Leveraging a deep learning neural network for classifying native and nonnative speakers based on acoustic features

Georgios P. Georgiou

University of Nicosia

georgiou.georg@unic.ac.cy

Abstract

The evolution of Artificial Intelligence (AI) has led to the development of sophisticated machine learning algorithms capable of tackling complex classification tasks, including distinguishing speakers based on linguistic features. This study evaluates the effectiveness of a deep learning algorithm in differentiating first language (L1) and second language (L2) speakers using specific acoustic features. The algorithm was trained on formant frequencies (F1, F2, F3) and vowel duration extracted from speech samples of adult native English speakers and Cypriot Greek speakers of English as an L2. The model was rigorously tested using k-fold cross-validation and optimized through a grid search over various hyperparameters. The findings revealed that the model achieved high metrics in terms of accuracy, precision, recall, F1 score, and area under the curve. Therefore, the deep learning classifier effectively identified and utilized the acoustic features that distinguish L1 from L2 speakers. Additionally, the results indicate the specific challenges L2 speakers face in producing L2 vowels, as evidenced by their divergence from L1 productions. These findings underscore the potential of deep learning algorithms to provide detailed insights into pronunciation difficulties encountered by L2 speakers. Such insights can be instrumental in developing more effective language learning strategies, tailoring pronunciation training to address specific issues faced by nonnative speakers. Moreover, these advancements can enhance language recognition technologies, making them more adaptable to the variations in speech patterns of L2 speakers. Overall, the study highlights the valuable role of AI in advancing our understanding of linguistic differences and improving language education and technology.

Keywords: deep learning; acoustic features; native and nonnative speakers.

Introduction

The challenges associated with producing nonnative or second language (L2) sounds are extensively documented in the literature (e.g., Georgiou & Themistocleous, 2021; Georgiou, 2021a, 2021b; Lee & Rhee, 2019; Piske et al., 2002). Difficulties in production are often attributed to learners' inability to accurately perceive target sounds. This difficulty may arise from age-related developmental processes that make speech acquisition mechanisms more specialized for processing first language (L1) input, or from a decline in perceptual sensitivity due to adults' extensive experience with their L1 (Iverson et al., 2003). Given these challenges, distinguishing between L1 and L2 speakers based on their speech characteristics remains a complex task. This study aims to address this challenge by exploring how a deep learning algorithm can differentiate between L1 and L2 speakers by analyzing the acoustic characteristics of their speech. By leveraging advanced computational methods, the research seeks to enhance our ability to identify and distinguish between native and nonnative speakers with greater accuracy.

Learners of a nonnative language face difficulties in acquiring sounds that do not exist or do not create contrast in their L1 (Bohn & Munro, 2007). For instance, the absence of the /ɪ – iː/ contrast in languages such as Catalan, Greek, Mandarin Chinese, and Spanish leads to perceptual and production challenges as both vowels are perceived as equivalents of a single vowel in the listeners' L1 (Georgiou, 2021c, 2022a; Morisson, 2008; Yang et al., 2015). More extensive and complex L2 vowel systems compared to the listeners' L1 systems may exacerbate speech acquisition problems, as speakers will attempt to map all L2 sounds onto the limited number of L1 sounds (e.g., Georgiou et al., 2020; Hacquard et al., 2007; Iverson & Evans, 2007). For example, Georgiou et al. (2020) found that Russian learners of English with low vocabulary sizes assimilated English /ɪ/ and /iː/ to Russian /i/, English /e/ and /æ/ to Russian /e/, and English /ʊ/ and /uː/ to Russian /u/, resulting in moderate discrimination of these sounds. However, not just the size and complexity but also the crosslinguistic acoustic similarity between L1 and L2 sounds can predict L2 speech acquisition (Alispahic et al., 2017; Elvin et al., 2021; Georgiou, 2023a; 2024). For example, Albanian speakers, whose L1 includes seven vowels covering all five Greek vowel qualities, produced the Greek vowel /e/ more fronted and the Greek vowel /u/ more backed. This is because Albanian /e/ is more fronted than Greek /e/, and Albanian /u/ is more backed than Greek /u/ (Georgiou & Kaskampa, 2024; Georgiou & Giannakou, 2024).

