INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Using Random Forests to classify language as a function of syllable timing in two groups: children with cochlear implants and with normal hearing

Mark Gibson[1], *Ferenc Bunta*[2], *Charles Johnson*[3]

[1]Universidad de Navarra
[2]University of Houston
[3]University of California Santa Barbara

mgibson@unav.es, fbunta@central.un.edu, charles.addisonj@gmail.com

Abstract

We trained a series of Random Forest models in a supervised learning environment on different temporal parameters related to syllable structure: voice onset time (VOT), vowel duration following simplex and complex onsets, and lateral duration in word initial (/lV) position and as the second consonant in a C1C2 cluster (where C means consonant). Capitalizing on previous work we trained the models on data from monolingual Spanish- and English-speaking adults. We asked whether the timing productions used by bilingual children with normal hearing (NH) and children with cochlear implants (CI) can be classified as pertaining to the same timing system (i.e. language), or whether the children are applying the same basic timing plan to two different languages. We also asked whether there were differences between the CI and NH groups. Our results indicate that the children from both groups produce qualitatively distinct timing plans for each language with no interference from the other language.

Index Terms: speech timing, syllable structure, Random Forest

1. Introduction

While a dearth of research in the past decades has addressed language interaction/interference in bilingual adults and children (including both simultaneous and sequential bilinguals) (see [1-7]), very few studies have explored the relationship between language and hearing loss in bilinguals, especially with regard to timing.

The timing of segments varies as a function of the position it occupies within a syllable. Segments appearing in word initial onsets, for example, are longer than the same segment in codas [8-10], and movement displacement is greater for onsets than codas [see among others 11,12]. Further, the complexity of the onset also modulates the temporal parameters of a segment. For example, a number of studies have shown that the duration of word initial /l/ is significantly longer than /l/ in a C1C2 clusters where C2 is /l/ ([13] and many studies since).

Less is known regarding the development of syllable timing systems, especially in bilinguals and atypical populations, such as children with hearing loss who use cochlear implants. In [14] and [15], the authors addressed the development of certain syllable timing parameters by bilingual English- and Spanish-speaking children with cochlear implants and their peers with normal hearing. However, these studies examined the timing productions for each language separately in order to address effects of hearing loss on the development of the timing parameters, and did not compare the results of the individual languages in order to address the degree of integration of the two phonological systems, which is our objective here.

We chose the specific parameters – voice onset time (VOT), vowel duration and lateral duration for our models for various reasons. First, VOT for voiceless stops is modulated by place of articulation in both Spanish and English [16-22], whereby velars (/k/) exhibit longer VOTs than labial (/p/) and coronal (/t/) stops. However, VOT for voiceless stops is generally known to be significantly longer in English than in Spanish (see [16]). Additionally, VOT in English has been shown to lengthen as a function of syllable complexity (i.e. VOT increases as more consonants are added to the onset) [23,24], whereas results from two different studies addressing the effects of onset complexity on VOT in Spanish reveal no effects [19,20]. In their studies with bilingual children, [14] and [15] found that both the CI and NH groups tended to lengthen the VOT in complex onsets, which is a non-attested pattern by Spanish adults.

With regard to vowel duration, adult patterns from both English and Spanish show vowel compression (a shortening of the vowel) following complex onsets as compared to vowels following singleton consonants (see [13], [25,26]). However, [14] and [15] (for the same subjects we are using for the Random Forest analyses) found a lengthening effect of the vowel following complex onsets, an unobserved pattern in both English and Spanish.

As regards lateral duration in both word initial singleton onsets and as the second consonant in stop+lateral clusters, previous studies in both Spanish and English (among other languages) have revealed a shortening, or compression, effect whereby laterals in complex onsets are shorter than laterals in word initial singleton onsets (see among others [27], [25], [28]). In their studies with bilingual children, [14] and [15] found that both the CI and NH groups in both languages produced the compression effect typical of these languages, though in both languages the CI group showed higher temporal stability (as measured by coefficients of variation) than the NH group, though there were magnitude differences across the languages.

