# ARTICULATORY COPY SYNTHESIS USING LONG SHORT-TERM MEMORY NETWORKS

*Yingming Gao, Peter Steiner, Peter Birkholz*

*Institute of Acoustics and Speech Communication, TU Dresden, Germany*
*yingming.gao@mailbox.tu-dresden.de, {peter.steiner, peter.birkholz}@tu-dresden.de*

**Abstract:** Investigating speech imitation, in particular articulatory copy synthesis, benefits the understanding of speech production and can improve speech recognition and synthesis. We proposed a framework for copy synthesis with an artificial neural network (LSTM regression model) and an articulatory speech synthesizer (VocalTractLab), which were responsible for the acoustic-to-articulatory mapping and the inversion, respectively. We used rule-based method to create gestural scores from texts, which were converted to articulatory trajectories and subsequently simulated to produce the corresponding acoustic signals. To make the subsequent mapping more robust, we expanded the acoustic and articulatory space by manipulating speaking effort, voice quality, pitch level, and vocal tract length of the created gestural scores or acoustic signals, producing 81 variants for each utterance. With acoustic features as input and articulatory trajectories as output, we trained the LSTM models to build the acoustic-to-articulatory inversion. For testing, we estimated the articulatory trajectories from acoustic features, thus obtaining the underlying articulatory process. The experiments showed that the correlation coefficients (between estimated articulatory trajectories from acoustic features and the real ones converted from gestural scores) ranged from 0.18 to 0.973 and the root mean square error (RMSE) ranged from 0.043 to 0.255 for the concerned 30 articulatory parameters of VocalTractLab. The estimated articulatory parameters were further fed into VocalTractLab, whose output speech achieved a word recognition accuracy of 17.24%.

## 1 Introduction

Investigating speech imitation, in particular articulatory copy synthesis, benefits the understanding of speech production and can improve speech recognition and synthesis [1]. Articulatory copy synthesis refers to a technique of reproducing human speech by modeling the corresponding articulation process. The research question can be abstracted as articulatory-to-acoustic mapping and its inversion. The existing methods usually suffer from one or more of the following limitations: (1) the construction of mapping models relies on recorded articulatory data [2], which is labor-intensive or invasive to speakers during data collecting, (2) the inversion is limited to short-utterances (either isolated vowels or simple 'CV[C]' syllables) [3][4], (3) the analysis-by-synthesis (ABS) based methods [1][4] are time-consuming and have no generality, i.e. the inversion has to be individually performed for each utterance, or (4) the mapping is based on speaker-dependent models [5], i.e., both the training and testing data come from the same speaker. In this paper, we propose a novel approach of articulatory copy synthesis tackling the above problems to some extent.

Articulatory copy synthesis is not only to reproduce the acoustic signal of given natural speech, but also to obtain the articulation. The difficulty of this topic comes from at least two

main aspects: (1) both the acoustic and articulatory parameters have temporal dependences, (2) the acoustic-to-articulatory inversion has the problem of non-uniqueness. Recently, long short-term memory (LSTM) neural networks, due to their power of modeling sequential data, have been successfully applied to many areas including handwriting recognition, language translation, speech recognition and so on. In the present study, we proposed a framework for copy synthesis with a neural network regression model and an articulatory speech synthesizer.

