日本語学習者発音能力評価システムの開発

高橋 恵利子（目白大学　外国語学部）畑佐 由紀子（広島大学　教育学研究科）
山元 啓史（東京工業大学　社会理工学研究科）
前川 眞一（東京工業大学　社会理工学研究科）　畑佐 一味（パデュー大学　人文学部）

本研究は日本語学習者の発音の自動評価システムの開発を目的としている．そのための基礎調査として，中国人日本語学習者の音声データと，それに対する母語話者の一対比較評価データから，課題文及び評価方法の妥当性について検討した． 評価者の属性に関わらず母語話者の評価はほぼ一致していたことから，一対比較による評価方法を用いれば，評価者の属性に関わらず，妥当な評価値が得られる可能性が指摘できる．今後，さらに評価対象とする音声データを増やして今回の結果を検証する必要がある．また，一対比較による膨大な評価作業における評価者の負担を軽減するため，一般母語話者を対象としたクラウドソーシングを採用することの意義と課題について言及する．

The development of an automatic evaluation system of L2 pronunciation in Japanese

Eriko Takahashi (Faculty of Foreign Language Studies, Mejiro University) /
Yukiko Hatasa (Graduate School of Education, Hiroshima University) /
Hilofumi Yamamoto (Graduate School of Decision Science and Technology, Tokyo Institute of Technology) /
Shin-ichi Maekawa (Graduate School of Decision Science and Technology, Tokyo Institute of Technology) /
Kazumi Hatasa (College of Liberal Arts, Purdue University)

The present project aims at developing an automatic evaluation system of non-native speakers' pronunciation in Japanese. In order to examine the quality of stimulus sentences and feasibility of the paired-comparison method to be used in our main study, a group of Japanese native speakers were asked to judge pairs of sentences read by Chinse leaners of Japanese. The results showed that native speakers' judgements were highly consistent, suggesting that the paired-comparison method can be used to establish an evaluation measure for a computer-based assessment. We plan to expand the data size by increasing the number of learners and by employing a crowdsourcing technique to obtain paired-comparison data from a large pool of native speakers.

1．まえがき

音声は，母語話者が非母語話者の日本語能力を推定する際の，もっとも簡便な材料である．文法的に正確であっても，発音が悪いと母語話者に正しく理解されなかったり，能力が低いと評価されることがある（土岐 1994：78）．グローバル人材交流が拡大する中で，日本語学習者にとって重要なのは，単に通じるだけでなく，より適切で好ましい話し方をすることである．しかし，自分の発音が一般母語話者にどう評価されるか，何をどう改善すればよいのかといった情報を提供できる自動評価システムは未だ存在しない．

従来の発音評価研究では，評価者が学習者の発話音声を聞き，項目ごと，あるいは全体的な印象について，4〜7 段階の尺度で評価する手法（芝他 1984：105-6）が取られているが，この方法では，同じ音声であっても評価者の属性や文脈によって評価値が異なることが指摘されている（小池 1998：151，小河原 1993：11，河野・松崎 1998：25，渡辺・松崎 2014：65，松崎 2007：303）．一方で，評価者が一般母語話者であるか音声の専門家であるかに関わらず，ほぼ同様の評価値になるという報告もあり（Schmid and Hopp 2014: 377; Warren et al. 2009: 94 ），評価者の訓練なしに尺度法を用いた場合の信頼性については意見が分かれている．また，この方法では尺度や段

階の解釈が，評価者間で一致しないことや，中心化傾向（中央を選ぶ傾向），寛大化傾向（何でもかんでもよい方を選ぶ傾向），ハロー効果（対象者に期待をするあまりそれに影響をうけてポジティブにもネガティブにも歪められる傾向）などが起こりやすい欠点が指摘されている（東他 1973：444-5）．

この解決方法として，音響分析を基にして自動的に発音を評価する手法も考えられるが，この手法は評価者の影響を受けないものの，音響分析から算出された音声の逸脱は人間が感じる逸脱と必ずしも一致しないという問題がある．学習者にとっては，評価者が誰であれ，一般的にどう評価されるのかということが問題であり，より統合的な評価を還元する必要がある．属性によって評価者を限定する必要がなくなれば，多くの評価者により，より多くの学習者を評価することが可能となる．本発表では，まず評価システム開発の事前調査として，評価者の均質性を検討することを目的とし，一対比較法（東他（1973）：445，507）を評価法として採用する．

一対比較法は，呈示される 2 刺激からより好ましい方を選択させる評価方法である．この手法は，類似度の高い刺激でも，その差異を詳しく評価分析することが可能なため，信頼性・妥当性が高い．その一方で，一対比較法は，刺激数が多いと作業時間がかかるというデメリットがある．

