INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Speaker-independent Speech Inversion for Estimation of Nasalance

Yashish M. Siriwardena[1], *Carol Espy-Wilson*[1], *Suzanne Boyce*[2], *Mark K.Tiede*[3], *Liran Oren*[2]

[1]University of Maryland College Park, Maryland, USA
[2]University of Cincinnati, Ohio, USA
[3]Haskins Laboratories, Connecticut, USA

yashish@umd.edu, espy@umd.edu, boycese@ucmail.uc.edu, mark.tiede@yale.edu, orenl@ucmail.uc.edu

Abstract

The velopharyngeal (VP) valve regulates the opening between the nasal and oral cavities. This valve opens and closes through a coordinated motion of the velum and pharyngeal walls. Nasalance is an objective measure derived from the oral and nasal acoustic signals that correlate with nasality. In this work, we evaluate the degree to which the nasalance measure reflects fine-grained patterns of VP movement by comparison with simultaneously collected direct measures of VP opening using high-speed nasopharyngoscopy (HSN). We show that nasalance is significantly correlated with the HSN signal, and that both match expected patterns of nasality. We then train a temporal convolution-based speech inversion system in a speaker-independent fashion to estimate VP movement for nasality, using nasalance as the ground truth. In further experiments, we also show the importance of incorporating source features (from glottal activity) to improve nasality prediction.
Index Terms: speech inversion, nasalance, source features, high-speed nasopharyngoscopy

1. Introduction

Speech is produced by the coordinated movement of articulators such as tongue, velum, and lips that shape the acoustic signal produced by the larynx, forming alternations of vocal tract constriction (for consonants) and opening (for vowels) [1]. These movement patterns can differ according to the language, dialect, abilities, and habits of the speaker, but the fact that the movements themselves overlap in time means that the evidence of their movement in the acoustic signal can be compressed, scattered across time, and sometimes obscured by co-occurring events. The result is that many linguistic phenomena that are hard to express in acoustic terms are more readily explained by differences in the timing and degree of vocal tract constriction [2, 3]. Systems that do speech inversion rely on ground truth articulatory variables; by using extracted acoustic features such as Mel Frequency Cepstral Coefficients (MFCCs), Mel-spectrograms, or the waveform itself as the input speech representation, the system can learn a mapping to the articulatory variables. However, none of the publicly available articulatory speech corpora have direct articulatory level data capturing the velar and glottal constrictions [4, 5]. Therefore, most of the available SI systems (trained on these datasets) are limited to estimating the articulatory level information pertaining to lip and tongue constrictions [6, 7, 8, 9, 10].

Acoustic-to-articulatory speech inversion inspired by Articulatory Phonology [11] maps the acoustic speech signal to the kinematic state of each constriction synergy (lips, tongue tip, tongue body, velum, and glottis) by its corresponding constriction degree and location coordinates, which are called vocal tract variables (TVs). In this work, we extend a speech inversion system based on TVs to estimate the activity of the velar constriction by collecting a dataset that can be effectively used in training a speaker-independent SI system. We choose 'Nasalance' as the ground-truth to capture nasality for two reasons. First, it is a non-invasive measure and can be easily collected from a larger population, which will be beneficial in building a more generalizable, speaker-independent SI system. However, nasalance measures the ratio of acoustical energy between the nasal and oral tract. Accordingly, as a variable it is dependent on the amount of energy flowing through the glottis, and thus has only an indirect relationship with VP articulation [12, 13]. Hence, the far reaching goal of the proposed SI system is not aimed at deriving aerodynamic relationships (such as nasalance) from the acoustic signal, but rather aimed at deriving VP articulatory movements. Our approach to achieve this goal was twofold. To investigate if nasalance is an accurate representation of velar constriction degree [11], we validated it with a more direct, invasive and accurate measure of VP activity called high-speed nasopharyngoscopy (HSN). To the best of our knowledge, this is the first time a SI system has been developed to estimate a proxy for a velar constriction degree TV, that will, in essence, capture the nasality in speech.

