INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

A Study on the Importance of Formant Transitions for Stop-Consonant Classification in VCV Sequence

Siddarth Chandrasekar[*1], *Arvind Ramesh*[*2], *Tilak Purohit*[3,4], *Prasanta Kumar Ghosh*[5]

[1]Indian Institute of Information Technology, Design and Manufacturing, Kancheepuram, India
[2]National Institute of Technology Tiruchirappalli, India
[3] Idiap Research Institute, Martigny, Switzerland
[4] École polytechnique fédérale de Lausanne (EPFL), Switzerland
[5] Indian Institute of Science (IISc), Bengaluru, India

siddarthc2000@gmail.com

Abstract

This study analyzes formant transitions in six English stop-consonants in vowel-consonant-vowel (VCV) sequences. We investigate whether natural speech preserves formant patterns, and if not, how it affects stop-consonant perception and automatic classification. We specifically ask three questions: 1) To what extent these formant transition patterns are preserved in naturally produced VCV sequences? 2) If not preserved, does it have any effect on the perception of the stop-consonant? 3) How does the classification of stop-consonants by automatic classifiers change when formant transition patterns are not preserved? We found that 33.56% of the corpus deviate from the formant transition pattern. The perception test reveals an Unweighted Average Recall (UAR) of 91.97% in identifying the stop-consonants in the VCV sequences when the pattern is not preserved compared to 93.54% when it is preserved. The best UAR from an automatic classifier is 68.35% and 77.5% in these two cases, respectively.

Index Terms: Formant transitions, Stop-consonants, Speech analysis

1. Introduction

Stop-consonants, also known as stops or plosives, are formed by obstructing the airstream in the vocal tract, resulting in occlusion in the oral cavity, followed by the release of the blocked air stream. These stops can be classified into two categories: voiced and unvoiced. While stop production requires transient release of air in the oral cavity, vocal cords vibrate during voiced stops (/b/, /d/, /g/) production unlike unvoiced stops (/p/, /t/, /k/). Despite their similar articulatory dynamics, voiced and unvoiced pairs (/b/ & /p/, /d/ & /t/, /g/ & /k/) are acoustically distinct. As a result, stop-consonants have been extensively studied to understand the relationship between motor movements and speech perception [1, 2, 3]. Moreover, they are found to provide crucial phonetic information about the words they are a part of [4, 5], thus playing an important role in word perception and recognition. In the following section, we discuss the previous research on stop-consonants in two parts, namely 1) perception studies in humans [6, 7, 8, 9, 10], and 2) recognition studies in machines [11, 12, 13, 14].

Perception of stop-consonants: Studies on the perception of stop-consonants began with exploring perceptual cues such as formants and their transitions [15, 6, 7, 16, 8]. Formant transitions, which are frequency shifts in vowel formants when they join with consonants, play a crucial role in perceiving vowel-consonant (VC) and consonant-vowel (CV) pairs. Since consonants are often short and weak, formant transitions are essential cues for perceiving them in a VC pair. It has been reported in the past that the relative dynamics of F2 with respect to F3 plays an important cue to place of articulation [6, 7, 8]. Previous literature has studied the relationship between speech articulation and formants. For instance, the height of the jaw and tongue affects F1 and F2, respectively [17], while the degree of mouth opening and tongue constriction affects F1 and F2 [18]. These investigations suggest that F1 and F2 play a significant role in the perception of stop-consonants, as they result from tongue and labial constriction [17, 18].

Since vocal tract configuration affects the formants [16], they are shown to follow a fixed trend across subjects for a given vowel-consonant combination [2, 6, 16, 7]. The widely used formant loci (transition trajectory) derived in [16] are based on spectrographic features of synthesized American-English CV utterances. These loci were hand-painted and then converted into sounds for the perception test. These formant loci detail the formant frequency and transition for a selected combination of stop-consonants and vowels. We resort to these formant loci as a standard pattern for the formant transition trajectory for this study. [2, 19] demonstrated that deviations in the vowel's formant frequency from the pattern induce insignificant changes in the auditory impression related to consonant perception.