本研究では，まず，評価すべき音声を総当たりで組み合わせ，それらを一対ごとの比較で評価を試み，評価者間の評価傾向に違いが見られるかを検討する．そのうえで，クラウドソーシングを取り入れることにより個々の評価者の作業負担を軽減し，人の集合知を利用した評価システムを開発する可能性について検討する．

2．方法

2.1　材料

発音課題用の短文として，比較的自然な状態で録音された材料（オーセンティックな材料）を採用することとした．テレビのインタビュー番組，ドキュメンタリーや，実際の会話から，長すぎないこと，意味理解が容易であること，音声が明瞭であることなどを条件に，10 の短文とその音声を採用した（表 1 参照）．

表 1　課題に用いられた短文リスト
SID はセンテンス ID

SID	課題文
S01	ここが玄関です.
S02	亡くなったおばあちゃんの写真です.
S03	大盛りです.
S04	冷凍食品はアメリカの方が多いですね.
S05	１２００円になります.
S06	これは冷蔵庫です.
S07	息子さんはおいくつですか.
S08	これが一番使いやすいですね.
S09	１万円お預かりいたします.
S10	中は何になりますか.

2.2　音声提供者

音声提供者（いわゆる発音を評価される「学生」に当たる）は 2 名の母語話者と 4 名の中国人日本語学習者である．尺度を網羅したデータを得るために，最高の技能を持つ者として母語話者（NS1, NS2），ほぼネイティブ水準の発音技能を持つ者（CH1,CH2），顕著な外国人訛りがある者（CM1,CM2）を採用した．

2.3　手続き

課題は読み上げ課題とリピート課題の 2 種類を用意した．音声提供者には，コンピュータ画面に呈示される文を読み上げることを求めた．続いて，聴覚提示される文を即座にリピート再生することも求めた．これは話速と文末イントネーションの理解を統制することを目的としている．読み上げ音声とリピート音声は 1 文ずつ交互に採取した．なお，課題文をパワーポイント（プレゼンテーション用呈示ソフトウェア，PPT）で呈示する際には，漢字には読み仮名をつけ，さらに中国語訳も添えて呈示した（図 1 参照）．各音声提供者から読み上げ課題，リピート課題の各 10 個の音声資料を得た．

評価用刺激として，1 文ごとに 6 名の音声を総当たりでペアにして，読み上げ課題の音声，リピート課題の音声，各 150 個（10×6×5／2）の音声刺激ファイルを作成した．全 300 個の刺激をランダムに配置し，5 名の評価者（教師 1 名，大学院生 2 名，学部生 2 名）にランダムで聴覚呈示した．評価者に音声学を専攻する者は含まれていない．また教師と大学院生は留学生との接触頻度が高いが，学部生は留学生と接する機会はほとんどない．評価者は任意 2 名の音声提供者による録音音声を聞き，直感的により自然だと思った方をキー入力で回答することが求められた．評価の所要時間は 1 人あたり 40 分～50 分程度であった．なお評価者の母方言は問わないこととした．

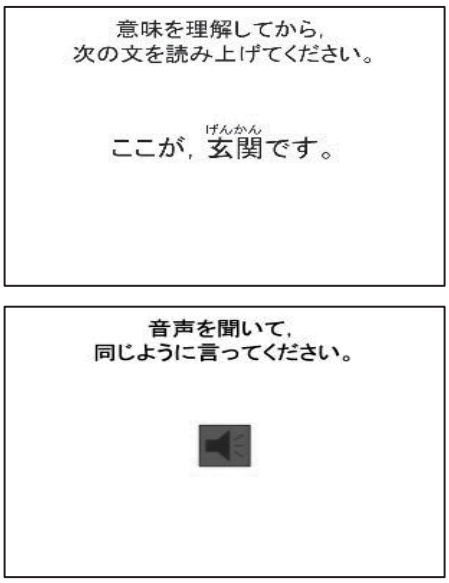

図 1　コンピュータによる読み上げ課題（左）とリピート課題（右）の教示画面

3．結果

一対比較のための音声を 5 名の評価者が評価し，その評価データを多次元尺度構成法の一つである MDPREF（Multidimensional Preference Scaling）の最尤解（Okubo and Mayekawa 2015）で計算した[1]．以下ではその結果を，評価者と課題文の観点から述べる．

3.1　評価者

評価者（E00001~E00005）と音声提供者（NS1,NS2,CH1,CH2,CM1,CM2）の関係を biplot で表したのが，図 2-a（読み上げ課題）図 2-b（リピート課題）である．

