確認してもらった．確認が済んだら，印刷原稿は回収する．ここで一回通常のシャドーイングをしてもらい，これを被験者の練習前のパフォーマンスとして録音した．被験者は上記の 3 条件のうち指定された一つの手法でその原稿を 3 回練習し，最後にもう一度通常のシャドーイングをして録音した．これが被験者の練習後のパフォーマンスデータとなる．この 5 回の練習を 1 セットとし，原稿と手法を変えて一人の被験者につき 3 セットを行った．実験における手法の順番と文章はラテン方格により決定し，順序と文章の差異による影響を考慮した．

実験の被験者には，日本で暮らす英語を母国語としない 20 − 30 歳の外国人留学生 11 名と日本人 1 名を採用した．手法の順序には 6 つのパターンがあるので，1 つのパターンにつき被験者 2 名で実験を行った．各手法の練習が終わった後にその手法に関するアンケートを記入してもらい，アンケートの回答内容に基づいたインタビューを行った．本実験では，実験の全過程を録音し，実験中のリアルタイムの認識結果とその時刻も記録した．後述の実験結果は，この記録を利用して事後解析を行ったものである．

4.2 実験結果

図 4 は，通常のシャドーイング（S），手動版（A），自動版（B）の 3 つの手法の練習前と練習後のシャドーイングの発話タイミングの向上率を示している．発話のタイミングは，モデル音声とシャドーイングの音声の各単語の時間的差分から，その平均値を求めることで算出する．提案手法 X における各単語の時間的差分の平均値の求め方は以下の数式によって定義されている．x_i と m_i はそれぞれ手法 X，モデル音声のシャドーイングにおいて第 i 個目の単語が発音された時点を示す．図 4 で示す百分率点は練習前と練習後における平均値の変化率である．左側の P1 - P12 は 12 名の被験者番号である．

Participants	English Level	Timing Improvement			Better than S	
		S	A	B	A	B
P1	TOEFL 101	12%	24%	7%	11.5	-5.2
P2	TOEFL 100	2%	16%	11%	14.8	9.7
P3	TOEIC 903	1%	23%	-7%	21.8	-8.4
P4	TOEFL 86	-18%	0%	-32%	18.0	-14.1
P5	TOEFL 88	0%	-10%	2%	-10.2	1.9
P6	TOEFL 81	10%	-11%	-7%	-20.9	-17.3
P7	TOEFL 101	11%	27%	17%	15.8	6.1
P8	TOEFL 100	15%	16%	2%	1.1	-13.4
P9	TOEIC 815	19%	10%	10%	-8.6	-8.6
P10	TOEFL 110	2%	21%	13%	18.5	11.3
P11	TOEFL 97	-24%	3%	15%	26.3	38.9
P12	TOEFL 107	-48%	21%	11%	68.8	58.4

図 4 被験者と実験結果
Fig. 4 Participants and Experiment Result

$$\Delta_X = \frac{\sum_{i=1}^{N}(x_i - m_i)}{N} \quad (1)$$

結果に示したように，発話のタイミングの向上に関しては A 手法（手動版システム）が一番有効であることが分かった．自動版と通常のシャドーイングは，人によって効果が異なっており，一般的な傾向が見られなかった．実験が終わって被験者にインタビューしたところ、6 人の被験者が「自動版ではミスした部分を重複練習させるため，その部分の練習は十分に行えるかもしれないが，文章全体のリズムやスピードの把握に影響がある」とコメントがあったため、それが一つの原因だと考えられる．

4.3 アンケート結果

表 1 はアンケート項目それぞれに対する結果を示している．各アンケートの質問項目は，Q1:「この手法で Shadowing をやるときのやり易さについて評価してください」，Q2:「この手法で Shadowing をやるときの時間的効率について評価してください」，Q3:「システムの詰まり判定はどのくらい自分の感覚に近いですか？」と設定した．時間的効率は「どのくらいの時間きちんと口が動いて発話の練習ができたか」と定義した．Q1 と Q2 は各手法について評価してもらい，Q3 は提案手法の自動版と手動版についてのみ評価してもらった．手法 S の通常のシャドーイングでは詰まり判定がないため Q3 は無効とした．各項目に対しての回答は，「悪い・難しい・遠い」とネガティブに評価した場合は 7 を，逆に「良い・やり易い・近い」とポジティブに評価した場合には 1 とした．回答における，中央値，標準偏差およびポジティブ側（4 以下）への回答割合 (表中では%と示す) の結果を，表 1 に示す．

このアンケート結果における Q1,Q2 について A と S の顕著な差により、提案手法の手動版が普通シャドーイングよりやり易くて練習の効率がよいとされることが分かった。また、被験者に「この三つの手法においてどれが好きですか？」と聞いたところ、9 人が提案手法の方を選んで、そこ中に手動版が一番好きのが 5 人で、4 人は自動版を選んだ。その原因について問いかけたところ、「普通のシャドーイングではうまく追っ掛けないのがよくあるが、それが発生したらその分の時間が無駄になってしまうし、一方、システムを使う場合だと簡単に再練習することができるところがいい」と言ったコメントが多かった。

