2-2020

Phonologically-Informed Speech Coding for Automatic Speech Recognition-based Foreign Language Pronunciation Training

Anthony J. Vicario
The Graduate Center, City University of New York

Phonologically-informed Speech Coding for Automatic Speech Recognition-based Foreign Language Pronunciation Training by

Anthony Vicario

A master's thesis submitted to the Graduate Faculty in Linguistics in partial fulfillment of the requirements for the degree of Master of Arts, The City University of New York

2020

Phonologically-informed Speech Coding for Automatic Speech Recognition-based Foreign Language Pronunciation Training

by

Anthony Vicario

This manuscript has been read and accepted for the Graduate Faculty in Linguistics in satisfaction of the thesis requirement for the degree of Master of Arts.

Date	Kyle Gorman Thesis Advisor
Date	Juliette Blevins Acting Executive Officer

The City University of New York

ABSTRACT

Phonologically-informed Speech Coding for Automatic Speech Recognition-based Foreign Language Pronunciation Training

by

Anthony Vicario

Advisor: Kyle Gorman

Automatic speech recognition (ASR) and computer-assisted pronunciation training (CAPT) systems used in foreign-language educational contexts are often not developed with the specific task of second-language acquisition in mind. Systems that are built for this task are often excessively targeted to one native language (L1) or a single phonemic contrast and are therefore burdensome to train. Current algorithms have been shown to provide erroneous feedback to learners (Neri et al., 2008) and show inconsistencies between human and computer perception (Thomson, 2011). These discrepancies have thus far hindered more extensive application of ASR in educational systems.

This thesis reviews the computational models of the human perception of American English vowels for use in an educational context; exploring and comparing two types of acoustic representation: a low-dimensionality “linguistically-informed" formant representation and more traditional Mel frequency cepstral coefficients (MFCCs). We first compare two algorithms for phoneme classification (support vector machines and long short-term memory recurrent neural networks) trained on American English vowel productions from the TIMIT corpus (Garofolo et al., 1993). We then conduct a perceptual study of non-native English vowel productions perceived by native American English speakers. We compare the results of the computational experiment and the human perception experiment to assess human/model agreement. Dissimilarities between human and model classification are explored. More phonologically-informed audio signal representations should

create a more human-aligned, less L1-dependent vowel classification system with higher interpretability that can be further refined with more phonetic- and/or phonological-based research. Results show that linguistically-informed speech coding produces results that better align with human classification, supporting use of the proposed coding for ASR-based CAPT.

ACKNOWLEDGMENTS

I would like to express my gratitude to all of the following people: Dr. Kyle Gorman for his advisement and patience throughout the entire process; Dr. Gita Martohardjono and the Second Language Acquisition Lab for allowing me to run my experiment and use their space; Dr. Marie Huffman at Stony Brook University for her help in gathering and analyzing the data used in the perceptual experiment, as well as the researchers who allowed me to access the data (NSF grant IBSS-1519908 to Drs. S. Brennan, E. Broselow, A. He, M. Huffman, J. Hwang and A. Samuel of Stony Brook University); all of my participants in the experiment as well as my colleagues at the Graduate Center who have helped me in this process.

Contents

List of Tables

List of Figures

1 Introduction

This section serves to give background on past studies involving ASR in CAPT as well as explore linguistic-based speech coding for use as features in the models proposed. We overview past studies that implement ASR in an educational context and past pedagogical studies on the effectiveness of ASR in teaching phonemic awareness.

The prospect of implementing Automatic Speech Recognition (ASR) technologies in Computer-Assisted Pronunciation Training (CAPT) is not a new one. However, the unreliability of ASR feedback hinders its use in current educational programs. Current algorithms have been shown to provide erroneous feedback to learners and show inconsistencies between human and computer perception. To address this, phoneme-specific classifiers have been proposed as a fallback to these system, despite the fact that the specificity needed in building such phoneme-specific classifiers is inconvenient.

Representations of the audio signal (speech codes) used to train these classifiers are often less "phonologically-informed" than the representations we propose in this thesis, potentially losing a large amount of research that could aid in a more human-aligned L1-independent classification system. We believe these systems should degrade the same way that humans do, and we hypothesize that phonologically-informed low-dimensionality speech coding may help with this.

1.1 ASR in CAPT

CAPT systems that make use of ASR focus mainly on the evaluation or grading of L2 speech, the most popular being Goodness of Pronunciation scoring (GOP; Witt et al., 1999; Witt and Young, 2000) and its various adaptations. GOP and similar scoring calculations are usually confidence scores from a Mel frequency cepstral coefficients (MFCCs) and Hidden Markov Model (HMM)-based system. The score indicates the certainty that a target sound was pronounced correctly,

meaning a lower confidence score suggests a higher chance of mispronunciation. However, systems based on MFCCs and GOP scores often do not align with human perception. To address this, researchers have suggested combinations of GOP scoring systems and phoneme-specific classifiers to more directly target difficult contrasts for L2 learners. There exists a large amount of research that is L1-dependent as that approach has yielded a higher accuracy than L1-independent approaches. However the benefits of an L1-independent approach would be significant.

Yoon et al. (2010) propose a pronunciation error detection method for L2 learners of English that is a combination of phone-level confidence scoring and landmark-based support vector machines (SVMs). The landmark-based SVMs were implemented to target those phones in which Korean learners of English make frequent errors (i.e. /f/ ~ /p/ and /ɪ/ ~ /i/). They found that the SVMs achieved superior performance over the GOP score in those select phonemes. However, this system is limited to a single L1 and targets only certain phoneme distinctions. They discuss that with this method, any unexpected pronunciation, such as mispronunciations of different phonemes or mispronunciations of /f/ that are not /p/, the SVMs may not achieve the same performance. They suggest that the two methods are complementary but find that the combination of the two did not improve the GOP score alone.

Strik et al. (2009) also suggest phoneme-specific classification, comparing four different methods of classifying a velar contrast (/x/ ~ /k/) that is deemed difficult for L2 learners of Dutch. They found that both of the proposed phoneme-specific classifiers outperformed GOP scores. They conclude that a classifier trained on acoustic-phonetic features often outperforms one trained on MFCCs. This is noteworthy as it demonstrates the efficiency of acoustic-phonetic features in phoneme classification while further reiterating the high specificity needed in such a system.

The fallback to phoneme-specific contrast classification is overly dependant on a student's L1, making it also dependent on the availability of data or research on the possible contrasts. Furthermore, such systems cannot handle unexpected input. One could imagine a scenario in which an L1 Korean speaker learning English, expected to produce the English /ɪ/ phoneme as /i/,

produces a sound closer to /ɛ/. Since the system was not trained to handle this and therefore the feedback may not be beneficial for said student. An ideal system would better align with human perception while also being L1-independent.

On the prospect of L1-independent classification, Espy-Wilson et al. (2007) highlight the benefit of using acoustic parameters instead of MFCCs in speech recognition. They conclude that the acoustic parameters are more invariant across databases, speaker, and recording conditions and may be more invariant across languages compared to MFCCs. More recently, Wang et al. (2018) propose an approach of evaluating L2 learners goodness of pronunciation based on phone embedding and Siamese networks. A pair of acoustic feature vectors of phone segments were encoded into phone embeddings by Siamese networks. They found that Siamese networks with hinge cosine similarity outperformed the other methods in their pronunciation errors verification task. This approach appears to be able to generalize phonemes irrespective to the student's L1. Such an approach using correlates to phonological features could prove transferable to different L1/L2 combinations. This is discussed further in Section 1.3.2.

1.2 Pedagogical aspects

The creation of an ASR system for CAPT would only be advantageous if students truly benefit from its use. Past research on the use of ASR in CAPT describes possible benefits to learners, although many result in improvement only in specific contexts, such as improvements in only one to two phonemes and/or in certain participant populations. Hincks (2003) concludes that ASR in CAPT benefits only those learners with an "intrusive" accent, with no significant improvement in other learners, suggesting that ASR would only be beneficial to beginner learners. However, Thomson (2011) notes that the improvements seen in beginner learners in these experiments were not compared in improving pronunciation quality.

Other researchers outwardly describe the unreliability of ASR and question the benefits it

would have to students. Neri et al. (2008) describe the difficulty in drawing conclusions from current available research, as the systems and experimental designs vary greatly. Additionally the algorithms used in the ASR systems are rarely reported, making it difficult to understand exactly how or why the system is performing poorly on the task and how it could be improved. Neri et al. (2008) state that "[...] state-of-the-art ASR technology is known to suffer from limitations that can result in the occasional provision of erroneous feedback to the learner, possibly compromising the learning process and outcome." This idea illustrates the need to better understand the errors of the ASR system. If the main hindrance to the application of ASR in CAPT systems is indeed erroneous feedback, the exploration of a system's architecture and training data is essential to move forward.

