外国語音声コミュニケーションにおける聴解崩れの計測とその可視化・モデル化

Measurement, visualization, and modeling of listening disfluency in foreign language oral communication

峯松　信明（Nobuaki MINEMATSU, Dr. Eng.）
東京大学大学院工学系研究科・電気系工学専攻・教授
(Professor, Graduate School of Engineering, The University of Tokyo)

電子情報通信学会、音響学会、音声学会、外国語教育メディア学会、情報処理学会、人工知能学会、IEEE、ISCA、他
受賞：外国語教育メディア学会論文賞(2023 年)、電子情報通信学会システムソサイエティ論文賞(2016 年)、音声学会学術奨励賞(2014 年)、他
著書：IT Text 音声認識システム、オーム社(2001 年)、韻律と音声言語情報処理、丸善出版(2006 年)、音声言語処理と自然言語処理、コロナ社(2018 年)、コグニティブインタラクション、オーム社(2022 年)、他
研究専門分野：音声工学　音声科学　外国語教育

あらまし

現代社会の営みは、分野が何であれ、国際化と切り離して考えることは難しい。この場合、各種営みに参与するメンバーが有する、英語で口頭コミュニケーションを図る能力は、時として、その営みの成果に大きく影響する。英語教育の世界で世界諸英語という言葉が広く使われるようになった。英語は母語話者の間ですら共有する発音様式（日本語で言えば東京方言、共通語）が存在せず、各自が方言を話しているのが現状である。また世界に目を向ければ、英語利用者の 3/4 は外国語として英語を話しており、誰もが母語の影響を受けた発音で語りかける。かつての英語音声教育は「母語話者のような発音」を目指す教育が主流であった。現在では、英語発音における母語の影響（所謂訛り）はその学習者の identity の一部であると理解され、「母語話者のような発音」を目指すのではなく、「多様な英語発音の中でコミュニケーションをとる能力」の養成が求められている。

筆者は音声工学者・科学者として長年、外国語音声教育支援に携わってきた。音声教育の技術支援と言うと、学習者の音声を分析し、発音の癖を検出し、それを矯正する手段を提供することが主目的であった。この場合、母語話者発音との比較が大前提であり、それはかつての「母語話者発音を目指す」音声教育に根ざす方法論と言える。英語発音の多様性を前提とした場合、学習者の「発音」学習を支援するのではなく、彼らの「聴取」学習を支援することがより重要であろう。英語を聞いている最中に、途中で単語が分からなくなり、頭が真っ白になる（「？」が浮かぶ）経験をした人は多いだろう。どうすれば、聴取の乱れを時系列として計測できるだろうか？もちろん、聴取行為そのものを音響現象として観測することはできない。科学的に妥当な時系列計測は脳計測に頼ることになるが、教室の学習者全員に脳計測装置を装着して計測することは、コスト的に、実現不可能であろう。

１．研究の目的

本プロジェクトではこの問題を、脳計測技術が広く使われるようになる前に、知覚心理学の世界で使われた「聴取過程を計測・分析するための」心理実験タスクである「シャドーイング」に着目して解決する。そして、1)聴取崩れの時系列計測手法、2)多様な世界諸英語発音に対する聴取崩れの可視化手法を構築した。更には、3)評価者が学習者音声を聴取した時の聴取崩れ（学習者音声に対する了解度、intelligibility）をモデル化することで、virtual shadowing rater, すなわち、intelligibility の自動予測技術を構築した。

２．研究の背景

外国語教育の分野で、提示された音声に対する学習者の聴取能力を計測する場合「書き取りテスト」が多用されてきた。「正しく書き取れない箇所」＝「正しく聞き取れない箇所」として解釈し、学習者の聴取能力とする方法である。この場合脳計測に頼らずとも安価に計測できるが、「聞いて書き起こした」結果は「聞いている時の様子」と一致するだろうか？

例えば１分の提示音声を聞いて書き起こす場合、一度の聴取では書き起こせないだろう。書く行為は時間がかかるからである。複数回の聴取を許せば、それはもう「聞き取りテスト」ではないだろう。更に「書きながらあれこれ考える」ことも起こるし、「スペルを忘れて、何度も書き直す」ことも起きるだろう。こうな

外国語音声コミュニケーションにおける聴解崩れの計測とその可視化・モデル化

Measurement, visualization, and modeling of listening disfluency in foreign language oral communication

