



UNIVERSITAT DE
BARCELONA

Department of Modern Languages and Literatures and
English Studies

M.A. Thesis

**Was That a Bag or a Bug?
Perceptual Measures, Euclidean Distance, Mahalanobis
Distance, and Pillai Scores in the Assessment of L2
Pronunciation**

James Waltz

Supervisor: Dr. Joan C. Mora

Academic year: 2020-2021

**Màster Oficial en Lingüística Aplicada
i Adquisició de Llengües en Contextos Multilingües
LAALCM**

Joan C. Mora com a supervisor/a del treball (Tesina de
(nom i cognoms)
Màster) presentat com a requeriment per a l'avaluació de l'assignatura **Projecte de**

Recerca en Lingüística Aplicada

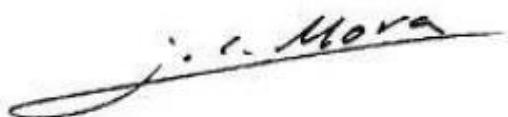
presentat per l'alumne/a: James Waltz
(nom i cognoms)

amb el títol de: Was that a Bag or a Bug? Perceptual Measures, Euclidean
Distances, Mahalanobis Distances, and Pillai Scores in the Assessment of L2
Pronunciation

certifico que he llegit el treball i l'aprovo perquè pugui ser presentat per a la seva
defensa pública.

I perquè consti i tingui els efectes oportuns signo aquest certificat en

Barcelona, a 02 de setembre de 2021



Dr/a. Joan C. Mora



UNIVERSITAT DE
BARCELONA

Facultat de Filologia i Comunicació
Dept. Llengües i Lit. Modernes i Estudis Anglesos

Gran Via de les Corts Catalanes, 585
08007 Barcelona, SPAIN
Tel. +34 934 035 686
Fax +34 933 171 249

**Official MA programme in
Applied Linguistics and Language Acquisition in Multilingual Contexts
(LAALCM)**

Universitat de Barcelona

Non-Plagiarism Statement

This form must be completed, dated and signed and must be included at the beginning of every copy of the MA Thesis you submit for assessment.

<i>Name and surnames:</i>	James Waltz
<i>MA Thesis title:</i>	Was that a Bag or a Bug? Perceptual Measures, Euclidean Distances, Mahalanobis Distances, and Pillai Scores in the Assessment of L2 Pronunciation
<i>Supervisor:</i>	Joan C. Mora

I HEREBY DECLARE THAT:

- This MA Thesis that I am submitting for assessment is entirely my own work and I have written it completely by myself.
- I have not previously submitted this work or any version of it for assessment in any other programme or institution.
- I have not used any other sources or resources than the ones mentioned.
- I have identified and included the source of all facts, ideas, opinions and viewpoints of others through in-text referencing and the relevant sources are all included in the list of references at the end of my work. Direct quotations from books, journal articles, internet sources or any other source whatsoever are acknowledged and the sources cited are identified in the list of references.

I understand that plagiarism and copying are serious offences. In case of proof that this MA Thesis fails to comply with this declaration, either as negligence or as a deliberate act, I understand that the examiner has the right to exclude me from the assessment act and consequently all research activities conducted for this course will be declared null and the MA Thesis will not be presented for public defense, thus obtaining the lowest qualification.

Date: 01/09/2021

Signature:

Abstract

Researchers employ a variety of techniques to measure accuracy of second-language pronunciation. Little research has been done on certain measures that have been used more in recent studies, such as Mahalanobis distance and Pillai scores, and how they compare to perceptual measures. Using pre- and post-test recordings of 23 Spanish/ Catalan learners of English that were obtained using a delayed word repetition task in a previous, high-variability phonetic training study on the English phonemes /æ/ and /ʌ/, this thesis examines the relationship between native-speaking judges' word identification and goodness ratings, Euclidean distances, Mahalanobis distances, and Pillai scores in their evaluation of pronunciation accuracy and improvement between test times. For each acoustic metric, measures between native- and non-native speakers' productions are taken as well as measures between non-native speakers' realizations of /æ/ and /ʌ/. An experimental way of computing perceptual ratings for items that are incorrectly identified by raters is also investigated and compared to existing measures.

Table of Contents

1. Introduction.....	1
2. Literature Review.....	2
2.1 Acquisition of L1 and L2 Phonology.....	2
2.2 Measuring Perception in L2 Speech Research.....	4
2.3 Measuring Pronunciation Accuracy in L2 Speech Research.....	5
2.3.1 Perceptual Measures of L2 Pronunciation.....	5
2.3.2 Acoustic Measures of L2 Pronunciation.....	6
2.3.2.1 Vowels, Formants, and Psychoacoustics.	7
2.3.2.2 Euclidean Distance.	9
2.3.2.3 Mahalanobis Distance.	10
2.3.2.4 Pillai Scores.	12
2.4 Comparing Acoustic and Perceptual Measures.....	14
2.5 Improving L2 Pronunciation.....	15
2.5.1 High Variability Phonetic Training.....	16
3. Research Questions	16
4. Methodology	17
4.1 Participants and Training.....	17
4.2 Raters.....	18
4.3 Perceptual Measures.....	19
4.4 Acoustic Measures.....	20
5. Results.....	23
5.1 Inter-Rater Reliability.....	23
5.2 Correlations Between AdjRat and Other Measures.....	23
5.3 Correlations Between All Measures.....	25
5.4 Pre- and Post-Test Differences.....	26
5.4.1 NNS-NS Acoustic Measures.....	26
5.4.2 Acoustic /æ/-/ʌ/ Measures and Perceptual Measures.....	27
6. Discussion.....	28
7. Conclusion.....	31
References.....	32
Appendices.....	37
Appendix A.....	37
Appendix B.....	38
Appendix C.....	39
Appendix D.....	40
Appendix E.....	49

1. Introduction

Learning a second language (L2) is not a simple undertaking. Even very closely related languages can vary widely in their vocabulary and grammar. Furthermore, L2 phonology, the sounds of the language and the rules that they follow, can be particularly difficult to acquire. When setting out to learn a new language, many learners find difficulty in hearing certain L2 sounds that do not exist in their first language (L1). This takes a toll not only on listening comprehension, but also the ability to produce those difficult sounds in one's own L2 speech. Even individuals who have spoken an L2 regularly for decades sometimes have an accent that is informed by their L1.

Because of this, it is no wonder that many people want to better acquire the phonology of the L2. For those language learners and L2 acquisition researchers alike, understanding the nature of these difficulties and finding which methods are effective at improving L2 pronunciation is an important task. Researchers must employ well-developed methods to measure L2 pronunciation accuracy to determine how accurate or accented pronunciation is and which interventions are most effective at improving L2 perception and pronunciation.

The goal of this thesis is to investigate how different acoustic and perceptual tools measure L2 pronunciation accuracy and how each of them measure the effects of phonetic training. This will be accomplished by using pre- and post-intervention speech samples from a group of Spanish-Catalan bilingual learners of English that were collected in a previous study by Mora et al. (under review). In that study, participants were trained on perceptual and productive differentiation of the English /æ/ and /ʌ/ contrast that is difficult for many Spanish speakers.

In this study, native English-speaking judges were recruited and an identification and rating task was created for those raters to evaluate the participants' productions. These ratings were then compared with acoustic measurements that had been analyzed by a variety of methods to see how the perceptual and acoustic data correlate and which capture pre- and post-test differences in L2 pronunciation accuracy.

2. Literature Review

To understand how an L2 learner acquires a new phonology, it is important, first, to understand how L1 phonology is acquired and how that affects L2 acquisition.

2.1 Acquisition of L1 and L2 Phonology

Infants begin with an ability to discriminate between virtually all speech sounds, but between the ages of 6 and 12 months, they already show improvement in discriminating between sounds of their L1 and a diminished ability with foreign language (FL) speech sounds (Kuhl et al., 2006). These L1 speech sounds form perceptual categories, known as phonological representations, to interpret the relevant acoustic qualities of the spoken language. For example, an infant being raised in an English-speaking environment will develop two phonological representations for the /r/ and /l/ phonemes that allow them to differentiate between these sounds, whereas an infant raised in a Japanese-speaking environment will develop a single phonological representation to handle all sounds in the same range of acoustic qualities, leading to decreased sensitivity to FL contrasts, as can be seen in Figure 1.

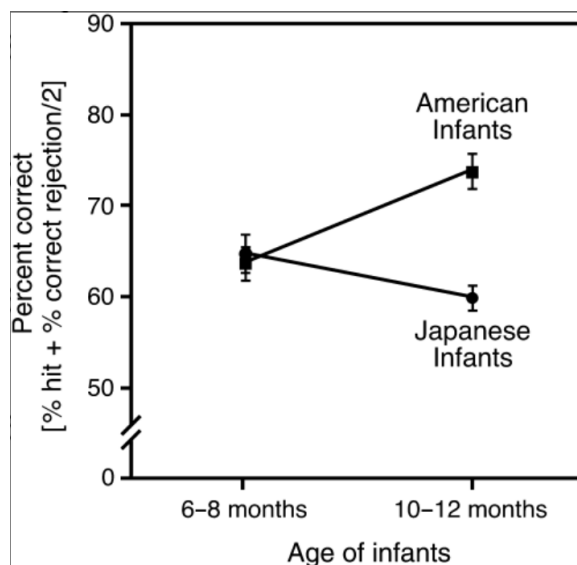


Figure 1. Discrimination of /ra/ and /la/ among American and Japanese Infants. From “Infants show a facilitation effect for native language phonetic perception

between 6 and 12 months” by P. K. Kuhl et al., 2006, *Developmental Science*, 9(2), p. F17.

These changes are not only in perception, but also in neural tissue and circuitry, a process referred to as *native language neural commitment* (NLNC) (Kuhl, 2004). The acquired phonological representations are also encoded into the vocabulary learned in the language so that one can recognize a word upon hearing it and produce it in speech. The mental representations of the sounds of words are called phono-lexical representations. Under normal circumstances, when an individual hears a word spoken in their L1, they have no trouble processing the acoustic information to activate the relevant phonological and lexical representations (Ramus et al., 2010). The matter is not as simple for processing speech in the L2.

The prominent models of L2 speech processing, PAM-L2 (Best & Tyler, 2007), NLM-e (Kuhl et al., 2008) SLM (J. E. Flege, 1995), and SLM-r (Flege et al., 2021), all propose that the structure of L1 phonology affects the perception of L2 and FL speech, creating difficulty in perceiving, encoding, and producing L2 speech accurately.

Not all L2 sounds are equally problematic. The relationship between the L2 sounds and the L1 categories affects how difficult a sound is to perceive and produce accurately. For example, L2 language learners and users demonstrate more difficulty when two sounds in the L2 are mapped onto a single sound category in the L1. Referring back to the example of L1 Japanese learners of L2 English, the difficulty with /r/ and /l/ stems from the fact that both sounds are mapped onto an existing Japanese sound category, /r/. Because of this, L1 Japanese learners of L2 English perceive both /r/ and /l/ realizations as falling within a single sound category, and so they struggle to discern the essential differences between the individual sounds as well as the differentiate relevant minimal pairs, such as “lake” and “rake.” Similarly, when using such words in speech, they are likely to produce something closer to the Japanese consonant, which falls somewhere between the English consonants, and sounds strongly accented to English speakers. While accurate perception of L2 sounds is not the

only factor that contributes to difficulty in the production of those sounds (Sakai & Moorman, 2018), it is widely asserted that accurate perception is a necessary condition for more accurate production in L2 speech (Flege, 1995).

While the difficulty faced by Japanese speakers when learning English is an easily recognizable and widely discussed and researched example, it is far from the only one. Similar difficulties exist for L1 Dutch learners of English for the sounds /ɛ/ and /æ/ (as in *bet* and *bat*) which are both perceived as the Dutch /ɛ/ (Escudero et al., 2008). The case that will be examined further in this thesis is for Spanish-Catalan bilingual learners of English and the sounds /æ/ and /ʌ/ (as in *cat* and *cut*)

These learners of English often demonstrate difficulty distinguishing the contrast between /æ/ and /ʌ/, both in perception and production (Aliaga-García & Mora, 2009; Carlet & Kivistö de Souza, 2018). A major reason for this difficulty is that both L2 vowels are often mapped onto the L1 vowel category /a/ (Rallo-Fabra & Romero, 2012). The English /æ/ is more similar to the Spanish /a/ than the English /ʌ/ is, both in terms of acoustic vowel quality and perceptual judgements by Spanish speakers, making it easier for Spanish speakers to identify the acoustic differences between /ʌ/ and /a/ than between /æ/ and /a/ (Cebrian et al., 2011). Recent studies have shown that, while the /æ/ and /ʌ/ contrast is particularly difficult for English learners with this background, some types of training are effective at improving both perception and production of these vowels (Carlet & Cebrian, 2019).

2.2 Measuring Perception in L2 Speech Research

For researchers examining the acquisition of L2 phonology, it is not sufficient to merely find that difficulties exist. Researchers must also measure accuracy of L2 perception and production to better understand the nature of those difficulties, as well as to determine which factors or interventions are most effective at overcoming them.

This thesis will focus on L2 production and measures thereof, but because of the importance of accurate L2 perception in developing accurate L2

production, it is worth taking a moment to mention one of the most commonly used perceptual measures: the ABX task. In a typical ABX task, a listener will hear two words, *A* and *B*, that differ in a single phoneme (e.g., *A* = cat, *B* = cut). They will then hear a third word, *X*, which will match one of the previous two (e.g., *X* = cat). The listener will then select whether *X* matched *A* or *B*. Typically an ABX task will have one or two contrasts under investigation and a number of distractor items that contain non-target contrasts so that participants do not become aware and focused on the specific contrast being studied. The more items above chance a participant can correctly identify, the more accurate their perception of the target contrast can be said to be.

2.3 Measuring Pronunciation Accuracy in L2 Speech Research

When measuring accuracy of L2 pronunciation, various methods exist. They can generally be divided into two categories: perceptual and acoustic measures.