To address the topic of whether the children access two qualitatively distinct timing systems, or apply the same timing scheme to two different languages with (or without) parametric modifications, we trained Random Forest models in a supervised learning environment with data (for clusters /bl, ɡl, pl, kl, tr, kr, pr/, and singleton consonants /p, t , k, b, d, ɡ, l/ for each language) from adult monolingual English and Spanish speakers (models will be explained in full in the following section 2). After training, we introduced the children's production data whereupon the model classified the children's

10.21437/Interspeech.2023-1931

input into groups (Adults vs children, cochlear implant group vs the normal hearing group, and English vs Spanish).

1.1 Speech materials for model training and inputs

Speech materials were collected from 22 (11 CI, 11 NH) bilingual English- and Spanish-speaking children in the Houston, Texas metropolitan area. The mean chronological age for the NH group was 5 years and two months, which was the mean hearing age of the CI group (hearing age is the time since implantation). The mean chronological age of the CI group was 6 years and 7 months.

The NH participants completed a hearing screening using pure tones at 0.5, 1, 2, and 4 kHz at 25 dB HL, bilaterally before data collection. Immediately prior to data collection, the cochlear implants of the CI group were reported to be functioning normally, with no issues reported throughout the course of data collection.

All cochlear implant participants were receiving or had received Auditory Verbal therapy (see [2]), a treatment approach that relies on oral communication as its sole communication mode. Only one mother reported limited sign language for her child, while the rest of the parents reported that their children used spoken language.

A corpus of 80 real English and Spanish words was elicited using a picture naming task. The word list was developed by [2] (among others) and was designed to gauge the phonological skills of children between the ages of 3 and 8 years. The target words depict items familiar to young children and have been used extensively with both monolingual and bilingual children.

Timestamps (marks along the x-axis that register the moment in time for a particular event) were marked on the acoustic file for the acoustic offset of C1 / acoustic onset of VOT (burst) (a in Figure 1), acoustic onset of voicing or acoustic onset of C2 (in the case of complex onsets, release of C2 (b in Figure 1), acoustic onset of vocalic target (c in Figure 1), and acoustic offset of vocalic target (d in Figure 1). Timestamps and labels are illustrated for a token 'glasses' in Figure 1.

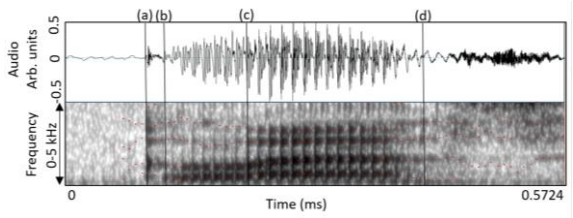

Figure 1. *Spectrogram (bottom panel) and waveform (top panel) of the word glasses [glæ.səz].*

2. Model Parameters

Random Forest classifiers are a machine-learning model that takes the consensus of a number of decision trees to determine the probability of a single instance of data belonging to a particular class. Each decision tree uses a set of Boolean conditions on features (such as /l/-duration < = 45) to classify data into one of a number (in this case 2) classes (EN/SP, CI/NH, Adults/children). They are a powerful machine-learning model that may be trained on relatively few data points when compared with other models such as Deep Neural Networks. Random Forests, like other machine learning models trained to perform data classification, partition the feature space into a number of categories. However, many other classification models (including a standard statistical analysis that relies upon the data being normally distributed within each class) do not have the flexibility of the Random Forest classifiers when partitioning the feature space, as they are often limited to partitioning via hyperplanes. Finally, Random Forest classifiers have a simple measure of feature importance. Since, there are a set number of decision trees in the Random Forest, and each has a set number of Boolean conditions, one may simply take the percentage of Boolean conditions related to each feature as that feature's importance in classification.

For our models, we used scikitlearn's RandomForestClassifer class to implement our Random Forests [29] in Python. Each of our Random Forest Classifiers consisted of 100 Decision trees, each of which had a maximum depth of 2 (only two Boolean conditions maximum were allowed to classify any piece of data).

