

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

Tilak Purohit, Prasanta Kumar Ghosh

SPIRE LAB, Electrical Engineering,
Indian Institute of Science (IISc), Bangalore, India



INTERSPEECH 2020



Table of Contents

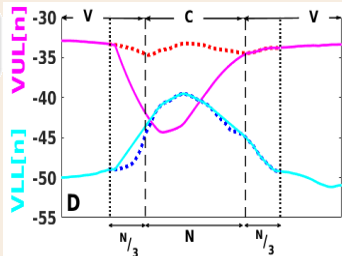
- 1 Introduction**
- 2 SPIRE-VCV Dataset
- 3 Key Research Questions
- 4 Results
- 5 Conclusion



Virtual Lip Trajectory

UL: Upper Lip **VUL:** Virtual Upper Lip
LL: Lower Lip **VLL:** Virtual Lower Lip

- ▲ A **VUL (VLL)** is a hypothetical trajectory below (above) the measured **UL (LL)** trajectory which could have been achieved by **UL (LL)** if **UL & LL** were not in contact with each other during bilabial stops and nasal.

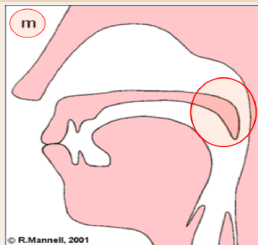
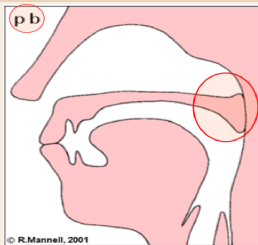


“A hypothetical trajectory where, to reach the virtual target, the lips would have to move beyond each other [1].”

[1] 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



Bilabial stops and nasal



- 🔥 Bilabial stop consonants (/p/, /b/) are produced when upper and lower lips come together creating a **complete closure** of the vocal tract causing an **occlusion** resulting in a block of airstream in the oral cavity, followed by a release of the blocked airstream (burst).
- 🔥 Bilabial nasal (/m/) is produced in a manner similar to the bilabial stop (/p/) but the block of airstream in the oral cavity caused due to complete closure of the vocal tract is **released via nose** instead.



Goals

- ▶ **Estimate the virtual lip trajectories** from the given UL & LL trajectories.
- ▶ Analyze the characteristics of virtual lip trajectories at **3 different speaking rates - slow, normal and fast.**



Motivation

- ▶ Could virtual lips reveal **specific articulatory planning** during bilabial stops & nasal production?
- ▶ Analyse the **lip displacement pattern**.
- ▶ Could the features derived via virtual lip trajectories help in **discriminating speech rates**?



Table of Contents

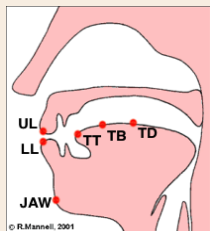
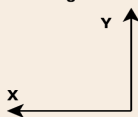
- 1 Introduction
- 2 SPIRE-VCV Dataset**
- 3 Key Research Questions
- 4 Results
- 5 Conclusion

Acoustic Articulatory Recording

- Articulatory movement data recorder: 3D Electromagnetic Articulograph (EMA) AG501.
- Microphone used: Audio-Technica ATM10a.
- Six sensors were connected to the subject to obtain twelve articulatory trajectories.
- Ten subjects (5 male + 5 female) in an age range of 18-22 years.
- All subjects were non-native speakers of English.



Upper Lip : UL
 Lower Lip : LL
 Jaw : JAW
 Tongue Tip : TT
 Tongue Body : TB
 Tongue Dorsum : TD





Dataset Details

- Utterance was of the format - **“Speak VCV Today”**.
- Each utterance was repeated thrice in each of the **three different speaking rates**, namely **slow**, **normal** and **fast**. Total nine utterances for one VCV sequence.
- 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/**. **Total of 450** (= 3-repetitions × 5-vowels × 3-rates × 10-subjects) recordings for every consonant.

Consonant (C)	Slow (sec)	Normal (sec)	Fast (sec)
/p/	0.22 ± 0.08	0.13 ± 0.04	0.08 ± 0.02
/b/	0.14 ± 0.05	0.09 ± 0.02	0.06 ± 0.01
/m/	0.20 ± 0.15	0.10 ± 0.03	0.06 ± 0.01



Data-processing and Annotation

- ▶ Recordings of the lip movements were done at a sampling rate of 250Hz in the midsagittal plane, i.e., UL_x, UL_y, LL_x, LL_y are used for the analysis in this work.
- ▶ The audio was recorded at 44.1KHz and later down-sampled to 16KHz.
- ▶ The VCV boundaries were manually marked by a team of four annotators.
- ▶ The boundaries were marked by observing the spectrogram, the raw waveform and the glottal pulses (obtained using Praat) simultaneously using a MATLAB based in-house annotation tool.

