Interspeech 2024
1-5 September 2024, Kos, Greece

Information-theoretic hypothesis generation of relative cue weighting for the voicing contrast

Annika Heuser[1], *Jianjing Kuang*[1]

[1]University of Pennsylvania, USA

aheuser@sas.upenn.edu, kuangj@sas.upenn.edu

Abstract

To learn the voicing contrast, children must identify which of the available perceptual cues are helpful in different contexts. Using Standard American English (SAE) as a case study, we generated hypotheses of which cues are the most informative for different contexts, such as onsets vs. codas. More specifically, we classified SAE obstruents as voiced vs. voiceless using decision trees trained and tested on TIMIT. We validated the feature importances of different contexts against the findings of previous perceptual studies and we gleaned more specific hypotheses to help design future experiments on children's acquisition of the voicing contrast.

Index Terms: phonemic perception, relative cue weighting, acquisition, voicing contrast

1. Introduction

Adults easily perceive the phonological contrasts of their native language, despite the multidimensional nature of the phonetic signal to phonological contrast mapping problem. This multidimensionality in part stems from many perceptual cues being available for phonological contrasts. For example, the voicing contrast for Standard American English (SAE) obstruents involves at least 16 acoustic cues [1]. Another major contributor to the multi-dimensionality is that for the same phonological contrast, cue weightings might change depending on the context. Continuing with SAE voicing examples: SAE voiced stops are often truly voiced (with pre-voicing) in intervocalic positions, but tend to be voiceless word initially [2]. Therefore, to fully acquire a phonological contrast, children need to acquire all its contextual variation. This presents a gap in the acquisition literature because the contextual variation is not typically studied. Accordingly, we must investigate how children learn a single phonological category from various phonetic realizations in different contexts. As a first step, we computationally model how contextually varied cue weightings can be learned from the speech signal.

By the same philosophy, Rhee et al. [3] computationally modeled the acquisition of cue weighting for lexical pitch perception. To weight F0 and spectral cues, they used three classification algorithms: linear discriminant analysis, support vector machines, and random forests. As opposed to determining the cue weights at just one time point, they tracked them over time with a corpus consisting of semi-spontaneous speech from children and adults. This study demonstrated a developmental curve of cue integration for language acquisition. Children acquire the primary cue at a younger age, and children older than 6 years old learn to utilize more phonetic cues to distinguish the contrast more efficiently.

In this paper, we use decision trees, which are closely related to the random forests used by Rhee et al. [3], in that a random forest combines many decision trees. While random forests tend to perform better, in terms of both accuracy and generalization, decision trees are more interpretable [4]. Because our goal is to make predictions that we can test via perception experiments, we opted for the decision tree algorithm. Building decision trees also requires fewer computational resources than random forests. If we are able to achieve sufficient classification performance and the cues picked out to be the most informative are supported by the literature, then decision trees are the most efficient tool for our purposes.

We focus on the SAE voicing contrast as a case study to validate this method. We investigate how the cue weights corresponding to adult voicing perception change across different contexts, such as syllable position (onset vs. coda) and for different manners of articulation (stops vs. fricatives). [5, 6, 7] provide evidence that these contexts impact the phonetic realization of the contrast. We chose this case study especially because of the substantial literature that we have only cited a small subset of in this section.

2. Methods

2.1. Corpus and acoustic features

How effective any classification algorithm can be depends on the data quality and features it has to work with. Our data is from TIMIT [8], a corpus of sentences read by 630 SAE speakers. We consider this data high-quality because the audio has minimal background noise and the TextGrids are hand-corrected. We only chose features that were available for all segments/contexts because we cannot effectively compare context-specific feature rankings to overall feature rankings if the context-specific rankings include additional features.