Although ASR is promising in these CAPT systems, there remains a lack of understanding as to the reliability of the systems. It has been shown that improvements can be achieved in students' acquisition of target phonemes by building an ASR system more specifically for the task of second-language education. Therefore, the research in second-language acquisition and speech perception could help produce understandable errors that can be better addressed by researchers and teachers.

1.3 Speech Coding

MFCCs are one of the most widely used speech codings in speech recognition (Shrawankar and Thakare, 2013). In this representation, a Fourier-based short-time spectral analysis is converted to the Mel-Frequency scale to roughly approximate the frequency sensitivity of the inner ear (Schutte and Glass, 2007). MFCCs are aimed at capturing important information in a speech signal for recognition and handling as little data as necessary (Tychtl and Psutka, 1999). There are advantages and disadvantages to the use of MFCCs; the disadvantages important to this thesis are poor interpretability. We therefore explore the use of a more phonologically-informed speech code to be more interpretable and to better represent human vowel perception. Human perception is robust and, as we will show, highly accurate; we question the extent, however, to which a classification

algorithm can be trained to approximate this accuracy. For our classification tasks we choose to focus on the perception of vowels as there exists extensive research as well as the fact that in the production of sounds, vowels have been shown to be more difficult to acquire than consonants (Jin and Liu, 2014). Vowels were also found to contribute more to the intelligibility of words than consonants (Bent et al., 2007), suggesting they should be prioritized pedagogically. The overall goal is to explore the use of acoustic and perceptive features for vowel classification, making more use of linguistic knowledge for this task.

1.3.1 Formants and variation

The base of our proposed speech code are the extensively-studied vowel formant measurements. The vocal tract has certain resonances, called formants, which are peaks in the energy in the vocal tract. In vowels, the height and backness of the tongue constructs different constrictions in the filter, causing different resonances, and therefore, different vowel productions (Pickett, 1999). Formant frequencies have been shown to be the most important acoustic cue in human vowel perception (Delattre et al., 1952). Although other audio signal representations (such as MFCCs) are more common in ASR, the use of formant measurements for vowel classification is seen often in landmark-based ASR research. Evanini et al. (2009) discuss the use of formants in ASR, stating:

> Despite the fact that other representations, such as MFCCs, are commonly used in ASR tasks, formants continue to prove useful to phoneticians because of their low dimensionality, their correspondence to articulatory gestures, their resistance to transmission channel effects, and their ability to characterize phonologically relevant vowel distinctions. (p. 1655)

This suggests that formants, with a deeper linguistic understanding of the representation, would prove useful in the task of vowel classification would prove especially useful for foreign-language pronunciation training. Formant frequencies for the same vowel, however, vary with speaker age,

sex, height, and with context, speech rate, stress, and more. There exist many vowel normalization techniques to computationally address speaker-intrinsic characteristics. Adank et al. (2004) summarize and compare twelve normalization methods with the goal of evaluating their efficiency in categorizing sounds into phonemic categories. They found that Lobanov's 1971 z-score normalization (Lobanov, 1971) performed best at preserving phonemic information and minimizing anatomical/physiological variation. Fabricius et al. (2009) also compare these methods with interest in the evaluation of visual cross-speaker mapping of vowel means with the "S-centroid" method proposed in Watt and Fabricius (2002). They found that Lobanov's z-score was "the most successful technique with regard to improving overlap and optimizing area ratios between pairs of speakers". Following these two findings we implement Lobanov's z-score normalization for our data.

To further address the variation of formants and to deter any context bias we consider the tenets of High Variability Pronunciation Training (HVPT) theory. This involves training language learners using a wide variety of contexts and speaker voices. HVPT has been shown to aid second-language-learners in vowel and consonant perception (Lambacher et al., 2005; Nishi and Kewley-Port, 2007, 2008). With this same idea, a classifier trained on many contexts should perform better at perception as more contexts are learned. This adds to the fact that computationally, more training data is beneficial to the training of a system. However, as the human perceptual system itself does not rely solely on formant frequencies, further features (outlined in section 1.3.2) are required to better model this system.

1.3.2 Acoustic correlates to distinctive features and landmark-based ASR

In order to encode phonological categories based on the formants discussed above, we must explore acoustic cues that correlate with said categories. For this goal, we use the well-established binary 'distinctive features'. In phonological theory, distinctive features are a set of features where each phoneme has a distinct set of binary values assigned to each feature (Jakobson et al., 1951),

where changing the value of one feature can potentially change the word. For example, /t/ and /d/ are distinguished by the feature [VOICE] where /t/ is [−VOICE] and /d/ is [+VOICE]. This feature therefore distinguishes between the two words /sit/ 'seat' and /sid/ 'seed'. The distinctive features are separated into three categories: (1) **Manner**: related to the configuration of the vocal tract, (2) **Place**: related to where the main constriction is located, (3) **Source**: related to the glottis/vocal folds (Pickett, 1999). We assume these features would provide the most robust classification for vowels as all vowels hold a unique set of binary features. As such, there exists ample literature on landmark-based speech recognition with the goal of using these features for phoneme classification. Features are not exhaustively considered as binary, though in this context we assume here that feature encoding is indeed binary. Meng et al. (1991) explore the use of distinctive features in ASR and report that the distinctive feature representation gives similar performance to direct vowel classification, with distinctive features possibly offering a more flexible mechanism for describing context dependency. They also find that the use of acoustic features can significantly reduce run-time computation for vowel classification with no cost to accuracy. Though they also expresses the lack of understanding between distinctive features and measurable acoustic representations, suggesting a need to better understand this connection for its use in classification.

Syrdal and Gopal (1986) propose use of the Bark scale for phonological feature mapping from acoustic properties, a scale originally proposed in Zwicker (1961) that is meant to more properly relate to the physical manner in which humans perceive sound. Zwicker proposed that an empirically defined critical band scale should be adapted to a standard tonality scale, the Bark scale, dividing the human auditory range below 16 kHz into 24 critical band units, or Barks (Syrdal and Gopal, 1986). They found that, computationally, vowel classification based on the Bark scale was significantly more accurate than classification based on unnormalized Hertz data (formants). They proposed that the critical Bark scale is able to construct a binary feature matrix similar to that of distinctive features. The F1−F0 dimension is meant to replicate the [HIGH] feature while the F3−F2 dimension is meant to replicate the [FRONT] feature for American English vowels, where a

value less than 3 critical bands represents the feature [+] and a value greater than 3 critical bands represents [−]. Table 1 shows their proposed feature matrix for ten American English vowels. We question the use of both the difference values as a continuous dimension and the binary features based on these critical value of 3 proposed.

Vowels	F1-F0 <3 Bark	F2-F1 <3 Bark	F3-F2 <3 Bark	F4-F2 <3 Bark	F4-F3 <3 Bark
i	+	-	+	+	+
ɪ	+	-	+	-	+
ɛ	-	-	+	-	+
æ	-	-	+	-	+
ɝ		-	+	-	-
ʌ	-	-	-	-	+
ɑ	-	+	-	-	+
ɔ		+	-	-	+
ʊ	+	-	-	-	+
u	+	-	-	-	+

Table 1: Vowel classification based on critical distance features in five Bark-difference dimensions after Syrdal and Gopal (1986).

They further discuss that the critical bark dimensions proposed may not transfer across languages. However, they should be sufficient in categorizing attempted productions of American English vowels.

Slifka (2006), in designing a landmark-based ASR system, used the F1-F0 Bark measure to encode vowel height, reporting 76% recognition accuracy for the feature [HIGH] and 78% for the feature [Low] with the F1-F0 Bark dimension as the only cue. She further encodes features with different approaches, using F1 and F2 slopes as a cue for the feature [TENSE]. F1 is expected to decrease in [+TENSE] vowels as the articulators move to a narrow constriction and [−TENSE] (lax) vowels are expected to show an off-glide toward a neutral vocal tract, measured by F2 slope (Slifka, 2003).

The use of these acoustic parameters and correlates to distinctive features could better represent

the human phonological system and aid in classification. The calculations and phonological feature mappings discussed here are further discussed in Section 2.2.1.2.

2 Modeling

In this section we test and compare the use of the previously-discussed landmark-based features and MFCC features as training for the task of vowel classification. We compare the performance of three models and discuss possible shortcomings. Section 4 later compares the models to the human perception experiment.

2.1 Data

The models were trained on American English vowel productions from the TIMIT Acoustic-Phonetic Continuous Speech Corpus (TIMIT; Garofolo et al., 1993). The TIMIT corpus contains 5.4 hours of recordings, consisting of 630 speakers of eight major dialects of American English. Each speaker read ten phonetically diverse sentences which were time-aligned with orthographic, phonetic, and word transcriptions. The corpus contains test and training subsets, balanced for phonetic and dialectal coverage. The training and test sets contain 4,620 and 1,680 utterances, respectively. Sex of the speaker is contained as metadata that was aligned with each utterance.