VCV Annotation Tool

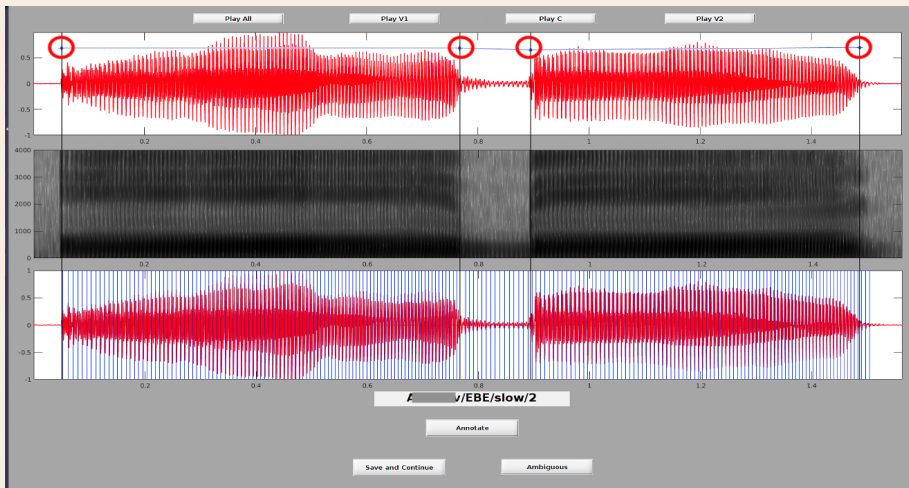




Table of Contents

- 1 Introduction
- 2 SPIRE-VCV Dataset
- 3 Key Research Questions**
- 4 Results
- 5 Conclusion

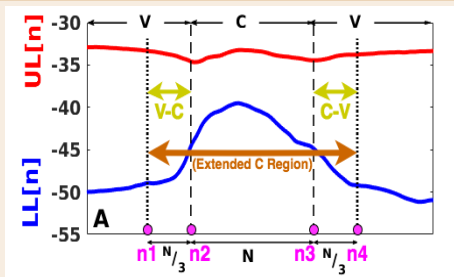


Objective

- ▶ Estimating the virtual upper lip (VUL) and virtual lower lip (VLL) trajectories.
- ▶ Deriving the representations/features from Virtual lip trajectories.
- ▶ Statistical analysis of the virtual lip trajectory features across 3 different speaking rates (slow, normal, fast).
- ▶ Carry out a classification experiment to check whether the virtual lip features help in the speech rate discrimination.



Methodology



Assumptions:

- ▲ The $UL[n]$ and $LL[n]$ are approximately related by an affine function: $UL[n] \approx \alpha_n LL[n] + \beta_n$.
- ▲ Relation between the VUL and VLL in the entire **extended C region** is similar to that between UL and LL in the **transition region**.
- ▲ The α_n and β_n in the case of VUL and VLL vary linearly from **V-C transition region** to **C-V transition region**.



Methodology (Cont.)

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:

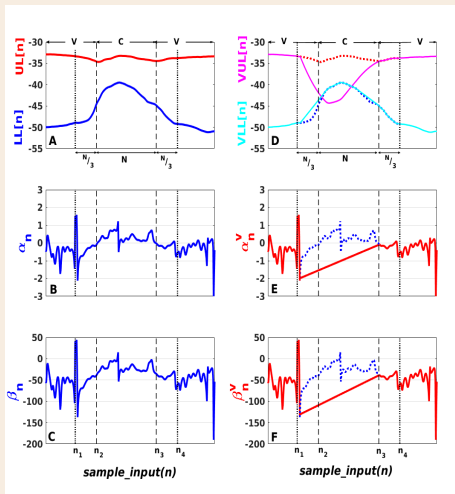
$$\alpha_n^v = \alpha_{m_1} + (n - m_1) \frac{\alpha_{m_2} - \alpha_{m_1}}{m_2 - m_1}, \quad (1)$$

$$\forall m_1 \leq n \leq m_2$$

$$\beta_n^v = \beta_{m_1} + (n - m_1) \frac{\beta_{m_2} - \beta_{m_1}}{m_2 - m_1}, \quad (2)$$

$$\forall m_1 \leq n \leq m_2$$

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 $n > m_2$.





Methodology (Cont.)

