Interspeech 2024
1-5 September 2024, Kos, Greece

Accent Conversion with Articulatory Representations

Yashish M. Siriwardena[1], *Nathan Swedlow*[2], *Audrey Howard*[2], *Evan Gitterman*[2], *Dan Darcy*[2], *Carol Espy-Wilson*[1], *Andrea Fanelli*[2]

[1]University of Maryland College Park, MD, USA
[2]Dolby Laboratories, CA, USA

yashish@terpmail.umd.edu, Nathan.Swedlow@dolby.com, audrey.howard@dolby.com, evan.gitterman@dolby.com, Dan.Darcy@dolby.com, espy@umd.edu, Andrea.Fanelli@dolby.com

Abstract

Conversion of non-native accented speech to native (American) English has a wide range of applications such as improving intelligibility of non-native speech. Previous work on this domain has used phonetic posteriograms as the target speech representation to train an acoustic model which is then used to extract a compact representation of input speech for accent conversion. In this work, we introduce the idea of using an effective articulatory speech representation, extracted from an acoustic-to-articulatory speech inversion system, to improve the acoustic model used in accent conversion. The idea to incorporate articulatory representations originates from their ability to well characterize accents in speech. To incorporate articulatory representations with conventional phonetic posteriograms, a multi-task learning based acoustic model is proposed. Objective and subjective evaluations show that the use of articulatory representations can improve the effectiveness of accent conversion.
Index Terms: accent conversion, tract variables, huBERT, multi-task learning, zero-shot

1. Introduction

Foreign Accent Conversion (FAC) is the process of transforming non-native English speech to match the accent or pronunciation pattern of a native English speaker, while retaining the speaker identity [1]. The defining step of every FAC pipeline is the extraction of meaningful speaker-independent speech embeddings, that disentangle every other speaker-specific vocal element from the accent. In most of the FAC pipelines, the Acoustic Model (AM) is the model pre-trained on native speech (eg. Librispeech [2]) to estimate such speech embeddings. Different versions of phonetic posteriograms (PPGs) have been previously used as speech embeddings for accent conversion [3, 4, 1], capturing how individual sounds are produced in speech over time. However, the higher dimensionality of PPGs, combined with their sparse nature, have led to the idea of using compact bottle neck features (BNFs) – features normally extracted from a layer of the pre-trained acoustic model – for the downstream accent conversion [1, 5, 6].

Techniques used for accent conversion can be broadly categorized into acoustic and articulatory methods [1]. Articulatory-based methods use speaker's articulatory trajectories – absolute x, y, z coordinates of the pallets placed on articulators, (e.g. lips, tongue, jaw) – to train a speaker-specific articulatory synthesizer. Here the synthesizer is trained to learn a mapping from the L2 speaker's articulatory data to an acoustic feature (e.g. Mel Cepstra). To perform accent conversion, the trained articulatory synthesizer is then provided with articulatory trajectories from a native speaker [7, 8]. Disentangling accent from voice identity in the articulatory domain is intuitive and effective, but it is impractical due to the challenges in collecting actual ground-truth articulatory data, which is expensive and requires specialized equipment [1]. On the other hand, acoustic methods are more applicable since they only require speech in acoustic form, but can be less effective in decoupling the accent from the other important vocal traits [1].

In this paper, we explore the possibility of using relative articulatory measures (Browman et al. [9]) in place of the absolute articulatory trajectories previously used for accent conversion. These relative articulatory measures, which are referred to as vocal tract variables (TVs), are more speaker-independent in nature when compared to the absolute articulatory trajectories. Because of that, they can be effective in developing zero-shot accent conversion systems without training any speaker-specific speech synthesizer. To circumvent the need of obtaining ground-truth articulatory data (TVs), we use a state-of-the-art acoustic-to-articulatory speech inversion system [10], originally trained with ground-truth articulatory data from the Wisconsin X-ray microbeam dataset [11]. One of the key contributions of this work is the development of a new acoustic model to extract speech embeddings for the accent conversion pipeline. In our model, the TVs extracted from the speech inversion system are used along with PPGs in a multi-task learning framework, to address whether utilizing PPGs or TVs alone, or a combination of both, can enhance the task of accent conversion.

