An investigation of the virtual lip trajectories during the production of bilabial stops and nasal at different speaking rates

Tilak Purohit, Prasanta Kumar Ghosh

Electrical Engineering, Indian Institute of Science, Bengaluru 560012, India

tilakpurohit@iisc.ac.in, prasantg@iisc.ac.in

Abstract

We propose a technique to estimate virtual upper lip (VUL) and virtual lower lip (VLL) trajectories during production of bilabial stop consonants (/p/, /b/) and nasal (/m/). A VUL (VLL) is a hypothetical trajectory below (above) the measured UL (LL) trajectory which could have been achieved by UL (LL) if UL and LL were not in contact with each other during bilabial stops and nasal. Maximum deviation of UL from VUL and its location as well as the range of VUL are used as features, denoted by VUL_MD, VUL_MDL, and VUL_R, respectively. Similarly, VLL_MD, VLL_MDL, and VLL_R are also computed. Analyses of these six features are carried out for /p/, /b/, and /m/ at slow, normal and fast rates based on electromagnetic articulograph (EMA) recordings of VCV stimuli spoken by ten subjects. While no significant differences were observed among /p/, /b/, and /m/ in every rate, all six features except VLL_MD were found to drop significantly from slow to fast rates. These six features were also found to perform better in an automatic classification task between slow vs fast rates compared to five baseline features computed from UL and LL comprising their ranges, velocities and minimum distance from each other.

Index Terms: Virtual lip trajectory, speaking rate, bilabial stops

1. Introduction

In articulatory phonetics, speech is defined as a series of unique articulatory gestures toward and out from the sound specific articulatory configurations resulting in a series of speech events [1]. In this work, we focus on articulatory configuration of bilabial stop consonants (/p/, /b/) and nasal (/m/), during which the upper and lower lips come together creating a closure [2].

Several works in the past have investigated the lip kinematics during labial stop production. Löfqvist [3] analyzed the lip kinematics in short and long stops in Japanese and Swedish speakers. Löfqvist [4] also examined the control of bilabial closure and release in stop consonants. Son et al. [5] reported reduction in lip aperture gesture in fast rate in bilabial stop /p/ in the context /a/-to-/a/. Lai et al. [6] examined the effect of aspiration and vowel context on lip movements during production of Cantonese bilabial plosives. M Son [7] examined how the upper and lower lips articulate to produce labial /p/ in Korean. Löfqvist et al. showed [3] that, for a bilabial stop production, the lips attain peak velocity at the moment of lip closure to create an airtight closure, and the impact due to this peak velocity results in a tissue compression thus creating an airtight seal. Furthermore, it was reported that during the bilabial stop consonant production, the lower lip (LL) pushes the upper lip (UL) in a vertically upward direction, not allowing upper lip to reach its lowest vertical position, thus proposing the idea of virtual targets for lip movements. A virtual articulatory target, thus, refers to a position beyond regular physiologically possible articulatory position. For example, during a bilabial stop production, the virtual target for the upper lip could be a position, vertically lower than lowest position achievable by the upper lip and similarly the virtual target for the lower lip could be a position, vertically higher than the highest position reachable by the lower lip. Understanding of virtual articulatory targets could shed light on understanding the motor planning reflected in articulatory motion during speech production, which, in turn, could be used to improve articulatory speech synthesizer [8, 9]. To the best of our knowledge, there is no prior work that quantifies the virtual lip targets in a data-driven manner.

With respect to the virtual lip targets during bilabial stop production, the dynamics of lips get deviated from its virtual target due to the tissue compression, when lips come in contact with each other. We hypothesize that the amount of deviation in lip movement due to labial constriction could be quantified if virtual lip trajectories during bilabial stop were known. In this work we propose a method to estimate virtual lip trajectories namely, virtual upper lip (VUL) and virtual lower lip (VLL) trajectories during the production of bilabial stop consonants (/p/, /b/) and nasal (/m/). This is done by posing it as an optimization problem to find the smoothest trajectories corresponding to both the lips satisfying constraints related to the lip dynamics before and after labial constriction. In addition to generating VUL, we compute the deviation between the UL and the estimated VUL, in particular the value and location of the maximum deviation and range of the virtual trajectory during the consonant in a vowel-consonant-vowel (VCV) sequence. These are done for VLL as well. With these representations derived from VUL and VLL, we address the following questions: 1) How do the VUL and VLL representations vary across /p/, /b/, /m/? 2) How do they vary with speaking rates? 3) How well these representations can discriminate different speaking rates?