2.3.1 Perceptual Measures of L2 Pronunciation

Perceptual measures are evaluations by human listeners. Judges, typically native speakers of the target language, listen to speech samples and evaluate them by criteria specified by the task they are given. Perceptual measures are preferred over acoustic measures by many researchers because they are considered more ecologically valid, or reflective of how speech would be perceived and understood in real-world communication (Munro, 2008). In some cases, these judges are so-called naïve judges, with no specific training (Flege et al., 1999; Muñoz & Llanes, 2014), while in others they are trained specialists such as linguists and L2 teachers (Carlet & Kivistö de Souza, 2018; Munro, 1993). Some researchers who used both expert and naïve judges have found expert judges to be more lenient and reliable when judging L2 accentedness (Kang, 2008; Thompson, 1991), although other studies have shown no significant differences between experts and naïve judges (Bongaerts et al., 1997).

The ways that judges evaluate the speech production of participants varies from study to study. The most widely used method involves judges listening to speech samples and rating them on an accentedness scale. Often, this scale is

given a label such as “Strong Foreign Accent” at one end and “No Foreign Accent” at the other (Piske et al., 2001). Scales of various sizes have been used, such as 5-point scales (e.g., Bongaerts et al., 1997; Mairano & Santiago, 2019; Thompson, 1991), 7-point scales (e.g., (Llanes et al., 2017), 9-point scales (e.g. Flege et al., 1999; Isaacs & Thomson, 2013), or even continuous scales, such as in Flege et al. (1995) in which raters moved a lever along a line which then created a score between 0 and 255. In a study using a 7-point scale and a direct magnitude estimation, Southwood and Flege (1999) suggest that 9- or 11-point scales may be better suited for measuring perceived foreign accent, although another study by Isaacs and Thomson (2013) found that raters struggle to choose among values in the center of the scale and that larger scales exacerbate this difficulty.

When examining particularly difficult sounds for L2 learners, only using ratings may be insufficient as the judges might not be able to correctly identify the produced word. In such cases, researchers sometimes use an identification task, in which the judge hears the word and then chooses from a number of options which word they heard. In this case, pronunciation accuracy can be measured by the percentage of judges who were able to correctly identify the word (Carlet & Kivistö de Souza, 2018; Iverson et al., 2012)

Identification tasks can also be paired with rating tasks. In a recent study, Carlet and Cebrian (2019) had judges perform a series of identification tasks with a 9-point rating scale from 1 (“difficult to identify as selected sound”) to 9 (“easy to identify as selected sound”). This study will adapt this method to calculate 3 means: percentage of correct identifications (ID), ratings of correctly identified words (RatCorr), as well as to test a novel scale in which ratings for incorrectly identified words are coded to indicate poorer pronunciation (AdjRat). These measures will be described in more detail in Section 4.3, Perceptual Measures of L2 Pronunciation.

Other methodological issues also complicate the use of human raters for measuring pronunciation accuracy. Rater decisions are unique to individuals, and raters with different backgrounds and expertise might provide different ratings, meaning that the use of rater judgements, however ecologically valid, are not fully

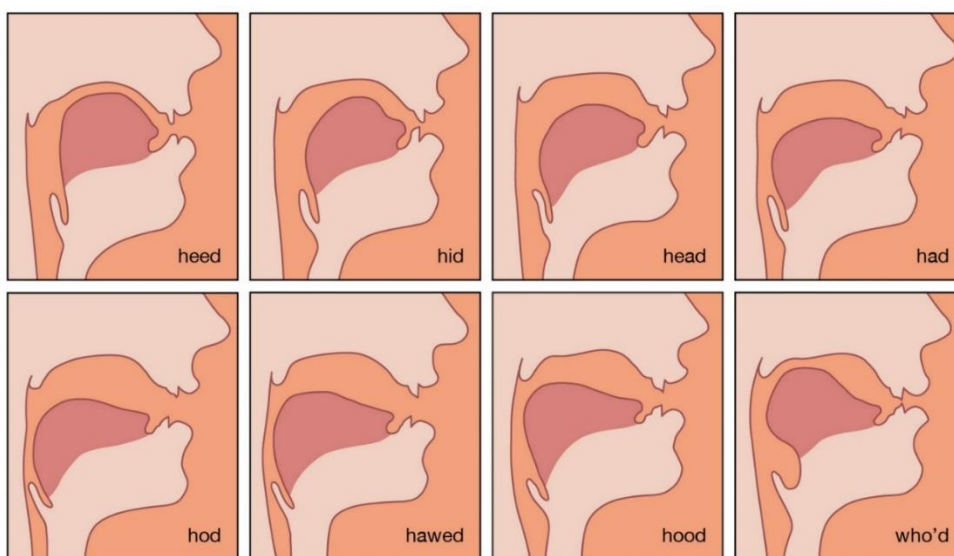
replicable (Kim, 2015). Use of raters to investigate specific L2 target sounds may be particularly problematic. Individual sounds are likely too short for raters to listen to and evaluate, but including the sounds context may lead to raters evaluating the overall accent of the speaker rather than the individual sound (Melnik-Leroy et al., 2021).

2.3.2 Acoustic Measures of L2 Pronunciation

Listener judgements are not the only way to evaluate accuracy of L2 pronunciation. Objective measures that look at the acoustic qualities of speech sounds are often used in research. Some of the most common acoustic measures taken to compare L1 and L2 speech are formant frequencies, voice-onset time, and peak intensity (Flege, 1987). As this thesis is focused on L2 English vowels, the discussion of acoustic measures will be limited to formants.

2.3.2.1 Vowels, Formants, and Psychoacoustics

Vowels are produced by vibrating the vocal cords while leaving the vocal tract unobstructed. The position of the tongue, the openness of the mouth, and the shape of the lips cause the sound produced by the vocal cords to create different resonant frequencies that determine the quality of the vowel, as shown in Figure 2. These frequencies are called *formants*. These formants can be most clearly seen and measured using a spectrogram as seen in Figure 3.



© Encyclopædia Britannica, Inc.

Figure 2. Tongue Position for Vowel Sounds. From *Encyclopedia Britannica*.
<https://www.britannica.com/science/phonetics/Vowels#/media/1/457255/3598>.

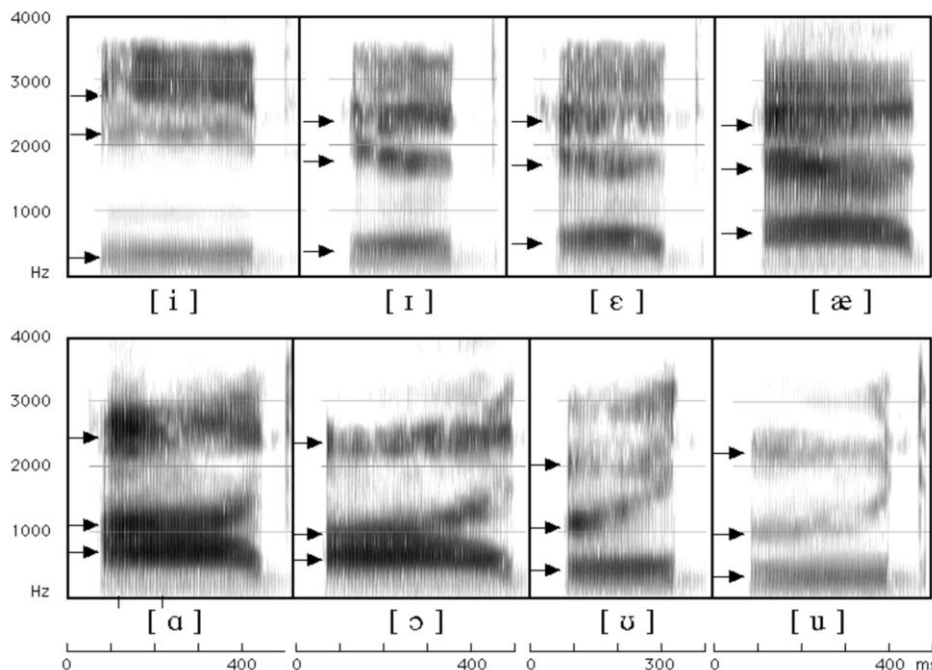


Figure 3. A Spectrogram of the Words “Heed,” “Hid,” “Head,” “Had,” “Hod,” “Hawed,” “Hood,” and “Who’d” with the Location of the First Three Formants Shown by Arrows. From *A Course in Phonetics* by P. Ladefoged, 2010, p. 194. Wadsworth Publishing Company.

The frequency of these formants, and particularly the first formant (F_1) and the second formant (F_2), are processed by the listener to identify vowel sounds and can be measured instrumentally by researchers. However, the relationship between frequency and perception is not amenable to being linearly partitioned, so it is necessary to convert the Hertz (Hz) measurements into a psychoacoustically valid scale, such as the widely used Bark (B) scale (Zwicker, 1961). Furthermore, individual and physiological differences, such as size of the larynx and vocal tract, affect how the formants are processed by listeners and measured by researchers, so a normalization procedure should be used, such as the Bark Difference method (Syrdal & Gopal, 1986).

With these measurements, it is possible to plot individual vowels on a formant chart, with the F_2 as the x-axis and the F_1 as the y-axis, as illustrated in Figure 4.

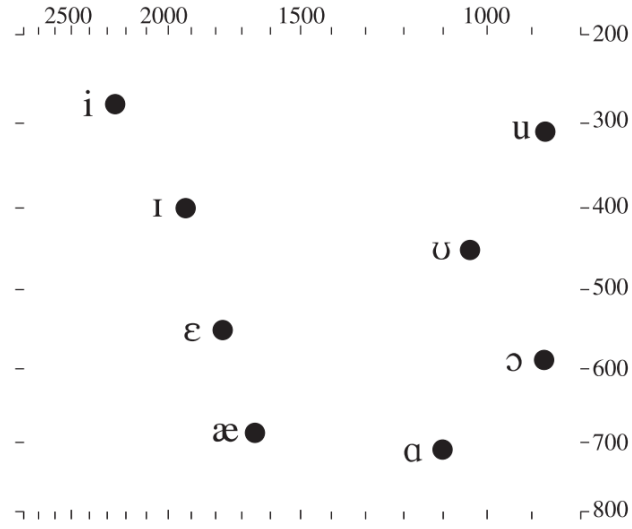


Figure 4. Formant Chart for Eight American English Vowels with the Scale Marked in Hz and Vowels Arranged at Bark Scale Intervals. From *A Course in Phonetics* by P. Ladefoged, 2010, p. 197. Wadsworth Publishing Company.

With the acoustic data processed, it is then possible to investigate how accurate pronunciation is for an L2 user. Different methods have been employed by researchers to do so. We will look at three: Euclidean distances, Mahalanobis distances, and Pillai scores.

2.3.2.2 Euclidean Distance

One method, and perhaps the most straightforward, to compare two vowel productions is to measure the Euclidean distance between them. A Euclidean distance is, in essence, the measurement of a straight line between two points. For example, to see whether /ε/ was more acoustically similar to /u/ or /æ/ in American English pronunciation, one could plot the vowels on a formant chart, draw a straight line between them, and compare them, as in Figure 5.

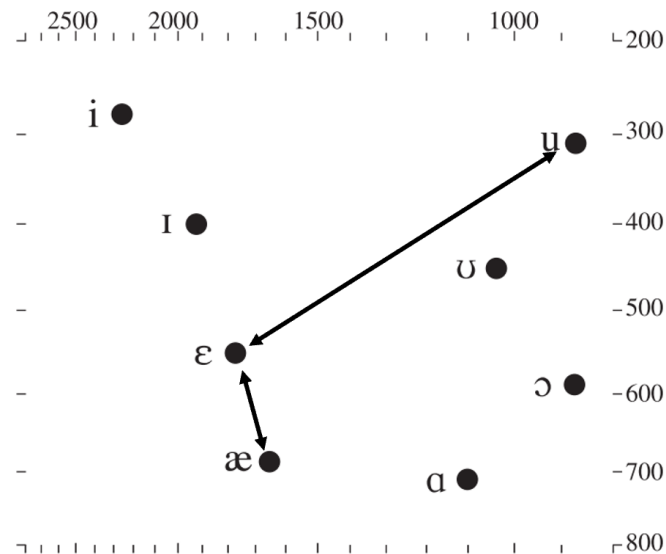


Figure 5. Formant Chart Illustrating Euclidean Distance. Adapted (lines added) from *A Course in Phonetics* by P. Ladefoged, 2010, p. 197. Wadsworth Publishing Company.

Of course, researchers do not draw and measure lines, but rather apply the Pythagorean theorem to determine the Euclidean distance between the two points using software capable of this arithmetical computation, such as Excel, SPSS, or R.

To apply this to accuracy of L2 pronunciation, one could measure the distance between a vowel as produced by an L2 speaker and the same vowel produced by a native speaker. The smaller the Euclidean distance, the more accurate the pronunciation would be. However, a single vowel will be produced with various F_1 and F_2 values by any speaker in different words and contexts, so, instead of using a single native-speaker token to compare to, it is more common to take various tokens of the same vowel produced by one or more native speakers and calculate the mean for the F_1 and F_2 , creating a centroid to which the Euclidean distances of the L2 tokens can be measured.

Various researchers have used Euclidean distance on its own, or among other metrics, to measure accuracy in the production of L2 vowels. It has been employed to measure the distance between non-native speaker (NNS) and native-speaker (NS) pronunciation of vowels (Aliaga-García & Mora, 2009; Flege et al., 1997), as well as to measure the distance between contrasting vowels (Kabakoff

et al., 2020; Mairano et al., 2019; Mairano & Santiago, 2019, 2020; Rallo Fabra, 2015).

2.3.2.3 Mahalanobis distance

A method that has seen more use in L2 research in recent years is the Mahalanobis distance. Originally devised to measure anthropometric distances between populations (Ghosh & Majumder, 1994), Mahalanobis found broader use in statistics for his method of measurement (Mahalanobis, 1936), and it has since been used widely in cluster analysis and classification.

While Euclidean distances measure the distance between two points in two-dimensional space, Mahalanobis distances measure the distance between a single point and the centroid of a distribution relative to the shape of that distribution. An illustration to clarify this can be found in Figure 6.

