Interspeech 2024
1-5 September 2024, Kos, Greece

A comparative study of the impact of voiceless alveolar and palato-alveolar sibilants in English on lip aperture and protrusion during VCV production

Chetan Sharma[1], *Vaishnavi Chandwanshi*[2], *Prasanta Kumar Ghosh*[1]

[1]Electrical Engineering, Indian Institute of Science (IISc), Bangalore-560012, India
[2]Rewa Engineering College, Rewa, Madhya Pradesh-486002, India

chetansharma@iisc.ac.in,vcchandwanshi@gmail.com, prasantg@iisc.ac.in

Abstract

Lip rounding and protrusion during sibilant production are known. Whether these features are discriminative among sibilants and how the discrimination changes in different vowels context are not well explored. In this work, we consider two voiceless sibilants, namely, /s/ (alveolar sibilant) and /ʃ/ (palato-alveolar sibilant) in English during VCV production and show that lip aperture (LA) and lip protrusion (LP) are significantly higher in case of /ʃ/ than /s/ irrespective of the vowels (/a/, /i/, /u/) context. Using the USC Speech MRI database comprising 74 subjects speaking VCV sequence, we also show that, when used for /s/ vs /ʃ/ automatic classification, LA provides the highest classification accuracy of 90.57%($\pm$5.91%) in case of /i/ followed by 85.09%($\pm$5.19%) and 82.38%($\pm$8.76%) in case of /u/ and /a/, respectively. The change in LA and LP from /s/ to /ʃ/ are seen as an effect of higher displacement of lower lip than that of upper lip for /a/ and /i/ unlike that for /u/.

Index Terms: sibilants, lip aperture, lip protrusion, vowel-consonant-vowel (VCV)

1. Introduction

Acoustic and articulatory contrasts of two English sibilants, namely voiceless alveolar sibilant /s/ (as in 'sip') and palato-alveolar sibilant /ʃ/ (as in 'ship'), have been a study of great interest in the literature [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]. In several previous studies, acoustic characterization of these sibilants has been done using various spectral features [11, 12], in particular, spectral moments [13], spectral energy, [14], intensity, centre of gravity, skewness, kurtosis [15]. As the nature and degree of constriction formed by the tongue determines the noise characteristics in these sibilants, several articulatory investigations have been carried out in the past [16]. These include distinctive features based on tongue blade [17], anatomical factors on tongue position variability [18], tongue shape [19], angle between tongue tip and parasagittal blade position [20], constriction location [21].

LP and rounding have been reported in the past during production of /ʃ/, a phenomenon often known as labialization [16, 22]. For example, "the alveolar sibilants tend to have a slightly spread lip shape, whereas the prepalatal sibilants tend to have a slightly protruded or rounded lip shape" [16]. As reported in [23], labiality is a relevant feature for realization of English consonants including sibilants. Labial feature values were found to be significant for sibilants /s/ and /ʃ/ [24], where it is reported that, for most people, lips move slightly in any word containing /s/ and quite considerably in any word containing /ʃ/. More lip rounding was observed in the production of /ʃ/ compared to that of /s/ [25]. Toda et al. [26] reported protrusion with large lip during English /ʃ/ production. It was also reported that more lip protrusion for /š/ was observed than for /s/, except for in the /u_u/ context among three vowels considered in [27]. Bonneau et al. [28] investigated the effect of /ʃ/ on French /i,y/. Five lip parameters were analysed: the height, width and area of lip opening, the distance between the corners of the mouth, as well as lip protrusion. Results showed that the carry-over effect of /ʃ/ can impede the opposition between /i/ and /y/ in the lip protrusion dimension. The effect on lip kinematics during sibilant production has not only been observed in English, but also in several other languages. For example, labialized sibilant was observed if /k/ was preceded by /s/ in Central Swedish [29]. Lip rounding was observed at the beginning of Korean /s/ if the round segment is in the domain of the syllable [30]. Whistled fricative in Changana was reported to be less labialized than the plane rounded /s/ [31].

Investigation of the labialization in the context of /s/ and /ʃ/, in the past, have used different number of speakers and different modalities of data [25, 26, 27, 28, 32]. The number of speakers in these studies vary from 3 to 10 including both male and female subjects [25, 26, 27, 28], while the study comparing adult and children in [32], 15 speakers in each category were used. It is not clear how the findings from these studies would generalize when applied on a larger set of speakers as articulation during speech production may vary depending on the shape, size and morphology of the speaker. No systematic comparison among different lip kinematic features have been done in the past particularly when /s/ and /ʃ/ are uttered in different vowels contexts. The previous studies reported either pictorial or graphical comparison or hypothesis testing to claim the extent of labialization. No thorough classification experiments were carried out to examine the discriminatory power of labial features between /s/ and /ʃ/ in different vowels contexts.