Experiments are carried out using articulatory movement recordings from ten subjects (5 male + 5 female) using electromagnetic articulograph and VCV stimuli at three different speaking rates (slow, normal, fast). Analyses reveal that while there is no significant difference across /p/, /b/ and /m/, the VUL and VLL representations differ significantly across rates. In particular, when these features are used for classification of slow vs fast speaking rates, they provide an average F1-score of 0.8 when data from /p/, /b/, and /m/ are combined.

2. Dataset

For this work, the lip movements were recorded using 3D Electromagnetic Articulograph (EMA) AG501 [10] from ten subjects (5 male + 5 female) of age range 18-22 years. The EMA data collection procedure and protocol was similar to the one outlined in [11]. All subjects were non-native speakers of English. The lip sensors were placed below and above the vermilion border of the upper lip and lower lip respectively. When the lips were in the closed position, the vertical separation between

the two sensors were noted to be approximately 1cm for all subjects. Sensors were also placed on the tongue tip, tongue body, tongue dorsum, jaw, left ear and right ear (last two for the head movement correction). None of the subjects reported any history of speech or hearing disorder. Each speaker was asked to speak the utterance of the format - "Speak VCV Today" where each utterance was repeated thrice in each of the three different speaking rates, namely slow, normal/moderate and fast making a total of nine utterances for one VCV sequence. Recordings of the lip movements at a sampling rate of 250Hz in the mid-sagittal plane, i.e., UL_x, UL_y, LL_x, LL_y are used for the analysis in this work. In fast speaking rate, the average duration of the consonant is 0.08±0.02 seconds, 0.06±0.01 seconds and 0.06±0.01 seconds for /p/, /b/, and /m/ respectivey. Those for normal rate are 0.13±0.04, 0.09±0.02 and 0.10±0.03 seconds. And for the slow rate, they are 0.22±0.08, 0.14±0.05 and 0.20±0.15 seconds. These duration values of consonants across rates suggest that subjects could follow the given instructions well during recording.

The list of VCV stimuli had all 15 possible combinations of three consonants (C) namely /p/, /b/, /m/ and five vowels (V) /a/, /e/, /i/, /o/ and /u/. Thus, we have a total of 450 (=3 repetitions$\times$ 5vowels $\times$ 3 rates$\times$ 10 subjects) recordings for every consonant. All recordings were done in a sound proof studio at the SPIRE Lab's Speech Production Facility, Indian Institute Science, Bangalore, India. The VCV boundaries were manually annotated by a team of four members. The boundaries were marked by observing the spectrogram, the raw waveform and the glottal pulses (obtained using Praat ([12]) simultaneously using an in-house built Matlab based annotation tool. The most challenging task was to mark the boundaries for fast cases specially for the labial nasal /m/.

3. Estimation of virtual lip trajectories and their representations

In this work, the vertical direction of the UL and LL (i.e., UL_y and LL_y) are only used for deriving the VUL and VLL. For simplicity we will denote UL_y and LL_y by UL and LL, respectively, from now onward.

3.1. Estimation of virtual upper and lower lip trajectories

Consider the UL and LL trajectories during a VCV production as shown in Figure 1A. The duration of C region (denoted by vertical dashed lines, from sample index n_2 to n_3) is N samples. We also mark a segment of duration N/3 samples immediately before and after the C region marked by vertical dotted lines. We assume these two segments of duration N/3 samples are V-C and C-V transition regions. The region between the two vertical dotted lines (of duration 5N/3 samples, sample index n_1 to n_4) is referred to as the extended C region. It is clear from Figure 1A that the UL trajectory goes down in V-C transition region while LL trajectory goes up during that time. Similarly, the UL trajectory goes up in the C-V transition region while LL trajectory goes down. Thus, in the transition regions, UL and LL follow an opposite trend. This trend is not present in the C region, when UL and LL come in contact with each other. Rather the interaction between UL and LL in C region is more complex than that in the transition regions. We hypothesize that the virtual upper and lower lip trajectories follow the trend of transition region throughout the extended C region.