2. Related Work

Foreign accent conversion with articulatory representations have been previously explored in [8, 12, 13], where speaker-specific articulatory synthesizers are first trained with ground-truth Electro Magnetic Articulograpy (EMA) data. As discussed in [12], these articulatory based synthesizers have struggled in synthesizing speech with good acoustic quality compared to that of acoustic feature based synthesizers (eg. MFCCs). The incomplete representation of the vocal tract, compounded with the difficulty in collecting articulatory data, has led the speech community to resort to purely acoustic based methods to perform accent conversion. However, recent work with vocal tract variables, an improved articulatory representation, has shown that they can be effectively used for high quality speech synthesis [14, 15]. This also suggests that effective articulatory representations with improved deep learning algorithms can enable accent conversion with articulatory representations.

Apart from categorizing the accent conversion pipelines to articulatory and acoustic based methods, they can also be discussed based on the use (or not) of native reference speech. Accent conversion systems developed over the years have used reference native speech to extract the native pronunciation pattern to perform accent conversion [16, 3, 4, 6]. Since the applications of this approach are limited, recent work has mostly explored ways to perform reference-free accent conversion

10.21437/Interspeech.2024-1416

[17, 1, 18, 5, 19]. In a recent work by Quamer et al., [6], an accent conversion pipeline with a transformer based seq2seq model was introduced. The accent conversion pipeline here uses a native reference and involves an acoustic model trained to estimate senone-PPGs. The work there is rather exploratory and shows that segmental and prosodic cues of non-native speech can be disentangled. We adopted the training paradigm and the seq2seq synthesizer architecture from this work and modified it to perform accent conversion with a newly designed acoustic model to incorporate articulatory representations.

3. Method

3.1. Articulatory representations from Speech Inversion

Acoustic-to-articulatory speech inversion (SI) aims to infer articulatory dynamics from spoken sounds [20]. While efforts to interpret articulatory movements from continuous speech signals have a long history [21], they have typically been limited to tracking specific parts of the vocal tract, like the upper and lower lips, tongue tip, and velum closure. However, it's essential not only to understand the primary effects of individual vocal tract movements, but also to grasp how these articulators interact. For instance, articulators such as the lips and jaw often cooperate to achieve specific vocal tract shapes [9]. Consequently, general SI systems prioritize understanding vocal tract constriction, estimating the degree and position of functional tract variables (TVs; from Articulatory Phonology in [9]), rather than solely focusing on individual articulator movement. During SI, acoustic features extracted from speech signals are used to predict these tract variables. This process involves learning an inverse mapping by training on a dataset containing matched acoustic and directly observed articulatory data.

To train the acoustic model discussed in section 3.2, TVs extracted from the SI system in [10] were used as the ground-truth. The SI system trained and evaluated in speaker-independent fashion estimates six functional TVs. Figure 1 shows how each of the 6 TVs can be visualized with respect to a vocal tract and the corresponding articulators involved.

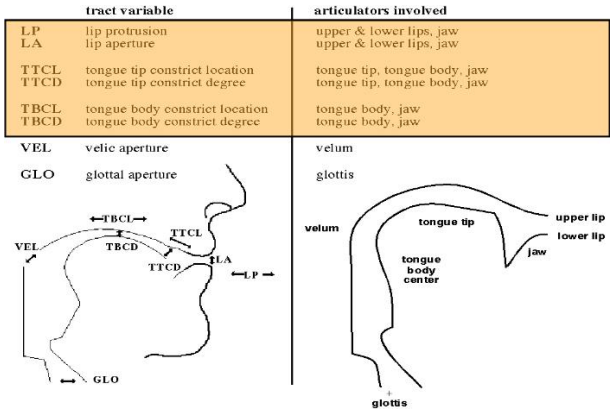

Figure 1: *Vocal tract variables and related articulators. The color shaded TVs are extracted from a speech inversion system*