In recent decades, the advancement of Artificial Intelligence (AI) has fostered powerful technologies capable of performing complex classifications and predictions. Machine learning, a significant component of AI, has found widespread application in linguistics across both typical and atypical populations (Georgiou, 2023b; Georgiou & Theodorou, 2023; Johnson & Kang, 2015; Xiong, 2023). Deep learning, a subset of machine learning algorithms, stands out by enabling computational models with multiple layers to learn data representations at varying levels of abstraction (LeCun et al., 2015). This capability has been leveraged in numerous studies where acoustic features serve as foundational elements for classification tasks (Dewa, 2016; Kobayashi & Wilson, 2020). For instance, Ferragne et al. (2019) employed deep learning techniques to classify speakers based on spectrograms of their production of the French vowel /ɑ̃/. The algorithm achieved a high accuracy rate of 85%, demonstrating its effectiveness in distinguishing between different speakers based on this specific acoustic feature. Themistocleous et al. (2018) utilized deep sequential neural networks to detect mild cognitive impairment by analyzing acoustic features such as fundamental frequency and formant frequencies of vowels. Their model demonstrated high accuracy in distinguishing patients from healthy individuals, suggesting potential for early disease diagnosis enhancement. While existing research underpins deep learning's efficacy in acoustic-based classification tasks, future investigations should aim to expand these methodologies to address classification challenges pertaining to L2 speakers.

This study aims to investigate the ability of a deep learning algorithm to classify L1 and L2 speakers of English on the basis of the acoustic characteristics of vowels. Specifically, the participants consisted of Cypriot Greek speakers of English as an L2 and L1 English speakers, who participated in controlled production tasks. As evidenced by several studies, Cypriot Greek speakers of English experience significant difficulties in perceiving and producing English vowels as a consequence of the influence of their L1 (Georgiou, 2019, 2022a, 2022b). To the best of our knowledge, the use of a deep learning classifier based on phonetic features has never been employed for this population. Taking into account the powerful capabilities of deep learning, it is expected that the algorithm will classify with success the L1 and L2 speakers.

1. Methodology

1.1. Participants

The study included 18 participants elicited from Georgiou and Savva (2024). Among them, eight were adult Cypriot Greek speakers of L2 English, aged between 18 and 21 (*M*age = 20.13; *SD* = 1.36). At the time of the study, these participants were pursuing a BA in English Language and Literature and came from moderate socioeconomic backgrounds. Their English proficiency, as indicated by their CEFR certificates, was at the C1 level, and none had lived in an English-speaking country. According to their self-reported questionnaire data, they began learning English at an average age of 8.13 years (*SD* = 0.99), listened to English for an average of 6.13 hours per day (*SD* = 3.31), and spoke English for an average of 3.38 hours per day (*SD* = 1.09). They rated their English-speaking skills at 4.25 out of 5 *(SD* = 0.27). The control group consisted of 10 Standard Southern British English speakers as reported by Georgiou (2024). All participants were female. All participants had normal vision and hearing and no history of cognitive or language disorders. They were informed about the study's objectives and their rights before participating and provided written consent in accordance with the Declaration of Helsinki. The characteristics of the participants are shown in Table 1.

Mean age in years(*SD*)	English onset age in years(*SD*)	English input in hours(*SD*)	English use in hours(*SD*)	English speaking skills; out of 5
20.13(1.36)	8.13(0.99)	6.13(3.31)	3.38(1.09)	4.25

Table 1: Participants' characteristics

1.2. Materials

The study materials included 11 monosyllabic English words within an /hVd/ context, each representing one of the English vowels /ɪ iː e ɜː æ ɑː ʌ ɒ ɔː uː ʊ/. These words were embedded in the carrier phrase "You say /hVd/ now".