## 2 Method

### 2.1 The framework of articulatory copy synthesis

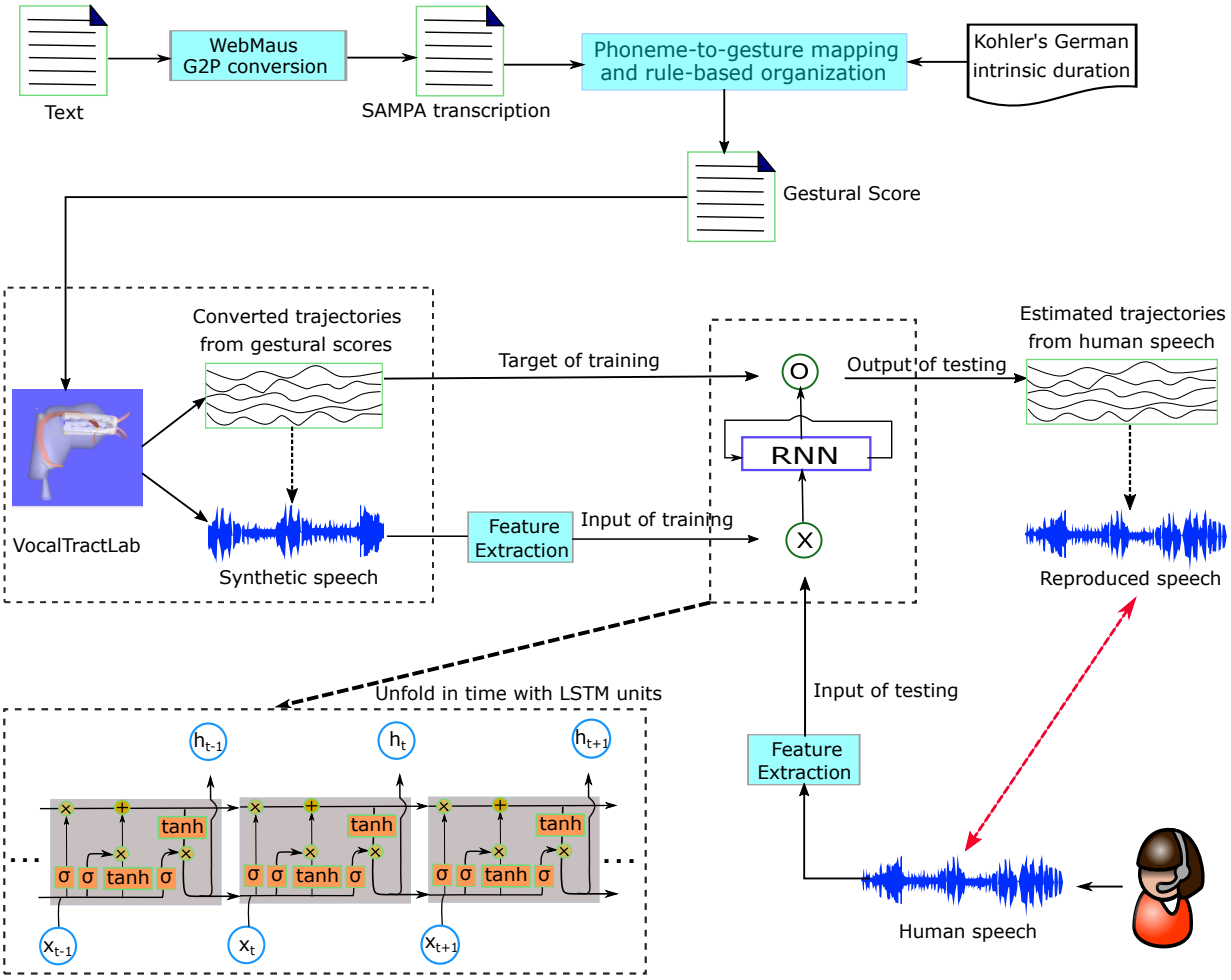


**Figure 1** – Schematic diagram of ariticulatory copy synthesis with Long Short-Term Memory Networks and VocalTractLab. The upper part shows the process of generating gestural scores. The middle part is the core of the system, modeling the acoustic-to-articulatory mapping and inversion. The bottom part illustrates the unfolded neural network model and testing the trained model with human speech.

Figure 1 illustrates the framework for copy synthesis. It can be divided into three main stages: creation of training data, modeling of mapping and inversion between acoustic and articulatory representations, and testing with human speech. Starting from text, we use a G2P tool to convert each word of it to a phoneme sequence, which is subsequently mapped into gestures with a rule-based method. All gestures constitute a gestural score. In the second stage, the cores are an articulatory speech synthesizer and an LSTM neural network (abstracted as a RNN neural network in Figure 1), which are responsible for the articulatory-to-acoustic mapping and

