Realistic Ultrasound Tongue Image Synthesis using Generative Adversarial Networks

Nadia Hajjej[1], Tamás Gábor Csapó[1,2]

[1] *Department of Telecommunications and Media Informatics,*
Budapest University of Technology and Economics
[2] *MTA-ELTE „Lendület" Lingual Articulation Research Group*

Abstract

Ultrasound Tongue Imaging (UTI) is a technique suitable for the acquisition of articulatory data, showing the motion of the tongue. When the subject is speaking, the ultrasound transducer is placed below the chin, resulting in mid-sagittal images of the tongue movement. The typical result of 2D ultrasound recordings is a series of grayscale images in which the tongue surface contour has a greater brightness than the surrounding tissue and air. UTI has been used for many years in phonetic research on speech production. However, these studies are mostly based on manually annotated articulatory data, and reliable extraction of high-level features from ultrasound data remains a challenge. In this paper, we propose a method to generate realistic ultrasound images from a database of midsagittal images of the tongue. First, we explain the principle of Generative Adversarial Networks (GAN), which is a subset of generative models, where deep neural networks are applied. Then, we detail our method, starting with the properties of the dataset, to the conception of the convolutional neural network model. The model consists of a generator and a discriminator network, which are trained against each other in the task of realistic image generation: the generator tries to fool the discriminator. The experiments demonstrate the efficiency of the GAN in creating realistic images when the training is run long enough, in order that the generator network can learn the properties of ultrasound images. The GAN-generated images were tested with a subjective test, and it supported our hypothesis that the synthesized ultrasound tongue images are of high quality and are difficult to distinguish from real images of the tongue. The results can be exploited for data augmentation, for predicting the next frame in a UTI sequence or for motion detection of tongue contours within images.

1. Introduction

Ultrasound tongue imaging (UTI) is a technique suitable for the acquisition of articulatory data. Stone (2005) summarized the typical methodology of investigating speech production using ultrasound. Usually, when the subject is

Email addresses: hajjej.nadia@gmail.com (Nadia Hajjej), csapot@tmit.bme.hu (Tamás Gábor Csapó)

Beszédtudomány – Speech Science 2020. 7–21. DOI: 10.15775/Besztud.2020.7-21

speaking, the ultrasound transducer is placed below the chin, resulting in midsagittal images of the tongue movement. The typical result of 2D ultrasound recordings is a series of gray-scale images in which the tongue surface contour has a greater brightness than the surrounding tissue and air (for a sample, see Figure 1). Although a large number of linguistic studies are applying 2D ultrasound (Stone, 2005), there are not many freely available databases with a large number of images. Eshky et al. (2018) introduced a database that is related to the Ultrax2020 project (`http://www.ultrax-speech.org/ultrasuite`), but it contains ultrasound images of children only. This UltraSuite repository currently contains tongue ultrasound data from 5–12 year old children who are typically developing or have a speech sound disorder. Another UTI dataset is related to Silent Speech Interfaces (Ji et al., 2018), but it contains processed ultrasound images and not the original raw data.

Ultrasound imaging of the tongue has been used for many years in research on speech production (Stone, 2005). For some relevant experiments of the MTA-ELTE „Lendület” Lingual Articulation Research Group, see Markó et al. (2017, 2018, 2019a) and Markó et al. (2019b). However, these studies are based on manually annotated articulatory data, and reliable extraction of high-level features from ultrasound data (e.g. automatic tongue contour tracking) remains a challenge (Csapó & Lulich, 2015; Csapó & Csopor, 2015; Xu et al., 2017b). The topic of the current study, i.e. the realistic synthesis of ultrasound images can be a starting point for exploiting the higher-level representation of the tongue in a variety of applications in speech research.

Ever since computers were developed, scientists and engineers thought of artificially intelligent systems that work and react like humans. In the past decades, the increase of generally available computational power provided a helping hand for developing fast learning machines. Meanwhile, the internet supplied an enormous amount of data for training. These two developments boosted the research on smart self-learning systems, with neural networks among the most promising techniques.

