

Research Article

DOI: <http://dx.doi.org/10.22192/ijamr.2025.12.12.006>

Operationalizing Speech Science in L2 English Pronunciation Instruction: Bridging the Technical Barrier for Research-Based Pedagogy in the Japanese EFL Context

Yoshimasa Uehara

Oklahoma City University, Oklahoma, USA

E-mail: uehara-yoshimasa-yu@alumni.osaka-u.ac.jp

Abstract

This study addresses the persistent gap between the theoretical necessity for research-based Second Language (L2) pronunciation instruction and the practical complexity of acoustic analysis. While the differing pronunciations of English native speakers (NE) and non-native speakers (NNE), particularly Japanese learners, have been extensively discussed, technical investigations visualizing these differences remain scarce. Therefore, this research compares the speech of an English native speaker and a non-native English speaker using spectral analysis and formant analysis to objectively clarify these acoustic deviations.

The findings confirm that measurable acoustic differences exist but highlight a significant methodological barrier: the high level of expertise required for current analytical tools. The ultimate methodological contribution proposed is the development of accessible systems capable of performing "simple spectrum analysis and simple formant analysis," thereby operationalizing rigorous acoustic findings for widespread application in English education.

Keywords

Speech Sciences,
Acoustic Phonetics,
Visualized Speech
Taring in English,
STEM Technique

Introduction

The study addresses the persistent gap between the theoretical necessity of research-based Second Language (L2) pronunciation instruction and the

practical complexity of acoustic analysis. Formant analysis is currently a key methodological trend in acoustic phonetics, and fundamental acoustic analysis is crucial for visualizing acoustic phonetics.

The core objective is supported by the foundational demand to replace subjective reliance, such as "individual teachers' intuition or feel," with objective, visually substantiated acoustic data. In addition, this study aims to prevent negative speech transfer from the teachers to students, and the suggestion provides evidence-based speech science instead of draining CD's sound in English at English language classes.

Literature Review

Acoustic analysis techniques are essential for rigorously quantifying speech properties, utilizing formant analysis to examine critical acoustic properties including formant frequency values (F1-F4). Research emphasizes that measurable acoustic improvement requires Form-Focused Instruction (FFI) with explicit Corrective Feedback (CF).

Challenges for Japanese learners of English are anchored in the Speech Learning Model (SLM), which hypothesizes that learners perceptually assimilate L2 sounds into their established L1 phonological space. This L1 transfer manifests as predictable articulatory deviations, particularly struggling with distinguishing the spectral quality of English low and mid vowels (e.g., /æ/, /ʌ/, /ə/, /ɔ/, and /ɑ/).

A robust body of literature highlights the necessity of shifting Second Language (L2) accent and pronunciation teaching toward a research-based approach, moving away from subjective methods that rely on "individual teachers' intuition or feel". Acoustic analysis techniques are essential for rigorously measuring and quantifying speech properties. Formant analysis has emerged as a key methodological trend in acoustic phonetics, providing a reliable means to examine critical acoustic properties (Grabowski, 2023), including formant frequency values (F1-F4), intensity, and duration, across L2 speech samples. Visual feedback, which involves comparing an L2 speaker's production to a native speaker's "target," enhances the learner's ability to notice acoustic differences. Crucially, research

by Saito and Lyster (2012) emphasizes that measurable acoustic improvement requires Form-Focused Instruction (FFI) with explicit Corrective Feedback (CF).

Derwing and Munro (2005) advocate for a fundamental shift toward a research-based approach to second language accent and pronunciation teaching. Their work establishes the theoretical necessity for instruction to be grounded in verifiable linguistic and pedagogical principles, moving away from intuitive or unproven methods. This directly supports the core objective of the present study: to replace the reliance on "individual teachers' intuition or feel" with objective, visually substantiated acoustic data.

The work implemented by Kartushina et al. (2015) examined the effect of phonetic production training with visual feedback on both perception and production of foreign speech sounds. Their research, and the broader literature they reference, suggests that while visual feedback is crucial for enhancing noticing, the effects of training are often modality-specific, meaning that production training may not significantly transfer to perception.