Classification of stop-consonants: Being one of the most vital cues, formant transitions were extensively utilized in classifying stop-consonants using statistical and machine learning algorithms [11, 12, 13, 14, 20, 21]. The earliest consonant recognition was carried out via a statistical approach of modelling the formants [11]. Rather than just plain formant transitions, a combination of multiple acoustic properties was shown to aid in identifying the consonant [12, 20, 21]. Time-Delay Neural Networks (TDNNs) [13] was one of the initial works to capture the spatial relationships in speech processing. They utilize time-delay windows (translation invariant filter) to capture the temporal relationship in the spectral coefficients to identify voiced stop-consonants /b/, /d/ and /g/. The ability of the TDNNs to capture the temporal relationship in acoustic features such as formant transition enabled it to surpass then state-of-the-art models such as Hidden Markov Models (HMMs).

Relevance of analyzing formants today: Although formants are no longer being explicitly modeled for tasks like Automatic speech recognition (ASR) or Text-to-Speech (TTS) however, they still provide important cues for fine course speech evaluation and perception studies. Recent study [22] have shown formant transitions to be a cue for distinguishing between fricative. [23] showed formant transition information helps in the estimation of place and manner of articulation for fricatives /f/, /s/, and /sh/. [24] extended study of formant transitions to alveolar and the bilabial nasal phonemes for the Indian language Telugu. Ongoing studies investigating sensorimotor adaptation

[*] work conducted while at SPIRE Lab, IISc, Bengaluru

10.21437/Interspeech.2023-1738

of speech [3, 25, 26] utilizes formant information such as formant shifts, transitions, and perturbations to reveal adaptation within the speech motor system. All these recent works show the importance of analysing speech formants.

In this study, we aim to analyze formant transitions in the context of VCV sequences and the perception and classification of the VCV sequences. Though few recent studies [27, 28, 29] discuss the effects of a few factors, such as spectral shape affecting the perception of stop-consonants, none of them have analyzed the impact of formant transition that deviate from the pattern. Our motivation to conduct this study is to address the following three questions:
1) Does the naturally produced VCV sequences always follow a formant trajectory similar to the known pattern [16] or are there some inherent variations occurring in formats?
2) If not, does it affect (if any) the perception of that stop-consonant?
3) How does the classification of stop-consonants by automatic classifiers alter when formant transition patterns are not upheld?

To conduct this study, we make use of the SPIRE VCV an acoustic-articulatory corpus [30]. Examining 900 VCV samples and their spectrograms, 66.44 % of them were found to follow the pattern and the rest do not. The 33.56% not following the pattern answers the first question in this study, that there are variations in the formant structure for the stop-consonant production. With the first answer in place, we conduct perceptual experiments with 52 listeners to study the effect of the formant transitions in the perception of stop-consonants. Further to answer our third question, classification experiments were conducted to evaluate the performance of machine learning models. We found that humans achieved a UAR of 93.54% and 91.97% in identifying the samples that follow the pattern and those that do not, respectively. In comparison, the best classifier recorded accuracies of 77.5% and 68.35%.

2. SPIRE VCV corpus

The SPIRE VCV corpus[30] consists of concurrent acoustic and articulatory data for symmetric vowel-consonant-vowel (VCV) sequences. The corpus consists of samples recorded at three different speaking rates, that is slow, normal, and fast, from ten non-native English speakers (five female and five male) of age 18 to 27 years. All the subjects were fluent in English. It has all 85 possible combinations of seventeen consonants (C), namely, /b/, /ch/, /d/, /f/, /g/, /jh/, /k/, /l/, /m/, /ng/, /n/, /p/, /r/, /s/, /t/, /v/ & /z/ and five vowels (V), /a/, /e/, /i/, /o/ & /u/. Hence, there are a total of 450 VCV recordings (= 3 rates $\times$ 3 repetitions per rate $\times$ 5 vowels $\times$ 10 subjects) for every consonant and a total of 7650 VCV recording samples (= 450 samples per C $\times$ 17 C). Initially, the audio was recorded at 48 kHz and then was downsampled to 16 kHz. All the utterances are of the format "Speak VCV Today." And the VCV boundaries were manually annotated by observing the wideband spectrogram, the raw waveform, and the glottal pulses. For this study we resort to the consonants /p/, /t/, /k/, /b/, /d/, /g/ with slow speaking rate, which accounts for 900 samples (= 6 consonants $\times$ 3 repetitions$\times$ 5 vowels $\times$ 1 rate$\times$ 10 subjects). The slow speaking rate is used as they provide a relatively wider transition period which permits a better evaluation and annotation of the formants [31, 32]. The mean and standard deviations of the duration of the consonant regions are reported in Table 1. The spectrogram and formants were extracted using Librosa [1] and PRAAT [2] respectively.