the inversion, respectively. The articulatory synthesizer converts the gestural scores generated in the first stage to articulatory trajectories and corresponding synthetic speech. After feature extraction, the acoustic features serve as the input of the neural network. Accordingly, the articulatory trajectories serve as the output during the training of the neural network regression model. The RNN with LSTM units can be unfolded into a series of networks (as represented in the bottom-left corner of Figure 1). Each memory unit is composed of a cell, an input gate, an output gate and a forget gate, to track the network state and the flow of information into and out of the cell. In the present work, the LSTM models the utterances as continuous trajectories in the acoustic as well as in the articulatory space. The property of recurrence of LSTM handles the temporal dynamic behavior of the parametric dependencies for entire sequences of data. The acoustic features and the articulatory trajectories serve as the input and output of the LSTM neural network, respectively. In the testing stage, after feature extraction, the acoustic features of human speech will serve as the input of the trained LSTM regression model whose output is the estimated articulatory trajectories. Using VTL again, we convert them to synthetic speech.

### 2.2 Articulatory speech synthesizer

VocalTractLab-2.2 (VTL) [6] is the articulatory speech synthesizer used in this work. It takes articulatory trajectories as input and simulates the acoustic signal as output. There are 33 control parameters in VTL (as listed in Table 1): 24 for the vocal tract model and 9 for the vocal fold model. To synthesize speech, users can directly specify values for them every 10 milliseconds. Alternatively, users can use a gestural score (an organized pattern of articulatory gestures) to indirectly control the articulation process. The realization of each phoneme is cooperatively governed by multiple gestures, each of which consists of three parameters [7]: a gesture *value*, a *duration*, and a *time constant*, which define target positions of articulators, their duration, and how quickly the participating articulators reach the targets (i.e., speaking effort), respectively. All involved gestures for the the realization of an utterance constitute its gestural score. In fact, the gestural score is internally converted to articulatory trajectories in VTL, because the motions of articulators in response to discrete gestures are controlled by linear dynamical systems, thus producing articulatory trajectories along the time-axis. Subsequently, acoustic signals are simulated by a time-varying branched acoustic tube system. VTL conducts the articulatory-to-acoustic conversion, thus preparing the training samples for the LSTM.

### 2.3 Creation of training samples

As shown in Figure 1, we first created gestural scores and then converted them to articulatory trajectories as well as corresponding acoustic signal with VTL. From the text, we first used the WebMAUS G2P service [8][9] to obtain the SAMPA sequence. Then, each phoneme was mapped to its participating gestures being specified by Kohler's German phoneme intrinsic duration and preferred time constants. Next, all gestures within an utterance were organized according to the time structure model of the syllable [10], thus producing a gestural score (denoted as "prototype", compared to its variants as introduced later). The details of this process are described in [11]. The only difference is the setting of $f0$. For each utterance, we combined every three syllables into one group, with 81.7, 79.7 and 77.7 semitones for values of pitch targets of the first, second, and third syllables, respectively. All created gestural scores were then fed into VTL to generate articulatory trajectories and corresponding synthetic speech. To assess the quality of the synthetic speech created with such a strategy, we examined the intelligibility with the WebMAUS automatic speech recognition (ASR) service [12]. We created 1681 gestural scores with the sentences from the BITS corpus [13], which were fed into VTL to synthesize

**Table 1** – articulatory parameters of VocalTractLab synthesizer

| Models | parameters | Description |
|---|---|---|
| Vocal tract model | $HX, HY$ | Horiz. and vert. hyoid positions |
| | $JX, JA$ | Jaw position and Jaw angle |
| | $LP, LD$ | Lip protrusion and vert. lip distance |
| | $VS, VO$ | Velum shape and velum opening |
| | $WC$ | Wall compilance |
| | $TCX, TCY$ | Horiz. and vert. tongue body center positions |
| | $TTX, TTY$ | Horiz. and vert. tongue tip positions |
| | $TBX, TBY$ | Horiz. and vert. tongue blade positions |
| | $TRX, TRY$ | Horiz. and vert. tongue root positions |
| | $MA1$ - $MA3$ | Minimal area for tongue body, tongue tip and teeth-lips |
| | $TS1$ - $TS4$ | Tongue side elevation (from the anterior to the posterior) |
| Vocal fold model | $f0$ | fundamental frequency |
| | $P_{sub}$ | subglottal pressure |
| | $X_{bottom}$ | lower displacement |
| | $X_{top}$ | upper displacement |
| | $chink\_area$ | chink area |
| | $lag$ | phase lag |
| | $rel\_amp$ | relative amplitude |
| | $double\_pulsing$ | double pulsing |
| | $aspiration\_strength$ | aspiration strength |

the corresponding acoustic signals. The ASR results reached a word recognition accuracy of 52.75% for the synthetic speech, compared to that of 76.6% for the original human-produced speech in BITS corpus.