Purpose

The research aims to objectively clarify acoustic deviations by comparing the speech of an English native speaker (NE) and a non-native English speaker (NNE) using spectral analysis and formant analysis. By conducting a foundational acoustic comparison, the study seeks to bridge the substantial methodological barrier that hinders the consistent application of research-based acoustic feedback.

Method

This research utilized Wavesurfer, a recognized Speech Acoustic Analysis Software Package (SAASP), to implement both spectral analysis and formant analysis. However, the current analytical tools require a high level of expertise in

disciplines including mathematics, sound engineering, programming, and statistics, making the process technically complex. Spectrum analysis mathematically involves converting curves into lines using Fourier transforms and applying Linear Predictive Coding (LPC).

The methodology requires advanced knowledge of core disciplines, including mathematics, sound engineering, programming, and statistics.

Fourier Transform and Frequency Metrics:

Mathematically, spectrum analysis involves converting curves into lines using Fourier transforms and applying Linear Predictive Coding (LPC). The resulting acoustic metrics derived from this process are inherently quantitative. Key frequency metrics measured include fundamental frequency (F0) and formant frequency values (F1-F4). F1 is used to track vowel height, and F2 is used to track tongue fronting/backing.

Statistical Metrics and Design:

The methodology relied on quantitative analysis using statistical principles.

1. Quantitative Distance Metric: The distinction between the approximants /l/ and /r/ was quantified using the F3-F2 distance, measured in Bark. This metric provides a specific quantitative cue for articulatory differentiation.

2. Experimental Validation: The efficacy of pedagogical application was validated using pretest-posttest experiments.

Ethics

Continued development of automated scoring and visual interfaces must adhere to stringent ethical safeguards, ensuring that technological solutions do not simply institutionalize existing acoustic biases.

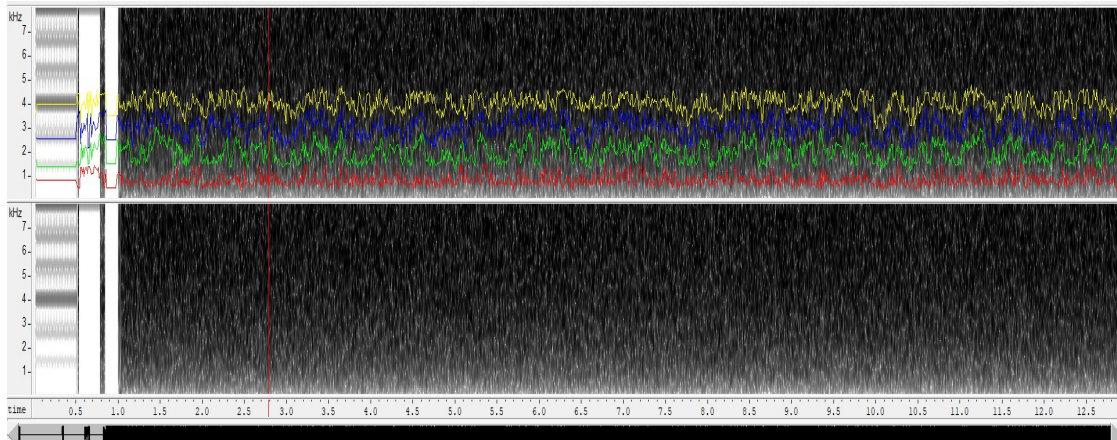
Analysis

The analysis focused on measuring quantitative acoustic metrics, including fundamental frequency (F0) and formant frequency values (F1-F4). Formant analysis tracks vowel height using F1 and tongue fronting/backing using F2. The distinction between the approximant's /l/ and /r/ was critically examined using the F3 value and the F3-F2 distance, measured in Bark.

