Measuring the Impact of Segmental Deviation on Perceptions of Accentedness using Gradient Phonological Class Features

Nitin Venkateswaran
Department of Linguistics
University of Florida
venkateswaran.n@ufl.edu

Rachel Meyer
Department of Linguistics
University of Florida
rmeyer2@ufl.edu

Ratree Wayland
Department of Linguistics
University of Florida
ratree@ufl.edu

Abstract

Using Phonet (Vásquez-Correa et al., 2019), a neural network-based model, we generate vector representations of speech segments consisting of phonological class probabilities and use these representations to quantify segmental deviations in the English of native Hindi speakers from American English (AE) and Indian English (IE) baselines, in order to explain how these deviations impact perceptions of accentedness by native AE speakers. The primary focus is on three AE phonemes and their realizations in Hindi English (HE) and Indian English: the labiovelar approximant /w/, often produced as the labiodental approximant [ʋ]; the alveolar stop /t/, commonly realized as the retroflex stop [ʈ]; and the rhotic approximant /ɹ/, rendered as the rhotic tap [ɾ]. Multinomial logistic regressions of Euclidean distances from HE segments to AE/IE baselines on accent ratings show that larger distances from AE baselines increase the likelihood of perceiving stronger accents while larger distances from IE baselines decrease the likelihood. Changes in the probability distributions of contrastive phonological classes are found to correlate with the strength of the perceived accent. These results offer valuable insights into the interplay between native phonology and the perception of accented speech.

1 Introduction

The growing prevalence of English as a global *lingua franca* has led to a diverse variety of Englishes shaped by local linguistic and cultural influences. Among these, Indian English occupies a unique position, with distinct phonological characteristics arising from substrate Indo-Aryan and Dravidian languages (for more, see Wiltshire, 2020). These characteristics often include systematic phonetic differences, which are perceived as accented speech by speakers of other varieties of English.

This study explores how phonetic variation in Hindi English, i.e. the English of native Hindi speakers, influences perceptions of accentedness by native speakers of American English. We focus on three American English phonemes: the labiovelar approximant /w/, often produced as the labiodental approximant [ʋ] in Hindi English (Sailaja, 2009; Wiltshire and Harnsberger, 2006; CIEFL, 1972); the alveolar stop /t/, commonly realized as the retroflex stop [ʈ] (Masica, 1991; Kachru, 1986); and the rhotic approximant /ɹ/, rendered as the rhotic tap [ɾ] (Wiltshire, 2015; Krishnamurti, 2003; Masica, 1991). We use Phonet (Vásquez-Correa et al., 2019), a neural network based on Gated Recurrent Units (GRU) (Chung et al., 2014), to train a single model on large speech corpora of both American and Indian English to infer the classification probabilities of phonological classes associated with the phone segments of both Englishes. The resulting probability vectors are treated as representations of the phone segments in a joint vector space spanning both Englishes. These representations are used to examine the relationship between perceived accent and the Hindi English segments' proximity to American and Indian English baselines in the joint vector space. The segments [ʋ], [ʈ], and [ɾ] are produced uniformly in similar contexts across the varieties of Indian English spoken in the Indian subcontinent (Wiltshire, 2020), including the English of native speakers of Hindi and other Indo-Aryan languages (Fuchs, 2019; Sirsa and Redford, 2013; Wiltshire and Harnsberger, 2006), which facilitates the use of Indian English baselines to study variations in accent perception driven by these segments in Hindi English speaker productions. Quantifying the degree of accentedness using explainable probability vector representations could also facilitate an empirical validation of theories of second language speech learning, in particular the Speech Learning Model (SLM/SLM-r; Flege and Bohn 2021) and the Perceptual Assimilation Model (PAM; Best 1995); the joint vector space of the trained Phonet model could be surmised as a *perceptual space* of segment representations to test theories of speech learning, with distances/similarities between the representations serving as indicators of how second language learners might assimilate the phonetic categories of the language being learned into their own native categories.

2 Related Work

There are a number of studies that investigate accent classification and native language identification using corpora of spoken English from the Indian sub-continent, employing both handcrafted feature-based and neural network-based methods. These studies have used a variety of inputs such as MFCC-based features, prosodic features, formant frequencies, and raw spectrogram-based features with a range of classification models (Guntur et al., 2019; Krishna and Krishnan, 2014; Cheng et al., 2013; Sharma et al., 2024; China Bhanja et al., 2022; Siddhant et al., 2017; Jiao et al., 2016). Feature-based approaches offer explainable results at the expense of hand-crafting time- and resource-intensive features, and neural network approaches are black-box mechanisms capable of automatically deducing key features from the data input. The use of Phonet in this study leverages the neural network's ability to automatically convert key aspects of the spectral speech input into explainable vector representations of speech segments, in order to facilitate an explainable framework of how accent perception relates to gradient phonetic variation.

Other computational methods have been instrumental in capturing gradient phonetic variation which, unlike Phonet, have relied on traditional machine-learning approaches. For example, Yuan and Liberman (2009) introduced a method for capturing nuanced variations, such as degrees of /l/-darkness in American English, using log probability scores from forced alignments instead of categorical phone labels. This method, extended in later work (Yuan and Liberman, 2011), demonstrated both categorical distinctions and gradient degrees of /l/-darkness across contexts. Support Vector Machines have been used to classify r-full and r-less tokens in English using MFCCs (McLarty et al., 2019). Random forest classification has also been employed to model sociophonetic variables (Villarreal et al., 2020), estimating variable realizations by comparing acoustic features with canonical pronunciations.

The approaches listed so far rely on segment-level features to analyze phonetic variation. Approaches that model phonological class probabilities as done in Phonet broaden the scope of analysis from individual segments to entire segment groups defined by the phonological classes. Probabilities of classes such as [continuant] and [sonorant] can be calculated to distinguish among stops, fricatives, and approximants in a gradient-based analysis, complementing traditional acoustic measures as shown in Tang et al. (2023). Such approaches have also been effective at measuring the degree of lenition (Wayland et al., 2023).

3 Methods

This section provides an overview of the Phonet model, its architecture, training methodology, and the datasets used for training. Additionally, the dataset consisting of the English of native Hindi speakers with accent annotations is described.

