INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Exploring a classification approach using quantised articulatory movements for acoustic to articulatory inversion

Jesuraj Bandekar, Sathvik Udupa, Prasanta Kumar Ghosh

Electrical Engineering, Indian Institute of Science, Bangalore, India, 560012
jesurajbandekar.661@gmail.com, sathvikudupa66@gmail.com, prasantag@gmail.com

Abstract

Acoustic to articulatory inversion (AAI) is the task of predicting articulatory trajectories from speech acoustics. An AAI model is typically optimised with regression-based objective functions on continuous articulatory movement targets. In this work, we explore an alternate approach by classifying bins of quantised articulatory movements. We extend it by utilising ordinal regression, along with a novel approach involving KL Divergence loss between a target Gaussian posterior and the predicted one. We train transformer AAI models with MFCC and TERA acoustic features, with various quantisation types (uniform vs non-uniform) and bins. Using 16 subjects' acoustic-articulatory data, we evaluate the results with correlation coefficient (CC) and root mean squared error on unseen utterances from seen and unseen speakers. While the quantization type did not alter the AAI performance, we find that the highest CC (0.8838) is achieved with TERA features using ordinal regression, also with the proposed KL divergence loss at 64 quantisation bins, which is found to be on par with the CC (0.8856) using the regression-based approach. In fact, reducing the number of quantisation bins to 16 does not significantly change the AAI performance.

Index Terms: acoustic to articulatory inversion, electromagnetic articulography, sequence to sequence learning

1. Introduction

Articulatory movements from Electromagnetic Articulography (EMA) denote the positions of various speech articulators. These are highly dynamic, present at a high frequency and depend on the morphology of the speaker. It is also a one-to-many mapping, identical speech can be produced by different vocal tract configurations [1]. Knowledge of articulatory movements is beneficial to various speech tasks such as speech recognition [2, 3], speaker verification [4], speech synthesis [5] and so on. In the absence of direct articulatory movements, it can be estimated from speech, this process is known as acoustic to articulatory inversion (AAI).

Various methods have been utilised in the past to learn the AAI mapping such as codebooks [6, 7], Gaussian Mixture Models (GMM) [8, 9, 10], Hidden Markov Models (HMM) [11]. Different deep learning-based sequence to sequence learning approaches have also been used, such as Convolutional neural networks (CNN) [12, 13], Recurrent neural networks (RNN) [14, 15, 16], wavenet [17], transformers [18, 19, 20] etc. While the networks used to learn feature representation are an important design choice to improve the results, the objective functions to train the networks are also equally important. The loss landscape can differ for different objective functions, hence understanding it better is important. In AAI, Mean squared error loss is commonly used to learn the mapping. This is convenient since articulatory trajectories are smoothly varying series of data which can be learnt through regression. Additionally, there have been approaches which augment the learning by adding auxiliary losses as in [21, 22, 23, 24].

In this work, we explore a different formulation of this problem by quantising articulatory movement targets. In the past, articulatory movements have been processed as hypercube codebooks and the required parameters retrieved based on the acoustic entry [25, 26]. Here, we use the sequence-to-sequence learning methodology and replace the continuous articulatory targets with their quantised bins. Since this process introduces quantisation error, we experiment with varying quantisation bins. Moreover, each speech articulator has a different density and range of motion, so we investigate if non-uniform or variable quantisation is necessary.

With quantised articulatory targets, we can use cross entropy over predictions on each frame, and each articulator to learn the correct quantised bin to predict. With this formulation, it lacks information on the closeness of the bins, due to which the performance can diminish. In order to alleviate this, there are approaches involving ordinal regression/classification [27] which computes an expectation as well with an L1 loss, so that the combined objective function will have a measure of the range of predicted values. We further extend this with a simple alternative by constructing a Gaussian probability density function centred around the target quantised bin and using KL Divergence to enforce it on the predictions. Further, we also validate whether quantisation can provide additional useful measures, such as uncertainty in the prediction.

