Interspeech 2022
18-22 September 2022, Incheon, Korea

Articulatory Synthesis for Data Augmentation in Phoneme Recognition

Paul K. Krug[1], *Peter Birkholz*[1], *Branislav Gerazov*[2], *Daniel R. van Niekerk*[3], *Anqi Xu*[3], *Yi Xu*[3]

[1]Institute of Acoustics and Speech Communication, Technische Universität Dresden, Germany
[2]Faculty of Electrical Engineering and Information Technologies, UKIM Skopje, R. N. Macedonia
[3]Department of Speech, Hearing and Phonetic Sciences, University College London, UK

paul_konstantin.krug@tu-dresden.de

Abstract

While numerous studies on automatic speech recognition have been published in recent years describing data augmentation strategies based on time or frequency domain signal processing, few works exist on the artificial extensions of training data sets using purely synthetic speech data. In this work, the German KIEL corpus was augmented with synthetic data generated with the state-of-the-art articulatory synthesizer VOCALTRACTLAB. It is shown that the additional synthetic data can lead to a significantly better performance in single-phoneme recognition in certain cases, while at the same time, the performance can also decrease in other cases, depending on the degree of acoustic naturalness of the synthetic phonemes. As a result, this work can potentially guide future studies to improve the quality of articulatory synthesis via the link between synthetic speech production and automatic speech recognition.
Index Terms: automatic speech recognition, phoneme recognition, articulatory speech synthesis, data augmentation

1. Introduction

In contrast to the current situation in the field of computer vision research, where state-of-the-art results could often be achieved through a systematic synthesis of training data [1–8], the use of synthetic data for speech recognition purposes [9–11] is rare, despite some recent success of data augmentation strategies [12–19]. This situation may seem perplexing, as numerous ways to improve speech recognition systems through the use of synthetic data are reasonable: (i) Specific synthesis of underrepresented phonemes in order to balance out and/or extend a training data set. (ii) Synthesis of specific domain related words that rarely occur in common language. (iii) Synthesis of underrepresented speaking styles, such as emotional speech or speech dialects. (iv) Synthesis of underrepresented voices, such as young speakers, old speakers or speakers with speech or voice disorders.

The low interest in synthetic augmentation methods may be explained by the lack of versatility and variability [11] in current state-of-the-art speech synthesis systems. While neural synthesis systems are often characterized by the fact that they provide remarkable synthesis results [20–22], they often leave little control over the synthesis process due to their end-to-end nature. All in all, these systems are not flexible enough to cope with the versatile applications of synthetic augmentation mentioned earlier. For each application, systems would either have to be newly developed, which is usually more costly than directly recording speech data for recognition, or, voice-conversion techniques [23–26] would have to be used in order to convert existing systems to domain-appropriate systems. However, such techniques may introduce unnatural speech conversion artifacts. An adequate solution to these problems would be to use a *true* speech synthesis system that allows full control over the complete speech production process. An ideal candidate for such a system is *articulatory speech synthesis* [27], which aims to replicate the human speech mechanism, i.e. to simulate the vocal tract dynamics. A disadvantage of articulatory synthesis is that it requires a precise knowledge of human speech production, and even the most advanced articulatory synthesizers are under continuous active development [28], as there are still open questions and uncertainties regarding the fundamental mechanisms of speech production. Nevertheless, the state-of-the-art articulatory synthesis software VOCALTRACTLAB [29] (VTL) provides a compelling tool to explore the use of synthetic data for speech recognition. The phoneme-to-speech functionality [28, 30] introduced in VTL version 2.3 is of particular interest as it allows the generation of large synthetic data sets in a simple way. This study makes the following contributions to the state of research:

(i) It is shown by single-phoneme recognition that natural speech data can be augmented with synthetic phonemes to increase the recognition rate of certain rare events.

(ii) It is shown that the recognition rate of certain phoneme classes can also decrease due to the augmentation, which points to directions in which the artificial speech generation must be improved in the future.