3.1 Phonet model

Phonet is a GRU-based neural network that estimates the posterior probabilities of the occurrence of phonological classes from speech signals. The signal is chunked into half-second segments, following which the log energy signal across 33 triangular filters along the Mel scale is calculated for each 25-ms window in the chunk. These log-energy feature sequences are processed by two bi-directional GRUs and a time-distributed dense layer, followed by separate dense layers for classifying each phonological class in a multi-task learning setup to calculate the probabilities of the classes associated with the input feature sequence. The probabilities are averaged across the frames to give a unique vector of the probabilities of phonological classes for each phone segment. The bi-directional GRU captures co-articulation effects by incorporating information from the previous and subsequent segments.

3.2 Phonological classes

Phonemes are grouped into phonological classes based on their shared phonetic features. One common distinction is between [+consonantal] and [-consonantal] phonemes. Consonantal phonemes, such as stops, fricatives, affricates, nasals, and liquids, involve constriction of the articulators in the vocal tract and are labeled [+consonantal]. In contrast, vowel and glide phonemes are typically labeled [-consonantal] because they do not involve the same level of constriction. An in-depth guide to phonological classes can be found in Hayes (2011).

For the American and Hindi English phonemes in this study, the labiovelar approximant /w/ is defined by the classes [+sonorant, +continuant, +approximant, +voice, +round, +labial, +dorsal +high, +back, +tense], while the labiodental approximant /ʋ/ is defined by [+sonorant, +continuant, +approximant, +voice, -round, +labial, +labiodental, -dorsal, -high, -back]. The alveolar /t/ is [+consonantal, +coronal, +anterior], but the retroflex /ʈ/ is [+consonantal, +coronal, -anterior]. Finally, the approximant /ɹ/ is [-consonantal, +sonorant, +continuant, +approximant, -tap, +voice, +coronal, +distributed], while the tap /ɾ/ is [+consonantal, +sonorant, +continunant, +approximant, +tap, +voice, +coronal, -distributed, +anterior]. The classes that contrast the /w/-/ʋ/, /t/-/ʈ/, and /ɹ/-/ɾ/ pairs are of particular interest for analyzing against accent ratings.

3.3 Training datasets

To train models on American English and Indian English speech data, we use the English language datasets of the Mozilla Common Voice Speech Corpus (Ardila et al., 2020) and select datasets tagged with `United States English` and `India and South Asia` accent tags. Data from the Librispeech-100 corpus (Panayotov et al., 2015), the L2-ARCTIC non-native English speech corpus (Zhao et al., 2018), and the Indic Text-To-Speech (TTS) corpus (Baby et al., 2016) are used to source additional data in both Englishes. Only the English data from native Hindi speakers is selected from the L2-ARCTIC and Indic TTS datasets; however, the Mozilla Common Voice corpus does not include the speaker's native language tag for Englishes from the Indian sub-continent and all the data with the `India and South Asia` accent tag from this corpus is consequently used, forming the bulk of the training set for the Indian English data. A total of approximately 150 hours of American English and 120 hours of Indian English data are used for training, which includes all the Indian English data available and a correspondingly balanced subset of American English data with a similar number of hours.

3.4 Hindi English dataset with accent ratings

The CSLU FAE (Foreign Accented English) Release 1.2 dataset (Lander, 2007) contains continuous speech in English by speakers of 22 languages, including samples from native Hindi speakers. The corpus consists of telephone-quality utterances with information about perceptual judgments of the accents in the utterances. The speakers were asked to speak about themselves in English for 20 seconds. Three native speakers of American English independently listened to each utterance and judged the speakers' accents on a 4-point scale: *1-negligible/no accent*, *2-mild accent*, *3-strong accent* and *4-very strong accent*. To facilitate investigation of the drivers of accent perception relative to the *no/negligible* accent baseline, the minimum accent rating of the three speakers is taken as the aggregate rating for each recording. The *very strong* accent rating is subsequently merged into the *strong* one, given only one recording is tagged with that rating after applying the aggregate measure. Table 1 shows the distributions of the three accents across the recordings of native Hindi speakers, and Table 2 shows the distribution of the target Hindi English phone segments by accent rating and word position. We refer to this subset of the CSLU FAE dataset containing native Hindi speakers as the Hindi English dataset in subsequent sections.

Accent Rating	No. Recordings
No/Negligible	17
Mild	194
Strong	137
Total	348

Table 1: Distribution of accent ratings in the Hindi English dataset using a minimum aggregate of the ratings of three independent raters.

	Initial	Medial	Final
No/Negligible	29	31	44
Mild	294	346	376
Strong	246	264	256

(a) Distribution of [ʋ]

	Initial	Medial	Final
No/Negligible	23	50	86
Mild	138	569	957
Strong	120	374	643

(b) Distribution of [ʈ]

	Initial	Medial	Final
No/Negligible	12	15	14
Mild	115	173	179
Strong	76	121	157

(c) Distribution of [ɾ]

Table 2: Distribution of target segments in the Hindi English dataset by word position and accent rating.

3.5 MFA pre-processing

The Montreal Forced Aligner (MFA) tool (McAuliffe et al., 2017) is used to force-align the audio and transcripts of the training and Hindi English datasets, with the resulting TextGrid files used to label the phonological classes of each audio frame during Phonet training, in conjunction with the mapping of phone segments to phonological classes described in section 3.6. The transcripts are transcribed into IPA segments using the pre-trained MFA grapheme-to-phoneme (G2P) models and existing pronunciation dictionaries for American and Indian English (McAuliffe and Sonderegger, 2023a,b, 2024a,c). Custom acoustic models for American and Indian English are trained to avoid potentially noisy output from the existing pre-trained model (McAuliffe and Sonderegger, 2024b), given that this model is trained on a variety of world Englishes.

3.6 Phonet training and inference

To learn the phonological classes associated with phone segments during training, and to generate probability distributions over the classes for segments during inference, a mapping between the IPA segments in the MFA pronunciation dictionaries and phonological classes is created for both American and Indian English phone sets. This mapping is created at the phonetic level, given that the learning of speech sounds in a second language occurs at the level of position-sensitive allophones and not at the phonemic level (Flege, 1995; Kohler, 1981).

A single Phonet model is trained on the combined American and Indian English training datasets to estimate the classification probabilities of phonological classes for segments of both languages in a joint vector space. The model can be said to incorporate the acoustic properties of both languages in its parameter weights; this means that, given a phone segment in the Hindi English data, the model can estimate whether the phonological class probabilities of that segment tend towards American English or Indian English baselines, or contain elements of both Englishes.