ると、「書き取りテスト」は「聞いた後の思考、試行の結果を観測するテスト」と言える。決して、聴取の様子を時系列で計測する方法ではないだろう。以上の問題は、端的に言えば、「聞き取った内容を、書かせることでデータ化する」ことに起因する問題である。

3．聴取崩れを時系列として安価に計測する

聞いた内容を「指で書く」複製行為を通してデータ化するのではなく、「口で話す」複製行為を通してデータ化することを考える。シャドーイングである。

シャドーイングは 1980 年代（脳計測が一般化する前）、知覚心理学、特に音声知覚研究において「聞いている様子を計測する」手段、即ち知覚実験タスクとして導入された。提示音声を、なるべく遅れずに復唱するタスクであり、シャドー音声を通して様々な議論が交わされた。その後脳計測が（研究環境において）一般化してくると、シャドーイングのような間接的な計測手段は用いられなくなった。

やがて 1990 年代になると、シャドーイングは外国語教育の分野で復活することになる。音声は揮発的なメディアであり（目の前に止まり続ける文字言語とは異なり）、提示後は、聞き手の頭の中に記憶としてしか残っていない。つまり聞いた後に、記憶の中にしっかり保持する癖をつけないと、聞いた音のイメージがこぼれ落ち、結局「読めば分かるが、聞いても分からない」状態になる。心理学用語で言えば「作動記憶（ワーキングメモリ）を強化し、聞いた音声を記憶の中にしっかり保持する癖をつける」ということになる。シャドーイングはこのための訓練として導入された。

本研究では、提示音声を聞き手がどこで聞き淀んだのか、聞き取り困難な様子を時系列計測するために、1980 年代の古典的計測を教育目的のために復活させた。そして「シャドー音声の崩れを聴取の崩れ」として解釈し、時系列データ化した。通常シャドーイング訓練は、初めに音源だけを提示して数回シャドーさせた後は、答え（テキスト）を見せてシャドーさせる（スクリプト・シャドーイング）。初回のシャドーイングは発話の崩れが計測されるが、テキストを見ながらシャドーすれば、極めて流暢な発声でシャドーできる。シャドー音声（S）とスクリプト・シャドー音声（SS）と

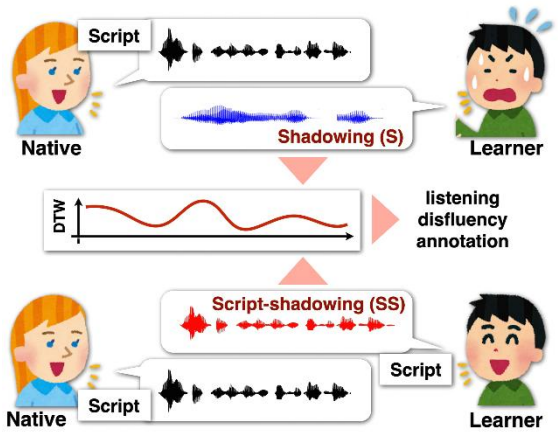

図 1　シャドーイングによる聴取崩れの計測

を時系列として比較すれば、シャドー崩れ（聴取崩れ）が時系列データとして取得できる。図 1 にその様子を示す。脳計測など行わずとも、マイク一本で、かつ教育的に妥当な方法で、頭の中の様子を覗き込む技術として外国語教育の分野では認識されるに至っている。

賢い読者は気づいたと思うが、シャドーイングは「聴取を計測するため」にも「聴取を強化するためにも」使えるタスクである。やればやるほど、聴取の癖を暴き出すことが可能で、また、どんどん聞けるようになる、一粒で二度美味しいタスクである。

4．あなたはどんな世界諸英語が聞き取れないのか？世界の誰があなたの英語発音に苦労するのか？

国際会議に参加する度に「聞き取りに苦労する」英語発音に出会う。あなたはどんな英語が聞き取れる／取れないのか？世界の誰があなたの英語を聞き取り易い／難いと感じるのか？その様子を可視化した。

東大院・工学系研究科は多くの留学生が在籍しているが、本研究科の日本語教室も様々な母語をもつ留学生が日本語を学んでいる。彼らに高校英語レベルの内容の文章（凡そ 30 秒）を読ませ（一人一人が異なる文章を読み上げて録音）、その後、一人が自分を含む全員の音声をシャドーイングし、また、全員がその学習者の英語音声をシャドーする、即ち、互いにシャドーしまくる「相互シャドーイング実験」を行った。この音声を分析すれば、「あなたはどんな世界諸英語が苦手なのか」「世界の誰があなたの英語発音に苦労するのか」が定量化され、可視化できることになる。

