DNN-based Ultrasound-to-Speech Conversion for a Silent Speech Interface

Tamás Gábor Csapó[1,2], *Tamás Grósz*[3], *Gábor Gosztolya*[3,4], *László Tóth*[4], *Alexandra Markó*[2,5]

[1]Department of Telecommunications and Media Informatics,
Budapest University of Technology and Economics, Budapest, Hungary
[2]MTA-ELTE Lendület Lingual Articulation Research Group, Budapest, Hungary
[3]Institute of Informatics, University of Szeged, Hungary
[4]MTA-SZTE Research Group on Artificial Intelligence, Szeged, Hungary
[5]Department of Phonetics, Eötvös Loránd University, Budapest, Hungary

csapot@tmit.bme.hu, {groszt, ggabor, tothl}@inf.u-szeged.hu, marko.alexandra@btk.elte.hu

Abstract

In this paper we present our initial results in articulatory-to-acoustic conversion based on tongue movement recordings using Deep Neural Networks (DNNs). Despite the fact that deep learning has revolutionized several fields, so far only a few researchers have applied DNNs for this task. Here, we compare various possible feature representation approaches combined with DNN-based regression. As the input, we recorded synchronized 2D ultrasound images and speech signals. The task of the DNN was to estimate Mel-Generalized Cepstrum-based Line Spectral Pair (MGC-LSP) coefficients, which then served as input to a standard pulse-noise vocoder for speech synthesis. As the raw ultrasound images have a relatively high resolution, we experimented with various feature selection and transformation approaches to reduce the size of the feature vectors. The synthetic speech signals resulting from the various DNN configurations were evaluated both using objective measures and a subjective listening test. We found that the representation that used several neighboring image frames in combination with a feature selection method was preferred both by the subjects taking part in the listening experiments, and in terms of the Normalized Mean Squared Error. Our results may be useful for creating Silent Speech Interface applications in the future.

Index Terms: articulatory, ultrasound, speech synthesis, deep neural networks, audiovisual speech processing

1. Introduction

During the past few years, there has been a significant interest in articulatory-to-acoustic conversion, which is often referred to as "Silent Speech Interfaces" (SSI) [1]. This has the main idea of recording the soundless articulatory movement, and automatically generating speech from the movement information, without the subject actually producing any sound. Such an SSI system can be highly useful for the speaking impaired (e.g. after laryngectomy), and for scenarios where regular speech is not feasible but information should be transmitted from the speaker (e.g. extremely noisy environments; military applications). For this automatic conversion task, typically electromagnetic articulography (EMA) [2, 3, 4], ultrasound [5, 6, 7, 8, 9, 10], permanent magnetic articulography [11], or multimodal approaches [12] are used. One of the first systems sought to find the relation between tongue movement recorded using ultrasound and spectral parameters of speech using a simple neural network [5], but the first results were not convincing, because the neural network was not able to handle this complex task. Later SSI systems applied the 'recognition-and-synthesis' approach; that is, they performed phone recognition based on articulatory movement, which was followed by text-to-speech synthesis [4, 6]. The drawback of this scenario might be that the error of the submodules adds up, which distorts the final speech generation output. Therefore, state-of-the-art SSI systems use the 'direct synthesis' principle, where the speech signal is generated without an intermediate step, directly from the articulatory data [2, 3, 7, 8, 9, 10, 11].