Recently, deep neural networks have produced high accuracy scores in speech and ultrasound-related tasks, such as articulatory-to-acoustic mapping (Csapó et al., 2017b), articulation-to-text mapping (Xu et al., 2017a; Tóth et al., 2018), articulation-to-F0 prediction (Grósz et al., 2018; Csapó et al., 2019), acoustic-to-articulatory inversion (Porras et al., 2019) and also edge (contour) detection (Csapó & Csopor, 2015; Xu et al., 2017b). For these problems, usually, regression models are used, which can be trained for mapping from the input to the target feature. Typical networks are fully-connected feed-forward deep neural networks, convolutional neural networks, and recurrent neural networks.

1.1. Generative models

A branch of deep learning methods deals with generative models, i.e. how to generate new data that is similar to the properties of the training data. In general, the goal of the generative models is to estimate or to learn the data distribution of the training data and generate new data points with some variations by modeling a distribution, which is as much as possible close to the real data distribution. Most of the generative models use the maximum likelihood method to define a model that estimates the parametrized probability distribu-

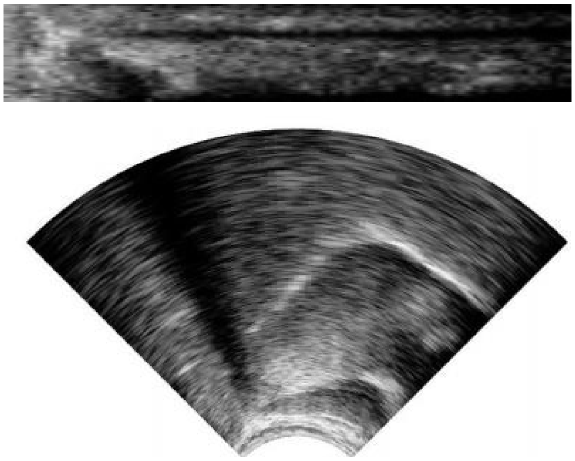


Figure 1: Ultrasound image of the tongue. Top: raw scanline data. Bottom: 'wedge' format (the tongue root is on the left, while the tongue tip on the right).

tion (Goodfellow, 2017). Those models differ mainly in the approximation of maximum likelihood. There are two main types: 1) models that aim to represent the probability distribution over the space where the data lies explicitly, and 2) models that interact implicitly with the probability distribution and try to generate samples from it. Since the introduction of Generative Adversarial Networks (GANs) by Goodfellow et al. (2014), they have proven a vast potential to automatically learn the natural features of a particular dataset, to mimic any data distribution and generate data like it. Typical examples include generated digits, flowers, realistic human faces (Karras et al., 2018), or speech data for emotion recognition (Chatziagapi et al., 2019).

In this paper, we aim to synthesize realistic ultrasound tongue images using Generative Adversarial Networks, that belong to the second type of the above generative models. The GAN-generated ultrasound images can be useful for data augmentation, which might be necessary for scenarios with limited data, e.g. for motion detection of tongue contours within images (Xu et al., 2017b) or articulatory-to-acoustic mapping (Csapó et al., 2017b).

2. Methods

2.1. GAN framework

The purpose of GANs is to create samples, which are able to deceive humans and even computers. Thus, the main idea of GAN is to set up a game between two players: the generator and the discriminator (Goodfellow et al., 2014; Goodfellow, 2017). The generator is the player that creates samples. Those samples are intended to come from the same distribution as the training data. The discriminator is the second player that examines samples to decide whether they are real or fake. The discriminator usually learns using traditional supervised learning techniques to divide inputs into two classes (real or fake). Figure 2 shows the block diagram of a GAN with sample real and generated ultrasound tongue images. Each of our players has its own differentiable function with respect to its parameters and its inputs. The discriminator has a function D

