



Enhancing Acoustic-to-Articulatory Speech Inversion by Incorporating Nasality

Saba Tabatabaee¹, Suzanne Boyce², Liran Oren³, Mark Tiede⁴, Carol Espy-Wilson¹

¹ Department of Electrical and Computer Engineering, University of Maryland College Park, USA

² Department of Communication Sciences and Disorders, University of Cincinnati, USA

³ Department of Otolaryngology-Head and Neck Surgery, University of Cincinnati, USA

⁴ Department of Psychiatry, Yale University, USA

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

Abstract

Speech is produced through the coordination of vocal tract constricting organs: lips, tongue, velum, and glottis. Previous works developed Speech Inversion (SI) systems to recover acoustic-to-articulatory mappings for lip and tongue constrictions, called oral tract variables (TVs), which were later enhanced by including source information (periodic and aperiodic energies, and F0 frequency) as proxies for glottal control. Comparison of the nasometric measures with high-speed nasopharyngoscopy showed that nasalance can serve as ground truth, and that an SI system trained with it reliably recovers velum movement patterns for American English speakers. Here, two SI training approaches are compared: baseline models that estimate oral TVs and nasalance independently, and a synergistic model that combines oral TVs and source features with nasalance. The synergistic model shows relative improvements of 5% in oral TVs estimation and 9% in nasalance estimation compared to the baseline models.

Index Terms: Speech Inversion, Tract Variables, Nasalance

1. Introduction

Speech articulation is a complex activity that requires finely timed coordination across articulators (lips, tongue, jaw, velum, and glottis) [1]. To recover this information from speech, a Speech Inversion (SI) system was developed that maps the speech signal not onto discrete articulator positions, but rather onto synergies of activity coordinated among articulators to achieve acoustic goals known as vocal Tract Variables (TVs) [2, 3]. The TVs are defined for the lips, tongue tip, tongue body, velum and glottis (see Figure 1 and Table 1 for more details). The kinematic state of each constrictor is defined by its corresponding degree of constriction and location coordinates, recovered as time-varying trajectories by the SI system.

Table 1: Constrictors and corresponding vocal tract variables.

Constrictor	Vocal Tract Variable
Lips	Lip Aperture (LA)
	Lip Protrusion (LP)
Tongue Tip	Tongue Tip Constriction Location (TTCL)
	Tongue Tip Constriction Degree (TTCD)
Tongue Body	Tongue Body Constriction Location (TBCL)
	Tongue Body Constriction Degree (TBCD)
Velum	Velum (VEL)
Glottis	Glottis (GLO)

The oral aspects of speech are predicted by oral TVs, which represent constrictions of the lips, tongue body, and tongue tip. Nasality is another crucial contrastive feature in speech

This work was supported by NSF Grant No. BCS2141413.

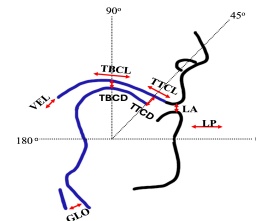


Figure 1: A visual representation of vocal tract variables (adapted from [4]).

production [5]. Nasality is controlled by constriction of the Velopharyngeal Port (VP), which regulates airflow and acoustic coupling between the nasal and oral cavities through the coordinated movement of the velum and pharyngeal walls. Direct observations of VP movement is invasive and costly because it requires trained specialists to administer. The measure nasalance, which derives from the proportional difference in acoustic energy emitted through the nose and mouth, is an alternative for indirect measure of VP movement. A recent study [6] validated the use of nasalance as ground truth for an SI system by demonstrating strong correspondence between nasalance and VP constriction degree as measured from high-speed nasopharyngoscopy images. Therefore, in this study, we refer to the estimates of the nasalance as VP TV, using it as a proxy for VP. Previous studies have developed SI systems to estimate either oral TVs [7, 8, 9, 10, 11, 12] or VP TV [6, 13, 14] separately, without leveraging their complementary information to create a comprehensive SI system. To bridge this gap, the present study introduces a novel SI system that simultaneously estimates oral TVs plus a VP TV from speech signals. To our knowledge, this is the first study to integrate these two critical speech features into a single, unified SI framework.

