Applying DNN Adaptation to Reduce the Session Dependency of Ultrasound Tongue Imaging-based Silent Speech Interfaces

Gábor Gosztolya[1,2], Tamás Grósz[2,3], László Tóth[2], Alexandra Markó[4,6], Tamás Gábor Csapó[5,6]

[1] MTA-SZTE Research Group on Artificial Intelligence of the Hungarian Academy of Sciences and University of Szeged, Tisza Lajos krt. 103, H-6720 Szeged, Hungary, ggabor@inf.u-szeged.hu

[2] Institute of Informatics, University of Szeged, Árpád tér 2, H-6720 Szeged, Hungary, groszt@inf.u-szeged.hu, tothl@inf.u-szeged.hu

[3] Department of Signal Processing and Acoustics, Aalto University, Otakaari 3, FI-02150 Espoo, Finland, tamas.grosz@aalto.fi

[4] Department of Applied Linguistics and Phonetics, Eötvös Loránd University, Múzeum krt. 4/A, H-1088 Budapest, Hungary, marko.alexandra@btk.elte.hu

[5] Department of Telecommuniations and Media Informatics, Budapest University of Technology and Economics, Magyar tudósok körútja 2, H-1117 Budapest, Hungary, csapot@tmit.bme.hu

[6] MTA-ELTE Lingual Articulation Research Group, Múzeum krt. 4/A, H-1088 Budapest, Hungary

Abstract: Silent Speech Interfaces (SSI) perform articulatory-to-acoustic mapping to convert articulatory movement into synthesized speech. Its main goal is to aid the speech handicapped, or to be used as a part of a communication system operating in silence-required environments or in those with high background noise. Although many previous studies addressed the speaker-dependency of SSI models, session-dependency is also an important issue due to the possible misalignment of the recording equipment. In particular, there are currently no solutions available, in the case of tongue ultrasound recordings. In this study, we investigate the degree of session-dependency of standard feed-forward DNN-based models for ultrasound-based SSI systems. Besides examining the amount of training data required for speech synthesis parameter estimation, we also show that DNN adaptation can be useful for handling session dependency. Our results indicate that by using adaptation, less training data and training time are needed to achieve the same speech quality over training a new DNN from scratch. Our experiments also suggest that the sub-optimal cross-session behavior is caused by the misalignment of the recording equipment, as adapting just the lower, feature extractor layers of the neural network proved to be sufficient, in achieving a comparative level of performance.

Keywords: Silent speech interfaces; articulatory-to-acoustic mapping; session dependency; Deep Neural Networks; DNN adaptation

1 Introduction

Over the past few years, there has been significant interest in articulatory-to-acoustic conversion research, which is often referred to as "Silent Speech Interfaces" (SSI) [5]. The idea is to record the soundless articulatory movement, and automatically generate speech from the movement information, while the subject is not producing any sound. Such an SSI system might be very useful for the speaking impaired (e.g. after a laryngectomy), and for scenarios where regular speech is not feasible, but information should be transmitted from the speaker (e.g. extremely noisy environments and/or military situations). For this automatic conversion task, typically electromagnetic articulography (EMA, [3, 19, 20]), ultrasound tongue imaging (UTI, [4, 14, 18, 28]), permanent magnetic articulography (PMA, [10]), surface Electromyography (sEMG, [6, 16, 22]), lip video [1, 7] and multimodal approaches are used [5]. Current SSI systems mostly apply the "direct synthesis" principle, where speech is generated without an intermediate step, directly from the articulatory data. This approach has the advantage compared to Silent Speech Recognition (SSR) that there is a significantly smaller delay between articulation and speech generation, and there are fewer error possibilities than in the case of the SSR + TTS (Text-to-Speech) approach, where first the articulatory movement is translated to a phoneme or word sequence, and then it is used to generate the speech signal via standard TTS techniques.