外国語音声コミュニケーションにおける聴解崩れの計測とその可視化・モデル化

Measurement, visualization, and modeling of listening disfluency in foreign language oral communication

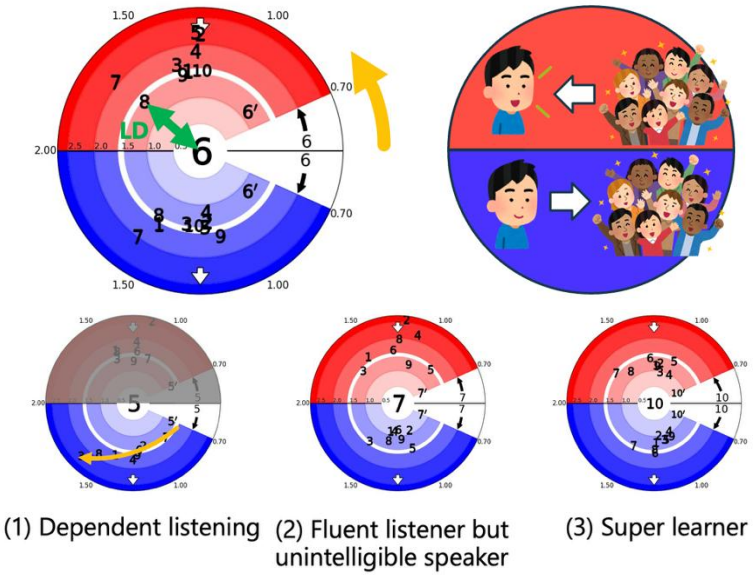

図２　Communicability chart（C-chart）

Communicability chart（C-chart）と命名した可視化手法を図２に示す。中心の番号が対象となる話者 ID であり、上の赤部は他人があなたを聞いた時の聴取困難さ、下の青部はあなたが他人を聞いた時の聴取困難さが、各々半径として表示される。角度はあなたの発音と他者の発音がどのくらい違うのか、発音差異を定量化したものである。

参加した留学生の数だけ C-chart を作成し、それらを分類すると、例えば、1)自分に近い発音は聞き取れるが、発音距離が大きい他者の聞き取りに苦労する学習者、2)様々な英語が聞けるが、他人はその学習者の英語発音に難儀している様子、3)世界中の誰にとっても聞き取り易い発音を有し、また、世界中のどんな英語も聞き取ってしまう super 学習者、など、様々なタイプが存在した。Super 学習者であるが、互いに面識のないハンガリー人２人がそうであった。初めての海外滞在先が日本で、英語圏での滞在経験のない参加者であった。ハンガリーという国の社会音声学的な特徴を考えると、「彼らがなぜそうなのか」について仮説を呈することができるが、ここでは省略する。

５．発音の了解度の自動予測

シャドーイングと言うと、外国語学習者が母語話者の音声を聞きながら即時復唱する行為を思い浮かべる人が多いだろう。しかし、シャドーイングを「聴取者がどこで聴取の乱れを感覚しているのか」を暴き出すためのタスクだと考えた場合、学習者以外にシャドーさせることで新たな技術が誕生する。

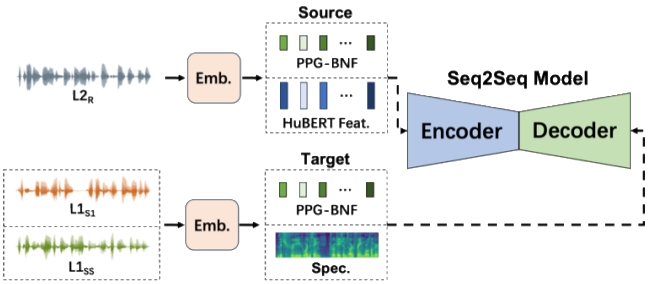

図３　Seq2Seq VC 技術を使ったバーチャルシャドワーの構築

学習者発音を評価する場合の評価尺度の一つとして了解度（intelligibility）がある。「聴取の正確さ」とも言われるが、提示された音声中の単語を、評価者がどれだけ正しく聞き取れたのか、に着眼した尺度である。学習者発音の了解度を求める場合、母語話者（あるいはそれに相当する評価者）が学習者音声を書き取り、正しく書き取ることができた単語の割合が評価点となる。この書き取り作業をシャドーに置き換えれば、評価者が学習者音声を聴取中に、どこで聴取が乱れたのかを計測することとなる。筆者の研究チームでは母語話者（あるいはそれに相当する）評価者に、数百人の学習者音声をシャドーさせ、評価者の聴取がどこでどの程度乱れたのかをコーパスとして構築した。

