DECEMBER 05 2022

Lenition measures: Neural networks' posterior probability vs. acoustic cues

Ratree Wayland ; Kevin Tang, Ph.D; Fenqi Wang ; Sophia Vellozzi ; Rahul Sengupta

Proc. Mtgs. Acoust 50, 060002 (2022)
https://doi.org/10.1121/2.0001728

Volume 50 http://acousticalsociety.org/

183rd Meeting of the Acoustical Society of America
Nashville, Tennessee
5-9 December 2022

Speech Communication: Paper 1pSC9

Lenition measures: Neural networks' posterior probability vs. acoustic cues

Ratree Wayland
Department of Linguistics, University of Florida, Gainesville, FL, 32611-5454; ratree@ufl.edu

Kevin Tang, Ph.D
Department of English Language and Linguistics, Heinrich-Heine-Universitat Dusseldorf, Dusseldorf, GERMANY; kevin.tang@hhu.de

Fenqi Wang
Department of Linguistics, University of Florida, Gainesville, FL, 32611-5454; fenqi@ufl.edu

Sophia Vellozzi and Rahul Sengupta
Department of Computer and Information Science, University of Florida, Gainesville, FL; s.vellozzi@ufl.edu; Aahulseng@ufl.edu

A phonologically informed neural network approach, Phonet, was compared to acoustic measurements of intensity, duration and harmonicity in estimating lenition degree of voiced and voiceless stops in a corpus of Argentine Spanish. Recurrent neural networks were trained to recognize phonological features [sonorant] and [continuant]. Their posterior probabilities were computed over the target segments. Relative to most acoustic metrics, posterior probabilities of the two features are more consistent, and in the direction predicted by known factors of lenition: stress, voicing, place of articulation, surrounding vowel height, and speaking rate. The results suggest that Phonet could more reliably quantify lenition gradient than some acoustic metrics.

Published by the Acoustical Society of America

 https://doi.org/10.1121/2.0001728

1. INTRODUCTION

Voiced stops /b, d, g/, in most if not all Spanish dialects, weaken and are produced as voiced stops [b, d, g] after a pause, a homorganic nasal, and in the case of /d/, after a lateral /l/,[15,20,36] but as voiced fricatives (i.e., voiced continuants) [β, ð, ɣ] in other contexts, including intervocalic and postvocalic, syllable-onset positions across, both within and across word boundaries. The weakening of an underlying stop consonant to a voiced continuant is commonly referred to as spirantization, a member of a broader class of phonological process known as lenition which also includes degemination, [tt → t]; deaspiration, [t^h] → [t]; voicing, [t] → [d]; flapping, [t, d] → [ɾ] debucalization, [t] → [ʔ, h]; gliding, [t] → [j]; and deletion or loss, [ʔ, h, j] → [ø].[14] The distribution of a continuant and a non-continuant variant of Spanish /b, d, g/ has been traditionally described as dichotomous and complementary. However, phonetic studies revealed a more varied and gradient distribution of the two surface realizations. For instance, continuant realizations, previously characterized as fricatives (i.e., produced with turbulent airflow) (17, 30, 33, 36), are phonetically closer to approximants [β̞, ð̞, ɣ̞] (i.e., produced without turbulent airflow) (5,40). Various factors, including surrounding vowel quality, stress, speaking rate (8, 37) and place of articulation[4,10,37,41] further increase phonetic variability and gradience, suggesting a continuum rather than a fixed degrees of lenition across environments. Like voiced stops, voiceless stops in some dialects of Spanish also undergo lenition (e.g., 3) and become partially or totally voiced (e.g., 29, 32).

Different acoustic properties have been used as measures of lenition including intensity, duration, periodicity (e.g., harmonic-to-noise ratio), etc. However, the intensity, typically calculated as a difference or as a ratio, is the most prevalent,[8,21,37,42] consistent with the view that lenition mainly affects the intensity of the target segment (see Harris, Urua and Tang[16] for a detailed discussion). For example, Martínez-Celdrán and Regueira,[32] Figueroa Candia and Evans,[11] and Broś et al.[3] used intensity difference between the target and its preceding segment (preceding segment's maximum intensity- minimum intensity of the target consonant) as a lenition marker. Similarly, Hualde, Shosted and Scarpace[21] measured the difference between the maximum intensity value during the vowel following the target consonant and the minimum value during the target consonant portion to quantify degree of lenition. For both relative intensity measures, the smaller the difference, the more open the constriction of the target consonant (i.e., the more lenited the target consonant) is inferred.

Another intensity measure used to quantify degree of lenition in some studies is the maximum rising velocity from the midpoint of the target consonant to the midpoint of the following vowel.[21,22,26] The smaller the maximum rising velocity value, the less abrupt the transition intensity and thus the more lenited the consonant is. Further, mean intensity of the target sounds could also proxy their degree of lenition: the higher the mean intensity, the more advanced their degree of lenition.

Relative duration of the target consonant (target sound duration divided by the total duration of the preceding sound + target sound + following sound) correlates negatively with the degree of consonant weakening and has been used as a reliable lenition measure (e.g., 9). This measure is usually used when the target consonant occurs intervocalically but can be adapted to other contexts. For example, Broś et al.[3] calculated relative duration ratio by dividing target sound duration by the total of the preceding sound + target sound because the segment following the target sound was not always a vowel in their data. The more lenited the target consonant, the shorter its duration, thus the smaller the duration ratio is expected.