As Deep Neural Networks (DNNs) have become dominant in more and more areas of speech technology, such as speech recognition [9, 13, 26], speech synthesis [2, 21] and language modeling [23, 24, 29], it is natural that recent studies have attempted to solve the ultrasound-to-speech conversion problem by employing deep learning, regardless of whether sEMG [17], ultrasound video [18] or PMA [10] is used as an input. Our team used DNNs to predict the spectral parameter values [4] and F0 [12] of a vocoder using UTI as articulatory input; in a later study we extended our method to include multi-task training [28].

A recent study [25] has summarized the state-of-the-art results in silent speech interfaces. Although there are lots of research findings on generating intelligible speech using EMA, UTI, PMA, sEMG, lip video and multimodal data, all the studies were conducted on relatively small databases and typically with just one or a small number of speakers [25]; while all of the articulatory tracking devices are obviously highly sensitive to the speaker. Another source of variance comes from the possible misalignment of the recording equipment. For example, for tongue-ultrasound recordings, the probe fixing headset has to be mounted onto the

speaker before use, and in practice it is impossible to mount it onto exactly the same spot as before. This inevitably causes the recorded ultrasound video to become misaligned compared to a video recorded in a previous session. Therefore, such recordings are not directly comparable. In the following, by "session" it is meant that the probe fixing headset is dismounted and mounted again onto the speaker.

There have already been some studies that use multi-speaker and/or multi-session articulatory data for SSI and SSR. Kim et al. investigated speaker-independent SSR using EMA and compared Procrustes matching-based articulatory normalization, feature-space maximum likelihood linear regression and i-vector experimentally on 12 healthy and two laryngectomized English speakers [19, 20]. The best results were achieved with a combination of the normalization approaches. For EMG-based recognition, a variety of signal normalization and model adaptation methods were investigated, as experiments revealed an across-sessions deviation of up to 5 mm [22]. From the nine different normalization and adaptation procedures, sharing training data across sessions and Variance Normalization and Feature Space Adaptation proved to be the most useful [22]. Janke et al. also studied session-independent sEMG: 16 sessions of a speaker were analyzed and the results indicated that the MCD (Mel-Cepstral Distortion) in the case of cross-session conversion is only slightly worse compared to the 500 sentence session-dependent result from the same speaker, confirming that sEMG is robust even with minor changes in the electrode placement or other influence [16]. Wand et al. utilized domain-adversarial DNN training for session-independent EMG-based speech recognition [30].

Unfortunately, for ultrasound-based SSI, there are no methods currently available for the alignment / adaptation / normalization of articulatory data recorded in different sessions or with different speakers. All the above-mentioned studies [16, 20, 22, 30] used EMA or sEMG for tracking articulatory movements; and although e.g. Maier-Hein et al., state that even slight changes in electrode positions affect the myoelectric signal [22], Janke et al. found that their sEMG-based framework employing GMMs virtually behaves session-insensitively without any form of adaptation [16]. In the ultrasound-based SSI systems, however, where slight changes in probe positioning can cause shifts and rotations in the image used as input (for an example, see Fig. 1), might not turn out to be ideal.

To this end, in this study we focus on the session dependency of the ultrasound-based direct speech synthesis process. Although we also consider speaker dependency to be a significant issue, here we will just concentrate on session dependency. Notice that using recordings from different speakers inevitably means using data from different sessions as well, but without the option of identifying and analyzing the negative effect of using different speaker data (e.g. F0, speaking style, oral cavity structure) and the effect of slight changes in the position of the recording equipment. To separate the effect from the two possible

error sources, in this study we shall focus on the session dependency of the ultrasound-based direct speech synthesis process. We will demonstrate experimentally, that a simple, yet, efficient, standard feed-forward DNN-based system displays clear signs of session dependency, to such an extent, that the synthesized utterances are practically unintelligible. Furthermore, we propose a simple session adaptation method, and show that it is more efficient than training a neural network from scratch using the adaptation data. We shall also examine the amount of training data required for successful DNN model adaptation. Of course, the applicability of the proposed approach for session adaptation (i.e. DNN model adaptation) is not necessarily limited to the UTI case, but it may be of interest for a broader audience as well.