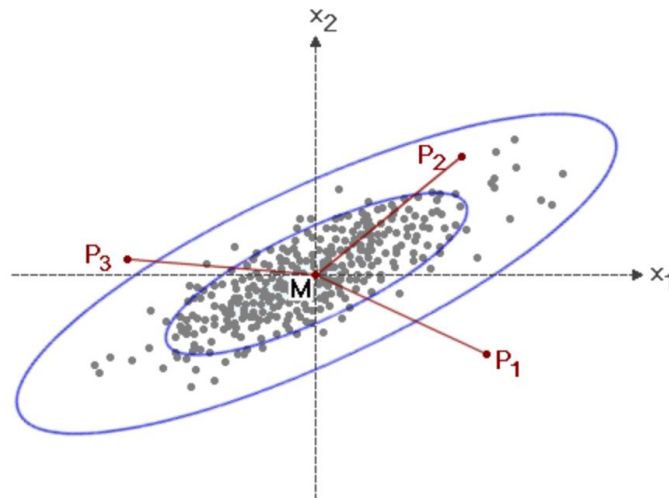


Figure 6. A Distribution with Probability Ellipses and 3 Points of Comparison.

From “Mahalanobis-Distanz” by *Grundlagen der Statistik*, 2013,

http://www.statistics4u.info/fundstat_germ/ee_mahalanobis_distance.html

In this figure, we see a cluster of grey points, blue lines showing the shape of their distribution at multivariate standard deviations, and 3 points to compare to this distribution (P_1 , P_2 and P_3). In this example, we can see that the distribution is not circular, but a flat ellipse. To see why this difference might be significant, consider two points: P_1 and P_2 . While P_1 and P_2 both share the same Euclidean distance from the centroid (M), P_2 falls between one and two standard deviations

and is much more likely to appear within the distribution. As such, it could be said to be more similar to the distribution than P_1 .

Applying this to phonetics, if these points were measurements of a spoken vowel, P_2 would be a more accurately produced vowel than P_1 , though Euclidean distances would obscure that fact. Depending on the axis of the line running from a point to the centroid, the distance will correspond with a different probability. Mahalanobis distances account for these differences, resulting in a measurement that fits with the specific distribution under consideration.

In the past decade, more studies have been conducted using Mahalanobis distances to measure pronunciation accuracy (Kartushina et al., 2015, 2016; Kartushina & Frauenfelder, 2014; Melnik-Leroy et al., 2021; Zhi & Li, 2021). In the case of Zhi and Li's (2021) research into L2 English among L1 Mandarin speakers, Mahalanobis distances showed significant correlations with perceptual measures for both intelligibility and accentedness as measured by British and American judges.

Much as with Euclidean distances, some studies have used Mahalanobis distances to measure distance between NNS productions and NS targets of vowels individually (Kartushina et al., 2015, 2016; Kartushina & Frauenfelder, 2014; Zhi & Li, 2021), while some have measured the distance between contrasting vowels (Melnik-Leroy et al., 2021).

While the advantages of Mahalanobis distances over Euclidean distances appear intuitive, there has not been a study comparing both with native-speaker judgements as of the time of writing.

2.3.2.4 Pillai Scores

Another method that has been used recently by phonetics researchers is the Pillai score, also known as the Pillai-Bartlett Trace. Proposed by K. C. S. Pillai (1955) and widely used to test hypotheses in multivariate analysis, it has recently been used in phonetics to measure vowel overlap.

While Euclidean distances only require two points to measure, Pillai scores plot several instances of both vowels in clusters and then measure the degree of overlap between them with a score of 0 to 1. A score of 0 would

indicate complete overlap whereas a score of 1 would indicate no overlap between the clusters, and scores in-between signifying varying degrees of overlap, as illustrated in Figure 7.



Figure 7. Example Plots for Different Pillai Scores. Adapted from “A Tutorial in Measuring Vowel Overlap in R” by J. Stanley, 2019.

<https://joeystanley.com/blog/a-tutorial-in-calculating-vowel-overlap> Copyright 2021 by Joey Stanley.

Pillai scores have their advantages. Because they take into account various tokens of the vowels, Pillai scores may be more ecologically valid for measuring how a speaker produces those vowels. One disadvantage of using Pillai scores is that they do not provide information on individual tokens, but only on the distribution of tokens produced by a speaker or group of speakers. This means that by-item data analysis is not possible when using Pillai scores to measure pronunciation accuracy.

Pillai scores were first and are most prominently used in phonetics to measure vowel merger and separation in a language (e.g., Hall-Lew, 2010; Hay et al., 2006). In recent years, their use to measure L2 accuracy has been promoted, particularly by P. Maraino, who has used them to evaluate how distinctly a speaker produces items of a difficult contrast in the L2 (Mairano et al., 2019; Mairano & Santiago, 2019, 2020). Pillai scores have been used in this fashion by other researchers (Perry & Tucker, 2019). Another potential application for L2 pronunciation research would be to calculate Pillai scores for the overlap between

NNSs' and NSs' productions of a target vowel. Greater acoustic differences between members of a difficult L2 contrast has been suggested to imply greater command of the L2 (Kartushina & Frauenfelder, 2014). Acoustic distinctness between members of a difficult contrast does not necessarily entail that the productions are more accurate vis-à-vis NS pronunciation. Calculating Pillai scores that measure the overlap between NNS and NS realizations may offer a better measure of target-likeness.

2.3.3 Comparing Acoustic and Perceptual Measures

Given the wide array of acoustic and perceptual methods that researchers can use to measure L2 pronunciation, one may well wonder if those measurements can be compared. While some studies employ both perceptual and acoustic measures (Birdsong, 2007; Chang & Yao, 2016; Llanes et al., 2017; Mairano & Santiago, 2019), there is much to be investigated in the relationship between them.

Munro investigated which acoustic qualities were most salient to expert judges when rating the accentedness of L1 Arabic speakers of L2 English (1993). He found that, while the L1 and L2 English speakers varied in their production of English vowels in a variety of acoustic measurements, including F_1 and F_2 , formant movement, and duration, the acoustic measures that appeared most salient to judges were the F_1 frequency and movement in the F_2 , although there was variety among the vowels.

More recently, some studies have been conducted comparing rater judgements and acoustic measures. Mairano et al. (2019) investigated correlations between several measures, including the perceptual measures of judges' ratings of intelligibility and comprehensibility, and acoustic measures, including Euclidean distances and Pillai scores measuring differentiation between vowel realizations for three vowel contrasts. Their study used French and Italian learners of L2 English. While they found significant correlations between perceptual measures and acoustic measures for nearly all vowel contrasts, they also identified the need for more research to confirm these findings. To date, there appears to be no study that has investigated the relationship between Euclidean distances, Pillai scores,

Mahalanobis distances, and perceptual measures both for differentiation of contrasting vowels and target-likeness of those vowels.

2.4 Improving L2 Pronunciation

For those learning an L2, measuring pronunciation accuracy may be interesting, but less-so than learning how to become more accurate in pronunciation. Given that L1 phonological categories can create problems in the accurate perception and production of L2 sounds, how can these difficulties be overcome? There certainly are examples of individuals who have native-like or near native-like speech in an L2 and many highly proficient bilinguals who, if not quite native-like, have learned to accurately encode, perceive, and produce L2 sounds despite these difficulties.

A belief common among many researchers (and laymen as well) is that one must start learning an L2 before a certain age to achieve native-likeness in the L2 (e.g., Johnson & Newport, 1989; Walsh & Diller, 1981). This notion, known as the Critical Period Hypothesis (CPH), has been proposed in various iterations and describing and responding to all of them is beyond the scope and focus of this thesis. Suffice it to say, more recent research on the nature of the acquisition of L2 phonology has suggested that the same mechanisms that are available for acquiring the L1 are still accessible in L2, but will be affected by the precision of L1 phonological categories, similarities and dissimilarities between the L1 and L2 phonology, and quantity and quality of L2 input (Flege & Bohn, 2021). Furthermore, research relating to the CPH focuses on native-likeness, and while successfully encoding L2 phonological categories may be a pre-requisite for reaching that end, it is also useful for L2 speakers who wish to communicate effectively in the L2 with or without achieving native-like status.

Extensive exposure and experience with high-quality, authentic L2 input has been shown to be effective at improving the perception and production of L2 sound categories (Piske et al., 2001). This is most achievable for individuals who live in a FL environment for an extended period of time and regularly engage with the target language, although shorter study abroad experiences have also been shown to lead to greater gains in L2 pronunciation and degree of foreign accent

when compared to at-home study programs (Llanes et al., 2017; Muñoz & Llanes, 2014).

For learners who study in an instructed, FL context, it is difficult to consistently access sufficient amounts of L2 input (Carlet & Kivistö de Souza, 2018; Tyler, 2019). In these cases, well developed, evidence-based interventions are particularly important.

2.4.1 High Variability Phonetic Training

One method language teachers and researchers have used to improve L2 pronunciation in instructed FL contexts is High Variability Phonetic Training (HVPT).

In the early 90's, researchers found that training difficult L2 sounds in a variety of contexts instead of isolation helps learners improve in their perception of those sounds (Logan et al., 1991). This led to the development of HVPT, a method in which exemplars of the target sounds are produced in a variety of phonetic contexts in minimal pairs of words or non-words in various perceptual or productive tasks, and participants receive immediate corrective feedback on their performance. As a technique, it has been shown to improve both perception and production of difficult sounds in the L2 (Aliaga-García & Mora, 2009; Barriuso & Hayes-Harb, 2018).

This thesis will use a selection of recordings from a study by Mora et al. (under review) in which HVPT techniques were used. Details of this study will be discussed in the Section 4, Methodology.

3. Research Questions and Hypotheses

This thesis will seek to answer three research questions related to perceptual and acoustic measures of L2 pronunciation accuracy.

RQ 1: Does the experimental, adjusted rating scale correlate with the ratings of items correctly identified with judges and the percentages of words correctly identified by raters? Does it show significant correlations with the Euclidean distances, Mahalanobis distances, and Pillai scores of NNS-NS and /æ/-/ʌ/ differences?

RQ 2: Which of the acoustic measures (Euclidean distances, Mahalanobis distances, and Pillai scores) show significant correlations with the perceptual measures (the novel rating scale, the rating scale of correct items, and percentages of correct identifications)?

RQ 3: Which of the perceptual and acoustic measures detect significant differences between pre- and post-test following a short-term HVPT of English /æ/ and /ʌ/ phonemes?

RQs 1 and 2 will be investigated through correlational analysis of the various measures, and RQ3 will be examined by comparing the main effects and interactions of the Time variable for different measures using a repeated-measure MANOVA.

4. Methodology

4.1 Participants and Training

The NNS and NS recordings were collected in a previous study as part of a Phonetics Training Conditions project (Mora et al., under review). In that study, Spanish-Catalan learners of English were divided into groups who participated in an HVPT under different conditions.

This study uses recordings from two of the groups studied. These two groups differed in only one training condition. Both were trained in silence with corrective feedback, but one group was trained using words (WD, $N = 12$) and the other with non-words (NWD, $N = 11$). One participant from the NWD group had to be excluded because of missing data relevant to the words that were examined in this study. These groups were chosen to get a sufficient amount of pre- and post-test data while maintaining as similar as possible training conditions for all participants.

Both groups joined four sessions and were tested on their ability to perceive and produce the English vowels /æ/ and /ʌ/ in a pretest by means of an ABX task, a delayed word repetition task (DWR), and a delayed sentence repetition task (DSR) in the first session. The same tasks were used as the post-test at the end of the final session.

In each session, the NNSs participated in a 30-minute phonetic training that consisted of AX discrimination (AX), identification (ID), and immediate repetition (IR) tasks. In the third session, L2 proficiency and vocabulary size were measured with an elicited imitation (EI) task and a receptive vocabulary test (X/Y Lex).

Table 1. Research Design from Mora et al. (under review)

	Week 1		Week 2	
	Session 1	Session 2	Session 3	Session 4
Pre-Test	ABX, DWR, DSR		(EI, X/Y_Lex)	
Training	AX, ID, IR	AX, ID, IR	AX, ID, IR	AX, ID, IR
Post-Test				ABX, DWR, DSR

To explore the RQs of this thesis, NNS productions of 24 words from 12 minimal pairs were selected from the DWR task (Table 2). Words were used instead of nonwords so that raters could rely on activation of phono-lexical representations for identification and rating.

Table 2. Words Selected from DWR

Back	Buck	Bad	Bud	Bag	Bug
Cap	Cup	Cat	Cut	Fan	Fun
Hat	Hut	Lack	Luck	Mad	Mud
Match	Much	Pan	Pun	Sack	Suck

To collect and calculate perceptual measures, those words were segmented, processed, and used in a task that will be described in the Section 4.3 below.

4.2 Raters

Ten NS raters were recruited to participate in the study (8 female, 2 male). All judges reported some knowledge of Spanish and time living in a Spanish-speaking country. Four raters were from the United Kingdom and 6 from the United States. All reported normal hearing.

Because of concerns about public health due to SARS-CoV-2, the raters were not asked to come to the SLA lab in person. Instead, meetings were held with each judge through video conferences lasting 15-30 minutes to provide them with instructions and address any questions they had. During this meeting, the participants were also recorded reading the same words chosen from the DWR task embedded in sentences (e.g., “I say ‘cat.’ I say ‘cat’ again.”). This was done to verify that the raters produced distinct differences between the target vowels and to look at potential effects of rater production on rater judgements, though due to time constraints acoustic analyses of these productions were not carried out.

Following the meeting, the relevant documents were shared with the raters, and they performed the tasks at their own pace and turned in their results digitally. They were also provided with text and video instructions for how to set up, perform, and share results, found in Appendix D.

4.3 Perceptual Measures of L2 Pronunciation

To measure accuracy of vowel production, an identification and rating task was created. A total of 24 CVC words from 12 minimal pairs containing the target contrast were labeled and extracted from the pre- and post-test recordings of the previously described participants. They were divided into groups of 3 minimal pairs to create four tasks for the raters (Appendix A). Because some words were minimal pairs, but not of the target contrast (e.g., “bug” and “buck”), those were separated into different tasks to assure choices were made based only on the /æ/ and /ʌ/ contrast.

Raters performed the task by running it in Praat. Words in each task were presented in a random order. The raters would select the word that best matched what they heard from a list of the six possible words on the screen. They would then rate the word for goodness of fit on a 9-point scale labeled “1=Very Bad Match” and “9=Very Good Match.” They had the option to replay each word twice. Each task included 3 practice items and 276 experimental items. An example of the task display can be found in Appendix B.

From the results, 3 different pronunciation measures were calculated. The first was an identification measure (ID), calculated as the percentage of judges who were able to correctly identify the word.