Some of the learning methodologies used in prior work such as deep mixture networks [28, 29, 30] are capable of evaluating prediction confidence. This is estimated with probability density over the predicted articulatory trajectories. Regression-based approaches are unable to provide such confidence measures, though they are able to estimate articulatory movements with better performance. By combining sequence-to-sequence learning and quantised target prediction, this line of work can be useful in measuring uncertainty in the prediction of articulatory movements.

Here are the contributions of our work,

- We explore the use of quantised articulatory movement targets for AAI and compare them with regression-based AAI
- We test out different extensions for quantisation and objective functions and report the results on models trained with two different input features - MFCC and TERA, and report on seen and unseen speaker performance.

We analyse the quantised posterior prediction to gain more in-

10.21437/Interspeech.2023-1593

sight into the uncertainty involved with AAI prediction. In the following sections, we first summarise the dataset, followed by the details involved in our proposed methodology, experimental setup and evaluation, ending with results and conclusions of our study.

2. Dataset

In this work, acoustic and articulatory data are collected simultaneously for 460 sentences from the MOCHA-TIMIT corpus [31]. The recordings from 10 subjects are used for training, comprising of 6 males and 4 females. Further, 6 subjects (3 Male, 3 Female) are taken for unseen subject evaluation. The speech is recorded using a microphone [32] and the articulatory movements are captured using Electromagnetic Articulagraph AG501 [33].

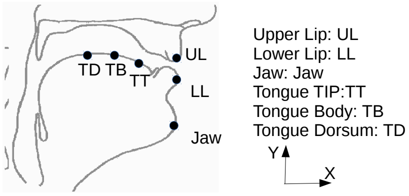

Figure 1: *Schematic diagram indicating the placement of EMA sensors [14]*

As shown in Figure 1, the data from 6 articulators are captured, resulting in 12 dimension articulatory trajectories - UL_x, UL_y , LL_x, LL_y , Jaw_x, Jaw_y , TT_x, TT_y , TB_x, TB_y , TD_x, TD_y. The training data covers around 4.3 hours of speech. To validate the generalisation of our methods, we report scores on unseen speakers as well.

3. Proposed methodology

3.1. Neural network architecture

We use a non-autoregressive encoder-only transformer neural network [34, 18]. For the baseline regression based AAI (r-AAI), the output features are projected to 12 dimensions to predict the articulatory trajectories (resulting in $N * 12$ prediction where N is the sequence length) and optimised with Mean Squared Error (MSE) loss. For quantised AAI prediction (q-AAI), the output features are projected to $12 * K$ dimensions where K is the number of bins available after quantisation (resulting in $N * 12 * K$ predictions). This is followed by softmax over each frame and articulator dimension and optimised as mentioned in Section 3.4. With q-AAI, the final articulatory movements are obtained after decoding the posterior, we have reported on argmax (greedy) decoding and decoding with expectation.

3.2. Acoustic features

We train with two types of input acoustic features - Mel Frequency Cepstral Coefficients (MFCC) and TERA [35]. We use MFCC since it has been consistently used in AAI literature, and TERA, since it has performed well in recent work [36].

3.3. Types of quantisation

We experiment with 3 types of quantisation.

3.3.1. Uniform constant range (UCR)

This is the default quantisation scheme where the data is uniformly quantised to K bins using the same lower and upper bounds for all articulators. Let K be the number of quantisation bins and $[a, b]$ be the quantisation range, such that $a = min\{EMA_{train}) - \Delta$ and $b = max\{EMA_{train}\} + \Delta$, where EMA_{train} is the set of all ground truth articulatory movements and Δ is a fixed offset to cover an additional range for unseen data. Thus, the quantisation process Q is a function of a, b and K, returning the quantised bin L for every value in every frame and every articulator in EMA ground truth data.

$$L = Q(EMA, a, b, K)$$

3.3.2. Uniform variable range (UVR)

This is the quantisation scheme where the data is uniformly quantised to K bins using articulatory specific lower and upper bounds. The procedure is the same as the previous one except that a and b are determined for each articulator. This is because the range of movements of each articulator is different and there might be advantages in modelling it separately.