3.2. Multi-task learning based Acoustic Model

The proposed multi-task learning based acoustic model (AM) contains two bidirectional long short memory (BiLSTM) layers followed by up-sampling, dropout and fully connected (Linear) layers as shared model layers. A fully connected layer with softmax activation is used to estimate PPG outputs and a fully connected layer with tanh activation is used to estimate TV outputs. Figure 2 shows details of the model architecture implemented in PyTorch. The model is optimized using a combined loss ($AMLoss_{combined}$) as defined in equation 1. A hyper-parameter (α) is used to weight the two individual losses incurred in estimating PPGs (PPG_{loss}) and TVs (TV_{loss}). The PPG_{loss} is the cross entropy loss between the estimated and the ground-truth senones. The TV_{loss} is the mean absolute error (MAE) between the estimated and the ground-truth TVs.

$$AMLoss_{combined} = \alpha \times TV_{loss} + (1 - \alpha) \times PPG_{loss} \quad (1)$$

Three values of alpha were explored: α = 1, which we will refer to as the 'TV only' model, α = 0.4, which we will refer to as the 'combined' model and α = 0, which we will refer to as the 'PPG only' model. The α for the combined model was chosen by doing a grid search across [0, 0.2, 0.3, 0.4, 0.5, 0.7, 1.0], and checking the average Pearson's product moment correlation (PPMC) scores for TV estimation and RMSE for PPG estimation. α = 0.4 gave the best compromise between the PPMC scores for estimated TVs and loss for estimating PPGs.

The learning rates to train the models were determined based on a grid search by testing all combinations from [1e-2, 1e-3, 1e-4, 3e-4] that resulted in 1e-4 as the best pick. A similar grid search was done to choose the batch size from [4, 8, 12, 16] and a batch size of 8 gave the best validation loss. The objective function was optimized using the ADAM optimizer with an 'ExponentialLR' learning rate scheduler and a decay of 0.5. All models were trained with an early stopping criteria monitoring the validation loss and using a patience of 6 epochs. Once the models are trained, the BNFs are extracted from the final layer of the 'shared model layers' as shown in figure 2.

3.2.1. Dataset and input acoustic features

Original train, dev and test splits (both -clean and -other) from the LibriSpeech dataset [2] was used to train all the acoustic model variants. All the audio files were first segmented to 2 second long segments and the shorter ones were zero padded at the end. As shown in figure 2, the acoustic model takes in Hidden-Unit BERT (HuBERT) [22] speech embeddings extracted from the pre-trained HuBERT-large model as the input speech representation. The HuBERT speech embeddings are sampled at 50 Hz and have a dimensionality of 1024. Tri-phone Phonetic Posteriograms (PPGs) are extracted from a pretrained model [3] as one of the target speech representations. The extracted PPGs are sampled at 100Hz and have a dimensionality of 5816. As the other target speech representation, TVs are extracted from a pre-trained acoustic-to-articulatory speech inversion system. The TVs are sampled at 100 Hz and contain 6 distinct variables as discussed in section 3.1.

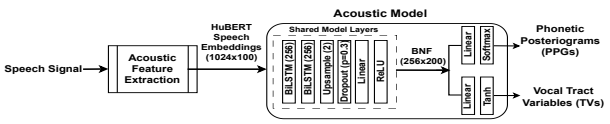

Figure 2: *Proposed Multi-task learning based Acoustic model from which PPG only, TV only and Combined BNFs are extracted for accent conversion*

3.3. Accent Conversion Pipeline

Our experiments were conducted using the ARCTIC [23] and L2-ARCTIC [24] datasets. This combined dataset comprises recordings from 28 speakers, each providing 1,132 utterances. Four speakers (NJS, YKWK, TXHC, and ZHAA) were excluded from the training set to serve as unseen speakers during testing. Same train and dev splits as in [6] were used. Speaker

BDL from the ARCTIC dataset was chosen as the reference native language (L1) speaker for all experiments. For each utterance, we extracted 80-dimensional Mel-spectrograms using a 25ms window and a 10ms shift. To convert Mel-spectrograms into waveforms, we utilized a pre-trained HiFi-GAN vocoder [25]. Figure 3 details the foreign accent conversion pipeline used. The model architecture and the training procedure is adopted from the work in [6] and comprises of a prosody encoder and a seq2seq synthesizer.