1.3. Procedure

1.3.1. Production task

Each participant completed the assessment individually in a quiet environment. They were given phrases on paper and instructed by the researcher to read them aloud as if they were speaking to a friend. The phrases were written in Standard British English orthography. Their spoken responses were recorded using a professional voice recorder at a 44.1 kHz sampling rate and saved as .wav files with 24-bit resolution. Each participant read the phrases twice, with the order of the phrases randomized for each participant. Before the test, it was ensured that participants knew the words and could correctly pronounce the target vowels by associating the words with other commonly used words containing the vowels being studied.

1.3.2. Acoustic analysis

The words from the speakers were isolated and analyzed using Praat software (Boersma & Weenink, 2024). By visually inspecting the spectrograms and waveforms, we were able to identify key acoustic features and measure vowel characteristics such as formant frequencies and duration. The analysis settings included a 0.025-second positive window length, 50 Hz pre-emphasis, and a spectrogram range of up to 5500 Hz (see Georgiou & Dimitriou, 2023). Formant frequencies were measured from the end of the preceding consonant /h/ to the start of the vowel (V), ending at the end of the vocalic period, and the start of the following consonant /d/. To reduce the influence of neighboring sounds, measurements were taken at the midpoint of each vowel segment. Vowel durations were manually to determine the start and end points for each vowel token. To accommodate the variation in F1, F2, and F3 values among all speakers, normalization was performed using the Lobanov *z*-score method (Lobanov, 1971).

1.3.3. Training of the deep learning algorithm

We utilized a deep learning algorithm for training the neural network mode using the h2o package (Fryda et al., 2024) in R (R Core Team, 2024). The dataset was then split into training and testing subsets with a 90-10 ratio to ensure a robust evaluation of the model's performance. A comprehensive grid search was employed to optimize the architecture of the deep learning model. The hyperparameters included varying the number of hidden units and epochs. We leveraged crossvalidation with five folds to ensure generalization and prevent overfitting. The best model from the grid search was identified, and its performance was evaluated on the testing dataset. Finally, the best-performing model from the hyperparameter grid was selected based on its evaluation metrics.

The model was trained with the following vowel speech measures: F1, F2, F3, and duration. Formants are the resonant frequencies of the vocal tract, and they are crucial in determining the quality of the sounds we hear. More specifically, F1 corresponds to the tongue height, that is, the distance of the tongue from the roof of the mouth. F2 corresponds to the tongue frontness, that is, how front is the tongue when we produce a vowel. F3 is associated with lip rounding during the articulation of vowels (Georgiou, 2020). Formants are measured in Hertz (Hz). Duration refers to the length of time a vowel sound is articulated during speech and is typically measured in milliseconds (ms). All these features are important for the perception and production of vowels across most of the languages. For a schematic architecture of the deep learning model, see Figure 1.