Five experiments were created to compare 1) English productions between NH and CI groups, 2) Spanish productions between NH and CI groups, 3) English productions of NH and CI groups with adult native English patterns, 4) Spanish productions of NH and CI groups with adult native Spanish patterns, and 5) English and Spanish (mixed) productions of NH and CI groups. Model 5 is of most interest here, as it classifies the test input into language (though results of all models will be addressed).

The models were trained on data from monolingual English published in [25], and [23] for VOT) and Spanish speakers published in [19, 20]. After the models were trained on the data for the respective languages, validation data from the children were introduced and the models searched for a two-category solution (for group and language). Different models were trained for C and C1C2 onsets. In addition to the variables VOT, vowel duration and C2 duration presented here, a fourth variable was introduced. As the Spanish productions were typified by an open transition (where there is a latency from the release of C1 to the onset of C2, say between /p/ and /l/ in a word *playa*, 'beach', for example), [20] were able to quantify values in ms for the interconsonantal, or interplateau interval (henceforth, ICI, or IPI), between C1 and C2 (as seen in Figure 2).

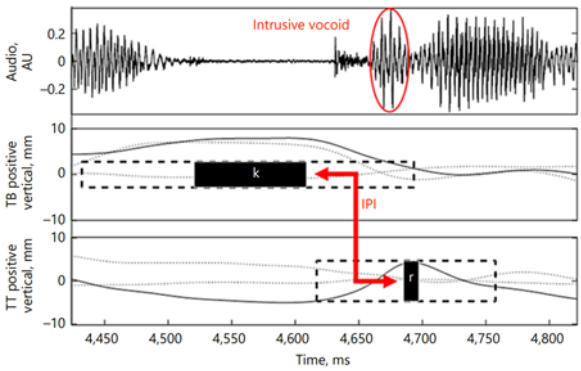

Figure 2. *From top to bottom are the waveform, tongue back (TB) and tongue tip (TT) signals for token "crema" The red ellipses delimits the intrusive vocoid. The black rectangles represent the consonants in the cluster. The lag from /k/ to /r/ in the bottom panel is the interconsonantal/interplateau interval, orICI/ IPI, as indicated in red.*

However, in English, the constriction plateaus for C1 and C2 overlap temporally (see among others, [25]), meaning there is no way to measure acoustically the interconsonantal interval, or latency, between say, a /p/ and /l/ in the word *play* in English. In Spanish, there is oftentimes an intrusive vocoid that appears between the consonants in clusters, and thus we could approximate the distance between the release of C1 and the onset of C2 using acoustic data. In our auditory signals for the children's English productions, we found no evidence for open transitions (where C1 ends before C2 begins), the end of C1 and the beginning of C2 were directly adjacent. As we could not directly measure how much overlap there was for the English productions using acoustic data, for our models all interconsonantal intervals/overlap for English clusters were labeled as -1 ms, which is the minimum that the two consonants could overlap, meaning that any error is maximally conservative.

3. Results

Results of the models were evaluated based on the best F-score, the weighted average of precision and recall at the optimal F-score threshold. The best F-score basically expresses how good the classifier was at categorizing the data. It is preferable to accuracy as it controls for asymmetric class representation in each experiment. Later, another evaluator, precision at best F-score represents the ratio of correct classifications to incorrect classifications (i.e., out of everything that was categorized as group A (hypothetical group), how many actually belonged to group A). Another model evaluator, recall at best F-score, represents the ratio of correct classifications to overall classifications (i.e., how much of group A (hypothetical group) could actually be classified). The accuracy at best F-score conveys the number of correct classifications divided by the number of total classifications. Finally, the relative importance values express, as a percentage, the frequency of use of a particular variable in categorization.