The harmonics-to-noise ratio (HNR), a measure of proportion of acoustic periodicity (harmonic) to aperiodicity (noise) of a given sound expressed in decibels (dB), is another lenition marker. An HNR of 0 dB indicates equal energy in harmonics (periodicity) and noise. A positive HNR value indicates a higher harmonic energy relative to noise energy while a negative HNR value indicates a higher noise energy relative to harmonic energy. The more lenited a segment, the more vowel-like it is, hence the higher the HNR.[3]

The goal of this study is to compare a computational approach, 'Phonet', to the traditional acoustic mea-

surements of lenition. Unlike the quantitative acoustic approach where acoustic values are direct estimates of lenition degrees, in Phonet, degree of lenition is estimated from the posterior probabilities of relevant phonological features, computed from the input signals by bidirectional recurrent neural networks (RNNs). The two relevant phonological features are [continuant] and [sonorant] since the lenition of Spanish voiced stops can be simplistically described as involving categorical feature changes from [-continuant] to [+continuant] and/or from[-sonorant] to [+sonorant]. However, to capture the highly varied degrees of lenition, it is necessary that we look beyond categorical manifestations of lenition changes and beyond the binary nature of phonological features. The basis for this approach is outlined in the following sections.

A. PHONOLOGICAL FEATURES AND LENITION

Phonemes are classified into classes based on their common phonetic features. For example, in most, if not all languages, phonemes are either [+consonantal] or [-consonantal] depending on whether their articulation involve constriction of articulators in the vocal tract. Stops, fricatives, affricates, nasals, and liquids are [+consonantal] while vowel and glide phonemes in most languages are [-consonantal]. The inverse correlate of [consonantal] is [syllabic]. [+syllabic] phonemes are the most sonorous segments and usually occupy the nucleus position of a syllable, while [–syllabic] phonemes are less sonorous and are typically not allowed in a syllable nucleus position. Vowels and syllabic consonants /ɻ̩, l̩, n̩, m̩, etc./ are [+syllabic] while other consonants including glides are [-syllabic].

[sonorant] and its inverse [obstruent] is another major class of phoneme. [+sonorant] phonemes are produced with little to no oral constriction while [-sonorant] or [+obstruent] phonemes are produced with complete or substantial airflow obstruction. Nasals, liquids, glides, and vowels are [+sonorant] while stops, fricatives and affricates which are produced with complete or substantial airflow obstruction are [-sonorant] or [+obstruent].

[continuant] feature describes sustainability of airflow through the oral cavity. An incomplete closure between articulators allows continual oral airflow for [+continuant] phonemes. Fricatives are [+continuant] because partial occlusion of the oral cavity during their production permits continuous oral flow. Other [+continuant] phonemes include liquids, glides, and vowels. For nasals, they are considered [-continuant] by some (because of airflow blockage through the oral cavity during their production), but [+continuant] by others (because continuous airflow is allowed through the nasal cavity). In this study, we specified them as [-continuant]. See (e.g., 18) for guides to other phonological features.

Phonemes that share a set of phonological features form a natural class and tend to pattern together when undergoing a phonological process. For example, /p, t, k/ are [-syllabic, -voice, -continuant, -sonorant, -delayed release] and form a natural class in English. They all become aspirated at the onset of a stressed syllable. Similarly, Spanish /b, d, g/ [-syllabic, +voice, -continuant, -sonorant, -delayed release] form a natural class. As discussed above, they are lenited and surface as fricatives [+continuant] or approximants [+sonorant] in intervocalic position.

B. POSTERIOR PROBABILITY AND PHONETIC GRADIENCE

Computational approaches have been used in studies of phonetic variations. Many of these studies have relied on forced alignment systems to determine pronunciation gradient (e.g., [dʒ]-[z] and [p^h]-[f] variations in Hindi English code-mixed speech,[38] ‘g’-dropping in English,[24,50] ‘th’-fronting, ‘td’-deletion, and ‘h’-dropping in English[1]). The forced alignment systems typically take word-level orthographic transcriptions as input, making reference to a pronunciation dictionary with phone-level transcription. Multiple pronunciations can be assigned to each word entry in the dictionary. For instance, to model ‘th’ fronting, two pronunciations, [θ] and [f], could be given to all words entries that may undergo ‘th’-fronting. Based on each word token’s acoustic properties, a trained forced aligner can automatically determine which of the

two pronunciations has the highest probability. However, since a forced alignment model contains an acoustic model for each phone type defined in the pronunciation dictionary, the degree of variation could not be determined beyond the granularity of the phone set (e.g., as either [θ] or [f]).

A novel and creative method to obtain a more gradient measure of variations (e.g., degrees of 'th'-fronting as opposed to simply coding a token as [θ] or [f]) was proposed by Yuan and Liberman.[48] In this study on degrees of /l/-darkness in American English, probability scores extracted during the forced alignment procedure were used as measure of variation instead of phone labels outputted by the forced alignment procedure. Probability score is the log probability (log probability density) of the aligned segment as a particular phone. More specifically, all /l/ tokens from a corpus of American English were forced aligned twice: first by a model trained on light /l/s (word-initial) and second by a model trained on dark /l/s (word-final and word-final consonant clusters). Degrees of /l/-darkness was indicated by the difference between the log probability scores from the two different alignments. The method was also used to examine finer variations of both types of /l/s by Yuan and Liberman.[49] In addition to revealing the categorical distinction between dark (in syllable coda) and light /l/ (in syllable onset), their results also revealed that intervocalic dark /l/ is less dark than canonical syllable-coda dark /l/, and that degrees of darkness depends on the stress of the flanking vowels. Magloughlin[31] used the same method to investigate gradient variation of /t/-/d/ affrication in English, measured by the log probability scores from the /tʃ, dʒ/ alignment and the /tɹ, dɹ/ alignment using acoustic models of /tʃ/ and /dʒ/, and /t/ and /d/, respectively.