To facilitate joint training, the phone set to phonological class mappings of the two Englishes are merged into a single mapping, shown in Table 6 in the Appendix. The training and Hindi English datasets are force-aligned using the custom acoustic models described in Section 3.5. An 80-20 train-test split is used for training; the range of accuracy and F1 scores across the phonological classes can be found in Table 5 in the Appendix. The model is trained for a maximum of 30 epochs with early stopping, using the Adam optimizer (Kingma and Ba, 2014) with a categorical cross-entropy loss function. Model hyperparameters include a size of 128 for the bidirectional GRU and hidden layer and a batch size of 16.

3.7 Statistical Analyses

In the vector space of phonological class probabilities defined by the Phonet model, Euclidean distances are calculated between instances of the target Hindi English phone segments and the centroids of all instances of the baseline segments in the American and Indian English training data. The baselines consist of 500 recordings randomly sampled from each of the American and Indian English training datasets. The distances are regressed on the accent ratings using a multinomial logistic regression, taking the *no/negligible* rating as the reference level. The general hypotheses are that, relative to a *no/negligible* accent rating, the odds of a *mild* or *strong* accent should increase with increasing distance from the American English baseline and decrease with increasing distance from the Indian English baseline. Interactions of distance with word position are also investigated, given that variations in the categorization of a speech segment can be driven by the position of the segment in the word sequence (Dmitrieva, 2019). Two-way ANOVA tests are conducted to analyze the effect of accent rating and word position on each of the class probabilities of the Hindi English target segments. Significant differences would be expected for phonological classes that are contrastive between the baseline American English and target Hindi English segments, and the direction of the difference should correlate with differences in accent strength, suggesting that the class probabilities have an impact on the strength of the accent perceived. We report results only for those phonological classes which show significant main effects of accent ratings, or interaction effects of accent ratings with word position, on the probabilities.

4 Results

Throughout this section, the terms AE and IE are used to refer to the American English and Indian English baselines respectively, with HE used to refer to the Hindi English dataset with accent ratings.

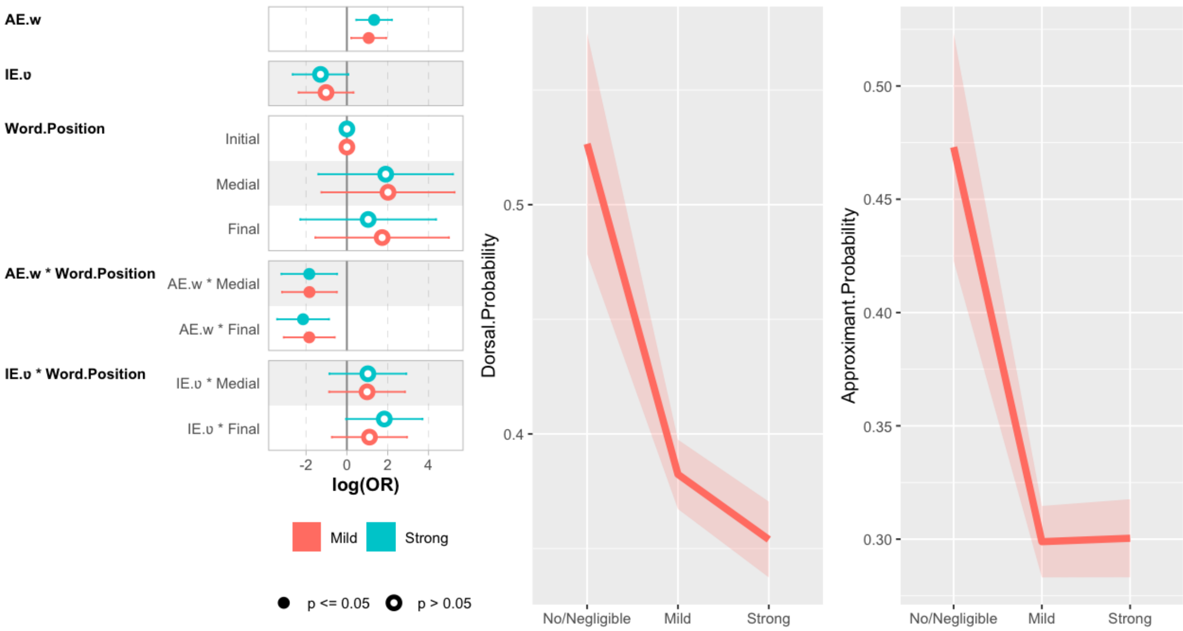

Figure 1: **Left:** Coefficient plots of multinomial logistic regression on accent ratings with reference level set to *no/negligible accent*, for the labiodental approximant [ʋ] in the Hindi English data. The interaction effect of Euclidean distance from AE [w] baseline with word position is significant, as is the main effect of distance from the AE baseline. **Center, Right:** Interaction plots of dorsal and approximant probabilities of the labiodental approximant [ʋ] in the Hindi English data by accent rating and initial word position (AE=American English; IE=Indian English).

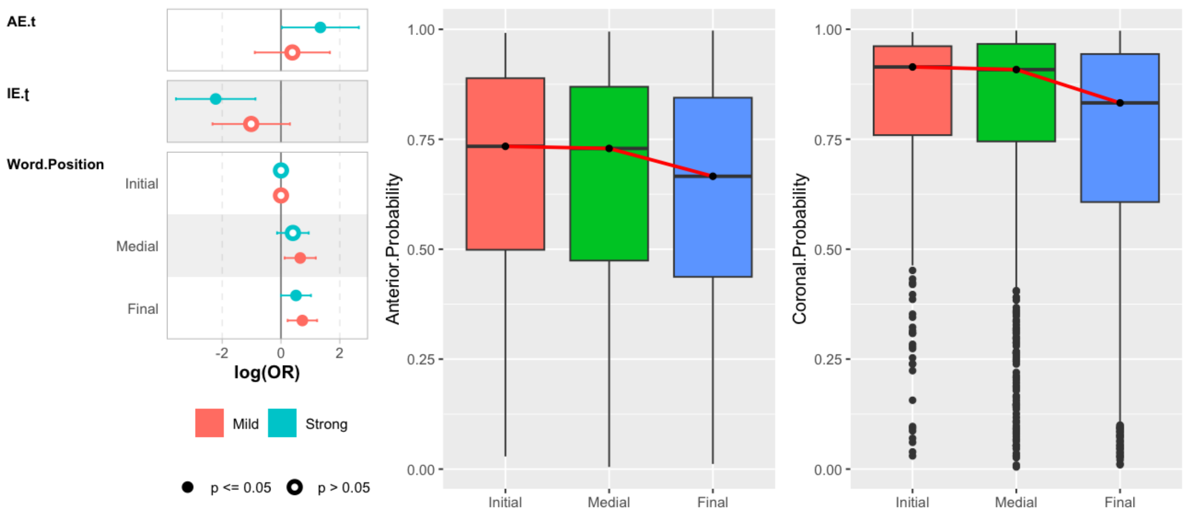

Figure 2: **Left:** Coefficient plots of multinomial logistic regression on accent ratings with reference level set to *no/negligible accent*, for the retroflex [ʈ] in the Hindi English data. The main effects of Euclidean distance from AE/IE baselines are significant, with increasing distance translating to higher/lower odds of strong accent perception. **Center, Right:** Distributions of anterior and coronal probabilities of retroflex [ʈ] in the Hindi English data by word position (AE=American English; IE=Indian English).