By and large, for each language, the models were able to clearly classify the adults from the children's productions as shown in Figure 3, and also distinguish between the NH and CI groups (as shown in the following Figure 3). For the first model which compares the children's English productions to address differences based on group, vowel duration (59%) (where % expresses importance value) and VOT (26%) exhibit the most relevant relative importance values for CV syllables, while /l/ duration (47%) is the most important distinguisher in /Cl/ syllables. This is not intuitively predictable from the results reported in [14] and [15], given the small intergroup variation for lateral duration in complex onsets (there was only a 2 ms difference across groups).

For CV syllables in Spanish, again vowel duration (69%) and VOT (18%) were the most relevant classifiers. For /Cl/ clusters, vowel duration (42%), followed by /l/ duration (37%) were the most important features in distinguishing the NH and CI groups' productions. However (in Table 1), for both the English and Spanish children's productions, the accuracy and precision at best f-scores are quite close to chance in some instances, meaning that on many accounts the two groups (NH and CI) are virtually indistinguishable within a specific language (though not so across languages). The following Figure 3 illustrates

precision and recall curves for the Spanish and English productions by group and Table 1 shows model ouputs.

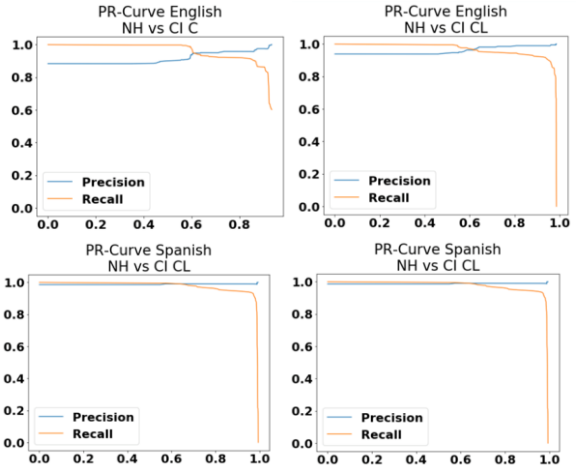

Figure 3. *Precision/recall plots showing the relationship between best precision and recall scores for the children's English (bottom left page) and Spanish (top right page) productions.*

Table 1: *Outputs of the five Random Forest models*

Experiment	Model evaluators for binary classifiers				Relative importance			
	precision at best f-score	recall at best f-score	best f-score	accuracy at best f-score	VOT	/l/ dur	V dur	ICI
English								
NH vs CI (CV)	0.6	0.87	0.71	0.65	0.26	0.15	0.59	0.0
NH vs CI (CIV)	0.68	1.0	0.81	0.74	0.30	0.47	0.23	0.0
Adult vs Youth (CV)	0.79	0.7	0.74	0.88	0.15	0.27	0.58	0.0
Adult vs Youth (CIV)	0.95	0.88	0.91	0.98	0.13	0.44	0.44	0.0
Spanish								
NH vs CI (CV)	0.57	0.98	0.72	0.61	0.18	0.13	0.69	0.0
NH vs CI (CIV)	0.8	0.88	0.84	0.81	0.21	0.37	0.42	0.0
Adult vs Youth (CV)	1.0	0.95	0.97	0.98	0.04	0.25	0.71	0.0
Adult vs Youth (CIV)	0.92	0.65	0.76	0.97	0.10	0.08	0.78	0.06
English/Spanish								
EN vs SP (CV)	0.64	0.98	0.77	0.67	0.61	0.34	0.06	0.0
EN vs SP (CIV)	0.97	0.91	0.94	0.95	0.03	0.41	0.19	0.36

Comparing the adult and children's groups, however, best f-score, accuracy and precision are all significantly higher than in the children-only models. For the English group, the important feature distinguishing the adult vs children's speech was vowel duration (58% for C onsets and 44% for C1C2 onsets), followed by /l/ duration (27% for C onsets and 44% for C1C2 onsets). For the Spanish adult-children's productions, the model produces similar results. Vowel duration following singleton and complex onsets (where C2 is a lateral) was the most significant feature. Notice, however, that the interconsonantal interval in the Spanish productions only showed a relative importance of 6%, meaning that the children's patterns and the adult patterns were fairly commensurate for this feature.