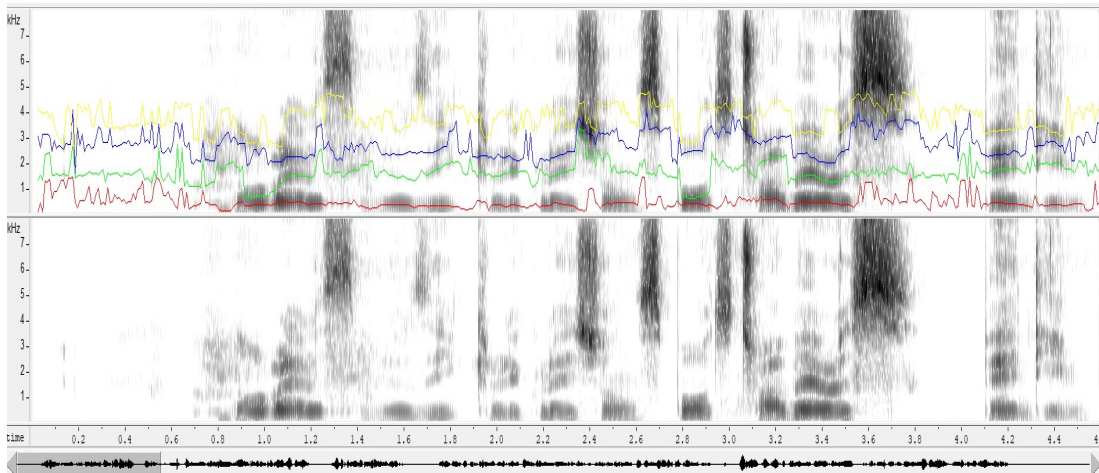
Results

The findings confirm that measurable acoustic differences exist between NE and NNE speech. The NE speaker's formants were stable, while the Japanese NNE speaker's formants exhibited big amplitudes. Acoustic research confirmed systematic formant space deviation in L2 production. Specifically, the L2 production of vowel /ʌ/ showed a significantly Higher F1, and vowel /ɑ/ showed a significantly Higher F2. The liquid /r/ production was characterized by an F3 that is consistently higher than NE values, which is the source of perceived accent. Furthermore, Japanese adults showed an overlapping, small F3-F2 distance (smaller than 6.0 Bark) for both /l/ and /r/, demonstrating insufficient articulatory differentiation.

English Native Speaker's spectrum and formant analysis



Japanese English Speaker's spectrum and formant analysis



Red Line=F1, Green Line=F2, Blue Line=F3, Yellow Line=F4

Discussion

While measurable acoustic differences were confirmed, the analysis highlighted a significant methodological barrier: the high level of expertise required for current analytical tools. This technical complexity prevents most classroom educators from applying critical acoustic feedback. Experimental evidence confirmed that only Japanese learners who received FFI with explicit CF achieved measurable acoustic improvement, validating the quantitative approach to pedagogical efficacy.

Conclusion

The analysis confirms that L1 transfer effects result in specific, predictable, and measurable

articulatory errors (e.g., high F1 for /ʌ/, high F3 for /r/) that visualization clarifies. The reliable pathway to move instruction beyond subjective judgments is the visualization of these acoustic deviations. The ultimate necessary methodological contribution is the development of robust, accessible systems capable of performing "simple spectrum analysis and simple formant analysis".

Implication

The operationalization via simplified systems enables Data-Driven Instruction based on objective data and ensures Effective Resource Allocation by prioritizing feedback based on empirically categorized difficulty, such as

flagging specific acoustically validated errors like high F3 values for /r/. Successful operationalization is evident in advancements in Japanese Computer-Assisted Pronunciation Training (CAPT) systems.

The system structure also allows for Scalability through the integration of Artificial Intelligence (AI) and Machine Learning (ML),

The following table demonstrates the diagnostic feedback provided by an operationalized acoustic visualization system, simulating the application of "simple spectrum analysis and simple formant analysis". This system is designed to provide **objective, visually substantiated acoustic data**, to identify the **measurable acoustic differences** and **predictable articulatory errors**, of Non-Native English (NNE) speakers.