Recently, deep neural networks have produced accuracy scores better than or equal to human performance in several different visual recognition tasks, such as object detection [13], image classification [14] and edge (contour) detection [15], and also in automatic speech recognition (ASR). In the area of Silent Speech Interfaces, not many solutions have investigated deep learning. Diener and his colleagues used surface electromyographic (EMG) speech synthesis in combination with a deep neural network [16]. EMG channels captured from the face were fed into a five-layer feed-forward neural network with a bottleneck layer topology, resulting in a feature-extraction-followed-by-mapping structure. The target of the DNN regression was a 25 dimension Mel Frequency Cepstral Coefficient vector of the speech signal. A standard Mel Log Spectrum Approximation (MLSA) vocoder was used for the synthesis of speech. In their experiments, a slight improvement was found compared to the Gaussian Mixture Model (GMM) based mapping technique. Jaumard-Hakoun et al. focused on the case of singing and were able to synthesize sung vowels based on ultrasound and video of the lips [10]. First, they applied a multimodal Deep AutoEncoder to extract features from the tongue and lips images. After this, multilayer perceptron (MLP) networks were used to predict spectral features represented by Line Spectral Frequency (LSF) parameters, with a separate MLP being trained for each of the 12 LSF features. As the last step, a vocal tract model was built and articulatory-based singing voice synthesis was developed. Although the results of these studies are encouraging, further research is necessary to develop high-quality and real-time SSI systems.

According to our literature survey, there is a lack of tailored deep learning methods for SSI regarding both the selection of the optimal feature set from ultrasound-based articulatory data and the speech generation step. Therefore the current paper focuses on these two areas. We show that with simple fully-connected deep neural networks are able to predict spectral features from raw ultrasound data. Furthermore, we also demonstrate that the quality of synthesized speech can be improved by feature selection and other feature engineering methods.

2. Methods

2.1. Data acquisition

One Hungarian female subject (42 years old) with normal speaking abilities was recorded while reading sentences aloud. The tongue movement was also recorded in midsagittal orientation using a "Micro" ultrasound system (Articulate Instruments Ltd.) with a 2-4 MHz / 64 element 20mm radius convex ultrasound transducer at 82 fps. During the recordings, the transducer was fixed using an ultrasound stabilization headset (Articulate Instruments Ltd.). The video of the lips was recorded at 59.94 fps (interlaced) from front with an NTSC microcamera that was attached to the helmet – but the lip video was not used in the current study. The speech signal was recorded with an Audio-Technica - ATR 3350 omnidirectional condenser microphone that was clipped approximately 20cm from the lips. Both the microphone signal and the ultrasound synchronization signals were digitized using an M-Audio – MTRACK PLUS external sound card at 22050 Hz sampling frequency. The ultrasound and the audio signals were synchronized using the frame synchronization output of the equipment with the Articulate Assistant Advanced software (Articulate Instruments Ltd.).

2.2. Preprocessing the speech signal

For the analysis and synthesis of speech, a standard vocoder was used from SPTK (`http://sp-tk.sourceforge.net`). First, the speech signal was lowpass filtered and resampled to 11 050 Hz. F0 was measured with the SWIPE algorithm [17]. Next, 12-order Mel-Generalized Cepstral analysis (MGC) [18] was performed with $\alpha = 0.42$ and $\gamma = -1/3$. MGCs were converted to Line Spectral Pair (LSP) representation as these have better interpolation properties. In order to make the result of the speech analysis to be in synchrony with the ultrasound images, the frame shift was chosen to be 1 / FPS (where FPS is the frame rate of the ultrasound). Together with the gain, the MGC-LSP resulted in a 13-dimension feature vector, which was used in the training experiments.

For the synthesis phase, we assumed that the F0 parameter can not be estimated from the articulatory images, so we used the original F0 extracted from the input. The predictions of the DNN served as the remaining MGC-LSP parameters required by the synthesizer. First, impulse-noise excitation was generated according to the F0 parameter. Afterwards, spectral filtering was applied using the MGC-LSP coefficients and a Mel-Generalized Log Spectral Approximation (MGLSA) filter [19] to reconstruct the speech signal.

2.3. Preprocessing the ultrasound signal

The original raw ultrasound images were 64×946 sized grayscale images. To remove some of the noise from the recordings and also to reduce the size of the dataset, we resized each image so as to have a new size of 64×119 using a bicubic interpolation. This reduction did not significantly affect the visual content of the images, and the DNNs trained on these reduced images achieved almost identical results.

