Interspeech 2024
1-5 September 2024, Kos, Greece

Highly Intelligible Speaker-Independent Articulatory Synthesis

Charles McGhee, Kate Knill, Mark Gales

ALTA Institute, Department of Engineering, University of Cambridge, UK

cgm43@cam.ac.uk, kmk1001@cam.ac.uk, mjfg100@cam.ac.uk

Abstract

An articulatory synthesiser which could accurately map vocal tract features to speech would enable novel evaluation of acoustic-to-articulatory inversion models beyond the small, typically monolingual, articulatory datasets available. However, current deep articulatory synthesisers and physical simulation-based synthesisers struggle to produce consistently intelligible speech, with Word Error Rates (WER) of around 20% for real or hand-crafted articulatory input. Additionally, deep learning methods have often only achieved this level of intelligibility when training and evaluating on the same speaker (speaker-dependent training). In this paper, we create a highly intelligible (WER $\sim$7% for real data and $\sim$10% for synthetic), speaker-independent articulatory synthesiser by training a deep synthesiser on a combination of high-quality real data and synthetic data generated by inversion. We then perform a multilingual evaluation of the joint inversion-synthesis system.

Index Terms: Articulatory Synthesis, Acoustic-to-Articulatory Inversion, Synthetic Data

1. Introduction

Acoustic-to-Articulatory Inversion (AAI) deconstructs the speech signal into the movements of the articulators which produced that speech. A perfect AAI model would provide an interpretable, universal speech encoding, with applications in speech-related machine learning fields, as well as areas such as second-language education and speech therapy. However, it can be difficult to evaluate the output of an AAI model. Typical evaluation criteria include some form of distance-based error between the predicted and ground truth articulatory features. This form of evaluation requires paired speech and articulatory data, but datasets in this area are often small and monolingual. There can also be large methodological differences in the collection of articulatory data [1], making cross-dataset comparisons difficult.

One alternative to distance-based comparisons would be to use the articulatory features as input for a task with datasets that span the domain of interest, such as phone recognition [2, 3]. The output features could also be analysed for specific speech inputs, such as rhotic and derhotic /ɹ/ [4]. Articulatory synthesis allows for both forms of evaluation, as we can analyse the output speech across a wide range of speakers and conditions, and we can analyse whether specific speech inputs are matched in the synthetic output. However, current articulatory synthesisers struggle to produce consistently intelligible speech [5, 6], with Word Error Rates (WERs) of around 20% for real or handcrafted articulatory input. This makes it difficult to use synthesis as an analytic tool for AAI as errors in the output speech may result from inaccurate synthesis rather than inaccurate inversion. Deep learning methods have also typically achieved this level of intelligibility by training and evaluating on the same speaker (speaker-dependent training) [6], which is insufficient for multispeaker, multilingual analysis of AAI. In this paper, we train a highly intelligible, speaker-independent deep articulatory synthesiser by pretraining the synthesiser on a large quantity of synthetic articulatory data (created through AAI) and then fine-tuning that model on high-quality real data. We then use this synthesiser to produce a novel multilingual evaluation of a joint inversion-synthesis system.

This research is funded by an EPSRC DTP and the Vice Chancellors Award. It is supported by Cambridge University Press & Assessment, a department of The Chancellor, Masters, and Scholars of the University of Cambridge.

2. Background and Related Works

An articulatory synthesiser uses an explicit model of the vocal tract and source parameters (e.g. describing frication) to synthesise speech. Approaches to articulatory synthesis can be broadly categorised into those where the synthesis is based on physical simulations of the vocal tract/source [7, 4] and those where source and vocal tract features are mapped to speech with deep neural networks [6, 8]. Physical simulation-based approaches to articulatory synthesis use the source-filter model to synthesise speech. Under this model the output speech signal is obtained by convolving an excitation signal which represents the source of the speech sound with the impulse response of the vocal tract transfer function [9]. Elaborate 3D models of the vocal tract geometry provide an interpretable way to design vocal tract transfer functions [10], however the intelligibility of these systems on connected speech can still be limited [5]. This may be due to the difficulty of modeling co-articulation across arbitrary phonetic contexts [7].

