Interspeech 2024
1-5 September 2024, Kos, Greece

Articulatory synthesis using representations learnt through phonetic label-aware contrastive loss

Jesuraj Bandekar, Sathvik Udupa, Prasanta Kumar Ghosh

Electrical Engineering Department, Indian Institute of Science (IISc), Bangalore-560012, India
jesurajbandekar.661@gmail.com, sathvikudupa66@gmail.com, prasantag@gmail.com

Abstract

Articulatory speech synthesis is a challenging task which requires mapping of time-varying articulatory trajectories and speech. In recent years, deep learning methods have been proposed for speech synthesis which have achieved significant progress towards human-like speech generation. However, articulatory speech synthesis is far from human-level performance. Thus, in this work, we further improve the results of articulatory speech synthesis to enhance synthesis quality. We consider a deep learning-based sequence-to-sequence baseline. We improve upon this network using a novel approach of label-aware contrastive learning using framewise phoneme alignment to learn better representations of the articulatory trajectories. With this approach, we obtain a relative improvement in Word Error Rate (WER) of 5.8% over the baseline. We also conduct mean opinion score (MOS) tests and other objective metrics to further evaluate our proposed models.

Index Terms: articulatory synthesis, speech production, label-aware contrastive learning

1. Introduction

In recent years, significant strides have been made in text-to-speech (TTS) using deep learning models [1, 2, 3, 4]. This is possible due to the availability of large-scale datasets and neural network architectures. Text-to-speech (TTS) mapping can be realised as mapping each character or phoneme to its corresponding sound in terms of speech features. This mapping is learnt implicitly with the data-driven learning of deep learning. On the other hand, the problem of articulatory synthesis is much harder. It is known that articulatory movements of speech have critical and non-critical articulators [5, 6, 7]. The presence of non-critical articulators may harm speech generation from time-varying articulatory movements. On the other hand, it may be counter-productive to ignore them since they may contain co-articulation information, which can be beneficial towards synthesis [8]. Additionally, phonemes sharing the same place of articulation are very hard to differentiate using just the articulatory trajectories [9]. Along with these considerations, articulatory to speech mapping datasets are quite limited. These factors present challenges in building high-performance articulatory speech synthesis systems.

Vocal tract modelling has been a prominent research direction towards articulatory synthesis. These works [8, 10, 11, 12] focus on building a physical model of the vocal tract based on articulatory movements for specific sounds (phonemes). These methods typically utilise various articulators such as lips, tongue, pharyngeal wall, velum and larynx offering advantages in terms of interpretability and control over the sound generation process. However, vocal tract models often require complex configurations and data from real-time magnetic resonance imaging (rtMRI) [13, 14] as the data source to capture the dynamic movements of the vocal tract during speech. Deep learning-based synthesis approaches are emerging as a powerful alternative. These methods leverage large amounts of speech data to learn the complex relationship between articulation and acoustics, offering greater generalizability and scalability.

Thus, researchers have proposed various approaches for machine-learning-based articulatory speech synthesis. Authors [15, 16] have proposed deep neural networks for articulatory speech synthesis. Authors have explored the use of excitation features [17] and estimation of speech parameters [18] using Long Short Term Memory (LSTM) models. EMA2S [19] is a multimodal network, which uses a deep learning-based vocoder for synthesising speech. In recent works [20, 21], authors used HiFI-Gan [22] based approaches to synthesise speech articulatory features.