Table 1: *Consonant region duration (in seconds) for the VCV combinations from the SPIRE VCV Dataset for slow speaking rate. A(B), A represents the mean duration, and B represents the standard deviation.*

	/b/	/d/	/g/	/p/	/t/	/k/
/a/	0.13(0.04)	0.10(0.04)	0.10(0.03)	0.22(0.06)	0.17(0.04)	0.18(0.08)
/e/	0.14(0.03)	0.14(0.07)	0.15(0.05)	0.22(0.06)	0.21(0.08)	0.22(0.08)
/i/	0.14(0.04)	0.16(0.09)	0.16(0.09)	0.24(0.09)	0.21(0.07)	0.23(0.09)
/o/	0.17(0.08)	0.17(0.09)	0.21(0.12)	0.23(0.07)	0.23(0.11)	0.26(0.15)
/u/	0.13(0.04)	0.13(0.06)	0.17(0.08)	0.23(0.09)	0.23(0.11)	0.23(0.09)

3. Study Protocol

3.1. Data Preprocessing

All 900 samples were manually inspected for their formant loci to evaluate the consistency of the human speech production system. The first two formants, F1 and F2, were extracted and overlaid on their spectrograms [3]. First, the formant points between the nuclei of the vowel and the consonants were manually corrected when the formants from PRAAT were erroneous, which is common in highly transient boundaries [33]. Furthermore, samples with noisy formants from PRAAT, which were irreparable due to noisy or unclear spectrogram, were discarded. This led to the removal of 85 samples, from which a sample is shown in Fig 1.

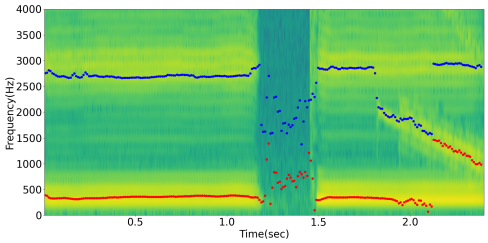

Figure 1: *The red and blue dots represents F1 and F2 respectively. Such samples were discarded due to their noisy formants.*

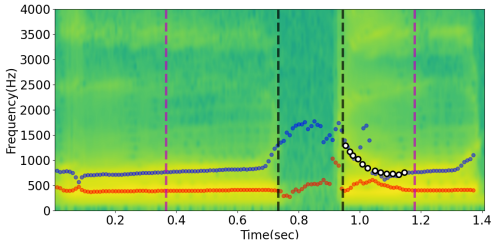

Figure 2: *The corrections were done between the vowel nuclei, represented by a magenta dashed vertical line, and the consonant boundary represented by a black dashed vertical line. While the blue dots represent the F2 formant trajectory from PRAAT, the white dots indicate the manually corrected trajectory.*

With their spectrogram as a reference, a total of 159 samples' formant trajectories were manually corrected for apparent errors. These errors include noisy jumps in the formant trajectories in the transient region, points of F2 being detected as

[1] https://doi.org/10.5281/zenodo.4792298
[2] http://www.praat.org/
[3] Supplementary Document: https://tinyurl.com/ybctp7tp

Table 2: *Consonant-wise distribution of the 900 samples. Disc denotes discarded samples, Corr denotes the corrected samples, and Orig refers to the samples which were not noisy/corrected.*

		/p/	/t/	/k/	/b/	/d/	/g/	Total
Disc		28	17	28	0	9	3	85
Mis-	Corr	5	1	6	0	0	0	12
Match	Orig	37	8	80	33	1	46	205
Match	Corr	18	21	4	33	36	35	147
	Orig	62	103	32	84	104	66	451