In addition to acoustic models in a forced alignment system, probability estimates from token classification can be obtained from other approaches. For instance, to examine the degree of r-lessness of postvocalic /ɹ/ in English, McLarty et al.[35] trained the Support Vector Machines (SVM) model to classify the canonical r-less tokens (oral vowels that are not preceding a liquid or nasal) and the canonical r-full tokens (prevocalic /ɹ/) using Mel-Frequency Cepstral Coefficients (MFCCs) as the acoustic representations. Once successfully trained (mean classification accuracy of 98.95%), the model was applied to ambiguous tokens (postvocalic /ɹ/) to obtain a probability estimate of being r-less vs. r-full. A similar method was used by Villarreal et al.[45] in their study of two English sociophonetic variables (non-prevocalic /ɹ/ and word-medial intervocalic /t/). It is important to note that the classification method used by most of these studies are trained on surface segments that are not necessary surface realizations of the segment undergoing variation of interest. It simply relies on acoustic similarities between these surface segments and the possible canonical realizations of a variation. For instance, in the case of 'th'-fronting, the model was trained to classify tokens that are either canonically [θ] or canonically [f] and these canonical tokens themselves are not subjected to 'th'-fronting. However, their acoustic characteristics would capture the range of possible surface realizations of 'th' fronting.

The potential of this approach to estimate the categorical manifestation of lenition was illustrated by the results of Cohen Priva and Gleason.[6] In this study, a range of lenition processes were modelled using a spoken corpus of American English. Three types of modeling methods differing in the underlying representation of the surface segments were examined. The first method compared the surface forms of two segment types (e.g., [t] and [d] for the lenition process /t/ → [d]) regardless of their underlying form (e.g., the [t] and [d] tokens do not need to share the underlying form /t/). The second method compared the surface forms of two segment types that share the same underlying form (e.g., /t/ is the underlying form for both [t] and [d]). The third method compared segments that surfaced unchanged, e.g., the [t] tokens realized from /t/ and the [d] tokens from /d/. The finding that all three modeling approaches yielded the same results, suggested that the various acoustic manifestations of a given lenition process (/t/ →[d] in this case) can be captured by comparing relevant pair of surface segments, regardless of their underlying form.

Unlike Cohen Priva and Gleason,[6] the Phonet approach targets a whole class of lenition, so we must go beyond classifying pairs of segments that are relevant to a lenition process to two groups of segments that are categorized by a binary phonological feature. We focus on the probability of the phonological feature [continuant] which differentiates stops from non-stops, and [sonorant] which differentiates stops

and fricatives from non-stops and non-fricatives because they capture the two categorical realizations of stop lenition in Spanish. A high [continuant] probability but a low [sonorant] probability would suggest a fricative-like realization, while a high [continuant] probability and a high [sonorant] probability would indicate an approximant-like realization of lenition. The degree of lenition is estimated from the probability of each phonological feature estimated from acoustic properties of the input signals.

C. PHONET

Phonet[44] is a bi-directional recurrent neural network model. It is trained to recognize input phones as belonging to different phonological classes defined by phonological features (e.g., sonorant, continuant). It is semi-automatic and only requires a segmentally aligned acoustic corpus (using forced alignment). Input to Phonet is log-energy distributed across triangular Mel filters computed from 25-ms windowed frames of each 0.5 second chunk of the input signal. Weighted categorical cross-entropy loss function was used. The weight factors for each class are based on the percentage of samples from the training set, that belong to each class. Adam optimizer[25] was used to train the model. Dropout and batch normalization layers were used to improve the generalization of the networks. The training lasted 81 epochs, with early stopping enabled (with a patience of 15 epochs). For more detail about the model and related procedures, see 44 and the publicly-available code at https://github.com/jcvasquezc/phonet. Once trained, posterior probabilities for different phonological features of the target segments can be computed by the model. It has been found to be highly accurate in quantifying degree of lenition in Spanish[43,47] and in intoxicated English speech,[46] and modelling the speech impairments of patients diagnosed with Parkinson's disease.[44] The architecture of Phonet is described in detail in Vásquez-Correa et al.[44] Phonet can be customized with different sets of phonological features and acoustic representations. In this study, we focus on the probability of the phonological features [continuant] and [sonorant] to capture degree of lenition.

2. THIS STUDY

This study compares Phonet to acoustic metrics of lenition. The reliability of the two approaches is evaluated against known effects and direction of lenition variables. Consistent and significant results in the predicted direction are taken as indicators of the effectiveness and the reliability of the approach.

A. METHODS

I. Materials

The Argentinian Spanish Corpus containing crowd-sourced recordings from 44 (31 female, 13 male) native speakers of Argentinian Spanish built by Guevara-Rukoz et al.[13] was used in this study. The male sub-corpus contains 2.4 hours of recording with 16,914 words (3,342 unique words) while the female sub-corpus contains 5.6 hours of recording with 35,360 words (4,107 unique words). For the study, word tokens with voiced and voiceless stops, /b, d, g, p, t, k/, occurring between two vowels with different degrees of openness were selected. Table 1 specifies the number of word tokens and word types by conditions – voicing (voiced or voiceless), place of articulation (bilabial, dental, and velar), preceding and following vowels (open, mid, and close).

II. Phonet Training Procedure

The Montreal Forced Aligner (version 2.0)[34] was performed on the corpus. Based on Hualde[23] grapheme-to-phoneme mapping in IPA, a phonemic pronunciation dictionary for the transcription of the corpus words was generated and used to train new acoustic models for the corpus and align the textgrids to the acoustic

Table 1: Word distribution by conditions – voicing, place of articulation, preceding vowel, and following vowel. The number left and right of the slash in each cell represents the number of word tokens and word types, respectively.