While recent methods have been obtaining significant performance improvements, they are not yet comparable to text-to-speech synthesis. Some works in deep learning-based articulatory synthesis use rtMRI. The rtMRI data has a problem with reverberated speech, which requires enhancement techniques to improve the performance. Additionally, while it contains representations of many articulators, it is high-dimensional data. Due to this, additional pre-processing steps or neural networks may be necessary to extract features. On the other hand, Electromagnetic articulography (EMA)-based data collection results in clean speech and low-dimensional articulatory representation. Recently, significant advancements have been made in acoustic-to-articulatory inversion with EMA [23, 24]. We use similar sequence-to-sequence neural networks as our high-performant baseline for articulatory speech synthesis. Unlike speech synthesis, in articulatory synthesis, we do not need duration modelling. This simplifies the model architecture required for learning the articulatory synthesis mapping. We further build upon the baseline based on label-aware contrastive loss [25]. Distinguishing phonemes that share identical places of articulation poses a fundamental challenge in synthesizing articulatory trajectories. Enhancing the model's learning of more refined representations involves employing frame-level phoneme alignment. This approach involves grouping representations corresponding to identical phonemes. Moreover, adopting the proposed loss function, as opposed to vanilla supervised contrastive learning or cross-entropy loss, addresses this challenge by weighing the loss according to the similarity of input data.

Thus, the contributions of our work are as follows -

- We propose a strong sequence-to-sequence baseline for articulatory synthesis
- We propose a novel approach to using label-aware contrastive loss to learn better representation using phonemes.

10.21437/Interspeech.2024-1756

- We validate our approach through various objective metrics of ASR, MCD and subjective score of MOS

2. Dataset

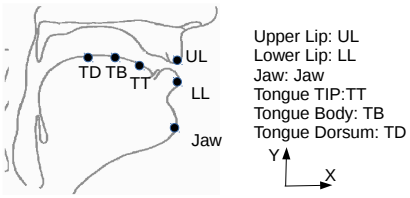

Figure 1: *Schematic diagram indicating the placement of EMA sensors [26]*

In this work, we work with articulatory speech synthesis using an internal dataset of acoustics and its corresponding articulatory data. It consists of utterances spoken on 460 English sentences present in the MOCHA-TIMIT dataset [27]. These sentences were spoken by 38 native Indian speakers, of the age range of 20 to 28. All subjects speak English fluently and have had no speech impairments in the past.

We use electromagnetic articulography (EMA) data as the time-varying articulatory movements. The data was recorded using Electromagnetic Articulograph AG501 [28], using 6 sensors glued on to different articulators as per previous work [29]. The data from the following 6 articulators are collected - Upper Lip (UL), Lower Lip (LP), Jaw, Tongue Tip (TT), Tongue Body (TB), and Tongue Dorsum (TD). We consider the articulatory motions in the midsaggital plane, which corresponds to the horizontal and vertical movement of the sensors (x and y axis). Thus, we have 12 dimensional time-varying articulatory trajectories for each speech utterance.

The dataset split in our experiments is as follows. We use 14,260 utterances from all 38 speakers for the train set. The validation set and test set have 1,610 utterances each. We make sure the sentences in all three sets are disjoint.

3. Proposed Methodology

In this section, we will discuss the background of the proposed method and describe the architecture used.

3.1. Supervised Contrastive Loss ($\mathbf{L}_{scl}$)

Contrastive loss deals with grouping samples that belong to the same class i.e. the positive samples together while pushing away the samples that don't belong to the same class or negative samples. In the case of supervised contrastive loss [30], we use the ground truth labels to know the positive and negative samples and use the loss function accordingly to train the model. Let us assume a data point x_i as the input to a model which gives the normalized output h_i. Let y_i be the corresponding label for the input x_i. For a batch size of K with $I = \{1, \ldots, K\}$ being the indices of the samples, we can write the supervised contrastive loss as

$$L_{\text{scl}} = -\frac{1}{K}\sum_{i=1}^{K}\frac{1}{|P(i)|}\sum_{j\in P(i)}\log\frac{\exp\left(\langle h_i, h_j\rangle/\tau\right)}{\sum_{k\in I/i}\exp\left(\langle h_i, h_k\rangle/\tau\right)} \tag{1}$$

where $P(i)$ is the set of positive pairs for x_i and τ is the temperature hyper-parameter.