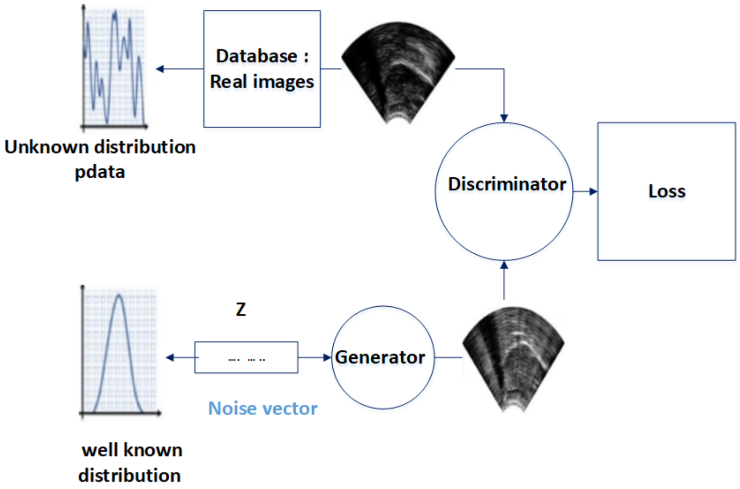


Figure 2: Block diagram of a GAN with discriminator and generator networks, used for the ultrasound tongue image synthesis task.

that takes x as input and uses θ_D as parameters. The generator is defined by a function G that takes z as input and uses θ_G as parameters. Both players have cost functions that are defined in terms of both players' parameters. This cost function is defined by the following equation:

$$min_{\theta_G} max_{\theta_D}(E_x log(D_{\theta_D}(x) + E_z log(1 - D_{\theta_D}(G_{\theta_G}(z)))$$

Now let us explain this relation. We have a real image x that will be examined by the discriminator D. For this image, it will give a value close to zero. Hence, for a fake image, it will give a higher value close to one. For the generator G, he will take a randomly generated vector from a very simple and well-known distribution and produce an image that will also be used to train the discriminator. The latter will be alternatively shown real and fake images. The generator's role is to minimize the output of D by providing more realistic images, while D tries to maximize the same thing. Each player's cost depends on the other player's parameter, but they can only control their own parameter. This scenario is most straightforward to describe as a 'minimax' game where the solution is a Nash equilibrium (Goodfellow et al., 2014; Goodfellow, 2017).

2.2. Dataset

Before building the required neural network, we chose the dataset from which we are aiming to generate similar samples. This dataset contains tongue-ultrasound images. For training the GAN, we used ultrasound tongue images from a previously collected database (Csapó et al., 2017a), which applied the ”Micro” ultrasound system (Articulate Instruments Ltd., UK). The database contains raw midsagittal ultrasound images of the scanline data (see Fig. 1 left) recorded at 82 fps from several speakers, of which we chose one Hungarian female speaker for our experiments and used 209 sentences altogether. Each pixel is stored as a 1-byte unsigned integer, which is actually a grayscale pixel intensity. Using the extracted raw ultrasound images, we can convert and visualize them as ultrasound frames in the ’wedge’ format (see Fig. 1 right). Those frames can be used to produce a video illustrating the movement of the tongue.

In our case, we are interested in using the raw ultrasound images, as they contain the information before any image processing. Therefore, after successfully extracting those images, we built a data set of 27 925 raw images having the dimension of 64×842 pixels. We split the image set into two groups. The first group is made of 2/3 of images, which is used for training, and the second one made of 1/3 of the dataset used for testing. Figure 1 illustrates a sample of raw ultrasound images and ultrasound frames as well.

2.3. Proposed method for ultrasound tongue image generation

As we have mentioned, the principle of Generative Adversarial Networks is managing a game between the two networks (i.e. the generator and the discriminator). We implemented this in Python using a DC-GAN implementation as a starting point (`https://github.com/carpedm20/DCGAN-tensorflow/`). For our project, we chose to use a deep convolutional neural network for both the generator and the discriminator. In our model, we fixed the number of hidden layers to nine for both discriminator and generator. As hyperparameter, we fixed 64 as batch size and 25 for the number of epochs. After every 100 iterations, we generated 64 images.