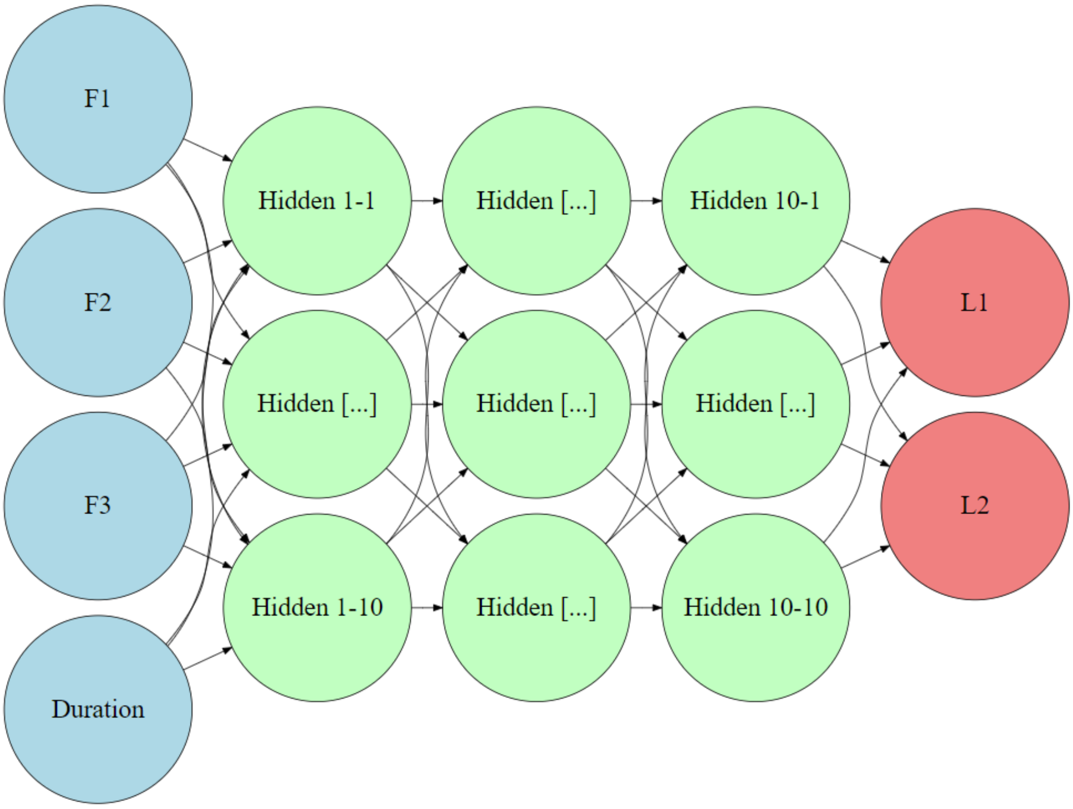

Figure 1: Schematic architecture of the deep learning model. Data flows from the input layer (F1, F2, F3, and Duration) through 10 hidden layers consisting of 10 neurons each to the output layer (L1 and L2)

The performance of the trained deep neural network model was evaluated using several key metrics on the testing subset. These metrics include accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic (ROC) curve (AUC). Accuracy measures the proportion of correctly classified instances among all instances in the dataset, providing an overall assessment of the correctness of the model. It is computed as (True Positives + True Negatives) / (Total Positives + Total Negatives + False Positives + False Negatives). Precision quantifies the proportion of true positive predictions out of all positive predictions made by the model, highlighting how well the model avoids false positives. It is calculated as True Positives / (True Positives + False Positives). Recall assesses the ability of the model to correctly identify all positive instances, measuring the proportion of true positive predictions out of all actual positive instances. It is calculated as True Positives / (True Positives + False Negatives). F1-score is a harmonic mean of precision and recall, offering a balanced evaluation of the model's performance by considering both false positives and false negatives. This metric is calculated using the formula 2 * (Precision * Recall) / (Precision + Recall). ROC curve is a graphical representation showing the model’s ability to distinguish between classes at various threshold settings. AUC (Area Under the Curve) quantifies the overall performance of the model in terms of its ability to differentiate between positive and negative instances. A perfect model would have an AUC of 1, while a random model would have an AUC of 0.5. A higher AUC value indicates a stronger model, whereas an AUC value closer to 0.5 suggests that the model's performance is not significantly better than random guessing.

2. Results

The results of the deep learning model indicated high classification scores for all metrics. These scores ranged from 0.83 to 1.00, indicating the high efficacy of the algorithm in assigning L1 and L2 speakers to their respective classes. More specifically, the highest scores were observed for precision and recall, followed by F1, accuracy, and AUC. Table 2 presents the scores of each metric, while Table 3 shows the metrics for each fold. Figure 2 illustrates the ROC curves of the training and testing datasets. The larger AUC score of the testing dataset compared to the corresponding AUC score of the training dataset indicates lower chances for overfitting.