Recent advancements in Self-Supervised Learning (SSL) have shifted the focus for developing SI systems towards more robust feature extraction techniques, surpassing traditional acoustic features such as Mel-frequency cepstral coefficients (MFCCs). Studies in [6, 15, 16] demonstrate that SSL-based speech representations consistently outperform traditional acoustic features in SI tasks. For instance, one study [6] demonstrated that a BiGRNN-HuBERT system, leveraging HuBERT-Large SSL representations, outperforms a Temporal Convolutional Network (TCN) model using spectrogram inputs for VP TV estimation. The SSL-based system also exhibited superior generalization across various corpora, underscoring the robustness of SSL-derived features. Additionally, a comparison between SI systems using HuBERT-Large vs. MFCC-based methods showed that the former improves oral TVs estimation compared to the latter [16]. Further work [15] demonstrated that WavLM-Large outperforms HuBERT-Large and other SSL

models like Wav2Vec2 and TERA, as well as traditional features, in SI tasks. Building on these findings, in this work we utilize SSL representations, including HuBERT-Large and WavLM-Large, to capture richer speech features, thereby enhancing the overall performance of the SI system.

Information in the acoustic signal is also carried by continuously varying glottal Source Features (SF). Integration of SF such as the relative amounts of non-periodic energy (aperiodicity), periodic energy (periodicity), and fundamental frequency (F0) into a SI system was shown to improve estimation of oral TVs [17]. Recent work in [14] showed that incorporation of these SF into the SI model significantly enhanced VP TV estimation. Building on these findings, this study integrates SF into the SI system to improve oral TVs and VP TV estimation. **Our key contributions** in the current work are as follows:

- Improvement of VP TV estimation by developing a Nasal-SI system with a larger training dataset, based on work in [6].
- Comparative analysis of the Nasal-SI system using two SSL representation models, HuBERT-Large and WavLM-Large.
- Introduction of a novel comprehensive SI system capable of simultaneously estimating oral TVs, VP TV, and the SF.
- Comparison of multi-task learning and single-task learning approaches in developing the SI system.
- Providing an ablation study to analyze the effects of incorporating VP TV and the three SF in the SI system on recovery of oral TVs relative to ground truth.

2. Dataset Description and Pre-Processing

2.1. Nasometry-EGG dataset

In this study, we implemented a custom nasometry setup based on work [6], which computes nasalance as the relative acoustic energy from separate oral and nasal microphones mounted on a plate serving to isolate the two sources. Audio signals were recorded at a 51.2 kHz sampling rate. We collected data from 24 healthy adult speakers (20 native English speakers, 3 French speakers, 1 Sinhala speaker). Electrolottography (EGG) data using electrodes placed at the thyroid prominence were collected concurrently. Participants read sections from well-known passages, including the "Grandfather Passage" [18], Harvard sentences [19], and other materials from [20] and [21]. The total duration was 2.1 hours, and the data were split into training (20 speakers), development (2 speakers), and test (2 speakers) sets using a speaker-independent approach.

2.2. XRMB dataset

The University of Wisconsin XRMB dataset [21] contains recordings of naturally spoken isolated sentences and short passages from 32 male and 25 female participants, paired with point source trajectories of pellets attached to the tongue and lips obtained using a rasterized X-ray microbeam tracking system. After excluding mistracked data, the dataset comprises 46 native English speakers (21 males, 25 females) with approximately 4 hours of speech. Using methods described in [22], we reconstructed much of the corrupted data, increasing the total duration to around 5.3 hours. The data were split into a training set (36 speakers) and separate development and test sets (5 speakers each, with 3 males and 2 females per set) in a speaker-independent manner. Anatomical differences between speakers cause variability in pellet positions; therefore, in this study the original X-Y coordinates are normalized for vocal tract shape by conversion into TVs using a geometric transformation from

[16]. The processed XRMB dataset includes six oral TVs, sampled at 100 Hz and normalized to the range of -1 to 1. These oral TVs are: LA, LP, TBCL, TBCD, TTCL, and TTCD (see Figure 1 and Table 1 for more details).