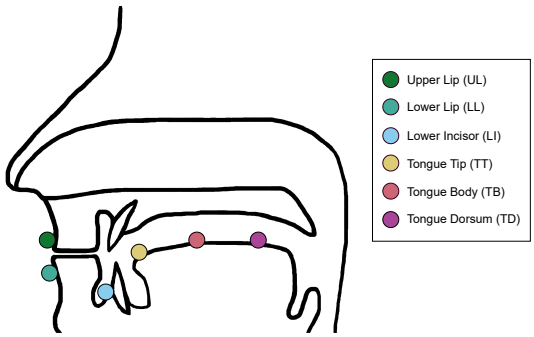

Figure 1: *Common sensor positions for EMA in the mid-sagittal plane.*

10.21437/Interspeech.2024-1160

Deep-learning methods by contrast learn a general mapping from articulatory and source features to speech. Neural networks can learn the non-linear dynamics involved in speech production [6], however they require paired speech and articulatory data to do so, which is limited in terms of quantity and quality. As detailed in [1], there isn't a standard technique for sensor placement in techniques such as Electromagnetic Articulography (EMA, see Figure 1 for common sensor placements), affecting speaker normalisation and cross-dataset comparison. Additionally, EMA sensors are attached to measurement equipment which is external to the mouth and can significantly alter the speech of some speakers [11]. For this reason, we use a different form of articulatory measurement, described in Section 3.1, for our training set. Articulatory datasets are also small, often containing less than 10 hours of paired speech and articulatory data in total. This forms the motivation for our use of synthetic data created through inversion.

We also need to find appropriate parameterisations of the source that complement our selected articulatory features. If we choose articulatory features which are similar to those found in EMA datasets, we need source features which can tell us whether speech is occurring (as opposed to another oral act, like swallowing), whether voicing is occurring and the fundamental frequency (F0) of that voiced sound (for intonation), and whether the sound is nasal as we have no velum closure information. In this paper, we opt for logarithmic F0 and energy as our source features, recognising that we will not have explicit velar information, but hoping the distinction between nasals and other voiced consonants will be captured by subtle differences in the source/articulatory features.

Recent deep articulatory synthesisers can be broken down into two sets of models, namely cascaded and end-to-end synthesisers. A cascaded articulatory synthesiser consists of two models. The first model uses articulatory and source features to predict an intermediate audio representation, such as mel spectrogram [8]. This intermediate representation is then fed into a vocoder which is trained separately to map the intermediate representation to a waveform. An end-to-end synthesiser would consist of a vocoder trained from scratch to accept the articulatory and source features as input [6]. If a pretrained vocoder is used, cascaded synthesisers can be faster and easier to train due to only needing to learn the comparatively simpler articulatory/source to intermediate feature mapping. In this paper, we use a cascaded style model for this reason.

3. Data

3.1. Articulatory Data

Wisconsin X-ray Microbeam Dataset

The main articulatory dataset we have chosen for our experiments is the Wisconsin X-ray microbeam dataset (XRMB) [12], distributed by UC Berkeley[1]. In XRMB, pellets glued to the articulators (in similar places to those shown in Figure 1) are tracked by an x-ray beam. Speaker-independent inversion performance on this dataset is usually higher than for EMA datasets [13], which may be the result of a more rigorous pellet placement methodology, or the fact that the pellets do not need to be attached to an external machine like the sensors in EMA We use x (anterior to posterior) and y (inferior to superior) coordinates measured from pellets on the first three tongue placements (T1-3), the upper lip, lower lip, and central incisor, roughly corresponding to the position of the EMA sensors in Figure 1. We use Point-Based (PB) features rather than Tract Variables (TV) [14] as we found minimal difference between the two sets in terms of inversion performance and PB features can be more easily visualised for applications like second-language education [15]. Our target features for inversion and input features to synthesis include the Δ and $\Delta\Delta$ of these PB features. All PB features are downsampled to 50Hz to match our inversion model output and low-pass filtered with a cutoff of 12Hz matching previous findings [16]. The PB features are then speaker-normalised to have zero mean and unit variance.

From the recordings, we select any tasks which involve speech, including consonant-vowel tasks, but excluding tasks such as swallowing. The recordings are segmented using the Montreal Forced Aligner (MFA) [17], by aligning the reference text to the speech, splitting utterances by silence and removing any sections with mistracked pellets. This leaves us with around 6 hours of paired articulatory and speech data across 48 speakers. We clean the audio samples by passing them through a deep speech enhancement model [18, 19, 20] [2] and then performing noise reduction with the `noisereduce` Python package [21, 22]. We split the speakers 40/8 into training and validation sets, giving the speakers included in the validation set in Table 1.