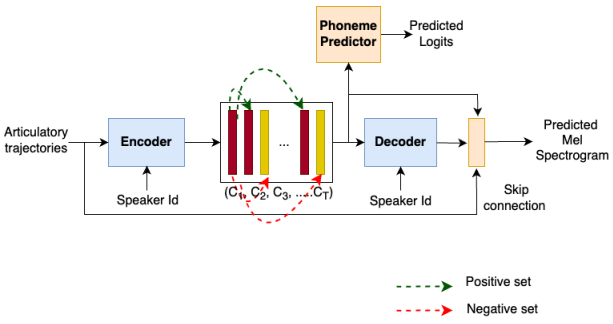

Figure 2: *Figure shows a block diagram of the architecture used. The two coloured arrows show positive and negative sets that are used for contrastive loss. The logits from the phoneme predictor are used as weights for label-aware contrastive loss*

3.2. Label-Aware Contrastive Loss ($\mathbf{L}_{lcl}$)

An issue with the supervised contrastive loss is that it treats all negative samples equally. This cannot always be the case because some of the classes might be harder to separate compared to others. A possible solution to this would be using the label-aware contrastive loss [25] written as

$$L_{\text{lcl}} = -\frac{1}{K}\sum_{i=1}^{K}\frac{1}{|P(i)|}\sum_{j\in P(i)}\log\frac{w_{i,y_i}\exp\left(\langle h_i, h_j\rangle/\tau\right)}{\sum_{k\in I/i} w_{i,y_k}\exp\left(\langle h_i, h_k\rangle/\tau\right)} \tag{2}$$

where w_{i,y_k} is the relationship between sample x_i and label y_k. For this value, we try to find the probability of predicting label y_k for the sample x_i. This prediction can be done by training a model using Cross-Entropy loss. So the weight $\mathbf{w_i}$, for sample x_i is given by

$$\mathbf{w_i} = \frac{\exp(h_i)}{\sum_{c=1}^{C}\exp(h_i)} \tag{3}$$

Here, $\mathbf{w}_i = \{w_{i,c}\}_{c=1}^{C}$ where C is the number of labels and for h_i which is the output of the model for input x_i

3.3. Baseline Architecture

For the baseline, we use an encoder block and a decoder block. Both these blocks are made up of several transformer neural networks [31]. Similar architecture has been used for Acoustic to Articulatory Inversion (AAI) tasks [23, 32]. The 12-dimension raw articulatory trajectories are the input and the model predicts the mel-spectrogram. For input $X = (x_1, x_2, ...x_T)$ where T is the length,

$$\hat{Y} = Model(X) \tag{4}$$

where $\hat{Y} = (\hat{y_1}, \hat{y_2}, ...\hat{y_T})$ is the predicted mel-spectrogram

The model is trained with the mean squared error loss (L_{mse}) between the predicted mel-spectrogram $\hat{Y}$ and the ground truth mel-spectrogram Y

$$L_{mse} = \frac{1}{T}\sum_{i=1}^{T}(y_i - \hat{y}_i)^2 \tag{5}$$

Both the encoder and decoder are conditioned using speaker IDs. We will be referring to the baseline model as **AS-B**.

3.4. Proposed Architecture

Similar to the baseline architecture the proposed architecture consists of transformer blocks as encoder and decoder. The articulatory trajectories are the inputs and both the encoder and