Single-phoneme recognition is thereby chosen over continuous phoneme recognition on the word or sentence level for several reasons. First, models that capture the temporal structure of utterances can learn both, acoustic recognition as well as language model-like pattern recognition from the frequency distributions of phonemes within utterances. However, the pure acoustic information is of particular interest in this case, as individual phonemes represent the smallest units of speech. Hence, synthetic phonemes must at least partially capture the degree of the natural acoustic realization in order to successfully augment natural data. If the augmentation does not work for certain phonemes, this is an indication of an unnatural realization of these sounds. This interplay of production and recognition can be a valuable tool for the development of articulatory synthesizers as it can point the direction in which the synthesis should be improved. Second, the synthesis of larger utterances involves further sources of unnaturalness, e.g. via speech flow [28], which should be avoided for now.

2. Methods

2.1. Kiel Data Set

As the main training data set, the *Read Speech* section of the *Kiel Corpus of Spoken German* [31, 32], *New Edition 2017* (KIEL) was used. The set contains speech data from 53 speakers

10.21437/Interspeech.2022-10874

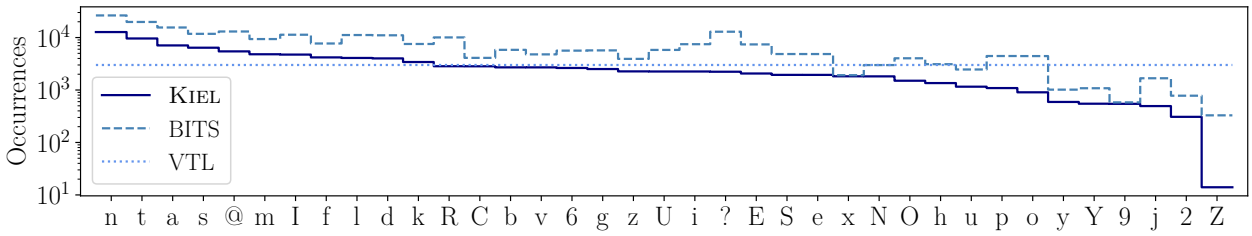

Figure 1: *The phoneme frequency distributions of the* KIEL, BITS *and* VTL *data sets, ordered based on the* KIEL *distribution.*

(27 male, 26 female) and has a total duration of 4.25 h (including silence), which makes it a low-resource corpus. The corpus contains segmental labelling, which means a manual time alignment of the actually realized phonemes is provided. This annotation was carried out by phonetically trained students under the supervision of phonetic experts. The audio files of the KIEL corpus have a native sampling rate of 16 kHz.

2.2. BITS Data Set

As the main validation and test data set, the *BAS Infrastructures for Technical Speech Processing* unit selection corpus [33] (BITS), version 1.7 was used. Even though this data set has a longer duration (approximately 13.5 h, including silence) compared to the KIEL corpus, it provides speech data from only 4 speakers (2 male, 2, female) due to its intended use in unit selection synthesis rather than speech recognition. Nevertheless, the BITS speech data is of high quality as the speakers were recruited in an elaborate process. The corpus provides a manual segmentation of the phonemes which were carefully annotated by phonetic experts in a multi-stage process. The audio files of the BITS corpus have a native sampling rate of 48 kHz.

2.3. Data Processing

First, the segment boundaries corresponding to the vowels[1,2] /a, e, i, o, u, E, I, O, U, 2, 9, y, Y, @, 6/, the plosives /p, t, k, b, d, g, ?/, the fricatives /f, v, s, z, S, Z, j, C, x, R, h/ and the nasals /m, n, N/ as well as the lateral /l/ were extracted from the KIEL and BITS data sets. The corresponding audio segments were extracted by extending the phoneme segments with a context duration of 32 ms in each direction, see Figure 2. Such a context is long enough to capture the relevant formant transitions that characterize certain phonemes, and short enough to ensure that only a small part of the preceding and succeeding phoneme is visible as the context in most cases. The extracted audio sections were then resampled to a sampling rate of 16 kHz and 16 ms of silence were concatenated at the beginning and at the end of each audio sample, respectively. Thereby, the signal was faded in and out using a cosine window in order to avoid discontinuities in the time signal. These transitions between silence and signal parts were realized at a length of 1 ms. Subsequently, 80 dimensional mel-scaled spectrograms were calculated from the audio samples as input features to the phoneme recognition model presented in Section 2.5. A window length of 256 samples (16 ms) and a hop length of 40 samples (2.5 ms) were used. The mel-spectrogram intensities were converted to the dB scale. The number of spectrogram time frames was required to be between 50 and 125, which corresponds to a restriction of the annotated phoneme length to a range between 29 ms and 216.5 ms. Spectrograms with less than 125 frames were accordingly right-padded with zero entries. The total signal durations of the processed data sets were 2.25 h (109,810 samples) and 5.7 h (256,877 samples) in case of KIEL and BITS, respectively (without silence and contexts). The BITS data was split into a validation set and a test set via a randomly shuffled and stratified split, with 10 % of the total samples in the validation set and 90 % in the test set. This way, the ratio of the KIEL training data to the BITS validation data is about 4:1. The phoneme frequency distributions of the final sets are visualized in Figure 1.