Table 1: *XRMB dataset statistics and validation set speakers.*

Num Speakers Train/Validation	Duration (hrs) Train/Validation	Validation Set Speakers
40/8	5.0/1.1	25, 26, 29, 33 40, 51, 56, 59

Haskins Production Rate Corpus

As a test set for inversion/synthesis we use the EMA-based Haskins Production Rate Corpus (HPRC). We expect that model performance will be significantly worse for HPRC than the validation set of XRMB, due to the mismatch in measurement equipment used and sensor/pellet placement methodologies. However, it is useful to determine whether or not our articulatory features and speech are well aligned, and we can compare our results with other State of the Art (SOTA) inversion methods that have used these two datasets [13]. We use the x/y coordinates of the sensors shown in Figure 1, i.e. TT, TB, TD, UL, LL and LI. The same speech enhancement of the HPRC audio data is performed as we did for XRMB. Dataset statistics, including which speakers were used in the test set, are given in Table 2.

Table 2: *HPRC dataset statistics and test set speakers.*

Num Speakers	Duration (hrs)	Test Set Speakers
4	3.9	F02, M01, M02, F03

3.2. Monolingual Speech Data

We intend to generate a large amount of synthetic articulatory data which can then be used to pre-train our synthesiser with ground truth source values. One dataset with an abundance of both clean audio and speakers is LibriTTS [23]. We opt to use both the train-clean-100/360 subsets to pretrain our synthesiser giving us $\sim$245 hours of audio and 1151 speakers to train on. All the audio goes through the same enhancement process as XRMB and HPRC.

[1] https://github.com/rsprouse/xray_microbeam_database

[2] https://huggingface.co/speechbrain/sepformer-dns4-16k-enhancement

3.3. Multilingual Speech Data

For our multilingual evaluation in Section 6, we select the English, German, Dutch, French, Italian, Korean and Japanese test sets from FLEURS [24]. We choose these languages as they are both high-resource, so Automatic Speech Recognition (ASR) performance should be high for the original speech, and they cover a wide range of phonological features. The English test set also gives us a good baseline to compare synthesis with real articulatory features from XRMB against synthesis with synthetic features from inversion.

4. Articulatory Inversion

Self-Supervised Learning (SSL) speech models, such as WavLM [25], have been shown to produce features that can be easily adapted to the inversion task [26]. However, to the best of the authors' knowledge, these sorts of models have not been directly adapted to perform inversion through fine-tuning. We explore fine-tuning a WavLM Large model in two settings, firstly fine-tuning all the weights of the model and secondly using parameter efficient fine-tuning in the form of Low Rank Adaptation (LoRA) [27]. The baseline model used is an adaptation of the model from [3] where features from the 10th layer of WavLM Large are extracted and then passed through a series of Bi-directional Gated Recurrent Units (GRUs). All models are trained to minimise the Mean Squared Error (MSE) between the predicted and ground truth PB values. We use a 1 cycle learning rate scheduler [28] combined with a learning rate range finder to determine the learning rate for each model and a batch size of 8. For the LoRA hyperparameters, we sweep ranks, r, (16, 32, 64, 128) and alpha (0.5r, r, 2r) values using the XRMB validation set, finding a combination of r=32, alpha=16 to perform the best. In Table 3, we report each model's performance on the XRMB validation and HPRC test set in terms of correlation and MSE between the predicted and ground truth PB values, not including the Δ and $\Delta\Delta$ values.

Table 3: *XRMB validation and HPRC test set results for each model, FT=Fine-Tuned*

	XRMB		HPRC	
Model	**Corr** $\uparrow$	**MSE** $\downarrow$	**Corr**	**MSE**
Baseline [3]	0.858	0.264	0.734	0.472
Full FT	**0.867**	0.257	0.738	0.468
LoRA FT	0.865	**0.251**	**0.754**	**0.435**

Both fine-tuned models outperform the baseline. Fully fine-tuning the model had comparable performance to LoRA fine-tuning on the validation set, but was noticeably worse on the test set, indicating the model may have overfit the XRMB training set. The LoRA fine-tuned model has similar performance to other inversion models on these datasets [13], although we cannot make a direct comparison as [13] used TV instead of PB features and did not detail which speakers were used for evaluation. We use the LoRA FT model to generate synthetic articulatory features for LibriTTS and as the inversion model in the system shown in Figure 2.