In this study, we carry out experiments to address some of these gaps present in the literature. In particular, we consider production of /s/ and /ʃ/ in the context of three vowels, namely, /a/, /i/, and /u/ during VCV production from 74 speakers, the largest corpus used in similar studies to the best of our knowledge. We use rtMRI videos during speech production as they clearly delineate the vocal tract boundaries including lip shapes required for labial feature extraction. We use lip aperture (LA) and lip protrusion (LP) as labial features in this study. In particular, we address the following questions: 1) Is LA (as well LP), significantly different between /s/ and /ʃ/ in different vowels contexts? 2) for a particular sibilant, are LA and LP significantly different across three vowels? 3) How accurately do LA and LP features classify /s/ vs /ʃ/? 4) What is the classification accuracy among /a/ vs /i/ vs /u/ when LA and LP during sibilant are used for classification?

Experiments using LA and LP from rtMRI video frames from 74 speakers speaking VCV with two sibilant consonants (/s/ and

Authors thank the Department of Science and Technology (DST), Govt of India for their support in this work.

10.21437/Interspeech.2024-2415

/ʃ/) and three vowels (/a/, /i/, /u/) reveal that both LA and LP are significantly higher during production of /ʃ/ compared to /s/. The highest LA value is achieved in case of vowel /i/ for both sibilants. However, when vowel /u/ is used, the highest LP value is observed indicating vowel specific labialization. The highest automatic /s/ vs /ʃ/ classification accuracy of 90.57%($\pm$5.91%) is obtained in case of vowel /i/ when LA is used as a feature indicating the discriminatory power of LA between the two sibilants in a vowel dependent manner.

2. Dataset

2.1. USC speech rtMRI database

We have utilized data from a subset of the USC 75-speaker speech rtMRI video database [33]. The videos were recorded at a frame rate of 83.277 frames/second [33], capturing subtle articulatory movements. Each frame of every video has a spatial resolution of 84 × 84 pixels with a spatial area of 2.4 × 2.4 mm^2 covered by every pixel. This resolution facilitated detailed visualization of lip movements and articulatory gestures during the production of /s/ and /ʃ/ sounds across different vowels contexts. We focused on a specific subset of this corpus, which comprises VCV sequences involving the /s/ and /ʃ/ spoken in the context of three different vowels, namely /u/ as in /usu/ and /uʃu/, /i/ as in /isi/ and /iʃi/, and /a/ as in /asa/ and /aʃa/. The compiled dataset enabled comprehensive analysis and comparison of LA and LP patterns associated with the pronunciation of /s/ and /ʃ/ in different vowels contexts. We considered 74 speakers (35 Males and 39 Females) among 75 available speakers as one female speaker did not have audio recordings of these stimuli of interest.

In order to extract video frame corresponding to the sibilants considered in this work, we marked the starting and ending points of the /s/ and /ʃ/ segments in every VCV for each subject across the three vowels contexts. This was done with help of researchers working in speech acoustics-phonetics using Audacity [34][35] observing both waveform and spectrogram. Subsequently, we identified the midpoint of these segments and the video frame closest to the midpoint is used as representative of the articulatory shape during /s/ and /ʃ/ production. This approach resulted in the extraction of 74 frames for each vowel (/u/, /i/, and /a/) and sibilant (/s/ and /ʃ/) combination.

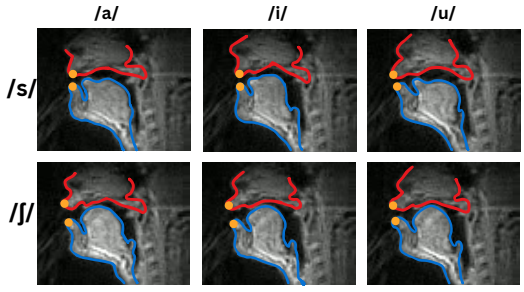

Figure 1: *Illustration of the UL and LL positions (yellow dots) during /s/ and /ʃ/ production with different vowel (/a/, /i/ and /u/) context during VCV production. Red and blue contours show the upper (C1) and lower (C2) contours, respectively.*

2.2. rtMRI video frame annotation

Manual annotation was carried out using a MATLAB-based Graphical User Interface (GUI), in a manner similar to that outlined in [36]. This involved the identification and delineation of Contour 1 (C1) (comprising, nose, upper lip, hard and soft palate, velum) and Contour 2 (C2) (comprising jaw, lower lip, tongue, epiglottis), along with the marking of landmark points representing key anatomical features, such as the upper lip (UL) and lower lip (LL), as shown in Fig.1. As manual marking of UL and LL may not be precise as they may not fall on the C1 and C2, respectively, the UL and LL points are programmatically updated to ensure the UL marks the part of upper lip with lowest X-coordinate value, i.e., the outermost part of UL, consistently for all rtMRI images of all speakers. The manually marked UL is used to identify the UL region from the C1 for the programmatic updation. Similarly, LL is also updated, shown by yellow cross in Fig. 2(A).