The second reason for using nasalance derives from this susceptibility to glottal source effects. Learning a mapping from an acoustic representation that is rich with source level information (eg. Melspectrograms, auditory spectrograms) to nasalance, along with source features (eg. voicing and pitch) may positively influence the SI system performance for nasality prediction. To investigate such effects of using source features, Electroglottography (EGG) was synchronously collected to extract a voicing parameter, and aperiodicity, periodicity and pitch extracted from an aperiodicity, periodicity and pitch (APP) detector [14] are also used as additional targets to further improve nasality prediction.

The content in the following sections of the paper is organized as follows. In section 2, we discuss the details of the dataset and explain the steps used to extract and validate the ground-truth nasalance parameter. In section 3, we highlight the details of the proposed SI system and the importance of using source features to estimate nasality. Finally in section 4, we discuss the key conclusions drawn from the experiments and possible future directions.

2. Dataset

This work is based on a subset of data from ongoing, collaborative data collection. The complete dataset, once collected, will be made public (subject to standard open source licensing agreements). One of the main goals of this dataset is to develop a speaker-independent speech inversion system to accurately estimate velar and glottal activity. The current dataset

10.21437/Interspeech.2023-2352

has been collected from 8 subjects (5 Female, 3 Male), and the demographic details of the speakers are listed in Table 1.

2.1. Ground-truth Nasalance Parameter

2.1.1. Background and Procedure for Data Collection

Table 1: *Dataset Description. SW: South-west, C: Central, W: White, B: Black, H: Hispanic, NH : Non-Hispanic*

Subject	Gender	Language	HSN status	Age (years)	Ethnicity/Race
1	M	English(SWOhio)	HSN	28	W, NH
2	F	English(STexas)	No HSN	24	W, H
3	F	English(SWOhio)	No HSN	31	W, NH
4	F	English(SWOhio)	No HSN	40	W, NH
5	F	English(CKentucky)	No HSN	28	B, NH
6	F	English(SWOhio)	HSN	34	W, NH
7	M	English(SWOhio)	No HSN	23	W, NH
8	M	English(SWOhio)	No HSN	35	W, NH

As noted above, nasalance is the relative proportion of nasal vs. oral acoustic output from two microphones (mic) mounted to the top and bottom of a separation plate located between the nose and upper lip to create an acoustic barrier. It is a simple, well-known, non-invasive and reliable technology for tracking VP constriction. We used a subset of speakers (subject 1 and subject 6) to synchronously collect a more direct but invasive measure of VP constriction using high-speed nasopharyngoscopy (HSN).

Figure 1 shows the setup used to collect the HSN and audio measurements to compute the nasalance parameter. For the HSN, a flexible scope (outer diameter: 2.2 or 3.6 mm) was connected to a video camera (MIRO 310; Vision Research, Inc., Wayne, New Jersey), and the images were captured at a rate of 1000 frames/second using 304 $\times$ 256 pixel resolution. To collect the audio data, 2 microphones (1/4", Type 4958, Bruel and Kjær, Duluth, Georgia) were connected to the top and the bottom of the separation plate made of aluminum. Windscreens were used to cover the microphones to prevent interference from airflow directed toward the microphones. The separation plate was placed against the participant's upper lip to create an acoustic barrier between the oral and nasal audio recordings. The acoustic data from the microphones were captured at 51.2 kHz using a data acquisition system (NI 9234, National Instruments, Austin, Texas) and customized LabVIEW code that digitized and converted the data to a ".wav" audio file. The initiation of the audio recording and imaging data (from the HSV nasopharyngoscopy) was synchronized using an input/output module (NI 9402; National Instruments) [15]

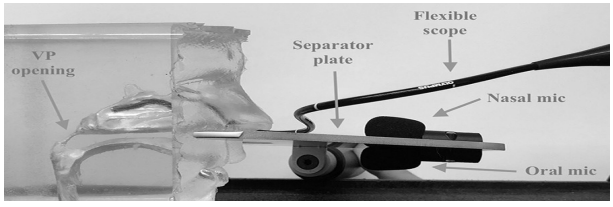

Figure 1: *Illustration of the experimental setup. HSN measurements were taken by connecting a flexible scope to a high-speed video camera (not shown). The figure is taken from [15] in The Cleft Palate-Craniofacial.*