5. Articulatory Synthesis

We follow [8] and use a cascaded system with the same 8-layer Conformer [29] model used to map the combined source and articulatory features to mel-spectrogram. We then use a pretrained BigVGAN [30] base 22kHz vocoder to map the generated mel-spectrogram to a waveform. Our combined source/articulatory input consists of PB features, their Δ and $\Delta\Delta$, log F0 (extracted using CREPE [31]) and log energy, totalling 38 features in all. These features are interpolated to match the sampling rate of the mel spectrogram input to the BigVGAN vocoder. The Conformer synthesis model is then trained to minimise the L1 distance between the generated and ground truth spectrogram. The same learning rate scheduler and finder procedure used for our inversion experiments is used for synthesis training with a batch size of 16. When pretraining, we train on the LibriTTS data for 3 epochs, before fine-tuning for 10 epochs on the ground truth data from XRMB. We also report results for the same Conformer model trained solely on XRMB for 10 epochs. To evaluate intelligibility, we use the Whisper [32] large-v3 model to transcribe the synthetic audio and score the synthetic audio transcription against the reference transcript. In Table 4, we show the results on XRMB validation set and HPRC test set for training solely on XRMB versus pretraining on synthetic LibriTTS data.

Table 4: *% WER results on XRMB validation set and HPRC test set for different pretraining settings.*

Pretraining	**XRMB**	**HPRC**
No	7.8	47.1
Yes	**6.6**	**39.8**

Pretraining the model on synthetic articulatory data and real source features reduced WER on both XRMB and HPRC by 15%. We can see a similar drop in synthesis performance as we saw for inversion when using the mismatched articulatory data from HPRC. Therefore, the fact that pretraining improved model performance on both XRMB and HPRC indicates that the improvement likely occurs as a result of greater exposure to source features across a large number of speakers. On our samples page [3] we give a number of samples from the XRMB and HPRC sets.

To dig deeper into the kinds of errors being made by our pretrained synthesiser, we fine-tune a WavLM Large model to perform phone recognition using TIMIT [33], achieving a Phone Error Rate (PER) of 7.8% on the standard TIMIT test set. We then generate a phonetic transcript for the synthetic and original speech in the XRMB validation set and score the synthetic against the original, giving an overall PER of 11.5%. In Table 5 we give the top 3 most common vowel and consonant substitution errors.

Table 5: *Top 3 vowel and consonant substitutions for the XRMB validation set.*

Rank	**Vowels**	**Consonants**
1	[ɪ]/[ə]	[s]/[z]
2	[i]/[ɪ]	[ŋ]/[n]
3	[ə]/[ɪ]	[d]/[n]

We can see that the most common substitutions occur for vowels which are fairly close together in terms of place and degree of constriction. As our source features will be comparable across vowels, it is likely these errors occur due to imperfect articulatory feature normalisation or modeling. For consonants

[3] insynthart.github.io

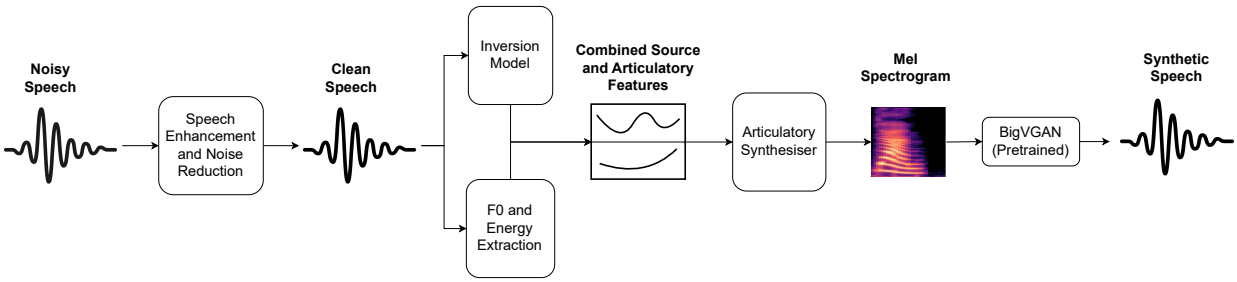

Figure 2: *Full inversion-synthesis system for re-synthesising any input speech.*

we see errors in voicing and nasality in the most common substitutions. Errors in voicing were the most common error on the TIMIT test set, however errors in nasality are likely a result of our current synthesiser input. In future iterations of this synthesiser, we may look to incorporate features such as nasalance [34], so as to improve nasal consonant synthesis. We use the pretrained synthesiser in the next section to perform a multilingual evaluation of the combined inversion-synthesis system.