The models were trained on ten monopthongal vowels, collapsing allophonic [ʉ] `ux` and [u] `uw` to phonemic /u/ `uw`, removing diphthongal American English /eɪ/ `ey` and /oʊ/ `ow`, and including schwa /ə/ `ax`.

A total of 35,981 training vowels were extracted from the TIMIT corpus. The vowels used are outlined in Table 2 in ARPABET and IPA. The number of data points for each vowel is included. The class imbalance, mainly seen in the lack of /ʊ/ `uh`, is discussed in further sections. This thesis will henceforth represent all vowels using the IPA. Training, Testing, and Development sets were split into 80%, 10%, 10% respectively for all classifiers.

ARPABET	IPA	Count
iy	i	5,214
ih	ɪ	4,532
eh	ɛ	3,612
ae	æ	2,420
aa	ɑ	2,477
ah	ʌ	2,475
ax	ə	3,924
ao	ɔ	2,089
uh	ʊ	565
uw	u	1,476

Table 2: Training vowels.

2.2 Preprocessing

2.2.1 Landmark-based features

The first four formant frequencies (F1, F2, F3, F4), their bandwidths, and vowel duration were extracted from each vowel utterance using Praat (Boersma and Weenink, 2009), a program for performing acoustic analysis. The formant and bandwidth measurements were extracted at three different time points in the vowel (10%, 50%, and 75% of the duration). Formant measurements returned as "unidentified" were replaced with the median formant measurement of the vowel class. Following Evanini et al. (2009), bandwidths were converted to log scale to make the distribution closer to Gaussian.

To address the natural variation in formant frequencies, the vowel space of a speaker must be normalized before measurement-based assumptions can be made. Following Adank et al. (2004), Lobanov's z-score transformation is used as it performs best at normalization while maintaining the most vowel information. Lobanov's z-score is meant to normalize while maintaining natural sociolinguistic variation. The transformation was calculated for the three formants. Lobanov's

equation can be seen in equation (1):

$$F_{ti} = \frac{F_{ti} - \mu_{ti}}{\delta_{ti}} \tag{1}$$

where μ_{ti} is the average formant frequency across the vowels for speaker t and δ_{ti} refers to the standard deviation for μ_{ti}. To reiterate, the standard deviation here is computed per-speaker. Speaker sex was also included as a feature in modeling to further address individual variation.

As discussed in section 1.3.2, the Bark scale attempts to replicate the feature matrix that phonological distinctive features provides. The critical Bark scale here is replicated as best as possible using the approximation proposed in Zwicker and Terhardt (1980), which is meant to correspond more accurately to the proposed critical band scale (Syrdal and Gopal, 1986). The critical band value in Bark is calculated as:

$$B = 13 \arctan 0.76f + 3.5 \arctan \frac{f}{7.5}^2 \tag{2}$$

The Bark scale was modified to that in Traunmüller (1981) to correct formants on the low-frequency end. Following the corrections given in Syrdal and Gopal (1986), formant frequencies below 150 Hz are raised to 150 Hz, frequencies between 150 Hz and 199 Hz are corrected with the formula:

$$f_c = f - 0.2\,(f - 150) \tag{3}$$

and frequencies between 200 Hz and 250 Hz with the formula:

$$f_c = f - 0.2\,(250 - f) \tag{4}$$

Critical bark values were calculated using the corrected formant output for all formant values extracted (*Bark Values*). Five bark-difference measures were calculated: Bark 1 (F1−F0), Bark 2 (F2−F1), Bark 3 (F3−F2), Bark 4 (F4−F2), Bark 5 (F4−F3). The critical distance of 3 Bark was

used to create a binary label for each Bark-difference measure (*Bark Binary*).

For the [TENSE] feature, F1 and F2 slope are calculated as the difference in F1 and F2 at the 10% and 75% measures for the vowel.

The segments before and after the vowel are also encoded as features to retain context information. The segment (consonant, vowel, silence) that occurs before and the segment that occurs after the vowel is encoded as 'before' and 'after', the 'context' features.

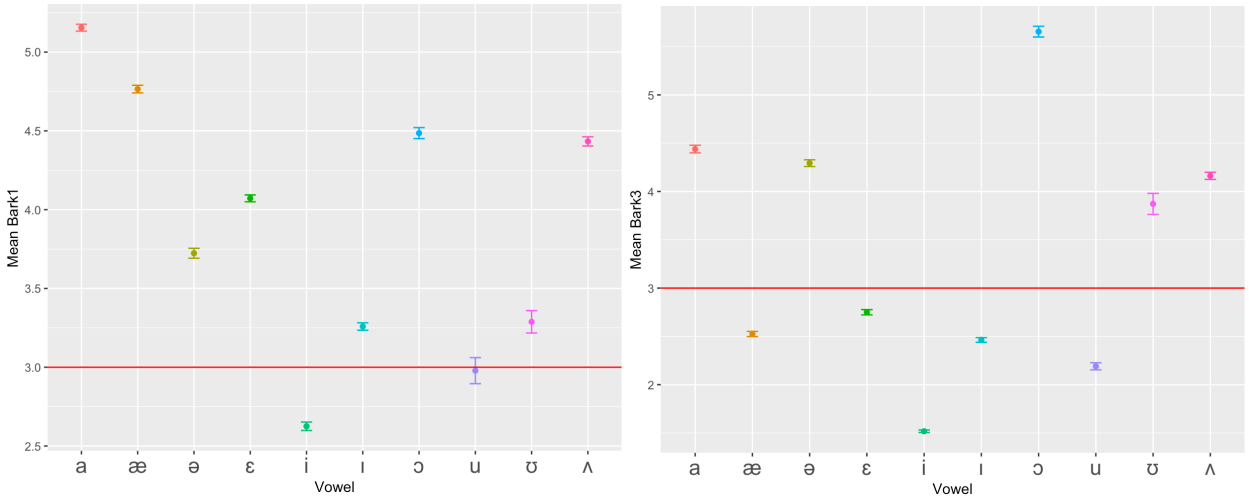
2.2.1.1 Feature exploration

In this section we explore the discriminability of TIMIT vowels in the Bark1 and Bark3 dimensions. Bark1 values by vowel are graphed in Figure 1a. The critical value of 3 is shown with a red line. We can see that the Bark1 dimension is able to replicate vowel height quite well in terms of the height continuum. Higher vowels have a lower Bark1, seen in /i/ having the highest value and /a/ having the lowest value. A one-way ANOVA showed a significant effect for vowel on Bark1 value, $F(9, 35971) = 2577$, $p < .01$. A Tukey post-hoc test at the 99% CI shows all but two vowel pairs as significantly different; those two pairs being /ʊ/ and /ɪ/ and /ɔ/ and /ʌ/, which is expected.

Vowels below the critical band of 3 are suggested to represent the [+HIGH] feature. We see that the only vowel definitively meeting this requirement is /i/, with the mean of /u/ exactly on the threshold. Additionally, the two high-lax vowels (/ɪ/ and /ʊ/) are not correctly identified as [+HIGH]. The data here would suggest a critical distance of approximately 3.5 if a binary feature is to be constructed. Fahey et al. (1996) also found that the critical distance of 3 is discriminating more the tense/lax distinction for the high vowels. As this may not extend to L2 vowel productions, we choose to encode the Bark1 as a continuous height value as well as construct the binary feature. In this way, height will be better encoded. This continuum will also better classify mid vowels.

The Bark3 dimension, meant to represent vowel frontness, is graphed in Figure 1b. We see that the front vowels are as expected all below the critical value of 3, indicating [+FRONT]. A one-way ANOVA showed a significant effect for vowel on Bark3 value, F(9,35971) = 6481, p < .01.

A Tukey post-hoc test at the 99% CI shows all but the /ɪ/ and /æ/ pair as significantly different. However, /u/ is clearly identified as [+FRONT] with a mean Bark3 of 2.19. It could be that this dimension is picking up the small constriction in the front of the mouth due to lip rounding during the production of /u/. It is important to note that fronting of /u/ before coronals in North American Dialects is very common and is annotated by TIMIT as `ux`. When removing this vowel from analysis, /u/ remains front under this dimension.



(a) Mean Bark1 by vowel (height).

(b) Mean Bark3 by vowel (frontness).

Figure 1: Bark dimensions by vowel for TIMIT.

2.2.2 MFCCs

Data was preprocessed using the Python Speech Features library[1] to extract 12 MFCCs from 26 filter-bank channels. The log-energy was also extracted from the signal, producing a feature vector of 25 coefficients per frame. The step size was 10 ms and the window size was 25 ms.

In attempt to retain contextual information in both directions, the training data take the shape XVX where V is the vowel of interest and X is any segment (consonant, vowel, silence) that occurs before or after the vowel. For example, Figure 1 shows the word 'steels' and its alignment extracted from a sentence in the TIMIT corpus.