F1, etc. These corrections were performed by people with a background in acoustics and phonetics. A sample correction is shown in Fig 2. Following the removal of 85 samples and correction of 159 samples, the 815 samples (159 corrected and 656 original samples) were manually inspected to determine whether the transition of the first two formants followed the pattern or not. Formant transitions at the VC and CV boundaries were inspected and matched with the pattern. If the sample was found to have a similar trend (rise or fall) to the pattern, it was classified as matching. Even if any one of the formant transitions was found to be not obeying the pattern, it was considered to be not matching. Out of the 815 samples, a total of 598 samples (73.37%) were found to follow the pattern (match), and the remaining 217 samples (26.63%) were found to be not following it (mismatch). The illustration of samples not following the pattern is shown in Fig 3. The final consonant-wise distribution of the 900 samples is shown in Table 2.

All the formant corrections were performed using a TKinter-based custom tool. Both the corrections and pattern-based segregation were cross-verified by other annotators. Further experiments were carried out utilizing the 815 samples.

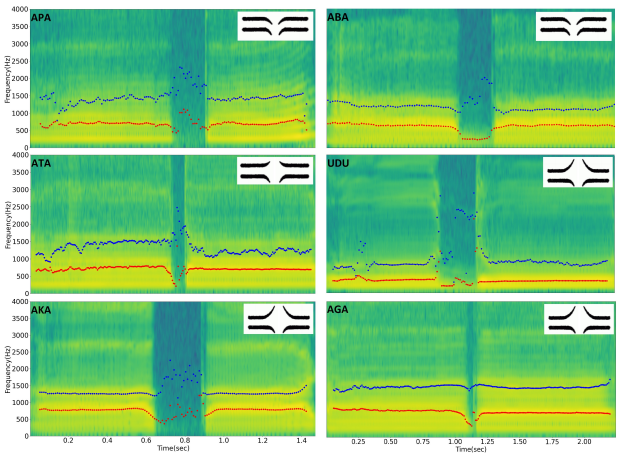

Figure 3: *A sample of each stop-consonant which does not follow the fixed pattern [16] is shown. The corresponding pattern is placed on the top-right of each sample.*

3.2. Human Perceptual Test

A total of 289 samples were selected to study the effects of the formant loci on the perception of consonants. These samples, 174 following the pattern and 115 not following, were carefully selected, covering samples of all subjects, consonants, and vowels. The distribution of the 289 samples across vowels and consonants is given in Table 3. For the perception test, a total of 52 participants, all native Indians with no hearing impairments, were employed. The perception test was carried out via customized google forms. The 289 samples were distributed among 10 forms(each consisting of 28-30 samples), and each form was perceived and evaluated by 5 participants, with 6 participants in two forms. In addition, six samples were duplicated in each form to ensure the consistency of participants so that each form would contain 34-36 samples. Hence, a total of 349 samples, including those repetitions, were used to carry out the perceptual test. A participant was considered consistent if he/she selected the same consonant for the repeated samples. Else he/she was considered inconsistent. The participants were informed of the audio format, 'Speak VCV Today' beforehand. Then, they were asked to listen to the audio samples as many times as desired and mark the perceived consonant. No particular device or headphone was recommended; listeners chose their own device as comfortable.

Table 3: *Vowel and Consonant wise distribution of samples used in perception test. A(B), where A represents samples not following the pattern, and B represents samples following the pattern.*

	/b/	/d/	/g/	/p/	/t/	/k/	Total
/a/	3(7)	0(10)	2(8)	4(6)	1(9)	10(0)	20(40)
/e/	0(10)	0(9)	3(7)	0(10)	4(4)	3(4)	10(44)
/i/	0(10)	0(9)	7(3)	2(6)	1(8)	8(1)	18(37)
/o/	7(3)	0(10)	6(4)	8(2)	1(9)	10(0)	32(38)
/u/	6(4)	1(9)	8(2)	10(0)	0(10)	10(0)	32(25)
Total	16(34)	1(47)	26(24)	24(24)	7(40)	41(5)	115(174)

Out of all responses, only the consistent ones were considered. Out of 52 participants, two participants, 1 from each of the forms with 6 entries, were found to be inconsistent. Hence, there were 5 consistent participants from each form (10 forms x 5 consistent participants = 50). The repeated samples, intended to check the consistency, were not considered while calculating the results. To conclude, all forms together having 349 samples (including duplicated samples) were filled by five consistent participants, accounting for 1745 responses (= 349 samples $\times$ 5 participants).