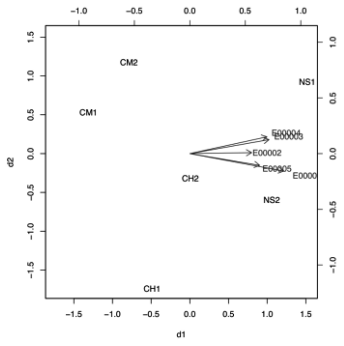

図 2-a 読み上げ課題における評価者 5 名と音声提供者 6 名の biplot

[1] MDPREF 用の R パッケージは，前川研究室の R パッケージのページから利用できる．
(http://www.ms.hum.titech.ac.jp/Rpackages.html)

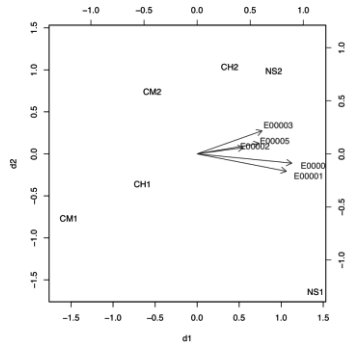

図 2-b　リピート課題における評価者 5 名と音声提供者 6 名の biplot

読み上げ課題については，全ての評価者同士の相関は 0.922 以上であった．リピート課題についても，E00001 と E00003 の評価者間の相関が 0.861 である他は，全て 0.936 以上の相関があった．MDPREF の d1 軸の総体的な寄与は，読み上げ課題で 0.970，リピート課題で 0.962 であり，いずれにおいても一元的な尺度と見なしてよい数値であった．

3.2　課題文

課題文（S01～S10）と音声提供者（NS1,NS2,CH1,CH2,CM1,CM2）の関係を biplot で表したのが，図 3-a（読み上げ課題），図 3-b（リピート課題）である．

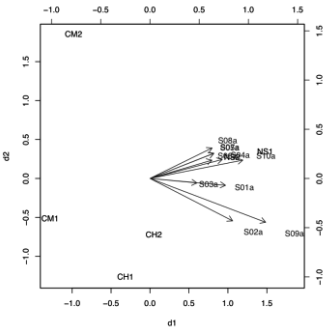

図 3-a　読み上げ課題における課題文 10 と音声提供者 6 名の biplot

読み上げ課題，リピート課題のいずれにおいても，センテンスごとの課題のベクトルは右向きを示しており，MDPREF の d1 軸の総体的な寄与は，読み上げ課題 0.892，リピート課題 0.917 であった．ただし，読み上げ課題とリピート課題では，CH1 の布置が大きく変わり，読み上げ課題では，NS1-NS2-CH2-CH1-CM2-CM1 の順に，リピート課題では NS1-NS2-CH2-CM2-CH1-CM1 の順になった．

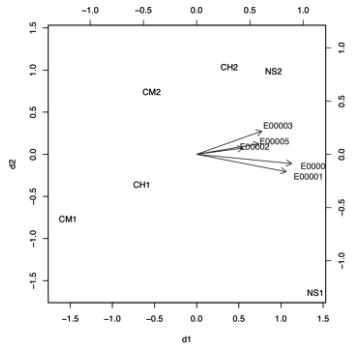

図 3-b　リピート課題における課題文 10 と音声提供者 6 名の biplot

読み上げ課題で顕著なベクトルは課題文 S09a である. S09a に類似する課題文は S02a であり，相関係数 0.994 であったが，発音特徴上の共通点は両者には見られない．S09a に直交する軸で音声提供者の布置を見ると，NS1-NS2-CH2-CH1-CM2-CM1 となった．これは，ネイティブ（NS）→ネイティブ水準（CH）→外国人訛りあり（CM）の順序と一致し，評定として妥当である．

リピート課題で顕著な課題文は S05b であり，それと最も相関が高いのは S06b（0.993）であった．逆に，S05b と最も相関が低いベクトルは S04b（0.344）であり，むしろ，S05b よりも S04b の方が他の課題文との相関は高かった（例えば，S04b と S09b は 0.959，S01b は 0.955，S02b は 0.976 など）．

課題文 S05b と S04b で直交する場合の音声提供者の布置を見てみると，S05b と直交する軸では，NS2-NS1-CM2-CH2-CM1-CH1 の順となった．CM2 がネイティブ（NS）に迫る高評価である一方，CH1 が外国人訛りあり（CM）よりもずっと低い評価となっている．S04b と直交する軸では，NS1-NS2-CH1-CH2-CM2-CM1 の順となり，ネイティブ（NS）→ネイティブ水準（CH）→外国人訛りあり（CM）の順序と一致する．