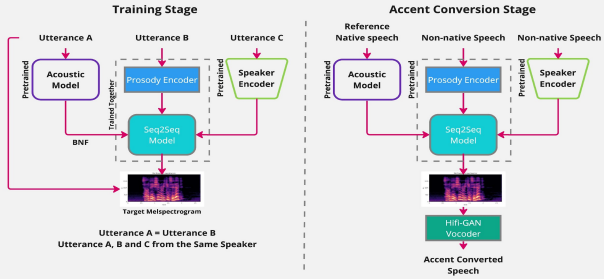

Figure 3: *Training and Accent conversion stages of the pipeline*

The acoustic model (AM) transforms an input 'utterance A' into a bottleneck feature (BNF) embedding, which encapsulates both the phonetic content and the articulatory constructs of the utterance. The seq2seq model takes in three inputs: BNFs from the AM, a speaker embedding representing the voice characteristics from 'utterance C' from a specific speaker, and a prosody embedding from a reference 'utterance B'. Utilizing these three sources of information, the seq2seq model tries to reconstruct the original Mel spectrogram. The prosody encoder and seq2seq model are trained synchronously in an unsupervised manner, similarly to an auto-encoder, while the speaker encoder [6] and acoustic model are pre-trained beforehand. Throughout training, the acoustic model and prosody encoder are provided with the same utterance from the same speaker, while the speaker encoder receives a different utterance C from the same speaker to generate a speaker embedding. This strategy ensures that the prosody encoder learns a distinct mapping from the speaker encoder, and the seq2seq model avoids inferring prosody solely from the speaker embedding.

Once the prosody encoder and the seq2seq synthesizer are trained, accent conversion is performed as shown in the right panel of Figure 3. Here, a reference speech from a native speaker (matching linguistic content with the L2 speech to be converted) is used to generate the BNFs from the pre-trained acoustic model. A mel-spectrogram extracted from the non-native speech utterance (which needs to be accent converted) is fed to the prosody encoder, and a speaker embedding extracted from the same utterance with the speaker encoder is fed to the seq2seq synthesizer. The synthesized mel-spectrogram as shown in figure 3 is then passed through the HiFi-GAN vocoder [25] to generate the accent converted audio waveform.

4. Results

This section summarises the objective and subjective evaluations of the accent conversion pipelines. Synthesized audio samples can be found in the web page[1].

4.1. Objective Evaluations

4.1.1. t-distributed Stochastic Neighbor Embedding (t-SNE) visualizations

Figure 4 shows the t-SNE visualization of the speaker embeddings generated by the speaker encoder for original L2 speech samples and the corresponding accent converted samples of the unseen test set data. Here the visualizations are only provided for the TV only variant since it reported the lowest average distance (9.25 $\pm$5.1) between the cluster centroids for accent converted and corresponding original L2 samples. Accent converted speech clustering closer to original speaker's samples suggests that the TV only model is preserving the speaker's identity noticeably well. It can also be seen that a better speaker identity transfer has happened with Spanish and Arabic speakers (NJS and ZHAA) compared to Korean and Chinese speakers (YKWK and TXHC).

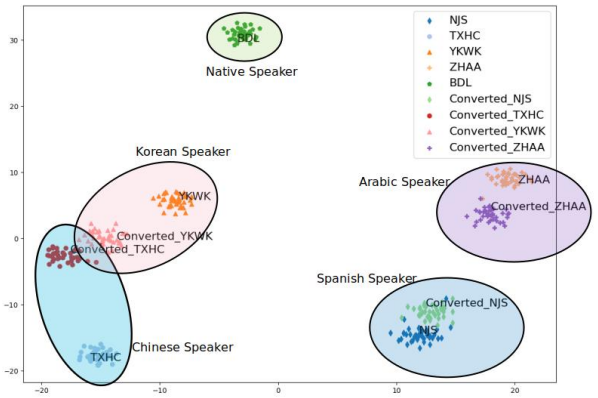

Figure 4: *t-SNE visualizations of speaker embeddings from original L2 speech and corresponding accent converted samples from TV only variant*

4.1.2. Mel Cepstral Distortion (MCD)

Table 1: *MCD results from the 3 acoustic model variants. Lower values suggest better synthesis quality*