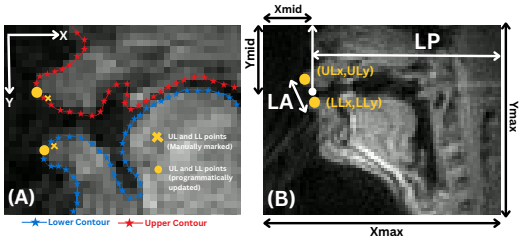

Figure 2: *(A) Illustration of UL & LL points manually marked (yellow cross) and UL& LL points programmatically marked (yellow dot), (B) Description of LA and LP calculation*

2.3. Illustration of articulatory shape during /s/ and /ʃ/ production

Fig. 1 shows the shape of the C1 and C2 during production of /s/ and /ʃ/ in VCV in the context of three vowels for the same speaker. It is clear that the tongue shape significantly changes from /s/ to /ʃ/. This is consistent with observation in [19], as tongue is the critical articulator for these two sounds [32][37]. From the UL an LL points, it is clear that lips are away with a higher distance during /ʃ/ production compared to that during /s/ production. This is observed for each of the three vowels considered for the speaker chosen. We further examine this quantitatively for all 74 speakers.

3. Experiments and results

There are several lip kinematics related features used in the past including LA, lip height, distance between lip corners, and LP. However, for the experiments in this study, we have calculated LA and LP, based on the manually (further programmatically updated) UL and LL points. The LA and LP calculation is depicted in Fig.2(B). In particular, LA $= \sqrt{(ULx - LLx)^2 + (ULy - LLy)^2}$ (following the definition in [38]). LP is calculated as follows: LP $= |X_{\text{mid}} - X_{\text{max}}|$, where $X_{\text{mid}} = \frac{ULx+LLx}{2}$, and X_{max}=84×2.4mm=201.6mm, the number of columns in an rtMRI video frame. Both statistical analysis and automatic classification are carried out to examine the extent to which LA and LP can discriminate /s/ and /ʃ/. This is also done to examine the discrimination power of LA and LP for three vowel classes, /a/, /i/, and /u/.

3.1. Statistical analysis using LA and LP

For each of the 74 subjects used in this study, we calculate six values of LA for /asa/, /aʃa/, /isi/, /iʃi/, /usu/, and /uʃu/. Similarly, six values of LP are also calculated for these cases. This allow us to do paired t-test to examine if the LA (as well as LP) value is significantly different between /s/ and /ʃ/ for each vowel separately. LA and LP may alter depending on the shape and size of the vocal tract of speakers. However, pair-wise test helps compare the change in LA and LP within a speaker while speaking /s/ and /ʃ/. Similarly, for each sibilant, the difference in LA (as well as LP) values across 3 vowel classes are examined.

3.1.1. *Statistical analysis using LA*

With 74 speakers in our database and three vowels, /a/, /i/, /u/, we obtain $74 \times 3 = 222$ LA values for /s/ and similarly 222 LA values for /ʃ/. The normalized histogram of the LA values is shown in Fig. 3(D) for both /s/ and /ʃ/. It is clear that, on average, the LA is significantly ($p < 0.01$) higher for /ʃ/ compared to that for /s/. Specifically, the average LA for /s/ is 7.26mm, while that for /ʃ/ is 14.78mm. This suggests that during /ʃ/ production, the articulatory configuration leads to a higher amount of opening of lips compared to that of /s/.

To investigate how the LA alters across the three vowels considered, we repeat the comparative plot separately for the three vowels as well. These are depicted in Fig. 3(A), (B), and (C) for vowels /a/, /i/, and /u/, respectively. It is evident from the figures that the average LA for /ʃ/ is significantly ($p < 0.01$) higher than that for /s/ for each of the vowels. This suggests that irrespective of the vowel used in VCV in the database considered, the LA is higher for /ʃ/ than /s/.

Interestingly, the LA value changes depending on the vowels in VCV for both /s/ and /ʃ/. The highest average LA is observed for /i/, followed by /a/ and /u/ when /s/ is considered. This is also true for /ʃ/ as well. When considering /s/, we found that the average LA in the case of vowel /i/ (i.e., 8.83mm) is not significantly higher than that of /a/ (i.e., 8.22mm). However, the average LA for both /i/ and /a/ is significantly ($p < 0.01$) higher than that for /u/ (i.e., 4.72mm). When /ʃ/ is considered, we found the average LA for vowel /i/ (i.e., 18.05mm), vowel /a/ (i.e., 15.96mm), and vowel /u/ (i.e., 10.32mm) are significantly ($p < 0.01$) different from each other.