2 Methods

2.1 Data Acquisition

A Hungarian female subject with normal speaking abilities was recorded while reading sentences aloud. Tongue movement was recorded in midsagittal orientation using the "Micro" ultrasound system of Articulate Instruments Ltd. at 82 fps. The speech signal was recorded with a Beyerdynamic TG H56c tan omnidirectional condenser microphone. The ultrasound data and the audio signals were synchronized using the tools provided by Articulate Instruments Ltd. (For more details, see our previous studies [4, 12, 28].) In our current experiments, the scanline data of the ultrasound recording was used. The original ultrasound images of 64×842 pixels were resized to 64×106 by bicubic interpolation, leading to 6784 features per time frame. To create the speech synthesis targets, the speech recordings (resampled to 22050 Hz) were analyzed using an MGLSA vocoder [15] at a frame shift of 1 / (82 fps), which resulted in F0, energy and 24-order spectral (MGC-LSP) features [27]. The vocoder spectral parameters (excluding F0) served as the DNN training targets.

Our data was collected in four sessions. The headset and the ultrasound probe were fitted each time using the same procedure; however, it cannot be guaranteed that the orientation of the probe remained "exactly" the same, across each session. In the first session we recorded 200 individual sentences (about 15 minutes in total), while in sessions two, three and four, we recorded 50 different sentences (less than 4 minutes each). In addition, in each session, the subject read the 9-sentence long Hungarian version of the short tale `The North Wind and the Sun'. We used the independent sentences for training purposes, while the utterances of "The North Wind and the Sun" were used as test sets. For more information about the four sessions, see Table 1.

Table 1
Key properties of the recordings used in our experiments; duration is expressed in terms of min:sec

Recording session	Individual Sentences (Train)		North Wind & Sun (Test)	
	Count	Duration	Count	Duration
Session #1	200	14:48	9	0:50
Session #2	50	3:44	9	0:49
Session #3	50	3:53	9	0:47
Session #4	50	3:41	9	0:48

Fig. 1 shows sample images taken from the four sessions with similar tongue positions. Although all four images are similar, there are visible positioning differences among them, which might lead a DNN trained on the first session to perform sub-optimally on the other sessions. We will demonstrate this sub-optimality experimentally in Section 3, and we will describe how we applied DNN adaptation to handle this issue in Section 4.

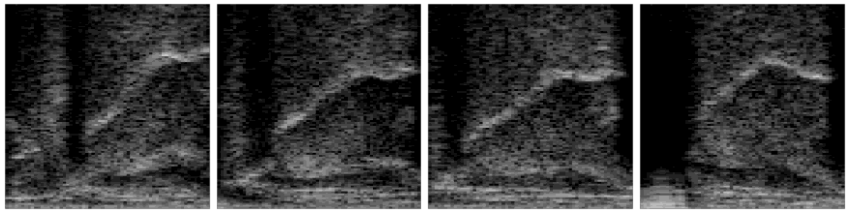

Figure 1
Sample ultrasound tongue images from the four sessions used. Note that all the images belong to the same speaker

2.2 DNN Parameters

We trained feed-forward, fully-connected DNNs with 5 hidden layers, each hidden layer consisting of 1000 ReLU neurons. The input neurons corresponded to the image pixels, while the output layer contained one linear neuron for each MGC-LSP feature and one for the gain (25 output parameters overall). To assist prediction, we presented a time slice of the ultrasound video (five consecutive frames) as input to the DNN, since in our previous studies [4, 12, 28] we found this technique to be beneficial. The input images consisted of 6784 pixels, meaning that the network had a total of 33920 input neurons.