[1] SAMPA notation is used throughout this work.

[2] Due to the small acoustic and articulatory contrast, the categories /a:/ and /E:/ were folded into /a/ and /E/, respectively.

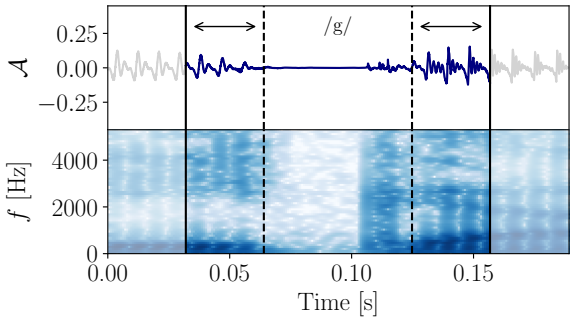

Figure 2: *Each extracted phoneme is extended by an adjacent context of* 32 *ms in both directions. The plot shows this process exemplified by a phoneme* /g/*. The annotated phoneme boundaries are shown as dashed lines, the boundaries extended by the context are shown as solid lines. The lower plot shows the spectrogram, in which the important formant transitions can be clearly seen in the context area.*

2.4. Synthetic Data

2.4.1. VocalTractLab

VTL features a one-dimensional aerodynamic-acoustic simulation of the vocal tract dynamics. It uses a realistic model of the human vocal tract geometry derived from magnetic resonance imaging (MRI) data and a geometric vocal fold model [34]. The VTL synthesis process can be driven by a high-level control via *gestural scores*. The dynamics of the gestural scores are governed by the TARGET-APPROXIMATION-MODEL [35, 36].

2.4.2. Data Generation

For the generation of synthetic speech data, the VTL synthesis backend was accessed via the PYTHON library

VOCALTRACTLAB-PYTHON[3] [37] (VTL-Python), which allows fast parallel synthesis via multiprocessing. Each of the 37 phonemes was generated $3 \cdot 10^3$ times using the VTL-Python speaker file *JD3*. This makes a total of $11.1 \cdot 10^4$ data samples, which is approximately similar to the number of natural samples from the KIEL data set. During generation, the phonemes of interest were embedded in a random context sequence /V, X_L, X, X_R, V/, where V means a random vowel with a duration of 96 ms. X denotes the phoneme of interest and X_L and X_R are randomly drawn from a uniform distribution of the 37 available phonemes. This is done to provide a well defined context for the phoneme of interest. The inner context is subject to the following constraints to avoid implausible combinations: Not more than two consonants must be in the sequence, voiced plosives must not follow their unvoiced counterparts and vice-versa, if the phoneme of interest is a glottal stop it must be succeeded by a vowel. For the three inner phonemes of the sequence, appropriate duration values were sampled from gamma distributions that were previously fitted to the BITS validation set duration distributions. Thereby, several different groups with different duration distributions were identified and fitted individually: the tense vowels, the lax vowels, the voiced plosives plus glottal stop, the unvoiced plosives, the nasals, the lateral, as well as /f, s, S/, /z, Z, C, x/ and /v, j, R, h/. Subsequently, gestural scores were calculated from the phoneme sequence and the respective duration information via the VTL phoneme-to-speech functionality. The gestural scores describe the dynamics of individual articulatory tiers during speech production. However, the default plain pitch tiers [28] were replaced with a pitch contour, specifically adjusted to augment the KIEL speech data. For each contour, four pitch targets were determined, whereby the duration of the first three targets was sampled from uniform distributions between 48 ms and a quarter of the total gestural score duration, while the duration of the last target was set to the remaining time until the end of the gestural score. The target offset parameters (see e.g. [38]) were determined by random sampling a mean fundamental frequency value $\bar{f}_0$ from a uniform distribution between 50 Hz and 250 Hz. Subsequently, the target offsets were drawn from a uniform distribution within the interval of the mean pitch $\pm$ 6 semitones. The corresponding target slope parameters were set to 0 and the time constants were set to 20 ms. Afterwards the synthetic speech data was generated using the respective VTL-PYTHON functionality. Finally, the phonemes of interest were extracted with a respective context duration as previously described in Section 2.3.