	Combined	**TV only**	**PPG only**
NJS(Spanish)	5.8574	5.8633	5.7314
TXHC (Mandarin)	6.5491	6.4276	6.3474
YKWK (Korean)	6.7301	6.6984	6.7331
ZHAA (Arabic)	5.6177	5.6040	5.4922
Average	6.3719	6.1614	**6.0918**

4.1.3. Word Error Rate (WER)

Table 2: *WER results for the 3 acoustic model variants. Lower scores suggest better recognition by the Whisper-large ASR system [26] and hence lesser accentedness*

	Combined	**TV only**	**PPG only**	**Original audio**
NJS(Spanish)	14.70	14.62	18.75	74.70
TXHC (Mandarin)	21.83	20.67	25.16	134.96
YKWK (Korean)	15.57	15.04	20.08	152.73
ZHAA (Arabic)	19.72	20.99	23.38	76.48
Average	17.96	**17.83**	21.84	134.72

4.2. Subjective Analysis

We conducted two experiments to evaluate the performance of the TV only, PPG only, and combined FAC systems. Each of these tests investigated a key attribute of interest including accentedness and acoustic quality respectively. Stimulus ordering was randomized and counter balanced across all subjects for both tests. All subjects who participated in these experiments are fluent English speakers and United States citizens. 20 subjects completed each assessment and all subjects were paid to participate. Each test was administered using an internal browser-based testing tool and subjects were instructed to perform each test from a quiet workspace of their choice using headphones for audio playback.

[1] https://yashish92.github.io/Accent-conversion-TVs/

Table 3: *Accent Conversion Test ratings (higher values for less perceived foreign accent) and Acoustic Quality Test ratings (higher scores for better quality) for the PPG only, TV only and combined systems compared to L2 and L1 references*

	Combined	TV only	PPG Only	Reference L2	Reference L1
Accent Conversion	**7.32 ±0.51**	7.21 ±0.43	6.78 ±0.43	2.07 ±0.37	8.56 ±0.14
Voice Quality (MOS)	2.45 ±0.33	2.59 ±0.36	**2.69 ±0.33**	3.36 ±0.43	3.46 ±0.18

4.2.1. Accent Conversion Test

Subjects were tasked with providing a score of accent conversion across 75 audio signals using a nine-point Likert scale, where a score of 9 corresponds to no foreign accent and a score of 1 corresponds to heavy foreign accent. Similar experimental methods have been performed to assess the efficacy of accent conversion by other FAC systems and in these experiments higher scores typically correspond to greater levels of accentedness [27, 5]. Given that our research evaluates the performance of accent conversion systems, we determined that higher subjective scores should represent the effect of accent conversion.

Each subject ranked 15 utterances per test system (PPG Only, TV Only, Combined) resulting in 45 total accent-converted test signals. Subjects also scored 15 L2 and 15 L1 reference utterances. We evaluated the performance of all three test systems relative to the L2 and L1 references using two-sample t-tests with a Bonferroni adjusted significance level ($\alpha = 0.005$). All three FAC systems scored significantly higher than the L2 reference ($p < 0.005$), meaning subjects perceived less foreign accent in the converted speech. Additionally, all three systems scored significantly lower than the L1 reference ($p < 0.005$), meaning greater accentedness was perceived in the test systems relative to the native English speaker reference. The TV only model and combined model scored slightly higher than the PPG only model, however these results were not statistically significant. Table 3 and figure 5 shows the results.

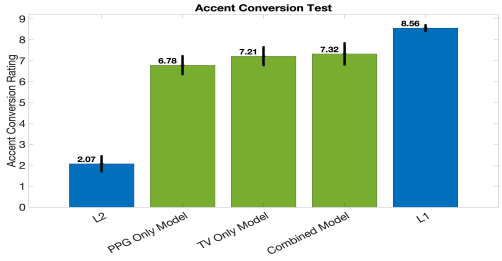

Figure 5: *Bars are mean rating for each system. References are shaded blue and test systems are shaded green. Error bars are 95% confidence interval*