In the simplest configuration, one ultrasound image serves as the input of the neural network, and the corresponding target vector is an estimate of the acoustic parameters of the speech signal at the given time instance. However, it is well known both in automatic speech recognition and in various image processing tasks that extending the feature vector with several neighbouring time frames markedly improves DNN performance. Due to this, we extended the input vector of our DNN to contain several consecutive images. This, however, also multiplies the number of DNN parameters, which means that both the training procedure and the synthesis will require more time, and the risk of overfitting also increases. To avoid these problems we decided to reduce the size of the input images further. On the one hand, we experimented with feature selection by using a novel, correlation-based feature selection method. On the other hand, we also utilized the Eigentounge feature transformation algorithm [20]. These methods significantly reduced the input vector, and this allowed us to train the DNNs on a larger temporal context, while retaining the same number of parameters.

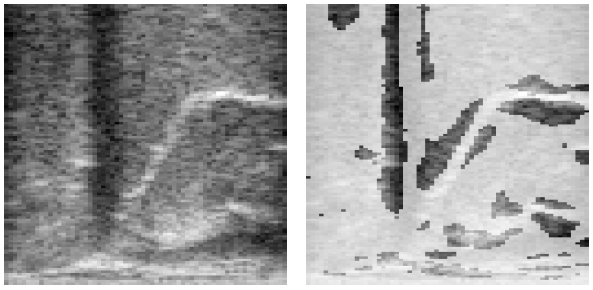

Figure 1: *A raw ultrasound image and the mask which was used in our correlation-based feature selection method (max., 20%).*

2.4. Correlation-based feature selection

After investigating the audio and video signals, we observed that the speech signal uttered depends only on specific parts of the ultrasound image. This means that large parts of the image are completely irrelevant and so the size of the DNN input vector could be reduced by discarding the corresponding pixels. Since the relevance of a pixel can roughly be captured by correlation, in the first applied feature selection method we utilized the correlation of the target scores and the pixels of the images. For this, we calculated the absolute value of the correlation for each pixel and each target score in the training set. Since we had 13 training targets, this process resulted in 13 correlation scores. As we planned to train one neural network to jointly predict all 13 parameters of the vocoder, we had to select only one, common feature subset instead of selecting a subset for each parameter. Because of this, we had to create one overall importance score for each pixel from the 13 correlation scores. We tested two solutions for this score aggregation by taking the mean and the maximum of the absolute correlation values. Then, in the last step of feature selection, we only retained $5, 10, \ldots, 25\%$ of the pixels with the largest importance scores. Fig. 1 shows an input ultrasound image and the parts which were retained after this feature selection approach.

2.5. Eigentounge feature extraction

The so-called 'Eigentongue' feature extraction method [20] is almost identical to the popular Eigenface method of Turk and Pentland [21], the only difference being the type of the input images. The eigentongue technique seeks to find a finite set of orthogonal images (called *eigentounges*), which constitute, up to a certain accuracy, a subspace for the representation of all likely tongue configurations. The standard way to extract the eigentongues is to apply PCA on the training data and define the eigenvectors obtained as the eigentounges. The eigentongue components extracted are supposed to encode the maximum

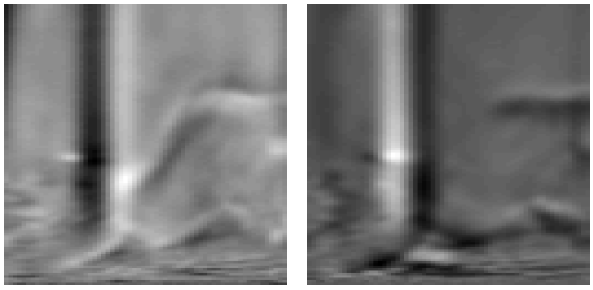

Figure 2: *The first two extracted Eigentongues.*

amount of relevant information present in the images, which mainly consist of tongue position, but other structures such as the hyoid bone and muscles are also present. Fig. 2 shows some examples of the eigentounge images. To reduce the dimension of the data it is standard practice to select a subset of the eigentounges corresponding to the highest eigenvalues, and only use these vectors to transform the data to a lower dimension.