		Voiced			Voiceless		
		Following vowel height					
Place	Preceding vowel	Close	Mid	Open	Close	Mid	Open
Bilabial	Close	19/1	3/0	8/1	7/1	0/0	6/1
	Mid	134/28	216/30	142/23	183/26	435/44	228/21
	Open	128/18	90/18	103/8	69/10	316/28	248/24
Dental	Close	0/0	20/3	0/0	0/0	26/1	38/2
	Mid	143/32	409/34	43/8	185/13	451/36	97/8
	Open	51/8	388/20	11/2	92/10	141/16	107/7
Velar	Close	0/0	0/0	0/0	4/0	40/4	21/1
	Mid	0/0	2/0	0/0	190/25	912/50	215/40
	Open	0/0	32/1	0/0	42/12	5776/40	307/32

signals. A tri-phone acoustic model in which the left and the right contexts of the target phone are used to adjust its alignment during the alignment procedure. The phone set parameter was set to IPA, which enabled extra decision tree modeling based on the specified phone set. All parameters were kept as the default. Of all the sentences spoken by the speakers, 80% of them were randomly chosen as the training set, and 20% as the test set. Each sentence contains one target word, therefore the target words were distributed proportionally in the train-test split. The decision to split the sentences by each speaker (i.e., the sentences of each speaker appear in both train and test set) was motivated by our need to examine the target words from all speakers. All the data (train and test) were used in the statistical analyses of all five acoustic parameters described below in addition to the two posterior probabilities. Since the surface realizations of the targets /b, d, g/, but not the targets /p, t, k/,[7] were expected to be ambiguous in their realizations of the two features of interest: continuant and sonorant, they were not included (i.e., silenced out) during training to avoid model contamination by the ambiguous tokens. In total, twenty-three phonological classes including syllabic, consonantal, sonorant, continuant, nasal, trill, flap, coronal, anterior, strident, lateral, dental, dorsal, diphthong, stress, voice, labial, round, close, open, front, back and pause were trained by twenty-three different Phonet models. Like Vásquez-Correa et al.,[44] one addition model was included to train the phonemes. However, in addition to the 18 phonemes from Vásquez-Correa et al.,[44] 7 additional phonemes including stressed /ˈa, ˈe, ˈi, ˈo, ˈu/, /ɲ/, /θ/ and /spn/ for speech-like noise were also included. Model training was performed on the NVIDIA GeForce RTX 3090 GPU.

Only the test data was used in the internal evaluation of the posterior probabilities generated by the Phonet model. The model was highly accurate showing unweighted average recall (UAR) ranges from 94%-98% across the different phonological classes. The sonorant and continuant features' UARS were 97% and 96%, respectively, suggesting a good model fit for our features of interest. The model was then applied to our target word tokens in both the train set and the test set with intervocalic voiced and voiceless stops, /b, d, g, p, t, k/. The predictions were computed for 25-ms windows every 10-ms. The average of the middle frame(s) was used as the prediction for phone tokens containing multiple frames. A sonorant posterior probability and a continuant posterior probability were obtained for each target stop.

III. Acoustic Parameters: HNR, Duration, and Intensity

To compare our model to the quantitative acoustic approach, five common acoustic parameters covering three broad acoustic dimensions of lenitions were selected for comparisons. Harmonic-to-noise ratio (HNR), relative duration, intensity difference (two types) and mean intensity were extracted from the target intervocalic voiced and voiceless stops, /b, d, g, p, t, k/. HNR was calculated as ten times the log10 ratio between the energy of harmonicity and noise. The mean HNR of the target segments was computed in Praat.

Relative duration for each target stop was obtained by taking the duration of a target stop and divided

it by the total duration of the preceding vowel + target consonant + following vowel. The duration of the segmental tokens was generated during the forced alignment (see section ii above). The more lenited the consonant is, the shorter the relative duration. Two intensity difference values were calculated for each target stop by subtracting minimum intensity of the target segment from the maximum intensity of (a) the preceding vowel, and (b) the following vowel. The assumption is that the smaller the intensity difference between the target stop and its flanking vowels, the less constricted and hence the more lenited it is. The maximum intensity values of the preceding and following vowel and the minimum intensity value of the target segment were calculated using the parabolic interpolation method in Praat. Finally, the mean intensity values of the target segments were calculated in Praat.

IV. Statistical Analyses

Values of the five acoustic parameters described above and the sonorant and continuant posterior probabilities generated by the Phonet model served as dependent variables in the linear mixed-effects regression models. The models' fixed variables were stress (stressed or unstressed), voicing (voiced or voiceless), place of articulation (bilabial, dental, and velar), preceding vowel height/openness (open, mid, and close), following vowel height (open, mid, and close), speaking rate, and word status (content or function). Speaking rate and word status were included as they are known to influence lenition. A higher degree of lenition is expected for a faster speaking rate relative to a slower speaking rate, and for function words compared to content words.[3,19,42] Similarly, a strong effect of stress on lenition has been reported, with a higher degree of lenition expected in unstressed syllables (or post-tonic) than in stressed syllables (or pre-tonic).[3,10,37] On the contrary, the influence of place of articulation and flanking vowel openness has been inconsistent.[8,26,27,29,37] Overall, velar stops are expected to be more lenited than labial and dental/alveolar stops, and the more open the flanking vowels, the greater the degree of lenition is expected. Regarding the effect of voicing, voiced stops are expected to be more lenited than voiceless stops.[3,7] Deviation coding was used for the categorical variables stress, voicing and word status, while forward difference coding was used for the variables place of articulation (bilabial > dental > velar), preceding vowel (close > mid > open), and following vowel (close > mid > open). The models were performed using the lmer function from the lme4 package[2] in R.[39] After comparing multiple model structures with maximum likelihood, the best-fit model structure for each variable was identified. Seven regression models were fitted with each of the five acoustic parameters and the two deep-learning-based features (the sonorant and the continuant phonological features) as the dependent variables. All models included different interactions terms but same random intercepts by speaker and word. The general formula of the model with three interaction terms is provided as follows:

DEPENDENT VARIABLES $\sim$ Stress + Voicing + Place of articulation + Preceding vowel + Following vowel + Speaking rate + Word status + Place of articulation:Preceding vowel + Place of articulation:Following vowel + Preceding vowel:Following vowel + (1 | Speaker) + (1 | Word).