We extracted consonant duration, which is tied to the distinction of voiced vs voiceless fricatives, and voiced vs voiceless postvocalic consonants [9]. In the case of all word-initial and some word-final stops, the consonant duration is just the voiced onset time (VOT), because the closure duration is separately segmented and labeled in the TIMIT TextGrids. The consonant duration of stops followed by other consonants (e.g. the "k" in "dark suit") does not correspond to VOT in TIMIT. VOT is well-known as an effective cue for initial stop voicing (e.g. [10, 11]); however, unlike the more general feature, consonant duration, it is often missing in some contexts (e.g. syllable coda position which sometimes only has closure duration). To effectively capture pre-voicing, we also extract the proportion of voicing during oral closure (partial voicing hereafter). To calculate partial voicing, we detect periodic pulses during the oral closure interval and calculate the proportion of periodic frames. It is important to note that our Praat script returns the

10.21437/Interspeech.2024-2340

inverse value, i.e. the proportion that is not voiced, but we still refer to it is as partial voicing. While we would only be able to measure VOT values for some, but not all stops, we can extract partial voicing values for consonants in every word position. Partial voicing may also be a more informative value because the amount of voicing realization in English depends on several factors, such as word position, stress patterns, and adjacent phonemes [12]. Another voicing cue only available for stops is closure duration [13]. To utilize the closure durations in the TIMIT TextGrids, we simply record values of 0 for non-stop obstruents, as well as for stops that are sentence initial, where the closure duration is impossible to detect.

We also extract f0 and formant transitions (at 5% vowel duration for onsets and 95% for codas), and the formant averages, of the adjacent vowels, and the vowel durations themselves. Onset f0 differs significantly between voiced and voiceless stops in English [14], so we expect this to be an informative cue at least for one of our contexts of interest. [7] and [15] provided evidence that formant transitions of initial stops covary with VOT and impact the perception of the voicing contrast in English. Finally, vowel duration is important to the perception of voiced and voiceless coda segments [16].

2.2. Data Pre-processing

We extracted each segment's phonemic label, start and end times based on the TextGrid, duration, partial voicing (max period factor= 1.3, max amplitude factor= 1.6, all other parameters are Praat Pitch defaults), average f0 (male range= 75-250Hz, female range= 100-300Hz, time step= 0.001s, otherwise all defaults), average F1-F3 (time step = 0.01s, max number of formants = 5, formant ceiling = 5500 Hz, window length = 0.025s, pre-emphasis from = 50 Hz), and f0 and F1-F3 at 5%, 50%, and 95% duration using Praat.

Next, we assigned syllable position values based on whether a segment aligned with the beginning or end of TIMIT's TextGrid word tier. In other words, we only labeled segments at the beginning or end of words, as onsets and codas, respectively. Not just the most word-initial and word-final consonants, but all consonants part of a complex consonant cluster at the beginning or end of a word, were assigned a syllable position value. Word-medial segments were all discarded because of evidence that word edges contain enough information for "phonological bootstrapping" of higher level categories, like lexical category [17, 18]. We also discarded words that did not contain a vowel, even if they did contain syllabic consonants, because we wanted all the consonants to have comparable adjacent vowel data. For consonants in the onset, we found the nearest vowel to the right, and appended the f0 and F1-F3 at 5% vowel duration. For consonants in the coda, we found the nearest vowel to the left, and appended the f0 and F1-F3 at 95% vowel duration. Regardless of whether a consonant has an onset or coda label, we appended the average f0 and F1-F3 of the nearest vowel to the right or left, respectively. We also discarded consonants that had any adjacent vowel values that were undefined. Finally, we excluded all segments that were not part of a voiced/voiceless stop or fricative pair. This ultimately left us with 34,043 data points.

In addition to the onset vs. coda label, we assign consonants a boolean value corresponding to whether the consonant is part of a complex consonant cluster. We do not differentiate between a consonant at the beginning, middle, or end of a complex consonant cluster–they are all assigned the same complex consonant cluster boolean value, as well as the same adjacent vowel values. Aside from these categorical values, we z-score normalized all features across each speaker. Normalizing the data is a standard pre-processing step for classification algorithms. However, our normalization across speakers is also justified by evidence that human listeners perform online speaker normalization (e.g. [19, 20, 21]). In our final dataset, 48.96% of consonants are voiced and 51.04% are unvoiced. This data balance is likely a byproduct of the fact that TIMIT was designed as a phonetically balanced speech corpus, which is another reason that is well-suited for this study. It is fortunate that the data is still so balanced after pre-processing, seeing as imbalanced data can pose a problem for classification algorithms [22].