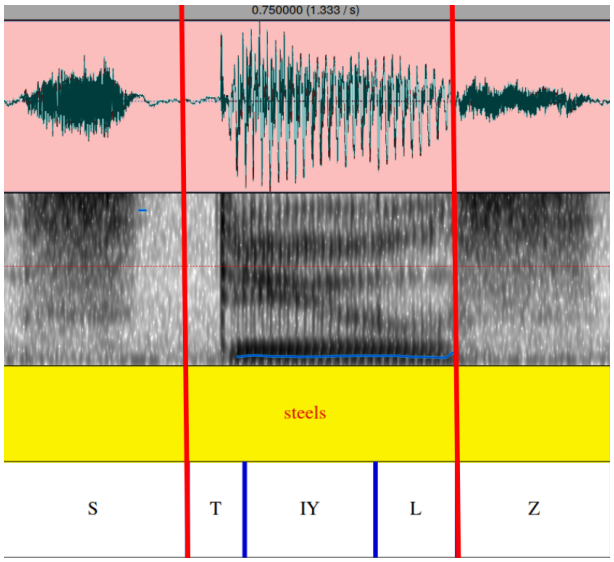


Figure 2: XVX context example.

With the target vowel being /i/ `IY` in this instance, the segment directly before and the segment directly after `IY` ('T' and 'L') are included in the extraction of the vowel, shown by the red lines. The audio segment within the red lines is extracted and labeled as `IY`. This process is repeated for all target vowels in the corpus.

The output of the model is the probability of each phoneme given the input signal. The

[1] `https://github.com/jameslyons/python_speech_features`

phoneme with highest probability was recorded as the model's prediction.

2.3 Models

In this section we introduce two phoneme classification techniques, support vector machines and long short-term memory recurrent neural networks. We discuss the models and features used to train the models in this experiment. The following section evaluates the performance of these classifiers on testing.

2.3.1 Support vector machines

Two models were trained using support vector machines (SVMs); one model trained on landmark-based features (SVM-PHONO) and one model trained on MFCC features (SVM-MFCC). All SVM models were trained using Scikit-learn (Pedregosa et al., 2011).

Support vector machines have been argued to show excellent generalization properties on vowel classification tasks (Wang and Paliwal, 2003). As per Juneja and Espy-Wilson (2008), SVMs outperform midden Markov models for phonetic feature detection (Niyogi et al., 1999; Keshet et al., 2001) and for phonetic classification from hand-transcribed segments (Clarkson and Moreno, 1999; Shimodaira et al., 2001).

The principle idea of SVMs (Cortes and Vapnik, 1995; Vapnik, 2013) is to find the ideal hyperplane that separates classes in some feature space. In attempting multi-class classification, the task becomes non-linear. As the classes are not separable in the feature space, the input vectors are mapped into a high-dimensional feature space through a non-linear mapping. A "linear decision surface" is then constructed in that space (Cortes and Vapnik, 1995). Here we use the radial basis function (RBF) kernels for this mapping.

RBF kernels requires two hyperparameters, C and γ. C trades off misclassification of training examples against simplicity of the decision surface. γ defines how much influence a single training

example has. The larger γ is, the closer other examples must be to be affected. We use Sklearn's grid search function to choose the best combination of these parameters. The combination that performs the best is a C of 1000 and a γ set to be the inverse of the product of the number of features and the variance of the observations (set using '`gamma=scale`').

The multiclass classification problem here must be reduced to multiple binary classifications. A vital choice in designing the classifier lies in the decision of the method of doing so. We consider the one-versus-all (OVA) and one-versus-one (OVO) methods.

The OVA method fits one classifier per class. For each classifier, the class is fitted against all the other classes. The OVO method constructs one classifier per pair of classes. The class which received the most votes is selected. In the event of a tie, the class with the highest aggregate classification confidence is selected by summing over the pair-wise classification confidence levels computed by the underlying binary classifiers (Pedregosa et al., 2011). In terms of complexity, OVO is usually slower than OVA, as it requires the fitting of $\frac{C(C-1)}{2}$ separate classifiers. However, OVO classifiers have been shown to outperform OVA classifiers in various tasks (Allwein et al., 2000; Weston et al., 1999). We therefore choose to use the OVO method. To address vowel class imbalance, class weights were set to inverse class frequencies (`class_weight=balanced` in Scikit-learn).

2.3.2 Long short-term memory recurrent neural networks

One model was trained using long short-term memory (LSTM) recurrent neural networks and MFCC features (LSTM-MFCC). The model was trained using PyTorch (Paszke et al., 2017). We attempt to replicate a model for phoneme classification described in Graves et al. 2005. The classification in the present experiment, however, is a ten-way classification task. Although either speech encoding can be used to train an LSTM, we choose to use the MFCC encoding only as to replicate the setup in Graves et al. 2005. The future work section discusses the training of an LSTM on the proposed PHONO features.

LSTMs (Hochreiter and Schmidhuber, 1997) are recurrent neural network variants which can store and retrieve information over long time periods with explicit gating mechanisms and a built-in constant error carousel. LSTMs are effective models for applications involving sequential data (Karpathy et al., 2015) including speech data. Here, the bidirectional LSTM (BLSTM) architecture for phoneme classification proposed in Graves et al. (2005) is replicated as closely as possible.

Equation 5 shows the function computed for each element in the input sequence for the LSTM:

$$\begin{aligned} i_t &= \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{(t-1)} + b_{hi}) \\ f_t &= \sigma(W_{if}x_t + b_{if} + W_{hf}h_{(t-1)} + b_{hf}) \\ g_t &= tanh(W_{ig}x_t + b_{ig} + W_{hg}h_{(t-1)} + b_{hg}) \\ o_t &= \sigma(W_{io}x_t + b_{io} + W_{ho}h_{(t-1)} + b_{ho}) \\ c_t &= f_t * c_{(t-1)} + i_t * g_t \\ h_t &= o_t * \tanh c_t \end{aligned} \tag{5}$$

where h_t is the hidden state at time t, c_t is the cell state at time t, x_t is the input at time t, $h_{(t-1)}$ is the hidden state of the layer at time $t-1$ or the initial hidden state at time 0, and i_t, f_t, g_t, o_t are the input, forget, cell, and output gates, respectively. σ is the sigmoid function, and * is the Hadamard product (Paszke et al., 2017). We chose to use the Adam optimization function (Kingma and Ba, 2014). Loss is calculated using negative log likelihood. For the output layer, a softmax activation function was used.

LSTM-MFCC: Experimental models tested models with 1 to 3 layers, hidden units of 30 to 93, LSTM with and without a dropout rate of .2 to .3, and a learning rate of .001 to .0003. The BLSTM model that performs best uses three bidirectional hidden layers with 93 hidden units, a dropout rate of .1, and a learning rate of .001. The classifier was trained for 30 epochs. Once again, class imbalance was addressed using inverse class frequencies as class weights.

2.4 Evaluation

This section reports the performance results of the computational models on the test data. As one goal of this thesis is to explore model performance on second-language vowel productions, all

models despite performance in this section, are analyzed further in section 4.

The classification task here is a ten-way multi-class classification task. Following the zero rule (predicting the majority class for all data points), baseline accuracy is 18.11%. We report both macro-accuracy (averaging the unweighted mean per label) and micro-accuracy (averaging the total true positives, false negatives and false positives) for all models. Table 3 reports macro- (and micro-, in parentheses) averages for all models tested.

Model	Accuracies
SVM-PHONO	.63 (.65)
SVM-MFCC	.60 (.63)
LSTM-MFCC	.63 (.70)

Table 3: Model accuracy.

The SVM-PHONO and LSTM-MFCC models have the highest macro-accuracies (63%) while the LSTM-MFCC model has the highest micro-accuracy (70%). The SVM-MFCC model has the lowest macro- and micro-accuracies. The confusion matrices for the SVM-PHONO and LSTM-MFCC models are shown in the tables below.

	ɑ	æ	ʌ	ɔ	ə	ɛ	ɪ	i̇	ʊ	u
ɑ	.72	.03	.10	.10	.03	.02	0	0	0	0
æ	.05	.76	.03	0	.01	.13	.02	0	0	0
ʌ	.10	.05	.39	.04	.21	.17	.03	0	0	0
ɔ	.28	.01	.07	.57	.05	0	.01	0	.01	0
ə	.02	0	.07	.02	.79	.04	.06	0	.01	0
ɛ	.01	.10	.07	0	.07	.58	.15	.01	0	0
ɪ	0	.02	.02	0	.07	.14	.60	.13	0	.02
i̇	0	0	0	0	.01	.01	.12	.83	0	.03
ʊ	.01	0	.07	.04	.32	.01	.26	.04	.17	.07
u	0	0	.01	0	.06	.03	.18	.22	.02	.48

Table 4: SVM-PHONO confusion matrix. Rows are true labels and columns are model predictions.