Using this setup, approximately 10 minutes of speech material per subject was recorded. This consisted of a mixture of short and long sentences and short paragraphs. For example, for nasality, the full set of prosodic nasal contrasts from Krakow et al. [16] was included, including e.g. "hoe me" vs. "home E", "seam ore" vs. "Seymour". For voicing, sentences contrasting words such as "Dodd" vs. "Todd" in a carrier phase were included. Sentences illustrating consonant cluster articulatory patterns were drawn from Zsiga et al. [17, 18]. For cross-dataset comparison, we also included some sentences from speech materials used in the U.W. x-ray microbeam corpus [4].

2.1.2. Nasalance Parameter

Oral and nasal mic signals collected from the nasometer setup were used to compute the nasalance parameter. The baseline wander was first removed from the two signals using a high pass filter (cutoff around 0.1Hz). The Root Mean Square (RMS) signals were then computed for both oral and nasal signals separately. During the RMS signal generation, both the squared signals were smoothed out using a moving average filter with a window size of 1000 ($\sim$ 20 ms) samples. Then a nasalance parameter (Nasalance$_{raw}$) was computed using the equation 1 based on Bunton et al. [19]. The Nasalance$_{raw}$ parameter was then downsampled to 100Hz and smoothed using a window of 10 samples (using Matlab function 'Fastsmooth' by [20]). The final nasalance parameter was then normalized to [-1,1] range to be used as the ground-truth for the speech inversion system.

$$Nasalance_{raw} = \frac{RMS_{nasal}}{RMS_{nasal} + RMS_{oral}} \quad (1)$$

2.2. Validating Nasalance with HSN

HSN was synchronously collected from subject 1 and subject 6 to assess the accuracy and agreement with the computed nasalance parameter. Here the temporal dynamics of the VP port is captured by summing the light intensity in the images (intensity of pixels) of the high-speed video (HSV) data. The resulting intensity trace has been shown to be an accurate measure for capturing the velum [15]. Figure 2 shows a sample HSV intensity trace and the corresponding HSV images at different points in time. Here an open VP port would be overall characterized by darker regions that come from the cavity of the VP port. On the other hand, a closed VP port would be characterized with brighter regions because of the increased amount of light reflecting off the tissue. It should be noted that the HSN parameter shows a trough (i.e. lower values) for nasal sounds in speech. This is in contrast to nasalance, which shows the opposite pattern of a peak.

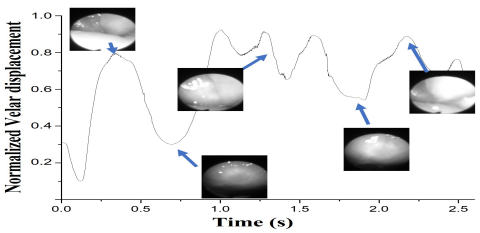

Figure 2: *HSV intensity trace for a male native speaker of American English from Cincinnati, OH producing "It's a see more, Sid. It's a seam ore, Sid". Images of the VP port at key time points are indicated by arrows.*

The HSV data has a sampling rate of 1kHz and the nasalance parameter as discussed earlier is sampled at 100Hz. To match the number of samples to compute the cross correlations, the nasalance parameter is linearly interpolated to match with the HSV intensity trace. The Pearson correlation coefficients are then computed for each sample data from the subject. The average correlation coefficients across the samples for subject 1 and subject 6 are -0.6081($p<0.001$) and -

0.5136($p<0.001$) respectively. These statistically significant negative correlations give an important validation on the accuracy of the computed nasalance parameter with respect to HSN.