Post-hoc comparisons of the interaction terms were carried out using emmeans (with Tukey HSD for *p*-value adjustment).[28] Results of the best-fit model for each dependent variable are reported in the next section. Due to space limitations, only the main effects are reported.

B. RESULTS

I. Acoustic Parameters: HNR, Duration, and Intensity

Figure 1 presents results of the main effects of the linear mixed-effects regression models for the acoustic parameters (x-axis; 1= relative intensity to the preceding vowel; 2 = relative intensity to the following vowel; 3 = mean intensity, 4 = duration ratio and 5 = mean HNR). Dependent variables are values of the five acoustic dimensions while stress, voicing, place of articulation, height of flanking vowels, speaking rate and word status are the predictors (shown as separate panels in the figure). All but speaking rate are categorical

variables. Reference levels for the categorical variables are stressed, voiceless, dental, velar, mid vowel, low vowel, and function words, respectively. Coefficient (β) values are represented on the y-axis. (*) indicates significant differences ($p<0.05$).

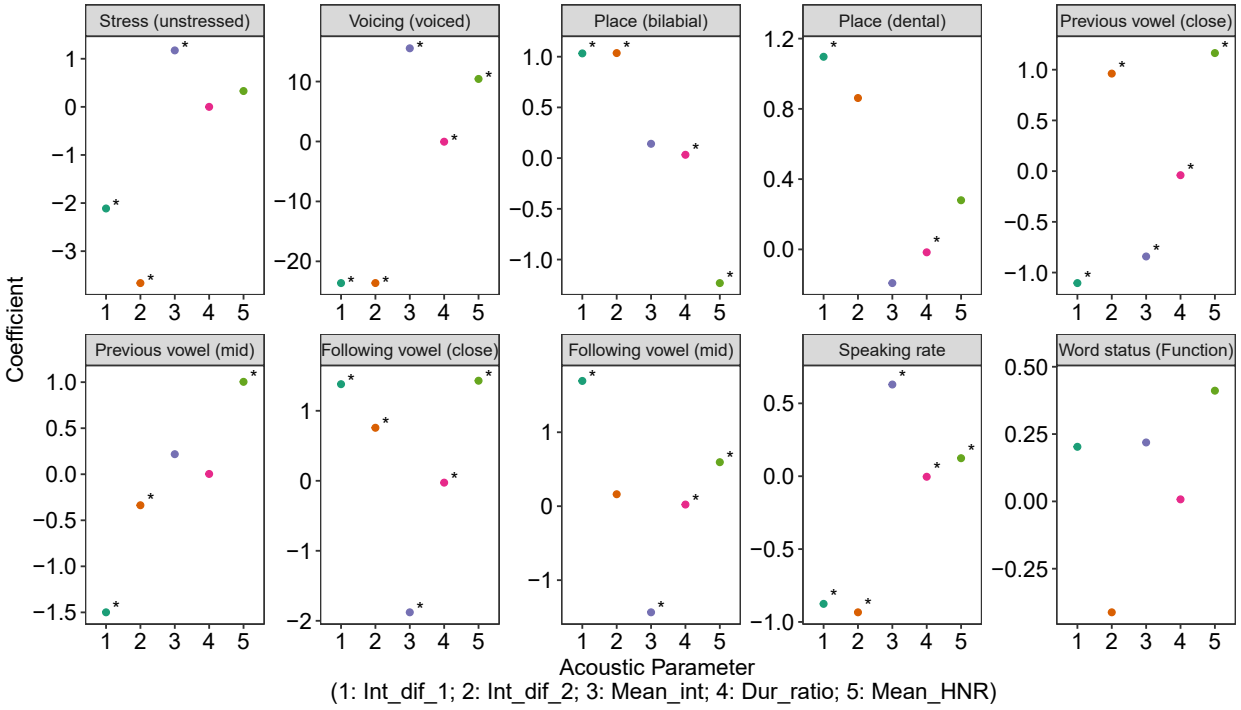

Figure 1: Results of the linear mixed-effects regression models for the acoustic measures.

From the first panel of this figure, we see that intensity difference between a target stop relative to a preceding vowel and a following vowel (labels 1, and 2, respectively, on the x-axis) was expectedly predicted to be significantly lower (negative coefficient values) in an unstressed relatively to a stressed syllable [βs=-2.113, -3.664; ps<0.001], suggesting a higher degree of lenition in an unstressed than a stressed syllable. Also expectedly, mean intensity was predicted to be significantly higher (positive coefficient values) in an unstressed syllable relatively to a stressed syllable [β=1.174, $p<0.001$]. Unexpectedly, however, a non-significant effect of stress on degree of lenition was suggested for duration ratio and mean HNR measures (Labels 4, 5 on the x-axis) [βs=0.000, 0.329; $p>0.05$]. The remaining results can be interpreted in a similar fashion and are summarized in Table 2.

II. Posterior Probability

Figure 2 visualizes results of the main effects of the linear mixed-effects regression models for the continuant (Con) and the sonorant (Son) posterior probabilities. The predictors for these models are the same as those for the acoustic parameters shown in Figure 1.

From the first panel of this figure, we see that, as expected, sonorant posterior probability was predicted to be significantly higher for a stop in an unstressed syllable relative to a stressed syllable [β=0.038, p=0.002]. For the continuant probability, the difference was in the same direction, but did not reach statistical significance [β=0.020, p=0.103]. Results for the remaining predictors can be interpreted in a similar fashion and are summarized in Table 2.