2.5. Phoneme Recognition Model

For the purpose of phoneme recognition, a deep recurrent neural network was used to capture the temporal structure of the input spectra. Thereby, 5 bidirectional gate recurrent unit (Bi-GRU) layers with 256 neurons each and *tanh* activation function were followed by a dense layer with 37 neurons and a *softmax* activation function, see Table 1. The recurrent layers were preceded by a masking layer (in order to make the network ignore the spectral padding) and a batch normalization layer. As a whole, this structure acts as a simple encoder which directly transforms the spectral time series into a 37 dimensional probability distribution where each dimension corresponds to a particular phoneme. The architecture of the recurrent neural network was broadly optimized with respect to the number of trainable model parameters, see Figure 3. Thereby, recurrent layers seemed to be beneficial over dense layers and gate recurrent units showed improved performance compared to long short term memory cells (probably due to less overfitting). As a consequence, the final model consisted only of Bi-GRU layers and has around $5 \cdot 10^6$ trainable parameters. As for the hyper parameter optimization, the batch size was varied between the values 16, 32, 64, whereby 32 was found to be the optimal value with respect to performance and training speed. The learning rate was varied between the values 10^{-3}, 10^{-4} and 10^{-5}, whereby 10^{-4} gave optimal performance.

[3] https://github.com/paul-krug/VocalTractLab-Python

Layer	Shape	N_{Params}
Input	(B, 125, 80)	0
Masking	(B, 125, 80)	0
Batch Normalization	(B, 125, 80)	320
Bi-GRU	(B, 125, 256)	519168
Bi-GRU	(B, 125, 256)	1182720
Bi-GRU	(B, 125, 256)	1182720
Bi-GRU	(B, 125, 256)	1182720
Bi-GRU	(B, 256)	1182720
Dense	(B, 37)	18981

Table 1: *Structure of the phoneme recognition network, ordered from the input (top) to the output (bottom). The columns describe the layer type, the shape of the corresponding output tensor and the number of model parameters provided by the layer, respectively. Thereby, B denotes the batch size.*

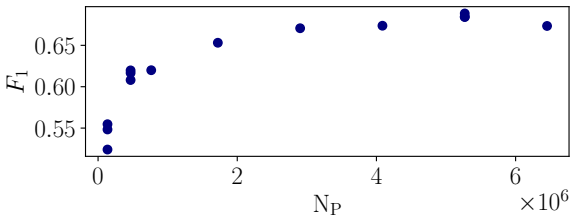

Figure 3: *Phoneme recognition performance measured via the F_1 score as a function of trainable model parameters.*

2.6. Experiments

Several experiments were conducted: (i) The phoneme recognition model was trained on the KIEL data and tested on BITS. (ii) The model was trained on the KIEL + VTL data and tested on BITS. (iii) The model was trained on the synthetic data only and tested on the natural data (KIEL + BITS test set). (iv) The model was trained on the natural data and tested on VTL data. Model performances were evaluated in terms of F_1 score and recall ($\mathcal{R}$). Each model was trained 5 times in order to obtain adequate statistical power, the mean and standard deviations over the individual instances are reported. Performances are also reported for the sub sets of male and female speakers. Each training was done for 25 epochs using early stopping on F_1 with a patience of 5 epochs and a minimum delta of 0.001.