Our fifth model uses a two-category solution based on language (English and Spanish) in oder to test whether the bilingual children produce two qualitatively different timing configurations, as opposed to applying one master configuration to two different languages. After training a model on both English and Spanish data we introduced the children's English and Spanish productions but did not make a specification for group (since we were not testing for effects of hearing loss here). The results of the models suggest that both groups produce qualitatively different patterns of syllable timing based on language. Overall, for pooled data across both groups, the model was able to classify, with relatively high

reliability, the productions for each language (though results were much better for complex onsets given the inclusion of ICI). For the singleton onset tokens, the best f-score was 77%, with recall at best f-score being 98% (though precision and recall were significantly lower than in /Cl/). In these cases, VOT and /l/ duration had the highest relative importance in classification (VOT = 61%; /l/ duration = 34%), while vowel duration was substantially less important (vowel duration = 6%). In /Cl/ clusters, best f-score, precision, recall and accuracy were very high, between 91%-100% for all measures. The relevant variable importance in complex clusters shows the highest values for /l/ duration (41%), followed closely by ICI (36%), then vowel duration (19%). VOT had the lowest relative importance. The following Figure 4 illustrates precision and recall curves for the Spanish and English productions by group (though we only include by-group graphs for illustrative purposes. The model did not distinguish by groups):

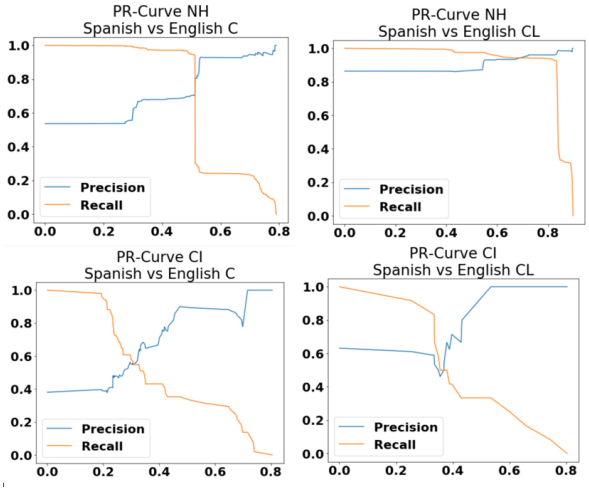

Figure 4. *Precision/recall plots showing the relationship between best precision and recall scores for the children's mixed English and Spanish productions of complex onsets and classification threshold.*

In sum, the models show differences between 1) the adult and children's productions as well as 2) differences between groups (NH and CI) for some features. Additionally, the models were capable of classifying 3) the children's productions by language. Based on the results of the models, it seems that both groups of children are applying two qualitatively different timing configurations based on language. We will revisit these results more fully in section 4.

4. Discussion

The results of our Random Forest models comparing the English and Spanish productions present a couple of interesting findings. For the individual languages, the models show a clear distinction between the adult and child speech in both English and Spanish, though the results distinguishing the children's groups are less robust. Vowel duration has the highest feature importance in distinguishing both adults from children and the NH group from the CI group. This is expected given the differences in raw measures for vowel duration between the two groups published in [14] and [15]. For English, VOT is also an important feature for distinguishing the adults from children and the individual groups' productions, yet for Spanish the importance is diminished. This result is also coherent with the idea that the voicing distinction based on VOT in English presents a more complex contrast system to produce than Spanish' simple true voice/short-lag VOT paradigm. Essentially, VOT is a biomechanical effect of stop articulation that follows from physical laws (air in areas of high pressure moves to low pressure areas by way of the Bernoulli principle). Some languages like English and German take advantage of this biomechanical effect to realize contrasts between stops. Spanish does not manipulate this physical effect to signal a contrast in voice. English speakers, on the other hand, must learn to modulate VOT to signal the correct phonological voice category. Hence, in this context it is entirely expected that VOT would be an important feature to distinguish groups (Adults vs. children and NH vs. CI) since the voicing distinction based on VOT is a phonologically encoded characteristic of the language.