The LSTM-MFCC model performs better than the SVM-PHONO on all vowels except /æ/. Both models perform best on /i/ at over 80% accuracy for the SVM-PHONO model and 90%

	ɑ	æ	ʌ	ɔ	ə	ɛ	ɪ	i	ʊ	u
ɑ	.76	.03	.04	.15	0	.01	0	0	0	.01
æ	.03	.71	.02	0	0	.17	.05	.01	0	0
ʌ	.09	.05	.53	.04	.14	.11	.03	0	0	0
ɔ	.24	0	.06	.69	.01	0	0	0	0	.01
ə	.01	0	.08	.01	.82	.03	.04	0	0	.01
ɛ	.01	.13	.05	.01	.04	.57	.18	.01	0	.01
ɪ	0	.01	.01	0	.06	.01	.68	.09	0	.03
i	0	0	0	0	0	0	.07	.90	0	.02
ʊ	.01	.01	.20	.04	.36	.02	.29	.01	0	.05
u	.01	.01	0	0	.03	.01	.14	.08	0	.74

Table 5: LSTM-MFCC confusion matrix. Rows are true labels and columns are model predictions.

accuracy for the LSTM-MFCC model. Schwa /ə/ is the second highest classified vowel for both models, followed by /æ/ for the SVM-PHONO model and /a/ for the LSTM-MFCC model.

Both models performed very poorly on /ʊ/ due to the class imbalance, both predicting neutral /ə/ at the highest rate. However, the SVM-PHONO model predicts /ʊ/ correctly 17% of the time, while the LSTM-MFCC model predicts /ʊ/ correctly 0% of the time.

2.5 Discussion

For specific phone-level accuracies, lower accuracies in the /u/ vowel could be due to the unexpected frontness seen in the analysis of our suggested dimensions in Section 2.2.1.1. With more understanding of this discrepancy, /u/ classification can be improved. Despite the models having the same macro accuracy and LSTM-MFCC having higher micro-accuracy, /ʊ/ predictions show that there is a higher understanding of underlying phonemic categories in the PHONO training rather than the MFCC. This would suggest that improving the accuracy for /ʊ/ would be easier with more PHONO features than it would for the MFCC features, which would most likely require more training data to learn the distinction. This is interesting since it is suggested that MFCCs are able to capture important information in speech signals with as little data as necessary (Tychtl and Psutka, 1999), though the MFCC model performs poorly compared to the PHONO model when

dealing with less training data.

If we assume that phonemes are alike when they differ by a small number of distinctive features, then across models, the error patterns are generally consistent in that errors are often seen between phonologically-like phonemes. For example, /i/ and /ɪ/ or /æ/ and /ɛ/) often differ by one feature phonologically. These errors are those that we call one ”phoneme-step” different in their respective dimension, meaning one sound away from a correct classification. Ignoring imbalance issues with /ʊ/, in the SVM-PHONO model /ɔ/ is most confused with /a/. These phonemes differ from each other by one step in the height dimension. Similarly, in the SVM-PHONO model /ʌ/ is most highly confused with /ə/, differing by one step in the frontness dimension. These facts suggest that the height and frontness dimensions can be better encoded. The vowels included in this experiment can be seen in the IPA vowel chart in Figure 3.

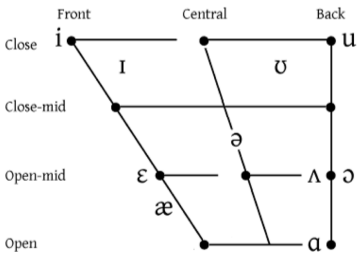


Figure 3: American English vowel chart after IPA chart (including only training vowels).

The highly confused pairs are often one phoneme different in height/frontness. This shows that for all algorithms vowel height/frontness are not perfectly separable by the features suggested. A better understanding of how to encode these features in the data could result in less of these errors. Likewise the addition of a one-step error penalty in training may also produce improved results. However, the interpretability of these errors potentially makes the phonologically-informed models easier to improve. A deeper exploration of the literature could potentially mitigate height/frontness discrepancies, more so than could MFCC features.

3 Behavioral Experiment

In order to compare the performance of human and computer vowel classification, human participants were asked to participate in a perceptual experiment. L1 American English participants were asked to listen to and identify vowel productions by L1 and L2 speakers of American English. In this section we describe the behavioral experiment and explore the results, discussing the human perceptual system and findings therein. In the following section we then compare the human and computer classifications to understand how the two differ and to further assess model performance.

3.1 Data

The stimuli used in this experiment come from the project *Communication in the Global University: A Longitudinal Study of Language Adaptation at Multiple Timescales in Native- and Non-Native Speakers*, a project led by The Center for Multilingual and Intercultural Communication at Stony Brook University.[2] This project looks at the communication between native English-speaking undergraduates (UGs) and non-native English-speaking graduate students working as International Teaching Assistants (ITAs). All ITAs included in the present experiment are native speakers of Mandarin.

ITAs and UGs were asked to read aloud words as they appeared on a screen in front of them. Those words were recorded to a full-session WAV file. Each stimulus was repeated three times in a randomized order. The present experiment uses the data of eight participants (two female ITAs, two male ITAs, two female UGs, and two male UGs). The audio from these participants was segmented into individual WAV files of each word production. From all 50 unique words available, we chose twelve target words and twelve filler words (repetitions 1 and 2) as stimuli based on vowel-type; focusing on minimal-pairs and contrasts deemed 'difficult' for native Mandarin speakers (Wang,

[2]I would like to gratefully acknowledge use of the data funded by NSF grant IBSS-1519908.

1997).

A total of 384 utterances were used in the experiment. 12 target utterances with 2 repetitions each for 8 participants, totalling to 192 target utterances. 12 filler utterances with 2 repetitions each for 8 participants totalling to 192 filler utterances. Target words are shown in Table 6 below with their corresponding (broad) IPA transcription.

Stimulus	**IPA**
pat	/pæt/
pet	/pɛt/
sit	/sɪt/
seat	/sit/
said	/sɛd/
sad	/sæd/
should	/ʃʊd/
shooed	/ʃud/
pot	/pɑt/
pod	/pɑd/
pick	/pɪk/
pig	/pɪg/

Table 6: Perception experiment stimuli.

3.2 Methodology

A forced-choice experiment was created and administered using PsychoPy (Peirce, 2007). Participants heard an audio file of the one-word production and were then presented with four choices marked by a number. Participants were asked to choose the word they heard by clicking the corresponding number. Each incorrect choice presented differed from the target word by one or two features: coda voicing (one change), vowel (one change), or both coda voicing and vowel (two changes).

3.2.1 Participants

Ten participants in total took part in the experiment. All participants were self-identified native speakers of American English. In addition, two participants reported 'high-proficiency' in Spanish and one participant reported 'fluency' in Romanian.

The age range of participants was 20 years old to 53 years old with an average age of 27.7 years old. All participants were from the New York Tri-State area. Participants participated voluntarily and did not receive monetary compensation. No participants reported a history of speech and/or hearing problems.

3.2.2 Data collection and analysis

Each word utterance was transcribed to find the 'true production' vowel, or the phonetic vowel the speaker produced in that utterance. The vowel produced ('true vowel'), the vowel target ('target vowel'), and the vowel chosen by experiment participants ('choice vowel') were compared for each utterance.

Vowel values that match are considered an accurate classification. True and target vowels are compared to find the 'true' phonetic classification of the vowels while true and choice vowels are compared to learn the perceptual classification of the vowels by humans in a forced-choice environment. The two values 'true vowel' and 'choice vowel' are considered here annotations of the utterance and are therefore compared using Cohen's κ for inter-annotator agreement.

3.3 Results

Cohen's κ for 'true vowel' and 'choice vowel' annotations are reported by and across stimulus group (native/non-native). Results can be seen below in Table 7. Our interpretations of Cohen's κ follow the well-known interpretations proposed in Landis and Koch (1977). On native stimuli the 'true vowel' and 'choice vowel' annotations agree with a Cohen's κ of .94 ("almost perfect") as

compared to non-native stimuli at a Cohen's κ of .65 ("substantial agreement").

Stimulus Group	True/Choice
Across	.80
Native	.94
Non-native	.65

Table 7: Cohen's κ for 'True' and 'Choice' vowels.

Overall mean accuracy is reported by and across stimulus group (native and non-native voices) and by vowel group (True, Target, Choice). Mean accuracy by vowel is also reported. Table 8 shows the overall mean accuracy scores while Table 9 reports mean accuracy by vowel.

Stimulus Group	True/Target	True/Choice	Choice/Target
Across	.86	.83	.88
Native	1.00	.95	.95
Non-native	.71	.71	.80

Table 8: Mean classification accuracy on experiment stimuli.