3. Results

The phoneme recognition results from the different experiments are listed in Table 2. One can see that the overall performance in terms of recall and F_1 score actually slightly decreases when

Training data	Test data	F_1 $[10^{-2}]$	$\mathcal{R}$ $[10^{-2}]$
KIEL	BITS	71.0 ± 0.5	70.3 ± 0.6
	└ (m)	74.2 ± 0.3	73.4 ± 0.6
	└ (f)	67.7 ± 0.7	67.2 ± 0.7
KIEL + VTL	BITS	70.3 ± 0.6	69.5 ± 0.9
	└ (m)	73.5 ± 0.8	72.5 ± 0.8
	└ (f)	66.8 ± 1.2	66.4 ± 1.5
VTL	KIEL + BITS	22.6 ± 0.4	26.1 ± 0.2
	└ (m)	23.8 ± 0.4	28.1 ± 0.4
	└ (f)	20.8 ± 0.3	24.0 ± 0.4
Kiel + BITS	VTL	42.0 ± 0.1	42.2 ± 0.3

Table 2: *The F_1 and $\mathcal{R}$ values are shown for different training and testing data scenarios. The data subsets with male and female speakers only are denoted with "m" and "f", respectively.*

the recognition model is trained on KIEL + VTL data compared to the case without VTL data. A more detailed look at the results reveals significant differences in case of certain phoneme classes, see Figure 4. For phonemes /o, O, Z/ an improved recognition rate was observed, however, in case of /y, s, h/ the performance decreased. With the exception of /s/, these phonemes are classes which are rather underrepresented in the KIEL data set, see Figure 1. This makes sense, since in these cases there is a preponderance of synthetic training data compared to natural data. Hence, the performance will be influenced accordingly into the positive or negative depending on the degree of acoustic naturalness among the synthetic samples. Strong changes were observed for the recognition rates of /y/ and /Z/, which saw an absolute decrease of 0.136 and an increase of 0.284, respectively. However, the results from experiments (iii) and (iv), involving models trained on synthetic or natural data only, see Table 2, indicate a significant lack of naturalness and variability within the synthetic data. In fact, when training was performed on VTL data only, strong confusion among the vowels, as well as among the plosives can be observed, see Figure 5 (left). While the recognition of fricatives /f, s, z, S, Z/ performs comparatively well, the recognition of nasals generalizes poorly to the natural data. A reason for this may be found within the spectral content of the phoneme samples. The top and bottom plots in Figure 6 show the mean spectra of all /Z/ and /N/ samples, respectively. In the former case, the synthetic spectral contour roughly coincides with the natural contour while strong deviations of up to 23 dB are observed in case of the synthetic and natural nasal spectra. Conversely, when training on the natural data and testing on the synthetic data, there is a much stronger concentration of events on the main diagonal of the confusion matrix, see Figure 5 (right). This implies that the natural data are diverse enough so that the recognition model at least roughly generalizes to the synthetic data and that the synthetic data may be considered outliers with regard to the distribution of natural data.

4. Discussion and Conclusion

In this study, it was shown that the recognition of certain underrepresented phonemes can be improved by augmenting the underlying training data via articulatory synthesis. However, it was also observed that the recognition rates decrease in some cases, probably due to an insufficient realization of acoustic naturalness. In addition, it was found that the recognition of natural phonemes was poor when models were trained only on synthetic data. While recognition was better when trained on

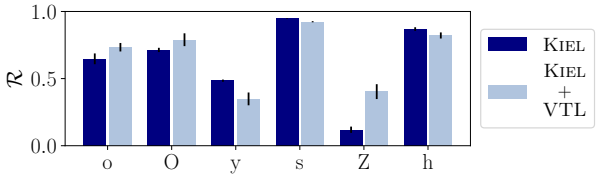

Figure 4: *Phoneme recognition performance measured via the recall score, shown for models trained on* KIEL *data (dark) and* KIEL + VTL *data (bright). Shown classes have $p < 0.1$ in respective t-tests between the* KIEL *and* KIEL + VTL *results.*

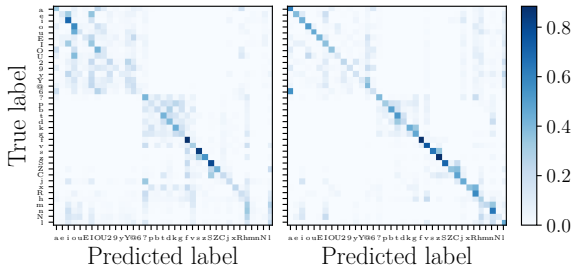

Figure 5: *Average confusion matrices, shown for models trained on* VTL *data and tested on natural data (left) vs. trained on natural data and tested on* VTL *data (right).*