3.3. Consonant classification using Machine Learning

This classification experiment aimed to find the effect of formant transition was performed in two settings: *Setting 1*: To identify the consonants from the samples where the consonant portion is removed *Setting 2*: To identify the consonants from the samples with the consonant portion included. While the first setting aided us in investigating the information contained in formant transition, *Setting 2* allowed us to compare the performance of the algorithms with human responses. In both settings, acoustic features of all frames from the first vowel nucleus to the second vowel nucleus are considered except that in the first setting features from frames corresponding to the consonant region were discarded. Multiple experiments were performed using four different classifiers, namely Logistic Regression (LR), Support Vector Machine (SVM) with an RBF Kernel, Random Forest (RF), XGBoost (XGB). Three different features were studied, including statistical features (functionals) of MFCCs alone, formants alone, and MFCCs and formants combined. Eight functionals per formant/MFCC were extracted: average, standard deviation, maximum and minimum value, argmax ratio (= argmax/length of formant vector), argmin ratio, kurtosis, and energy. These functionals were employed in classifying the six stop-consonants. All the models were tested on the 289 samples utilized in the perception test and trained on the remaining 526 samples.

We use scikit-learn to train the automated classifiers, and use the default hyper-parameters. More details could be found in the supplementary document.

4. Results and Discussion

4.1. Data Preprocessing

From Table 2, it is evident that the human speech production system indeed does not always follow a fixed formant transition for producing VCV with a stop-consonant. Out of 815 samples, 26.63% were found to be not abiding by the pattern. There is a subtle difference between the distribution of voiced & unvoiced stop-consonants. The loci proposed in [16] was based on voiced stop-consonants and were generalized to unvoiced stops. But in comparison, in this experiment, a more significant fraction of the unvoiced stops were not following the pattern than their voiced counterparts (cognates), as clear from Table 2. In addition, most of the alveolar stops followed the pattern compared with the bilabial and velar stops.

4.2. Human Perceptual Test

We resort to Unweighted average recall (UAR) as our evaluation metric, which exhibits the mean of the percentage correctly classified in the confusion matrix diagonal. Our reason for considering UAR is the unbalance instances of classes we deal with in the classification task, as is suggested in [34].

From a total of 1445 responses, a UAR of 92.91% was recorded. A random grouping of the responses was done to assess the scores of the 289 samples, i.e., without the five recurrences. Randomly 1 response per sample was selected, which was repeated for all 289 unique samples. This random grouping of 289 samples made up one group. In such a manner, 30 such groups were created, and their average UAR and standard deviation are reported in Table 4. The statistical test yielded a t-value of 3.1370 and a p-value of 0.0027 from the t-test, which mismatched and mismatched samples were considered.

Table 4: *Average UAR (%) of the 30 groups Match-wise with their standard deviation.*

	Match	**Mismatch**	**Total**
Mean(Std Dev)	93.54(1.6)	91.97(1.97)	92.91(1.38)

Note that samples following the pattern were perceived more correctly, on average, than those that did not follow the pattern. The deviation of formant transitions from the pattern induces a minor yet significant change in the auditory impression of the stop-consonant. From Table 5, it is observed that, out of 124 responses that were misclassified in the perception test, 72 samples of them were samples from female subjects, and the remaining 52 were samples from male subjects. Seventeen each from 72 & 52 are to be unique subjects. It was observed that most of the participants faced more difficulty in classifying audio samples uttered by females than males. Table 5 depicts the gender-wise and match-wise (i.e., mismatch and match) distribution among the wrongly chosen stop-consonants from the perceptual test. 50.8 % of total wrong samples constitute the matching category, whereas 49.2 % constitute the mismatch category.