6. Multilingual Evaluation

We use the inversion-synthesis pipeline in Figure 2 to re-synthesise speech for the FLEURS test sets. As above, we use Whisper large-v3 to produce an orthographic transcript for the synthetic speech and cleaned original speech, scoring both transcripts against the reference provided in the dataset and giving the results in Table 6.

Table 6: *% WER/CER results on FLEURS test set for synthetic and cleaned original speech; CER languages are italicised.*

Language	WER/CER Synthetic	WER/CER Original
French	55.0	5.6
Dutch	47.0	5.7
German	26.2	5.3
Japanese	17.2	4.5
Korean	12.7	2.9
Italian	11.4	2.4
English	10.4	4.1

We were able to re-synthesise examples in every language that gave 0% WER/CER. We include these examples on our samples page (as "Intelligible" samples) alongside examples which were judged as less intelligible ($\geq$ 50% WER/CER). The 10% WER for the English test set is comparable to the 7% WER on the XRMB validation set. This shows that our inversion model not only achieves a low distance-based measure (as shown in Table 3), but also produces articulatory features which reasonably reflect the speech that is being analysed. However, the joint system does seem to be sensitive to recording conditions. The ranked error rates for the synthetic speech correlate with the error rates we get from the transcript generated from the clean original speech (excluding English). In addition to this, when we examined the performance of the system on French speakers, we found one particular speaker who performed significantly worse than the others in terms of WER, as shown in Figure 3.

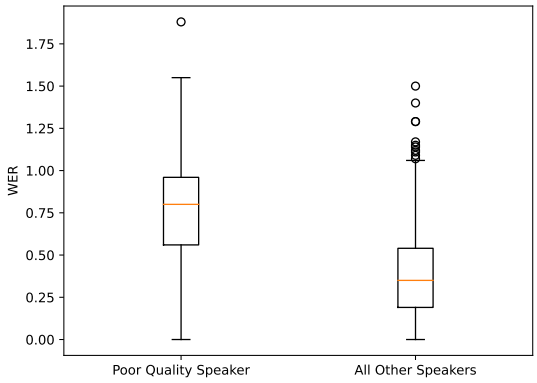

Figure 3: *WER performance between a single French speaker with poor recording quality and all other French speakers.*

Listening to the recordings for this speaker, we notice that even after enhancement the speaker is quite muffled with sharp bursts of energy for sibilants (Example 1 on the samples page). The reliance on energy as a feature to cover aspects of speech such as aperiodicity, nasality and silence may be a major factor of this degradation in model performance. We have also found that the system seems to be incapable of replicating some sounds that are not found in English, such as the trill /r/ in Italian, as demonstrated in Example 2 on our samples page. More analysis, however, is required before we can determine which speech sounds the model is capable of producing.

7. Conclusions

In this paper, we have shown that it is possible to build a highly intelligible articulatory synthesiser using a high-quality articulatory dataset and synthetic articulatory data created through inversion. We used the resulting synthesiser to perform a novel evaluation of the output of a combined inversion-synthesis system, generating intelligible speech in languages which were not seen during training. In future work, we will continue to analyse the capability of the joint system in producing speech sounds which are not found in English and carry out human evaluations of intelligibility to reinforce the results presented in this work. We will also look to improve the synthesis model by reviewing the input source features and the inversion model through semi-supervised learning.

8. References

[1] T. Rebernik, J. Jacobi, R. Jonkers, A. Noiray, and M. Wieling, "A review of data collection practices using electromagnetic articulography," *Laboratory Phonology*, vol. 12, no. 1, p. 6, 2021.

[2] A. Narwekar and P. K. Ghosh, "A comparative study of articulatory features from facial video and acoustic-to-articulatory inversion for phonetic discrimination," in *Proc. 2016 International Conference on Signal Processing and Communications (SPCOM)*, 2016, pp. 1–5.

