INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Are IP initial vowels acoustically more distinct? Results from LDA and CNN classifications

Fanny Guitard-Ivent[1], *Gabriele Chignoli*[1], *Cécile Fougeron*[1], *Laurianne Georgeton*[2]

[1]Laboratoire de Phonétique et Phonologie (UMR7018, CNRS – Sorbonne Nouvelle)
[2]French Forensic Police Office (SCPTS)

{fanny.ivent, gabriele.chignoli, cecile.fougeron}@sorbonne-nouvelle.fr,
laurianne.georgeton@interieur.gouv.fr

Abstract

Past results have suggested that initial strengthening (IS) effects target the contrastive phonetic properties of segments, with a maximization of acoustic contrasts in initial position of strong prosodic domains. Here, we investigate whether IS effects translate into a better acoustic discriminability within the French oral vowels system. Discriminability is assessed on the basis of classification results of two types of classifiers: a linear discriminant analysis (LDA) based on the four formants frequencies, and a deep convolutional neural network (CNN) based on spectrograms. The test set includes 720 exemplars of /i, y, e, ɛ, a, x, u, o, ɔ/ (with /x/=/ø, œ/) produced in a labial context, either in intonational phrase initial (IPi) or word initial (Wi) position. Classifiers were trained using a set of 4500 vowels extracted from a large read speech corpus. Results show a better discriminability of vowels (overall better classification rate) in IPi than in Wi with the two methods. Less confusion in IPi is found between rounded and unrounded, and between back and front vowels, but not between the vowels along the four-way height contrast. Less confusion between peripheral and central vowels also expresses a maximization of contrasts within the acoustic space in IPi position.

Index Terms: Initial Strengthening, vowel discriminability, French, acoustic, automatic classifications, LDA, CNN

1. Introduction

Initial strengthening (IS) refers to phonetic variation undergone by segments in initial position of prosodic domain. Domain-initial segments have been described as showing a spatial and/or temporal expansion, which is proportional to the strength of the prosodic boundary - i.e. the stronger, the boundary, the stronger, IS effects (see [3] for a review). The effects of IS on the phonetic properties of consonants are very well documented. The results of studies conducted in different languages (such as French, English, Korean, Taiwanese and Japanese, for instance) and involving different types of articulatory structures (lingual, nasal or glottal) agree that IS increases the consonantal properties of domain initial oral or nasal consonants [5, 6, 7, 8, 14, 20]. There are fewer studies on IS effects on vowel properties but the existing studies indicate that the phonetic properties of domain-initial vowels are also modified [6, 7, 9, 10, 11, 15, 18].

Results of different studies have suggested that IS effects target the contrastive phonetic properties of segments, with a maximization of phonetic contrasts in initial position of strong prosodic domains. For instance, in IPi position, consonants have been found to be more consonantal, e.g. more distinct in their nasal, oral or glottal features (see [3, 4] for a review). In their studies on the effect of IS on the set of oral French vowels, Georgeton and colleagues [11, 12] also showed that IS-induced variations affect the way in which vowel contrasts are realized phonetically. Even though both rounded and unrounded vowels are produced with a larger lip opening, the increase is larger for the unrounded vowels. As a consequence, the lip opening distinction between unrounded and rounded vowels is maximized in strong prosodic position (IPi). Variation in tongue position and acoustic correlates of the four-way height contrast among front vowels was also found to result in a maximization of the distinctions between /e/ and /ɛ/, and between /ɛ/ and /a/, but not between /i/ and /e/ [12]. Finally, the acoustic contrast between back and front vowels was also maximized in IPi position: in IPi, front vowels had higher F2 values and back vowels had lower F2 and F2-F1 values. Finally, Georgeton [10] reports an increase of the F1/F2 acoustic space in IPi position, with corner vowels becoming more peripheral.

In the present study, we investigate whether the previously reported phonetic variation in IPi indeed translate into a better acoustic discriminability among vowels. We are particularly interested to test if IS effects on phonetic properties of vowels is beneficial for the discrimination between vowel categories within a crowded set of contrasts, such as the oral vowels system in French. To this end, rather than testing discriminability with human listeners in a time-intensive perceptual experiment, we use the results of automatic classification as an index of the potential effects of IS on perception. We test whether classification rates of vowels are better in initial position of intonational phrase (IPi) rather than in initial word position (Wi), and whether confusions are comparable in the two positions.