Lateral duration (both as singleton word initial /l/ and /l/ in clusters) is an important feature in distinguishing adult from children's speech in English, but not so in Spanish. We suspect that this is due to the fact that in English, the children must master two laterals, while only one (clear) lateral exists in Spanish.

5. Conclusions

Our results suggest that the children produce distinct timing configurations for each language as it has been found by [2] regarding singleton stop VOT. Comparing our results of the English production task with the previously reported results of their Spanish productions, we find that the production of VOT, the short-lag/long-lag paradigm for voiced stops and lateral reduction all point to a clear distinction between the two languages (as did the interconsonantal interval introduced in our Random Forest models). English and Spanish both permit complex syllable onsets (structure), but differ in the phonetic realization of certain parameters such as VOT and the voicing implementation of voiced stops. Our Random Forest analyses were able to categorize with substantial accuracy the children's productions in each language, suggesting that the children do indeed produce separate timing configurations when producing each language. Whether or not they are accessing two gradient dimensions of the same timing plan as per [1], or accessing two different timing plans altogether remains an open empirical question we intend to examine in moving forward. Specifically, in order to better understand the interaction between timing plans for the different languages we intend to look at productions involving code-switching, where it has been shown that language interference, even when absent in language-isolated production tasks, may emerge when the two languages are coerced into contact in a laboratory setting ([1] and [30]). Equally, manipulating the elicitation task may also prove fruitful in perturbing the specific timing plans of the two languages.

6. Acknowledgements

This work was generously supported by a grant [ref. PID2019-105929GA-I00] from the Spanish Ministry of Science and Innovation (Ministerio de Ciencia e Innovación).

7. References

[1] M. Antoniou, C. Best, M. Tyler and C. Kroos, "Inter-language interference in VOT production by L2-dominant bilinguals: Asymmetries in phonetic code-switching," *Journal of Phonetics*, vol. 39, no. 4, pp. 558–570, 2010.

[2] F. Bunta, C.E. Goodin-Mayeda, A. Procter and A. Hernandez, "Initial stop voicing in bilingual children with cochlear implants and their peers with normal hearing," *Journal of Speech, Language, and Hearing Research*, vol. 59, pp. 686–698, 2016.

[3] A. Caramazza, G,H. Yeni-Komshian, E.B. Zurif and E. Carbone, " The acquisition of a new phonological contrast: The case of stop consonants in French–English bilinguals," *Journal of the Acoustical Society of America*, vol 54, pp. 421–428

[4] J.E. Flege, "Interactions between the native and second-language phonetic systems,". in P. Burmeister, T. Piske, and A. Rohde (Eds.), *An integrated view of language development: Papers in honor of Henning Wode*, pp. 217–243, 2002. Trier, Germany: Wissenschaftlicher Verlag.

[5] J.E. Flege, I.R.A MacKay and T Piske, "Assessing bilingual dominance," *Applied Psycholinguistics*, vol 23, pp. 567–598, 2002.

[6] J.E. Flege and W. Eefting, "Cross-language switching in stop consonant perception and production by Dutch speakers of English," *Speech Communication*, vol 6, pp. 185–202, 1987b.

[7] J. Paradis, "Do bilingual two-year-olds have separate phonological systems?," *International Journal of Bilingualism*, vol. 5, no. 1, pp. 19–38, 2001.

[8] R.A. Krakow, "The Articulatory Organization of Syllables: A Kinematic Analysis of Labial and Velar Gestures¡," PhD dissertation, Yale University, New Haven, CT, 1989.

[9] V. Fromkin, "Some Phonetic Specifications of Linguistic Units: an Electromyographic Investigation," *Working Papers in Phonetics 3*, UCLA, 1965.

[10] I. Lehiste, "Segmental and Syllabic Quantity in Estonian," *American Studies in Uralic Linguistics*, vol. 1. Bloomington: Indiana University.

[11] M. Macchi, "Labial articulation patterns associated with segmental features and syllable structure in English," *Phonetica*, vol. 45, pp. 109–121.12, 1988.