3.3.3. Non-Uniform constant range (NUCR)

This is the quantisation scheme where the data is non-uniformly quantised to K bins using the bins obtained after Llyod-max quantisation[1]. Here, we use global a and b for all articulators.

3.4. Optimisation types

In q-AAI, the neural networks predict $12 * K$ logits for each frame, on which softmax is applied, with K quantisation bins. This acts as the posteriors which have to be optimised. Let N be the sequence length and M the number of articulatory features. Let F_{ij} are the logit and L_{ij} are the true quantisation bin for that frame i and articulatory j.

3.4.1. Cross Entropy loss (CE)

A straightforward approach is to apply cross entropy over each frame and each articulator to classify the correct quantisation bin.

$$Loss_{CE} = \frac{1}{N * M} \sum_{i=1}^{N} \sum_{j=1}^{M} CrossEntropy(F_{ij}, L_{ij})$$

3.4.2. Expectation loss (EXP)

We further extend it by applying ordinal regression as used in [27], which calculates the deviation from the expectation

$$Loss_{exp} = \frac{1}{N * M} \sum_{i=1}^{N} \sum_{j=1}^{M} SmoothL1(L_{ij}, \sum_{k=1}^{K} P(k)_{ij} \cdot k)$$

Where $SmoothL1$ is the smooth L1 loss [2] which is a mixture of L1 and MSE loss. This loss can be directly combined with cross-entropy such that,

$$Loss = Loss_{CE} + Loss_{exp}$$

With this approach, the output can be decoded with the expectation $\sum_{k=1}^{K} P(k)_{ij} \cdot k$

[1] https://www.mathworks.com/help/comm/ref/lloyds.html

[2] https://pytorch.org/docs/stable/generated/torch.nn.SmoothL1Loss.html

Table 1: *This table represents the correlation coefficient (CC) for Uniform constant range (UCR), Uniform variable range (UVR), and Non-Uniform constant range (NUCR) quantisation types, with different objective functions. The standard deviation across articulators is shown in brackets. The results for decoding with argmax and expectation are shown. All models are trained with 64-bin quantised articulatory movement targets. The baseline r-AAI results are CC of* **0.8581***(0.056) for MFCC and* **0.8856***(0.048) for TERA.*

		CE		CE+EXP		EXP		CE+KL		KL	
		argmax	EXP	argmax	EXP	argmax	EXP	argmax	EXP	argmax	EXP
UCR	MFCC	0.8335 (0.07)	-	0.8416 (0.063)	0.8534 (0.057)	0.5717 (0.074)	0.8577 (0.057)	0.8357 (0.065)	0.8484 (0.06)	0.8577 (0.055)	0.8601 (0.054)
	TERA	0.8633 (0.061)	-	0.8677 (0.056)	0.8786 (0.051)	0.6148 (0.091)	**0.8839** (0.051)	0.8626 (0.059)	0.8743 (0.054)	**0.8804** (0.049)	**0.8828** (0.049)
UVR	MFCC	0.8332 (0.065)	-	0.8367 (0.064)	0.8482 (0.059)	0.6111 (0.117)	0.8544 (0.058)	0.8406 (0.065)	0.8513 (0.06)	0.8491 (0.059)	0.8522 (0.059)
	TERA	0.8614 (0.059)	-	0.869 (0.057)	0.8784 (0.052)	0.6466 (0.098)	**0.8805** (0.051)	0.8612 (0.06)	0.8724 (0.055)	0.8724 (0.054)	0.8759 (0.054)
NUCR	MFCC	0.8388 (0.067)	-	0.842 (0.063)	0.8547 (0.057)	0.7248 (0.07)	0.854 (0.059)	0.8341 (0.069)	0.846 (0.064)	0.8499 (0.062)	0.8487 (0.062)
	TERA	0.8599 (0.061)	-	0.8677 (0.057)	0.8782 (0.052)	0.7531 (0.063)	**0.8805** (0.052)	0.8603 (0.061)	0.8706 (0.056)	0.877 (0.054)	0.876 (0.054)