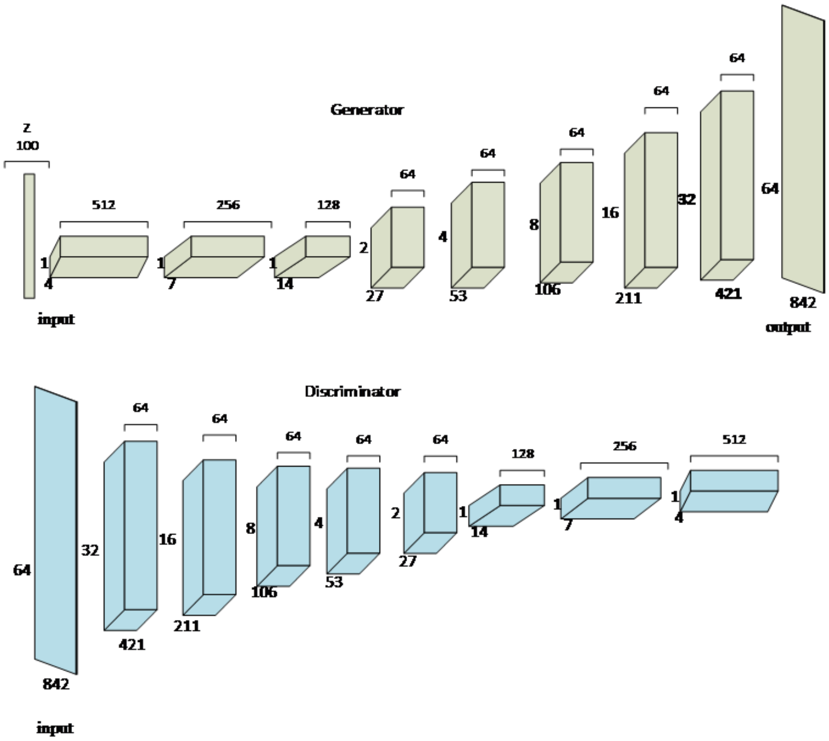
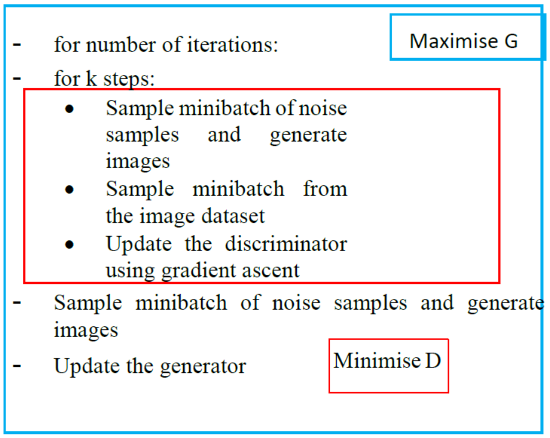


Figure 3: The architecture of the Generator (top) and Discriminator (bottom) networks within the GAN.



```
- for number of iterations:                         Maximise G
- for k steps:
    • Sample minibatch of noise samples and generate images
    • Sample minibatch from the image dataset
    • Update the discriminator using gradient ascent
- Sample minibatch of noise samples and generate images
- Update the generator                              Minimise D
```

Figure 4: Pseudo code of GAN training.

Figure 3 illustrates the GAN network architecture. The discriminator (Fig. 3 bottom) is a downsampling network using strided convolutional layers. Its role is to check real images and save variables in order to use them in fake image checking. The discriminator has a last convolutional layer before applying the cross-entropy. On the other hand, the generator (Fig. 3 top) is an upsampling neural network that takes as input a vector z randomly generated from a known distribution. After linearly transforming z, it will be fed through the layers in order to get as a result an image with the same size as our original image. The generated image will be the input of the discriminator, and according to its result, the networks will improve their performance. There is one discriminator update per each generator update. The process can be summarized with the following pseudo-code shown in Figure 4.

Obviously, the quality of the GAN-generated images was low in early epochs and continuously improved during the training until the final epochs, because as we have written, during the training, the generator improves its performance.