[12] C. Browman and L. Goldstein, "Articulatory phonology: an overview," *Phonetica*, vol. 49, no. 3-4, pp. 155–80, 1992.

[13] D. Byrd, "Influences on articulatory timing in consonant sequences," *Journal of Phonetics*, vol. 24, no. 2, pp. 209–244, 1996.

[14] M. Gibson, F. Bunta, E. Goodin-Mayeda and A. Hernández, "The acquisition of syllablelevel timing contrasts by English- and Spanish-speaking bilingual children with normal hearing and English- and Spanish-speaking bilingual children with cochlear implants," *Journal of Phonetics*, vol. 71, pp. 98–112, 2018.

[15] M. Gibson, F. Bunta, C. Johnson, and M. Huárriz, "Early productions of syllable-level timing by bilingual English- and Spanish-speaking children with cochlear implants and their peers with normal hearing," *Journal of Phonetics*, vol. 95, pp. 1–16, 2022.

[16] S. A. 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

[17] T. H. Crystal and A. S. House, "Segmental durations in connected ‐ speech signals: Current results," *The Journal of the Acoustical Society of America*, vol. 83, pp. 1553, 1988.

[18] G.J. Docherty, *The timing of voicing in British English obstruents*, No. 9. Walter de Gruyter, 1992.

[19] M. Gibson, A.M. Fernández Planas, A. Gafos and Ramirez, E. "Consonant duration and VOT as a function of syllable complexity and voicing in a sub-set of Spanish clusters," in Proc. INTERSPEECH 2016– *15 Annual Conference of the International Speech Communication Association*, Dresden, German, Sept. 2016, pp. 1690–1694.

[20] M. Gibson, S. Sotiropoulou, S. Tobin and A. Gafos, "Temporal aspects of word initial single consonants and consonants in clusters in Spanish," *Phonetica*, vol. 76, pp. 448–478, 2019.

[21] L. Lisker and A.S. Abramson 'A Cross-Language Study of Voicing in Initial Stops: Acoustical Measurements', *Word*, vol. 20, no.3, pp 384–422, 1964.

[22] L. Williams, "he perception of stop consonant voicing by Spanish–English bilinguals," *Perception & Psychophysics*, 21, pp. 289–297, 1977.

[23] D.H. Klatt, "Voice onset time, frication, and aspiration in word-initial consonant clusters," *Journal of Speech and Hearing Research*, vol. 18, no. 4, pp. 686–706, 1975.

[24] P. Menyuk and M. Klatt, "Voice onset time in consonant cluster production by children and adults," *Journal of Child Language*, vol 2, no. 2, pp. 223-231, 1975.

[25] S. Marin and M. Pouplier, "Temporal organization of complex onsets and codas in American English: Testing the predictions of a gestural coupling model," *Motor Control*, vol. 14, no. 3, pp. 380-407, 2010.

[26] S. Sotiropoulou, M. Gibson and A. Gafos, "Temporal stability patterns of stop-liquid and stop-rhotic clusters in Spanish," *Journal of Phonetics*, vol. 82, pp. 1–22, 2020.

[27] D. O'Shaughnessy, "Consonant durations in clusters," *Proceedings of the IEEE, Transactions on Acoustics, Speech, and Signal Processing*, vol. 22, no. 4, pp. 282–295, 1974.

[28] P. Hoole, L. Bombien, B. Kühnert and C. Mooshammer, "Intrinsic and prosodic effects on articulatory coordination in initial consonant clusters", Frontiers in Phonetics and Speech Science, pp 275–286, 2009

[29] F. Pedregosa, G. Varoquaux and A. Gramfort, *Journal of Machine Learning Research*, vol. 12, pp. 2825. http://www.jmlr.org/papers/v12/pedregosa11a.html

[30] J. Magloire and K.P. Green, "A cross-language comparison of speaking rate effects on the production of voice onset time in English and Spanish," *Phonetica*, vol. 56, 158–185, 1999.