To examine the extent by which the LA changes from /s/ to /ʃ/, we have computed the δ by subtracting LA for /s/ from LA for /ʃ/ for every speaker. The normalized histogram along with average values for δ are shown in the respective subfigures of Fig. 3. It is clear from the histograms that the δ is positive valued for every speaker with the highest average δ observed for /i/ (9.21mm), followed by /a/ (7.74mm) and /u/ (5.60mm), while the average δ is 7.52mm when all vowels are considered together. This indicates that, the LA, during /ʃ/ production, is more than that during /s/ production for each of the 74 speakers considered.

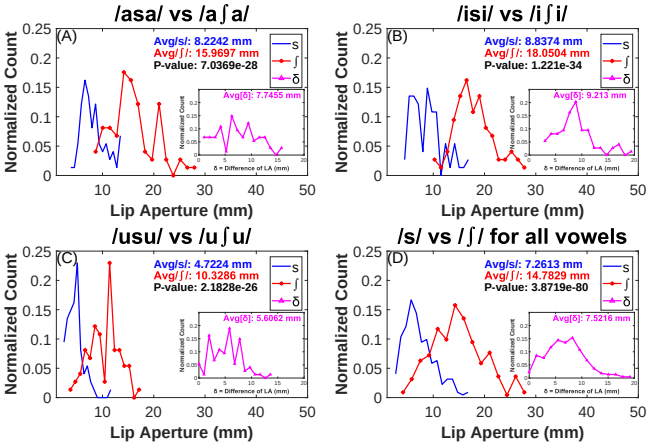

Figure 3: *Comparision of Lips Aperture (LA) for /s/ and /ʃ/ with respect to vowel (A)/a/ (B)/i/ (C)/u/ and (D)All vowels*

3.1.2. *Statistical analysis using LP*

A similar analysis is carried out using LP. The normalized histograms of LP values for each of the three vowel and all vowels together are shown in Fig. 4 for both /s/ and /ʃ/.

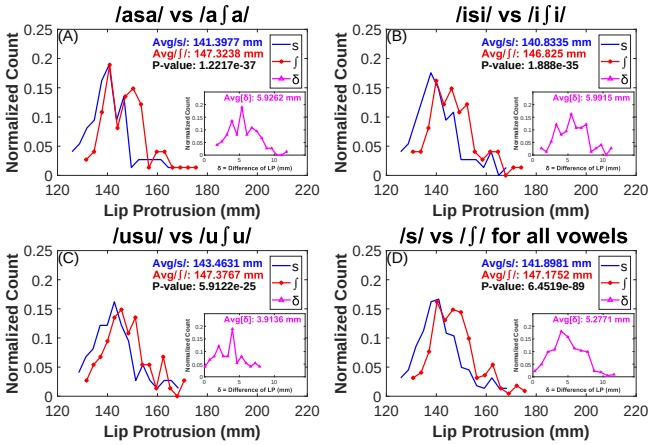

Figure 4: *Comparision of Protrusion(LP) for /s/ and /ʃ/ with respect to vowel (A)/a/ (B)/i/ (C)/u/ and (D)All vowels*

It is evident from the figures that the average LP for /ʃ/ is significantly ($p < 0.01$) higher than that for /s/ for each of the vowels as well as when all vowels are considered together. This suggests that irrespective of the vowel used in VCV in the database considered, the LP, just like LA, is higher for /ʃ/ than /s/.

However, the absolute value of LP is not the same across vowels for both /s/ and /ʃ/. The highest average LP is observed for /u/, followed by /a/ and /i/, when /s/ is considered. This is also true when /ʃ/ is considered. When considering /s/, we found that the average LP in the case of vowel /i/ (i.e., 140.83mm), /a/ (i.e., 141.39mm) and /u/ (i.e., 143.46mm) are significantly ($p < 0.01$) different from each other. When /ʃ/ is considered, we found the average LP for vowel /i/ (i.e. 146.82mm), vowel /a/ (i.e., 147.32mm), and vowel /u/ (i.e., 147.37mm) are not significantly ($p < 0.01$) different from each other. This suggests that the LP tends to achieve similar values irrespective of the vowel when /ʃ/ is produced, which is not the case when /s/ is produced.

In a manner similar to that of LA, we have computed the δ for LP between /s/ and /ʃ/ for every speaker. The δ is found to be positive for every speaker in case of each vowel. The highest average δ is observed for /i/ (5.99mm), followed by /a/(5.92mm) and /u/(3.91mm). The normalized histograms of the δ values for each vowels and all vowels together are shown in respective subplots of Fig. 4.