However, the LSTM regression model trained with such data is not robust against different acoustic variations. For one thing, the duration and time constants used are fixed for specifying the same gestures. Obviously, they vary a lot with their contexts. For another thing, all data were produced by a single speaker (the VTL model speaker), thus having a similar speaking style and voice quality. To tackle these limitations, we expanded the acoustic space and/or articulatory space by introducing several variants for each utterance in several manners. First, we increased or decreased the values of time constants by 30%, with all else being unchanged, for all gestures in the "prototype" gestural scores. Second, we overall increased or decreased the values of pitch targets by 3 semitones on the basis of the "prototype" gestural scores. Third, we substituted the "modal" phonation setting with "slightly-pressed" or "slightly-breathy" settings. Each combination of such operations resulted in a variant of the gestural score such that its corresponding articulatory and acoustic representations formed a new training sample. Last, on the basis of synthetic speech, we further manipulated the vocal tract length, by setting factors equal to 0.8 and 1.2 with the "change gender" functionality of Praat, while keeping its articulatory trajectories unchanged. Therefore, we created 81 training samples (three time constants $\times$ three pitch levels $\times$ three phonation settings $\times$ three vocal tract lengths) for each utterance.

## 3 Experiments and Results

### 3.1 Dataset

To create the training samples for the LSTM, we selected the first 2,000 sentences from a language model corpus used in an speech recognition system [14]. For each sentence, we created 81 variants with the strategy described in Sec. 2.3. We obtained 16,000 training utterances

(~591.8 hours speech). The data was randomly split into the actual training set (95%) and evaluation set (5%). For the test set, we used both the synthetic and human speech. The first 120 of 1681 synthetic utterances of BITS texts are used as synthetic test samples (40,901 frames, ~0.11 hours). In addition, we tested some human utterances to examine the generalization capacity of the trained model, including 80 sentences from the BITS corpus [13], 10 sentences from the Berlin Emotional Database [15], 30 sentences from the MMASCS database [16], and 10 words produced in a carrier sentence. These are in total 130 natural utterances produced by 7 speakers (5 males and 2 females), containing 47,580 frames (~0.13 hours speech).

We extracted 42 acoustic features from acoustic signals: 13 Mel-Frequency Cepstral Coefficients (MFCC) and 1 voiced/unvoiced probability extracted using STRAIGHT [17] as well as their first and second order derivatives. These features were extracted from a 20-millisecond-length window shifted every 10 milliseconds. VTL synthesizes acoustic signals with articulatory trajectories every 10 milliseconds. We assumed that one frame of the acoustic signal is controlled by the corresponding frame of articulatory parameters so that they constitute one training sample of the LSTM networks. The acoustic features were first z-score normalized per sentence during feature extraction, and then they as well as articulatory trajectories were further normalized to the range [0, 1] for the whole dataset when applied to LSTM networks.

### 3.2 Experimental setup

Some settings of the neural network regression model parameters are listed in Table 2. The input and output of the LSTM are 42 acoustic features and 30 articulatory control parameters[1], respectively. Each training batch had 20,000 training samples/frames. The length of time steps was 100 frames, thus one second of speech. The mean square error (MSE) between outputs and targets was used as the loss function. We trained the model for 100 epochs. Figure 2 shows the MSE loss on evaluation and test sets after each epoch. To compare the performance due to model architecture, we developed three systems: system-1 (indicated with red solid line) has 2 hidden layers and 256 nodes per layer, system-2 (indicated with blue dotted line) has 3 hidden layers and 256 nodes per layer, and system-3 (indicated with black dashed line) has 2 hidden layers and 512 nodes per layer. The system-3 performed best and was used to further analyze the results.