2.3. Patterns of timing for Nasality

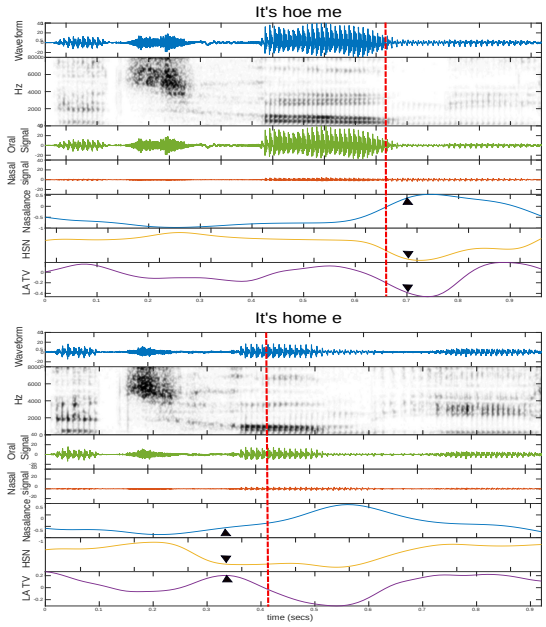

Figure 3: *The vertical red dash lines in the top and bottom panels mark the onset of bilabial contact for the /m/. The black triangles mark velum lowering offset and the coordinated event in the lower lip (lip raising onset or offset)*

A number of studies have shown that American English shows different patterns of velum raising and lowering (i.e. VP port constriction) according to syllabic organization [21]. As shown in Krakow et al. [21] an example of this pattern for "home E" vs "hoe me" is that the velum moves earlier and the VP port stays open longer when the /m/ is in the rime (home) than when the /m/ is in the onset of the following word (me). The lip-velum coordination during the syllable-initial and -final nasal was also observed in Krakow et al. [21], where it has been noted that there is close temporal proximity between the end of velum lowering and the beginning of lip raising for the syllable-initial and a large offset between the end of velum lowering and the end of lip raising for syllable-final.

To see if the nasalance parameter will also showcase such patterns (word-initial vs word-final /m/) with respect to the HSN and lip movement, the words 'hoe me' and 'home e' were analyzed. Figure 3 shows the data for 'It's hoe me' and 'It's home e' collected from subject 1 in the dataset. To analyze the lip movement pattern, the lip aperture tract variable (LA TV) was extracted from the articulatory speech inversion system in [9]. Both the HSN and nasalance patterns shown in Figure 3 replicate the timing patterns described in Krakow et al.[21] with respect to the LA TV. Data from a larger group of subjects is needed to further verify the pattern.

2.4. Voicing parameter: EGG envelope

Electroglottography (EGG) is a well-established technology for tracking vocal fold oscillation, using the degree of electrical conductance across the glottal gap between electrodes placed on the two parallel outer sides of the throat. In this study, EGG data was also collected (from all the subjects) synchronously with the other HSN and audio measurements in section 2.1.1.

The EGG signal is sampled at 51.2 KHz, and to compute a parameter which can capture the voicing activity of speech, the envelope of the EGG signal was extracted. As with the nasalance parameter, we first high pass filtered the signal to remove the baseline wander. Then the magnitude of the Hilbert transform [22] was computed as the envelope of the EGG signal. The envelope was downsampled to 100 Hz and smoothed and normalized the same way to the nasalance parameter to generate the final voicing parameter.

3. Speech Inversion System

3.1. Input Audio Representation

The audio recorded by the oral and nasal mic signals were mixed together to create a combined audio signal. The combined signal was then downsampled to 16kHz and segmented to 2 second long segments. The shorter, remaining segments were zero padded at the end. The segmentation was done mainly to increase the number of audio samples to train the DNN based SI system and to have input acoustic representations of fixed dimensionality to the input layer of the DNN model.

We used auditory spectrograms (Audspec) [23] as the input speech representation for the SI system. The auditory spectrograms have a logarithmic frequency scale and provide a unified multi-resolution representation of the spectral and temporal features likely critical in the perception of sound [23].

3.2. Model Architecture and Training

3.2.1. Model Architecture

We developed a Temporal Convolution Network (TCN) based SI system inspired by the work in [9]. The model was optimized using the Mean Squared Error (MSE) loss computed between the predicted parameters and the ground truth. The SI system was implemented in PyTorch with 1-D convolutional (CNN) layers. Figure 4 shows the proposed model architecture with its sub-modules used for pre-processing and dilated TCN. The pre-processing module contains two 1-D CNN layers with 1×1 kernels (C1, and C2), which have 128 filters each. The d1, d2 and d3 dilated CNN layers have a kernel size of 3 with 1,4 and 16 dilation rates respectively. Upsampling (window size 4) was done after C4 layer and average pooling (window size 5) was done after C5 layer along with BatchNorm layers after every CNN layer in the TCN network. The upsampling and average pooling operations take care of matching the time dimension of the input spectrograms to the target time dimension of TVs.