We capture the interaction between the UL and LL using a time-varying affine function. Let $UL[n]$ and $LL[n]$ denote the

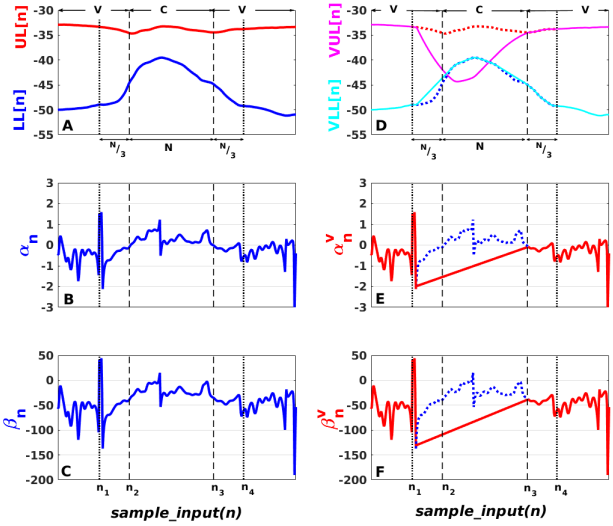

Figure 1: *Illustration of the original UL and LL trajectories (left column) and virtual UL and LL trajectories (right column). Last two rows show the parameters of the affine function relating upper and lower lips movements.*

UL and LL values, respectively, at the n-th sample. We assume that the $UL[n]$ and $LL[n]$ are approximately related by an affine function: $UL[n] \approx \alpha_n LL[n] + \beta_n$, where α_n and β_n are the time varying slope and intercept at sample index n. α_n and β_n are estimated assuming locally linear relation between UL and LL around sample n. M samples before and after sample index n are considered and a straight line is fit to the $2M+1$ pairs of UL and LL values. The slope and intercept of this line are used as estimates of α_n and β_n, respectively. The choice of M depends on the sampling rate of the UL and LL trajectories. As the UL and LL recordings are available at a sampling rate F_s=250Hz, the UL and LL trajectories may not satisfy locally linear relationship for a large choice of M. Hence, we upsample the UL and LL at a sampling rate $F_u > F_s$, from where $2M+1$ samples are chosen around the target sample index n at F_s sampling rate.

The estimated α_n and β_n are shown in Figure 1B and 1C, respectively, for the UL and LL trajectories shown in Figure 1A. It is clear from Figure 1B that α_n is negative in the transition regions indicating the opposite trends in the UL and LL dynamics.

We assume that a relation between the VUL and VLL in the entire extended C region is similar to that between UL and LL in the transition region. We further assume that the variation of α_n and β_n in the case of VUL and VLL is linear from V-C transition region to C-V transition region. Thus, for a sample index m_1 in the V-C transition region and a sample index m_2 in the C-V transition region, the linearly interpolated α_n^v and β_n^v for VUL and VLL are obtained as follows:

$$\alpha_n^v = \alpha_{m_1} + (n-m_1)\frac{\alpha_{m_2}-\alpha_{m_1}}{m_2-m_1},\ \forall m_1 \le n \le m_2$$
$$\beta_n^v = \beta_{m_1} + (n-m_1)\frac{\beta_{m_2}-\beta_{m_1}}{m_2-m_1},\ \forall m_1 \le n \le m_2 \quad (1)$$