4．考察

リピート課題では，課題文によって発音評価順位に変動があり，ニアネイティブ水準の CH1 の評価が低くなる現象が見られた．これは，課題の性質，つまり実際の発話に基づいたリピートを求めたことによると思われる．リピート課題では，「音声を聞いて同じように言ってください」という教示を与えたが，CH1 は教示に従い，声色を含めた忠実な模倣を試みていた．実際の会話やインタビューから切り出したモデル音声は，発話者の癖やその場の発話意図も含むものであった．しかし母語話者が評価する際はモデルとなった音源は呈示されないため，そうした音声的特徴は余剰情報となり，CH1 の忠実な模倣が却って奇異な印象を与えた可能性が考えられる．

しかし全体的には，読み上げ課題とリピート課題の間のそれぞれの評価において，大きな差は見られなかった．したがって，読み上げ課題によって得られた音声だけでも評価が可能であると考えられる．

5 名の評価者に関しては，年齢も留学生との接触頻度も母方言も異なっていたが，評価者間に高い相関が認められた．これにより，日本語母語話者であれば，留学生との接触頻度に関わらず，ほぼ同じような結果が得られることも分かった．これは，母語話者には共通する一定の評価原理が存在することを示唆しており，二者択一を迫る一対比較法では，この原理に基づいた評価の差が現れやすいと考えられる．

なお本研究では，評価者による誤用箇所の特定や，判断理由に関するインタビューなどは行っていない．専門家は発音の誤用箇所を具体的に指摘することができるが，一般の母語話者は，発音の巧拙の印象評価はできても，誤用箇所やタイプを特定し説明することが困難で，「なんとなく変だった」「不自然」といったあいまいなコメントに留まることが多い（松崎 2007 : 303）．さらに複合的な誤用に関しては判断や表現が一層困難になる．本研究はむしろ，母語話者の直感による印象評価を重視するため，評価の理由や，発音上の問題点に関する情報提供を求めなかった．しかし，今後，学習者へのフィードバックや，評価に関わる音声的特徴を洗い出すためには，発音の誤用分析は不可欠であると考える．本調査では，5 名の評価が分かれる刺激文があったことから，概ねの評価は一致するものの，優先評価項目が評価者間で部分的に異なる可能性が考えられる．それが誤差の範囲に留まるものであるか否か，今回のデータからは判断できない．今後，さらに音声提供者および評価者データを蓄積し，評価に影響を与える発音上の特徴について分析を進めるとともに，母語話者の主観的評価原理を明らかにする必要がある．

5．今後の課題

5.1　クラウドソーシングの活用

今後，規模を拡充してさらにデータを収集するために，集合知の活用，および評価作業の負担軽減という 2 点からクラウドソーシング（crowdsourcing）の活用を検討している．

クラウドソーシングは，専門家・非専門家を問わず不特定多数の人々に作業への介入を促し，課

題を処理するシステムである．作業をネット上で公開・共有することで，多種多様な人々の集合知を集積することができる（永田 2014：475，永田 2013：373）．通常の教室指導でフィードバックされるのは，一人の教師による評価であるが，本研究の最終的な目的は，学習者の発音に対する一般母語話者の評価をフィードバックすることである．それは不特定多数の母語話者の集合知によって得られるものであり，クラウドソーシングの原理にかなうものである．

また，不特定多数の人物による分担作業という側面も本研究にとって重要である．一対比較法は，本研究の実験でも確認されたように，厳格で安定した評価値が得られる点で優れている．しかし，総当たりペアの比較が条件であるため，課題文と評価すべき学習者の数が増えると，比較すべきペアの数が激増してしまう．今回は，評価者の負担を考慮して課題文の数を 10 に限定したが，評価システムには，母語の音声的特徴などを考慮に入れながら，発音誤用を網羅的に抽出できる課題文セットが必要となる．そこで次のステップとして，英語母語話者に特徴的な発音の問題を含む課題文 40 文を 40 名の英語母語話者に音読させた音声データを構築した．この膨大なデータの一対比較を一人の人間が行うことは，評価作業の負担を考えると事実上不可能である．

この問題の解決策としてクラウドソーシングの活用が考えられる．従来，発音の評価といった作業は，音声学や言語学の専門家を必要としていたため，大量のデータ処理が困難であった．しかし，評価者の均質性が保証できれば，大量の音声データを複数の評価者で分担評価し，結果を統合することが可能となる．今回の一対比較の調査結果から，誰が評価しても安定した評価が得られることがひとまず確認できた．これにより，日本語母語話者であれば誰でも妥当な評価が可能であると仮定できる．