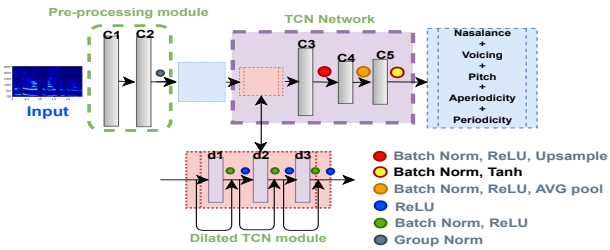

Figure 4: *Model architecture (SI system). Here C1-C5 represent 1D-CNN layers and d1-d3 represent 1D dilated CNN layers*

3.2.2. Model Training

All the model parameters were randomly initialized with a seed (=7) for reproducibility. Table 3 lists the hyper-parameters and the corresponding values considered to fine-tune the model.

Table 2: *PPMC scores (mean and .std across 8 trials) for the SI systems trained with and without source features as additional targets to estimate nasalance.*

	Nasalance	**Voicing**	**Perio.**	**Aperio.**	**Pitch**	**Average**
SI-SF	**0.7341(0.02)**	0.80541(0.01)	0.9008(0.03)	0.8257(0.02)	0.7995(0.03)	0.8131(0.03)
SI-noSF	0.6967(0.02)	-	-	-	-	-

Table 3: *Hyperparameter Tuning for the TCN model*

Parameter	**Possible Values**	**Chosen Values**
Learning Rate	[1e-4, 3e-4, 1e-3, 1e-2]	1e-3
Batch size	[16,32,64,128]	64
Optimizer	ADAM, RMSprop, SGD	ADAM
Rate scheduler	ExponentialLR, PolynomialLR	ExponentialLR

A grid search was performed when fine-tuning the hyper-parameters and the best parameters were chosen based on the validation loss. All the models were implemented with PyTorch machine learning framework and trained with NVIDIA TITAN X GPUs. The best performing model has around 1 million trainable parameters, takes around 8 minutes (± 2) to converge, and can be found in a Github repository[1]

The dataset was divided into training, validation and testing splits, so that the training set has utterances from 6 speakers (4 females, 2 males). The validation and testing splits have data from 2 speakers (1 male, 1 female) with 1/2 of the data from each speaker in the validation split and the other half in the test split. None of the data from the speakers in the validation and test splits were included in the training split and hence all the models are trained in a 'speaker-independent' fashion. The splits also ensured that around 70% of the total number of utterances were present in training (1 hour of speech), and all the allocations were done in a completely random manner.

3.3. Results of Speaker-independent Speech Inversion

Two speech inversion systems were trained to estimate the nasalance parameter from the input auditory spectrograms. Pearson Product Moment Correlation (PPMC) score is used as the metric to evaluate the predictions by the SI systems. Table 2 shows the PPMC scores for correlations between the estimated and ground-truth nasalance parameter for the systems trained with additional source features as targets (SI-SF) and the one with nasalance parameter as the only target (SI-noSF).

Figure 5 shows sample nasalance estimation by the SI-SF and SI-noSF models for an utterance in the test set. The utterance, 'Say tube again' contains a nasal consonant [n] around 1.15-1.25 seconds which is captured by both the SI systems. However, it is important to note that the nasalance parameter estimated by the SI-SF model has better agreement with the ground-truth compared to the SI-noSF model.

4. Discussion and Conclusion

The results of correlation analysis in the section 2.2 gives a general, but an important validation for the nasalance parameter with respect to the more direct HSV intensity trace. The fact that we found known patterns of timing for nasality (discussed in section 2.3) further supports the validity of using nasalance as a proxy variable for velopharyngeal constriction. This work highlights the performance of our SI system in estimating velopharyngeal movement dynamics for unseen speaker data. It also shows that incorporating source features as additional targets improves the estimation accuracy of the velopharyngeal movement parameter. This is consistent with the observations made in [9] with conventional acoustic-to-articulatory speech inversion, and could also suggest that the TCN model is particularly sensitive to source/VP interactions.