2.6. DNN setup

We trained DNNs with 5 hidden layers, with each hidden layer consisting of 1000 rectified neurons [22]. The input layer consisted of 7 616 neurons. The output layer was a linear one, with one neuron for each MGC-LSP feature and one for the gain. We also trained separate DNNs for each of the output features, similar to that in [10] to see which approach is better. The feature space reduction methods allowed us to present five consecutive images as input to the DNN, while the size of input layer did not change, meaning that the number of parameters in the network remained the same as before.

3. Discussion

Our system is quite similar to that of Jaumard-Hakoun et al. [10] in the sense that we are using a 2D ultrasound input to predict spectral coefficients of a vocoder, and the prediction is based on deep neural networks. Their paper was the inspiration for using 11 050 Hz to sample the speech signal and 12-order MGC-LSP coefficients.

We expect that the feature selection strategies presented in sections 2.4 and 2.5 are superior compared to those that just use raw ultrasound data. We also hypothesize that using multiple consecutive images as input can increase the accuracy of the regression. We tested these hypotheses in the following experiments.

4. Experimental results

4.1. Objective measurements

To objectively measure the performance of the DNNs we chose two widely used metrics, namely the Normalized Mean Square Error (NMSE) and R^2. As the training target values varied at different scales, we had to use the normalized version of MSE, otherwise the output having the largest range (in our case, the gain) would have dominated the MSE error. Furthermore, as the NMSE just measures the distance between predictions and the expected outputs, we also used the (mean) R^2 metric, which measures how well the predictions fit the expected curves.

Table 1 lists the NMSE and (mean) R^2 values got for the different approaches tested (note that our goal is to minimize NMSE and maximize R^2). It is clear that training a shared DNN model for all 13 parameters was beneficial, as our predictions were more accurate this way using both metrics. When retaining only 20% of the attributes, we can see an increase in the error. The two correlation-based feature selection techniques behave very similarly, while the Eigentongue approach proved to be slightly better. Yet, when we utilize the preceding and subsequent two images as well, the Eigentongue feature set becomes the worst of the three tested methods. Overall, we see that using our correlation-based feature selection approach and ranking the pixels by the maximum of their correlation values, we can slightly outperform the baseline approach that used the whole actual ultrasound image. Of course, our final aim is to produce the parameters which lead to the most naturally sounding speech, hence we used our objective metrics (NMSE and R^2) primarily to reduce the list of models which would be evaluated by subjective listening tests.

Table 1: *NMSE and mean R^2 scores on the development set*

Type	NMSE	Mean R^2
DNN (separate models)	0.409	0.597
DNN (joint model)	0.384	0.619
DNN (feature selection (max.), 20%)	0.441	0.562
DNN (feature selection (avg.), 20%)	0.442	0.561
DNN (Eigentongue, 20%)	0.432	0.577
DNN (feature sel. (max.), 5 images)	0.380	0.625
DNN (feature sel. (avg.), 5 images)	0.388	0.615
DNN (Eigentongue, 5 images)	0.402	0.608

4.2. Subjective listening tests

In order to determine which proposed system is closer to natural speech, we conducted two online MUSHRA (MUlti-Stimulus test with Hidden Reference and Anchor) listening tests [23]. The advantage of MUSHRA is that it allows the evaluation of multiple samples in a single trial without breaking the task into many pairwise comparisons. Our aim was to compare the natural sentences with the synthesized sentences with the baseline, the proposed approaches and a benchmark system. In the benchmark system, the natural F0 and gain parameters were used, whereas the other 12 spectral features were constant (extracted from a /swa/ vowel). In the tests, the listeners had to rate the naturalness of each stimulus in a randomized order relative to the reference (which was the natural sentence), from 0 (highly unnatural) to 100 (highly natural).