- 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:

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

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

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

and $x_{m_1} = LL[m_1]$, $x_{m_2} = LL[m_2]$ (3)

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 3. The best choices of m_1 and m_2 are selected by running the optimization (eq 3) for all possible combinations of m_1 and m_2 and selecting the one which results in the least objective function value.



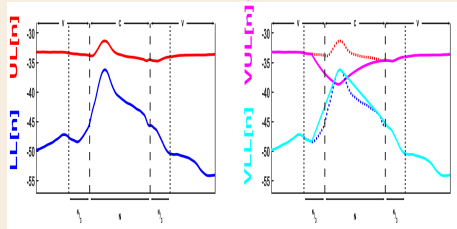
Table of Contents

- 1 Introduction
- 2 SPIRE-VCV Dataset
- 3 Key Research Questions
- 4 Results**
- 5 Conclusion

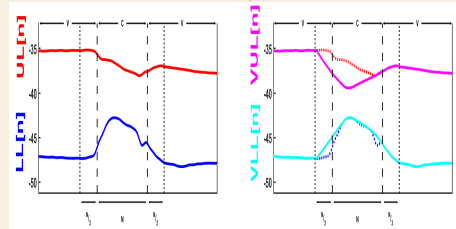


Samples of Virtual UL & LL Trajectories

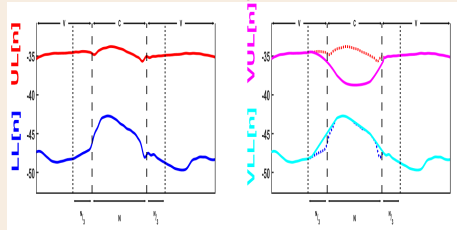
Example: 1



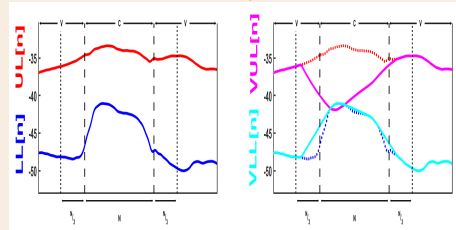
Example: 2



Example: 3



Example: 4



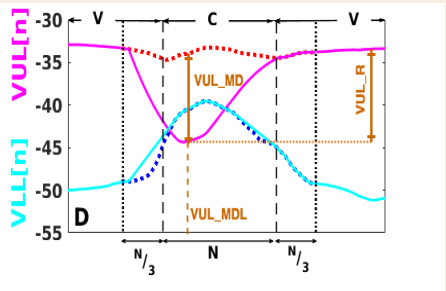
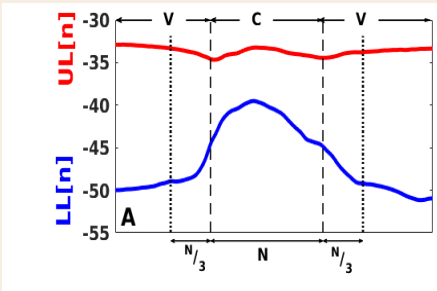


Objective

- ▶ ~~Estimating the virtual upper lip (VUL) and virtual lower lip (VLL) trajectories.~~
- ▶ Deriving the representations/features from Virtual lip trajectories.
- ▶ Statistical analysis of the virtual lip trajectory features across 3 different speaking rates (slow, normal, fast).
- ▶ Carry out a classification experiment to check weather the virtual lip features helps in the speech rate discrimination.



Virtual Lip trajectory features



- **VUL_MD**: Maximum deviation between UL & VUL.
 - **VUL_MDL**: Location of maximum deviation between UL & VUL.
 - **VUL_R**: Range of VUL.
- (Similarly for VLL, features derived are: VLL_MD, VLL_MDL, VLL_R).

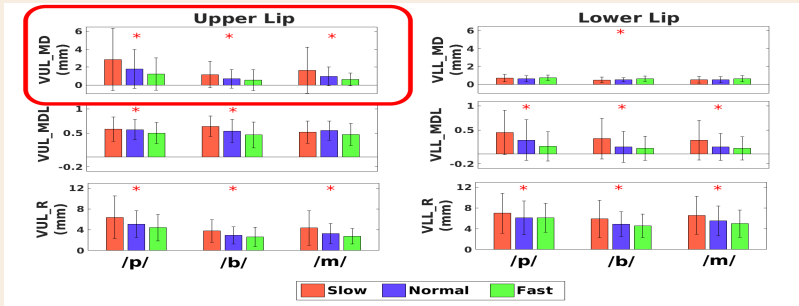


Objective

- ▶ Estimating the virtual upper lip (VUL) and virtual lower lip (VLL) trajectories.
- ▶ Deriving the representations/features from Virtual lip trajectories.
- ▶ Statistical analysis of the virtual lip trajectory features across 3 different speaking rates (slow, normal, fast).
- ▶ Carry out a classification experiment to check weather the virtual lip features help in the speech rate discrimination.