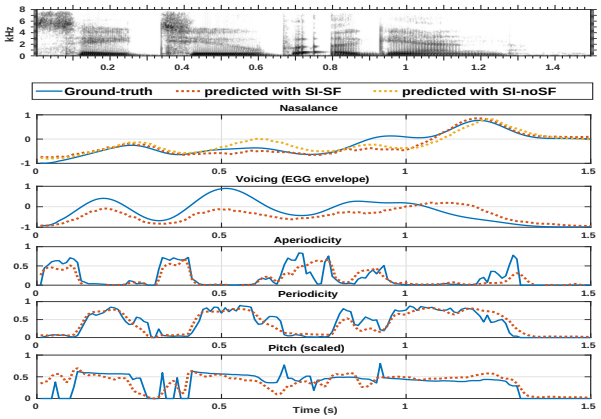

Figure 5: *Nasalance and source features for the utterance 'Say tube again' estimated by the SI-SF model and nasalance estimated by the SI-noSF model with respect to the ground-truth . Solid blue Line - ground truth, red dotted line - predictions by the SI-SF, yellow dotted Line - predictions by SI-noSF.*

In future work, the authors plan to improve the performance and generalizability of the current SI system by training on data from a larger group of subjects (from the ongoing data collection). More emphasis will also be made on validating and fine tuning the nasalance parameter as a proxy to the velar TV. Further experiments will also be done to understand what the DNN models are actually picking as source-filter interactions that are ultimately helping the overall SI task.

To summarize, in this work we present the details on a dataset collected to estimate the velar and glottal activity in speech. We particularly looked into estimating a validated nasalance parameter (as a proxy to a velar TV) using a speaker-independent SI system. It should be noted, that having a SI system to estimate parameters directly related to the velar (and glottal) constrictions can be hugely beneficial, since it gives an almost complete articulatory level representation of speech which can be useful in diverse speech applications (eg. articulatory speech synthesis [24, 25]). An accurate, validated speech inversion system would also be a significant breakthrough for researchers with little or no ability to collect articulatory data directly, e.g. scholars without well-equipped phonetics laboratories, scholars doing field studies in dispersed communities. While speech inversion data is not equivalent to direct observation, it may enable hypothesis formation and testing that will motivate more targeted studies.

5. Acknowledgement

This work was supported by the NSF grant BCS2141413

[1]https://github.com/Yashish92/TCN-SI-tool-Nasality

6. References

[1] K. N. Stevens, *Acoustic phonetics*. MIT press, 2000, vol. 30.

[2] T. Cho and P. Keating, "Effects of initial position versus prominence in english," *Journal of Phonetics*, vol. 37, no. 4, pp. 466–485, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0095447009000497

[3] J. Krivokapić, "Gestural coordination at prosodic boundaries and its role for prosodic structure and speech planning processes," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 369, no. 1658, p. 20130397, 2014. [Online]. Available: https://royalsocietypublishing.org/doi/abs/10.1098/rstb.2013.0397

[4] J. R. Westbury, "Speech Production Database User ' S Handbook," *IEEE Personal Communications - IEEE Pers. Commun.*, vol. 0, no. June, 1994.

[5] M. Tiede, C. Y. Espy-Wilson, D. Goldenberg, V. Mitra, H. Nam, and G. Sivaraman, "Quantifying kinematic aspects of reduction in a contrasting rate production task," *The Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. 3580–3580, 2017. [Online]. Available: https://doi.org/10.1121/1.4987629

[6] A. Illa and P. K. Ghosh, "Low Resource Acoustic-to-articulatory Inversion Using Bi-directional Long Short Term Memory," in *Proc. Interspeech 2018*, 2018, pp. 3122–3126.