For the first test, 10 sentences were randomly chosen which were not included in the training of the DNNs. We chose six types of the DNN configurations listed in Table 1. Together with the natural, benchmark, and vocoded samples, 90 utterances were included in the test (1 speaker · 9 types · 10 sentences). For the second test, 15 sentences were chosen, but only with the following approaches: 'DNN (joint model)' , 'DNN (Eigentongue, 5 images)', and 'DNN (feature selection (maximum), 5 images)'. Altogether, 90 utterances were included (1 speaker · 6 types · 15 sentences). The samples of the listening test are available online: `http://smartlab.tmit.bme.hu/interspeech2017_ssi`.

4.2.1. Results of listening test #1

The first test was performed by the authors of the paper, in order to pre-select the potentially best approaches for the main listening test. The means and standard deviations of the naturalness results are shown in Table 2. The 'DNN (joint model)'

Table 2: *Naturalness scores of listening test #1*

Type	Mean	Std.dev.
Natural	98.00	3.56
Benchmark	0.47	1.40
Vocoded	67.92	15.90
DNN (separate models)	31.80	13.86
DNN (joint model)	33.60	15.14
DNN (feature selection (max.), 20%)	29.40	11.70
DNN (Eigentongue, 20%)	29.50	12.98
DNN (feature sel. (max.), 5 images)	35.30	16.59
DNN (Eigentongue, 5 images)	33.02	13.32

outperformed the 'DNN (separate models)', indicating that it is more useful to train one neural network with all the data, rather than separate neural nets for the 13 MGC-LSP spectral features. Independent of the feature selection method (correlation-based and EigenTongue), using several consecutive ultrasound frames always resulted in more natural synthesized sentences. These trends are similar to those found in the objective measurements.

According to the results, we selected 3 out of the 6 feature representation approaches for a second listening test. The goal of the second test was to rank the three best DNN-based approaches of listening test #1 with a larger number of subjects.

4.2.2. Results of listening test #2

Altogether 23 listeners participated in the main test (20 females, 3 males). All of them were native speakers of Hungarian, and 21 of them were university students. The subjects were between 19–32 years (mean: 21 years). The authors did not participate in the main test. On average the whole test took 15 minutes to complete. The MUSHRA scores of the listening test are presented in Fig. 3 for the natural, benchmark, vocoded reference, and the three DNN models. In general, 'Natural' sentences should yield 100% in MUSHRA type tests. However, we were using 11 050 Hz sampling frequency even for the natural sentences, meaning that the lack of high-frequency components was audible for the listeners. The 'Benchmark' type ranked the lowest, as these utterances were unintelligible. The 'Vocoded' references achieved naturalness scores of 56%, as a standard vocoder with impulse-noise excitation was used, and this is clearly different from natural speech. The utterance types in which the spectral features were predicted based on ultrasound were ranked around 30%, indicating that they are roughly half-way between the vocoded references and the benchmark.

The ratings of the listeners were compared by Mann-Whitney-Wilcoxon ranksum tests as well, with a 95% confidence level, indicating that the DNN-based synthetic signals significantly differ from the natural, benchmark, and vocoded, but do not significantly differ from each other. However, Fig. 3 shows improvements in the 'DNN (Eigentongue, 5 images)' representation compared to 'DNN (joint model)'. Also, 'DNN (feature selection (max.), 5 images)' was slightly preferred over 'DNN (Eigentongue, 5 images)'. This means that using feature selection methods in combination with consecutive ultrasound images was helpful in synthesizing more natural sentences. According to the preference of the subjects of the listening test, the 'DNN (feature selection (maximum), 5 images)' approach ranked the best.