4.2.2. Acoustic Quality Test

Subjects ranked the acoustic quality of 75 utterances using a 5-point mean opinion score (MOS), where higher scores represented improved acoustic quality while lower scores represented poorer quality (5 - Excellent, 4 - Good, 3 - Fair, 2 - Poor, 1 - Bad). Each subject ranked 15 utterances per test system and an additional 15 utterances per reference system. Two-sample t-tests were performed between all three test systems and both reference systems respectively. A Bonferroni adjusted significance level ($\alpha = 0.005$) was applied to all tests. The combined model was the only system to score significantly lower than the L2 reference ($p < 0.005$). Both the PPG only and TV only systems scored lower than the L2 reference ($p = 0.01, p = 0.006$ respectively) however this difference did not reach statistical significance. All three FAC systems scored significantly lower than the L1 reference. Additionally, two-sample t-tests were used to evaluate differences between the three test systems. The PPG only method, TV only method, and combined method performed comparably meaning there was not a statistically significant difference in MOS score between the three systems under test. Results are shown in table 3 and figure 6.

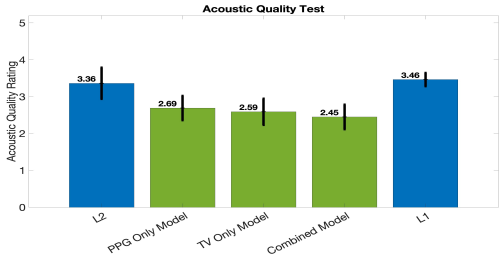

Figure 6: *Bars are MOS mean for each system. References are shaded blue and test systems are shaded green. Error bars are 95% confidence interval*

5. Discussion and Future Work

This work proposes a multi-task learning based acoustic model architecture to incorporate TVs from a speech inversion system, with conventionally used PPGs, to extract a compact speech representation (BNFs) for the downstream task of accent conversion. Three acoustic model variants (PPG only, TV only and Combined) were developed to extract BNFs and then used for training a speaker conditioned seq2seq model architecture in unsupervised fashion. The pre-trained seq2seq model is then used to perform accent conversion on a held out unseen speaker set in zero-shot fashion. The resulting accent converted samples were evaluated based on objective and subjective measures to understand the feasibility of incorporating articulatory features in performing foreign accent conversion.

The MOS scores from the Acoustic Quality Test and the MCD scores from the objective evaluations suggest that the PPG only acoustic model variant has slightly better synthesis quality in accent converted samples. However, the WER scores and the Accent Conversion Test scores suggest that the 'TV only' and 'Combined' acoustic model variants are better at removing the foreign accent compared to the PPG only variant. Additionally, the t-SNE analysis reveals that the speaker identity is better preserved in the TV-only model, which is a crucial aspect of accent conversion. Overall, these findings assure the feasibility of incorporating articulatory features in an accent conversion pipeline, and the potential improvements gained with articulatory representations derived from an acoustic-to-articulatory speech inversion system.

It is important to note that the accent conversion pipeline used in this work requires native reference speech at the time of inference. Even though this limits the use of the proposed system in real-world applications, this work proposes an important direction towards improving the current accent conversion pipelines by incorporating articulatory representations. Moreover, the authors plan to circumvent the need for native reference speech by training a separate translator model (another seq2seq model) in the future. Here the translator model will learn a mapping from non-native BNFs generated from the acoustic model to corresponding native speech's BNFs generated from the same acoustic model. The pre-trained translator model can then be used to generate corresponding native speech's BNFs (left branch of Figure 3 in Accent conversion stage) to replace the need for reference native speech.

6. References

[1] G. Zhao, S. Ding, and R. Gutierrez-Osuna, "Converting foreign accent speech without a reference," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2367–2381, 2021.

[2] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:2191379

[3] G. Zhao, S. Sonsaat, J. Levis, E. Chukharev-Hudilainen, and R. Gutierrez-Osuna, "Accent conversion using phonetic posteriorgrams," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5314–5318.

[4] G. Zhao and R. Gutierrez-Osuna, "Using phonetic posteriorgram based frame pairing for segmental accent conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 10, pp. 1649–1660, 2019.

[5] W. Quamer, A. Das, J. Levis, E. Chukharev-Hudilainen, and R. Gutierrez-Osuna, "Zero-Shot Foreign Accent Conversion without a Native Reference," in *Proc. Interspeech 2022*, 2022, pp. 4920–4924.