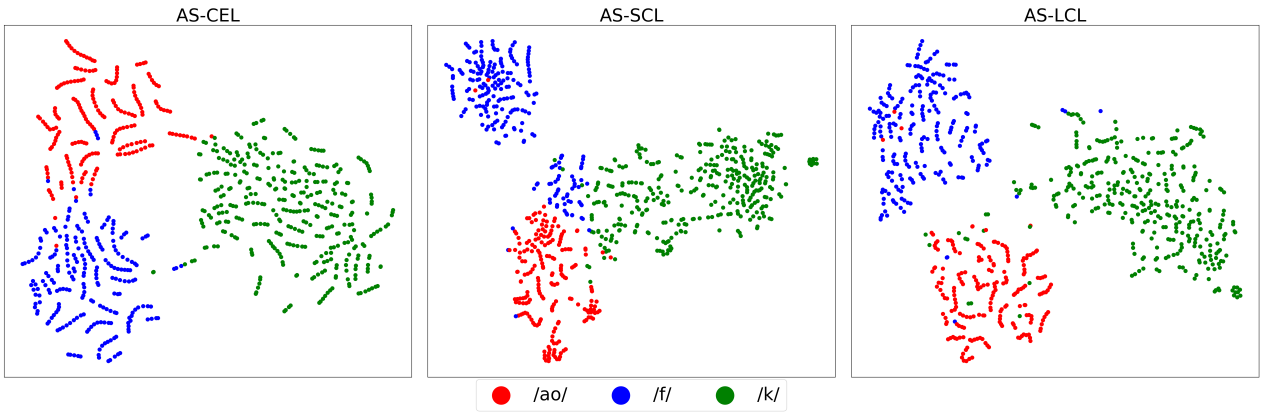

Figure 3: *Figure shows the visualization of the encoder output for the 3 configurations of the proposed model for phonemes /ao/, /f/ and /k/ using t-SNE*

decoder are conditioned with speaker IDs as done in the baseline. Additionally, to improve pronunciation quality and intelligibility of the synthesized audio we try to make the output of the encoder learn the phoneme representation. We achieve this by using the label-aware contrastive loss from equation (2). Let us consider the output of the encoder to be $C = (c_1, c_2, ...c_T)$ for the input X i.e.

$$C = Encoder(X) \tag{6}$$

We get the frame-level phoneme alignment using the Kaldi toolkit [33]. Using this alignment we apply the label-aware contrastive loss in the following manner. For each frame, we consider all the frames that belong to the same phoneme as the positive set and the rest of the frames as the negative set. Additionally, we also try to predict the phoneme for each frame to get the normalized logits that can be used as weights mentioned in equation (3). But instead of using a separate block to predict the phonemes like in [25], we use a linear layer (LL_p) which takes C as input to predict the phoneme.

$$h_c = LL_p(C) \tag{7}$$

We obtain the weights with h_c and train using the Cross-Entropy loss as mentioned in equation (3). Let us call this loss as L_{ce}. The C is then passed on to the decoder. We also use a concatenated skip connection by concatenating the X and C to the output of the decoder. A linear layer is used after this to get the predicted mel-spectrogram. We use the L_{mse} loss from equation (5). Mel-spectrograms are chosen as the prediction target because they provide a stable and lower-dimensional representation of audio signals, and open-source state-of-the-art vocoders can effectively synthesize waveforms from mel-spectrograms, facilitating efficient and high-quality audio generation. So the total loss, L can be written as

$$L = L_{mse} + \alpha L_{lcl} + \beta L_{ce} \tag{8}$$

where α and β are hyperparameters. The HiFi-gan vocoder [22] is employed to synthesize audio for all baselines and proposed models. We use the pre-trained model made available by Speechbrain [34] on huggingface [1]. Additionally, to match the length of the input to the length of the mel-spectrogram required for audio synthesis, we downsample the articulatory trajectories to 62.5 Hz. This downsampling enables us to use the configuration of the mel-spectrogram needed by the vocoder.

[1] https://huggingface.co/speechbrain/tts-hifigan-libritts-16kHz

4. Experimental setup

In this section, we will discuss the different configurations of the proposed models, details of the hyperparameters and the evaluation metrics used.

4.1. Training with different losses

To verify the efficacy of the label-aware contrastive loss we train the proposed model with different combinations of losses.
1) Training with label-aware contrastive loss (AS-LCL): This configuration involves utilizing the label-aware contrastive loss outlined in Section 3.
2) Training with supervised contrastive loss (AS-SCL): This setup utilizes the supervised contrastive loss from equation (1) instead of the label-aware contrastive loss.
3) Training with cross-entropy loss (AS-CEL): In this configuration, no contrastive loss is employed for the content encoder. Only frame-wise phoneme prediction is conducted, utilizing cross-entropy loss.

4.2. Model Configuration

Both the encoder and decoder employ a transformer encoder with 6 layers, each featuring single-head attention. Input and output dimensions are set to 386, with feedforward dimensions of 1536 and an attention dimension of 64. The parameter count for the proposed architecture is 45.25M. Adam optimizer is utilized with an initial learning rate of 0.0001. All proposed models are trained for 500 epochs and the best checkpoint is chosen. The hyperparameter α is set to 0.005, β is set to 0.5 and the temperature for contrastive loss, τ is set to 0.1. We train all our models with the PyTorch framework using a single GPU. All codes are open-sourced here[2] towards facilitating research and reproducibility.