3.2. Automatic classification using LA and LP

We carry out two different types of classification experiments using LA and LP: 1) /s/ vs /ʃ/ for each vowels and all vowels together, 2) /a/ vs /i/ vs /u/ for each sibilant. This is done to examine the nature of discriminatory cues LA and LP provide across sibilants and vowels in different contexts. Due to small dataset size, we carry out K-Nearest Neighbours (KNN) based classification[39] in a five-fold cross validation setup. Experiments with other classifiers do not provide better results. In every fold, 80% data is used for training and 20% testing. The test set contains class-balanced data from speaker unseen to the training set. Best choice of K (denoted by $K^\star$) in KNN classifier is decided from a set of values $\{3, 5, 7, 9, 11\}$ based on the performance on a dev set which is set as 20% of the training set. Accuracy is used as the performance metric.

/s/ vs /ʃ/ classification using LA: Fig. 5(A) shows fold-wise (columns correspond to fold) confusion matrix, $K^\star$ and accuracy of classification between /s/ and /ʃ/. Each row corresponds to

	(A) /s/ vs /ʃ/ using LA						(B) /s/ vs /ʃ/ using LP						(C) /s/ vs /ʃ/ using LA and LP Jointly					
	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Overall	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Overall	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Overall
/a/	[14 1; 1 14] K* =3 Acc = 93.33%	[14 1; 4 11] K* = 3 Acc = 83.33%	[6 9; 0 15] K* = 3 Acc = 70.00%	[14 1; 3 12] K* = 5 Acc = 86.66%	[11 3; 3 11] K* = 3 Acc = 78.57%	82.38% (±8.76%)	[12 3; 10 5] K* =11 Acc = 56.66%	[11 4; 7 8] K* = 3 Acc = 66.33%	[8 7; 3 12] K* = 5 Acc = 66.66%	[11 4; 3 12] K* = 7 Acc = 76.66%	[7 7; 5 9] K* = 7 Acc = 57.14%	64.09% (±8.19%)	[13 2; 2 13] K* = 3 Acc = 86.66%	[15 0; 5 10] K* = 3 Acc = 83.33%	[8 7; 0 15] K* = 3 Acc = 76.66%	[14 1; 3 12] K* = 3 Acc = 86.66%	[13 1; 4 10] K* = 11 Acc = 82.14%	83.09% (±4.11%)
/i/	[15 0; 0 15] K* = 3 Acc = 100%	[14 1; 3 12] K* = 3 Acc = 86.66%	[11 4; 0 15] K* = 3 Acc = 86.66%	[14 1; 3 12] K* = 5 Acc= 86.66%	[12 2; 0 14] K* = 7 Acc = 92.85%	90.57% (±5.91%)	[10 5; 9 6] K* = 7 Acc = 53.33%	[10 10; 5 5] K* = 7 Acc = 50.00%	[4 11; 2 13] K* = 7 Acc = 56.66%	[10 5; 8 7] K* = 7 Acc= 56.66%	[8 6; 7 7] K* = 3 Acc = 53.57%	54.04% (±2.77%)	[13 2; 0 15] K* = 9 Acc = 93.33%	[14 1; 2 13] K* = 11 Acc = 90.00%	[10 5; 0 15] K* = 7 Acc = 83.33%	[14 1; 1 14] K* = 9 Acc= 93.33%	[12 2; 1 13] K* = 5 Acc = 89.28%	89.85% (±4.09%)
/u/	[14 1; 3 12] K* = 3 Acc = 86.66%	[15 0; 2 13] K* = 3 Acc = 93.33%	[10 5; 0 15] K* = 7 Acc = 83.33%	[12 3; 3 12] K* = 3 Acc = 80.00%	[13 1; 0 10] K* = 5 Acc = 82.14%	85.09% (±5.19%)	[10 5; 9 6] K* = 5 Acc = 53.33%	[11 4; 9 6] K* = 5 Acc = 56.66%	[4 11; 9 6] K* = 5 Acc = 33.33%	[7 8; 7 8] K* = 3 Acc = 50.00%	[6 8; 8 6] K* = 11 Acc = 42.85%	47.23% (±9.30%)	[13 2; 1 14] K* = 5 Acc = 90.00%	[15 0; 3 12] K* = 3 Acc = 90.00%	[10 5; 1 14] K* = 3 Acc = 80.00%	[13 2; 2 13] K* = 5 Acc = 86.66%	[13 1; 3 11] K* = 3 Acc = 95.71%	86.47% (±4.10%)
All Vowels	[40 5; 9 36] K* = 5 Acc = 84.44%	[40 5; 13 32] K* = 9 Acc = 80.00%	[30 15; 2 43] K* = 5 Acc = 81.11%	[39 6; 9 36] K* = 5 Acc = 83.33%	[34 8; 9 33] K* = 5 Acc = 79.76%	81.73% (±2.07%)	[29 16; 26 19] K* = 5 Acc = 53.33%	[34 11; 27 18] K* = 11 Acc = 57.77%	[17 28; 17 28] K* = 9 Acc = 50.00%	[26 19; 23 22] K* = 5 Acc = 53.33%	[22 20; 28 14] K* = 5 Acc = 42.58%	51.46% (±5.54%)	[41 4; 10 35] K* = 5 Acc = 84.44%	[39 6; 17 28] K* = 3 Acc = 74.44%	[23 22; 4 41] K* = 3 Acc = 71.11%	[36 9; 13 32] K* = 3 Acc = 75.55%	[36 6; 11 31] K* = 11 Acc = 79.76%	77.06% (±5.15%)