Exemplary α_n^v and β_n^v are illustrated in Figure 1E and 1F respectively for the α_n and β_n in Figure 1B and 1C. It should be noted that α_n^v and β_n^v are, respectively, identical to α_n and β_n for $n < m_1$ and $n > m_2$. With this linear variation of slope

and intercept between m_1 and m_2, we pose the estimation of VUL and VLL as an optimization problem, where $VUL[n] = \alpha_n^v VLL[n] + \beta_n^v$, as follows:

$$\{VLL[n], m_1 \leq n \leq m_2\} = \arg\min_{\{x_m\}} \frac{1}{m_2 - m_1} \sum_{m=m_1+1}^{m_2} (x_m - x_{m-1})^2$$

$$\text{such that } LL[m] \leq x_m \leq \max_{n_1 \leq k \leq n_4} UL[k]\,,$$

$$\min_{n_1 \leq k \leq n_4} LL[k] \leq \alpha_n^v x_m + \beta_n^v \leq UL[m], \ \forall m_1 \leq m \leq m_2$$

$$\text{and } x_{m_1} = LL[m_1], \ x_{m_2} = LL[m_2] \quad (2)$$

The objective function ensures that the estimated $VLL[n]$ is smooth and low pass in nature similar to a typical LL trajectory, as the optimization minimizes the energy of the first order difference of the optimization variable sequence. The constraints in the above optimization ensures that the VLL lies above LL and VUL lies below UL by the definition of a virtual lip trajectory. An upper limit on the VLL is used as the maximum value of UL in the extended C region. Similarly a lower limit on the VUL is used as the minimum value of the LL in the extended C region. The boundary (equality) constraints ensure that the estimated VLL matches with LL at m_1 and m_2 to make sure that the VLL matches with LL outside the $m_1 \leq m \leq m_2$ as the virtual lip trajectory is estimated around the bilabial stop only. VUL is obtained from estimated VLL using $VUL[n] = \alpha_n^v VLL[n] + \beta_n^v$. Exemplary estimates of the VUL and VLL are shown in Figure 1D for the UL and LL in Figure 1A. It is clear that the estimated trajectories are smooth in nature and satisfy all constraints.

For every choice of $m_1 (n_1 \leq m_1 \leq n_2)$ and $m_2 (n_3 \leq m_2 \leq n_4)$, VLL can be estimated using eq 2. The best choices of m_1 and m_2 are selected by running the optimization (eq 2) for all possible combinations of m_1 and m_2 and selecting the one which results in the least objective function value.

3.2. Representations derived from VUL and VLL

Three features are extracted from estimated VUL. Two of these features are based on the deviation of UL from VUL. The maximum deviation (MD) and the corresponding location (MDL) normalized by the consonant region duration are used as two features as follows: VUL_MD = $\max_{n_2 \leq k \leq n_3} |UL[k] - VUL[k]|$, VUL_MDL = $\frac{\eta - n_2}{n_3 - n_2}$, where $\eta = \arg\max_{n_2 \leq k \leq n_3} |UL[k] - VUL[k]|$. The third feature is the range of the VUL as follows: VUL_R=$\max_{n_1 \leq k \leq n_4} VUL[k] - \min_{n_1 \leq k \leq n_4} VUL[k]$. Similarly three features from VLL are also computed. They are denoted by VLL_MD, VLL_MDL and VLL_R, respectively.

4. Experimental Setup

The VUL and VLL trajectories are computed for every recording of VCV separately with F_u=5kHz. The window size $(2M + 1)$ for computing the α_n and β_n values are varied from 5 (1ms) to 51 (10ms). As varying the window size could result in different estimates of α_n and β_n, we vary the window size from 1ms to 10ms to examine how it, in turn, causes variations in the estimated VUL and VLL. As, in the proposed computation of VUL and VLL, the α_n and β_n trajectories are assumed to vary linearly, it may not guarantee a solution of the optimization problem (eq 2) always. It turns out that 14% of all the recordings (7%, 12%, 21% of all slow, normal and fast recordings) used in this work do not yield any estimate of the VUL and VLL. These are excluded from further analysis. VUL_MD, VUL_MDL, VUL_R, VLL_MD, VLL_MDL, VLL_R are calculated from the original UL, LL and the estimated VUL, VLL trajectories for each VCV recording. The parameters are pooled from all subjects in each speaking rate separately in order to carry out speaking rate specific analysis of these parameters. Welch's t-test [13] is performed to find out if each of these parameters differs significantly from slow to fast speaking rates. We also investigate the power of each of these six parameters for discriminating slow vs fast rate. As baseline features, the range, velocity (in the extended consonant region) of UL and LL as well as minimum distance between them (in the consonant region) are considered. The classification is carried out in a ten fold cross validation setup. For this purpose, the data from all ten subjects are pooled in each rate, randomly shuffled and divided into ten groups. In each fold, nine groups are used for training and the remaining one group is used as the test set in a round robin fashion. The SVM classifier with radial basis kernel has been trained using the python sklearn package [14][15]. Default values of the soft margin constant (C) and width of the Gaussian kernel (γ) are used. The VUL and VLL parameters separately as well as their various combinations are used for classification tasks. F1-score [16] is used as an evaluation metric to compare merits of different features.