4.3. Evaluation Metrics

To show the comparison of the different models proposed in this work we use 3 metrics namely, Speech Recognition (ASR) score, Mel-Cepstral distortion (MCD) and Mean Opinion Score (MOS). All the metrics are reported on the test set.

ASR: We report the Character Error Rate (CER) and Word Error Rate (WER) between ground truth text and transcripts of

[2] https://github.com/coding-phoenix-12/articulatory-synthesis.git

Table 1: *Table shows CER and WER of the synthesized audios for different models using 3 different ASR systems. 95% Confidence Interval is also reported in brackets*

Models	w2v2 with KenLM		w2v2 ASR		Google API	
	CER (%)	WER (%)	CER (%)	WER (%)	CER (%)	WER (%)
AS-B	13.43 (13.10 - 14.55)	23.94 (22.82 - 25.27)	16.02 (15.81 - 17.11)	40.04 (39.79 - 42.51)	28.40 (26.67 - 29.22)	48.78 (47.07 - 50.24)
AS-LCL	**12.95 (12.48 - 13.96)**	**22.55 (21.42 - 23.82)**	**15.37 (14.95 - 16.24)**	**38.35 (37.57 - 40.14)**	**26.08 (25.19 - 27.62)**	**46.47 (45.30 - 48.41)**
AS-CEL	13.62 (12.83 - 14.28)	24.18 (22.23 - 24.62)	15.87 (15.14 - 16.42)	39.59 (38.06 - 40.63)	27.25 (26.25 - 28.75)	47.42 (46.14 - 49.21)
AS-SCL	13.99 (13.58 - 15.10)	24.22 (23.26 - 25.76)	16.27 (15.91 - 17.26)	39.89 (39.25 - 41.92)	27.54 (25.91 - 28.37)	48.32 (46.53 - 49.61)

the synthesized audios. We extract the transcripts using 3 Automatic Speech Recognition (ASR) models. The first model uses a wav2vec2 [35] model fine-tuned on IndicTimit [36] along with a KenLM [37] language model which has seen sentences from the test set. The second ASR model uses only the wav2vec2 acoustic model without the language model. These models are trained using fairseq [38]. For the third ASR model, we use an Indian English ASR model from Google Chirp API to get transcription for synthesized audios and calculate CER and WER.

MCD: The mel-cepstral distortion is calculated between the synthesized audio and the ground truth audio.

MOS: 40 randomly picked audios from the test set, are scored by 10 validators on a scale of 1-5 based on how natural the audio sounds 1 being the worst score and 5 being the best.

4.4. Encoder Representations

To better understand the efficacy of the proposed method, we get the output of the encoder i.e. C and try to visualize them using t-SNE [39]. We select a few audios and get the encoder embeddings from the 3 models AS-CEL, AS-SCL and AS-LCL as we are trying to learn phoneme-dependent embeddings in these models, unlike the baseline. For visualization, we pick the embeddings that belong to the 3 phonemes, /ao/, /f/ and /k/ using the frame level alignment we obtain from the Kaldi toolkit.

5. Results and discussion

In this section, we discuss the results of our experiments using the evaluation metrics mentioned before and compare the different models.

5.1. CER/WER

We use ASR models to check the quality of the synthesized audio. Table 1 shows the scores with different ASR systems. The proposed AS-LCL gives a relative improvement in WER by 5.8 % when an ASR system consisting of both acoustic and language models is used. Note that the language model has seen sentences from the test set. We also calculate WER using only the acoustic model and see a similar relative improvement in WER by 5.9 %. To further verify the performance we use Google API to get the synthesized audio transcribed and calculate CER and WER and see a similar improvement in performance with the proposed model.