3. Experiments and results

As we have previously mentioned, after every 100 iterations during the GAN training, we generated 64 ultrasound images. The raw images were converted to the 'wedge' format for visualization. In order to assess the quality of our images, we created an internet-based test (`http://leszped.tmit.bme.hu/gan2018/`). In this test, we used 100 hand-selected samples, including 20 real and 80 generated images, chosen by visual inspection to ensure that there are different images in the experiment. The latter is made of 20 'early' samples created after the early iterations of the training, and 60 'late' ones generated from the last iterations. The task of the participants was to assess the quality and reality of the images on a scale between 0–100, without knowing whether they are real or generated. Thus, as a result, we will have numbers associated to images describing their quality. Figures 5–7 show several examples from the subjective evaluation process, while Figure 8 presents a sample image with the question provided.

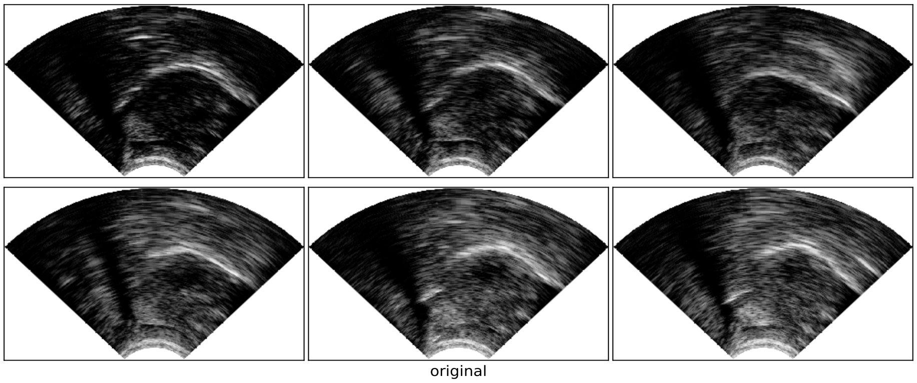


Figure 5: Original Ultrasound Tongue Images.

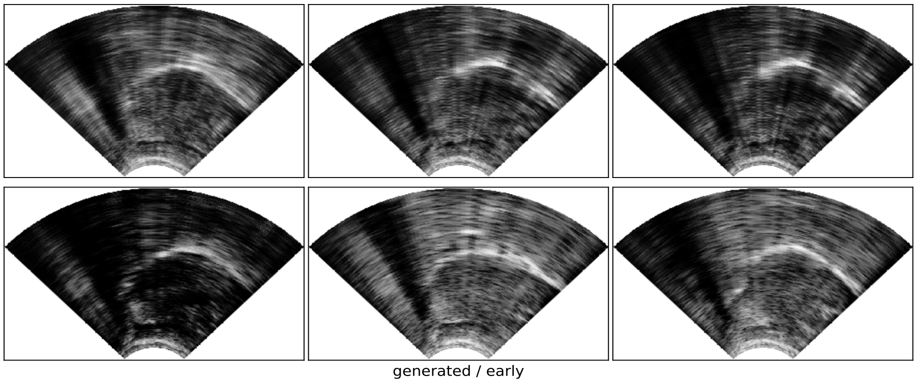


Figure 6: Generated Ultrasound Tongue Images from early epochs.