The subjects of the listening test were allowed to leave textual feedback. Some of them noted that they perceived various artefacts on the samples, e.g. low-frequency distortions, robotic voice, and a 'smudged' feeling. Others wrote that most of the

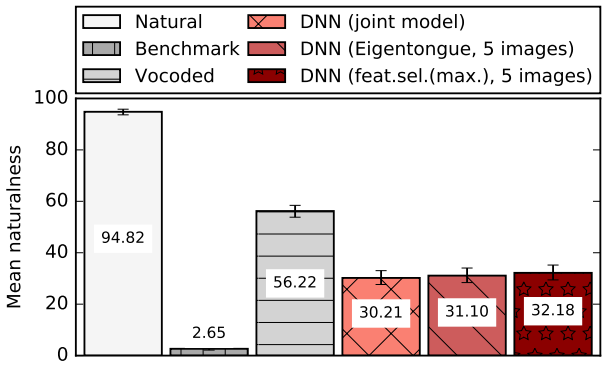

Figure 3: *Results of the listening test #2 concerning naturalness. The errorbars show the 95% confidence intervals.*

samples were intelligible, especially the shorter utterances.

5. Conclusions

Here, we described our initial experiments in articulatory-to-acoustic mapping. Raw 2D ultrasound of the tongue was used as input to a fully-connected feed-forward rectified deep neural network. The DNNs had to predict Mel-Generalized Cepstrum features in Line Spectral Pair representation, which was used to synthesize speech with a vocoder. We investigated several types and combinations of feature representations, including 1) a baseline approach where a joint model was used to predict all 13 MGC-LSP features, 2) separate models for predicting the 13 spectral features (as suggested by [10]), 3) two variants of a correlation-based feature selection, 4) Eigentongue feature selection to reduce the size of ultrasound images [20], and 5) the feature selection methods combined with using several consecutive ultrasound frames.

We found that our hypothesis was supported by the evaluations: the representation that used five neighboring image frames in combination with a correlation-based feature selection method was preferred both in terms of the Normalized Mean Squared Error and by the subjects taking part in the listening experiments. Although we did not test intelligibility directly, the results of the current study are encouraging, as we were able to convert raw tongue-ultrasound data (using the original F0 estimate) to intelligible speech using deep neural networks. Mapping from articulatory data to F0 is a challenge, but there has been some research on voiced/unvoiced prediction [7].

In the future, we plan to investigate other neural network types (e.g. AutoEncoders and convolutional neural networks). Adding multimodal articulatory data (e.g. video of the lips) is also expected to increase the naturalness of synthesized samples. We also intend to use more advanced vocoders which make the synthesized speech samples sound less robotic (e.g. [24]). Finally, we intend to record silent speech (as suggested by [9]) and study the differences compared to regular speech, in terms of articulatory features. Our results may prove to be useful for creating Silent Speech Interface applications.

6. Acknowledgements

Tamás Gábor Csapó was partly supported by the VUK (AAL-2014-1-183) and the EUREKA / DANSPLAT projects. Tamás Grósz was supported by the ÚNKP-16-3 New National Excellence Programme of the Ministry of Human Capacities. We would like to thank the listeners who participated in the test.

7. References

[1] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270–287, 2010.

[2] F. Bocquelet, T. Hueber, L. Girin, P. Badin, and B. Yvert, "Robust Articulatory Speech Synthesis using Deep Neural Networks for BCI Applications," in *Proc. Interspeech*, 2014, pp. 2288–2292.

[3] F. Bocquelet, T. Hueber, L. Girin, C. Savariaux, and B. Yvert, "Real-Time Control of an Articulatory-Based Speech Synthesizer for Brain Computer Interfaces," *PLOS Computational Biology*, vol. 12, no. 11, p. e1005119, nov 2016.

[4] J. Wang, A. Samal, and J. Green, "Preliminary Test of a Real-Time, Interactive Silent Speech Interface Based on Electromagnetic Articulograph," in *Proceedings of the 5th Workshop on Speech and Language Processing for Assistive Technologies*, 2014, pp. 38–45.