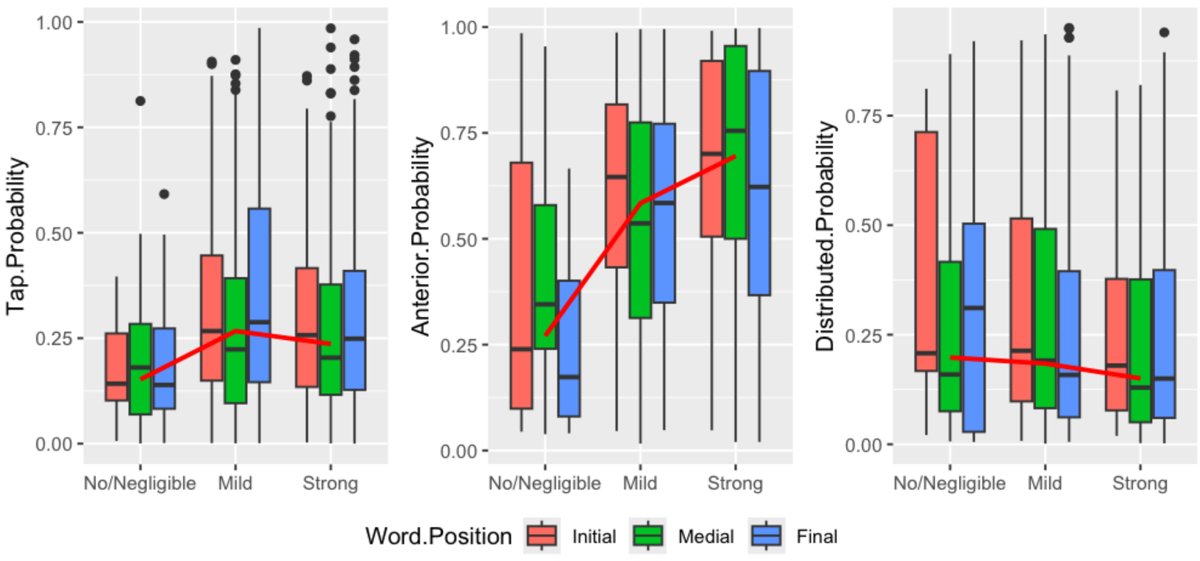

Figure 3: Distribution of tap, anterior, and distributed probabilities of rhotic tap [ɾ] in the Hindi English data by accent rating and word position. The differences in distributions across the accent ratings of all classes taken together suggest that speakers with the *strong* accent are producing the rhotic tap [ɾ] and those with *no/negligible* accent the rhotic approximant [ɹ].

Segment	Accent	Effect	β-coef.	p-val
[ʋ]	Mild	AE Dist.	1.069	.0144
		AE*Medial	-1.833	.007
		AE*Final	-1.844	.0038
	Strong	AE Dist.	1.334	.0027
		AE*Medial	-1.843	.008
		AE*Final	-2.146	.00093
[ʈ]	Mild	Medial Pos.	0.654	.0153
		Final Pos.	0.727	.0045
	Strong	AE Dist.	1.339	.0446
		IE Dist.	-2.22	.0013
		Final Pos.	0.510	.0497
[ɾ]	Mild	AE Dist.	3.567	9.2e-07
		IE Dist.	-3.041	1e-06
	Strong	AE Dist.	4.618	5.6e-09
		IE Dist.	-3.179	6.1e-07

Table 3: Log-odds coefficients (β-coef) of selected variables with accent rating as dependent, taking the *no/negligible* accent as reference level. Only significant effects are reported (p< .05). Positive log-odds coefficients suggest increased likelihood of the accent rating per unit increase in the regressor, relative to the reference accent. Negative coefficients suggest a decreased likelihood. (AE=American English; IE=Indian English).

4.1 Labiodental approximant [ʋ]

Figure 1 shows the coefficient plot of the multinomial logistic regression model described in Section 3.7, and Table 3 includes the β-coefficients for significant regressors with associated p-values. Interaction effects between distance from AE baseline and word position are significant both word medially and word finally. The main effect of distance from AE baseline is also significant. This amounts to higher odds of accent perception both word initially and medially for both *mild* and *strong* accent ratings: for every unit increase in Euclidean distance from the AE baseline, the corresponding increase in the sum of the log-odds coefficients across main and interaction effects is higher word-initially and medially than word finally. There are no main nor interaction effects with distance from the IE [ʋ] baseline, suggesting that accent perception is driven by listeners' unmet expectations of perceiving the labiovelar approximant [w].

Looking at the two-way ANOVA tests, the interaction effects of accent rating and word position on dorsal and approximant probabilities are significant (dorsal: $F_{4,1877}$=3.121, p=.0143; approximant:$F_{4,1877}$=3.899, p=.0037). Tukey post-hoc tests reveal significant differences in average dorsal probabilities word-initially between the *no/negligible* and *strong* accent ratings (p=.02), as well as significant differences in average approximant probabilities word-initially between the *no/negligible* and *mild* and *strong* accent ratings (*mild*: p=.0263; *strong*: p=.0315). The interaction plots are shown in Figure 1 for both phonological classes. The plots show that the dorsal and approximant probabilities decrease with increasing accent strength in word initial position, suggesting that speakers with stronger accents are using the [ʋ] instead of the [w] word-initially.

4.2 Retroflex stop [ʈ]

Starting with the logistic regression, the results indicate that there are no significant interaction effects between distances from baselines and word position on accent ratings for the retroflex stop [ʈ]. There are significant main effects of distance from baselines for the *strong* accent rating (Table 3), with larger distance from AE/IE baseline resulting in higher/lower odds of the *strong* accent. Word position of the retroflex [ʈ] is significant medially and finally with the odds of perceiving an accent higher in those positions.