Figure 5: */s/ vs /ʃ/ classification accuracy using (A) LA (B) LP (C) LA and LP jointly. The fold-specific confusion matrix, best value of K and accuracy are listed in each column. The column with heading 'Overall' shows the average(±SD) of the accuracy across all folds.*

the data corresponding to one vowel in the VCV (e.g., first row for /asa/ vs /aʃa/) except for the last row which is for all vowels together. Although $K^{\star}$ alters across folds, $K^{\star}$=3 is a common choice except for the all vowel case where $K^{\star}$=5 occurs for 4 out of 5 folds. The last column show the average (with standard deviation (SD) in bracket) classification accuracy across five folds. Comparison among three vowels show that the average accuracy is the highest for vowel /i/, which is 90.57%(±5.91%). This is followed by the classification accuracy for vowel /u/ and /a/. Performance drops when all vowels are combined. This result is interesting as it shows cues from lips can provide high degree of discrimination between /s/ and /ʃ/ although lip is not a critical articulator for these sibilants.

/s/ vs /ʃ/ classification using LP: Similar to the Fig. 5(A), classification performance between /s/ and /ʃ/ using LP are shown in Fig. 5(B) for each vowel separately as well as all vowels together. $K^{\star}$=5 or 7 are common choice in different folds when LP is used for classification. It is clear from these figures that the performance using LP is lower than that using LA for each of the three vowels as well as in the case of all vowels. This demonstrates the superior discrimination capacity of LA over LP for /s/ vs /ʃ/ classification. Interestingly, classification accuracy using LP is found to be the highest, i.e., 64.09%(±8.19%) in the case of vowel /a/ among three vowels, unlike when LA is used for classification.

/s/ vs /ʃ/ classification using LA and LP jointly: Fig. 5(C) shows the classification results for /s/ vs /ʃ/ classification when both LA and LP are used as features. Comparing performance among three vowels, it is clear that the classification accuracy is the highest when vowel /i/ is used, similar to that observed when LA only is used for classification. Interestingly, in the case of vowel /i/, the average classification accuracy using LA and LP does not improve over that using LA alone, although this is not the case for vowel /a/ and /u/. This indicates that the complementarity of LA and LP features for /s/ and /ʃ/ is vowel dependent. When data from all vowels are taken together, the accuracy does not improve by using LA & LP jointly over LA alone.

/a/ vs /i/ vs /u/ classification: As average LA and LP were found to be different across various vowels with the difference being significant for most of the vowel pairs, we explore the discriminability among /a/ vs /i/ vs /u/ using automatic classification experiment. Considering LA during /s/ production, we obtain a classification accuracy of 54.92% (±6.24%) among /a/ vs /i/ vs /u/ using a five-fold cross validation setup. The accuracy values are 29.36%(±9.37%) and 56.76%(±10.24%) when LP and LA & LP together are used, respectively. In these classification experiments, the best K value is found to be 7 for most of the folds. The accuracy values are found to be 54.66% (±8.54%), 30.19%(±3.70%), and 56.73%(±3.08%), when the LA, LP and LA & LP together are used, respectively, in case of sibilant /ʃ/. These results suggest that the LA during the sibilant is influenced by the vowel during the VCV production and could provide cues for discrimination across three vowels used in this study.

4. Discussion

LA and LP values being significantly higher in case of /ʃ/ than /s/ is consistent with findings from the studies in the past although this study for the first time show that this is true for three different vowels contexts /a/, /i/, /u/ in a VCV sequence. Vowel dependent LA and LP values in case of both /s/ and /ʃ/ provide an evidence that due to coarticulation, vowels leave their signatures in the lip kinematics during sibilant production leading to a maximum of ∼56% classification accuracy among /a/, /i/ and /u/. With a 74 speakers dataset, this study also demonstrates that the increase in LA and LP values from /s/ to /ʃ/ is the highest when vowel /i/ is considered among three vowels used in this study. During /a/ production lip opens up and during /u/ production significant lip protrusion is observed [40]. It could be that due to this lip position during /a/ and /u/ production, coarticulation effect on the lip movement during sibilant production in a VCV sequence is different in case of /a/ and /u/ vowels compared to that for /i/ vowel, for which LA is relatively lower [40]. Both LA and LP are influenced by position of UL and LL. The largest change in LA and LP are observed for vowel /i/ (δ in Fig. 3 and 4). It is found that, in case of vowel /i/, UL moves by 2.95mm(±1.15mm) and LL moves by 3.71mm(±1.00mm) leading to an average δ value of 9.21mm and 5.99mm, for LA and LP, respectively. Similarly, UL moves by 3.03mm(±1.08mm) and LL moves by 3.24mm(±1.00mm) in case of vowel /a/. In case of both these vowels /i/ and /a/, LL, on average, moves more than UL. However, this is not the case for vowel /u/, for which UL moves by 3.08mm(±1.07mm) and LL moves by 1.34mm(±0.80mm). This could be because of significant lip protrusion in case of vowel /u/.