Table 2 summarizes the main effects shown in Figures 1 and 2. Significant effects (i.e., those marked

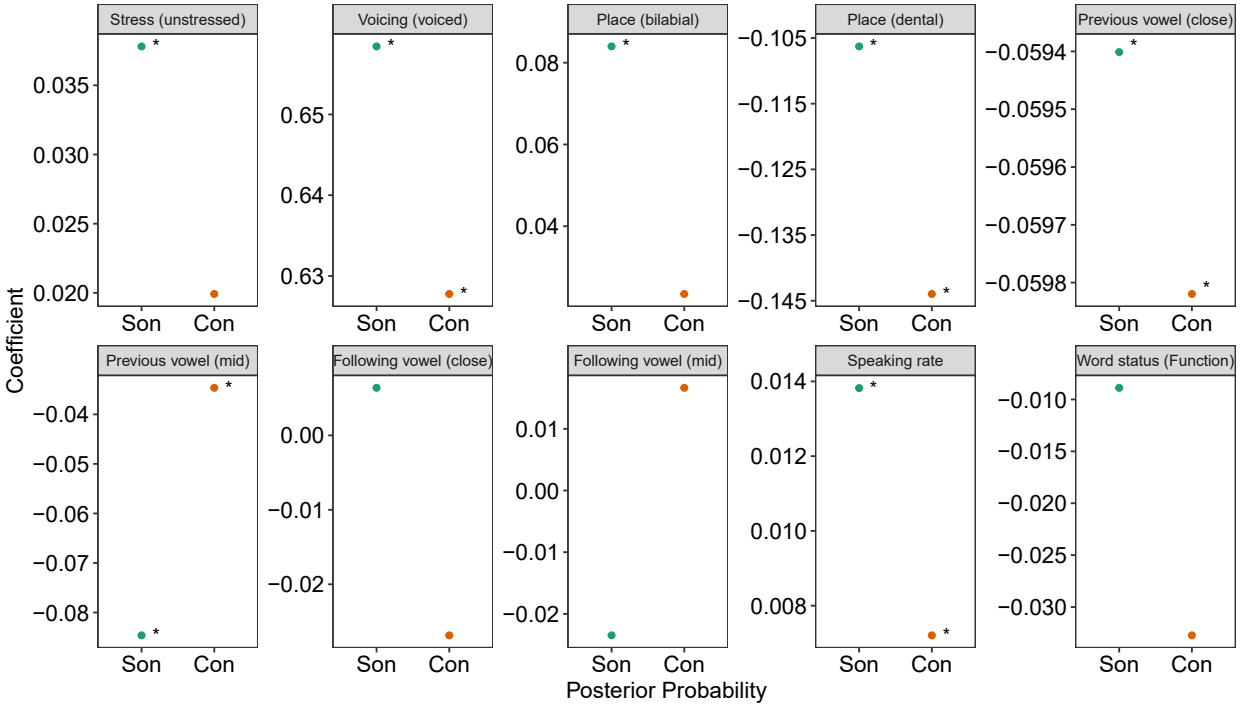

Figure 2: Results of the linear mixed-effects regression models for the sonorant (Son) and continuant (Con) posterior probability.

with * in Figures 1 and 2) are indicated with a (✓) while the (+) and (-) signs indicate whether the results are expected or unexpected based on previous findings on known lenition factors. That is, a higher degree of lenition is expected for a voiced stop compared to a voiceless stop; in an unstressed syllable relative to a stressed syllable; in a faster speaking rate compared to a slower speaking rate, and for a function word compared to a content word. As for the effects of place of articulation, we follow Fowley[12] and predicted the following hierarchy: bilabial<dental<velar. For the effects of vowel height, we predicted the following hierarchy: close<mid<open. Non-significant results are left blank.

Several generalizations can be made from this table. First, no effects of word status on degree of lenition were predicted based on either the acoustic or the posterior probability measures. Second, a higher degree of lenition is predicted for a faster than a slower speaking rate by all measures. Third, among the acoustic parameters, only mean intensity values are consistently in the predicted direction for most lenition factors. Fourth, HNR values are the least consistent with the lenition degree predicted by known lenition factors. Specifically, the expected HNR values were predicted for voicing (higher for a voiced than a voiceless stop) and speaking rate (higher for a fast relative to a slow speaking rate), but not for place of articulation or openness of the surrounding vowels. Similar results were found for relative duration. Surprisingly, relative intensity measures are not in the predicted direction for some factors. Fourth, relative to most acoustic parameters, the effects of known factors on degrees of lenition are more consistent and in the expected direction for sonorant and continuant probabilities. For sonorant probability, the effects of all but two predictors are significant. Relatively fewer predictors significantly and expectedly predicted continuant probability, including voicing, place (a dental is significantly less lenited than a velar), preceding vowel height and speaking rate.

Table 2: Summary of the significant main effects (✓), (+) = expected, (-) = unexpected.

	Relative intensity (re: preceding vowel)	Relative intensity (re: following vowel)	Mean intensity	Relative duration	HNR	Sonorant probability	Continuant probability
Stress	✓+	✓+	✓+			✓+	
Voicing	✓+	✓+	✓+	✓+	✓+	✓+	✓+
Place: bi<den	✓+	✓+		✓+	✓+	✓-	
den<vel	✓+			✓-		✓+	✓+
Preceding vowel: hi<mid	✓-	✓+	✓+	✓-	✓-	✓+	✓+
mid <open	✓-	✓-			✓-	✓+	✓+
Following vowel: hi<mid	✓+	✓+	✓+	✓-	✓-		
mid<open	✓+		✓+	✓+	✓-		
Word status							
Speaking rate	✓+	✓+	✓+	✓+	✓+	✓+	✓+

3. CONCLUSION

Lenition is a gradient phenomenon, affecting different target consonants to a varying degree depending on their place of articulation, position in a prosodic domain, surrounding segments, speaking rate, etc. Traditionally, gradient degree of lenition has been directly quantified along several acoustic dimensions. In this study, a computational approach to measure lenition is introduced and compared to the acoustic metrics. Unlike the acoustic approach where lenition degree is directly reflected on the values of the acoustic measurement, posterior probabilities of relevant phonological features computed by a deep learning neural network is an estimate of lenition degree in the new approach (Phonet). Both approaches were tested on Argentinian stop consonants. Reliability of each approach was measured against lenition patterns predicted by known lenition factors including, voicing, stress, height of flanking vowels, place of articulation, speaking rate and word status (content or function).