[3] C. McGhee, K. Knill, and M. Gales, "Towards Acoustic-to-Articulatory Inversion for Pronunciation Training," in *Proc. 9th Workshop on Speech and Language Technology in Education (SLaTE)*, 2023, pp. 66–70.

[4] N. R. Benway, Y. M. Siriwardena, J. L. Preston, E. Hitchcock, T. McAllister, and C. Espy-Wilson, "Acoustic-to-Articulatory Speech Inversion Features for Mispronunciation Detection of /ɹ/ in Child Speech Sound Disorders," in *Proc. INTERSPEECH 2023*, 2023, pp. 4568–4572.

[5] P. K. Krug, S. Stone, and P. Birkholz, "Intelligibility and naturalness of articulatory synthesis with VocalTractLab compared to established speech synthesis technologies," in *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, 2021, pp. 102–107.

[6] 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.

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

[8] M. Kim, Z. Piao, J. Lee, and H.-G. Kang, "Style modeling for multi-speaker articulation-to-speech," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[9] R. Blandin, S. Stone, A. Remacle, V. Didone, and P. Birkholz, "A comparative study of 3D and 1D acoustic simulations of the higher frequencies of speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[10] R. Blandin, M. Arnela, S. Félix, J.-B. Doc, and P. Birkholz, "Efficient 3D acoustic simulation of the vocal tract by combining the multimodal method and finite elements," *IEEE Access*, vol. 10, pp. 69 922–69 938, 2022.

[11] N. Meenakshi, C. Yarra, B. K. Yamini, and P. K. Ghosh, "Comparison of speech quality with and without sensors in electromagnetic articulograph AG 501 recording," in *INTERSPEECH 2014*, 2014, pp. 935–939.

[12] J. R. Westbury, G. Turner, and J. Dembowski, "X-ray microbeam speech production database user's handbook," *University of Wisconsin*, 1994.

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

[14] R. S. McGowan, "Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary model tests," *Speech Communication*, vol. 14, no. 1, pp. 19–48, 1994.

[15] A. Suemitsu, T. Ito, and M. Tiede, "An electromagnetic articulography-based articulatory feedback approach to facilitate second language speech production learning," in *Proc. of Meetings on Acoustics*, vol. 19, no. 1. AIP Publishing, 2013.

[16] 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.

[17] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using Kaldi." in *Proc. INTERSPEECH 2017*, 2017, pp. 498–502.

[18] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," *CoRR*, vol. abs/2106.04624, 2021.

[19] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[20] H. Dubey, V. Gopal, R. Cutler, S. Matusevych, S. Braun, E. S. Eskimez, M. Thakker, T. Yoshioka, H. Gamper, and R. Aichner, "ICASSP 2022 Deep Noise Suppression Challenge," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[21] T. Sainburg, M. Thielk, and T. Q. Gentner, "Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires," *PLoS computational biology*, vol. 16, no. 10, p. e1008228, 2020.

[22] T. Sainburg, "timsainb/noisereduce: v1.0," Jun. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3243139

[23] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech," *CoRR*, vol. abs/1904.02882, 2019.

[24] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, "FLEURS: Few-shot learning evaluation of universal representations of speech," in *Proc. 2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 798–805.

[25] 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.

[26] C. J. Cho, P. Wu, A. Mohamed, and G. K. Anumanchipalli, "Evidence of vocal tract articulation in self-supervised learning of speech," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[27] E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "LoRA: Low-rank adaptation of large language models," in *Proc. International Conference on Learning Representations (ICLR)*, 2021.

[28] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Artificial intelligence and machine learning for multi-domain operations applications, Proc. SPIE Defense + Commercial Sensing*, vol. 11006. SPIE, 2019, pp. 369–386.

[29] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for speech recognition," in *Proc. INTERSPEECH 2020*, 2020, pp. 5036–5040.

[30] S.-g. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, "BigVGAN: A Universal Neural Vocoder with Large-Scale Training ," in *Proc. ICLR*, 2023.

[31] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "CREPE: A convolutional representation for pitch estimation," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 161–165.

[32] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. International Conference on Machine Learning*, 2023, pp. 28 492–28 518.

[33] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, p. 27403, 1993.

[34] Y. M. Siriwardena, C. Espy-Wilson, S. Boyce, M. Tiede, and L. Oren, "Speaker-independent Speech Inversion for Estimation of Nasalance," in *Proc. INTERSPEECH 2023*, 2023, pp. 4743–4747.