[5] B. Denby and M. Stone, "Speech synthesis from real time ultrasound images of the tongue," in *Proc. ICASSP*. Montreal, Quebec, Canada: IEEE, 2004, pp. 685–688.

[6] T. Hueber, E.-L. Benaroya, G. Chollet, G. Dreyfus, and M. Stone, "Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips," *Speech Communication*, vol. 52, no. 4, pp. 288–300, 2010.

[7] T. Hueber, E.-l. Benaroya, B. Denby, and G. Chollet, "Statistical Mapping Between Articulatory and Acoustic Data for an Ultrasound-Based Silent Speech Interface," in *Proc. Interspeech*, Florence, Italy, 2011, pp. 593–596.

[8] T. Hueber, G. Bailly, and B. Denby, "Continuous Articulatory-to-Acoustic Mapping using Phone-based Trajectory HMM for a Silent Speech Interface," in *Proc. Interspeech*, Portland, OR, USA, 2012, pp. 723–726.

[9] T. Hueber and G. Bailly, "Statistical conversion of silent articulation into audible speech using full-covariance HMM," *Computer Speech and Language*, vol. 36, pp. 274–293, 2016.

[10] A. Jaumard-Hakoun, K. Xu, C. Leboullenger, P. Roussel-Ragot, and B. Denby, "An Articulatory-Based Singing Voice Synthesis Using Tongue and Lips Imaging," in *Proc. Interspeech*, 2016, pp. 1467–1471.

[11] J. A. Gonzalez, R. K. Moore, J. M. Gilbert, L. A. Cheah, S. Ell, and J. Bai, "A silent speech system based on permanent magnet articulography and direct synthesis," *Computer Speech and Language*, vol. 39, pp. 67–87, 2016.

[12] J. Freitas, A. J. Ferreira, M. A. T. Figueiredo, A. J. S. Teixeira, and M. S. Dias, "Enhancing multimodal silent speech interfaces with feature selection," in *Proc. Interspeech*, Singapore, Singapore, 2014, pp. 1169–1173.

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," jun 2015. [Online]. Available: http://arxiv.org/abs/1506.01497

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[15] S. Xie and Z. Tu, "Holistically-Nested Edge Detection," in *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, dec 2015, pp. 1395–1403.

[16] L. Diener, M. Janke, and T. Schultz, "Direct conversion from facial myoelectric signals to speech using Deep Neural Networks," in *2015 International Joint Conference on Neural Networks (IJCNN)*, jul 2015, pp. 1–7.

[17] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 124, no. 3, pp. 1638–1652, sep 2008.

[18] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis - a unified approach to speech spectral estimation," in *Proc. ICSLP*, Yokohama, Japan, 1994, pp. 1043–1046.

[19] S. Imai, K. Sumita, and C. Furuichi, "Mel Log Spectrum Approximation (MLSA) filter for speech synthesis," *Electronics and Communications in Japan (Part I: Communications)*, vol. 66, no. 2, pp. 10–18, 1983.

[20] T. Hueber, G. Aversano, G. Chollet, B. Denby, G. Dreyfus, Y. Oussar, P. Roussel, and M. Stone, "Eigentongue feature extraction for an ultrasound-based silent speech interface," in *Proc. ICASSP*, Honolulu, HI, USA, 2007, pp. 1245–1248.

[21] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, Jan. 1991.

[22] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier networks," in *Proceedings of AISTATS*, 2011, pp. 315–323.

[23] "ITU-R Recommendation BS.1534: Method for the subjective assessment of intermediate audio quality," 2001.

[24] T. G. Csapó, G. Németh, M. Cernak, and P. N. Garner, "Modeling Unvoiced Sounds In Statistical Parametric Speech Synthesis with a Continuous Vocoder," in *Proc. EUSIPCO*, Budapest, Hungary, 2016, pp. 1338–1342.