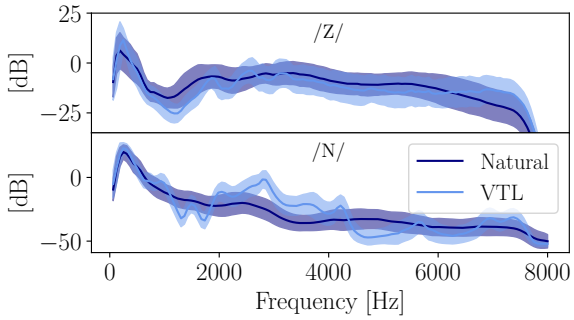

Figure 6: *Average spectra of the phonemes /Z/ and /N/ extracted from* KIEL + BITS *(Natural) and* VTL *data.*

natural data only and tested on synthetic data, overall, a large gap between recognition rates of natural and synthetic data was still observed. These findings may be explained by two causes: (i) The lack of variation among the synthetic data and (ii) the intrinsic unnaturalness of the synthetic speech sounds. Solutions to the former problem may be, first, to create inter-speaker variability either by MRI recordings of new vocal tract geometries or by acoustic-data driven methods such as vocal learning simulations [39–41], and second, to increase intra-speaker variability, e.g., by stochastic variation of preset vocal tract shapes, e.g., using PCA transformations [41], as well as by changing voice quality [42], target-time constants, etc. To solve the second problem, the physical modeling of the vocal tract itself must be improved. Starting points are the improved modeling of fricative noise sources, loss mechanisms (for more realistic formant bandwidths), sound radiation, as well as articulatory dynamics.

5. Acknowledgements

This work has been funded by the Leverhulme Trust Research Project Grant RPG-2019-241: "High quality simulation of early vocal learning".

6. References

[1] E. Wood *et al.*, "Rendering of eyes for eye-shape registration and gaze estimation," in *Proc. ICCV*, 2015, pp. 3756–3764.

[2] Y. Movshovitz-Attias *et al.*, "How useful is photo-realistic rendering for visual learning?" in *Proc. ECCV*, 2016, pp. 202–217.

[3] S. R. Richter *et al.*, "Playing for data: Ground truth from computer games," in *Proc. ECCV*, 2016, pp. 102–118.

[4] G. Ros *et al.*, "The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Proc. CVPR*, 2016, pp. 3234–3243.

[5] G. Varol *et al.*, "Learning from synthetic humans," in *Proc. CVPR*, 2017, pp. 4627–4635.

[6] F. Lin *et al.*, "Facial expression recognition with data augmentation and compact feature learning," in *Proc. ICIP*, 2018, pp. 1957–1961.

[7] S. Jain *et al.*, "Synthetic data augmentation for surface defect detection and classification using deep learning," *J. Intell. Manuf.*, pp. 1–14, 2020.

[8] E. Wood *et al.*, "Fake it till you make it: Face analysis in the wild using synthetic data alone," in *Proc. ICCV*, 2021, pp. 3681–3691.

[9] A. Rosenberg *et al.*, "Speech recognition with augmented synthesized speech," in *Proc. ASRU*, 2019, pp. 996–1002.

[10] N. Rossenbach *et al.*, "Generating synthetic audio data for attention-based speech recognition systems," in *Proc. ICASSP*, 2020, pp. 7069–7073.

[11] G. Wang *et al.*, "Improving speech recognition using consistent predictions on synthesized speech," in *Proc. ICASSP*, 2020, pp. 7029–7033.

[12] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," *IEEE Speech Audio Process.*, vol. 6, no. 1, pp. 49–60, 1998.

[13] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (vtlp) improves speech recognition," in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, vol. 117, 2013, p. 21.

[14] A. Ragni *et al.*, "Data augmentation for low resource languages," in *Proc. Interspeech*, 2014, pp. 810–814.

[15] T. Ko *et al.*, "Audio augmentation for speech recognition," in *Proc. Interspeech*, 2015, pp. 3586–3589.