[7] Y. M. Siriwardena, A. A. Attia, G. Sivaraman, and C. Espy-Wilson, "Audio data augmentation for acoustic-to-articulatory speech inversion using bidirectional gated rnns." arXiv 2022. [Online]. Available: https://arxiv.org/abs/2205.13086

[8] A. S. Shahrebabaki, S. M. Siniscalchi, G. Salvi, and T. Svendsen, "Sequence-to-Sequence Articulatory Inversion Through Time Convolution of Sub-Band Frequency Signals," in *Proc. Interspeech 2020*, 2020, pp. 2882–2886. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1140

[9] Y. M. Siriwardena and C. Espy-Wilson, "The secret source : Incorporating source features to improve acoustic-to-articulatory speech inversion," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[10] S. Udupa, A. Roy, A. Singh, A. Illa, and P. K. Ghosh, "Estimating Articulatory Movements in Speech Production with Transformer Networks," in *Proc. Interspeech 2021*, 2021, pp. 1154–1158.

[11] C. P. Browman and L. Goldstein, "Articulatory Phonology : An Overview *," *Phonetica*, vol. 49, pp. 155–180, 1992.

[12] A. Kochetov, "Research methods in articulatory phonetics ii: Studying other gestures and recent trends," *Language and Linguistics Compass*, vol. 14, no. 6, p. e12371, 2020. [Online]. Available: https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/lnc3.12371

[13] P. Rong, R. Shosted, and C. Carignan, "The relationship between velopharyngeal opening and place of articulation: An aerodynamic and epg investigation," in *9th International Seminar on Speech Production (ISSP)*, 06 2011.

[14] O. Deshmukh, C. Y. Espy-Wilson, A. Salomon, and J. Singh, "Use of temporal information: detection of periodicity, aperiodicity, and pitch in speech," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 776–786, 2005.

[15] L. Oren, M. Rollins, S. Padakanti, A. Kummer, E. Gutmark, and S. Boyce, "Using high-speed nasopharyngoscopy to quantify the bubbling above the velopharyngeal valve in cases of nasal rustle," *The Cleft Palate-Craniofacial Journal*, vol. 57, no. 5, pp. 637–645, 2020, pMID: 31867995. [Online]. Available: https://doi.org/10.1177/1055665619894183

[16] R. A. Krakow, P. S. Beddor, L. M. Goldstein, and C. A. Fowler, "Coarticulatory influences on the perceived height of nasal vowels," *J Acoust Soc Am*, vol. 83, no. 3, pp. 1146–1158, Mar. 1988.

[17] E. C. Zsiga, "Acoustic evidence for gestural overlap in consonant sequences," *Journal of Phonetics*, vol. 22, no. 2, pp. 121–140, 1994. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0095447019301895

[18] E. C. Zsiga and R. Nitisaroj, "Perception of thai tones in citation form and connected speech," *The Journal of the Acoustical Society of America*, vol. 116, no. 4 Supplement, pp. 2628–2628, 10 2004. [Online]. Available: https://doi.org/10.1121/1.4785480

[19] K. Bunton and B. H. Story, "The relation of nasality and nasalance to nasal port area based on a computational model," *Cleft Palate Craniofac J*, vol. 49, no. 6, pp. 741–749, Oct. 2011.

[20] T. O'Haver, "Fast smoothing function," *MathWorks*, 2017. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/19998-fast-smoothing-function

[21] R. A. Krakow, "Physiological organization of syllables: a review," *Journal of Phonetics*, vol. 27, no. 1, pp. 23–54, 1999. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S009544709990089X

[22] M. Feldman, "Hilbert transforms," in *Encyclopedia of Vibration*, S. Braun, Ed. Oxford: Elsevier, 2001, pp. 642–648. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B0122270851000576

[23] K. Wang and S. Shamma, "Self-normalization and noise-robustness in early auditory representations," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 3, pp. 421–435, 1994.

[24] P. Wu, S. Watanabe, L. Goldstein, A. W. Black, and G. K. Anumanchipalli, "Deep Speech Synthesis from Articulatory Representations," in *Proc. Interspeech 2022*, 2022, pp. 779–783.

[25] Y. M. Siriwardena, C. Espy-Wilson, and S. Shamma, "Learning to compute the articulatory representations of speech with the mirrornet," 2022. [Online]. Available: https://arxiv.org/abs/2210.16454