2.3. Decision Tree Analysis

We built every decision tree on 90% of the data, in order to hold out 10% for testing. Using sklearn's DecisionTreeClassifier, we could choose between more than one split criterion, and tried both Gini impurity and Shannon entropy. Given that both criteria determine how homogeneous the resulting two groups are, with respect to the target classification category, we did not expect them to make a significant difference in testing accuracy or to change the rankings based on feature importance. We verify this in section 3. We also explored different maximum tree depth constraints. When unconstrained by such a parameter, the algorithm will continue splitting the training data until it has achieved 100% accuracy. This is often an overfit of the training data, resulting in sub-optimal performance on the held-out test data. We searched for the split criterion and depth that allowed us to achieve the best test accuracy. We averaged the test accuracies for each criterion/depth combination over 10 random train-test data split iterations. Using the combination we found to produce the best tree in terms of test accuracy, we extracted the normalized total reduction of the split criterion for each feature (e.g. partial voicing, consonant duration, average adjacent vowel f0, etc.). We recorded these values for each of 10 random train-test splits and averaged across the individual features.

To model the contexts that result in different phonetic realizations of the voicing contrast, we built trees in which we manually specified the first split to be based on one of our categorical features, e.g. syllable position, manner of articulation, etc. We use splitting on syllable position as our example to explain how we manually set a feature for our first split. We first created our 90-10% train-test split and then divided the training and test data into two groups consisting of only onset data or only coda data. We then run the same sklearn DecisionTreeClassifier algorithm based on just the onset or just the coda data. The preprocessed data is made up of 56.81% onset and 43.19% coda consonants, 47.57% stop and 52.43% fricative consonants, and 34.75% consonants that are part of complex consonant clusters and 65.25% that are not. This means that the syllable position and manner of articulation initial splits are likely to be relatively even, and the resulting subtrees will be built based roughly the same amount of training data.

However, because the test data is not as likely to break into even subgroups based on the categorical feature, we weighted the accuracies of the two sub-trees based on the percentage of the test data that they classify. To clarify this statement, we again precede with the syllable position initial split example: If $\frac{2}{3}$ of the test data was from an onset, then we would calculate the onset decision tree's accuracy from $\frac{2}{3}$ of this data and the coda decision tree's accuracy from the other $\frac{1}{3}$. The test accuracies of the onset and coda sub-trees would then be multiplied by $\frac{2}{3}$ and $\frac{1}{3}$, respectively, and summed. While the test accuracy of the

individual sub-trees is also interesting, we weight them to be able to verify that they are about as accurate as decision trees in which we did not specify the initial feature (or two). In contrast, we are interested in how the rankings of the remaining features differ between only onset data and only coda data, so we report the raw feature importances of each sub-tree, without weighting or otherwise aggregating them.

We can manually specify the second tree level in much the same way that we specify the first. After setting the initial split to be based on syllable position, for example, we can then set the second to be based on manner of articulation. To do this, we divide the onset and coda train and test groups into two further groups, namely stops and fricatives. This results in eight groups, half train and half test, from which we generate four sub-trees: 1) onset-stop, 2) onset-fricative, 3) coda-stop, and 4) coda-fricative. We weight the test accuracies of each sub-tree and take the raw feature importances as before. We report the results of implementing this very example in section 3.

3. Results

Figure 1 demonstrates that the split criterion (Gini impurity vs. Shannon entropy) does not make a difference in the test accuracy, but the tree depth constraint does.