**Table 2** – The settings of neural network regression model parameters.

| **Parameters** | Input_dimension | Output_dimension | Batch_size | Sequence_length | Epoch |
|---|---|---|---|---|---|
| **Values** | 42 | 30 | 20,000 | 100 | 100 |

### 3.3 Results

We evaluated the performance of the copy synthesis system in terms of three metrics: the correlation coefficients and RMSE of articulatory parameters (between estimated ones from speech and the real ones), and the ASR accuracy. The experiment with synthetic testing utterances showed that, for the concerned 30 articulatory parameters of VTL, the correlation coefficients ranged from 0.18 to 0.973 and the RMSE ranged from 0.043 to 0.255. The estimated articulatory parameters were further fed into VTL, whose output speech achieved a word recognition accuracy of 17.24%. Figure 3 shows the comparison of originally synthesized utterance and its

[1]The parameters wall compliance ($WC$), phase lag ($lag$), and fundamental frequency ($f0$) were excluded as they do not change withing the dataset.

reproduced synthetic counterpart for the word “keinerlei”. They are similar in terms of both acoustic and articulatory representation. The human test utterances were evaluated in terms of ARS accuracy, achieving 7.6% word accuracy.

A further examination in detail revealed that the trained LSTM model reproduced most articulatory parameters well, except for velum shape ($VS$), tongue root vertical positon ($TRY$) and subglottal pressure ($P_{sub}$). Especially, the $P_{sub}$ defined from gestural score always equaled 800 Pascal in non-silent part and gradually increased from or decreased to 0 Pascal in the utterance initial and final part, respectively, while the estimated $P_{sub}$ fluctuated a lot, thus leading to a very low correlation coefficient of 0.18. What happened due to the unsmooth subglottal pressure trajectory was the introduction of noise during the acoustic simulation. This may partially explain why the human-produced BITS speech has a little better quality than originally synthetic BITS speech in terms of recognition accuracy (76.6% versus 52.75%) while the reproduced synthetic speech has only a word recognition accuracy of 17.24%. The worst case occurred for the reproduced human speech from estimated articulatory trajectories. In addition to the noise introduced by unsmooth trajectories, the speaker variation may be another factor. The reason is that the synthetic speech both in training set and test set was produced by the same speaker (i.e. VTL), thus resulting in similar acoustic characteristics. However, the acoustic difference between VTL speaker and human speakers affects the performance.

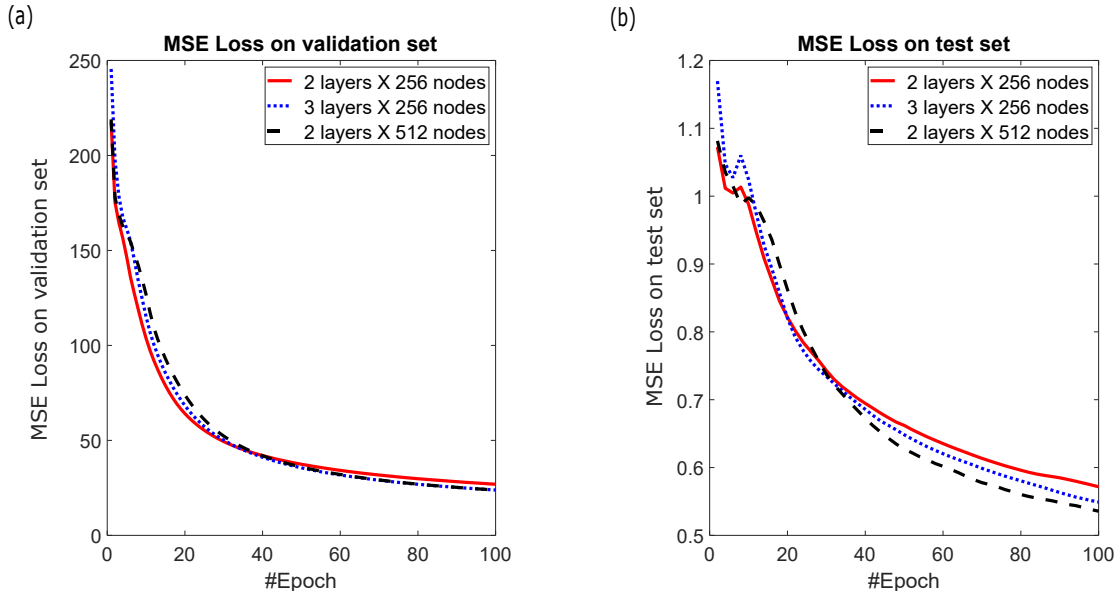


**Figure 2** – Mean square error (MSE) between outputs and targets after each epoch on: (a) the validation set and (b) the test set. The performance of the neural network regression model is compared with different settings for the number of hidden layers and nodes per layer.