Metric	Score
Accuracy	0.84
Precision	1.00
Recall	1.00
F1	0.85
AUC	0.83

Table 2: Scores for each metric

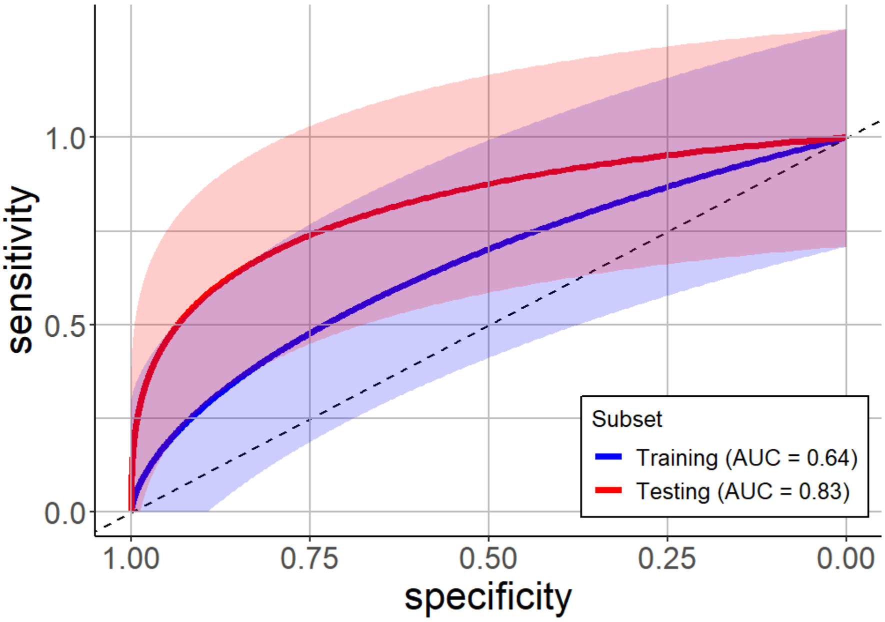

Figure 2: Mean ROC curves and AUC values of the training (blue) and testing (red) subsets. The dashed gray diagonal line shows the baseline. The shaded area indicated ± 1 standard deviation from the mean for two curves.

Metric	mean	SD	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Accuracy	0.661538	0.090956	0.641791	0.756757	0.539474	0.622951	0.746667
Precision	0.584571	0.100389	0.566038	0.71875	0.460317	0.531915	0.645833
Recall	0.910818	0.10798	0.967742	0.71875	0.966667	0.961538	0.939394
F1	0.701411	0.05216	0.714286	0.71875	0.623656	0.684932	0.765432
AUC	0.720117	0.028864	0.691756	0.717262	0.697826	0.72967	0.764069

Table 3: Metrics for each fold together with the means and SDs during crossvalidation

3. Discussion

The study implemented a deep learning neural network algorithm to classify L1 and L2 speakers of English. The classifier was trained with speech features such as F1, F2, F3, and duration of English vowels elicited from controlled productions from both L1 and L2 speakers. The goal was to gather various metrics from the algorithm to evaluate its performance; this would allow us to explore how L2 speech differentiates from the respective L1 speech.

Our evaluation metrics demonstrated strong overall performance for the deep learning model in distinguishing between L1 and L2 speakers. This aligns with previous research suggesting the high efficacy of deep learning in distinguishing groups based on acoustic features (e.g., Themistocleous et al., 2018). We implemented rigorous validation techniques to ensure the reliability of our results. Initially, we employed established methods like k-fold crossvalidation to evaluate the generalization performance of the model. Additionally, we conducted a grid search over a range of hyperparameters to optimize the model's architecture, focusing on the number of hidden units and epochs. The dataset was divided into two subsets: the training subset and the testing subset, comprising 90% and 10% of the data, respectively. The high metric values observed suggest that the model generalizes well to new, previously unseen data. Generalization is a key indicator of a model's ability to perform accurately in real-world scenarios beyond the training data. The success of the model on the testing subset indicates that it has effectively captured the underlying patterns in the data during training without overfitting, thus enhancing its reliability and utility.

The results suggest that L2 speakers encounter difficulties with the accurate production of the L2 vowels. This research expands prior findings regarding the challenges Cypriot Greek speakers face in the acquisition of English sounds (e.g., Georgiou, 2019; 2022b; Georgiou et al., 2024). Although participants were advanced speakers of English, they were not able to produce the L2 vowels in a native-like manner. This might be due to the fact that explicit teaching of pronunciation does not typically occur in Cypriot Greek classrooms, where other linguistic levels are prioritized (Georgiou, 2019). In addition, as participants had never lived in an English-speaking country, chances for exposure to qualitative input in the L2 are minimal.