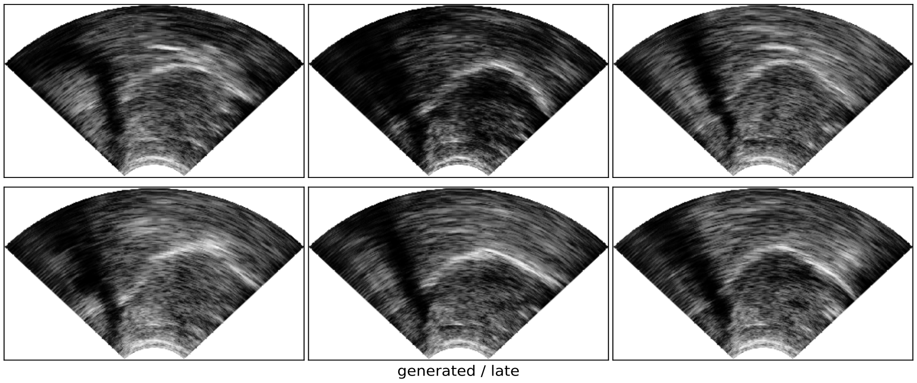
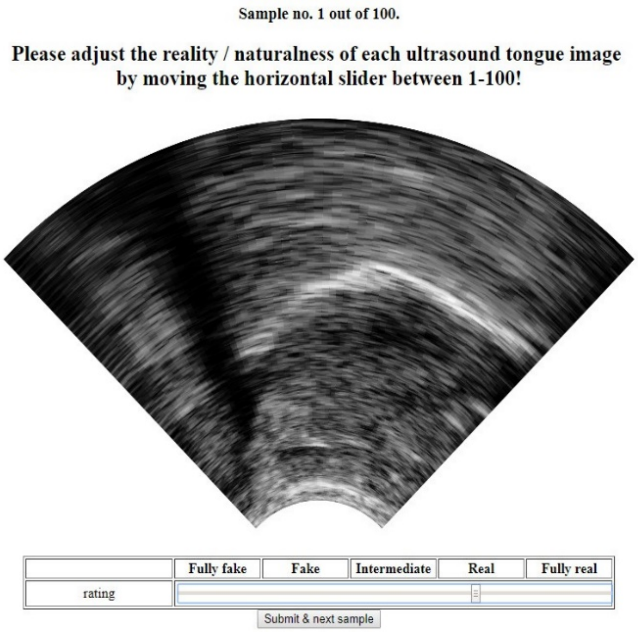


Figure 7: Generated Ultrasound Tongue Images from late epochs.



Sample no. 1 out of 100.

Please adjust the reality / naturalness of each ultrasound tongue image by moving the horizontal slider between 1-100!

	Fully fake	Fake	Intermediate	Real	Fully real
rating					

Figure 8: A sample generated image from the subjective test.

In Fig. 6, the 'early' generated images can be mostly distinguished from the 'late' images of Fig. 7, as in the early epochs, the generator within the GAN was not able to produce realistic images. On the other hand, the 'late' generated images (Fig. 7) are visually close to the original ultrasound images of the tongue (Fig. 5). In the figures, the shapes of the tongues are different (as they are isolated examples of real or synthetic images), and would produce different sounds. This means that while the model generates realistic looking ultrasound images, they are not constrained linguistically, which could be addressed in future work.

A total of 8 subjects, blinded to the approaches, participated in the subjective test, three of them being speech researchers and the remaining five being university students. The test took, on average, 13 minutes to complete. The test results are summarized in Table 1. Analyzing the results, we can see that the images generated from the 'early' epochs were evaluated with low scores (around 29%). In contrast, the tongue images from the 'late' epochs reached 59%, which

Table 1: Mean and standard deviation result of the subjective test for the 'reality / naturalness' question. Higher scores are better.

	early images	late images	real images
No. of images	20	60	20
Average result (experts)	17.57 (14.63)	63.28 (23.54)	68.77 (25.09)
Average result (non-experts)	35.65 (29.69)	56.79 (22.13)	68.06 (19.29)
Average result (all)	28.87 (26.61)	59.23 (22.89)	68.33 (21.66)

is close to the quality of the real ultrasound images (being 68%). The three experts were more strict: they evaluated the 'early' images with lower scores, and the 'late' images with higher scores than the non-experts. Therefore, we can say that the GANs are efficient in ultrasound tongue image generation and deceive humans, as the results showed how hard it is to differentiate between real and generated images.

4. Discussion

Generative Adversarial Networks, being a subfield of generative models within machine learning, are suitable to synthesize new images which are similar to the training data. Typical example uses of GANs include generated digits, flowers, or human faces (Karras et al., 2018). In this study, we presented a pioneering work in ultrasound tongue image synthesis using GANs.

According to the experiments, the Generative Adversarial Networks are able to generate realistic tongue ultrasound images. Therefore, the results can be useful for data augmentation. This might be important for scenarios with limited data, e.g. for motion detection of tongue contours within images (Xu et al., 2017b) or articulatory-to-acoustic mapping (Csapó et al., 2017b). One potential issue is that errors in synthetic data can propagate into future models trained on this data, something we need to be careful about not just in medical applications but in general. Besides, a conditional GAN with a similar architecture could be used for predicting the next frame in a UTI sequence (Wu et al., 2018), or can be useful for acoustic-to-articulatory inversion (Porras et al., 2019).