## 4 Conclusion and future work

We proposed a framework for copy synthesis with an artificial neural network (LSTM regression model) and an articulatory speech synthesizer (VocalTractLab), which are responsible for the acoustic-to-articulatory mapping and the inversion, respectively. With a rule-based method and VTL, we created the gestural scores which were converted to articulatory trajectories and subsequently simulated to the corresponding acoustic signal. They constituted the training samples for the LSTM regression model. From the acoustic features of test speech, the trained model can estimate articulatory parameters with high correlation coefficients and low RMSE except for some specific parameters. However, the estimated articulatory trajectories are not as

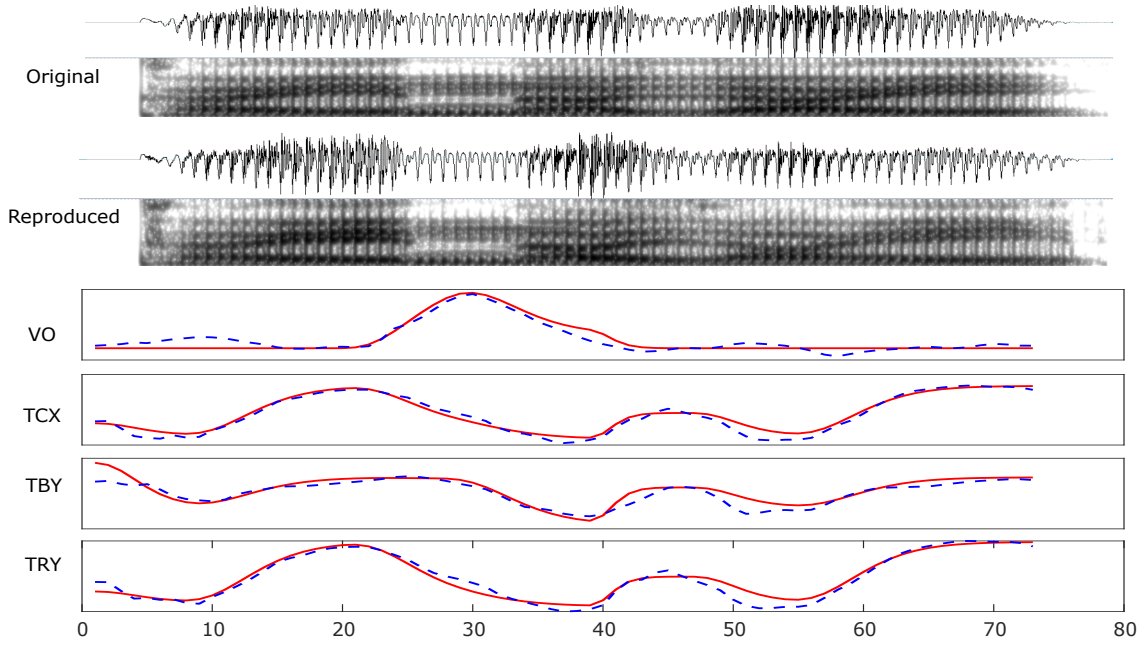


**Figure 3** – The comparison of original and reproduced utterances, for the word "keinerlei", form the articulatory and acoustic aspects. The top panel shows the oscillogram and spectrogram of testing speech originally synthesized with VocalTractLab. Below it is the the counterparts of the reproduced speech from estimated articulatory trajectories with VTL. The rest part of this figure are the VO, TCX, TBY, TRY trajectories with red solid lines for reference trajectories of originally synthesized speech and blue dashed lines for estimated ones.

smooth as the ones converted from original gestural scores, thus making the quality of reproduced synthetic speech not very high. Punishing the parameter fluctuation in the loss function is worth considering in future work. Besides, although we employed multiple variants of gestural scores or the speech signal to alleviate the speaker variation, the generalization capacity of trained model is limited. In the present work, we only manipulated some variables with several discrete values, the resulting training samples did not cover the full articulatory and acoustic space. Therefore, we will widen the articulatory and acoustic space by manipulating variables with random values, for example, sampled from Gaussian distributions.