3. Methodology

3.1. Developing Nasal-SI System

We developed the Nasal-SI system to estimate VP TV from speech signals by integrating four parameters, including the EGG envelope (EGG-env) and three SFs, which are aperiodicity (Ap), periodicity (Per), and F0, based on research in [14]. The ground truth values for the three SF were extracted using the APP detector [23]. Building on the approach in [6], the Nasal-SI system was developed using a Bidirectional Gated Recurrent Neural Network (BiGRNN) based model architecture, as illustrated in Figure 2.

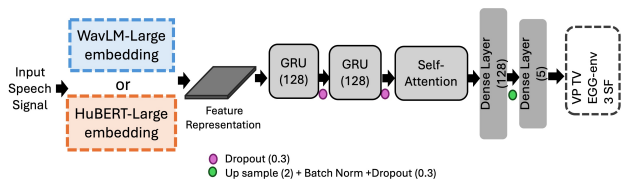


Figure 2: Proposed model architecture for the Nasal-SI system.

The system was initially developed using HuBERT-Large [24] from the SpeechBrain toolkit [25]. Subsequently, we developed the Nasal-SI system using WavLM-Large [26] embedding. Both embeddings were evaluated to assess their impact on the performance of the Nasal-SI system, with the best-performing embedding model being proposed as the optimal input. The Nasal-SI model consists of two bi-directional layers of 128 Gated Recurrent Units (GRUs) with dropout layers to prevent overfitting. After the second GRU layer, a self-attention layer is applied to enhance contextual information and capture long-range dependencies. The output is processed through a dense layer with 128 hidden units, upsampled from 50 Hz (the sampling rate of WavLM-Large and HuBERT-Large) to 100 Hz to match the target output sampling rate, and then passed through another dense layer with five hidden units to generate the outputs for VP TV, EGG-env, and three SF. The ADAM optimizer is used with a learning rate of $5e-4$ and batch size of 8, selected via grid search. The nasalance measure used as ground truth for estimating the VP TV parameter was computed from recorded oral and nasal microphone signals by first applying a high-pass filter with a 20 Hz cutoff to remove low-frequency background noise. The acoustic energy was then calculated using the root mean square of the signals and smoothed with a 25 ms rectangular moving average filter. The nasalance measure was then computed using the nasal acoustic energy (AEnasal) and oral acoustic energy (AEoral) with equation 1:

$$\text{Nasalance} = \frac{AEnasal}{AEnasal + AEoral} \quad (1)$$

The nasalance was downsampled to 100 Hz and normalized to the range of -1 to 1. The ground truth for EGG-env parameter was computed using the EGG signal. The EGG signal recorded at 51.2 kHz sampling rate, was first high-pass filtered at 20 Hz to remove baseline wander and noise. The EGG envelope was then extracted by computing the magnitude of the Hilbert transform. This envelope was downsampled to 100 Hz and normalized to the range of -1 to 1.

3.2. Developing STL-SI and MTL-SI Systems

The STL-SI system was developed using a Single-Task Learning (STL) approach to estimate 10 parameters, including 6 oral TVs, VP TV, and 3 SF. The MTL-SI system was developed using a Multi-Task Learning (MTL) approach, where the first task estimates the 6 oral TVs, and the second task estimates VP TV and 3 SF with a separate final layer. Both approaches were evaluated to identify the best method for developing the synergistic model that estimates these parameters simultaneously. The ground truth values for the 6 oral TVs were obtained using the method outlined in Section 2.2. For the 3 SF, the ground truth values were obtained using the APP detector [23]. We used the best-performing Nasal-SI system (with WavLM-Large embedding) to obtain ground truth estimates for VP TV for XRMB audios to develop the STL-SI and MTL-SI systems.