Two classification experiments are conducted: a linear discriminant analysis (LDA) based on the four formants frequencies; and a classification involving a deep convolutional neural network (CNN) based on spectrogram pictures. While LDA is often used in the field of phonetics, the use of CNN is more recent (see for example [1, 13]). It will be interesting to see if CNN offers satisfactory and comparable results to those obtained with the LDA. Indeed, while CNN requires a large set of training data, the spectrogram-based classification has an interesting potential in phonetic research: classification is based on a wider range of information than a pre-selected set of acoustic features (e.g. formants, MFCC…), and it has the advantage of not relying on time- and expertise-consuming measurements.

The analyses proposed here are done on the data collected in [10, 11, 12]. We compare classification results of the French

http://dx.doi.org/10.21437/Interspeech.2019-2153

oral vowels /i, y, e, ɛ, a, x, u, o, ɔ/ (with /x/=/ø, œ/) produced in either IPi and Wi positions. According to the results described above, several predictions can be made. First, we expect a better classification of IPi vowels. Second, confusion patterns are expected to reflect the maximization of acoustic contrasts along the distinctive dimensions of the system: height, backness, rounding and peripherality. Less confusion are predicted between rounded and unrounded vowels but also between front, central and back vowels and between mid-close and mid-open vowels or mid-open and open vowels. Conversely more confusions between /i/ and /e/ are expected in IPi according the greater spectral proximity between them found in IPi [12]. Finally, we expect less confusion between peripheral vowels with the central vowel /x/ (with /x/=/ø, œ/).

2. Method

2.1. Speech material

To test the effect of prosodic position on vowel classification, we explored a corpus built by Georgeton [10] to investigate initial strengthening effect on French oral vowels. The 10 oral French vowels (V) /i, e, ɛ, a, y, ø, œ, u, o, ɔ/ were produced in [ip#VC] sequences where V is the initial segment of a fake first name and "#" represent a prosodic boundary, either Intonational Phrase boundary (IP) or Word boundary (W). The two flanking consonants were always /p/ except for the three mi-open vowels /ɛ/, /œ/, /ɔ/ that have a particular distribution in French. For these vowels, following consonants was a /v/ for /ɛ/, a /f/ for /œ/ and a /ʁ/ for /ɔ/ (see [11] for a detailed description of the corpus). Four female speakers of standard French were recorded reading 16 repetitions of 20 sentences containing these sequences (2 prosodic boundaries×10 vowels types) in random order.

From this corpus, we selected 80 exemplars of each vowel category (40 in IPi and 40 in Wi positions) in order to obtain 10 exemplars per vowel categories produced by each speaker in each prosodic position. We used duration criterion to do this selection favoring tokens near vowel category mean duration. Note that in this study, we defined a single open-mid central vowel category labelled /x/ containing 50% of the /ø/ and 50% of the /œ/ of the original corpus. This was done in order to match the categories available in the training set and match what is usually done in automatic transcription systems, since this contrast is unstable in French. At the end of this procedure, we have a test set of 720 observations (40 exemplars x 2 prosodic boundaries x 9 vowel categories).

2.2. Classification

2.2.1. Training dataset

To build the models to be used in the two classification methods, we decided to resort to external corpora for two reasons. First, our dataset was small and we did not want to reduce it anymore with a split into training and test sets. Second, we wanted the model to be built on realizations of vowels in various prosodic positions. Therefore, we used a large corpus of read speech, the BREF corpus [16] composed of read part of the newspaper *Le Monde* by non-professional speakers.

To get as close as possible to test dataset, we extracted vowels with durations ranging from 45 to 250 ms, also produced in a labial context. Right and left bilabial contexts were preferred but to increase the number of exemplars we also include vowels preceded or followed by a labial consonant among /p, b, m, f, v/. 500 exemplars of each vowel category produced by 43 different female speakers constituted our training set (500 x 9 vowels = 4500 observations).

2.2.2. Linear Discriminant Analysis

The LDA was trained and tested using MASS package [22] with R software [21]. The nine vowel categories /i, y, e, ɛ, a, x, u, o, ɔ/ (with /x/ = /ø, œ, ə/) were the a priori categories and the four first formant values measured at 50% of the vowel duration were our discriminant variables. Formant values were extracted using the Burg algorithm of Praat software [2] on a window length of 25 ms. The detection of amplitude peaks was determined in a band lower than 5.5 kHz.