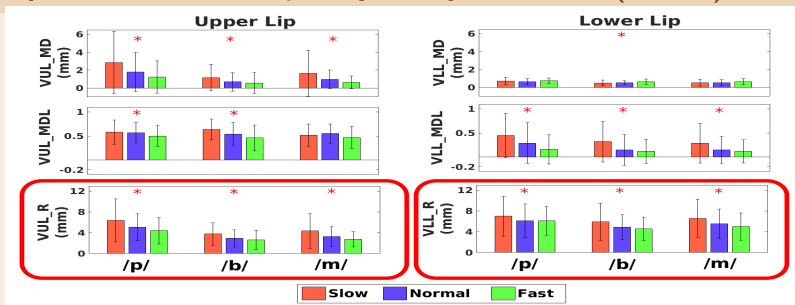




Analysis of Virtual Lip trajectory features



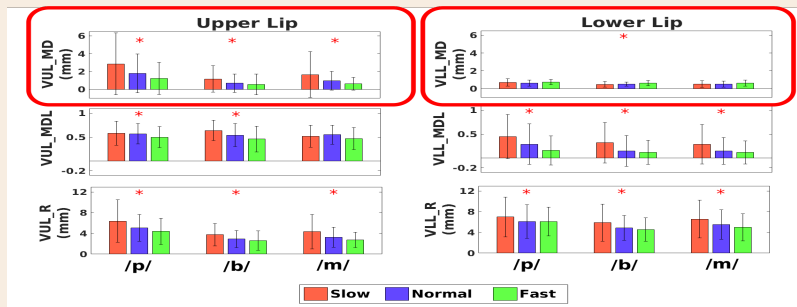
🔥 A significant drop in VUL_MD value from slow to fast suggests that in the fast rate the measured UL trajectory is more close to the VUL trajectory compared to that in the slow rate.

Analysis of Virtual Lip trajectory features (Cont.)



-  A significant drop from slow to fast in VUL_R and VLL_R, suggests that as the speaking increases, the range of the virtual motion of the UL and the LL decreases.
-  In spite of the increased velocity in fast case, VUL_R reduces compared to its slow rate counterpart.

Analysis of Virtual Lip trajectory features (Cont.)



🔥 In the slow rate, the VUL_MD is significantly higher than VLL_MD for /p/, /b/ and /m/. This is true for normal but not for fast rate.



Objective

- ▶ Estimating the virtual upper lip (VUL) and virtual lower lip (VLL) trajectories.
- ▶ Deriving the representations/features from Virtual lip trajectories.
- ▶ Statistical analysis of the virtual lip trajectory features across 3 different speaking rates (slow, normal, fast).
- ▶ Carry out a classification experiment to check weather the virtual lip features help in the speech rate discrimination.



Experiment details and Baseline

- ▶ The derived virtual lip parameters, separately as well as their various combinations from all 10 subjects are pooled and used as the features in order to carry out speaking rate classification task.
- ▶ The SVM classifier with radial basis kernel has been trained and the classification is carried out in a ten fold cross validation setup.
- ▶ 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.
- ▶ F1-score is used as an evaluation metric to compare merits of different features.



Classification Results

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)

- ▲ The values show that in case of all bilabial stops, the **MD features perform the best** followed by Range followed by the MDL features.
- ▲ Combining features **improves F1-score** over their respective individual cases.
- ▲ These results suggest that the virtual lip features **provide better discrimination** between slow and fast rates compared to the baseline features.



Table of Contents

- 1 Introduction
- 2 SPIRE-VCV Dataset
- 3 Key Research Questions
- 4 Results
- 5 Conclusion**



Key Findings / Takeaways

- Virtual lip trajectories during bilabial stop computed by the proposed approach in this work are found to **significantly vary across speaking rates**.
- The representations derived from Virtual Lips 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.



Future Works

- ▶ The motion of LL is partly contributed by the jaw movement. Thus, normalizing the LL movement by **removing the effect of jaw** may provide insight into the nature of motor control for the lip motion.
- ▶ We would also like to explore ways of **relaxing the linear variation** of α_n and β_n for the formulation of the optimization problem.
- ▶ We would like to study the **vowel specific trends** for the same set of experiments.

Acknowledgement



- 🔥 We would like to thank **Shankar Narayanan** for his help with generating figures and the **Department of Science & Technology (DST), Govt. of India** for their support !!

THANK YOU

For queries write to us at: spirelab.ee@iisc.ac.in