True/Target accuracy suggests that native speakers' true production matched the target vowels at 100% accuracy while the non-native speakers' productions matched the target vowels at just over 70% accuracy. True/Choice accuracy shows that human-perceived and transcribed vowels matched at 95% accuracy for native speaker vowels and 70% accuracy for non-native speaker vowels. Of all vowel groups, choice/target accuracy has the highest accuracy across stimulus group and the highest accuracy for the non-native stimulus group. Accuracy is equal for the native stimulus group in true/choice and choice/target vowel groups.

Choice/target accuracy on the native stimulus group is over 95% while accuracy on the non-native is just over 80%. A chi-square test of independence was performed to examine the relation between nativeness and correct classification. The relation between these variables was significant ($X^2(1, N = 10) = 117.73$, $p < .001$), suggesting the native stimuli are more likely to have an accurate classification by English-native listeners.

Target	Stimulus Group	True	Choice
ɑ	Across	1.00	.97
	Native	1.00	.98
	Non-native	1.00	.96
ɪ	Across	.85	.96
	Native	1.00	.99
	Non-native	.71	.93
ʊ	Across	.69	.93
	Native	1.00	1.00
	Non-native	.38	.86
ɛ	Across	.88	.88
	Native	1.00	.98
	Non-native	.75	.79
i	Across	.94	.84
	Native	1.00	.96
	Non-native	.88	.71
æ	Across	.77	.75
	Native	1.00	.87
	Non-native	.53	.63
u	Across	.80	.55
	Native	1.00	.78
	Non-native	.50	.20

Table 9: Mean classification accuracy by vowel and stimulus group.

Table 9 shows mean classification accuracy by vowel and stimulus group. The highest true/target across-group accuracy is seen in /ɑ/ at 100% accuracy and the highest choice/target across-group accuracy is also seen in /ɑ/ at 97%. The lowest true/target across-group accuracy is seen in /ʊ/ at 69% while the lowest choice/target across-group accuracy is seen in /u/ at 55%. Native and non-native choice/target accuracy were both lowest in /u/, though high in /ʊ/, which were given in minimal pairs in this instance.

The confusion matrix of participant choice vowel and target vowel is shown in Table 10 below. As previously discussed, if we assume that phonemes are "like" when they differ by a small number of distinctive features then the errors seen here are, as expected, highest between like phonemes (/æ/ and /ɛ/, /ɪ/ and /i/, and /u/ and /ʊ/).

	ɑ	æ	ɛ	ɪ	i	ʊ	u
ɑ	310	0	0	0	0	10	0
æ	0	241	77	2	0	0	0
ɛ	0	35	282	3	0	0	0
ɪ	0	0	3	461	16	0	0
i	0	0	0	23	134	0	0
ʊ	0	0	0	0	0	149	11
u	0	0	0	0	0	67	93

Table 10: Choice/Target confusion matrix.

3.4 Discussion

Cohen's κ suggests that the participant choices in the experiment agreed with the phonetic transcription of the vowel at "substantial" agreement rates. Agreement on non-native vowels was significantly lower than that of native vowels which is expected.

The choice/target accuracy shown in Table 8 being significantly greater than that of choice/true shows that, on non-native stimuli, participants were more likely to classify a vowel as target-like than they were to classify a vowel as its 'true' production. This is seen further in Table 9 where the accuracy of true/target is often lower than choice/target. The greatest difference here is seen in /ʊ/ accuracies. However, /u/ is oppositely lower in choice/target accuracy. We relate this fact to well-known word frequency effects in the human perceptual system that humans have a prior bias in favor of common words, especially on correct responses (Broadbent, 1967). In looking at the experiment stimuli, the word *shooed* is less frequent than the word *should*, causing a bias in favor of *should*; seen again in /u/ having the lowest accuracy of all native vowels and native /ʊ/ productions having 100% accuracy. Non-native accuracy for /ʊ/ is also significantly higher than that for /u/, suggesting that participants are biased to choose that they heard a mispronunciation of *should* rather than the target-like *shooed* that was produced.

Accuracy is significantly higher in native-voice stimuli than in non-native-voice stimuli, as is expected. The high accuracy of /ɑ/ in non-native stimuli can be attested to the Mandarin vowel

system also having the /ɑ/ vowel, causing little production error on the side of the non-native speakers.

The results from the experiment are as-expected and provide a valid gold-standard from which to compare the computational models. In the case of word bias, a similar bias to the human perception should not be seen in the computer classification. With this in mind, we assume computer accuracy to resemble more the true/choice comparisons, or accuracy of the choice/target without word frequency biases.

4 Model and Human Comparisons

As previously discussed, current algorithms in CAPT have shown to reflect inconsistencies between human and computer perception. This is problematic as feedback from such a system should approximate the feedback from a teacher; meaning a system that provides more human-like feedback would be considered more accurate. Therefore to assess the models proposed in this thesis, we compare vowel classifications by the human participants in the behavioral experiment to the proposed models in the computational experiment. Inconsistencies between the humans and models will show how the speech codes and algorithms used for classification compare to the human perceptual system.

4.1 Methodology

The stimuli presented to the participants in the behavioral experiment (ITA Data) were used as a validation task for the computational models. The model classifications were compared to the human classifications from the experiment in Section 3. In comparing the two, discrepancies between human and computational classifications can give insight as to where the models are providing feedback unaligned to human perception.

The ITA data were preprocessed following the steps outlined in section 2.2.1 for the PHONO model and those outlined in 2.2.2 for the MFCC models. The data were first re-sampled to a 16000 Hz sampling rate to match that of the TIMIT corpus, the sampling rate the algorithms saw in training.

For each utterance, the true vowel is compared to the human and computer classified vowel. A classification that does not match the true vowel is considered an 'error'.

4.2 Results

Overall mean accuracy was calculated for both humans and computational models. Human and computational performance (mean accuracy) on the ITA data are reported in Table 11 below. Human results are the same as those reported in Table 8.

	Human	SVM PHONO	SVM MFCC	LSTM MFCC
Across	.83	.57	.31	.21
Native	.95	.57	.35	.24
Non-native	.70	.56	.28	.20

Table 11: Computational and human classification performance on ITA data true vowels.

Human classification has a significantly higher accuracy by and across stimulus group, nearing 100% accuracy for native stimuli. For computational accuracies, SVM-PHONO has the highest accuracy by and across stimulus group. The SVM-PHONO performs similarly for native and non-native stimuli while the SVM-MFCC has a significantly higher accuracy for native stimuli than for non-native stimuli. LSTM-MFCC has the lowest accuracies for all groups.

The Cohen's κ for the human and computer classifications are shown below in Table 12. SVM-PHONO showed 'fair' agreement for non-native stimuli and 'moderate' agreement for native stimuli. SVM-MFCC showed 'fair' agreement for native stimuli and 'slight' agreement for non-native stimuli.

	SVM PHONO	SVM MFCC	LSTM MFCC
Across	.44	.16	.15
Native	.49	.21	.19
Non-native	.40	.10	.11

Table 12: Cohen's κ for human and model classifications.

4.3 Discussion

Results show that human classification is far more robust than that of the computational models, with human classification of native stimuli nearing 100% accuracy and non-native stimuli at 70% accuracy. This is compared to the highest performing computational model (SVM-PHONO) reaching only 57% accuracy on native stimuli and 56% on non-native stimuli.

The errors seen in the human classification are attested to a lexical bias creating a discrepancy in the classification of /ʊ/ and /u/ not seen in the computational models. SVM-PHONO classifies /ʊ/ at higher accuracy than the humans classification due to this. Additionally, LSTM-MFCC had 0% accuracy in classifying /ʊ/, suggesting that the PHONO model is able to encode some understanding in the features separating /ʊ/ and /u/ that the MFCC models are unable to without more training data.