なおクラウドソーシングの構築に当たっては，上述のような目的から，評価者属性（年齢，性別，出身地，職業など）を確認し，評価用音声ファイルを割り当てることとする．

5.2 システム構築

クラウドソーシングを実施するに先立って，まずは評価刺激の作成を行う．収集した学習者の課題文読み上げデータをサーバー上に保管し，2 刺激 1 対の評価用音声ファイルを自動作成する．評価協力者には，サーバー上に保管された評価用音声ファイルへのアクセスを求める．このとき，評価協力者に音声ファイルを 20～30 程度呈示し，それぞれを一対比較で評価させ，結果をサーバー上で回収する．評価者各自が端末からサーバーにアクセスできれば，場所や時間に関わらず，評価作業を実施できる．（図 4 参照）．

図 4　クラウドソーシングイメージ図

大量の分担作業が実現するまでは，本発表のような限定された条件で，読み上げ音声データを組み合わせ，評価者に一対評価をしてもらう．この評価データについて，MDPREF による評価値の比較分析を行い（Okubo & Mayekawa 2015），学習者の音声に対する母語話者の評価に影響を与える音声的特徴とその関連性を分析し，評価を下げる音声的特徴を抽出する．これを基に，既存の音声データに新しい音声を追加しながら音声刺激を組み合わせる．そして，そこで得られた結果を更新する形で，極限法（市川（1991）：158）の一種である上下法（up-and-down method）を組み合わせるなどして，評価結果の自明なペアを自動的に判断し省略する仕組みを構築する．これにより，評価対象刺激数を抑えることが可能となる．

今回は 6 名の音声データに対する 5 名の評価者データを扱い，評価が妥当なものであることを確認した．しかし，評価対象はネイティブ，ネイティブ水準，外国人訛りありという，違いの分かりやすい音声であったため，評価が一致しやすかったという側面は否定できない．今後はその中間にあるような音声も含めた上で，網羅的に扱っていく必要がある．また評価者に関しても，年代や職業，母方言など幅を広げながらデータを増やし評価者属性に偏りが生じないことを検証していく．

参考文献リスト

東洋・大山正・詫摩武俊・藤永保（編）：心理用語の基礎知識：概念の正確な理解と整理，有斐閣（1973）.

市川伸一：心理測定法への招待：測定からみた心理学入門，サイエンス社（1991）.

小河原義朗：外国人の日本語の発音に対する日本人の評価，東北大学文学部日本語学科論集，No.3，pp.1-12（1993）

小池真理：学習者の会話能力に対する評価に見られる日本語教師と一般日本人のずれ：初級学習者の到達度試験のロールプレイに対する評価，北海道大学留学生センター紀要，No.2，pp.138-156（1998）.

河野俊之・松崎寛：一般日本人と日本語教師の音声評価の差異，日本語教育方法研究会誌，Vol.5，No.2，pp.24-25（1998）.

芝祐順・渡部洋・石塚智一：統計用語辞典，新曜社（1984）.

土岐哲：聞き手の国際化，日本語学，Vol.13，pp.74-80（1994）.

永崎研宣：日本語クラウドソーシング翻刻に向けて：特集デジタル時代の日本語，情報の科学と技術，Vol.64，No.11，pp.475-480（2014）.

永崎研宣：人文学分野とサイバーインフラストラクチャー：デジタル・ヒューマニティーズにおける現状と課題，情報の科学と技術，Vol.63，No.9，pp.369-376（2013）.

松崎寛：発音評価研究に関する覚書，大学における日本語教育の構築と展開，ひつじ書房，pp.297-309（2007）.

渡辺裕美・松崎寛：発音評価の相違：日本人教師・ロシア人教師・一般日本人の比較，日本語教育，No.159，pp.61-75（2014）.

Okubo, Tomoya and Shin-Ichi Mayekawa: Modeling viewpoint shifts in probalistic choice, *Psychometrika*, Vol.80, No.2, pp.412-427(2015).

Schmid, Monika. S. and Holger Hopp: Comparing foreign accent in L1 attrition and L2 acquision: Range and rater effects, *Language Testing*, Vol.31, pp.367-388(2014).

Warren, Paul, Irina Elgort, and David Crabbe: Comprehensibility and prosody ratings for pronunciation software development, *Language Learning and Technology*, Vol.13, pp.87-102(1984).