2.2.3. Convolutional Neural Network

For the deep learning classification a CNN was trained with the final model constructed as a slightly adapted LeNet [17] architecture and Adam optimizer with batch size of 32 in 120 epochs. As input to the CNN model, vowel spectograms were extracted from the test and training datasets. These pictures were obtained using the Praat default settings (5-ms analysis frames and 2-ms hop size) in order to capture a broad-band representation with a 16 kHz sampling rate. We established a duration range for all speech segments between 45 and 250 ms, the obtained spectrograms were padded resulting in all samples having the same width and height. A resize was performed in addition to 8-bits and grayscale conversion so the GPU memory would handle mini-batches of sufficient size for learning to take place. The images size chosen was 1:3 of the original ones or 500 pixels for the time domain corresponding to 2,5 ms and 300 pixels on frequency dimension equal to 47,82 Hz.

3. Results

The effects of prosodic position on vowel discrimination are tested a posteriori with an analysis of the classification scores for the vowel exemplars produced in IPi vs. Wi positions.. We will first report overall, and by vowel, correct classification rates. Then we propose an analysis on the misclassification rates for groups of vowels along different contrasting dimensions in the French vowel system.

3.1. Overall and by vowel results

Overall, better classification rate was found for vowels produced in IPi than in Wi position, for both methods as shown in Table 1. Both the LDA and the CNN classification were found to be more sensitive with IPi vowels, will a higher global accuracy. Indeed 68% of the vowel tokens produced in IPi position were accurately classified against 63% of the token produced in Wi position for the LDA, and 69% against 60% in the CNN.

A better recognition was found in IPi for the vowels /i, y, a/ with both methods. Vowel /i/ was nonetheless well recognized in both prosodic positions (98% in IPi for the two classifiers, and 90% and 85% in Wi for the LDA and CNN classifiers respectively). For /y/, a larger difference in recognition rate was found according to prosodic position especially with the CNN technique (+23%). For /a/, the effect of prosodic position was the most drastic. It was shown by the two classification methods, with nonetheless a larger difference in recognition rate shown by the LDA technique

(+58% vs. +33% with CNN) IPi = 98% vs. Wi=40%, vs. CNN: IPi = 98% vs. Wi=65%). A graphical representation of the confusion matrices is shown in Figure 1 for the LDA method and Figure 2 for the CNN method. As can be seen in these figures, a better recognition of /a/ in IPi resulted from less confusions with /x/ in the LDA method, and with /x/, /ɔ/ and /u/ in the CNN method.

Table 1: *Correct classification rates obtained for each vowel by prosodic position (IPi and Wi) with LDA (left) and CNN (right). % on 40 test tokens*

	LDA		CNN	
	IPi	Wi	IPi	Wi
Overall	68	63	69	60
/i/	98	90	98	85
/y/	90	83	88	65
/e/	55	65	88	75
/ɛ/	65	68	45	48
/a/	98	40	98	65
/x/	55	55	78	60
/u/	15	20	100	100
/o/	80	73	0	0
/ɔ/	58	75	25	45

A slightly better recognition in IPi position was also observed for vowels /x, e/ with the CNN classifier, and for /o/ with the LDA method. For /e/, the reduction of confusions with /y/ in IPi explained the better results observed in this position with the CNN model. For /o/, it is explained by less confusions with /u/ in IPi position with the LDA method.

For some vowels however, results showed a better classification rate in Wi position. This concerned the back mid-open /ɔ/ which was more often misclassified as /a/ in IPi than Wi position with the LDA method. This is consistent with the fact that [ɔ]s in IPi position have a higher F1, closer to that of /a/, as found by [10]. With the CNN method, [ɔ]s were also less misclassified as /a/ in Wi position but also as /o/. A better classification of [e]s uttered in Wi was also found in the LDA classification, due to less confusion with /i/ in this position. This also echoes the acoustic results of Georgeton [10] showing a greater acoustic proximity between /i/ and /e/ in IPi.