5. 関連研究

音声認識を用いた外国語習得サポートについての前例は多くある。Project LISTEN[6] は音声認識を用いた読書サポートシステムで、学習者の読み上げ音声を聞いて評価し、学習者がミスをしたり言い淀みのがあったらワンクリックでサポートが得られる。SPELL[7] は自動的に学習者の発音を評価するシステムである。他に、音声認識を用いた発音トレーニングシステムや [8][9] 音声言語インタラクションシステム [10] も提案されていた。音声認識用いシャドーイングへのサポートシステムについて前例の多くは発音評価に注目するものが多かったが [11][12][13]、シャドーイングのインタラクション自体を変えることでシャドーイングの練習をサポートするシステムに関してはこのシステムが初だと考える。

6. 結論と今後の展望

本論文では，WithYou という，利用者の発話進度に合わせてモデル音声の再生進度を調整するシステムを提案した．また，練習に失敗したことの判断を利用者に任せるかどうかによって，学習者自身が失敗したと感じたときにボタンを押して音声を戻す手動版と，自動的に学習者のミスを検出し音声戻す自動版を実装した．評価実験により，発音のタイミングの向上の点では手動版が最も有効であることが分かった．

インタビューから，シャドーイングをするときは自分の発話進度をモデル音声の進度と合わせなくてはならないため，お手本との違い（例えば is と was）に注意を払うより

表 1 アンケート項目に対する結果
Table 1 Result of Subjective Evaluation

	S			A			B		
	Mean	SD	%	Mean	SD	%	Mean	SD	%
Q1	3.75	1.28	5/12	2.91	1.50	8/12	4.16	1.40	5/12
Q2	3.75	1.60	6/12	2.66	1.30	9/12	3	1.53	7/12
Q3	X	X	X	3.5	1.62	5/12	3.75	1.48	7/12

も進度合わせを優先してしまい，ボタンを押さないことがあるということも分かった．これらのことから，練習が中断することなく終わるという性質によって，手動版の方がタイミングのサポートに関する評価が高かったと考えられる．一方，自動版では学習者の操作が一切いらないため，ジョギングや散歩など，様々なシチュエーションでの練習への応用が期待できる．

今後は，今回の実験で得た知見を取り入れてシステムを改善し，新しい実験によってその有効性を確認していく予定である．

7. 参考文献

参考文献

[1] Hinkel, E.: Current Perspectives on Teaching the Four Skills, *TESOL Quarterly* (2006).
[2] Nazara, S.: Students Perception on EFL Speaking Skill Development, *Journal of English Teaching* (2011).
[3] Wilson, M.: SPEECH SHADOWING AND SPEECH COMPREHENSION, *Speech Communication 4*, pp. 55–73 (1985).
[4] 智子, 堀.: Exploring shadowing as a method of English pronunciation training, PhD Thesis, 関西学院大学 (2008).
[5] Satsuki, K. and Soichi, O.: Shadowing,Dictationand ReadingAloud:WhichisEffective?, *The Japan Association of College English Teachers(JACET)* (2012).
[6] Mostow, J., Roth, S. F., Hauptmann, A. G., Kane, M., Mostow, J., Roth, S., Hauptmann, A. and Kane, M.: A prototype reading coach that listens, *Proceedings of the National Conference on Artificial Intelligence*, JOHN WILEY & SONS LTD, pp. 785–785 (1994).
[7] Hiller, S., Rooney, E., Laver, J. and Jack, M.: SPELL: An automated system for computer-aided pronunciation teaching, *Speech Communication*, Vol. 13, No. 3, pp. 463–473 (1993).
[8] Mak, B., Siu, M., Ng, M., Tam, Y.-C., Chan, Y.-C., Chan, K.-W., Leung, K.-Y., Ho, S., Chong, F.-H., Wong, J. et al.: PLASER: pronunciation learning via automatic speech recognition, *Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing-Volume 2*, Association for Computational Linguistics, pp. 23–29 (2003).
[9] Russell, M., Series, R. W., Wallace, J. L., Brown, C. and Skilling, A.: The STAR system: an interactive pronunciation tutor for young children, *Computer Speech & Language*, Vol. 14, No. 2, pp. 161–175 (2000).
[10] Morton, H. and Jack, M. A.: Scenario-based spoken interaction with virtual agents, *Computer Assisted Language Learning*, Vol. 18, No. 3, pp. 171–191 (2005).
[11] Luo, D., Shimomura, N., Minematsu, N., Yamauchi, Y. and Hirose, K.: Automatic pronunciation evaluation of language learners' utterances generated through shadowing., *INTERSPEECH*, Citeseer, pp. 2807–2810 (2008).
[12] LUO, D., Qiao, Y., MINEMATSU, N., YAMAUCHI, Y. and HIROSE, K.: Analysis and Utilization of Speaker Adaptation Techniques for Shadowing and Read-Speech Pronunciation Evaluation, *IEICE technical report. Speech*, Vol. 109, No. 99, pp. 51–56 (online), available from ⟨http://ci.nii.ac.jp/naid/110007340424/en/⟩ (2009).
[13] Luo, D., Minematsu, N., Yamauchi, Y. and Hirose, K.: Analysis and comparison of automatic language proficiency assessment between shadowed sentences and read sentences., *SLaTE*, pp. 37–40 (2009).
[14] 汎用大語彙連続音声認識エンジン Julius: , available from ⟨http://julius.osdn.jp/⟩ (accessed 2015-09-25).