The second was the mean of the judges' ratings for items that had been correctly identified by all judges (RatCorr). These ratings were inverted so that a rating of 1 represented a very good fit and 9, a very bad fit. Of 1104 items, 416 (37.7%) were correctly identified by all 10 raters and had rating scores collected.

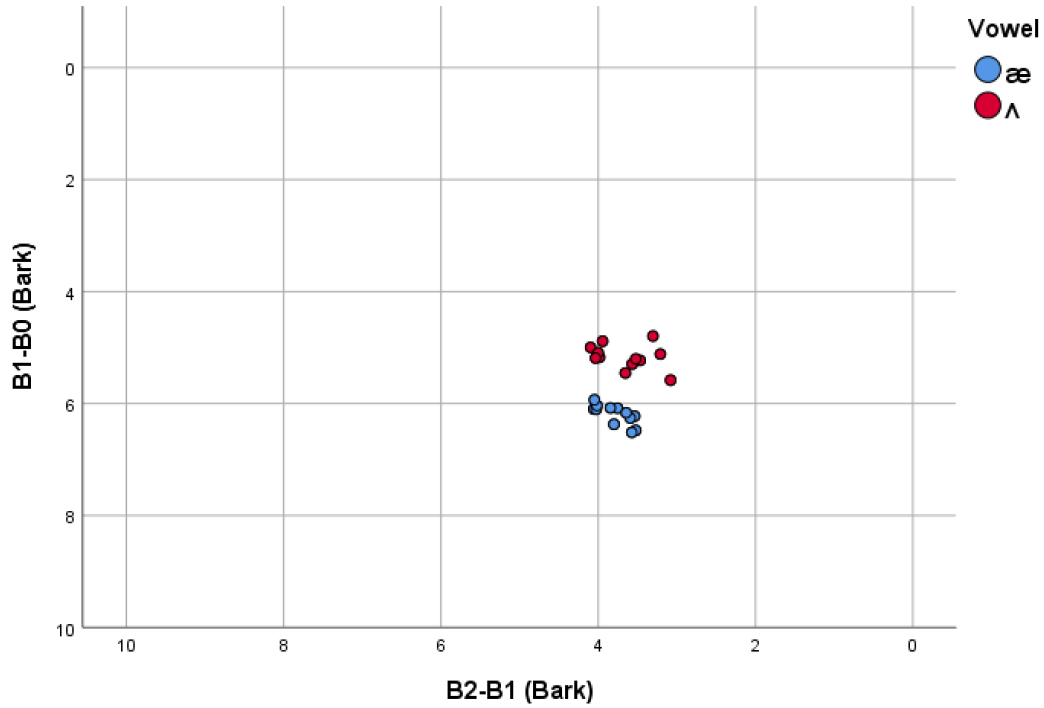
In order to retain as much data as possible an experimental scale (AdjRat) was also calculated. In this measure, the ratings for incorrectly identified words became negative rating values. Following the process described for the ratings for only correctly identified words, the scale was then inverted, becoming an 18-point scale where a lower score meant a better fit for the correct word. Appendix C illustrates this transformation.

Beyond a desire to retain and use as much data as possible, there are also reasons that this scale might provide useful insight. A vowel in an incorrectly identified word that the rater said was very bad fit with the rater's selection is likely to be in the middle ground between the two vowel categories, whereas if the rater identifies the incorrect word but feels that it was a very good fit, the vowel must be well into the contrasting category. Besides this rationale for acoustic validity, there is likely ecological validity for this kind of scale as well. A word that a listener struggles to understand is likely catch the listener's attention and elicit clarification, whereas a word that the listener believes they understand but have misidentified may be more likely to lead to misunderstandings and further confusion.

4.4 Acoustic Measures

The Acoustic data used in this study were those collected in Mora et al. (under review). In that study, time stamps and formant values (pitch (f_0), F1, and F2) were taken in Hertz (Hz) manually and extracted using a Praat script (Appendix E). They were then converted to Bark (B) measures and normalized through a Bark-difference procedure (Syrdal & Gopal, 1986). The difference in Bark between F1 and f_0 were used as an estimate of vowel height, and the

difference between F2 and F1 were used as an estimate of vowel frontness (Flege et al., 1997; Mora et al., 2015). The researchers followed the same procedure for the native-speaker productions used in with the testing material to calculate NS targets. Figures 8 and 9 visualize these data in Bark-Difference measures.



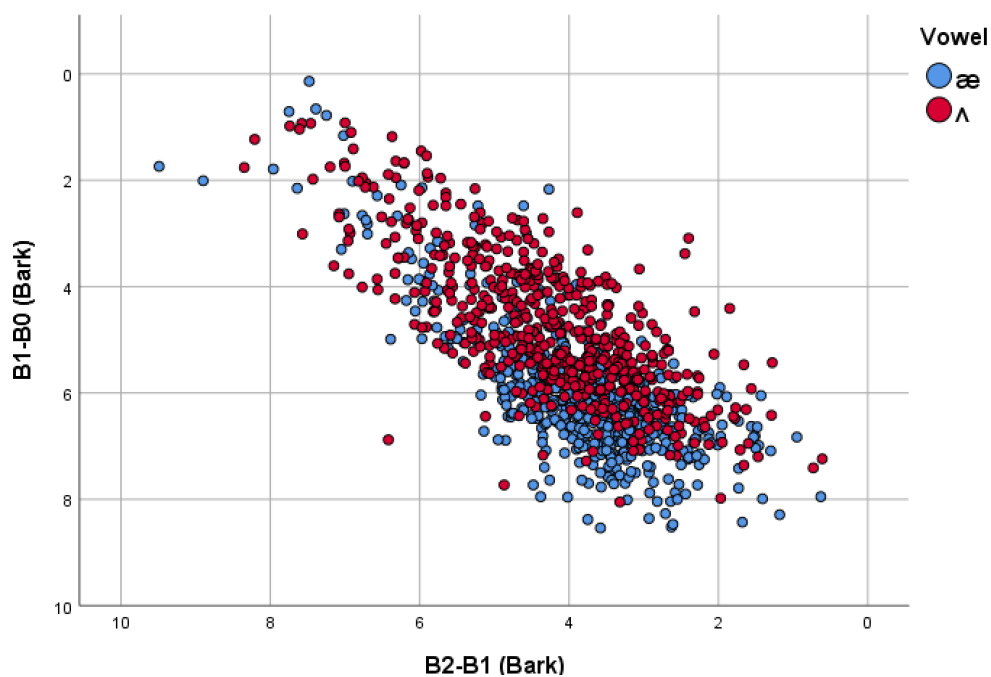


Figure 9. NNS Realizations of /æ/-/ʌ/. Values on X- and Y- axes reversed to better align with typical Formant Charts.

While Figure 9 contains all NNS productions, Figure 10 provides plots for one participant at pre- and post-test for visual clarity.

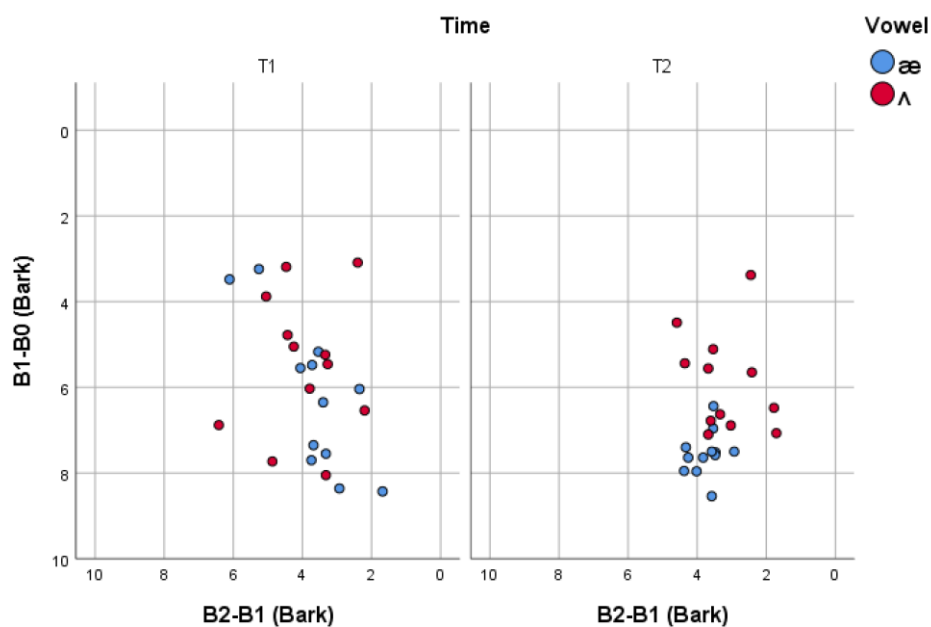


Figure 10. Realizations of /æ/-/ʌ/ for One Participant at Pre- and Post-Test. Values on X- and Y- axes reversed to better align with typical Formant Charts.

For this study, Euclidean distances, Mahalanobis distances, and Pillai scores for NNS-NS and /æ/-/ʌ/ comparisons were computed using the previously collected NNS and NS measures using an R script that is being developed by Mora and Borràs (in preparation).

5. Results

5.1 Inter-Rater Reliability

The ratings in the RatCorr variable were submitted to a reliability analysis with an intra-class correlation coefficient (ICC) using a two-way random model with a level of “absolute agreement.” This showed strong inter-rater reliability, with Cronbach’s $\alpha = 0.83$.

The same process was followed for the AdjRat variable, also resulting in strong inter-reliability, Cronbach’s $\alpha = 0.71$.

5.2 Correlations Between AdjRat and Other Measures

Correlations were computed to see if the AdjRat scale would correlate significantly with the other two perceptual measures as well as the Mahalanobis distances, Euclidean distances, and Pillai scores (RQ1). The acoustic measures were calculated both between NNS and NS productions of vowels and between /æ/-/ʌ/ productions within subjects. Pillai scores can only be calculated by-subject, so, for consistency of comparisons, all metrics are from by-subject data.

Three variables in the data were not normally distributed: ID, RatCorr, and the Pillai scores of overlap between NNS and NS vowel productions (PillaiBtwn). New variables were computed to normalize these distributions in SPSS. ID values were normalized through an *arcsine* transformation, creating the new variable IDArc, $W(92) = .986, p = 0.409$. RatCorr values were normalized with a *Log10* transformation creating the new variable RatCorrLG10, $W(92) = .987, p = .529$, as were PillaiBtwn values (PillaiBtwnLG10), $W(92) = .975, p = 0.069$. A description of all measures and their abbreviations can be found in Table 3.

Table 3. Measures and their Abbreviations

Measure	Abbreviation
Euclidean Distances, NNS-NS	EucBtwn
Euclidean Distances, /æ/-/Λ/	EucWthn
Mahalanobis Distances, NNS-NS	MahBtwn
Mahalanobis Distances, /æ/-/Λ/	MahWthn
Pillai Scores, NNS-NS (Log10-transformed)	PillaiBtwnLG10
Pillai Scores, /æ/-/Λ/	PillaiWthn
Percentages of NNS Words Correctly Identified by Raters (Arcsine-transformed)	IDArc
Ratings of Words Correctly Identified by all Raters (Log10-transformed)	RatCorrLG10
Rating scale with Ratings for Incorrectly Identified Words Calculated as Less Target-Like	AdjRat

A correlation matrix for all the variables will be used to illustrate the results, Table 4.

Table 4. Correlation Matrix of Perceptual and Acoustic measures

		Correlations								
		EucBtwn	EucWthn	MahBtwn	MahWthn	PillaiBtwnLG10	PillaiWthn	IDArc	RatCorrLG10	AdjRat
EucBtwn	Pearson Correlation		.xxx	.xxx**	.xxx	.xxx**	.xxx	.xxx	.xxx	.xxx
	Sig. (2-tailed)		-	-	-	-	-	-	-	-
EucWthn	Pearson Correlation	.189		.xxx	.xxx**	.xxx	.xxx**	.xxx**	.xxx	.xxx**
	Sig. (2-tailed)	.071		-	-	-	-	-	-	-
MahBtwn	Pearson Correlation	.618**	.050		.xxx	.xxx**	.xxx	.xxx	.xxx	.xxx
	Sig. (2-tailed)	.000	.636		-	-	-	-	-	-
MahWthn	Pearson Correlation	.199	.381**	.076		.xxx	.xxx**	.xxx	.xxx	.xxx*
	Sig. (2-tailed)	.057	.000	.473		-	-	-	-	-
PillaiBtwnLG10	Pearson Correlation	.558**	.203	.417**	.095		.xxx	.xxx	.xxx	.xxx
	Sig. (2-tailed)	.000	.052	.000	.367		-	-	-	-
PillaiWthn	Pearson Correlation	.064	.480**	-.112	.587**	.131		.xxx**	.xxx	.xxx**
	Sig. (2-tailed)	.542	.000	.288	.000	.214		-	-	-
IDArc	Pearson Correlation	-.106	.308**	-.057	.200	.034	.294**		.xxx*	.xxx**
	Sig. (2-tailed)	.314	.003	.592	.056	.745	.004		-	-
RatCorrLG10	Pearson Correlation	.062	-.054	.144	.014	.011	-.126	-.224*		.xxx**
	Sig. (2-tailed)	.558	.610	.170	.892	.917	.230	.032		-
AdjRat	Pearson Correlation	.119	-.301**	.073	-.225*	.054	-.344**	-.825**	.366**	
	Sig. (2-tailed)	.260	.004	.491	.031	.612	.001	.000	.000	

** . Correlation is significant at the 0.01 level (2-tailed).

* . Correlation is significant at the 0.05 level (2-tailed).

Note. For all correlations, $N = 92$. Significant correlations ($p < 0.05$) shown in grey.

AdjRat showed significant relationships with the other perceptual measures: a strong, negative correlation with IDArc ($r(92) = -0.825$, $p < 0.01$), and a weak, positive correlation with RatCorrLG10 ($r(92) = 0.366$, $p > 0.01$). This

is not surprising since they are all computed from the same identifications and ratings, it but does provide some small support for the hypothesis that inverting the incorrectly identified ratings to measure pronunciation accuracy did not radically change the data.