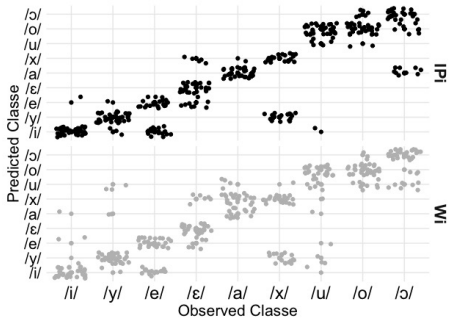

Figure 1: *Results of LDA classification presented by vowel category (x axis presents observed classes and y axis the predicted classes) as function of prosodic position (IPi in black on the upper panel vs. Wi in gray on the lower panel).*

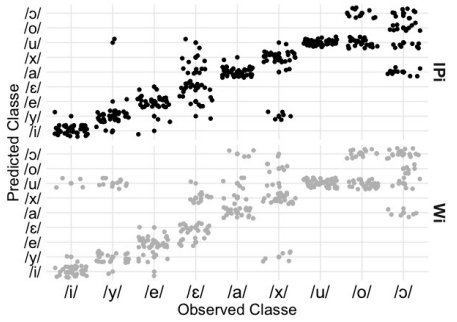

Figure 2: *Results of CNN classification presented by vowel category (x-axis presents observed classes and y-axis the predicted classes) as function of prosodic position (IPi in black on the upper panel vs. Wi in gray on the lower panel).*

In the other cases, vowels were equally recognized in the two prosodic positions. Similar classification scores were found for /ɛ/ in the two prosodic positions with the two methods. Nonetheless, confusions looked different in the two positions in the CNN method, with confusions being more distributed over the various vowel categories in IPi vs. Wi (where /ɛ/ is mostly confused as /e/ and /x/). No effect of prosodic position on vowel classification was also observed for /o/ with the CNN classifier and for /x, u/ with the LDA classifier. For /o/, the CNN classification was disappointing since all exemplars were misclassified (as /u/ or /ɔ/ in the two positions). Overall the acoustic proximity of the back vowels leaded to a lot of confusion within the back vowel groups. For the vowel /u/, classification rates with the LDA were poor in both positions due to frequent confusions /o/ as illustrated Figure 1. For /x,/ results obtained with LDA indicated relatively high confusions with /y/ regardless of the prosodic position as shown Figure 1.

3.2. Analysis by contrast type

To identify on which dimensions vowels are more distinct in IPi position, we analyzed the confusions along five contrasting dimensions at play in the French system of oral vowels:

- ROUNDING contrast between rounded /y, x, u, o, ɔ/ and unrounded vowels /i, e, ɛ, a/
- PLACE OF ARTICULATION (POA) contrasts between front /i, y, e, ɛ, a/, central /x/ and back /u, o, ɔ/.
- HEIGHT contrasts between closed /i, y, u/, mid-closed /e, o/, mid-open /ɛ, ɔ/ and open /a/. For this analysis, the /x/ category was discarded because it groups mid-close /ø/ and mid-open / œ/.
- TO CENTER: here, we analyze confusions of peripheral vowels /i, y, e, ɛ, a, u, o, ɔ/ with the center of vocalic space /x/.
- FROM CENTER: conversely, here we focus on confusions of the central vowels /x/ with peripheral ones /i, y, e, ɛ, a, u, o, ɔ/.

As illustrated in Figure 3, for the two methods, less confusion between rounded and unrounded (ROUNDING PANEL), and between front, central and back vowels (POA PANEL) were observed in IPi position compared to Wi. An increased acoustic and/or articulatory distinction between vowels along these dimensions was also found in [10, 11, 12] (see introduction). Our classification results show that these

changes in phonetic properties indeed yielded better discrimination of vowel tokens when uttered in IPi position.

Misclassifications of the peripheral vowels /i, y, e, ɛ, a, u, o, ɔ/ with the central vowel /x/ were also reduced in IPi position in both methods (TO CENTER panel). For the CNN analysis only, central /x/ vowels were also less often confused with peripheral vowels in IPi position (FROM CENTER panel). These results are consistent with the findings in [10] showing a larger F1/F2 acoustic space in IPi position with less centralization. However, this trend is not captured by the LDA classifier for which almost half of the [x]s were misclassified (mostly as /y/, see above) , independently of their position.