Diagnostic Feedback Example from the Acoustic Visualization System

Acoustic Feature Analyzed	Native English (NE) Target/Cue	Non-Native English (NNE) Measured Deviation	Diagnostic Interpretation & Instructional Prioritization
Spectrogram and Formant Waveform,	Formants are stable.	Formants exhibit big amplitudes; spectrogram is strong.	Visualization clarifies measurable acoustic differences not readily apparent when listening. This supports Data-Driven Instruction, replacing subjective reliance with objective data.
Formant F1 (Vowel Height),	Lower F1 for /ʌ/ (mid-vowel height).	Vowel /ʌ/ showed a significantly Higher F1,	Identifies specific, predictable, and measurable articulatory errors resulting from L1 transfer effects,
Formant F2 (Tongue Fronting/Backing),	Lower F2 for /ɑ/ (back-vowel tongue position).	Vowel /ɑ/ showed a significantly Higher F2.	Used for automated instructional prioritization of difficult items, such as high F2 values for /ɑ/.
Formant F3 (Liquid /r/ Cue),	Critically Low F3.	F3 is consistently higher than NE values.	F3 height is recognized as the acoustic source of perceived accent. The system must flag high F3 values for /r/, which requires Form-Focused Instruction (FFI) with explicit Corrective Feedback (CF) for measurable improvement.
F3-F2 Distance (Bark)	Typically, greater than 4.0 Bark for /l/ and smaller than 4.0 Bark for /r/.	Overlapping, small F3-F2 distance, measuring smaller than 6.0 Bark for both /l/ and /r/.	Demonstrates insufficient articulatory differentiation. This quantitative data enables Effective Resource Allocation by prioritizing feedback based on empirically categorized difficulty.

In addition, Automated Assessment Models is required to construct the visualized application. For objective judgment in practical application (e.g., CAPT systems), the methodology integrates quantitative assessment based on statistical models, such as Hidden Markov Model (HMM) likelihood scores. The integration of the Hidden Markov Model (HMM) into the methodology is central to the operationalization of acoustic phonetics in pedagogical tools.

The HMM is utilized as a statistical model within advanced Japanese Computer-Assisted Pronunciation Training (CAPT) systems. The primary purpose of integrating HMM is to facilitate Objective Judgment. By providing a statistical foundation for assessment, the HMM framework supports the study's core objective of replacing subjective reliance on "individual teachers' intuition or feel" with objective, visually substantiated acoustic data. The broader methodological development aims to transform complex, established acoustic techniques, which require advanced knowledge in disciplines including statistics, into accessible and reproducible methods below.

The classic speech recognition pipeline is integrated into the following four structural layers:

- 1. Signal Processing Layer (Time/Frequency/Cepstral Domains):** This involves the digitization of the continuous signal, short-time framing, windowing, and conversion into the frequency domain via the Fast Fourier Transform (FFT).
- 2. Feature Extraction Layer (MFCC, Formants):** Based on the FFT results, this layer performs spectrogram analysis, formant extraction using Linear Predictive Coding (LPC), and perceptually optimized processing (MFCC derivation) using the Mel scale. The result of this process serves as the Observation Vector for the HMM.

- 3. Statistical Modeling Layer (HMM States/Transitions):** HMM parameters are learned from the observed data, modeling the states of phonemes or sub-phonemes and their probabilistic transitions.

- 4. Decoding Layer (Viterbi Path):** The Viterbi algorithm is utilized to efficiently search and restore the most probable sequence of hidden states, i.e. the recognized phoneme sequence, from the observation sequence.

The performance of HMM-based systems relies heavily not only on the statistical sophistication of the acoustic modeling but also on the quality of the observation features generated by the signal processing. It is critically important to use features that are perceptually optimized based on acoustic linguistics, such as MFCCs and formants, rather than raw spectral information.

Specifically, MFCCs achieve high robustness against signal variations and noise by applying a non-linear transformation (the Mel scale) that closely approximates human auditory characteristics. This enhancement improves the accuracy of the HMM's Observation Probability, thereby maintaining a high overall recognition rate for the system.