このコーパスを用いて、学習者音声を入力として、評価者のシャドー音声へと変換することを考える。話者 A の内容を変えずに話者 B の音声とする技術を声質変換（話者変換）と呼ぶが、学習者音声→評価者シャドー音声は（シャドー音声は時々崩れるため）、発話内容も一部変わる声質変換と言える（図３参照）。

伝わるだろうと思った発声が（母語訛りのために）「どうも伝わらない」という体験をした人は少なくないだろう。学習者の音声には、常に母語訛りが混入するが、どういう母語訛りが相手の聴取を乱すのかは、通常、学習者は知る由もない。日本で英語を学ぶ学習者は、日常生活の中で英語を話す機会は少ないとは言え、探せばいくらでもあるだろう。しかしベトナムで（日本に労働者として入国する目的で）日本語を学ぶ

外国語音声コミュニケーションにおける聴解崩れの計測とその可視化・モデル化

Measurement, visualization, and modeling of listening disfluency in foreign language oral communication

図４ Special Training for English Academic Communication (STEAC)の flyer

学習者を考えた場合、ベトナムでの日常生活の中で日本人と会い、日本語を話す機会は殆どない。非国際語を対象として学ぶ場合、その対象語が話される環境で学ぶ以外は（日本で日本語を学ぶ以外は）、教室以外で話す機会はほぼない。これは、学習者自身の発音がどこで、どのくらい相手の聴取を乱しているのかを、体験できないことを意味する。我々が構築した技術は、そのような状況で生きる技術だと考えている。どこで聞き崩れたのかを正直に教えてくれるバーチャルリスナーを構築していることに相当する。

６．将来展望

筆者は工学部で音声・言語技術を教える一方、英語教員として工学部全学生の口頭運用能力向上のための授業を展開している（図４参照）。夏休みや春休みの２ヶ月間、毎日 30 分の（音声技術、AI 技術満載の）オンライン音声課題を課し、聴取力、発音力、そして会話力の強化を図っている。毎日、録音音声を提出させているが、学生には好評である。夏／春休み明けには、教職員にも公開し、キャンパス全体の英語化にも寄与している。卒業直前に参加した学生から「卒業後もお金を払うから継続したい」と言われている。

さて、来年度（2026 年度）から工学系、情報系の２大学院において授業が基本英語化されることが決定している。本研究で培った技術は、学部生の「聴取力」向上を実現する技術として活用されている。

今後であるが、まずは、個々の要素技術の精度向上を今後も継続していく。夏休み／春休みの授業では毎回数百名の学生の英語音声を集めており、それを使って、次回の授業で用いる技術を精緻化している。所謂、human-in-the-loop 型の開発環境を、英語音声教育という文脈において実現している。

また、昨今の大学事情（国際卓越研究大学の発足など）を考えると、各々の分野に内容的に適合した英語音源を揃え、「聴取力」「発音力」「会話力」を鍛え、国際会議でどのような英語話者と遭遇しても、円滑にコミュニケーションを図れる人材を養成するインフラに発展させたい。将来的には、複数の大学、研究機関にて利用可能な、共用インフラを構築したい。

おわりに

SCAT からの支援によって、聴取の様子を教育学的に妥当な方法で計測し、可視化し、モデル化する技術を開発することができた。ここに感謝したい。

参考文献

[1] Kunihara, T., Zhu, C., Minematsu, N., Nakanishi, N. (2022) Gradual Improvements Observed in Learners' Perception and Production of L2 Sounds Through Continuing Shadowing Practices on a Daily Basis. Proc. Interspeech,1303-1307

[2] Tomita, Y., Gao, Y., Minematsu, N., Nakanishi, N., Saito, D. (2024) Analysis and Visualization of Directional Diversity in Listening Fluency of World Englishes Speakers in the Framework of Mutual Shadowing. Proc. Interspeech, 4024-4028

[3] Geng, H., Saito, D., Minematsu, N. (2024) A Pilot Study of Applying Sequence-to-Sequence Voice Conversion to Evaluate the Intelligibility of L2 Speech Using a Native Speaker's Shadowings, Proc. APSIPA, 1-6

この研究は、令和３年度ＳＣＡＴ研究助成の対象として採用され、令和４～６年度に実施されたものです。