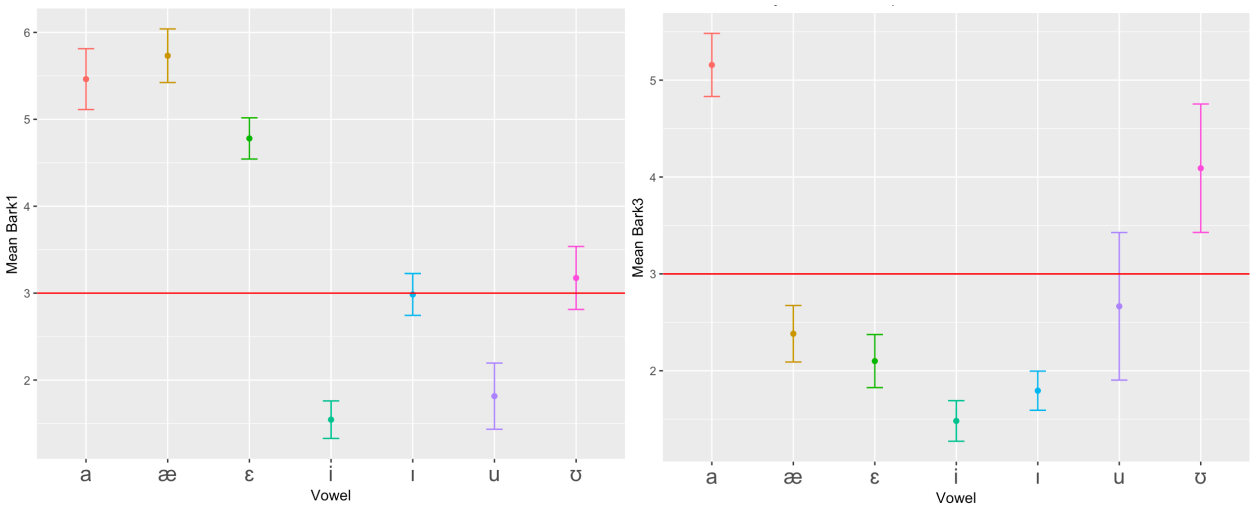
Agreement between human and model classifications is quite low for all models, however the agreement with the SVM-PHONO is significantly higher than the other two models. The SVM-PHONO and SVM-MFCC models performed similarly in testing, though agreement with the human participants is significantly higher for the SVM-PHONO model. The generally low agreement could be attested to the lexical bias, where the SVM-PHONO model correctly identifies /ʊ/ and the human classifications do not. Without this lexical bias we would expect the agreement to be higher. Humans perform much better at those distinctions discussed in Section 2.5 that are difficult for the computational models. These vowels are the main point at which the human and computer classifications diverge.

4.3.1 Bark

This section explores the Bark difference dimensions for the ITA data. The exploration here suggests that the proposed Bark dimensions do properly discriminate L2 vowels into the expected phonological features.

However, L2 vowels are less discriminable by the Bark1 and Bark3 dimensions alone than are L1 vowels. Bark features for the TIMIT data conflicted tensity with the dimension meant to represent height, creating two cues for tensity and aiding in classification. There is less discriminability for L2 vowels as the Bark dimensions *more properly* capture the phonological features. This suggests that the proposed the Bark features do properly encode height and backness for L2 data.

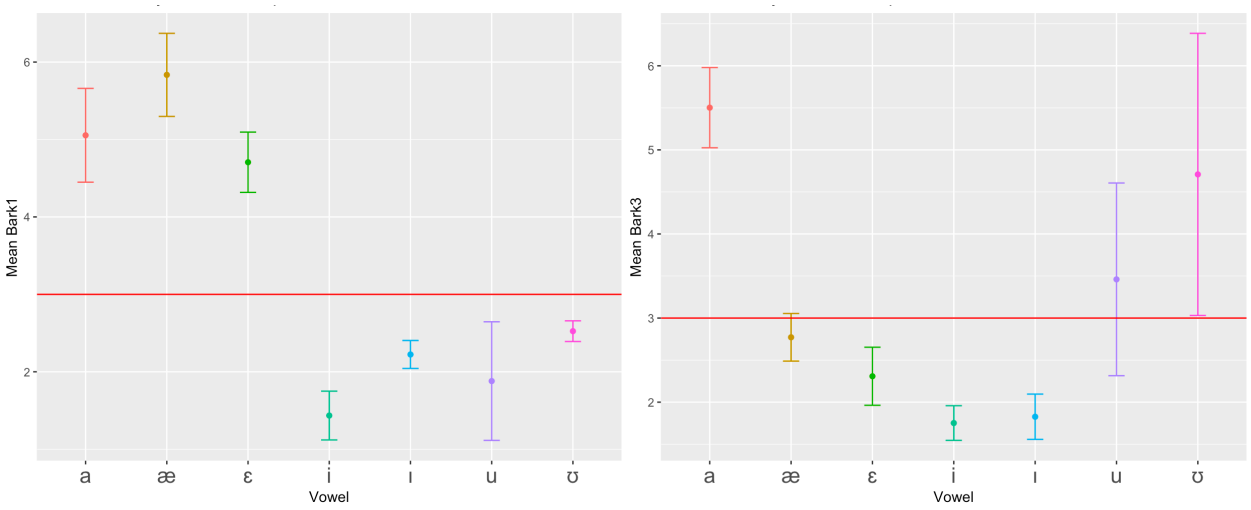
Mean Bark1 by vowel for the UG and ITA productions are graphed in Figure 4 and the ITA productions alone are graphed in Figure 5.



(a) Mean Bark1 by vowel.

(b) Mean Bark3 by vowel.

Figure 4: Bark dimensions by vowel for ITA and UG Productions.



(a) Mean Bark1 by vowel.

(b) Mean Bark3 by vowel.

Figure 5: Bark dimensions by vowel for ITA Productions.

The Bark1 and Bark3 dimensions capture vowel height and frontness similar to that seen in the TIMIT data. There is less distinction in that for L2 data, these Bark dimensions capture the phonological features, though the difference between means are less distinct than in L1 data. /æ/ and /ɛ/ are indistinguishable by these two dimensions themselves for the L2 data while they *are* distinguishable with these two dimensions alone for the L1 data.

In Figure 4a we see that in acoustic realization the high vowels are definitively higher than the low vowels. However, the high lax vowels are much closer to being below the critical distance of 3 than in the TIMIT data. We can possibly attest this to the difference in read speech and isolated speech. Where the TIMIT data is read full sentences, the ITA data is read words in isolation. /u/ was at the critical distance with a mean of 3 in TIMIT but is clearly below that distance in the ITA data. The potential of a higher /u/ produced by L2 learners should be explored further. Figure 4b shows that the Bark3 dimension properly identifies front and back vowels as expected. /u/ was seen in TIMIT unexpectedly as strongly fronted, which is also seen in the data here. In comparison, the confidence interval for /u/ does span above the critical distance, which was not seen in the TIMIT data. We next look at only the productions of the ITAs to understand the influence of the L2 productions on this dimension.

Figure 5 shows the Bark1 and Bark3 dimensions for the ITA productions only. We see that most assumptions made by the suggested dimensions are met for the L2 productions alone, while they were not for the L1 productions. In Figure 5a we see that all high vowels are properly identified as high based on the critical value. This dimension for TIMIT was unable to identify the high lax vowels as high. However, while this dimension was able to distinguish between the high tense and high lax vowels for native speakers, it cannot for the L2 speakers. For this we have to look to the tensity feature. The L1 data therefore had two cues that were able to distinguish the high lax and high tense vowels while the L2 data had only one. We also see that in 5b the /u/ productions are more distinguished by the critical bands. The mean Bark3 for /u/ is properly above the critical value while it was below for across group analysis.

5 Conclusion

This thesis aimed to explore and support the creation of an L1-independent human-aligned vowel classification system using a more phonologically-informed speech coding; especially for use in an educational context. We trained and analyzed three models using two different speech codings (MFCC and the proposed PHONO) and two machine learning algorithms (SVM and LSTM). We also conducted a human vowel perception study on native American-English speakers' perception of native and non-native vowel productions. We believe evaluation on vowels as perceived by humans is a more naturalistic setting and better represents use of the models as an educational tool as compared to previous studies. Our results show that a phonologically-informed speech coding performs better (more human-like) for vowel classification, supporting the use of such a speech coding in CAPT systems.

Our analysis of the models showed that the LSTM-MFCC and SVM-PHONO models perform similarly in macro-accuracy (.63 for both models) while the LSTM-MFCC performs higher in micro-accuracy (.70 for LSTM-MFCC and .65 for SVM-PHONO). This would suggest that the LSTM-MFCC model slightly outperforms the proposed PHONO features. However, in comparing model classification to human classification in the perception experiment, we see that agreement scores with human classification is significantly higher in the SVM-PHONO model (Cohen's κ of .44 for the SVM-PHONO and .16 for the LSTM-MFCC). These results show that the SVM-PHONO model classification is more human-like with confusions between similar phones, as would be expected for human classification.

We addressed the class imbalance seen in TIMIT vowels by applying different class weights to our algorithms in training, which directly affected classification of /ʊ/. We were able to more accurately classify /ʊ/ with the PHONO features than with MFCC features, which suggests that the PHONO features were better able to encode phonological categories. A difficulty in the classification of /ʊ/ was also seen in the perception experiment where well-known word-frequency biases in

the human system showed human preference to the /u/ phoneme. This illustrates that human perception is difficult to approximate, though the proposed PHONO features are better than MFCCs in attempting to do so.

It is not clear if model performance is up to standards but we believe we have shown a great deal of improvement using the proposed PHONO features. A gap still exists between human and model performance. For future development of the PHONO models, this gap can be addressed with further investigation into relevant phonological research, testing more features to improve performance. Additionally, we can implement improved extraction techniques such as the pinpoint measurements techniques used by Evanini et al. (2009) in FAVE-extract. We would also suggest training an LSTM model with the proposed PHONO features, as the LSTM performed well in our experiments. We believe implementing a custom error function to learn highly confuseable phonemes to this model would also be beneficial.