[6] W. Quamer, A. Das, and R. Gutierrez-Osuna, "Decoupling Segmental and Prosodic Cues of Non-native Speech through Vector Quantization," in *Proc. INTERSPEECH 2023*, 2023, pp. 2083–2087.

[7] D. Felps, C. Geng, and R. Gutierrez-Osuna, "Foreign accent conversion through concatenative synthesis in the articulatory domain," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 8, pp. 2301–2312, 2012.

[8] S. Aryal and R. Gutierrez-Osuna, "Data driven articulatory synthesis with deep neural networks," *Computer Speech & Language*, vol. 36, pp. 260–273, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230815000200

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

[10] A. A. Attia, Y. M. Siriwardena, and C. Espy-Wilson, "Improving speech inversion through self-supervised embeddings and enhanced tract variables," 2023.

[11] J. R. Westbury, "Speech Production Database User's Handbook," *IEEE Personal Communications*, vol. 0, no. June, 1994.

[12] S. Aryal and R. Gutierrez-Osuna, "Comparing articulatory and acoustic strategies for reducing non-native accents," in *Proc. Interspeech*, 2016. [Online]. Available: https://psi.engr.tamu.edu/wp-content/uploads/2018/01/aryal2016interspeech.pdf

[13] D. Felps, C. Geng, and R. Gutierrez-Osuna, "Foreign accent conversion through concatenative synthesis in the articulatory domain," *IEEE Transactions on Audio, Speech and Language Processing*, 2012. [Online]. Available: https://psi.engr.tamu.edu/wp-content/uploads/2018/01/felps2012taslp.pdf

[14] 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.

[15] Y. M. Siriwardena, C. Espy-Wilson, and S. Shamma, "Learning to Compute the Articulatory Representations of Speech with the MIRRORNET," in *Proc. INTERSPEECH 2023*, 2023, pp. 5137–5141.

[16] S. Ding, G. Zhao, and R. Gutierrez-Osuna, "Accentron: Foreign accent conversion to arbitrary non-native speakers using zero-shot learning," *Computer Speech & Language*, vol. 72, p. 101302, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230821001029

[17] S. Liu, D. Wang, Y. Cao, L. Sun, X. Wu, S. Kang, Z. Wu, X. Liu, D. Su, D. Yu, and H. Meng, "End-to-end accent conversion without using native utterances," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6289–6293.

[18] Z. Wang, W. Ge, X. Wang, S. Yang, W. Gan, H. Chen, H. Li, L. Xie, and X. Li, "Accent and speaker disentanglement in many-to-many voice conversion," in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2021, pp. 1–5.

[19] M. Jin, P. Serai, J. Wu, A. Tjandra, V. Manohar, and Q. He, "Voice-preserving zero-shot multiple accent conversion," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[20] G. Sivaraman, V. Mitra, H. Nam, M. Tiede, and C. Espy-Wilson, "Unsupervised speaker adaptation for speaker independent acoustic to articulatory speech inversion," *The Journal of the Acoustical Society of America*, vol. 146, no. 1, pp. 316–329, 2019. [Online]. Available: https://doi.org/10.1121/1.5116130

[21] G. Papcun, J. Hochberg, T. R. Thomas, F. Laroche, J. Zacks, and S. Levy, "Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data." *The Journal of the Acoustical Society of America*, vol. 92, no. 2 Pt 1, pp. 688–700, aug 1992. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/1506525

[22] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 29, p. 3451–3460, oct 2021. [Online]. Available: https://doi.org/10.1109/TASLP.2021.3122291

[23] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Proc. 5th ISCA Workshop on Speech Synthesis (SSW 5)*, 2004, pp. 223–224.

[24] G. Zhao, S. Sonsaat, A. Silpachai, I. Lucic, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna, "L2-ARCTIC: A Non-native English Speech Corpus," in *Proc. Interspeech 2018*, 2018, pp. 2783–2787.

[25] J. Kong, J. Kim, and J. Bae, "Hifi-gan: generative adversarial networks for efficient and high fidelity speech synthesis," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS'20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[26] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. ICML'23. JMLR.org, 2023.

[27] G. Zhao, S. Ding, and R. Gutierrez-Osuna, "Converting foreign accent speech without a reference," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2367–2381, 2021.