3.4.3. KL loss

We define Gaussian probability density functions over quantisation bins, parameterised by the ground truth quantised articulatory movement as mean and the standard deviation as a hyperparameter, as shown below -

$$f(x)_{ij} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-L_{ij}}{\sigma})^2}$$

This formulation defines Gaussian centred around every quantised articulatory movement with σ as a hyperparameter, we use $\sigma = 4$. This is evaluated at $x = k$ where $k = (1, 2, 3, .., K)$. This is repeated for all $i = (1, 2, 3, .., N)$ and $j = (1, 2, 3, .., M)$. Let $P(x)_{ij}$ be the distribution from softmax on neural network output. We compute the Kullback-Leibler divergence (KLD) objective, with pointwise KLD in pytorch[3].

$$Loss_{KL} = \frac{1}{N*M*K} \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=1}^{K} f(k)_{ij} (log f(k)_{ij} - log P(k)_{ij})$$

This loss can be combined with ordinal or cross-entropy. We share an efficient implementation of this loss as part of the codebase, with which the training time is comparable to r-AAI baseline.

4. Experimental setup

We train separate models with MFCC (13-dimensional) and TERA (features extracted from the last layer) as input acoustic features. MFCC is extracted with librosa [4] and TERA is extracted with s3prl [5]. The input features as well as the target articulatory movements are pre-processed to be at 100Hz. Articulatory movements are standardised for each utterance, separately for each articulator. Zero padding is performed at the batch level and padding values are masked for transformer self-attention to train the AAI models. We use adam optimizer with a learning rate of 0.0001, with early stopping based on validation loss.

Table 2: *This table represents the correlation coefficient (CC) and root mean squared error (RMSE) for MFCC and TERA input features, for varying quantisation bins for models with KL configuration for UCR.*

		8	16	32	64	128
MFCC	CC	0.8047 (0.07)	0.8534 (0.056)	0.8546 (0.058)	0.8601 (0.054)	0.8505 (0.056)
	RMSE	1.845 (0.0590)	1.203 (0.235)	1.193 (0.271)	1.155 (0.207)	1.163 (0.204)
TERA	CC	0.8443 (0.059)	0.8799 (0.051)	0.8829 (0.049)	0.8828 (0.049)	0.8742 (0.05)
	RMSE	1.819 (0.060)	1.132 (0.241)	1.1308 (0.285)	1.090 (0.223)	1.1109 (0.211)

We use data from 10 subjects for training AAI models. 80% of the sentences for each subject are used for training, 10% for validation and 10% for testing. This consists of seen speaker evaluation. Further, we also evaluate on 10% of unseen sentences for 6 unseen speakers. We evaluate our models with correlation coefficient (CC)[37] and also report root mean squared error (RMSE). All models are trained using pytorch on one NVIDIA RTX 2080ti GPU. All the codes will be made available publicly.

In the next section, we will be showing the following results

- Use r-AAI models as the baseline with TERA and MFCC as input
- Train q-AAI models in various configurations - varying quantisation bins, quantisation types, and objective functions for classification, and validate if it performs as well as r-AAI.
- Report correlation coefficient for all different configurations, for seen and unseen speakers
- Visualise the posterior of different objective functions to demonstrate additional functionality of this approach.

5. Results

5.1. Quantisation types and objective functions

Table 1 shows the correlation coefficient for different types of quantisation and objective functions. Compared to r-AAI, we find that there is a reduction in performance using cross entropy objective, for both MFCC and TERA. Further, we see that by

[3] https://pytorch.org/docs/stable/generated/torch.nn.KLDivLoss.html

[4] https://librosa.org/doc/main/generated/librosa.feature.mfcc.html

[5] https://github.com/s3prl/s3prl

applying the ordinal classification approaches (EXP and KL), the metrics are on par with r-AAI. With EXP ordinal classification, we find that decoding with expectation leads to better performance than argmax decoding. This is expected since the expectation objective lets the logits shift away from argmax outputs, making the posterior asymmetric around the final prediction. This is also brought up in the work which uses EXP loss [27]. We then observe that our alternative - KL based ordinal loss also performs well. This has the additional benefit of the posterior scores centring around the argmax prediction.