AdjRat also correlated significantly with some, but not all, of the acoustic measures. It showed significant negative, weak correlations with EucWthn ($r(92) = -0.301, p = 0.004$), MahWthn ($r(92) = -0.225, p = 0.031$), and PillaiWthn ($r(92) = -0.344, p = 0.001$). IDArc showed significant weak, positive correlations with two acoustic measures, EucWthn ($r(92) = 0.308, p = 0.003$) and PillaiWthn ($r(92) = 0.294, p = 0.004$), and the weak, positive correlation with MahBtwn approached significance ($r(92) = 0.200, p = 0.056$). Both AdjRat and IDArc showed more significant relationships with the acoustic metrics than RatCorrLG10, which did not show a significant relationship with any of the acoustic measures.

Based on this, AdjRat and ID appear to have similar correlations with the acoustic measures of the /æ/ and /ʌ/ contrast in English for these participants. They correlated significantly or approaching significantly with the same acoustic metrics and the strength of those correlations also appeared comparable. Application of the AdjRat scale with other contrasts and languages would be necessary to confirm its validity as a measure of pronunciation accuracy.

5.3 Correlations between All Acoustic Measures and Perceptual Measures

We will refer to the same correlation matrix used in the analysis of RQ1 to investigate how the NNS-NS and /æ/-/ʌ/ metrics for Mahalanobis distances, Euclidean distances, and Pillai scores correlate with the three perceptual measures (RQ2).

First, let us consider the NNS-NS measures: EucBtwn, MahBtwn, and PillaiBtwn. None of the correlations between these metrics and any of the perceptual measures met significance (all $ps > 0.17$) and all were weak (all $rs < 0.144$).

The story is different for the measures between /æ/ and /ʌ/. None of the correlations between these measures and the RatCorrLG10 metric reached significance (all $ps > 0.23$). However, EucWthn showed a significant weak,

negative correlation with RatAdj ($r(92) = -0.301$, $p = 0.004$) and a significant weak, positive correlation with IDArc ($r(92) = 0.308$, $p = 0.003$). Similarly, PillaiWthn had a significant weak, negative correlation with RatAdj ($r(92) = -0.344$, $p = 0.001$) and significant weak, positive correlation with IDArc ($r(92) = -0.294$, $p = 0.004$). MahWthn similarly showed a significant weak, negative correlation with RatAdj ($r(92) = -0.225$, $p = 0.031$) and approached significance in the weak, positive correlation with IDArc ($r(92) = -0.200$, $p = 0.056$).

In summary, none of the NNS-NS measures had significant relationships with any of the perceptual measures. The RatCorrLG10 did not significantly correlate with any of the acoustic measures. Measures of acoustic differences between /æ/ and /ʌ/ either reached or approached significant correlations with both the AdjRat and IDArc metrics. Potential interpretations of these results will be discussed in Section 6.

5.4 Pre- and Post-Test Differences

To assess how the different metrics measured the main effects of training and whether acoustic measures would detect larger effects than the perceptual measures (RQ3), a repeated-measure within-subjects MANOVA was used, with *Time* (Pre- and Post-Test) and *Vowel* (/æ/ and /ʌ/) as within-subjects variables, with all aforementioned perceptual and acoustic measures as dependent variables, with the exception of PillaiWthn. Because the PillaiWthn metric computes one score for both vowels, it was submitted a repeated-measure, within-subjects ANOVA with *Time* (Pre- and Post-Test) as the within-subjects variable.

The results will be presented in two parts, one for the NNS-NS measures and one for the /æ/-/ʌ/ measures and perceptual measures.

5.4.1 NNS-NS Measures

Table 5. MANOVA results for NNS-NS Acoustic Measures

Metric	<i>Time</i> , Main Effect, F(1,22)	η_p^2	<i>Vowel</i> , Main Effect, F(1,22)	η_p^2	<i>Time x Vowel</i> Interaction, F(1,22)	η_p^2
EucBtwn	0.099 (0.756)	.004	29.427 ($<.001$)	.572	4.675 (.042)	.175

MahBtwn	1.696 (0.206)	.072	24.319 ($<.001$)	.525	0.014 (.907)	.001
PillaiBtwnLG10	0.107 (0.747)	.005	2.603 (.121)	.106	0.017 (.898)	.001

Note. All p -values in parentheses, significant values ($p > .05$) in grey.

None of the NNS-NS acoustic metrics found any main effect of *Time* (all $ps > .206$), however EucBtwn and MahBtwn found main effects of *Vowel* ($F(1,22) = 29.471, p < .001$; $F(1,22) = 24.139, p < .001$, respectively), and EucBtwn found a significant interaction between the two ($F(1,22) = 4.675, p = .042$). For both EucBtwn and MahBtwn measures, NNSs' /æ/ productions were found to be significantly closer to NSs' than was the case for /ʌ/ ($t(45) = -4.375, p < .001$; $t(45) = -5.635, p < .001$, respectively). The *Time x Vowel* interaction in EucBtwn appears to be due to the fact that /æ/ distances slightly increased (0.04, 1.02 Bark at pre-test, 1.06 Bark at post-test), and the /ʌ/ distances slightly decreased (-0.07, 1.25 Bark at pre-test, 1.18 Bark at post-test), though neither time change reached significance (all $ps > .151$).

5.4.2 Acoustic /æ/-/ʌ/ Measures and Perceptual Measures

Table 6. MANOVA and ANOVA Results for /æ/-/ʌ/ and Perceptual Measures

Metric	<i>Time</i> , Main Effect, $F(1,22)$	η_p^2	<i>Vowel</i> , Main Effect, $F(1,22)$	η_p^2	<i>Time x Vowel</i> Interaction, $F(1,22)$	η_p^2
EucWthn	0.266 (0.611)	.012	2.555 (.124)	.104	0.221 (.643)	.010
MahWthn	4.81 (0.039)	.179	9.347 (.006)	.298	1.025 (.322)	.045
PillaiWthn*	9.641 (0.003)	.176	---		---	
AdjRat	8.285 (0.009)	.274	1.592 (.22)	.067	1.212 (.283)	.052
IDArc	8.493 (0.008)	.279	0.529 (.475)	.023	0.6 (.447)	.027
RatCorrLG10	1.531 (0.229)	.065	1.261 (.274)	.054	0.992 (.33)	.043

Note. All p -values shown in parentheses, significant values ($p > .05$) in grey.

* PillaiWthn is a single measure for both vowels, so no *Vowel* effects can be tested.

Two /æ/-/ʌ/ acoustic measures found main effects of time: MahWthn ($F(1,22) = 4.81, p = .039$), PillaiWthn ($F(1,22) = 8.285, p = .003$). MahWthn and PillaiWthn both showed significantly more distance and less overlap between /æ/ and /ʌ/ realizations (i.e., better differentiation) between pre- and post-test ($t(45) = -2.431, p = .019$; and $t(45) = -3.105, p = .003$, respectively). MahWthn also detected a main effect of Vowel ($F(1,22) = 9.387, p = .006$). This showed that /æ/ realizations tended to be closer to the distribution of /ʌ/ than the other way around ($t(45) = -2.109, p = .038$). EucWthn detected no significant main effects or interactions (all $ps > .124$).

Two perceptual measures found main effects of time as well: AdjRat ($F(1,22) = 8.285, p = .009$), and IDArc ($F(1,22) = 8.493, p = .008$). AdjRat and IDArc similarly showed lower ratings/ higher percentages (i.e., more target like production) between test times ($t(45) = 2.684, p = .01$; and $t(45) = -2.693, p = .01$, respectively).

RatCorrLG10 did not detect any significant main effects (all $ps > .229$), and none of the measures detected any significant interactions (all $ps > .283$).

6. Discussion

Based on the correlational analysis, it appears that the experimental scale, AdjRat, correlated well with the other perceptual measures as well as the acoustic measures of contrast between /æ/ and /ʌ/ (RQ1). While it did not correlate significantly with the acoustic measures of NNS-NS differences, neither did the other perceptual measures. Like IDArc, MahWthn, and PillaiWthn, it also identified a significant main effect of *Time* in the MANOVA analysis (RQ3).

With these findings, it seems that it functioned as a valid scale of NS perceptions of L2 pronunciation accuracy with these data. This should be understood with the caveat that there may be particularities to Spanish-Catalan speakers' productions of the English /æ/ and /ʌ/ that allow this measure to work that might not be true when investigating other contrasts. Before being used on its own, this measure should continue to be tested with other vowels, contrasts, speakers, and languages.

The results of the correlational analysis and MANOVA also suggest that acoustic measures that directly measure distance or overlap between NNS and NS did not compare well with perceptual measures, whereas measures of the degree of separation in NNS realizations of difficult contrasts did (RQ2, 3). It is consistent with previous studies that judges' evaluations of pronunciation accuracy corresponds with how distinctly they produce vowel contrasts (Mairano & Santiago, 2020).

What is more surprising is that none of the measures based on the raters' evaluations correlated significantly with the EucBtwn, MahBtwn, or PillaiBtwn metrics, despite the fact that those measures should indicate accuracy as in native-likeness in a similar way that the raters' goodness scale would. Mahalanobis distances between NNS and NS vowel productions in particular have been shown in other cases to correlate with perceptual ratings by Zhi and Li (2021) in their study involving multiple vowels and Chinese L1 speakers of L2 English. Let us examine a few possible reasons this might be the case.

One possible reason for this discrepancy could be that, while this study only considered one pair of vowels, their study considered 10 English monophthongs together. The MANOVA showed that MahBtwn and EucBtwn metrics detected a significant main effect of *Vowel*. It is possible that, in the study by Zhi and Li, there were sufficient differences in the NNS-NS distances and ratings between all the vowels to increase the strength and significance of the correlation. If more vowels had been included in this study, a similar correlation might have appeared, though would need further research to confirm.

Another possibility may explain both the lack of consistency between NS judgements and the NNS-NS measures as well as the consistency of those judgements with the measures of /æ/ and /ʌ/ contrast. It may be that, despite the instructions, the judges' evaluated productions not on goodness of fit but on ease of identification. Should this be the case, a participant who exaggerated the differences between the vowels would increase the distinctness of their /æ/ and /ʌ/ contrast in a way that increased ease of identity for the judges and measures of the distinctness of those vowels, even if the production of both /æ/ and /ʌ/ did not

improve in production in vis-à-vis NNS – NS comparisons. This possibility could be investigated by measuring the distance between both the target vowel and the contrasting vowel. If the distance between the item and the target vowel shows no appreciable difference, but the distance from the non-target vowel does, it could be the case that the participant has modified their pronunciation considerably, but overshot the target in a way that the NNS – NS acoustic measures used here cannot capture.

Another possibility is that the judges, who all reported some time living in Spanish speaking environments and competence with L2 Spanish, may have phonological and phonolexical representations that are activated by Spanish-accented speech as a consequence. Including raters with more limited knowledge of Spanish and exposure to Spanish-accented English and comparing their results with the other raters could shed some light on this possibility.

A final possibility is that the participants who have differentiated the /æ/ and /ʌ/ contrast, even if their F1 and F2 values haven't become more target-like, have encoded other acoustic features that successfully activate the raters' phonological representations. Other features that differentiate these vowels, such as duration, may have been produced more successfully than native-likeness of the frontness and highness of the vowels. This, like the other possibilities, would also need to be studied further to be supported, and, of course, other possibilities still exist that could explain this apparent feature of the data.

While this study has provided some new data and further avenues for research, it is not without limitations. Some of these limitations are related to only meeting with raters virtually. Because of this, the raters did not perform the tasks in a controlled environment. Differences in noise conditions, hardware, length of time spent on tasks, and more could have affected the raters' judgements. It also meant that the raters may have been less likely to reach out with immediate questions they may have had while performing the tasks, as to do so would require using email and waiting for a response.

An even larger limitation this created is that not all background data have been fully turned in to date. While this report was able to include some

background information, more data, such as detailed lengths of residency in Spanish-speaking countries and familiarity with Spanish-accented English, would have been useful.

7. Conclusion

Overall, this study showed that, even for short-term phonetic training interventions, native-speaker judgements of pronunciation accuracy appear to work well at detecting pre- and post-test differences. It was also found that these judgements correspond well with acoustic measures of how much differentiation an L2 speaker is able to produce of a difficult vowel contrast, but not necessarily with acoustic measures of how closely vowels are produced to the NS targets, though this is a topic that deserves further study with other participants, languages, and contrasts.

Furthermore, it provides some provisional evidence that, when using ID and Rating tasks with NS judges, a valid scale can be calculated using goodness ratings for misidentified words as a mirror of goodness ratings for the correctly identified words. This claim, too, requires further study to test its validity, but could offer measure that captures more nuance than using percentages of correct identification, and preserve more data than using mean ratings for correctly identified words.

As with all things, there are still many things to learn and much work to do.

8,789 Words

References

- Aliaga-García, C., & Mora, J. C. (2009). Assessing the effects of phonetic training on L2 sound perception and production. In *Recent research in second language phonetics/phonology: Perception and production* (pp. 2–31). Cambridge Scholars Publishing.
- Barriuso, T. A., & Hayes-Harb, R. (2018). High Variability Phonetic Training as a Bridge from Research to Practice, *CATESOL Journal*, 2018. *CATESOL Journal*, 30(1), 177–194. <https://eric.ed.gov/?id=EJ1174231>
- Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception. In Murry J Munro & O.-S. Bohn (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 13–34). John Benjamins. <https://doi.org/10.1075/LLLT.17.07BES>
- Birdsong, D. (2007). Nativelike pronunciation among late learners of French as a second language. In O.-S. Bohn & M. J. Munro (Eds.), *Language Experience in Second Language Learning* (pp. 99–116). John Benjamins. <https://doi.org/10.1075/LLLT.17.12BIR>
- Bongaerts, T., van Summeren, C., Planken, B., & Schils, E. (1997). Age and Ultimate Attainment in the Pronunciation of a Foreign Language. *Studies in Second Language Acquisition*, 19(4), 447–465. <https://doi.org/10.1017/S0272263197004026>
- Carlet, A., & Cebrian, J. (2019). Assessing the Effect of Perceptual Training on L2 Vowel Identification, Generalization and Long-term Effects. In M. Hejrná, A. Højen, A. Jerspersen, A. M. Nyvad, & M. Sørensen (Eds.), *A Sound Approach to Language Matters--In honor of Ocke-Schwen Bohn* (pp. 91–119). Aarhus University.
- Carlet, A., & Kivistö de Souza, H. (2018). Improving L2 Pronunciation Inside and Outside the Classroom: Perception, Production and Autonomous Learning of L2 Vowels. *Ilha Do Desterro*, 71(3), 99–123. <https://doi.org/10.5007/2175-8026.2018V71N3P99>
- Cebrian, J., Mora, J. C., & Aliaga-García, C. (2011). Assessing crosslinguistic similarity by means of rated discrimination and perceptual assimilation tasks. In K. Wrembel, M. Kul, & K. Dziubalska-Koaczyk (Eds.), *Achievements and Perspectives in the Acquisition of Second Language Speech: New Sounds 2010* (Vol. 1, pp. 41–52). Peter Lang.
- Chang, C., & Yao, Y. (2016). Toward an Understanding of Heritage Prosody. *Heritage Language Journal*, 13(2), 134–160. <https://doi.org/10.46538/HLJ.13.2.4>
- Escudero, P., Hayes-Harb, R., & Mitterer, H. (2008). Novel second-language words and asymmetric lexical access. *Journal of Phonetics*, 36(2), 345–360. <https://doi.org/10.1016/J.WOCN.2007.11.002>
- Flege, J. E. (1987). The Instrumental Study of L2 Speech Production: Some Methodological Considerations. *Language Learning*, 37, 285–296.