5. Results and Discussion

5.1. Analysis of VUL and VLL parameters at different rates

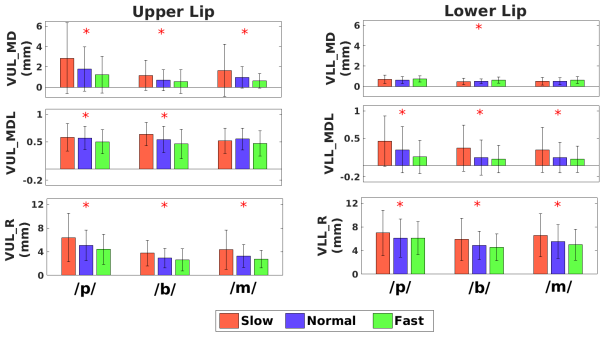

Figure 2: *Comparison of different representations from VUL and VLL across slow, normal and fast rates separately for /p/, /b/, and /m/*

Figure 2 shows the bar plots of six proposed parameters computed using window size of 51 samples at slow, normal and fast rates using red, blue and green bars, respectively, separately for /p/, /b/, /m/. The errorbar shows the standard deviation. A red star indicates that the parameter in the respective case is significantly ($p < 0.01$) different between slow and fast rates. It is clear from the figure that except for VLL_MD, every parameter value, on average, reduces with increasing speaking rate.

A significant drop in VUL_MD value from slow to fast suggests that in the fast rate the measured UL trajectory is more close to the VUL trajectory compared to that in the slow rate. This happens for all /p/, /b/ and /m/ cases. In fast speaking rate, the duration of the labial stop (as given in Section 2) is smaller than their counterparts in slow rate. This, in turn, means that the lips stay in contact for a longer time to create closure in the slow rate than that in the fast rate. This causes more mechanical interaction between lips in slow than fast rate causing larger deviation in the former than the latter.

Features	/p/	/b/	/m/	/p/,/b/,/m/
Baseline	0.68(.05)	0.71(.05)	0.65(.08)	0.75(.04)
VUL_MD	0.64(.07)	0.70(.09)	0.63(.13)	0.66(.04)
VLL_MD	0.58(.09)	0.62(.13)	0.61(.08)	0.62(.03)
VUL_MDL	0.48(.15)	0.57(.08)	0.40(.11)	0.51(.05)
VLL_MDL	0.63(.10)	0.53(.08)	0.54(.08)	0.57(.05)
VUL_R	0.52(.06)	0.63(.09)	0.63(.08)	0.62(.07)
VLL_R	0.51(.12)	0.57(.08)	0.51(.11)	0.56(.06)
Range	0.68(.03)	0.70(.07)	0.62(.10)	0.63(.05)
MD	0.70(.05)	0.76(.07)	0.74(.11)	0.74(.04)
MDL	0.64(.10)	0.60(.08)	0.54(.11)	0.64(.05)
All	0.78(.08)	0.83(.08)	0.72(.04)	0.80(.04)

Table 1: *F1-score from the slow vs fast rate classification using baseline features and features from proposed virtual lip trajectories. (·) indicates the standard deviation across 10 folds.*