The two-way ANOVA tests show significant main effects of word position on both anterior ($F_{2,2951}$=5.327, *p*=.00491) and coronal ($F_{2,2951}$=25.980, *p*=6.6e-12) probabilities. Tukey post-hoc tests show lower average anterior probabilities word finally than in both initial (*p*=.02) and medial (*p*=.0397) positions, with word final coronal probabilities also lower than in initial (*p*<.001) and medial (*p*<.001) positions, as the probability distributions in Figure 2 show. However, there are no significant interaction effects word-finally between accent ratings and word position on the probabilities of either phonological class, nor are there significant main effects of accent ratings on the probabilities, suggesting that the anterior and coronal probabilities have no association with the strength of the accent rating for the retroflex [ʈ].

4.3 Rhotic tap [ɾ]

Results for the rhotic tap [ɾ] indicate that there are no interaction effects in the logistic regression between distances from baselines and word position. Significant main effects are observed for distance from baselines (Table 3), with larger distance from AE/IE baselines resulting in higher/lower odds of accent perception. The two-way ANOVA tests show significant main effects of accent ratings on anterior ($F_{2,853}$=26.08, *p*=1.02e-11), distributed ($F_{2,853}$=4.056, *p*=.0176) and tap ($F_{2,853}$=5.798, *p*=.00316) probabilities, and significant main effects of word position on tap probabilities ($F_{2,853}$=4.369, *p*=.01295). Tukey post-hoc tests reveal significant differences in average anterior probabilities between all accent rating pairs, with the largest differences between the *strong* and *no/negligible* (*p*<.001) and *mild* and *no/negligible* (*p*<.001) ratings. Differences in average distributed probabilities between *strong* and *mild* accent ratings are also significant (*p*=.03). Differences in tap probabilities between *mild* and *no/negligible* ratings are significant (*p*=.005) as well as between final and medial positions (*p*=.0093). These distributions are shown in Figure 3. Given that the tap, anterior and distributed classes between the tap [ɾ] and approximant [ɹ] rhotics are contrastive, when taken together the higher anterior and tap probabilities and lower distributed probabilities for *strong* and *mild* accents relative to the *no/negligible* accent could indicate that speakers in the HE dataset vary between the tap [ɾ] and the approximant [ɹ] in their productions, with strongly accented speakers tending towards the rhotic tap.

5 Discussion

5.1 Alignment with theories of second language speech learning

The results empirically show that instances of the Hindi English segments that are farther from the American (Indian) English baselines are associated with higher (lower) odds of an accent. These results align with predictions from contemporary theoretical models of cross-language speech learning, such as the Perceptual Assimilation Model (PAM; Best, 1995) and its extension (PAM-L2; Best and Tyler, 2007), which state that a second language learner's ability to perceptually distinguish speech categories in the language being learned (L2) depends on the categories' perceived similarity to the closest categories in the speaker's native language (L1). The Speech Learning Model (SLM; Flege, 1995) posits that learners at the initial stages of language learning subconsciously map L2 categories to their most similar L1 categories, and new L2 categories are eventually created in the learners' mental representations independent of their L1 categories as learners are exposed to more input distributions in the L2.

The existence of the labiovelar approximant [ʋ], retroflex stop [ʈ], and rhotic tap [ɾ] in the English of L1 Hindi speakers could be the result of transfer effects from learners' L1 language (Sharma, 2017; Kachru, 1986) or learners' exposure to productions from other speakers of Hindi English or Indian English (Sirsa and Redford, 2013). The transfer hypothesis is supported by the existence of the phonemic categories /ʋ/, /ʈ/ and /ɾ/ in Hindi, which also lacks the /w/, /t/ and /ɹ/ phonemes from General American English (Ohala, 1999; Masica, 1991; Giegerich, 1992). The realizations of the /w/, /t/ and /ɹ/ categories as [ʋ], [ʈ] and [ɾ] respectively in

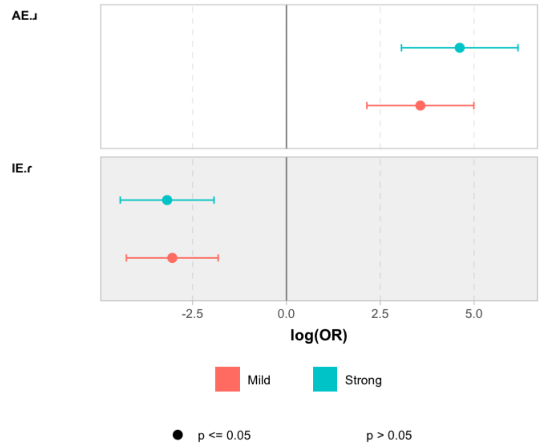

Figure 4: Coefficient plots of multinomial logistic regression on accent ratings with reference level set to *no/negligible* accent, for the rhotic tap [ɾ]. The main effects of Euclidean distance from AE/IE baselines are significant (AE=American English; IE=Indian English).

the Hindi English data are supported by the Single-Category assimilation model from PAM/PAM-L2, which predicts poor discrimination of the American English categories when they are perceived by learners to be similar to their L1 Hindi categories. The SLM also predicts the realization of the L1 Hindi categories in speech in place of the American English categories once learners subconsciously map the American English categories to their most similar L1 Hindi categories. To get an approximate similarity measure, the cosine similarities between the baseline American English categories and the L1 Hindi categories in the Hindi English data are computed in the joint vector space of the Phonet model, using their probability vector representations. Only the set of speakers with a *strong* accent rating is used for the calculation, given that speakers with *no/negligible* or *mild* accents may be producing American English-like categories in their speech in line with the SLM hypothesis described. The cosine similarities between the category pairs are strong ([w]-[ʋ]: μ=0.70, σ=0.14; [t]-[ʈ]: μ=0.81, σ=0.12; [ɹ]-[ɾ]: μ=0.74, σ=0.07), which supports the predictions of the PAM/PAM-L2 and SLM models.

Also consistent with the SLM model is the finding that the perceived degree of accentedness varies depending on the position of the segment within the word, as the mapping of L2 to L1 sounds occurs at the level of position-sensitive allophones. For example, larger distances from the American English labiovelar approximant [w] baseline are more prominent word-initially, and the retroflex [ʈ] segment has a greater impact on accentedness perception word-medially and finally, possibly because the category /t/ is realized in American English as retroflex [ʈ] primarily in word-initial positions and particularly before the rhotic approximant [ɹ] as in 'try' (Polka, 1991).