5.2. Effect of number of quantisation bins

For Table 2, we show the results for models trained with KL loss. The table shows various quantisation bins being used ranging from 8 (3 bits) to 128 (7 bits). We find that there is a degradation in CC with 8 bins. On the other hand, the results are fairly consistent for both MFCC and TERA number of bins $\geq$ 16. We also report the RMSE for these models (they are not reported for all configurations due to space constraints).

5.3. Unseen speaker performance

Evaluating AAI models on unseen speakers is important to ensure the generalisability of AAI. In table 3, we report r-AAI as well as q-AAI models in different configurations, trained 64 bins. We find that the results are on a similar range of baseline for unseen speakers as well.

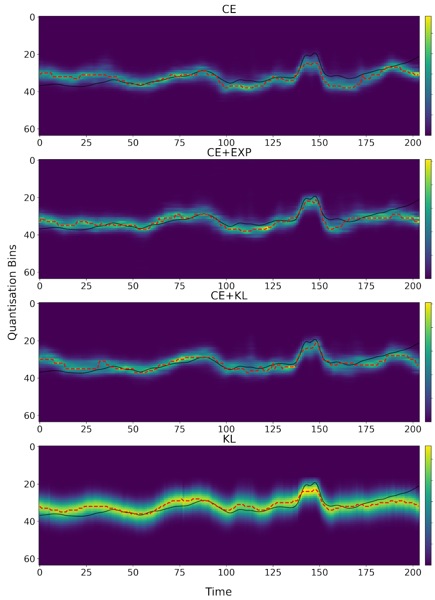

Figure 2: *Figure shows the posterior values from models (TERA input,* 64 *quantisation bins, UCR) with different objective functions, for LL_y trajectory for a seen male speaker from test utterance. The black line corresponds to the ground truth articulatory trajectory and the red line corresponds to the argmax predicted output. Note that the ground truth articulatory movement is shifted to be shown on the posterior values.*

5.4. Evaluating posterior heatmaps

Figure 2 shows the posterior values (output after softmax) as heatmaps of different models. Each output is of shape $64 * 210$ with 64 quantisation bins and 210 frames. Note that the intensity range of each model is different, denoted on the right. We observe that for models with CE and KL, there are few regions with high probability, apart from which the probability mass is quite evenly distributed. This is not ideal since it doesn't reflect confident predictions. On the other hand, for KL model, we see that there is a relatively higher value over the argmax prediction, with the mass decaying on moving away from argmax. This is due to applying a fixed Gaussian over the labels. This approach shows promise on identifying uncertainty, such as training with learnable variance. We will expand on these in our future works.

Table 3: *This table represents the correlation coefficient (CC) for 6 unseen speakers, for CE, EXP and KL configurations (UCR), with MFCC and TERA features. The r-AAI baseline is also reported. We find that both EXP and KL based classification approaches are on par with r-AAI in unseen speaker evaluation.*

		MFCC	TERA
baseline		0.7041(0.114)	0.7546(0.103)
CE	argmax	0.7048(0.114)	0.7402(0.099)
CE + EXP	argmax	0.6975(0.113)	0.7389(0.103)
	EXP	0.7087(0.109)	0.7479(0.102)
KL	argmax	0.7062(0.107)	0.7494(0.101)
	EXP	0.7093(0.107)	0.7523(0.102)

5.5. Smoothing

Since we estimate movements with quantised bins, a question arises whether the predicted trajectories are smooth enough. The earlier methods of AAI with codebooks required post-processing of smoothing. To test this, we use different post-processing such as quadratic and Kalman smoothing and find that there is no improvement in CC or any noticeable difference in the trajectories for 64 bin models. Hence, we report all results without any post-processing.