For the STL-SI and MTL-SI systems, we used the pre-trained WavLM-Large model to extract representations from input speech signals (Figure 3). The representations from 25 hidden layers of the WavLM-Large embedding were stacked, and a 2D convolutional layer computed a weighted sum of these representations, resulting in a single-layer output. This was processed through three bidirectional GRU layers (two 512-unit and one 256-unit GRU, all with a 0.3 dropout rate). This was followed by a dense layer with 128 hidden units, upsampling by a factor of 2, batch normalization, and a 0.3 dropout rate. For the final step, two approaches were explored; STL-SI used the STL approach with a dense layer with 10 units to estimate 6 oral TVs, one VP TV and 3 SF (Figure 3, Part A). MTL-SI used the MTL approach with two final layers, one for estimating the 6 oral TVs (a dense layer with 6 units) and the other for estimating VP TV and 3 SF (a dense layer with 4 units) (Figure 3, Part B).

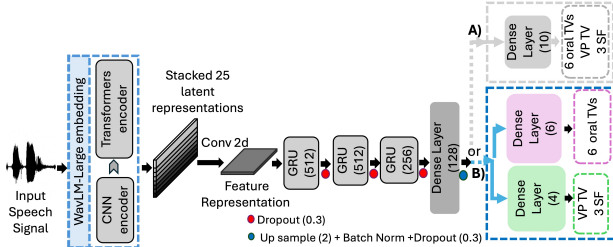


Figure 3: *Proposed model architecture for the SI systems: A) STL-SI system B) MTL-SI system.*

STL-SI and MTL-SI were trained using the Adam optimizer with a learning rate of $5e-4$ and a batch size of 8, both selected via grid search. Early stopping with a patience of 8 epochs was used to prevent overfitting. The loss function to develop STL-SI and MTL-SI is shown in Equation 2, which combines Pearson Correlation (PC) and Root Mean Square Error (RMSE), with α empirically set to 0.8 for optimal performance:

$$\text{Loss} = \alpha(1 - \text{PC}) + (1 - \alpha) \cdot \text{RMSE} \quad (2)$$

4. Results and Discussion

4.1. Results for Nasal-SI System

To evaluate the Nasal-SI system, we used Pearson Product-Moment Correlation (PPMC) to measure the similarity between the ground truth and estimated parameters. Building on the approach in [6], the training set was expanded by incorporating

recordings from four additional speakers, to improve the VP TV estimation. The performance of the Nasal-SI system was compared with the BiGRNN-HuBERT model from [6], which served as the baseline, with both models evaluated on the same test set for consistency. In [6], PPMC results for the BiGRNN-HuBERT model were reported after segmenting the test set audio into fixed 2-second duration intervals. Therefore, we evaluated the Nasal-SI system under two scenarios: first, the audio in the test set was segmented into 2-second durations, and the PPMC score was computed on these intervals for the Nasal-SI system. These results were then compared to those of the BiGRNN-HuBERT model. Second, the test set audio was used in its unsegmented form, with the PPMC score calculated for complete utterances. We also obtained PPMC results for the BiGRNN-HuBERT model under this scenario and compared them to those of the Nasal-SI system.

Table 2: *PPMC scores of the Nasal-SI models on the Nasometry-EGG test set. H: HuBERT-Large, W: WavLM-Large.*

Model	Embed.	Seg.	VP	EGG-env	Per	Ap	F0
[6]	H	-	0.8757	0.8258	0.7309	0.5307	0.7644
Nasal-SI	H	-	0.8904	0.8331	0.7320	0.5348	0.7669
Nasal-SI	W	-	0.9152	0.8403	0.6683	0.5392	0.6957
[6]	H	2	0.8115	0.8330	0.8373	0.8542	0.8562
Nasal-SI	H	2	0.8533	0.8145	0.7648	0.5687	0.7214
Nasal-SI	W	2	0.8663	0.8425	0.7673	0.6111	0.7209

As shown in Table 2, the Nasal-SI system with the WavLM-Large model outperforms alternative configurations, achieving the highest PPMC score across both experimental scenarios. Specifically, the relative improvements compared to the baseline BiGRNN-HuBERT are 6.75% and 4.51% for segmented and unsegmented audio conditions, respectively. Moreover, the results show that using WavLM-Large embeddings improves performance in VP TV estimation compared to Hubert-Large embeddings when developing the Nasal-SI system. This finding aligns with work in [26], which highlighted the superior performance of WavLM-Large embeddings over Hubert-Large embeddings in ASR and other tasks. In this study, we extend this conclusion by demonstrating that WavLM-Large significantly boosts performance in SI systems, particularly for VP TV estimation, a domain not explored in [26].