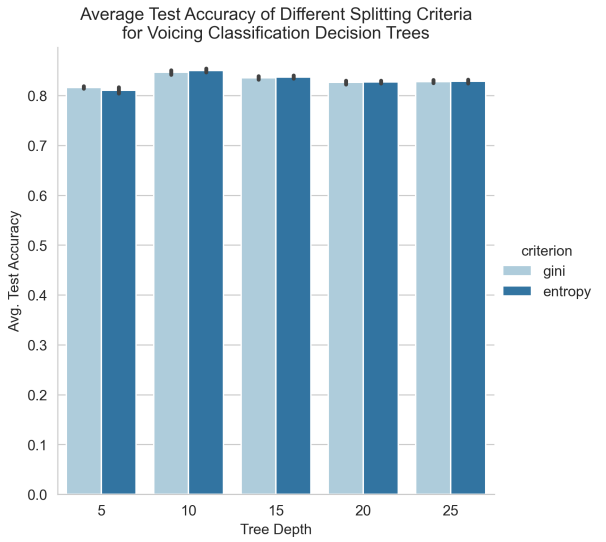

Figure 1: *Decision trees with depths of about 10 have the best test accuracy.*

We searched more precisely between depths 6 and 15, and found the optimal depth to be 10, resulting in an average test accuracy of 85.02% with the Shannon entropy split criterion. It is important to note that the best depth and criterion change depending on the random seeds, meaning that neither is crucial to the results so long as the depth is around 10. This is also true when we manually specify the first feature or two. In fact, we found the optimal depth of the subtrees from first splitting on syllable position and then manner of articulation to be 7. Seeing as there are two manually specified levels above the subtrees, the total maximum depth is 9. The average test score weighted across the four subtrees is 85.81%, which is very similar to that of the optimal tree when we did not manually specify anything.

The criterion does not affect the feature rankings, as shown in figure 2. Figure 2 also demonstrates that partial voicing (re-

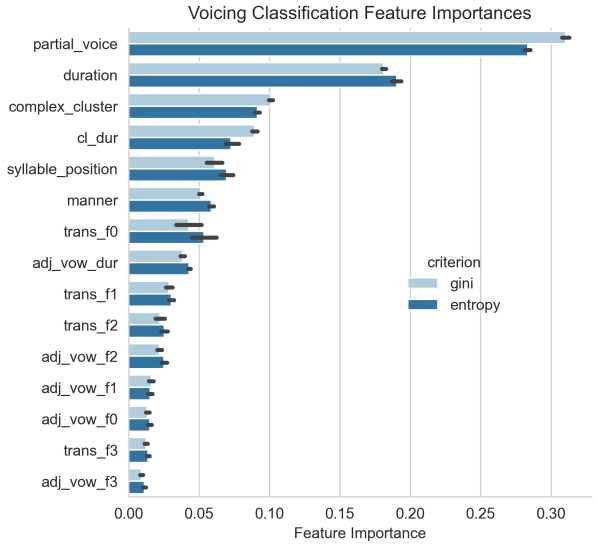

Figure 2: *The feature importance ranking does not change if we use a different split criterion.*

ferred to as `partial_voice` in the figures) is consistently an important cue for phonemic voicing classification, as we would expect, seeing as it is a measure of phonetic voicing. The complex consonant cluster boolean (`complex_cluster`) is also a high ranking feature, suggesting that it may split the data into different contexts across which children learn separate cue weightings. We therefore decided to determine the feature rankings for solo consonants vs. consonants part of a complex consonant cluster, and discuss them later in this section.

For the sake of validating the decision tree method as a means of generating hypotheses of contextual variation of cue weighting, we calculated the feature rankings for four contexts. These contexts corresponded to each sub-tree resulting from manually specifying the first two levels to 1) syllable position and 2) manner of articulation. Many perceptual experiments only correspond to one of these four contexts, making the sub-tree feature rankings the best means of validation. Looking to the onset sub-plot, we find consonant duration (referred to as `duration` in all figures) to be by far the most important feature for both stops and fricatives. The consonant duration of an onset stop in TIMIT is simply its VOT, which is a well-known predictor of voicing in stops [23, 24]. For fricatives, consonant duration significantly differs between voiced and voiceless onset fricatives [25]. We expect the overall consonant duration to heavily correlate with frication duration which has been found to significantly affect voiced/voiceless perception in adults [26]. It therefore makes sense that consonant duration is also the second most important feature for coda fricatives.