- Flege, J. E. (1995). Second language speech learning: Theory, findings and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 233–277). York Press.
- Flege, J. E., Aoyama, K., & Bohn, O.-S. (2021). The Revised Speech Learning Model (SLM-r) Applied. In *Second Language Speech Learning* (pp. 84–118). Cambridge University Press.
<https://doi.org/10.1017/9781108886901.003>
- Flege, J. E., & Bohn, O.-S. (2021). The Revised Speech Learning Model (SLM-r). *Second Language Speech Learning*, 3–83.
<https://doi.org/10.1017/9781108886901.002>
- Flege, J. E., Bohn, O.-S., & Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels Cite this paper Effects of experience on non-native speakers' production and perception of English vowels. *U.S.A. Journal of Phonetics*, 35294, 437–470.
- Flege, J. E., Yeni-Komshian, G. H., & Liu, S. (1999). Age Constraints on Second-Language Acquisition. *Journal of Memory and Language*, 41, 78–104.
<http://www.idealibrary.com>
- Flege, James Emil, Munro, M. J., & Mackay, I. R. A. (1995). Factors affecting strength of perceived foreign accent in a second language. *Journal of the Acoustical Society of America*, 97, 3125–3134.
- Ghosh, J. K., & Majumder, P. P. (1994). Letter to the Editor. *Annals of Human Biology*, 21(3), 287–289. <https://doi.org/10.1080/03014469400003292>
- Hall-Lew, L. (2010). Improved representation of variance in measures of vowel merger. *159th Meeting of the Acoustic Society of America/NOISE-CON*, 9, 2–10. <https://doi.org/10.1121/1.3460625>
- Hay, J., Warren, P., & Drager, K. (2006). Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics*, 34(4), 458–484. <https://doi.org/10.1016/J.WOCN.2005.10.001>
- Isaacs, T., & Thomson, R. I. (2013). Rater Experience, Rating Scale Length, and Judgments of L2 Pronunciation: Revisiting Research Conventions. *Language Assessment Quarterly*, 10(2), 135–159.
<https://doi.org/10.1080/15434303.2013.769545>
- Iverson, P., Pinet, M., & Evans. B. G. (2012). Auditory training for experienced and inexperienced second-language learners: Native French speakers learning English vowels. *Applied Psycholinguistics*, 33(1), 145–160.
<https://doi.org/10.1017/S0142716411000300>
- Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, 21(1), 60–99.
[https://doi.org/10.1016/0010-0285\(89\)90003-0](https://doi.org/10.1016/0010-0285(89)90003-0)
- Kabakoff, H., Go, G., & Levi, S. V. (2020). Training a non-native vowel contrast with a distributional learning paradigm results in improved perception and production. *Journal of Phonetics*, 78.
<https://doi.org/10.1016/J.WOCN.2019.100940>

- Kang, O. (2008). Ratings of L2 Oral Performance in English: Relative Impact of Rater Characteristics and Acoustic Measures of Accentedness. *Spain Fellow Working Papers in Second or Foreign Language Assessment*, 6.
- Kartushina, N., & Frauenfelder, U. H. (2014). On the effects of L2 perception and of individual differences in L1 production on L2 pronunciation. *Frontiers in Psychology*, 0(NOV), 1246. <https://doi.org/10.3389/FPSYG.2014.01246>
- 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. <https://doi.org/10.1121/1.4926561>
- Kartushina, N., Hervais-Adelman, A., Frauenfelder, U. H., & Golestani, N. (2016). Mutual influences between native and non-native vowels in production: Evidence from short-term visual articulatory feedback training. *Journal of Phonetics*, 57, 21–39. <https://doi.org/10.1016/J.WOCN.2016.05.001>
- Kim, H. J. (2015). A Qualitative Analysis of Rater Behavior on an L2 Speaking Assessment. *Language Assessment Quarterly*, 12(3), 239–261. <https://doi.org/10.1080/15434303.2015.1049353>
- Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nature Reviews Neuroscience* 2004 5:11, 5(11), 831–843. <https://doi.org/10.1038/nrn1533>
- Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: New data and native language magnet theory expanded (NLM-e). In *Philosophical Transactions of the Royal Society B: Biological Sciences* (Vol. 363, Issue 1493, pp. 979–1000). Royal Society. <https://doi.org/10.1098/rstb.2007.2154>
- Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science*, 9(2). <https://doi.org/10.1111/J.1467-7687.2006.00468.X>
- Ladefoged, P. (2010). *A Course in Phonetics* (6th ed.). Wadsworth Publishing Company.
- Llanes, À., Mora, J. C., & Serrano, R. (2017). Differential effects of SA and intensive AH courses on teenagers' L2 pronunciation. *International Journal of Applied Linguistics*, 27(2), 470–490. <https://doi.org/10.1111/IJAL.12151>
- Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: a first report. *The Journal of the Acoustical Society of America*, 89(2), 874–886. <https://doi.org/10.1121/1.1894649>
- Mahalanobis, P. K. (1936). On the Generalised Distance in Statistics. *Proceedings of the National Institute of Sciences of India*, 49–55. <https://doi.org/10.1007/S13171-019-00164-5>

- Mairano, P., Bouzon, C., Capliez, M., & De Iacovo, V. (2019). Acoustic distances, Pillai scores and LDA classification scores as metrics of L2 comprehensibility and nativelikeness. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *International Congress on Phonetics Sciences*. <https://hal.archives-ouvertes.fr/hal-03046802>
- Mairano, P., & Santiago, F. (2019, November). On the link between L2 learner's vocabulary knowledge and pronunciation accuracy: a corpus-based study. *Journées Internationales de Linguistique de Corpus*. <https://hal.archives-ouvertes.fr/hal-03046797>
- Mairano, P., & Santiago, F. (2020). What vocabulary size tells us about pronunciation skills: Issues in assessing L2 learners. *Journal of French Language Studies*, 30(2), 141–160. <https://doi.org/10.1017/S0959269520000010>
- Melnik-Leroy, G. A., Turnbull, R., & Peperkamp, S. (2021). On the relationship between perception and production of L2 sounds: Evidence from Anglophones' processing of the French /u/–/y/ contrast: *Second Language Research*. <https://doi.org/10.1177/0267658320988061>
- Mora, J. C., Keidel, J. L., & Flege, J. E. (2015). Effects of Spanish use on the production of Catalan vowels by early Spanish-Catalan bilinguals. In *The Phonetics - Phonology Interface* (pp. 33–54). <https://doi.org/10.1075/CILT.335.02MOR>
- Mora, J. C., Ortega, M., Mora-Plaza, I., & Aliaga-García, C. (under review). *Training the pronunciation of L2 vowels under different conditions: the use of non-lexical materials and masking noise*.
- Muñoz, C., & Llanes, À. (2014). Study Abroad and Changes in Degree of Foreign Accent in Children and Adults. *The Modern Language Journal*, 98(1), 432–449. <https://doi.org/10.1111/J.1540-4781.2014.12059.X>
- Munro, M. J. (1993). Productions of english vowels by native speakers of arabic: Acoustic measurements and accentedness ratings. *Language and Speech*, 36(1), 39–66. <https://doi.org/10.1177/002383099303600103>
- Munro, M. J. (2008). Foreign accent and speech intelligibility. In J. G. Hansen Edward & M. L. Zampini (Eds.), *Phonology and Second Language Acquisition* (pp. 193–218). John Benjamins. <https://doi.org/10.1075/sibil.36.10mun>
- Perry, S. J., & Tucker, B. V. (2019). L2 Production of American English Vowels in Function Words by Spanish L1 Speakers. *Canadian Acoustics*, 47(3), 94–95. <https://jcaa.caa-aca.ca/index.php/jcaa/article/view/3328>
- Pillai, K. C. S. (1955). Some New Test Criteria in Multivariate Analysis. *The Annals of Mathematical Statistics*, 26(1), 117–121. <https://doi.org/10.1214/AOMS/1177728599>
- Piske, T., Mackay, I. R. A., & Flege, J. E. (2001). Factors affecting degree of foreign accent in an L2: a review. *Journal of Phonetics*, 29, 191–215. <https://doi.org/10.006/jpho.2001.0134>

- Rallo Fabra, L. (2015). Can nonnative speakers reduce English vowels in a native-like fashion? Evidence from L1-Spanish L2-English bilinguals. *Phonetica*, 72(2–3), 162–181. <https://doi.org/10.1159/000430920>
- Rallo Fabra, L., & Romero, J. (2012). Native Catalan learners' perception and production of English vowels. *Journal of Phonetics*, 40(3), 491–508. <https://doi.org/10.1016/J.WOCN.2012.01.001>
- Ramus, F., Peperkamp, S., Christophe, A., Jacquemot, C., Kouider, S., & Dupoux, E. (2010). A psycholinguistic perspective on the acquisition of phonology. *Laboratory Phonology*, 10, 311–340.
- Sakai, M., & Moorman, C. (2018). Can perception training improve the production of second language phonemes? A meta-analytic review of 25 years of perception training research. *Applied Psycholinguistics*, 39(1), 187–224. <https://doi.org/10.1017/S0142716417000418>
- Southwood, H. M., & Flege, J. E. (1999). Scaling foreign accent: direct magnitude estimation versus interval scaling. *Clinical Linguistics & Phonetics*, 13(5), 335–349. <https://doi.org/10.1080/026992099299013>
- Syrdal, A. K., & Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *The Journal of the Acoustical Society of America*, 79(4), 1086–1100. <https://doi.org/10.1121/1.393381>
- Thompson, I. (1991). Foreign Accents Revisited: The English Pronunciation of Russian Immigrants. *Language Learning*, 41(2), 177–204. <https://doi.org/10.1111/J.1467-1770.1991.TB00683.X>
- Tyler, M. D. (2019). PAM-L2 and phonological category acquisition in the foreign language classroom. In *A Sound Approach to Language Matters – In Honor of Ocke-Schwen Bohn* (pp. 607–630). Dept. of English, School of Communication & Culture, Aarhus University.
- Walsh, T. M., & Diller, K. C. (1981). Neurolinguistic Considerations on the Optimum Age for Second Language Learning. In K. Diller (Ed.), *Individual Differences and Universals in Language Learning Aptitude*. Newbury House. <https://doi.org/10.3765/BLS.V5I0.2157>
- Zhi, ., & Li, A. (2021). Acoustic salience in the evaluations of intelligibility and foreign accentedness of nonnative vowel production. *Lingua*, 256. <https://doi.org/10.1016/J.LINGUA.2021.103069>
- Zwicker, E. (1961). Subdivision of the Audible Frequency Range into Critical Bands (Frequenzgruppen). *The Journal of the Acoustical Society of America*, 33, 248. <https://doi.org/10.1121/1.1908630>

Appendices

Appendix A

Words by Task

Task 1

Back	Buck	Fan	Fun	Match	Much
------	------	-----	-----	-------	------

Task 2

Bad	Bud	Cap	Cup	Lack	Luck
-----	-----	-----	-----	------	------

Task 3

Bag	Bug	Cat	Cut	Pan	Pun
-----	-----	-----	-----	-----	-----

Task 4

Hat	Hut	Mad	Mud	Sack	Suck
-----	-----	-----	-----	------	------

Appendix B

Task Display, example from Task 1

1 / 276

Click on the word you hear.
Rate how well it matches.
Click NEXT to hear the next word.

back buck fan fun match much

1=Very Bad Match **play again** 9=Very Good Match

1	2	3	4	5	6	7	8	9
---	---	---	---	---	---	---	---	---

Appendix C***Relationship between Initial Ratings, Identification, and the Experimental Adjusted Rating Measure***

Initial Rating	Identification	Adjusted Rating
9	Correct	1
8	Correct	2
7	Correct	3
6	Correct	4
5	Correct	5
4	Correct	6
3	Correct	7
2	Correct	8
1	Correct	9
1	Incorrect	10
2	Incorrect	11
3	Incorrect	12
4	Incorrect	13
5	Incorrect	14
6	Incorrect	15
7	Incorrect	16
8	Incorrect	17
9	Incorrect	18

Appendix D

Rater Instructions

Thank you for participating in our study!

Below you will find instructions on how to complete all the steps in the study. If anything is unclear or you have any questions, do not hesitate to email me at: orionbluewaltz@gmail.com

Video instructions are also available here:
https://drive.google.com/file/d/1sLwp6KcVpgNrp_XsfzCmCdAg9FXYJIKS/view?usp=sharing

Download and Install Praat

If you do not have Praat, you will need to download it. It is free to download and quick to set up.