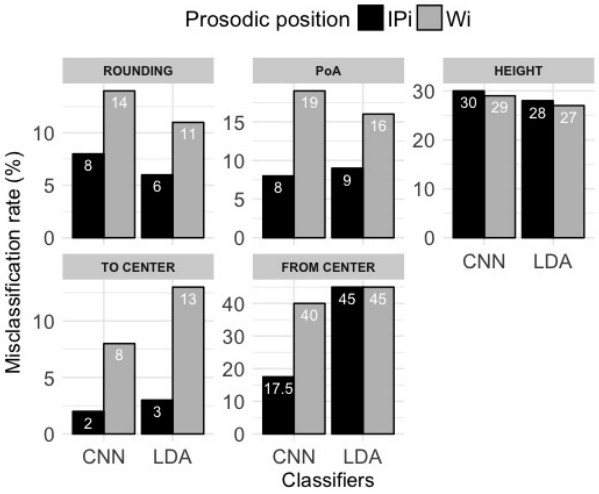

Figure 3: *Misclassification rate (%) for each type of contrast (rounding, PoA, height, to center and from center) obtained with each classifier (CNN and LDA) as function of prosodic position (IPi in black and Wi in gray).*

Concerning the contrast in vowel height (HEIGHT panel), no effect of prosodic position was found: misclassification rate between open, mid-open, mid-closed and closed vowel was quite high (~29%) in both prosodic positions and for both classifiers. According to the observations of tongue contours in [12], we expected less confusion between /e/ and /ɛ/, and /ɛ/ and /a/ in IPi position, but this was not the case. For /i/-/e/, however, more confusions were predicted by the same authors due to the closing of /e/ in IPi position, and this was confirmed in our data. Moreover, more confusions of /ɔ/ with /a/ were found in IPi.

4. Discussion & Conclusion

Whether prosodically induced phonetic variation has implication on speech perception is a long-standing question. Kim & Cho [15] and Mitterer, Cho and Kim [19] showed that the perception of a contrast between consonants (marked by VOT cues) is modulated with respect to the fact that its phonetic implementation is affected by prosodic position. In this study, we tackled this question using automatic classification results as a proxy of human perception to test whether vowels uttered in initial position of intonational phrase (IPi) are better discriminated than vowels uttered in word initial position (Wi). Results indeed show that IP initial vowels are acoustically more distinct: they are overall better classified as tokens of their vowel category and confusions along most of the dimensions of contrast in the systems are reduced. However, it was also shown that this trend does not affect all vowels equally (the mid vowels /ɛ, ɔ/ did not show a better classification in IPi, for instance) and that discrimination of vowels along the four-way height contrast dimension was not affected by prosodic position. These results echo those of previous acoustic or articulatory descriptions showing that prosodically-induced variation are segment-dependent and not systematic [e.g. 7]. Nonetheless, the overall better discrimination found for IP initial vowels is in line with the idea that initial strengthening effects on the phonetic make-up of vowel contributes to a clearer realization of phonetic contrasts that the one produced in IP medial (W initial) position.

If prosodically induced phonetic variations are strong enough to be captured by automatic classifiers fed with either formant values or a more global acoustic picture (a spectrogram), implications for human perception are not direct. It has been previously shown [19] that native English listeners and Korean learners of English use temporal cues induced by prosodic boundary to categorize voiced and voiceless stops. Listeners accepted stops with long VOT as voiced when they occurred after a strong prosodic boundary. More important, the authors revealed that temporal cues of prosodic boundary were treated differently than speaking-rate modulations. Further studies are indeed needed to understand how and whether the properties of vowels realized in IP initial position may contribute to the decoding of the segmental or prosodic content of speech.

A second objective in this study was the comparison of two classifiers: a classical formant based LDA and a relatively new classification technic based on a CNN. The two classifiers showed similar overall results in terms of performances and the CNN method better captured the differences between prosodic positions when results were split by vowels. Although it is quite difficult to understand which information is extracted for the neural networks classification, there is a larger spectrum of information available on a spectrogram to capture discriminant information, than a simple set of 4 formants. These results are promising for further studies relying on this technique, particularly when the extraction of acoustic features such as formants is problematic (for instance when dealing with pathological speech).

The major problem with neural network technique concerns the amount of material needed, and not always available, to construct classifier models. In our study, we had the opportunity to use an external and large corpus as a training set, but we are indeed conscious that it is not always the case.