5. Conclusion

Results in this comparative study quantifies the secondary articulation (i.e., labialization) during two sibilants, namely /s/ and /ʃ/ production in three different vowels contexts using LA and LP on data from 74 speakers. LA was found to classify two sibilants with the highest classification accuracy of 90.57%(±5.19%) in case of vowel /i/ among all sibilant-vowel combination during VCV production. This finding shows that external articulator, like lip, could provide discriminative cues between /s/ and /ʃ/ although tongue (internal articulator) is critical for creating constriction during these sibilant production. Similar studies using other sibilants is required to understand secondary articulation in different vowels contexts. These are parts of our future works.

6. References

[1] A. Nogita, "The/s/-/ʃ/ confusion by Japanese ESL learners in grapheme-phoneme correspondence: bias towards [s] and <s>," *Working Papers of the Linguistics Circle*, vol. 26, no. 1, pp. 45–57, 2016.

[2] G. Bailey, S. Nichols, D. Turton, and M. Baranowski, "Affrication as the cause of/s/-retraction: Evidence from manchester English," *Glossa: a journal of general linguistics*, vol. 7, no. 1, 2022.

[3] V. Bukmaier and J. Harrington, "An analysis of post-vocalic/s-ʃ/neutralization in augsburg german: Evidence for a gradient sound change," *Frontiers in psychology*, vol. 5, p. 86732, 2014.

[4] M. J. Jones and K. McDougall, "The acoustic character of fricated/t/in australian english: A comparison with/s/and/ʃ," *Journal of the International Phonetic Association*, vol. 39, no. 3, pp. 265–289, 2009.

[5] K. Johnson and M. Babel, "On the perceptual basis of distinctive features: Evidence from the perception of fricatives by dutch and english speakers," *Journal of phonetics*, vol. 38, no. 1, pp. 127–136, 2010.

[6] S. Dufour, S. Kriegel, and N. Nguyen, "The perception of the french/s/-/ʃ/contrast in early creole-french bilinguals," *Frontiers in Psychology*, vol. 5, p. 109225, 2014.

[7] D. T. Francisco and H. F. Wertzner, "Differences between the production of [s] and [ʃ] in the speech of adults, typically developing children, and children with speech sound disorders: An ultrasound study," *Clinical linguistics & phonetics*, vol. 31, no. 5, pp. 375–390, 2017.

[8] K. L. Haley, E. Seelinger, K. C. Mandulak, and D. J. Zajac, "Evaluating the spectral distinction between sibilant fricatives through a speaker-centered approach," *Journal of Phonetics*, vol. 38, no. 4, pp. 548–554, 2010.

[9] M. Stevens, J. Harrington, and F. Schiel, "Associating the origin and spread of sound change using agent-based modelling applied to/s/-retraction in English," *Glossa: a journal of general linguistics*, no. 1, 2019.

[10] F. Li, J. Edwards, and M. E. Beckman, "Contrast and covert contrast: The phonetic development of voiceless sibilant fricatives in English and Japanese toddlers," *Journal of phonetics*, vol. 37, no. 1, pp. 111–124, 2009.

[11] K. Iskarous, C. H. Shadle, and M. I. Proctor, "Articulatory–acoustic kinematics: The production of American English/s/," *The Journal of the Acoustical Society of America*, vol. 129, no. 2, pp. 944–954, 2011.

[12] M. Toda, S. Maeda, K. Honda, S. Fuchs *et al.*, "Formant-cavity affiliation in sibilant fricatives," *Turbulent sounds: An interdisciplinary guide*, vol. 21, pp. 343–374, 2010.

[13] A. Jongman, R. Wayland, and S. Wong, "Acoustic characteristics of english fricatives," *The Journal of the Acoustical Society of America*, vol. 108, no. 3, pp. 1252–1263, 2000.

[14] A. Bladon and F. Seitz, "Spectral edge orientation as a discriminator of fricatives," *The Journal of the Acoustical Society of America*, vol. 80, no. S1, pp. S18–S18, 1986.

[15] S. Carralero-Fernandez, "Speaker-specific information in the acoustic characteristics of english fricatives," *IJFL (International Journal of Forensic Linguistic)*, vol. 3, no. 1, pp. 105–115, 2022.