## 5 Acknowledgements

This research work is partially sponsored by China Scholarship Council. We thank the Center for Information Services and High Performance Computing (ZIH) at TU Dresden for generous allocations of computing resources.

## References

[1] GAO, Y., S. STONE, and P. BIRKHOLZ: *Articulatory copy synthesis based on a genetic algorithm. Proc. Interspeech 2019*, pp. 3770–3774, 2019.

[2] ERICSDOTTER, C.: *Articulatory copy synthesis: Acoustic performance of an MRI and X-ray based framework.* In *Proceedings of the XVth ICPhS*, pp. 2909–2912. 2003.

[3] PHILIPPSEN, A. K., R. F. REINHART, and B. WREDE: *Learning how to speak: Imitation-based refinement of syllable production in an articulatory-acoustic model.* In *4th International Conference on Development and Learning and on Epigenetic Robotics*, pp. 195–200. IEEE, 2014.

[4] DANG, J. and K. HONDA: *Estimation of vocal tract shapes from speech sounds with a physiological articulatory model. Journal of Phonetics*, 30(3), pp. 511–532, 2002.

[5] ELIE, B. and Y. LAPRIE: *Copy synthesis of running speech based on vocal tract imaging and audio recording*. In *ICA 2016 - 22nd International Congress on Acoustics*. 2016.

[6] BIRKHOLZ, P.: *Modeling consonant-vowel coarticulation for articulatory speech synthesis. PloS one*, 8(4), p. e60603, 2013.

[7] BIRKHOLZ, P., I. STEINER, and S. BREUER: *Control concepts for articulatory speech synthesis*. In *Proceedings of the 6th ISCA Workshop on Speech Synthesis, Bonn, Germany*, pp. 5–10. 2007.

[8] REICHEL, U. D.: *Perma and balloon: Tools for string alignment and text processing*. In *Interspeech, 13th Annual Conference of the International Speech Communication Association*, pp. 1874–1877. 2012.

[9] REICHEL, U. D. and T. KISLER: *Language-independent grapheme-phoneme conversion and word stress assignment as a web service. Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2014*, pp. 42–49, 2014.

[10] XU, Y. and F. LIU: *Tonal alignment, syllable structure and coarticulation: Toward an integrated model. Italian Journal of Linguistics*, 18(1), pp. 125–159, 2006.

[11] GAO, Y., H. DING, P. BIRKHOLZ, R. JÄCKEL, and Y. LIN: *Perception of German tense and lax vowel contrast by chinese learners. Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2019*, pp. 25–32, 2019.

[12] KISLER, T., U. REICHEL, and F. SCHIEL: *Multilingual processing of speech via web services. Computer Speech & Language*, 45, pp. 326–347, 2017.

[13] ELLBOGEN, T., F. SCHIEL, and A. STEFFEN: *The BITS speech synthesis corpus for German*. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, vol. 47, p. 40. European Language Resources Association (ELRA), Lisbon, Portugal, 2004.

[14] RADECK-ARNETH, S., B. MILDE, A. LANGE, E. GOUVÊA, S. RADOMSKI, M. MÜHLHÄUSER, and C. BIEMANN: *Open source German distant speech recognition: Corpus and acoustic model*. In *International Conference on Text, Speech, and Dialogue*, pp. 480–488. Springer, 2015.

[15] F. BURKHARDT, M. R. W. F. S., A. PAESCHKE and B. WEISS: *A database of German emotional speech.* In *Interspeech*, vol. 5, pp. 1517–1520. 2005.

[16] D. SCHABUS, M. P. and P. HOOLE: *The MMASCS multi-modal annotated synchronous corpus of audio, video, facial motion and tongue motion data of normal, fast and slow speech.* In *LREC*, pp. 3411–3416. 2014.

[17] KAWAHARA, H., M. MORISE, T. TAKAHASHI, R. NISIMURA, T. IRINO, and H. BANNO: *Tandem-straight: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation*. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3933–3936. IEEE, 2008.