A significant drop from slow to fast is also seen when VUL_R and VLL_R are considered. This suggests that as the speaking rate increases, the range of the virtual motion of the upper and lower lip decreases. The lips, in its virtual motion, has to reach its target at an instant within a consonant region starting from its position in vowel region and again returning back to post-consonant vowel position completing its virtual movement cycle. In the fast rate, the range of VUL and VLL become small as lips get less time to complete the movement cycle. Therefore, it compromises on the extent of the movement from its position during vowels on either side of the consonant. However, in slow rate, the planning for virtual lip movement could exploit longer duration to result in a relatively larger range for VUL and VLL. It is interesting that such significant change in VUL_R and VLL_R happens although the velocity of the upper lip increases significantly ($p < 0.01$) from 0.09 mm/s (during /p/), 0.10 mm/s (during /b/), and 0.09 mm/s (during /m/) in slow rate to 0.16 mm/s (during /p/), 0.14 mm/s (during /b/), and 0.14 mm/s (during /m/) in fast rate. This is true for the lower lip as well. In spite of the increased velocity in fast case the range of VUL reduces compared to its slow rate counterpart. This suggests that such a reduced range of VUL is probably a result of the articulatory planning during bilabial stop production at different rates. We also find that the factor by which the range of virtual lip trajectory changes from that of the measured lip trajectory decreases with increasing speaking rates.

From the plots in the middle row in Figure 2, it is clear that the deviation of the observed UL from the estimated virtual UL trajectory is maximum nearly in the middle of the consonant region. Unlike this, the maximum deviation in the case of LL occurs in the initial part of the consonant segment, particularly (within first 1/5-th of the consonant segment) in the case of fast speaking rate. Comparing the VUL_MD and VLL_MD, it turns out that, in the slow rate, the VUL_MD is significantly ($p < 0.01$) higher than VLL_MD for /p/, /b/ and /m/. This is true for normal but not for fast rate. In other words, this indicates that LL matches the VLL more closely than what UL does. As during the consonant segment, the UL and LL come in contact with each other and interact to deviate their trajectory from the respective VUL and VLL, it appears that LL exerts more pressure on the UL causing UL to deviate more from its virtual target, similar to the finding by Löfqvist et al. [3].

5.2. Slow vs fast rate classification using features from VUL and VLL

Apart from the statistical analysis, we examine the extent to which various features from VUL and VLL (computed using window size 51) provide cues for classification of slow vs. fast rates. For comparison, we derive features from measured UL and LL trajectories and use them as baseline features. The ranges of the UL and LL trajectories are known to decrease with increasing speaking rate. The amount of pressure between two lips during closure also varies across rates which may be captured by the distance between UL and LL , i.e., lip aperture, as reported by Son [5]. Thus, range of both lips and their minimum distance in consonant region is included in the baseline features. Lip velocities also vary with speaking rates [17]. Hence, the mean velocities of UL and LL (in the extended consonant region) are included in the baseline features.

F1-score of the two class (slow vs fast) classification are shown in Table 1 using baseline features as well various combinations of six proposed features. It is clear from Table 1 that among six proposed features (cyan colored rows), VUL_MD, on average, performs the best for all bilabial stops. We also examine, which among MD, MDL and Range features perform the best. For this purpose, we combine each of these features from both VUL and VLL and report the F1-score in 8-th to 10-th rows (green rows) in Table 1. Combining features this way improves F1-score over their respective individual cases. It is clear that in case of all bilabial stops, the MD features perform the best followed by Range features followed by the MDL features. Finally, when all six features are combined together (pink colored row in Table 1) they result in the best F1-score, in particular an F1-score of 0.8 when all data from /p/, /b/, and /m/ are combined. This is significantly ($p < 0.05$) better than the F1-score obtained by the baseline features (gray colored row in Table 1) and the best performing single feature, namely VUL_MD. These classification results suggest that the virtual lip trajectories computed by the proposed technique provide better discrimination between slow and fast rates compared to features from measured lip movements.

When the window size is varied for computing VUL and VLL and features from them, we observe that there is minimal changes in α_n and β_n and, hence, estimated VUL and VLL and the six representations do not change much. Thus, the results reported in this section hold good for different window sizes.