5. Conclusion and future work

In this paper, we have shown in detail our method aiming to synthesize realistic ultrasound images. First, we have explained the principle of Generative Adversarial Networks. Then, we have detailed our method, starting with the creation of the dataset, to the conception of the network model, and finally, the investigation of the obtained results.

The performance shown by the GANs in generating realistic tongue ultrasound images encourages us to improve the used model by taking into consideration the time dimension, to be able to predict the next input for the generator (e.g. as a form of a recurrent neural network) which may enhance the performance and get better and accurate results. In the future, we plan to train generative networks conditioned on the linguistic content, and test the GAN-based methods on other types of articulatory data (e.g. vocal tract MRI and lip images).

Acknowledgement

The authors were partially funded by the National Research, Development and Innovation Office of Hungary (FK 124584 and PD 127915) and by the MTA „Lendület" program. We would like to thank the subjects taking part in the subjective experiment. The Titan X GPU used for this research was donated by NVIDIA Corporation.

References

Chatziagapi, A., Paraskevopoulos, G., Sgouropoulos, D., Pantazopoulos, G., Nikandrou, M., Giannakopoulos, T., Katsamanis, A., Potamianos, A., & Narayanan, S. (2019). Data Augmentation Using GANs for Speech

Emotion Recognition. In *Proc. Interspeech* (pp. 171–175). Graz, Austria. URL: `http://www.isca-speech.org/archive/Interspeech{_}2019/abstracts/2561.html`. doi:`10.21437/Interspeech.2019-2561`.

Csapó, T. G., Al-Radhi, M. S., Németh, G., Gosztolya, G., Grósz, T., Tóth, L., & Markó, A. (2019). Ultrasound-based Silent Speech Interface Built on a Continuous Vocoder. In *Proc. Interspeech* (pp. 894–898). Graz, Austria. doi:`10.21437/Interspeech.2019-2046`. `arXiv:1906.09885`.

Csapó, T. G., & Csopor, D. (2015). Ultrahangos nyelvkontúr követés automatikusan: a mély neuronhálókon alapuló AutoTrace eljárás vizsgálata [Automatic tongue contour tracking based on ultrasound: investigation of the Deep Neural Network based AutoTrace method] (in Hungarian). *Beszédkutatás 2015 [Speech Research 2015]*, *1*, 177–187.

Csapó, T. G., Deme, A., Gráczi, T. E., Markó, A., & Varjasi, G. (2017a). Synchronized speech, tongue ultrasound and lip movement video recordings with the "Micro" system. In *Challenges in analysis and processing of spontaneous speech*.

Csapó, T. G., Grósz, T., Gosztolya, G., Tóth, L., & Markó, A. (2017b). DNN-Based Ultrasound-to-Speech Conversion for a Silent Speech Interface. In *Proc. Interspeech* (pp. 3672–3676). Stockholm, Sweden. URL: `http://dx.doi.org/10.21437/Interspeech.2017-939`. doi:`10.21437/Interspeech.2017-939`.

Csapó, T. G., & Lulich, S. M. (2015). Error analysis of extracted tongue contours from 2D ultrasound images. In *Proc. Interspeech* (pp. 2157–2161). Dresden, Germany.

Eshky, A., Ribeiro, M. S., Cleland, J., Richmond, K., Roxburgh, Z., Scobbie, J. M., & Wrench, A. (2018). UltraSuite: A Repository of Ultrasound and Acoustic Data from Child Speech Therapy Sessions. In *Proc. Interspeech* (pp. 1888–1892). Hyderabad, India: ISCA. URL: `http://www.isca-speech.org/archive/Interspeech{_}2018/abstracts/1736.html`. doi:`10.21437/Interspeech.2018-1736`.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27 (NIPS 2014)* (pp. 2672–2680). URL: `https://papers.nips.cc/paper/5423-generative-adversarial-nets`.