5. Acknowledgements

This work was supported by the program "Investissements d'Avenir" ANR-10-LABX-0083 (Labex EFL) and by the ANR-17-CE39-0016 (VoxCrim).

6. References

[1] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn and D. Yu, "Convolutional Neural Networks for Speech Recognition," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, 2014.

[2] P. Boersma, and D. Weenink, "Praat: doing phonetics by computer" (Version 5.4.04) [Computer program]. Retrieved from: http://www.praat.org/, 2014.

[3] T. Cho, “The continuum companion to phonology,” Laboratory Phonology, pp. 343-368, 2011.
[4] T. Cho, “Prosodic boundary strengthening in the phonetics–prosody interface,” Language and Linguistics Compass, vol. 10, no. 3, pp. 120-141, 2016.
[5] T. Cho, and P. Keating, “Effects of initial position versus prominence in English,” *Journal of Phonetics*, vol. 37, no. 4, pp. 466-485, 2009.
[6] C. Fougeron, “Variations articulatoires en début de constituants prosodiques de différents niveaux en français,” PhD thesis, University of Sorbonne Nouvelle - Paris 3, Paris, 1998.
[7] C. Fougeron, “Articulatory properties of initial segments in several prosodic constituents in French,” *Journal of Phonetics*, vol. 29, no. 2, pp. 109-135, 2001.
[8] C. Fougeron, and P. A. Keating, “Articulatory strengthening at edges of prosodic domains,” *The journal of the acoustical society of America*, vol. 101, no. 6, pp. 3728-3740, 1997.
[9] C. Gendrot, K. Gerdes, and M. Adda-Decker, “Impact of prosodic position on vocalic space in German and French,” in *ICPhS 2011 – 17th International Conference of Phonetics Science, Hong-Kong, China, Proceedings,* 2011, pp. 731-734.
[10] L. Georgeton, “Renforcement des voyelles orales du français en position initiale de constituants prosodiques : interaction avec les contrastes phonologiques,” PhD thesis, University of Sorbonne Nouvelle - Paris 3, Paris, 2014.
[11] L. Georgeton, and C. Fougeron, “Domain-initial strengthening on French vowels and phonological contrasts: evidence from lip articulation and spectral variation,*” Journal of Phonetics*, vol. 44, pp. 83-95, 2014.
[12] L. Georgeton, T. K. Antolík, and C. Fougeron, “Effect of domain initial strengthening on vowel height and backness contrasts in French: Acoustic and ultrasound data,*” Journal of Speech, Language, and Hearing Research*, vol. 59, no. 6, pp. 1575-1586, 2016.
[13] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, K. Wilson, “CNN Architectures for Large-Scale Audio Classification,*”* In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 131-135, 2017.
[14] P. Keating, T. Cho, C. Fougeron, and C. Hsu, “Domain-initial articulatory strengthening in four languages,” *Laboratory Phonology*, vol. 6, pp. 143–161, 2003.
[15] S. Kim, and T. Cho, “Prosodic strengthening in the articulation of English /æ/,” *Studies in Phonetics, Phonology and Morphology*, vol. 18, no. 2, pp. 321-337, 2012.
[16] L. Lamel, J. Gauvain, and M. Eskenazi, “BREF, a Large Vocabulary Spoken Corpus for French,” in *EUROSPEECH 1991 – The 2nd European Conference on Speech Communication and Technology, September 24-26,* Genova, Italy, *Proceedings,* 1991.
[17] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, 1998.
[18] H. Lehnert-LeHouillier, J. McDonough and S. McAleavey, “Prosodic strengthening in American English domain-initial vowels,” *Proceedings of Speech Prosody 2010-Fifth International Conference,* 2010.
[19] H. Mitterer, T. Cho, and S. Kim, S, “How does prosody influence speech categorization?,” *Journal of Phonetics*, vol. 54, pp. 68-79, 2016.
[20] A. Onaka, “Domain-initial strengthening in Japanese: An acoustic and articulatory study,” In *ICPhS 2003 - The 15th international congress of phonetic sciences*, *Barcelona, Spain, Proceedings,* 2003, pp. 2091-2094.
[21] R. C. Team, “R: A language and environment for statistical computing,” 2013.
[22] W. N. Venables, and B. D. Ripley, *Modern Applied Statistics with S*. Springer Science & Business Media, New York, 2002.