4.2. Results for the STL-SI and MTL-SI Systems

4.2.1. Oral TVs Estimation in STL-SI and MTL-SI Models

The SI system in [16], which employs HuBERT-Large SSL representations, serves as the baseline for comparison with the STL-SI and MTL-SI systems, since it applies the same geometric transformations to generate ground truth values for the oral TVs as those used in the STL-SI and MTL-SI systems. Additionally, to ensure a fair comparison, we used the same train, test, and development splits. As shown in Table 3, developing the SI system with the MTL approach resulted in the best performance for oral TVs parameter estimation, outperforming both the STL-SI system and the baseline model. Specifically, MTL-SI achieved a 4.70% relative improvement in the average PPMC score for oral TVs compared to the baseline model.

Table 4 presents an ablation study of the MTL-SI system, the top-performing model, to identify the effect of SF and VP TV on oral TVs estimation. In the first row, we excluded the VP TV and 3SF parameters, estimating only the 6 oral TVs, which created a single-task setup similar to the baseline model. Compared to the baseline, the average PPMC score for 6 oral TVs increased from 0.8141 to 0.8411, highlighting the effectiveness

Table 3: PPMC score for estimated parameters of SI models on XRMB test set.

Model	Embedding	VP	LA	LP	TBCL	TBCD	TTCL	TTCD	Per	Ap	F0	AVG. oral TVs
[16]	HuBERT-Large	-	0.8902	0.7142	0.7361	0.8180	0.8032	0.9229	-	-	-	0.8141
STL-SI	WavLM-Large	0.9420	0.9026	0.7376	0.7591	0.8509	0.8307	0.9403	0.9326	0.8759	0.7551	0.8372
MTL-SI	WavLM-Large	0.9462	0.9104	0.7594	0.7981	0.8626	0.8360	0.9478	0.9403	0.8815	0.7470	0.8524

of the proposed SI system design. In the subsequent steps, we estimated either the VP TV or 3SF along with the 6 oral TVs and trained the MTL-SI model in each case. In both scenarios, the PPMC scores improved, demonstrating that incorporating additional speech information, whether 3SF or nasalance, enhances the estimation of oral TVs. Finally, the best performance in oral TVs estimation was achieved when 6 oral TVs were estimated together with VP TV and 3SF. This demonstrates that integrating complementary phonetic information into the MTL framework enhances the accuracy of oral TVs estimation, which is consistent with recent works [8, 27].

Table 4: Ablation study of the MTL-SI system: PPMC scores for different parameter exclusions on the XRMB test Set.

Excluded Param	VP	AVG. oral TVs	Per	Ap	F0
VP, 3 SF	-	0.8411	-	-	-
3 SF	0.9503	0.8485	-	-	-
VP	-	0.8489	0.9437	0.8894	0.7678
-	0.9462	0.8524	0.9403	0.8815	0.7470

4.2.2. VP TV Estimation in the STL-SI and MTL-SI Models

We evaluated the STL-SI and MTL-SI systems for VP TV estimation using the Nasometry-EGG test set, with ground truth derived from recorded oral and nasal signals. We evaluated the STL-SI and MTL-SI systems in two scenarios: using 2-second audio segments and the original unsegmented audios from the test set. The results of the best-performing Nasal-SI system (with WavLM-Large) for both scenarios are also presented for comparison in Table 5. The results in Table 5 indicate that, for each of the two scenarios, both the STL-SI and the MTL-SI systems outperform the Nasal-SI system. The MTL-SI system outperforms STL-SI in VP TV estimation, highlighting that MTL not only improves oral TVs estimation (as shown in Table 3) but also enhances VP TV estimation. The MTL-SI model outperformed the BiGRNN-HuBERT [6] (used as the baseline model in Table 2) by relative improvements of 5.82% in the unsegmented audio scenario and 9.08% in the segmented audio scenario. Overall, the MTL-SI system achieved high PPMC scores for both oral TVs and VP TV. This suggests that the MTL-SI system, by consolidating the estimation of VP TV and oral TVs into a single model, is more powerful compared to SI systems that estimate them separately.