Goodfellow, I. J. (2017). NIPS 2016 Tutorial: Generative Adversarial Networks. *CoRR*, *abs/1701.0*. URL: `http://arxiv.org/abs/1701.00160`. `arXiv:1701.00160`.

Grósz, T., Gosztolya, G., Tóth, L., Csapó, T. G., & Markó, A. (2018). F0 Estimation for DNN-Based Ultrasound Silent Speech Interfaces. In *Proc. ICASSP* (pp. 291–295). Calgary, Canada.

Ji, Y., Liu, L., Wang, H., Liu, Z., Niu, Z., & Denby, B. (2018). Updating the Silent Speech Challenge benchmark with deep learning. *Speech Communication*, *98*, 42–50. doi:`10.1016/j.specom.2018.02.002`. `arXiv:1709.06818`.

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018). Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. URL: `https://openreview.net/forum?id=Hk99zCeAb`.

Markó, A., Bartók, M., Csapó, T. G., Deme, A., & Gráczi, T. E. (2019a). The effect of focal accent on vowels in Hungarian: Articulatory and acoustic data. In *Proc. ICPhS* (pp. 2715–2719). Melbourne, Australia.

Markó, A., Bartók, M., Gráczi, T. E., Deme, A., & Csapó, T. G. (2018). Mondathangsúlyos és hangsúlytalan helyzetű magánhangzók néhány artikulációs és akusztikai jellemzője a magyarban. *Beszédkutatás*, *26*, 85–109.

Markó, A., Csapó, T. G., Deme, A., Gráczi, T. E., & Bartók, M. (2019b). Gyermekek lingvális artikulációjának variabilitása magánhangzós nyelvkontúrok alapján. In *Az anyanyelv-elsajátítás folyamata hároméves kor után* (pp. 165–190). ELTE Eötvös Kiadó.

Markó, A., Csapó, T. G., Deme, A., Gráczi, T. E., & Varjasi, G. (2017). A gyermeki artikuláció vizsgálata – Új lehetőségek a hazai kutatásban. In *Új utak a gyermeknyelvi kutatásokban* (pp. 65–95). Budapest, Hungary: ELTE Eötvös Kiadó. URL: http://www.eltereader.hu/media/2017/11/Bona{_}Gyermeknyelv{_}READER.pdf.

Porras, D., Sepúlveda-Sepúlveda, A., & Csapó, T. G. (2019). DNN-based Acoustic-to-Articulatory Inversion using Ultrasound Tongue Imaging. In *International Joint Conference on Neural Networks* (pp. N–19221). Budapest, Hungary. URL: http://arxiv.org/abs/1904.06083. arXiv:1904.06083.

Stone, M. (2005). A guide to analysing tongue motion from ultrasound images. *Clinical Linguistics & Phonetics*, *19*, 455–501. doi:10.1080/02699200500113558.

Tóth, L., Gosztolya, G., Grósz, T., Markó, A., & Csapó, T. G. (2018). Multi-Task Learning of Phonetic Labels and Speech Synthesis Parameters for Ultrasound-Based Silent Speech Interfaces. In *Proc. Interspeech* (pp. 3172–3176). Hyderabad, India. doi:10.21437/Interspeech.2018-1078.

Wu, C., Chen, S., Sheng, G., Roussel, P., & Denby, B. (2018). Predicting tongue motion in unlabeled ultrasound video using 3D convolutional neural networks. In *Proc. ICASSP*. Calgary, Canada.

Xu, K., Roussel, P., Csapó, T. G., & Denby, B. (2017a). Convolutional neural network-based automatic classification of midsagittal tongue gestural targets using B-mode ultrasound images. *The Journal of the Acoustical Society of America*, *141*, EL531–EL537. doi:10.1121/1.4984122.

Xu, K., Roussel, P., & Denby, B. (2017b). Is Speckle Tracking Feasible for Ultrasound Tongue Images? *Acta Acustica united with Acustica*, *103*, 365–368. doi:10.3813/AAA.919065.