4.3. Consonant classification using Machine Learning

Table 6 showcases the results for the automated classification. The best UAR of 54% in setting 1 was achieved by training a RF model on the formant functionals. Whereas in setting 2, the best UAR of 71.68% was achieved via an RF model trained on the combined functionals of formants and MFCC. The inclusion of the MFCC functionals of the consonant region increased the score by 17.68%. Through this improvement, it is understood that, while the formant transitions contain information on the intervocalic consonant, it alone isn't sufficient to identify the consonant precisely. Breaking down the model's performance in setting 2, it achieves a UAR of 77.5% in identifying the samples following the pattern and 68.35% in identifying the samples that do not follow the pattern. This difference of 9.15% could be attributed to the fewer training samples that do not follow the pattern. In the training set of 526 samples, 424 follow the pattern, and 102 samples do not. Due to the lesser count of the latter, the model is probably unable to generalize to new samples that do not follow the pattern. Indeed when the discarded samples were added to the training set, the model's overall UAR increased by 2%, generating a better-generalized model.

Table 5: *Gender & Match wise distribution of misclassified responses from the perceptual test.*

	Match	**Mismatch**	**Total**
Male	24	28	52
Female	39	33	72
Total	63 (50.80%)	61(49.20%)	124

Table 6: *UAR (%) of Consonant classification using Random Forest Classifier. (·) indicates the standard deviation.*

	Match	**Mismatch**	**Total**
Setting 1	58.96 (4)	43.84 (9)	54.00 (1.9)
Setting 2	77.50 (3)	68.35 (2)	71.68 (2)

5. Conclusion

In this study, we evaluate the role of formant transitions in the perception of intervocalic stop-consonants. For this purpose, we utilize the SPIRE VCV corpus. We find that the human speech-production system does not always follow a similar formant pattern as suggested by [16] for producing the same stop-consonant. We investigate the difference in the perception of the samples that follow the pattern and those that do not. In addition, we utilize machine learning classifiers to study the effects of the formant transitions in identifying the stop-consonants. Through these, we show that the deviation of formant transitions from the pattern induces a minor yet significant change in the auditory impression of the stop-consonant. While the formants had little impact on the perception of stop-consonants, they had a larger impact on the machine learning classifiers. Eventually, further analysis is required to appraise the appropriateness concerning the formant transition observed here, including the analysis of various speaking rates of samples. Future studies could explore formant transitions in other languages, such as Czech, which has palatalized voiced stops.

6. Acknowledgements

This work was partially funded by the Department of Science & Technology (DST), Govt of India, and the Swiss National Science Foundation through the Bridge Discovery project EMIL: Emotion in the loop - a step towards a comprehensive closed-loop deep brain stimulation in Parkinson's disease (grant no. $40B2-0\_194794/1$).

7. References

[1] A. M. Liberman and I. G. Mattingly, "The motor theory of speech perception revised," *Cognition*, 1985.

[2] S. E. G. Öhman, "Coarticulation in VCV utterances: Spectrographic measurements," *Journal of the Acoustical Society of America*, 1966.

[3] A. K. Namasivayam *et al.*, "Speech sound disorders in children: An articulatory phonology perspective," *Frontiers in Psychology*, 2020.

[4] J. Jakimik and R. Cole, "Role of stop consonants in word recognition," *Journal of the Acoustical Society of America*, 1977.

[5] M. J. Owren and G. C. Cardillo, "The relative roles of vowels and consonants in discriminating talker identity versus word meaning." *Journal of the Acoustical Society of America*, 2006.

[6] F. S. Cooper *et al.*, "Some experiments on the perception of synthetic speech sounds," *Journal of the Acoustical Society of America*, 1952.

[7] A. M. Liberman *et al.*, "The role of consonant-vowel transitions in the perception of the stop and nasal consonants." *Psychological Monographs: General and Applied*, 1954.

[8] A. Liberman *et al.*, "Tempo of frequency change as a cue for distinguishing classes of speech sounds," *Journal of experimental psychology*, 1956.

[9] M. Just *et al.*, "Acoustic cues and psychological processes in the perception of natural stop consonants," *Perception & Psychophysics*, 1978.