Turning now to the coda subplot, we highlight that adjacent vowel duration (`adj_vow_dur`) is particularly important for coda stops. Peterson & Lehiste [27] found that vowel duration before a voiced consonant was consistently longer for both stops and plosives. However, while vowel duration may be a decently predictive feature for fricative codas as well, consonant duration outranks it. Also consistent with previous perception studies, closure duration is a very weak cue for coda voicing [28]. Altogether, the results presented in figure 3 are largely in line with the findings from perceptual studies on the SAE voicing contrast. Therefore, our model successfully simulated the

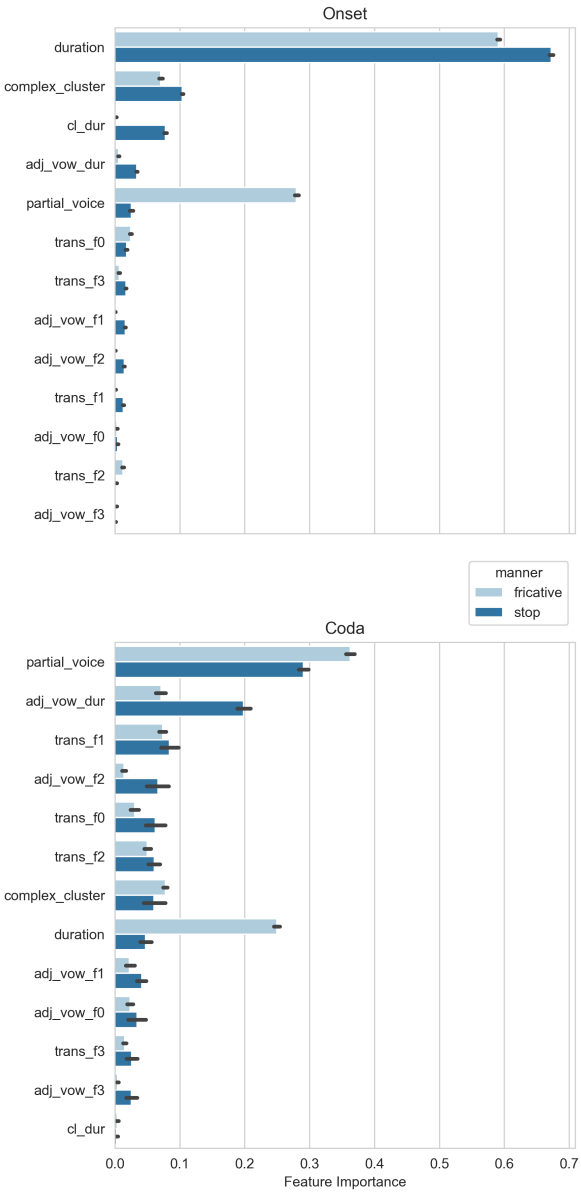

Figure 3: *The relative feature importances from first splitting on syllable position and then on manner of articulation.*

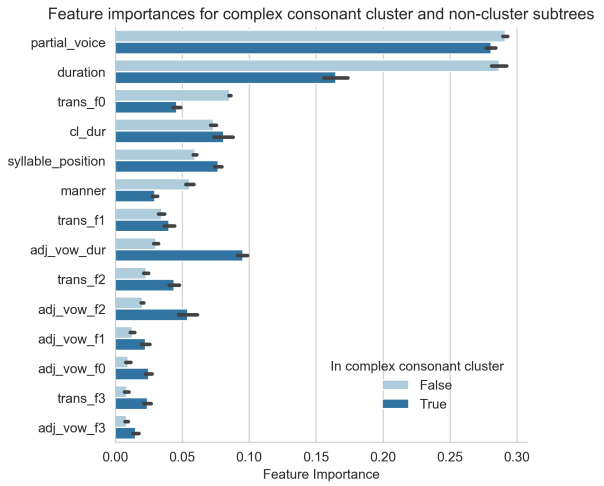

Figure 4: *The relative feature importances from splitting on the complex consonant cluster boolean.*

contextual variation of perceptual cue weighting for the voicing contrast. Unlike typically well-controlled perception experiments, which can only focus on a very limited set of cues (2-3 cues at most), we were able to test a more comprehensive set of cues. Our model also generates a number of novel findings and presents an opportunity for experimental verification. To name a few here, we found that for both fricatives and stops, partial voicing is consistently the most important cue for coda voicing contrast, outranking the well-studied cues such as vowel duration and closure duration. Formant transitions are more informative for coda voicing than for onset voicing. It is also interesting that fricatives and stops have similar cue weightings across onsets and codas, suggesting that they may have the same contextual effect on voicing.