5.2. MCD and MOS

Table 2 provides the details of the MCD and MOS. We can see that when it comes to MCD, all the models are on par with the baseline with the AS-LCL model performing slightly better than the baseline. Similarly, the MOS score shows a similar trend of all the models being on par with the baseline with AS-LCL performing slightly better.

5.3. Visualization of the encoder representations

From Figure 3 we can see that the AS-LCL model that uses the label-aware contrastive loss formed better distinctly separated clusters in comparison to the other two models, specifically for the phonemes. This visualization supports the claim that the proposed loss function indeed does a better job at separating similar phonemes.

Table 2: *Table shows MCD and MOS scores of the synthesized audios for different models. Standard deviations for the MOS are mentioned in the brackets*

Models	MCD	MOS
AS-B	12.64	3.81 (1.12)
AS-LCL	**12.60**	**3.90 (1.08)**
AS-CEL	12.62	3.88 (1.10)
AS-SCL	12.74	3.79 (1.09)

6. Conclusions

In this work, we propose phonetic label-aware contrastive loss for articulatory speech synthesis. The proposed phonetic label-aware contrastive loss outperforms the baseline when it comes to the quality of the synthesized audio. The WER and CER obtained using different ASR models support this claim. Additionally, the MOS further supports the improvement in synthesis quality of the performance of the proposed models. Among the three configurations of the proposed models, the AS-LCL does better than AS-SCL and AS-CEL. We also show a visualization of the learnt representations which show that the AS-LCL model forms comparatively better clusters. This enforces the claim that label-aware contrastive loss indeed helps with learning better phonetic representations, especially with phonemes that share critical articulators, compared to supervised contrastive loss or cross-entropy loss. In the future, we plan to work more on articulatory synthesis, using more information regarding critical and non-critical articulators.

7. Acknowledgement

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

8. References

[1] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, *FastSpeech: fast, robust and controllable text to speech*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[2] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *International Conference on Learning Representations*, 2020.

[3] J. Kong, J. Park, B. Kim, J. Kim, D. Kong, and S. Kim, "VITS2: Improving Quality and Efficiency of Single-Stage Text-to-Speech with Adversarial Learning and Architecture Design," in *Proc. INTERSPEECH*, 2023, pp. 4374–4378.

[4] M. Le, A. Vyas, B. Shi, B. Karrer, L. Sari, R. Moritz, M. Williamson, V. Manohar, Y. Adi, J. Mahadeokar, and W.-N. Hsu, "Voicebox: Text-guided multilingual universal speech generation at scale," in *NeurIPS*, vol. 36, 2023, pp. 14 005–14 034.

[5] P. J. Jackson and V. D. Singampalli, "Statistical identification of critical, dependent and redundant articulators," *Journal of the Acoustical Society of America*, vol. 123, no. 5, pp. 3321–3321, 2008.

[6] Philip J.B. Jackson et al., "Statistical identification of articulation constraints in the production of speech," *Speech Communication*, vol. 51, no. 8, 2009.

[7] PK Anusuya et al., "A data driven phoneme-specific analysis of articulatory importance," in *International Seminar On Speech Production*, 2020.

[8] P. Birkholz, "Modeling consonant-vowel coarticulation for articulatory speech synthesis," *PloS one*, vol. 8, no. 4, p. e60603, 2013.

[9] K. N. Stevens, *Acoustic phonetics*. Cambridge: MIT Press, 1998.

[10] P. Birkholz, S. Kürbis, S. Stone, P. Häsner, R. Blandin, and M. Fleischer, "Printable 3d vocal tract shapes from mri data and their acoustic and aerodynamic properties." *Scientific Data*, vol. 7, no. 1, pp. 255–255, 2020.

[11] Arnela, Marc et al., "Influence of vocal tract geometry simplifications on the numerical simulation of vowel sounds," *The Journal of the Acoustical Society of America*, vol. 140, no. 3, pp. 1707–1718, 2016.

[12] V. Ribeiro, K. Isaieva, J. Leclere, P.-A. Vuissoz, and Y. Laprie, "Automatic generation of the complete vocal tract shape from the sequence of phonemes to be articulated," *Speech Communication*, vol. 141, pp. 1–13, 2022.