6. Conclusions

In this work, we view AAI task as a classification of bins corresponding to quantised articulatory movements. We test this approach with various quantisation methods, quantisation bits and objective functions. We compare with AAI baseline of mean squared error objective and report the scores for seen and unseen speakers. We find that the results of q-AAI are consistent with varying quantisation schemes. The default classification method has a drop in performance. This degradation can be removed by involving ordinal classification, where we also propose a novel training scheme with Gaussian labels, optimised with KL Divergence. We experiment with a varying number of quantisation bins and observe that 6 bits of EMA are sufficient to reach baseline AAI results. For all configurations, we report the correlation coefficient on MFCC and TERA input features and find that with ordinal classification approaches, the performance is on par with baseline AAI. Further, we show that the quantisation posterior could act as a good representation of the uncertainty involved in articulatory movement prediction in AAI. While this approach does not provide any further benefit in terms of performance when compared to the baseline, this formulation shows promise across various new methods that can be applicable to AAI. We shall cover these in our future works.

7. References

[1] B. Lindblom and J. F. Lubker, "Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation," *Journal of Phonetics*, vol. 7, pp. 147–161, 1979.

[2] P. K. Ghosh and S. Narayanan, "Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion," *The Journal of the Acoustical Society of America*, vol. 130, no. 4, pp. EL251–EL257, 2011.

[3] S. Hu, S. Liu, X. Xie, M. Geng, T. Wang, S. Hu, M. Cui, X. Liu, and H. Meng, "Exploiting cross domain acoustic-to-articulatory inverted features for disordered speech recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6747–6751.

[4] M. Li, J. Kim, A. Lammert, P. K. Ghosh, V. Ramanarayanan, and S. Narayanan, "Speaker verification based on the fusion of speech acoustics and inverted articulatory signals," *Computer speech & language*, vol. 36, pp. 196–211, 2016.

[5] Z.-H. Ling, K. Richmond, J. Yamagishi, and R.-H. Wang, "Integrating articulatory features into HMM-based parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1171–1185, 2009.

[6] J. N. Larar, J. Schroeter, and M. M. Sondhi, "Vector quantization of the articulatory space," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 36, pp. 1812–1818, 1988.

[7] B. Atal, J. J. Chang, M. V. Mathews, and J. W. Tukey, "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique." *The Journal of the Acoustical Society of America*, vol. 63 5, pp. 1535–53, 1978.

[8] K. Richmond, "Mixture density networks, human articulatory data and acoustic-to-articulatory inversion of continuous speech." in *https://era.ed.ac.uk/handle/1842/1143*, 2001.

[9] T. Toda, A. W. Black, and K. Tokuda, "Acoustic-to-articulatory inversion mapping with gaussian mixture model," in *INTERSPEECH*, 2004.

[10] ——, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," *Speech Communication*, vol. 50, no. 3, pp. 215–227, 2008.

[11] L. Zhang and S. Renals, "Acoustic-articulatory modeling with the trajectory HMM," *IEEE Signal Processing Letters*, vol. 15, pp. 245–248, 2008.

[12] A. Illa and P. K. Ghosh, "Representation learning using convolution neural network for acoustic-to-articulatory inversion," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5931–5935.

[13] A. S. Shahrebabaki, S. M. Siniscalchi, G. Salvi, and T. Svendsen, "Sequence-to-sequence articulatory inversion through time convolution of sub-band frequency signals," *database*, vol. 1, p. 5, 2020.

[14] A. Illa and P. K. Ghosh, "Low resource acoustic-to-articulatory inversion using bi-directional long short term memory," in *Proc. Interspeech*, 2018.

[15] P. Liu, Q. Yu, Z. Wu, S. Kang, H. Meng, and L. Cai, "A deep recurrent approach for acoustic-to-articulatory inversion," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4450–4454.