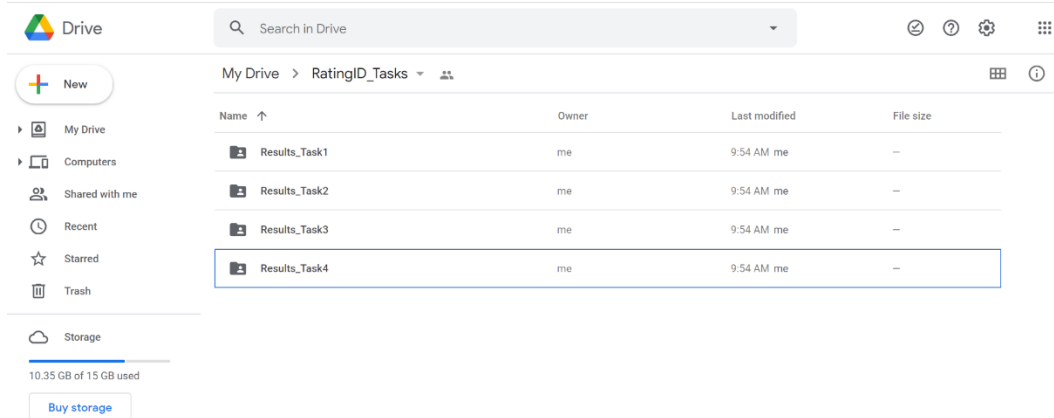
Windows:
https://www.fon.hum.uva.nl/praat/download_win.html

Mac:
https://www.fon.hum.uva.nl/praat/download_mac.html

Either the 64-bit or 32-bit edition will work for our study, so choose whichever you prefer!

Download the Rating and Identification Tasks

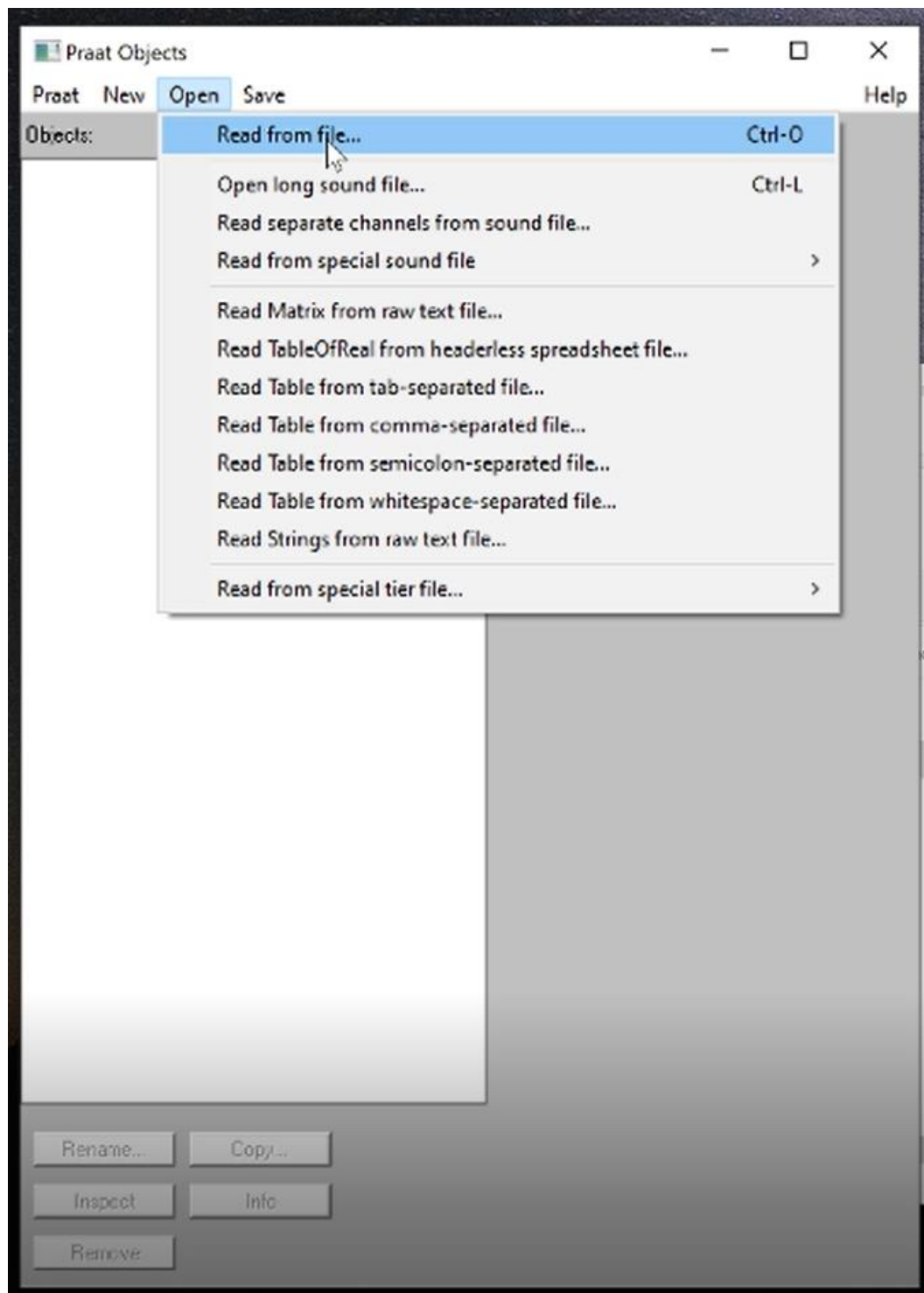
You will find four tasks in compressed folders on the study's Google Drive here:
<https://drive.google.com/drive/folders/1FKEDpgCKVM4Du3AKk55EpW9JBW8xcp77?usp=sharing>



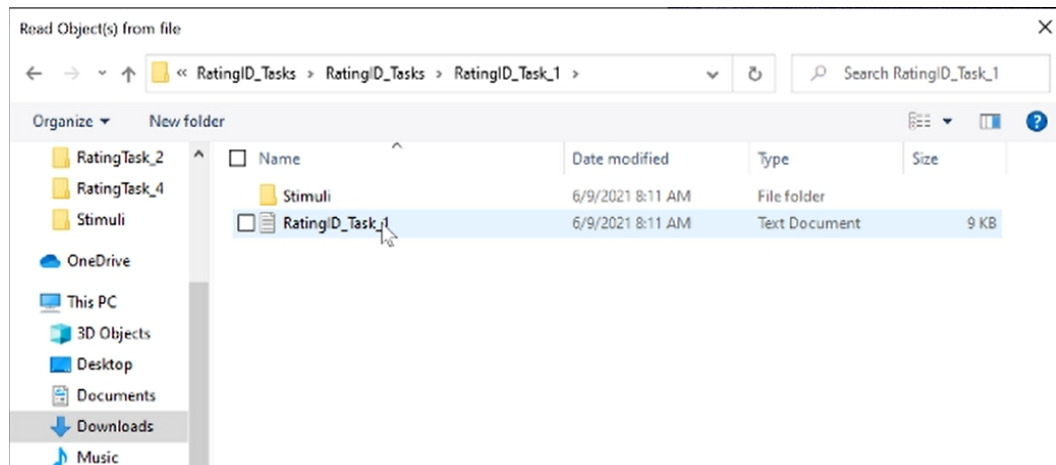
Download and extract the folders wherever is convenient for you. The task folders can be placed anywhere, but it is important that the task files and stimuli folders stay as they are now.

Open the Task in Praat

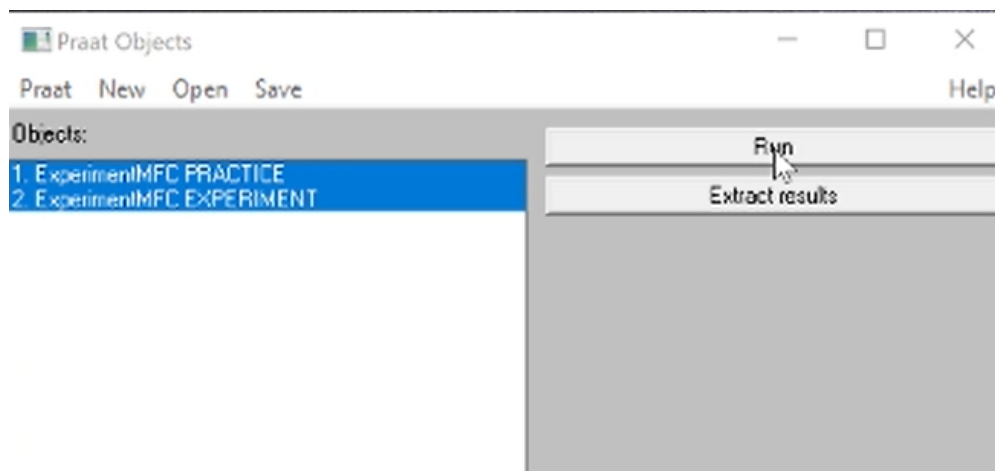
First, open the Praat application. Then, go to Open > Read from File



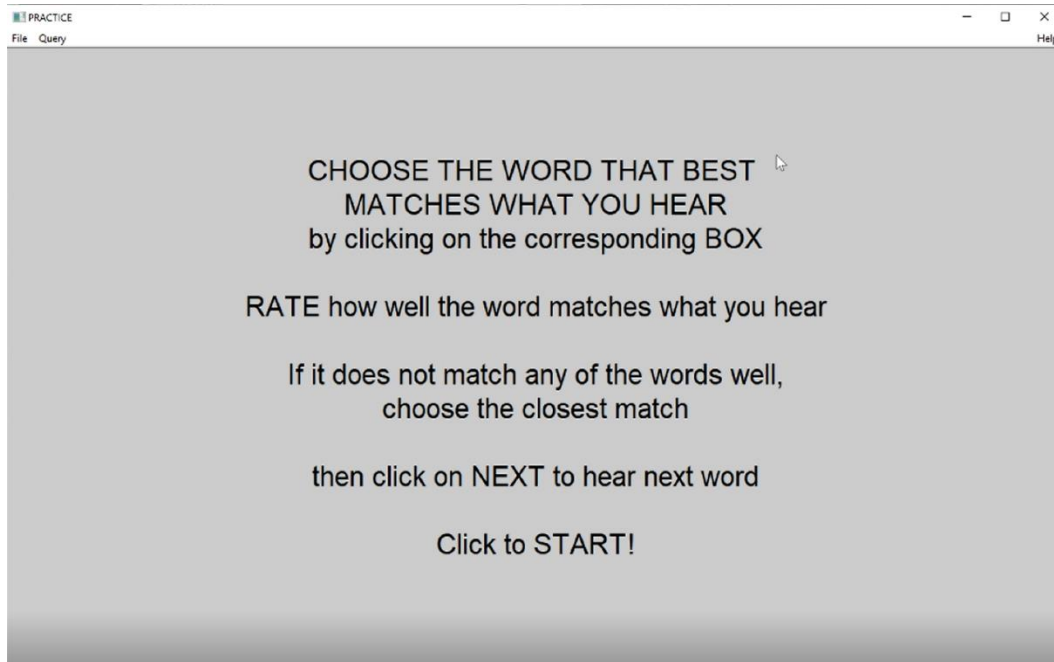
And choose the task text file



Select both the Practice and Experiment items in Praat and click “Run”

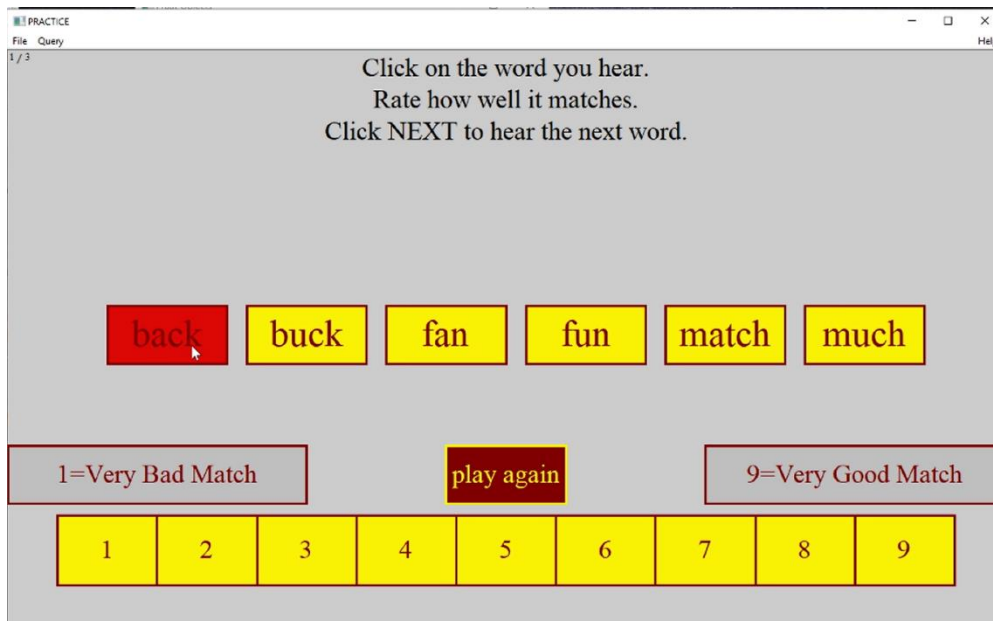


Now you should see the instructions screen



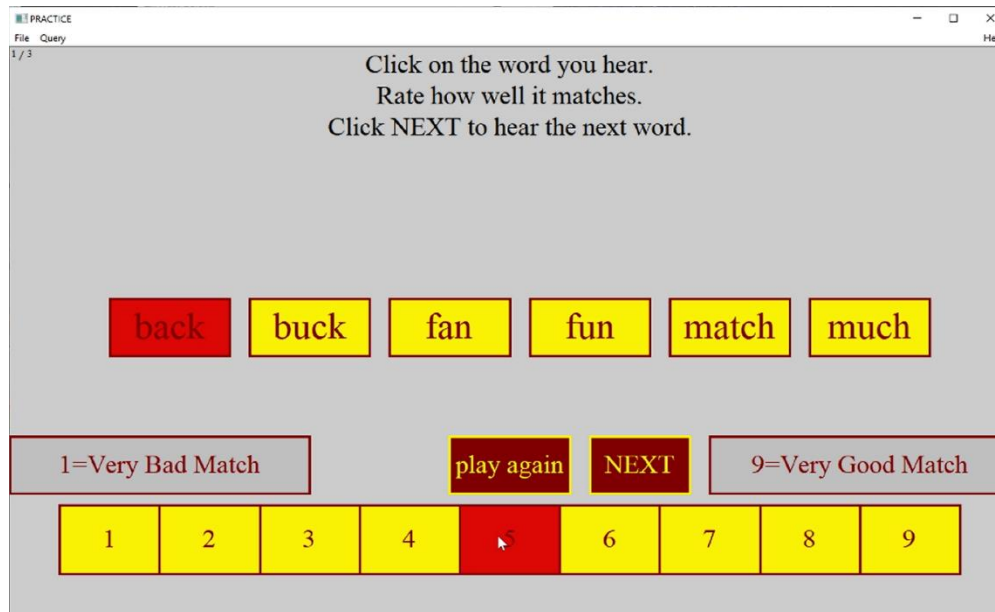
Identify and Rate the Words

Once you start the task, you will hear a word said aloud and you will see several words in front of you. First identify which of the options the word sounds like:



Next, rate how well the word you heard matches your selection. A 9 would be a word that was immediately recognizable and sounded just like the word, and a 1 would be a word that didn't match the word you selected well and you are not

confident if the selection is correct, but it is your best guess. You don't need to worry about choosing the same number as other judges. Do try to be consistent with yourself and try to use the full scale.