[10] M. Halle, G. W. Hughes, and J. A. Radley, "On acoustical cues for stop consonants," *Journal of the Acoustical Society of America*, 1956.

[11] J. Forgie, C. D. Forgie, and E. P. Dickey, "A recognition program for english fricative consonants," *Journal of the Acoustical Society of America*, 1961.

[12] T. J. Edwards, "Probabilistic vector model for voicing mode identification of intervocalic stop consonants," *Journal of the Acoustical Society of America*, 1977.

[13] A. Waibel *et al.*, "Phoneme recognition using time-delay neural networks," *IEEE Trans. Acoust. Speech Signal Process.*, 1989.

[14] X. Kong *et al.*, "Landmark-based consonant voicing detection on multilingual corpora," *Acoustical Society of America Journal*, 2017.

[15] A. Liberman, P. Delattre, and F. Cooper, "Study of one factor in the perception of the unvoiced stop consonants," *Journal of the Acoustical Society of America*, 1952.

[16] P. Delattre, A. Liberman, and F. Cooper, "Acoustic loci and transitional cues for consonants," *Journal of the Acoustical Society of America*, 1954.

[17] D. Erickson, "Articulation of extreme formant patterns for emphasized vowels," *Phonetica*, 2002.

[18] K. N. Stevens and A. S. House, "Development of a quantitative description of vowel articulation," *Journal of the Acoustical Society of America*, 1955.

[19] K. Menon, P. Rao, and R. Thosar, "Formant transitions and stop consonant perception in syllables," *Language and Speech*, 1974.

[20] M. Plauché and M. Sönmez, "Machine learning techniques for the identification of cues for stop place," in *Proc. of Interspeech*, 2000.

[21] A. M. A. Ali, J. V. der Spiegel, and P. Mueller, "Acoustic-phonetic features for the automatic classification of stop consonants," *IEEE Trans. Speech Audio Process.*, 2001.

[22] E. C. Kapnoula, J. Edwards, and B. McMurray, "Gradient activation of speech categories facilitates listeners' recovery from lexical garden paths, but not perception of speech-in-noise." *Journal of Experimental Psychology: Human Perception and Performance*, 2021.

[23] K. S. Nataraj, P. C. Pandey, and H. Dasgupta, "Effect of frication duration and formant transitions on the perception of fricatives in vcv utterances," in *Proc. of ICASSP*, 2020.

[24] V. K. R. Maddela and B. Peri, "Durational and formantshift characteristics of telugu alveolar and bilabial nasal phonemes," *2021 IEEE Mysore Sub Section International Conference (MysuruCon)*, 2021.

[25] E. Kearney *et al.*, "A simple 3-parameter model for examining adaptation in speech and voice production," *Frontiers in psychology*, 2020.

[26] D. R. Lametti *et al.*, "Robust sensoel speech," *Current Biology*, 2018.

[27] D. Fogerty *et al.*, "Simultaneous and forward masking of vowels and stop consonants: Effects of age, hearing loss, and spectral shaping." *Journal of the Acoustical Society of America*, 2017.

[28] B. D. Samuels and B. Vaux, "Silence-cued stop perception: Split decisions," *Oslo Studies in Language*, 2021.

[29] S. Phatak and K. W. Grant, "Effects of temporal distortions on consonant perception with and without undistorted visual speech cues." *Journal of the Acoustical Society of America*, 2019.

[30] T. Purohit *et al.*, "SPIRE VCV: An Acoustic-Articulatory Corpus with Three Different Speaking Rates," in *Proc. of 24th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, 2021.

[31] M. Pitermann, "Effect of speaking rate and contrastive stress on formant dynamics and vowel perception." *Journal of the Acoustical Society of America*, 2000.

[32] T. Gay, "Effect of speaking rate on vowel formant movements." *Journal of the Acoustical Society of America*, 1978.

[33] M. Hunt, "A robust formant-based speech spectrum comparison measure," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 1985.

[34] B. Schuller *et al.*, "The interspeech 2014 computational paralinguistics challenge: cognitive & physical load," in *Proc. of Interspeech*, 2014.