For the behavioral experiment, future research would aim to collect data specifically for the task of evaluating an educational tool, rather than improvising such data collection. Additionally, future research would recruit non-native participants from more diverse language backgrounds as the goal of supporting L1-independence is somewhat neglected with all non-native stimuli being from Mandarin L1 speakers. One thing we have not controlled for are L2 sociolinguistic factors (such as length of stay in the US, length of study, language experiences, etc.), which could be useful to explore.

References

Adank, P., Smits, R., and Van Hout, R. (2004). A comparison of vowel normalization procedures for language variation research. *The Journal of the Acoustical Society of America*, 116(5):3099–3107.

Allwein, E. L., Schapire, R. E., and Singer, Y. (2000). Reducing multiclass to binary: A unifying approach for margin classifiers. *Journal of machine learning research*, 1:113–141.

Bent, T., Bradlow, A., and Smith, B. (2007). *Phonemic errors in different word positions and their effects on intelligibility of non-native speech: All's well that begins well*, pages 331–348. John Benjamins Publishing Company.

Boersma, P. and Weenink, D. (2009). Praat: doing phonetics by computer (version 5.1.13).

Broadbent, D. E. (1967). Word-frequency effect and response bias. *Psychological review*, 74 1:1–15.

Clarkson, P. and Moreno, P. J. (1999). On the use of support vector machines for phonetic classification. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99*, volume 2, pages 585–588. IEEE.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Delattre, P., Liberman, A. M., Cooper, F. S., and Gerstman, L. J. (1952). An experimental study of the acoustic determinants of vowel color; observations on one-and two-formant vowels synthesized from spectrographic patterns. *Word*, 8(3):195–210.

Espy-Wilson, C. Y., Pruthi, T., Juneja, A., and Deshmukh, O. (2007). Landmark-based approach to speech recognition: An alternative to HMMs. In *Eighth Annual Conference of the International Speech Communication Association*.

Evanini, K., Isard, S., and Liberman, M. (2009). Automatic formant extraction for sociolinguistic analysis of large corpora. In *Tenth Annual Conference of the International Speech Communication Association*.

Fabricius, A. H., Watt, D., and Johnson, D. E. (2009). A comparison of three speaker-intrinsic vowel formant frequency normalization algorithms for sociophonetics. *Language Variation and Change*, 21(3):413–435.

Fahey, R. P., Diehl, R. L., and Traunmüller, H. (1996). Perception of back vowels: Effects of varying F1-F0 Bark distance. *The Journal of the Acoustical Society of America*, 99(4):2350–2357.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L. (1993). DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM.

Graves, A., Fernández, S., and Schmidhuber, J. (2005). Bidirectional lstm networks for improved phoneme classification and recognition. In *International Conference on Artificial Neural Networks*, pages 799–804. Springer.

Hincks, R. (2003). Speech technologies for pronunciation feedback and evaluation. *ReCALL*, 15(1):3–20.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Jakobson, R., Fant, C. G., and Halle, M. (1951). Preliminaries to speech analysis: The distinctive features and their correlates.

Jin, S.-H. and Liu, C. (2014). Intelligibility of American English vowels and consonants spoken by international students in the United States. *Journal of Speech, Language, and Hearing Research*, 57(2):583–596.

Juneja, A. and Espy-Wilson, C. (2008). A probabilistic framework for landmark detection based on phonetic features for automatic speech recognition. *The Journal of the Acoustical Society of America*, 123(2):1154–1168.

Karpathy, A., Johnson, J., and Fei-Fei, L. (2015). Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*.

Keshet, J., Chazan, D., and Bobrovsky, B.-Z. (2001). Plosive spotting with margin classifiers. In *Seventh European Conference on Speech Communication and Technology*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lambacher, S. G., Martens, W. L., Kakehi, K., Marasinghe, C. A., and Molholt, G. (2005). The effects of identification training on the identification and production of American English vowels by native speakers of Japanese. *Applied Psycholinguistics*, 26(2):227–247.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Lobanov, B. M. (1971). Classification of Russian vowels spoken by different speakers. *The Journal of the Acoustical Society of America*, 49(2B):606–608.

Meng, H. M., Zue, V. W., and Leung, H. C. (1991). Signal representation, attribute extraction and, the use of distinctive features for phonetic classification. Technical report, Massachusetts Inst of Tech Cambridge Lab for Computer Science.

Neri, A., Cucchiarini, C., and Strik, H. (2008). The effectiveness of computer-based speech corrective feedback for improving segmental quality in L2 Dutch. *ReCALL*, 20(2):225–243.

Nishi, K. and Kewley-Port, D. (2007). Training Japanese listeners to perceive American English vowels: Influence of training sets. *Journal of Speech, Language, and Hearing Research*.

Nishi, K. and Kewley-Port, D. (2008). Nonnative speech perception training using vowel subsets: Effects of vowels in sets and order of training. *Journal of Speech, Language, and Hearing Research*.

Niyogi, P., Burges, C., and Ramesh, P. (1999). Distinctive feature detection using support vector machines. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, volume 1, pages 425–428. IEEE.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic Differentiation in PyTorch. In *NIPS Autodiff Workshop*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peirce, J. (2007). Psychopy - psychophysics software in python. *J Neurosci Methods*, pages 162(1–2):8–13.

Pickett, J. M. (1999). *The acoustics of speech communication: Fundamentals, speech perception theory, and technology*. Allyn and Bacon Boston.

Schutte, K. and Glass, J. (2007). Features and classifiers for robust automatic speech recognition. *Research Abstracts–2007, Research Project. MIT CSAIL Publications and digital archives*.

Shimodaira, H., Noma, K.-i., Nakai, M., and Sagayama, S. (2001). Support vector machine with dynamic time-alignment kernel for speech recognition. In *Seventh European Conference on Speech Communication and Technology*.

Shrawankar, U. and Thakare, V. M. (2013). Techniques for feature extraction in speech recognition system: A comparative study. *arXiv preprint arXiv:1305.1145*.

Slifka, J. (2003). Tense/lax vowel classification using dynamic spectral cues. In *Proceedings of 15th International Conference of Phonetic Sciences, Barcelona, Spain*, pages 921–924.

Slifka, J. (2006). Acoustic cues, landmarks, and distinctive features: a model of human speech processing. *ECTI Transactions on Computer and Information Technology (ECTI-CIT)*, 2(2):91–96.

Strik, H., Truong, K., De Wet, F., and Cucchiarini, C. (2009). Comparing different approaches for automatic pronunciation error detection. *Speech communication*, 51(10):845–852.

Syrdal, A. K. and 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.

Thomson, R. I. (2011). Computer assisted pronunciation training: Targeting second language vowel perception improves pronunciation. *Calico Journal*, 28(3):744.

Traunmüller, H. (1981). Perceptual dimension of openness in vowels. *The Journal of the Acoustical Society of America*, 69(5):1465–1475.

Tychtl, Z. and Psutka, J. (1999). Speech production based on the mel-frequency cepstral coefficients. In *Sixth European Conference on Speech Communication and Technology*.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

Wang, X. (1997). *The acquisition of English vowels by Mandarin ESL learners: A study of production and perception*. PhD thesis, Theses (Dept. of Linguistics)/Simon Fraser University.

Wang, X. and Paliwal, K. K. (2003). Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition. *Pattern recognition*, 36(10):2429–2439.

Wang, Z., Zhang, J., and Xie, Y. (2018). L2 mispronunciation verification based on acoustic phone embedding and siamese networks. In *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 444–448. IEEE.

Watt, D. and Fabricius, A. (2002). Evaluation of a technique for improving the mapping of multiple speakers vowel spaces in the F1˜F2 plane. *Leeds working papers in linguistics and phonetics*, 9(9):159–173.

Weston, J., Watkins, C., et al. (1999). Support vector machines for multi-class pattern recognition. In *Esann*, volume 99, pages 219–224.

Witt, S. M. et al. (1999). *Use of speech recognition in computer-assisted language learning*. PhD thesis, University of Cambridge Cambridge, United Kingdom.

Witt, S. M. and Young, S. J. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech communication*, 30(2-3):95–108.

Yoon, S.-Y., Hasegawa-Johnson, M., and Sproat, R. (2010). Landmark-based automated pronunciation error detection. In *Eleventh Annual Conference of the International Speech Communication Association*.

Zwicker, E. (1961). Subdivision of the audible frequency range into critical bands (Frequenzgruppen). *The Journal of the Acoustical Society of America*, 33(2):248–248.

Zwicker, E. and Terhardt, E. (1980). Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. *The Journal of the Acoustical Society of America*, 68(5):1523–1525.