The retroflex [ʈ] segments in word-final position in the Hindi English data have lower anterior and coronal probabilities than in initial and medial positions, suggesting a higher degree of retroflexion word-finally. The lack of significant effects of accent ratings on anterior and coronal probabilities, together with the significant effect of word-final position on accent strength and the high degree of word-final retroflexion suggest that while the production of the retroflex [ʈ] segment is significant, there may be other acoustic differences between the [t]/[ʈ] segments that are more salient to the perception of accentedness. This finding lines up with research showing that American English speakers have difficulty distinguishing retroflex from dental stops in Hindi (Pruitt et al., 2006; Polka, 1991), suggesting a lack of sensitivity to retroflexion.

The significant difference in average dorsal and approximant probabilities between the *no/negligible* and *strong* accents for the labiodental approximant [ʋ] segments in the data suggests that English speakers of Hindi realize the segment as a labial sound without the accompanying tongue back approximation toward the velum. Moreover, the constriction at the lips is too narrow to achieve the typical resonance of an approximant. For the rhotic tap [ɾ] segment, higher anterior and tap probabilities for *mild* and *strong* accents indicate a forward articulation consistent with a tap rather than the retracted, posterior articulation of the American English [ɹ]. Lower distributed probabilities for *mild* and *strong* accents suggest a reduced tongue contact spread, characteristic of the localized articulation of the tap and contrasting with the broader tongue configuration typical of the approximant [ɹ].

5.2 Investigating Phonet's probability-based representations for accent classification

We investigate whether the phonological class probability vectors generated by Phonet for the segments in this study can differentiate among accent ratings relative to two baseline representations: the log Mel-filterbank (MFCC) transformations described in Section 3.1 that serve as input to the Phonet model, and pre-trained embeddings from the final transformer layer of the WavLM architecture, using the `wavlm-large` model (Chen et al., 2022). The MFCC and WavLM representations are derived for the target Hindi English segments in this study by averaging across all frames for the segment. We run two types of accent classification models that take the representations as input: a linear support vector classifier (SVC) with L2 regularization, with a grid search determining the optimal regularization parameter for each model, and a neural network classifier (NNet) with a single dense layer of size 512 that uses a ReLU activation, followed by a softmax classification layer. The network is trained using cross-entropy loss with the Adam optimizer and a dropout value of 0.5. The Phonet probabilities, like the MFCC representations, are log-transformed. An 80-20 train-test split is used with results averaged across three seeds.

		***F*-score**	
Segment	**Features**	**SVC**	**NNet**
[ʋ]	MFCC	51.28	51.74
	Phonet	45.93	52.43
	WavLM	62.14	68.34
[ʈ]	MFCC	50.19	47.64
	Phonet	49.96	52.44
	WavLM	69.96	79.3
[ɾ]	MFCC	52.56	57.41
	Phonet	52.39	55.65
	WavLM	61.37	67.24

Table 4: F-scores from support vector (SVC) and neural network (NNet) accent classifiers using features from different segment representations as input. Results are averaged across three seeds.

The results in Table 4 show that the WavLM representations, as expected, discriminate the accent ratings best across all segments and classifier types. The nonlinear neural network classifiers trained using Phonet representations show noticeable improvements in the F-score compared to the linear SVC classifiers, and the improvements are seen across all segments. The improvement is particularly visible with the labiodental approximant [ʋ]: the biased linear SVC classifier does worse with Phonet representations compared with MFCC-based ones whereas the nonlinear neural network classifier shows comparable performance between the two representations. The MFCC-based neural network classifiers, in contrast, only show improvement over the linear SVC classifiers for the rhotic tap [ɾ] segment, with worse results for the retroflex [ʈ] segment possibly due to overfitting. These findings indicate that the Phonet-based representations may be richer than the MFCC-based ones in the sense that they contain more non-linear relationships and interactions that can be unlocked by more complex models; however, they do not rival the pre-trained WavLM representations which contain more information to better discriminate accents, at the cost of reduced explainability.

6 Conclusion and Future Directions

This study demonstrates the use of a neural network model, Phonet, to capture gradient phonetic variation which reveals nuanced patterns of L2 mispronunciation that align with and extend second-language speech theories. These findings align with theoretical models of second language speech learning such as the Perceptual Assimilation Model and the Speech Learning Model, particularly in demonstrating the influence of L1 phonological systems on L2 production and the positional sensitivity of speech articulation. The study highlights how gradient phonetic variation offers deeper insights into the articulatory and perceptual mechanisms underlying accentedness, bridging theoretical predictions and empirical observations. Beyond validating second-language speech models, this approach unveils fine-grained articulatory details, advancing our understanding of L2 speech learning and providing a robust foundation for future research in cross-language speech perception and production.

Future research could explore observed patterns of L2 English mispronunciation and positional sensitivity for other L1 languages using neural network-based vector representations to see if generalizations are present. Analyzing co-articulatory effects and dynamic speech variations could further bridge theoretical models and real-world speech patterns, offering deeper insights into second-language acquisition.

Acknowledgments

This research is partially funded by NSF-National Science Foundation (SenSE). Award No. 852 2037266 – SenSE. We also thank the reviewers for their helpful feedback.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben

Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Arun Baby, Anju Leela Thomas, N. L. Nishanthi, and TTS Consortium. 2016. Resources for Indian languages. In *CBBLR – Community-Based Building of Language Resources*, pages 37–43, Brno, Czech Republic. Tribun EU.

C. T. Best and M. D. Tyler. 2007. Nonnative and second-language speech perception: Commonalities and complementarities. In M. J. Munro and O. S. Bohn, editors, *Language experience in second language speech learning: In honor of James Emil Flege*, pages 13–34. Amsterdam:Benjamin.

Catherine T. Best. 1995. *A Direct Realist View of Cross-Language Speech Perception*. Speech perception and linguistic experience: Issues in cross-language research, Strange, Winifred [Ed], Timonium, MD: York Press, Inc, 1995, pp 171-204. York Press, Inc.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Jian Cheng, Nikhil Bojja, and Xin Chen. 2013. Automatic accent quantification of Indian speakers of English. In *Interspeech*, pages 2574–2578.

Chuya China Bhanja, Mohammad Azharuddin Laskar, Rabul Hussain Laskar, and Sivaji Bandyopadhyay. 2022. Deep neural network based two-stage Indian language identification system using glottal closure instants as anchor points. *Journal of King Saud University - Computer and Information Sciences*, 34(4):1439–1454.

Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*.

CIEFL. 1972. *The Sound System of Indian English*. CIEFL, Monograph 7. Hyderabad.

Olga Dmitrieva. 2019. Transferring perceptual cue-weighting from second language into first language: Cues to voicing in Russian speakers of English. *Journal of Phonetics*, 73:128–143.

James E. Flege. 1995. Second language speech learning: Theory, findings, and problems. *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, 92:233–277.

James Emil Flege and Ocke-Schwen Bohn. 2021. *The Revised Speech Learning Model (SLM-r)*, page 3–83. Cambridge University Press.

Robert Fuchs. 2019. Almost [w] anishing: The elusive/v/-/w/contrast in educated indian english. In *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia*, pages 1382–1386.

Heinz J. Giegerich. 1992. *English Phonology: An Introduction*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Radha Krishna Guntur, Krishnan Ramakrishnan, and Vinay Kumar Mittal. 2019. Non-native Accent Partitioning for Speakers of Indian Regional Languages. In *Proceedings of the 16th International Conference on Natural Language Processing*, pages 65–74, International Institute of Information Technology, Hyderabad, India. NLP Association of India.

Bruce Hayes. 2011. *Introductory Phonology*. John Wiley & Sons.

Yishan Jiao, Ming Tu, Visar Berisha, and Julie M Liss. 2016. Accent Identification by Combining Deep Neural Networks and Recurrent Neural Networks Trained on Long and Short Term Features. In *Interspeech*, pages 2388–2392.

Braj B Kachru. 1986. The Indianization of English. *English Today*, 2(2):31–33.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

K. Kohler. 1981. Contrastive phonology and the acquisition of phonetic skills. *Phonetica*, 38:213–226.

G Radha Krishna and Raghava Krishnan. 2014. Influence of mother tongue on English accent. In *Proceedings of the 11th International conference on Natural Language Processing*, pages 63–67.

Bhadriraju Krishnamurti. 2003. *The Dravidian Languages*. Cambridge University Press, Cambridge.

T. Lander. 2007. *CSLU: Foreign Accented English Release 1.2 LDC2007S08. Web Download*. Linguistic Data Consortium, Philadelphia.

Colin Masica. 1991. *The Indo-Aryan languages*. Cambridge University Press, Cambridge.

M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Proc. Interspeech 2017*, pages 498–502.

Michael McAuliffe and Morgan Sonderegger. 2023a. English (India) MFA G2P model v3.0.0. Technical report.

Michael McAuliffe and Morgan Sonderegger. 2023b. English (US) MFA G2P model v3.0.0. Technical report.

Michael McAuliffe and Morgan Sonderegger. 2024a. English (India) MFA dictionary v3.1.0. Technical report.

Michael McAuliffe and Morgan Sonderegger. 2024b. English MFA acoustic model v3.0.0. Technical report.

Michael McAuliffe and Morgan Sonderegger. 2024c. English (US) MFA dictionary v3.1.0. Technical report.

Jason McLarty, Taylor Jones, and Christopher Hall. 2019. Corpus-Based Sociophonetic Approaches to Postvocalic R-Lessness in African American Language. *American Speech*, 94(1):91–109.

Manjari Ohala. 1999. Hindi. In *Handbook of the International Phonetic Association*, pages 100–103. Cambridge University Press, Cambridge.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Linda Polka. 1991. Cross-language speech perception in adults: Phonemic, phonetic, and acoustic contributions. *The Journal of the Acoustical Society of America*, 89(6):2961–2977.

John S. Pruitt, James J. Jenkins, and Winifred Strange. 2006. Training the perception of Hindi dental and retroflex stops by native speakers of American English and Japanese. *The Journal of the Acoustical Society of America*, 119(3):1684–1696.

Pingali Sailaja. 2009. *Indian English*. Edinburgh University Press, Edinburgh.

Ch. Rahul A. N. Sharma, Harsh Kumar Singh, H.Suhas Prabhu, Aniketh V. Jambha, C Jyotsna, and Peeta Basa Pati. 2024. Accent Detection in Indian Languages through Convolutional Neural Network based Spectrogram Analysis. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–5.

Devyani Sharma. 2017. *Chapter 16: English in India*, pages 311–329. De Gruyter Mouton, Berlin, Boston.

Aditya Siddhant, Preethi Jyothi, and Sriram Ganapathy. 2017. Leveraging native language speech for accent identification using deep Siamese networks. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 621–628.

Hema Sirsa and Melissa A. Redford. 2013. The effects of native language on Indian English sounds and timing patterns. *Journal of Phonetics*, 41 6:393–406.

Kevin Tang, Ratree Wayland, Fenqi Wang, Sophia Vellozzi, Rahul Sengupta, and Lori Altmann. 2023. From sonority hierarchy to posterior probability as a measure of lenition: The case of Spanish stops. *The Journal of the Acoustical Society of America*, 153(2):1191–1203.

D. Villarreal, L. Clark, J. Hay, and K. Watson. 2020. From categories to gradience: Auto-coding sociophonetic variation with random forests. *Laboratory Phonology*, 11(1):6.

J.C. Vásquez-Correa, Philipp Klumpp, Juan Rafael Orozco-Arroyave, and Elmar Nöth. 2019. Phonet: A Tool Based on Gated Recurrent Neural Networks to Extract Phonological Posteriors from Speech. In *Interspeech 2019*, pages 549–553.

Ratree Wayland, Kevin Tang, Fenqi Wang, Sophia Vellozzi, and Rahul Sengupta. 2023. Quantitative acoustic versus deep learning metrics of lenition. *Languages*, 8(2).

Caroline Wiltshire. 2015. Dravidian varieties of Indian English. In G. K. Panikkar, B. Ramakrishna Reddy, K. Rangan, and B. B. Rajapurohit, editors, *Studies on Indian Languages and Cultures (V. I. Subramoniam Commemoration Vol. II)*, pages 49–63. International School of Dravidian Linguistics: Thiruvananthapuram.

Caroline R. Wiltshire. 2020. *Uniformity and Variability in the Indian English Accent*. Elements in World Englishes. Cambridge University Press.

Caroline R. Wiltshire and James D. Harnsberger. 2006. The influence of Gujarati and Tamil L1s on Indian English: A preliminary study. *World Englishes*, 25(1):91–104.

J. Yuan and M. Liberman. 2009. Investigating /l/ variation in English through forced alignment. In *Proc. Interspeech 2009*, pages 2215–2218.