[16] M. C. Pennington, *Phonology in English language teaching: An international approach*. Routledge, 2014.

[17] M. Pennington, "Toward phonetically grounded distinctive features. part ii: Experimental evidence for blade features," *IULC Working Papers*, vol. 11, no. 1, 2011.

[18] K. Rudy and Y. Yunusova, "The effect of anatomic factors on tongue position variability during consonants," pp. 343–374, 2013.

[19] N. Zharkova, "Differentiating tongue shapes for alveolar-postalveolar and alveolar-velar contrasts," *Speech Communication*, vol. 113, pp. 15–24, 2019.

[20] M. Tiede, W.-r. Chen, and D. H. Whalen, "Taiwanese mandarin sibilant contrasts investigated using coregistered ema and ultrasound," *Proceedings of ICPhS*, pp. 427–431, 2019.

[21] M. Clayards and T. Knowles, "Prominence enhances voicelessness and not place distinction in English voiceless sibilants." in *ICPhS*, 2015.

[22] M. Yavas, *Applied English Phonology*. John Wiley & Sons, 2020.

[23] R. M. Voigt, "Labialization and the so-called sibilant anomaly in Tigrinya," *Bulletin of the School of Oriental and African Studies*, vol. 51, no. 3, pp. 525–536, 1988.

[24] P. Ladefoged, K. Johnson, and P. Ladefoged, *A course in phonetics*. Thomson Wadsworth Boston, MA, 2006, vol. 3.

[25] M. Proctor, C. Shadle, and K. Iskarous, "An MRI study of vocalic context effects and lip rounding in the production of English sibilants," in *Proceedings of the 11th Australasian International Conference on Speech Science and Technology*, 2006, pp. 307–312.

[26] M. Toda, S. Maeda, A. J. Carlen, and L. Meftahi, "Lip protrusion/rounding dissociation in French and English consonants:/w/ vs. /ʃ/ and /ʒ/," in *Proccedings of the 15-th International Congress of Phonetic Sciences*, 2003, pp. 1763–1766.

[27] A. Faber, "Lip protrusion in sibilant production," *The Journal of the Acoustical Society of America*, vol. 86, no. S1, pp. S113–S113, 1989.

[28] A. Bonneau, J. Busset, and B. Wrobel-Dautcourt, "Contextual effects on protrusion and lip opening for/i, y/," in *10th Annual Conference of the International Speech Communication Association-Interspeech*, 2009.

[29] K. Nilsson, "The development of Sibilants in Swedish," *Phonetica*, vol. 13, no. 3, pp. 177–183, 1965.

[30] H.-S. Kang, "Korean Sibilant/s/before a High Front and a Round Segment," *Phonetics and Speech Sciences*, vol. 2, no. 4, pp. 59–65, 2010.

[31] R. K. Shosted, "Articulatory and acoustic characteristics of whistled fricatives in Changana," in *Selected Proceedings of the 40th Annual Conference on African Linguistics: African Languages and Linguistics Today*. Cascadilla Somerville, MA, 2011, pp. 119–29.

[32] N. Zharkova, "Ultrasound and acoustic analysis of sibilant fricatives in preadolescents and adults," *The Journal of the Acoustical Society of America*, vol. 139, no. 5, pp. 2342–2351, 2016.

[33] Y. Lim, A. Toutios, Y. Bliesener, Y. Tian, S. G. Lingala, C. Vaz, T. Sorensen, M. Oh, S. Harper, W. Chen *et al.*, "A multispeaker dataset of raw and reconstructed speech production real-time MRI video and 3D volumetric images," *Scientific data*, vol. 8, no. 1, p. 187, 2021.

[34] F. Y. Chong, "The acoustic and perceptual effects of single-microphone noise reduction in hearing aids on Mandarin fricatives and affricates," Ph.D. dissertation, University of British Columbia, 2016.

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

[36] R. Mannem and P. K. Ghosh, "A deep neural network based correction scheme for improved air-tissue boundary prediction in real-time magnetic resonance imaging video," *Computer Speech & Language*, vol. 66, p. 101160, 2021.

[37] N. Zharkova, W. J. Hardcastle, and F. E. Gibbon, "The dynamics of voiceless sibilant fricative production in children between 7 and 13 years old: An ultrasound and acoustic study," *The Journal of the Acoustical Society of America*, vol. 144, no. 3, pp. 1454–1466, 2018.

[38] S. Schötz, J. Frid, and A. Löfqvist, "Development of speech motor control: Lip movement variability," *The Journal of the Acoustical Society of America*, vol. 133, no. 6, pp. 4210–4217, 2013.

[39] C. M. Bishop, *Neural networks for pattern recognition*. Oxford university press, 1995.

[40] V. Fromkin, "Lip positions in american english vowels," *Language and speech*, vol. 7, no. 4, pp. 215–225, 1964.