[13] A. Niebergall, S. Zhang, E. Kunay, G. Keydana, M. Job, M. Uecker, and J. Frahm, "Real-time MRI of speaking at a resolution of 33 ms: undersampled radial FLASH with nonlinear inverse reconstruction," *Magn Reson Med*, vol. 69, no. 2, pp. 477–485, Apr. 2012.

[14] S. Narayanan et al., "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (tc)," *The Journal of the Acoustical Society of America*, vol. 136, no. 3, pp. 1307–1311, 2014.

[15] S. Aryal and R. Gutierrez-Osuna, "Data driven articulatory synthesis with deep neural networks," *Computer Speech and Language*, vol. 36, no. C, pp. 260–273, 2016.

[16] F. Bocquelet, T. Hueber, L. Girin, P. Badin, and B. Yvert, "Robust articulatory speech synthesis using deep neural networks for bci applications," in *Interspeech 2014-15th Annual Conference of the International Speech Communication Association*, 2014.

[17] Z.-C. Liu, Z.-H. Ling, and L.-R. Dai, "Articulatory-to-Acoustic Conversion with Cascaded Prediction of Spectral and Excitation Features Using Neural Networks," in *Proc. Interspeech*, 2016, pp. 1502–1506.

[18] F. Taguchi and T. Kaburagi, "Articulatory-to-speech Conversion Using Bi-directional Long Short-term Memory," in *Proc. Interspeech*, 2018, pp. 2499–2503.

[19] Y.-W. Chen, K.-H. Hung, S.-Y. Chuang, J. Sherman, W.-C. Huang, X. Lu, and Y. Tsao, "Ema2s: An end-to-end multimodal articulatory-to-speech system," in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2021, pp. 1–5.

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

[21] P. Wu, T. Li, Y. Lu, Y. Zhang, J. Lian, A. W. Black, L. Goldstein, S. Watanabe, and G. K. Anumanchipalli, "Deep Speech Synthesis from MRI-Based Articulatory Representations," in *Proc. INTERSPEECH*, 2023, pp. 5132–5136.

[22] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.

[23] S. Udupa, S. C, and P. K. Ghosh, "Improved acoustic-to-articulatory inversion using representations from pretrained self-supervised learning models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[24] Y. M. Siriwardena, G. Sivaraman, and C. Espy-Wilson, "Acoustic-to-articulatory Speech Inversion with Multi-task Learning," in *Proc. Interspeech*, 2022, pp. 5020–5024.

[25] V. Suresh and D. Ong, "Not all negatives are equal: Label-aware contrastive loss for fine-grained text classification," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 4381–4394.

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

[27] "A multichannel articulatory speech database and its application for automatic speech recognition," in *Proc. 5th seminar on speech production: models and data*, 2000.

[28] "3d electromagnetic articulograph," available online: http://www.ar ticulograph.de/, last accessed: 4/2/2020.

[29] P. Zhu, L. Xie, and Y. Chen, "Articulatory movement prediction using deep bidirectional long short-term memory based recurrent neural networks and word/phone embeddings," in *INTERSPEECH*, 2015.

[30] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18 661–18 673, 2020.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[32] J. Bandekar, S. Udupa, and P. K. Ghosh, "Exploring a classification approach using quantised articulatory movements for acoustic to articulatory inversion," in *Proc. INTERSPEECH*, 2023, pp. 5147–5151.

[33] P. Daniel, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *ASRU*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[34] Mirco Ravanelli et al., "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[35] A. B. al, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *NeurIPS*, 2020.

[36] C. Yarra, R. Aggarwal, A. Rajpal, and P. K. Ghosh, "Indic timit and indic english lexicon: A speech database of indian speakers using timit stimuli and a lexicon from their mispronunciations," in *O-COCOSDA*. IEEE, 2019, pp. 1–6.

[37] K. Heafield, "Kenlm: Faster and smaller language model queries," in *Proceedings of the sixth workshop on statistical machine translation*, 2011, pp. 187–197.

[38] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

[39] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.