By identifying specific acoustic features that distinguish L1 and L2 speakers, the study enhances our understanding of how nonnative speakers produce speech differently from native speakers. The findings can pinpoint specific areas where L2 speakers struggle, providing valuable insights for linguists and language educators to develop targeted interventions to help learners improve their pronunciation. Insights from the study can inform the creation of pronunciation curricula that address the specific needs of L2 learners, using data-driven methods to focus on the most challenging aspects of pronunciation. Educators can be trained to recognize and address the common pronunciation issues faced by L2 learners, using evidence-based techniques to help students achieve better outcomes. The development of classifiers that accurately distinguish between L1 and L2 speakers can improve speech recognition systems by allowing them to adapt to the specific characteristics of nonnative speech, thereby increasing their accuracy and usability for diverse user groups.

4. Conclusion

While the deep learning algorithm showed promising outcomes, future efforts should expand the participant pool to bolster the reliability of the results. Furthermore, upcoming research could enhance the algorithm by including more predictor variables, such as other acoustic measures extracted from speech samples (see Georgiou & Kaskampa, 2024). Integrating these additional variables has the potential to enhance the ability of the model to predict accurately, leading to improved accuracy in distinguishing between L1 and L2 speakers.

Acknowledgments

The study was supported by the Phonetic Lab of the University of Nicosia. Informed consent was obtained from all participants according to the Declaration of Helsinki. The author has no competing interests to disclose.

References

Alispahic, S., Mulak, K. E., & Escudero, P. (2017). Acoustic properties predict perception of unfamiliar Dutch vowels by adult Australian English and Peruvian Spanish listeners. *Frontiers in Psychology, 8*, 52.

Bohn, O. S., & Munro, M. J. (Eds.). (2007). *Language experience in second language speech learning: In honor of James Emil Flege* (Vol. 17). John Benjamins Publishing.

Dewa, C. K. (2016). Javanese vowels sound classification with convolutional neural network. In *2016 International Seminar on Intelligent Technology and Its Applications (ISITIA)* (pp. 123-128). IEEE.

Ferragne, E., Gendrot, C., & Pellegrini, T. (2019). Towards phonetic interpretability in deep learning applied to voice comparison. In *ICPhS* (pp. ISBN-978).

Fryda, T. et al. (2024). *h2o: R Interface for the 'H2O' Scalable Machine Learning Platform*. R package version 3.44.0.3.

Georgiou, G. P. (2019). 'Bit' and 'beat' are heard as the same: Mapping the vowel perceptual patterns of Greek-English bilingual children. *Language Sciences, 72*, 1-12.

Georgiou, G. (2020). *An Introduction to Issues in General Linguistics*. Cambridge Scholars Publishing.

Georgiou, G. P. (2021a). Interplay between perceived cross-linguistic similarity and L2 production: Analyzing the L2 vowel patterns of bilinguals. *Journal of Second Language Studies, 4*(1), 48-65.

Georgiou, G. P. (2021b). Effects of phonetic training on the discrimination of second language sounds by learners with naturalistic access to the second language. *Journal of Psycholinguistic Research, 50*(3), 707-721.

Georgiou, G. P. (2021c). Toward a new model for speech perception: The Universal Perceptual Model (UPM) of Second Language. *Cognitive Processing, 22*(2), 277-289.

Georgiou, G. P. (2022a). The acquisition of /ɪ/–/iː/ is challenging: Perceptual and production evidence from Cypriot Greek speakers of English. *Behavioral Sciences, 12*(12), 469.

Georgiou, G. P. (2022b). The impact of auditory perceptual training on the perception and production of English vowels by Cypriot Greek children and adults. *Language Learning and Development, 18*(4), 379–392.

Georgiou, G. P. (2023a). Speakers of different L1 dialects with acoustically proximal vowel systems present with similar nonnative speech perception abilities: Data from Greek listeners of Dutch. *Speech Communication, 150*, 32-40.