Consequently, the performance improvement of HMMs is achieved through the synergy of engineering optimization in signal processing and the mathematical framework of statistical modeling.

Classic HMM-based systems achieved significant success in early speech recognition, and their computational efficiency and high interpretability are still valued. However, HMMs have several fundamental limitations.

One major limitation is the HMM's reliance on the Markov assumption (first-order independence of state transitions) and the assumption of frame-to-frame independence of observation

probabilities. Actual speech exhibits strong correlations between adjacent frames and long-term dependencies across the entire utterance, which the HMM cannot directly model.

Furthermore, the expressive power of the Gaussian Mixture Models (GMMs) used to model the Observation Probability was limited. To overcome these constraints and capture more complex acoustic phenomena and contextual dependencies, speech recognition technology transitioned into the era of deep learning.

This evolution specifically involved the development of DNN-HMM hybrid models, where the GMM component of the HMM was replaced by a Deep Neural Network (DNN), and further progressed to End-to-End (E2E) models that integrate the role of the HMM into the DNN, often utilizing Connectionist Temporal Classification (CTC) to maintain label alignment.

Future Perspectives

Although HMM-based systems have yielded their position to high-performance deep learning models, the fundamental techniques of signal processing continue to be the backbone of speech recognition. The efficient frequency transformation provided by the FFT and the use of the Mel scale, which mimics human auditory characteristics, remain core components of acoustic input pre-processing even in modern, high-performance E2E systems utilizing Transformers and RNNs.

A crucial takeaway from the classic pipeline is that the philosophy of reducing redundancy and abstracting information in a manner optimal for human perception (MFCC derivation) when processing complex signals holds enduring value across technological evolution. Fundamental theories of acoustic phonetics and insights from statistical modeling will continue to provide

important guidance for designing new architectures and pursuing research focused on enhancing robustness.

Acknowledgement

My sincere thanks go to Gina, G. (Singer and Guitar of Dirt Fisherman, Idaho, USA) and Associate Professor, Yusaku, M. to help with recording for this study.

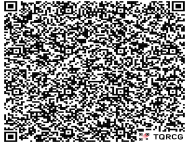
Script used in this study:

Novels, music, recreation, sports, travels, painting, performing arts and so on are included in the category of life-long learning. These activities are considered as “soft.” Examples are participation in open lectures at various schools and in various organizations, internet communication, and pre- and in-service job training etc. Also, volunteer activities and NPO/NGO activities are contained in life-long learning in order to make a learning society.

References

- Derwing, T. M., & Munro, M. J. (2005). Secondlanguage accent and pronunciation teaching: A research-based approach. *TESOL Quarterly*, 39, 379–397.
- Grabowski, E. 2023 Methodological trends in acoustic phonetic analysis, *ICPhS2023*, pp.833-837
https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2023/full_papers/984.pdf
- Kartushina, N., Hervais-Adelman, A., Frauenfelder, U. H., & Golestani, N. (2015). The effect of phonetic production training with visual feedback on the perception and production of foreign speech sounds. *The Journal of the Acoustical Society of America*, 138(2), 817–832.

Saito, K., & Lyster, R. (2012). Effects of form-focused instruction and corrective feedback on L2 pronunciation development of /r/ by Japanese learners of English. *Language Learning*, 62(2), 595–633.

Access this Article in Online	
	Website: www.ijarm.com
	Subject: Pedagogy
Quick Response Code	
DOI: 10.22192/ijamr.2025.12.12.006	

How to cite this article:

Yoshimasa Uehara. (2025). Operationalizing Speech Science in L2 English Pronunciation Instruction: Bridging the Technical Barrier for Research-Based Pedagogy in the Japanese EFL Context. *Int. J. Adv. Multidiscip. Res.* 12(12): 49-56.

DOI: <http://dx.doi.org/10.22192/ijamr.2025.12.12.006>