We present additional novel findings based on figure 4. Interestingly, adjacent vowel duration (`adj_vow_dur`) is much stronger for consonants in complex consonant clusters than for solo consonants. For solo consonants, the partial voicing and consonant duration have roughly the same feature importance, while consonant duration is much weaker for consonants in complex consonant clusters. We might then hypothesize that partial voicing and consonant duration, but neither alone, allows for sufficient voicing perception of solo consonants, for example. This could explain why children more quickly learn to classify non-cluster segments [29] because they don't need to integrate as many cues. This can be proven or dismissed with a well-designed perceptual experiment.

4. Discussion

Decision trees perform well enough on classifying the SAE voicing contrast, based on the mean test accuracy of ∼85% generally, and across all the contexts we tested. The features they pick out as important for voicing classification are corroborated by the literature. There are, of course, more contexts that we could have investigated, like place of articulation, intervocalic, etc., in order to generate additional hypotheses that can already be confirmed or denied by the literature. Nonetheless, the feature rankings of the contexts we did investigate gave rise to novel predictions on the well-studied SAE voicing contrast.

There are many opportunities for further research with this approach, now that we have demonstrated that our decision trees generate cue weightings consistent with human perception. For example, we could quantify the effects of adding or subtracting features on the classification accuracy. If we only provided the model with the top feature whose importances sum to at least 0.5, for example, could we achieve comparable accuracy to using all the features? In future work, we will answer this question and determine how many cues are necessary to model human voicing perception in various contexts. Another direction of future research involves generating hypotheses for languages with different voicing contrasts, like Spanish, a true voicing language [30]. Because building decision trees does not require many computational resources, we can use this tool to easily simulate human perception in a wide range of contexts and across many languages.

5. References

[1] L. Lisker, ""Voicing" in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees," *Language and speech*, vol. 29, no. 1, pp. 3–11, 1986.

[2] L. Davidson, "Variability in the implementation of voicing in American English obstruents," *Journal of Phonetics*, vol. 54, pp. 35–50, 2016.

[3] N. Rhee, A. Chen, and J. Kuang, "Going beyond f0: The acquisition of Mandarin tones," *Journal of Child Language*, vol. 48, no. 2, pp. 387–398, 2021.

[4] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random forests and decision trees," *International Journal of Computer Science Issues (IJCSI)*, vol. 9, no. 5, p. 272, 2012.

[5] T. Cho, D. H. Whalen, and G. Docherty, "Voice onset time and beyond: Exploring laryngeal contrast in 19 languages," *Journal of Phonetics*, vol. 72, pp. 52–65, 2019.

[6] J. Steffman, "Phrase-final lengthening modulates listeners' perception of vowel duration as a cue to coda stop voicing," *The Journal of the Acoustical Society of America*, vol. 145, no. 6, pp. EL560–EL566, 2019.

[7] J. R. Benkí, "Place of articulation and first formant transition pattern both affect perception of voicing in English," *Journal of Phonetics*, vol. 29, no. 1, pp. 1–22, 2001.

[8] J. S. Garofolo, "Timit acoustic phonetic continuous speech corpus," *Linguistic Data Consortium, 1993*, 1993.

[9] D. H. Klatt, "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence," *The Journal of the Acoustical Society of America*, vol. 59, no. 5, pp. 1208–1221, 1976.

[10] D. B. Pisoni and J. Tash, "Reaction times to comparisons within and across phonetic categories," *Perception & psychophysics*, vol. 15, no. 2, pp. 285–290, 1974.