Jiahong Yuan and Mark Liberman. 2011. /l/ variation in American English: A corpus approach. *Journal of Speech Sciences*, 1(2):35–46.

G. Zhao, S. Sonsaat, A. Silpachai, I. Lucic, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna. 2018. L2-ARCTIC: A Non-native English Speech Corpus. In *Proc. Interspeech 2018*, pages 2783–2787.

A Appendix

A.1 Phonet accuracy and F1 scores

Table 5 shows the Phonet model's accuracy and F1 classification scores for each phonological class.

A.2 Phone to phonological class mapping

Table 6 shows the merged mapping between the MFA phonesets from McAuliffe and Sonderegger (2024a,c) and the phonological classes from Hayes (2011).

Phonological Class	**Accuracy**	**F1 score**
syllabic	91.07	91.23
consonantal	91.55	91.59
long	86.69	88.8
sonorant	93.68	93.68
continuant	92.50	92.50
delayed release	91.98	92.57
approximant	92.86	92.9
tap	97.31	98.33
nasal	91.83	92.98
voice	93.2	93.2
spread glottis	95.66	96.81
labial	87.65	88.8
round	90.4	92.42
dental	96.15	97.33
coronal	88.65	89.02
anterior	88.08	88.79
distributed	87.56	90.31
strident	95.11	95.52
lateral	92.9	94.8
dorsal	90.97	91.01
high	87.56	88.61
low	91.37	92.41
front	90.26	90.99
back	90.33	92.01
tense	86.84	90.98
constr glottis	99.99	99.99

Table 5: Accuracy and F1 scores for classification of phonological classes by the Phonet model.

Phonological Class	**Phone List**
syllabic	a aj aw aː eː ej i iː oː ow æ ɐ ɑ ɑː ɒ ɒː ɔj ə ɚ ɛ ɛː ɜ ɜː ɝ ɪ ʉ ʉː ʊ
consonantal	b bʲ c cʰ cʷ d dʒ dʲ d̪ f fʲ h j k kʰ kʷ l m mʲ m̩ n n̩ p pʰ pʲ pʷ s t tʃ tʰ tʲ tʷ t̪ v vʲ z ç ð ŋ ɖ ɟ ɟʷ ɡ ɡʷ ɫ ɫ̩ ɱ ɲ ɾ ɾʲ ɾ̃ ʃ ʈ ʈʲ ʈʷ ʎ ʒ ʋ ʔ θ
long	aː ɑː ɒː iː ɛː ɜː eː oː ʉː
sonorant	a aː aj aw ɐ æ ɑ ɑː ɒ ɒː ɛ ɛː ɜ ɜː ɝ eː ej ɪ i iː oː ow ɔj ʉ ʉː ʊ ə ɚ l ɫ ɫ̩ ʎ j ɾ ɾʲ ɾ̃ ɹ m m̩ mʲ ɱ ŋ ɲ n̩ n ʋ w
continuant	a aː aj aw ɐ æ ɑ ɑː ɒ ɒː ɛ ɛː ɜ ɜː ɝ eː ej ɪ i iː oː ow ɔj ʉ ʉː ʊ ə ɚ ð θ f fʲ j ɾ ɾ̃ ɾʲ ɹ ʃ ʒ v vʲ ç l ɫ ɫ̩ ʎ h s z ʋ w
delayed release	f fʲ ʃ ʒ ç v vʲ tʃ dʒ h s z ð θ
approximant	a aː aj aw ɐ æ ɑ ɑː ɒ ɒː ɛ ɛː ɜ ɜː ɝ eː ej ɪ i iː oː ow ɔj ʉ ʉː ʊ ə ɚ j ɾ̃ ɾ ɾʲ ɹ l ɫ̩ ɫ ʎ ʋ w
tap	ɾ ɾ̃ ɾʲ
nasal	m mʲ m̩ ɱ n n̩ ŋ ɲ
voice	a aː aj aw ɐ æ ɑ ɑː ɒ ɒː ɛ ɛː ɜ ɜː ɝ eː ej ɪ i iː oː ow ɔj ʉ ʉː ʊ ə ɚ ð d dʲ ɖ d̪ ɾ ɾ̃ ɾʲ ɹ j ɟ ɟʷ ʒ dʒ v vʲ m mʲ m̩ ɱ n ŋ n̩ ɲ b bʲ l ɫ̩ ʎ ɡʷ ɡ z ʋ w
spread glottis	h
labial	p pʲ pʰ pʷ f fʲ v vʲ ʋ ɱ mʲ m m̩ b bʲ
round	ɒ ɒː ow oː ɔj ʉ ʉː ʊ
dental	t̪ d̪ ð θ
coronal	c cʰ cʷ ç ɾ ɾ̃ ɾʲ ɹ ʃ ʒ dʒ tʃ ʈ ʈʲ ʈʷ t tʰ tʷ tʲ t̪ n̩ n ɲ ʎ d ɖ d̪ dʲ l ɫ ɫ̩ s z θ ð
anterior	ɾ ɾ̃ ɾʲ t tʷ tʰ tʲ t̪ d dʲ d̪ n̩ n l ɫ ɫ̩ s z θ ð
distributed	c ç cʰ cʷ ɟ ɟʷ tʃ dʒ ʃ ʒ ɹ ʎ ɲ θ ð
strident	s z tʃ dʒ ʃ ʒ
lateral	l ɫ ɫ̩ ʎ
dorsal	a aː aj aw ɐ æ ɑ ɑː ɒ ɒː ɛ ɛː ɜ ɜː ɝ eː ej ɪ i iː oː ow ɔj ʉ ʉː ʊ ə ɚ c cʰ cʷ ç k kʷ ɡ ɡʷ ŋ ɲ ɫ ɫ̩ ʎ w
high	ɪ i iː ʉ ʉː ʊ c cʰ cʷ ç k kʷ ɡ ɡʷ ʎ ŋ ɲ w
low	a aː aj aw ɑ ɑː ɒ ɒː æ
front	æ ɛ ɛː ɪi iː c cʰ cʷ ç eː ej j ɟ ɟʷ ɲ ʎ
back	ɑ ɑː ɒ ɒː ɜ ɜː ɝ oː ow ɔj ʊ ɫ ɫ̩ w
tense	eː ej i iː ʉ ʉː oː ow ə ɚ j w
constr. glottis	ʔ

Table 6: Mapping between MFA phonesets and Hayes' phonological classes for Phonet modeling.