The results obtained indicated that, compared to the acoustic measures, sonorant and continuant posterior probability values estimated by Phonet are more consistently in the direction predicted by well-established effects of lenition factors. As expected, a significantly higher sonorant probability, hence a more advanced degree of lenition, is predicted when a stop occurs in an unstressed relative to a stressed syllable. Also as expected, a voiced stop is predicted to be more lenited than a voiceless stop based on the sonorant and continuant probabilities. Significant and expected results were also obtained for other factors including preceding vowel height and speaking rate. Among the acoustic metrics, mean intensity measure is the most consistent and in the expected direction while HNR is the least consistent and largely unexpected. Most but not all predicted results were obtained for relative duration and relative intensity measures. Overall, lenition patterns predicted by the sonorant and continuant posterior probabilities are more consistent with intensity measures than duration measure, or HNR. This is not surprising given that inputs to the Phonet model that generate the sonorant and continuant posterior probabilities are feature sequences based on log-energy of the input signals.

All measures (acoustic and posterior probabilities) indicated that lenition degree increases with speaking rate. In contrast, no difference in lenition was predicted for a stop in a function versus a content word. This is inconsistent with a strong effect of word status on lenition that Broś et al.[3] reported for the Spanish of Grand Canaria. Further investigation is needed to investigate whether dialectal differences or other factors are responsible for this result discrepancy.

In sum, except for the mean acoustic intensity measurement, expected gradient lenition patterns predicted by known variables are more consistently revealed by the posterior probabilities of the sonorant and continuant phonological features generated by Phonet, suggesting that it could more reliably quantify fine-grained degree of lenition than most acoustic metrics.

ACKNOWLEDGMENTS

This research was funded by NSF-National Science Foundation (SenSE). Award No. 2037266 – SenSE.

REFERENCES

[1] George Bailey, *Automatic detection of sociolinguistic variation using forced alignment*, University of Pennsylvania Working Papers in Linguistics: Selected Papers from New Ways of Analyzing Variation (NWAV 44), York, 2016, pp. 10–20.

[2] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker, *Fitting linear mixed-effects models using lme4*, Journal of Statistical Software **67** (2015), no. 1.

[3] Karolina Broś, Marzena Żygis, Adam Sikorski, and Jan Wołłejko, *Phonological contrasts and gradient effects in ongoing lenition in the Spanish of Gran Canaria*, Phonology **38** (2021), no. 1.

[4] Patricio Carrasco, José I Hualde, and Miquel Simonet, *Dialectal differences in Spanish voiced obstruent allophony: Costa Rican versus Iberian Spanish*, Phonetica **69** (2012), no. 3, 149–179.

[5] Eugenio Martínez Celdrán et al., *Cantidad e intensidad en los sonidos obstruyentes del castellano: hacia una caracterización acústica de los sonidos aproximantes*, Estudios de fonética experimental (1984), 71–129.

[6] Uriel Cohen Priva and Emily Gleason, *The causal structure of lenition: A case for the causal precedence of durational shortening*, Language **96** (2020), no. 2, 413–448.

[7] Laura Colantoni and Irina Marinescu, *The scope of stop weakening in Argentine Spanish*, Selected proceedings of the 4th Conference on Laboratory Approaches to Spanish Phonology (Austin, TX, USA), Cascadilla Press, september 2010, pp. 100–114.

[8] Jennifer Cole, José Ignacio Hualde, and Khalil Iskarous, *Effects of prosodic and segmental context on /g/-lenition in Spanish*, Proceedings of the fourth international linguistics and phonetics conference, vol. 2, The Karolinum Press Prague, 1999, pp. 575–589.

[9] Christina Villafaña Dalcher, *Consonant weakening in Florentine Italian: A cross-disciplinary approach to gradient and variable sound change*, Language variation and change **20** (2008), no. 2, 275–316.

[10] David Eddington, *What are the contextual phonetic variants of in colloquial Spanish?*, Probus **23** (2011), no. 1, 1–19.

[11] Mauricio Figueroa and Bronwen G. Evans, *Evaluation of segmentation approaches and constriction degree correlates for spirant approximant consonants.*, ICPhS, 2015.

[12] James Foley, *Foundations of theoretical phonology*, Cambridge University Press Cambridge, 1977.

[13] Adriana Guevara-Rukoz, Isin Demirsahin, Fei He, Shan Hui Cathy Chu, Supheakmungkol Sarin, Knot Pipatsrisawat, Alexander Gutkin, Alena Butryna, and Oddur Kjartansson, *Crowdsourcing Latin American Spanish for low-resource text-to-speech*, (2020).

[14] Naomi Gurevich, *Lenition*, vol. 3, ch. 66, John Wiley & Sons, Ltd, Chester, UK, 2011.

[15] Robert M. Hammond, *The sounds of Spanish: Analysis and application (with special reference to American English).*, ERIC, 2001.

[16] John Harris, Eno-Abasi Urua, and Kevin Tang, *A unified model of lenition as modulation reduction: gauging consonant strength in Ibibio*, Phonology (To appear), (Preprint on PsyArXiv).

[17] J.W. Harris, *Spanish phonology*, M.I.T. Press research monographs, M.I.T. Press, 1969.

[18] Bruce Hayes, *Introductory phonology*, vol. 7, John Wiley & Sons, 2008.

[19] Patrick Honeybone, *Lenition in English*, The Oxford Handbook of the History of English, Oxford University Press, 11 2012.

[20] José Ignacio Hualde, *The sounds of Spanish with audio cd*, Cambridge University Press, 2005.

[21] José Ignacio Hualde, Ryan Shosted, and Daniel Scarpace, *Acoustics and articulation of Spanish /d/ spirantization.*, ICPhS, 2011, pp. 906–909.