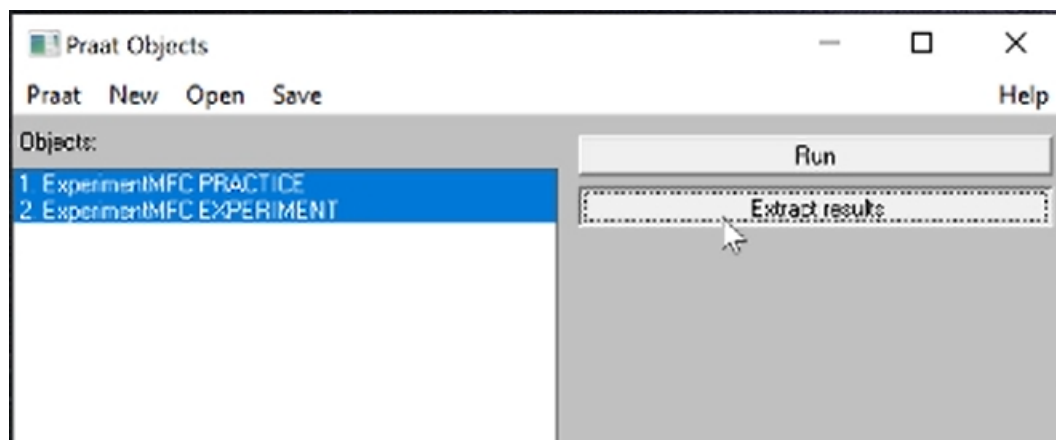
Then press “NEXT” to hear the next item.

If you want to hear a word again, you can click the “play again” button up to two times.

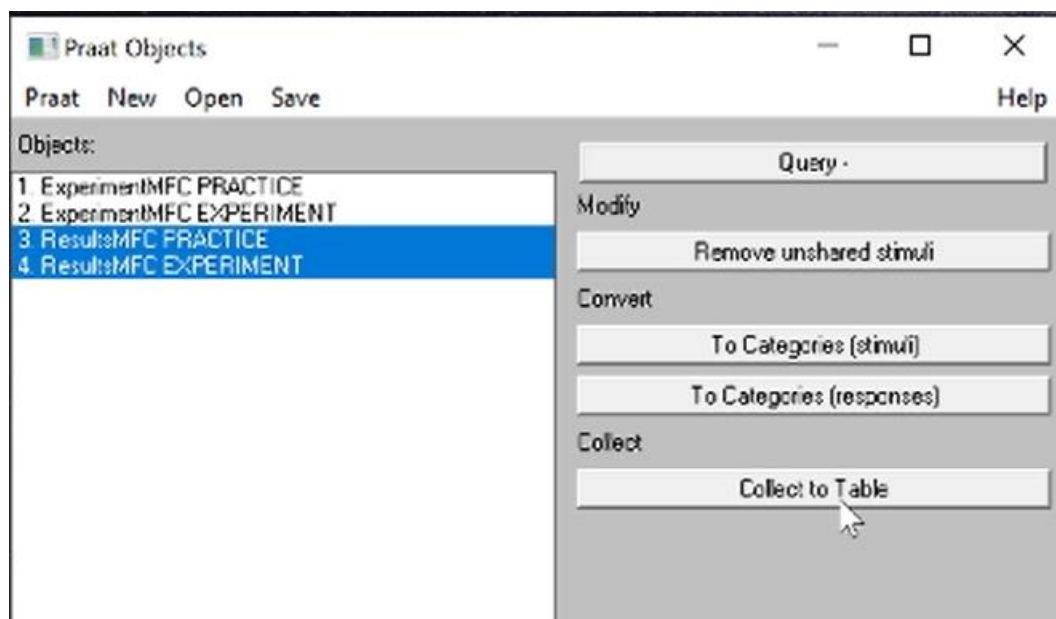
Follow this procedure for all of the words in each task without closing the task or Praat. Halfway through, you will see a prompt for an optional break, but remember to leave everything open or you will lose your progress.

Save and Upload Results

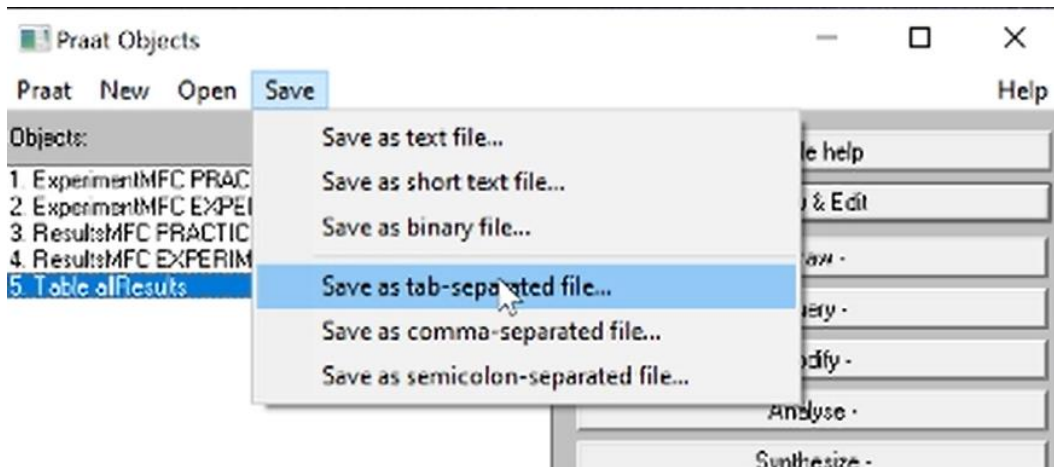
Extract the results by closing the task window, and clicking the “extract results” button in the objects window



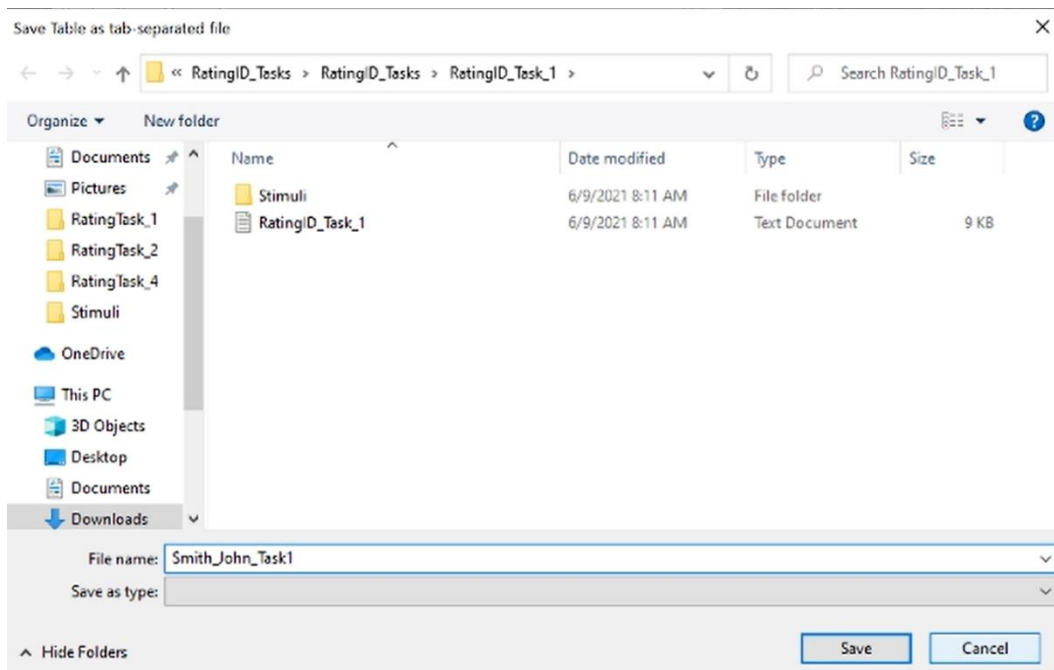
Select the results and click “Collect to Table”



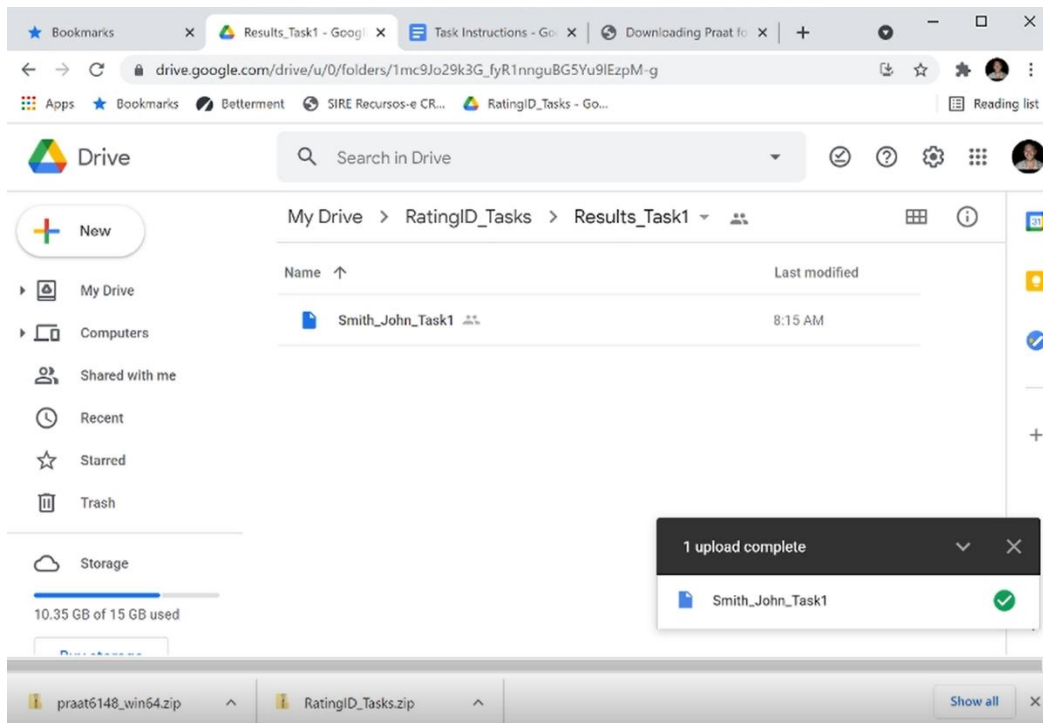
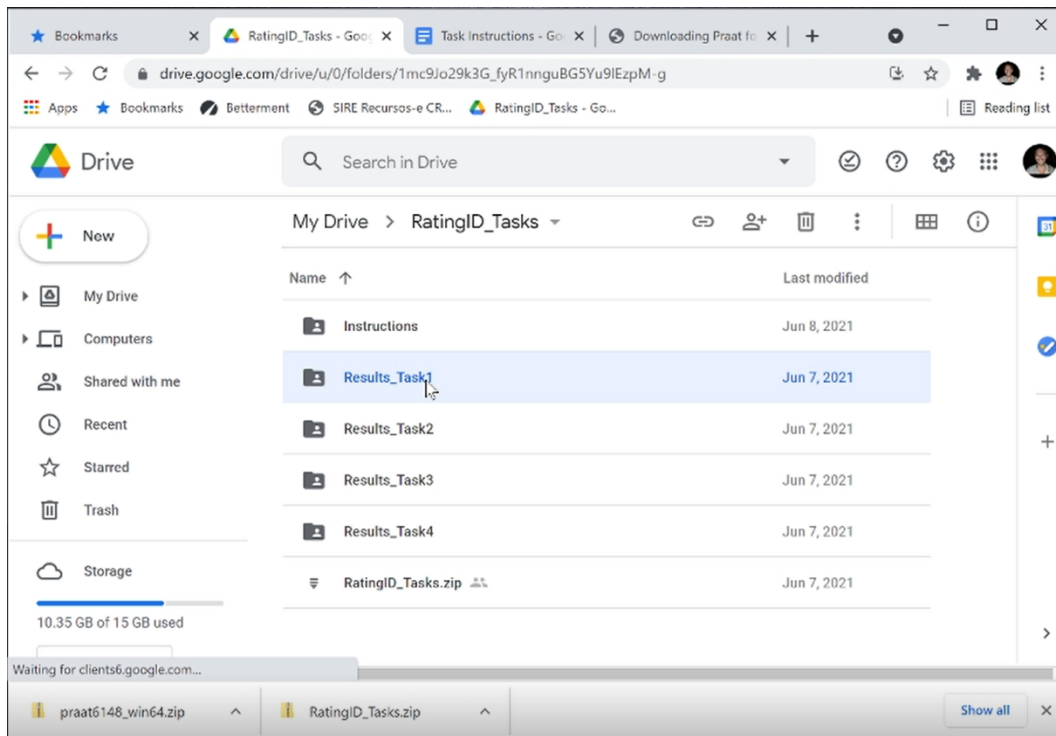
Select the table, then click Save > Save as Tab-Separated File



Label the file with your last name, first name, and the task number separated by underscores, using this format: "LastName_FirstName_Task#", e.g. "Smith_John_Task1"



Upload the file to the corresponding folder in Google Drive



Appendix E

Praat Script for Pitch, F1, F2, and Time Stamps

```
use_sound$ = selected$ ("Sound")
```

```
editor Sound 'use_sound$'
```

```
begin = Get begin of selection
```

```
end = Get end of selection
```

```
duration = end-begin
```

```
duration = duration*1000
```

```
endeditor
```

```
f$ = use_sound$ + "_Dur.txt"
```

```
fileappend "'f$'" 'f$' 'tab$' 'duration:2' 'tab$' 'begin' 'tab$' 'end' 'newline$'
```