[16] 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 preprint arXiv:2205.13086*, 2022.

[17] N. Bozorg and M. T. Johnson, "Acoustic-to-articulatory inversion with deep autoregressive articulatory-wavenet," *Networks (CNNs)*, vol. 22, p. 23, 2020.

[18] S. Udupa, A. Roy, A. Singh, A. Illa, and P. K. Ghosh, "Estimating articulatory movements in speech production with transformer networks," *arXiv preprint arXiv:2104.05017*, 2021.

[19] S. Udupa, A. Illa, and P. K. Ghosh, "Streaming model for acoustic to articulatory inversion with transformer networks," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2022. International Speech Communication Association, 2022, pp. 625–629.

[20] S. Udupa, P. K. Ghosh *et al.*, "Improved acoustic-to-articulatory inversion using representations from pretrained self-supervised learning models," *arXiv preprint arXiv:2210.16871*, 2022.

[21] X. Xie, X. Liu, and L. Wang, "Deep neural network based acoustic-to-articulatory inversion using phone sequence information." in *Interspeech*, 2016, pp. 1497–1501.

[22] A. S. Shahrebabaki, G. Salvi, T. Svendsen, and S. M. Siniscalchi, "Acoustic-to-articulatory mapping with joint optimization of deep speech enhancement and articulatory inversion models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 135–147, 2021.

[23] Y. M. Siriwardena, G. Sivaraman, and C. Espy-Wilson, "Acoustic-to-articulatory speech inversion with multi-task learning," *arXiv preprint arXiv:2205.13755*, 2022.

[24] 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," *arXiv preprint arXiv:2302.06774*, 2023.

[25] S. Ouni and Y. Laprie, "Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion," *The Journal of the Acoustical Society of America*, vol. 118, no. 1, pp. 444–460, 2005.

[26] S. Panchapagesan and A. Alwan, "A study of acoustic-to-articulatory inversion of speech by analysis-by-synthesis using chain matrices and the maeda articulatory model," *The Journal of the Acoustical Society of America*, vol. 129, no. 4, pp. 2144–2162, 2011.

[27] Z. Qin, P. Zhang, and X. Li, "Ultra fast deep lane detection with hybrid anchor driven ordinal classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–14, 2022.

[28] K. Richmond, S. King, and P. Taylor, "Modelling the uncertainty in recovering articulation from acoustics," *Computer Speech & Language*, vol. 17, no. 2-3, pp. 153–172, 2003.

[29] K. Richmond, "A trajectory mixture density network for the acoustic-articulatory inversion mapping," in *Ninth International Conference on Spoken Language Processing*, 2006.

[30] B. Uria, I. Murray, S. Renals, and K. Richmond, "Deep architectures for articulatory inversion," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[31] A. Wrench, "MOCHA-TIMIT," Department of Speech and Language Sciences, Queen Margaret University College, Edinburgh, speech database, 1999. [Online]. Available: http://sls.qmuc.ac.uk

[32] "EM9600 shotgun microphone," avaliable online: http://www.tbone-mics.com/en/product/information/details/the-tbone-em-9600-richtrohr-mikrofon/, last accessed:4/2/2020.

[33] "3d electromagnetic articulograph," available online: http://www.articulograph.de/, last accessed: 4/2/2020.

[34] 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.

[35] A. T. Liu, S.-W. Li, and H.-y. Lee, "Tera: Self-supervised learning of transformer encoder representation for speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2351–2366, 2021.

[36] S. Udupa, P. K. Ghosh *et al.*, "Improved acoustic-to-articulatory inversion using representations from pretrained self-supervised learning models," *arXiv preprint arXiv:2210.16871*, 2022.

[37] P. K. Ghosh and S. Narayanan, "A generalized smoothness criterion for acoustic-to-articulatory inversion," *The Journal of the Acoustical Society of America*, vol. 128, no. 4, pp. 2162–2172, 2010.