[22] José Ignacio Hualde, Miquel Simonet, and Marianna Nadeu, *Consonant lenition and phonological recategorization*, Laboratory Phonology **2** (2012), no. 1, 301–329.

[23] José Ignacio Hualde, *Los sonidos del español: Spanish language edition*, Cambridge University Press, Cambridge, 2013.

[24] Tyler Kendall, Charlotte Vaughn, Charlie Farrington, Kaylynn Gunter, Jaidan McLean, Chloe Tacata, and Shelby Arnson, *Considering performance in the automated and manual coding of sociolinguistic variables: Lessons from variable (ING)*, Frontiers in Artificial Intelligence **4** (2021).

[25] Diederik P Kingma and Jimmy Ba, *Adam: A method for stochastic optimization*, 2014.

[26] John Kingston, *Lenition*, Selected Proceedings of the 3rd Conference on Laboratory Approaches to Spanish Phonology, Cascadilla Press, september 2008, pp. 1–31.

[27] Lisa M. Lavoie, *Consonant strength: Phonological patterns and phonetic manifestations*, Garland, New York., 2001.

[28] Russell V. Lenth, Paul Buerkner, Maxime Herve, Jonathon Love, Hannes Riebl, and Henrik Singman, *emmeans: Estimated marginal means, aka least-squares means [R package]*, 2021.

[29] Anthony Murray Lewis, *Weakening of intervocalic /p, t, k/ in two Spanish dialects: Toward the quantification of lenition processes*, University of Illinois at Urbana-Champaign, 2001.

[30] Maria Del Carmen Lozano, *Stop and spirant alternations: Fortition and spirantization processes in phonology*, Indiana University, 1978.

[31] Lyra Magloughlin, */tɹ/ and /dɹ/ in North American English: Phonologization of a coarticulatory effect*, Ph.D. thesis, Université d'Ottawa/University of Ottawa, Ottawa, may 2018.

[32] Eugenio Martínez-Celdrán and Xosé Luís Regueira, *Spirant approximants in Galician*, Journal of the International Phonetic Association **38** (2008), no. 1, 51–68.

[33] Joan Mascaró and M Aronoff, *Continuant spreading in Basque, Catalan, and Spanish*, Language Sound Structure. Studies in Phonology Presented to Morris Halle by his Teacher and his Students, 1984, pp. 287–298.

[34] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger, *Montreal Forced Aligner: Trainable text-speech alignment using Kaldi.*, Interspeech, vol. 2017, 2017, pp. 498–502.

[35] Jason McLarty, Taylor Jones, and Christopher Hall, *Corpus-Based Sociophonetic Approaches to Postvocalic R-Lessness in African American Language*, American Speech **94** (2019), no. 1, 91–109.

[36] Tomás Navarro Tomás, *Manual de pronunciación española, 1st*, 1977.

[37] Marta Ortega-Llebaria, *Interplay between phonetic and inventory constraints in the degree of spirantization of voiced stops: Comparing intervocalic/b/and intervocalic/g*, Laboratory Approaches to Spanish Phonology (Timothy L. Face, ed.), De Gruyter Mouton, Berlin, 2004, pp. 237–253.

[38] Ayushi Pandey, Pamir Gogoi, and Kevin Tang, *Understanding forced alignment errors in Hindi-English code-mixed speech – a feature analysis*, Proceedings of First Workshop on Speech Technologies for Code-switching in Multilingual Communities 2020 (held online), INCOMA Ltd., october 2020, pp. 13–17.

[39] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022.

[40] Joaquin Romero Gallego, *Gestural organization in Spanish: An experimental study of spirantization and aspiration*, (1995).

[41] Miquel Simonet, José I Hualde, and Marianna Nadeu, *Lenition of /d/ in spontaneous Spanish and Catalan*, Thirteenth Annual Conference of the International Speech Communication Association, 2012.

[42] Antonia Soler and Joaquín Romero, *The role of duration in stop lenition in Spanish*, Proceedings of the 14th International Congress of Phonetic Sciences, vol. 1, The Regents of the University of California Oakland, CA, 1999, pp. 483–486.

[43] Kevin Tang, Ratree Wayland, Fenqi Wang, Sophia Vellozzi, Rahul Sengupta, and Lori Altmann, *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** (2023), no. 2, 1191–1203.

[44] Juan Camilo Vásquez-Correa, Philipp Klumpp, Juan Rafael Orozco-Arroyave, and Elmar Nöth, *Phonet: A tool based on gated recurrent neural networks to extract phonological posteriors from speech.*, Interspeech, 2019, pp. 549–553.

[45] Dan Villarreal, Lynn Clark, Jennifer Hay, and Kevin Watson, *From categories to gradience: Auto-coding sociophonetic variation with random forests*, Laboratory Phonology **11** (2020), no. 1.

[46] Ratree Wayland, Kevin Tang, Fenqi Wang, Sophia Vellozzi, and Rahul Sengupta, *Measuring gradient effects of alcohol on speech with neural networks' posterior probability of phonological features*, Submitted.

[47] Ratree Wayland, Kevin Tang, Fenqi Wang, Sophia Vellozzi, Rahul Sengupta, and Lori Altman, *Lenition measures: Neural networks' posterior probability versus acoustic cues*, The Journal of the Acoustical Society of America **152** (2022), no. 4, A59–A59.

[48] Jiahong Yuan and Mark Liberman, *Investigating /l/ variation in English through forced alignment*, Proceedings of Interspeech 2009 (Brighton, UK), International Speech Community Association (ISCA), september 2009, pp. 2215–2218.

[49] _____, *Automatic detection of "g-dropping" in American English using forced alignment*, 2011 IEEE Workshop on Automatic Speech Recognition & Understanding, IEEE, 2011, pp. 490–493.

[50] _____, */l/ variation in American English: A corpus approach*, Journal of Speech Sciences **1** (2011), no. 2, 35–46.