Georgiou, G. P. (2023b). Comparison of the prediction accuracy of machine learning algorithms in crosslinguistic vowel classification. *Scientific Reports, 13*, 15594

Georgiou, G. P. (2024). Classification of English vowels in terms of Cypriot Greek categories: the role of acoustic similarity between L1 and L2 sounds. *Canadian Journal of Linguistics, 69*(1), 46-62.

Georgiou, G. P. & Dimitriou, D. (2023). Perception of Dutch vowels by Cypriot Greek listeners: to what extent can listeners' patterns be predicted by acoustic and perceptual similarity? *Attention, Perception, and Psychophysics*, 85, 2459–2474.

Georgiou, G. P, & Giannakou A. (2024). Acoustic Characteristics of Greek Vowels Produced by Adult Heritage Speakers of Albanian. *Acoustics, 6*(1), 257-271.

Georgiou, G. P., & Kaskampa, A. (2024). Differences in voice quality measures among monolingual and bilingual speakers. Ampersand, 12, 100175.

Georgiou, G. P., & Savva, E. (2024). Exploring acoustic overlap in second language vowel productions (submitted).

Georgiou, G. P., & Themistocleous, C. (2021). Vowel Learning in Diglossic Settings: Evidence from Arabic-Greek Learners. *International Journal of Bilingualism, 25*(1), 135-150.

Georgiou, G. P., & Theodorou, E. (2023). Detection of developmental language disorder in Cypriot Greek children using a machine learning neural network algorithm. *arXiv preprint* arXiv:2311.15054.

Georgiou, G. P., Giannakou, A., & Alexander, K. (2024). Perception of second language phonetic contrasts by monolinguals and bidialectals: a comparison of competencies. *Quarterly Journal of Experimental Psychology.* https://doi.org/10.1177/17470218241264566

Georgiou, G. P., Perfilieva, N, & Tenizi, M. (2020). Vocabulary size leads to better attunement to L2 phonetic differences: Clues from Russian learners of English. *Language Learning and Development, 16*(4), 382-398.

Hacquard, V., Walter, M. A., & Marantz, A. (2007). The effects of inventory on vowel perception in French and Spanish: An MEG study. *Brain and language*, *100*(3), 295-300.

Iverson, P., & Evans, B. G. (2007). Learning English vowels with different first-language vowel systems: Perception of formant targets, formant movement, and duration. *The Journal of the Acoustical Society of America, 122*(5), 2842–2854.

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, *87*(1), B47-B57.

Johnson, D. O., & Kang, O. (2015). Automatic prominent syllable detection with machine learning classifiers. *International Journal of Speech Technology*, *18*, 583-592.

Kobayashi, A., & Wilson, I. (2020). Using deep learning to classify English native pronunciation level from acoustic information. In *SHS Web of Conferences* (Vol. 77, p. 02004). EDP Sciences.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444.

Lee, S., & Rhee, S. C. (2019). The relationship between vowel production and proficiency levels in L2 English produced by Korean EFL learners. *Phonetics and Speech Sciences*, *11*(2), 1-13.

Lobanov, B. M. (1971). Classification of Russian vowels spoken by different speakers. *Journal of Acoustical Society of America, 49*, 606–608.

Morrison, G. S. (2008). L1-Spanish Speakers' Acquisition of the English/i/—/I/Contrast: Duration-based perception is not the initial developmental stage. *Language and Speech, 51*(4), 285–315.

Piske, T., Flege, J. E., MacKay, I. R., & Meador, D. (2002). The production of English vowels by fluent early and late Italian-English bilinguals. *Phonetica*, *59*(1), 49-71.

R Core Team (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Themistocleous, C., Eckerström, M., & Kokkinakis, D. (2018). Identification of mild cognitive impairment from speech in Swedish using deep sequential neural networks. *Frontiers in neurology*, *9*, 975.

Xiong, W. (2023). A Study on the Recognition of English Pronunciation Features in Teaching by Machine Learning Algorithms. *Journal of Computing Science and Engineering*, *17*(3), 93-99.

Yang, X., Shi, F., Liu, X., & Zhao, Y. (2016). Learning styles and perceptual patterns for English/i/and/ɪ/among Chinese college students. *Applied Psycholinguistics, 37*(3), 673-701.