6. Conclusions

Virtual lip trajectories during bilabial stop computed by the proposed approach in this work are found to significantly vary across speaking rates. In fact, representations derived from them are found to yield an F1-score of 0.8 for a slow vs fast rate classification task. The rate specific variation in the virtual lip trajectories obtained using the proposed approach could reveal speaking rate specific articulatory planning for the production of bilabial stops and nasal. The motion of LL is partly contributed by the jaw movement. Thus, normalized LL movement by removing the effect of jaw may provide insight into the nature of motor control for lip motion. This is part of our future work. We would also like to explore ways (e.g., relaxing the linear variation of α_n and β_n) of formulating the optimization problem so that the solutions of VUL and VLL exist for any given UL and LL trajectory.

7. Acknowledgement

Authors thank Shankar Narayanan for his help with generating figures and the Department of Science & Technology (DST), Govt of India for their support.

8. References

[1] S. Tilsen and L. Goldstein, "Articulatory gestures are individually selected in production," *Journal of phonetics*, vol. 40, no. 6, pp. 764–779, 2012.

[2] L. R. Rabiner and R. W. Schafer, "Introduction to digital speech processing," *Foundations and trends in signal processing*, vol. 1, no. 1, pp. 1–194, 2007.

[3] A. Löfqvist and V. L. Gracco, "Lip and jaw kinematics in bilabial stop consonant production," *Journal of Speech, Language, and Hearing Research*, vol. 40, no. 4, pp. 877–893, 1997.

[4] A. Löfqvist and V. L. Gracco, "Control of oral closure in lingual stop consonant production," *The Journal of the Acoustical Society of America*, vol. 111, no. 6, pp. 2811–2827, 2002.

[5] M. Son *et al.*, "Coordinative movement of articulators in bilabial stop /p/," *Phonetics and Speech Sciences*, vol. 10, no. 4, pp. 77–89, 2018.

[6] T. Y. Lai, M. N. Wong, M. L. Ng, and E. T. Tong, "Effects of aspiration and vowel context on lip and jaw kinematics in cantonese bilabial plosive production," *Speech, Language and Hearing*, vol. 18, no. 4, pp. 212–218, 2015.

[7] M. Son *et al.*, "Word-boundary and rate effects on upper and lower lip movements in the articulation of the bilabial stop /p/ in korean," *Phonetics and Speech Sciences*, vol. 10, no. 1, pp. 23–31, 2018.

[8] P. Birkholz, B. J. Kröger, and C. Neuschaefer-Rube, "Articulatory synthesis and perception of plosive-vowel syllables with virtual consonant targets," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[9] B. J. Kröger and P. Birkholz, "A gesture-based concept for speech movement control in articulatory speech synthesis," in *Verbal and Nonverbal Communication Behaviours*. Springer, 2007, pp. 174–189.

[10] [Online]. Available: https://www.articulograph.de/

[11] A. Illa and P. K. Ghosh, "The impact of speaking rate on acoustic-to-articulatory inversion," *Computer Speech & Language*, vol. 59, pp. 75–90, 2020.

[12] B. Paul and W. David. Praat: doing phonetics by computer. [Online]. Available: http://www.fon.hum.uva.nl/praat/

[13] B. L. Welch, "The generalization of 'Student's' Problem when several different population variances are involved," *Biometrika*, vol. 34, no. 1-2, pp. 28–35, 01 1947. [Online]. Available: https://doi.org/10.1093/biomet/34.1-2.28

[14] scikit-learn developers. Support vector machines. [Online]. Available: https://scikit-learn.org/stable/modules/svm.html

[15] C.-C. Chang and C.-J. Lin, ""LIBSVM : A library for support vector machines"," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, May 2011. [Online]. Available: https://doi.org/10.1145/1961189.1961199

[16] wikipedia.org. F1_score. [Online]. Available: https://en.wikipedia.org/wiki/F1_score

[17] T. Gay and H. Hirose, "Effect of speaking rate on labial consonant production," *Phonetica*, vol. 27, pp. 44–56, 1973.