Table 5: PPMC scores for VP TV estimation of MTL-SI and STL-SI models on the Nasometry-EGG test set.

Model	Embedding	Segment	VP
Nasal-SI	WavLM-Large	-	0.9152
STL-SI	WavLM-Large	-	0.9193
MTL-SI	WavLM-Large	-	0.9267
Nasal-SI	WavLM-Large	2 Second	0.8663
STL-SI	WavLM-Large	2 Second	0.8802
MTL-SI	WavLM-Large	2 Second	0.8852

4.2.3. Cross-corpus evaluation of the proposed MTL-SI system

To further test generalization, we compared the outputs from the MTL-SI system for an utterance ‘‘Say packed memos’’ taken

from a different dataset for which nasometry and Electromagnetic Articulography (EMA) were used to simultaneously collect audio and the X-Y movements of point-source sensors placed on the lips, Tongue Tip (TT) and Tongue Body (TB). Using the oral and nasal signals, we obtained the nasalance as defined in equation 1. Figure 4 shows that the MTL-SI system accurately estimates the constriction patterns for each consonant, closely matching the ground truth. Furthermore, for the two instances of /m/ in the ‘‘memos’’, the VP TV estimated by the MTL-SI system exhibits two peaks, consistent with the nasalance and co-located with the lip closure for the two /m/. These results suggest that the MTL-SI system can generalize effectively to a corpus that was not encountered during training.

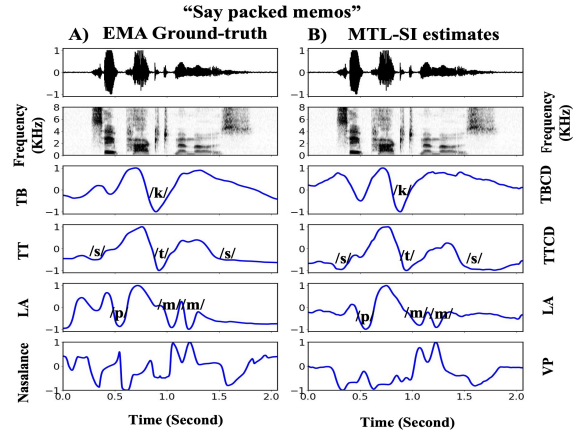


Figure 4: Waveforms, spectrograms, followed by a comparison of the ground-truth EMA and nasalance with the corresponding estimates from the MTL-SI system for the utterance ‘‘Say packed memos’’. The second row shows the spectrogram, visualizing the signal’s frequency content over time with color intensity representing frequency strength. The primary constriction for each consonant is labeled. Note that the peaks in VP match the labial constrictions for the two /m/ in ‘‘memos’’.

5. Conclusions And Future Work

In this paper, a novel SI system was proposed that simultaneously estimates VP TV, along with oral TVs and three source features. The results demonstrated that the proposed synergistic SI model improves the estimation of VP TV and oral TVs, underscoring the complementary nature of these characteristics. Additionally, the results highlighted the effectiveness of the multi-task learning framework and the use of WavLM-Large self-supervised representations to boost the performance of the SI system. In particular, the MTL-SI model outperformed other SI systems that estimate VP TV and oral TVs separately, showcasing the advantages of integrating multiple tasks into a single model. The MTL-SI system, by estimating multiple speech parameters within a single framework, has clinical potential for effective monitoring and intervention, particularly for craniofacial disorders. Future work will focus on its application in the clinical diagnosis of velopharyngeal port dysfunction.