[16] J. Fainberg *et al.*, "Improving children's speech recognition through out-of-domain data augmentation." in *Interspeech*, 2016, pp. 1598–1602.

[17] C. Kim *et al.*, "Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in google home," in *Proc. Interspeech*, 2017, pp. 379–383.

[18] D. S. Park *et al.*, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, 2019, pp. 2613–2617.

[19] ——, "SpecAugment on large scale datasets," in *Proc. ICASSP*, 2020, pp. 6879–6883.

[20] A. van den Oord *et al.*, "WaveNet: A generative model for raw audio," *CoRR*, vol. abs/1609.03499, 2016.

[21] J. Shen *et al.*, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. ICASSP*, 2018, pp. 4779–4783.

[22] Y. Ren *et al.*, "FastSpeech: Fast, robust and controllable text to speech," in *Proc. NeurIPS*, vol. 32, 2019, pp. 3171–3180.

[23] F. Fang *et al.*, "High-quality nonparallel voice conversion based on cycle-consistent adversarial network," in *Proc. ICASSP*, 2018, pp. 5279–5283.

[24] H. Kameoka *et al.*, "StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *Proc. SLT*, 2018, pp. 266–273.

[25] A. Polyak and L. Wolf, "Attention-based WaveNet autoencoder for universal voice conversion," in *Proc. ICASSP*, 2019, pp. 6800–6804.

[26] K. Zhou *et al.*, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," in *Proc. ICASSP*, 2021, pp. 920–924.

[27] C. H. Shadle and R. I. Damper, "Prospects for articulatory synthesis: A position paper," in *Proc. SSW 4*, 2001.

[28] P. K. Krug *et al.*, "Intelligibility and naturalness of articulatory synthesis with VocalTractLab compared to established speech synthesis technologies," in *Proc. SSW 11*, 2021, pp. 102–107.

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

[30] P. K. Krug *et al.*, "Modelling microprosodic effects can lead to an audible improvement in articulatory synthesis," *J. Acoust. Soc. Am.*, vol. 150, no. 2, pp. 1209–1217, 2021.

[31] K. J. Kohler *et al.*, "From scenario to segment. the controlled elicitation, transcription, segmentation and labelling of spontaneous speech," *Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Christian- Albrechts-Universität Kiel (AIPUK)*, vol. 29, 1995.

[32] ——, "From the acoustic data collection to a labelled speech data bank of spoken standard german," *Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel (AIPUK)*, vol. 32, pp. 1–29, 1997.

[33] F. Schiel *et al.*, "Die BITS Sprachsynthesekorpora – Diphon-und Unit Selection-Synthesekorpora für das Deutsche," in *Proc. KONVENS*, 2006, pp. 121–124.

[34] P. Birkholz *et al.*, "Perceptual optimization of an enhanced geometric vocal fold model for articulatory speech synthesis." in *Proc. Interspeech*, 2019, pp. 3765–3769.

[35] Y. Xu and Q. E. Wang, "Pitch targets and their realization: Evidence from Mandarin Chinese," *Speech Commun.*, vol. 33, no. 4, pp. 319–337, 2001.

[36] S. Prom-On *et al.*, "Modeling tone and intonation in Mandarin and English as a process of target approximation," *J. Acoust. Soc. Am.*, vol. 125, no. 1, pp. 405–424, 2009.

[37] P. K. Krug, "A window-based method for target estimation," *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2022*, pp. 196–203, 2022.

[38] P. K. Krug *et al.*, "TargetOptimizer 2.0: Enhanced estimation of articulatory targets," *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2021*, pp. 145–152, 2021.

[39] D. R. van Niekerk *et al.*, "Finding intelligible consonant-vowel sounds using high-quality articulatory synthesis," in *Proc. Interspeech*, 2020, pp. 4457–4461.

[40] A. Xu *et al.*, "Model-based exploration of linking between vowel articulatory space and acoustic space," in *Proc. Interspeech*, 2021, pp. 3191–3195.

[41] P. K. Krug *et al.*, "Efficient exploration of articulatory dimensions," *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2022*, pp. 51–58, 2022.

[42] P. Birkholz *et al.*, "Synthesis of breathy, normal, and pressed phonation using a two-mass model with a triangular glottis," in *Proc. Interspeech*, 2011, pp. 2681–2684.