[11] P. A. Keating, M. J. Mikoś, and W. F. Ganong III, "A cross-language study of range of voice onset time in the perception of initial stop voicing," *The Journal of the Acoustical Society of America*, vol. 70, no. 5, pp. 1261–1271, 1981.

[12] L. Davidson, "Patterns of voicing in American English voiced obstruents in connected speech." in *ICPhS*, 2015.

[13] L. Lisker, "Closure duration and the intervocalic voiced-voiceless distinction in English," *Language*, vol. 33, no. 1, pp. 42–49, 1957.

[14] O. Dmitrieva, F. Llanos, A. A. Shultz, and A. L. Francis, "Phonological status, not voice onset time, determines the acoustic realization of onset f0 as a secondary voicing cue in Spanish and English," *Journal of Phonetics*, vol. 49, pp. 77–95, 2015.

[15] K. N. Stevens and D. H. Klatt, "Role of formant transitions in the voiced-voiceless distinction for stops," *The Journal of the Acoustical Society of America*, vol. 55, no. 3, pp. 653–659, 1974.

[16] L. J. Raphael, "Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English," *The Journal of the Acoustical Society of America*, vol. 51, no. 4B, pp. 1296–1303, 1972.

[17] A. Christophe Teresa Guasti Marina Nespor, "Reflections on phonological bootstrapping: Its role for lexical and syntactic acquisition," *Language and cognitive processes*, vol. 12, no. 5-6, pp. 585–612, 1997.

[18] L. Onnis and M. H. Christiansen, "Lexical categories at the edge of the word," *Cognitive Science*, vol. 32, no. 1, pp. 184–221, 2008.

[19] P. Ladefoged and D. E. Broadbent, "Information conveyed by vowels," *The Journal of the acoustical society of America*, vol. 29, no. 1, pp. 98–104, 1957.

[20] K. Johnson and M. J. Sjerps, "Speaker normalization in speech perception," *The handbook of speech perception*, pp. 145–176, 2021.

[21] M. J. Sjerps, H. Mitterer, and J. M. McQueen, "Listening to different speakers: On the time-course of perceptual compensation for vocal-tract characteristics," *Neuropsychologia*, vol. 49, no. 14, pp. 3831–3846, 2011.

[22] D. Ramyachitra and P. Manikandan, "Imbalanced dataset classification and solutions: a review," *International Journal of Computing and Business Research (IJCBR)*, vol. 5, no. 4, pp. 1–29, 2014.

[23] A. S. Abramson and D. H. Whalen, "Voice Onset Time (VOT) at 50: Theoretical and practical issues in measuring voicing distinctions," *Journal of phonetics*, vol. 63, pp. 75–86, 2017.

[24] M. A. Zlatin, "Voicing contrast: perceptual and productive voice onset time characteristics of adults," *The Journal of the Acoustical Society of America*, vol. 56, no. 3, pp. 981–994, 1974.

[25] S. R. Baum and S. E. Blumstein, "Preliminary observations on the use of duration as a cue to syllable-initial fricative consonant voicing in English," *The Journal of the Acoustical Society of America*, vol. 82, no. 3, pp. 1073–1077, 1987.

[26] R. A. Cole and W. E. Cooper, "Perception of voicing in English affricates and fricatives," *The journal of the acoustical society of America*, vol. 58, no. 6, pp. 1280–1287, 1975.

[27] G. E. Peterson and I. Lehiste, "Duration of syllable nuclei in English," *The Journal of the Acoustical Society of America*, vol. 32, no. 6, pp. 693–703, 1960.

[28] J. Penney, F. Cox, and A. Szakay, "Effects of glottalisation, preceding vowel duration, and coda closure duration on the perception of coda stop voicing," *Phonetica*, vol. 78, no. 1, pp. 29–63, 2021.

[29] S. McLeod, J. Van Doorn, and V. A. Reed, "Consonant cluster development in two-year-olds," 2001.

[30] L. Williams, "The perception of stop consonant voicing by Spanish-English bilinguals," *Perception & Psychophysics*, vol. 21, no. 4, pp. 289–297, 1977.