6. References

- [1] K. N. Stevens, *Acoustic phonetics*. MIT press, 2000, vol. 30.
- [2] C. P. Browman and L. M. Goldstein, "Towards an articulatory phonology," *Phonology*, vol. 3, pp. 219–252, 1986.
- [3] E. L. Saltzman and K. G. Munhall, "A dynamical approach to gestural patterning in speech production," *Ecological psychology*, vol. 1, no. 4, pp. 333–382, 1989.
- [4] C. Y. Espy-Wilson, A. C. Lammert, N. Seneviratne, and T. F. Quatieri, "Assessing neuromotor coordination in depression using inverted vocal tract variables," in *Interspeech*, 2019, pp. 1448–1452.
- [5] E. Fry, "Phonics: A large phoneme-grapheme frequency count revised," *Journal of Literacy Research*, vol. 36, no. 1, pp. 85–98, 2004.
- [6] Y. M. Siriwardena, S. E. Boyce, M. K. Tiede, L. Oren, B. Fletcher, M. Stern, and C. Y. Espy-Wilson, "Speaker-independent speech inversion for recovery of velopharyngeal port constriction degree," *The Journal of the Acoustical Society of America*, vol. 156, no. 2, pp. 1380–1390, 2024.
- [7] N. Seneviratne, G. Sivaraman, and C. Y. Espy-Wilson, "Multi-corpus acoustic-to-articulatory speech inversion," in *Interspeech*, 2019, pp. 859–863.
- [8] P. Wu, L.-W. Chen, C. J. Cho, S. Watanabe, L. Goldstein, A. W. Black, and G. K. Anumanchipalli, "Speaker-independent acoustic-to-articulatory speech inversion," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [9] S. Udupa, A. Illa, and P. K. Ghosh, "Streaming model for acoustic to articulatory inversion with transformer networks," in *INTER-SPEECH*, 2022, pp. 625–629.
- [10] Q. Fang, "On the performance of ema-synchronized speech and stand-alone speech in acoustic-to-articulatory inversion," in *Proc. Interspeech 2024*, 2024, pp. 3110–3114.
- [11] T. Yan, K. Maekawa, Y. Nota, and M. Hirata, "Combining language corpora in a japanese electromagnetic articulography database for acoustic-to-articulatory inversion," in *Proc. INTER-SPEECH*, vol. 2023, 2023, pp. 1464–1467.
- [12] N. Seneviratne, G. Sivaraman, V. Mitra, and C. Y. Espy-Wilson, "Noise robust acoustic to articulatory speech inversion," in *Interspeech*, 2018, pp. 3137–3141.
- [13] R. Feng, Y.-A. Chen, Y.-L. Liu, J.-H. Yuan, and Z.-H. Ling, "Wav2nas: An exploratory approach to nasalance estimation in speech," in *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2024, pp. 1–5.
- [14] Y. M. Siriwardena, C. Espy-Wilson, S. Boyce, M. K. Tiede, and L. Oren, "Speaker-independent speech inversion for estimation of nasalance," *arXiv preprint arXiv:2306.00203*, 2023.
- [15] C. J. Cho, P. Wu, A. Mohamed, and G. K. Anumanchipalli, "Evidence of vocal tract articulation in self-supervised learning of speech," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [16] A. A. Attia, Y. M. Siriwardena, and C. Espy-Wilson, "Improving speech inversion through self-supervised embeddings and enhanced tract variables," in *2024 32nd European Signal Processing Conference (EUSIPCO)*. IEEE, 2024, pp. 306–310.
- [17] 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)*. IEEE, 2023, pp. 1–5.
- [18] F. Darley, A. Aronson, and J. Brown, "Motor speech disorders, saunders w," *B, Philadelphia*, pp. 171–975, 1975.
- [19] E. H. Rothauer, "Ieee recommended practice for speech quality measurements," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, no. 3, pp. 225–246, 1969.
- [20] R. A. Krakow, *The articulatory organization of syllables: A kinematic analysis of labial and velar gestures*. Yale University, 1989.
- [21] J. R. Westbury, "Speech production database user's handbook," *IEEE Personal Communications-IEEE Pers. Commun.*, vol. 0, no. 1994.
- [22] A. A. Attia and C. Y. Espy-Wilson, "Masked autoencoders are articulatory learners," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [23] 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.
- [24] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhota, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.
- [25] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong et al., "Speechbrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.
- [26] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao et al., "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.
- [27] Y. M. Siriwardena, G. Sivaraman, and C. Espy-Wilson, "Acoustic-to-articulatory speech inversion with multi-task learning," *arXiv preprint arXiv:2205.13755*, 2022.