2.3 Evaluation

As estimating the parameters of the synthesizer is a simple regression problem, the most suitable evaluation metric is the Pearson correlation; or, in our case, as we have 25 speech synthesis parameters to predict, we will take the mean of the 25 correlation values. In our earlier studies, [28], we also used this evaluation metric. In our last experiments, however, in order to determine which proposed system is closer to natural speech, we also conducted an online MUSHRA (MUlti-Stimulus test with Hidden Reference and Anchor) listening test [31]. 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 of the baseline, the proposed approaches (various session adaptation variants) and a benchmark system (the latter being cross-session synthesis without adaptation). In the test, 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 (very unnatural) to 100 (very natural). We chose sentences from 4-layer adaptation and full training, and tested two adaptation data sizes (20 and 50 sentences). Altogether 96 utterances were included in the test (12 sentences x 8 variants). In the MUSHRA evaluation, each configuration was evaluated by 12 native Hungarian speakers with normal hearing abilities.

3 Results with Single-Session DNN Training

3.1 The Effect of the Amount of Training Data

In our first experiments, we examined how the amount of training data affects the performance of the DNN model. For this, we trained our neural network on the recordings of the same session that we used for testing. We used N = 1, 5, 10, 20 and 50 sentences for training, and evaluated our models on the 9 sentences of `The North Wind and the Sun' from the same session. Since for Session #1 we had more utterances in the training data, there we also experimented with N = 100, 150 and 200.

The mean correlation values obtained this way have been plotted in Fig. 2. Clearly, the correlation scores vary to a great extent among the different sessions, though at this point we did not perform any cross-session experiments: DNN training and evaluation were performed by using recordings taken from the same session. We can also see that, by increasing the number of training sentences, the correlation values increased, as expected. Also note that, when we used more than N = 100 sentences (roughly 7 minutes of recordings), there is a slight

improvement only, although we had only one session with enough training data to confirm this.

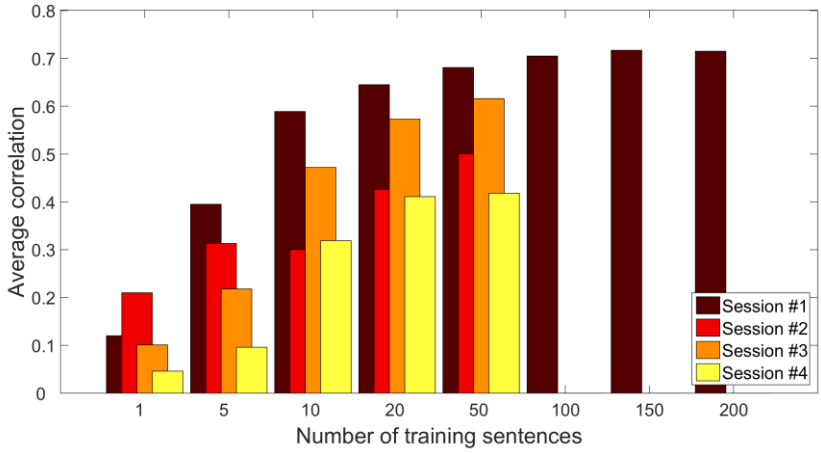

Figure 2
Average correlation values obtained for the four sessions as a function of the number of sentences used for training

Examining our sample images (see Fig. 1), it is hard to see any difference among the sessions which might explain the significant difference in the average correlation scores observed in Fig. 2. Perhaps the only exception is the large dark area in the posterior region (on the left hand side of the image) in session #4, where not only the hyoid bone blocked the ultrasound waves (as it did on the other images), but also there was probably insufficient amount of gel between the transducer and the skin, limiting the visibility in that particular direction. However, for session #2 we got similarly low correlation scores, while the ultrasound video contained no such artifact. Since we fitted the recording equipment following the same procedure for each session, these results alone, in our opinion, indicate that UTI-based SSI systems are session-sensitive even without using data taken from multiple speakers.

3.2 Cross-Session Results

In our next experiment, we sought to examine how the misalignment of input images affects the performance of the neural network. To this end, we trained our DNN on all the 200 sentences of the first session, and evaluated it on the utterances of `The North Wind and the Sun' recorded in the remaining three sessions.

Table 2
Average correlation scores obtained for the recordings of `The North Wind and the Sun' depending on the DNN training data

Training Data		Average correlation for sessions			
Session	Size	#2	#3	#4	Avg.
Session #1	200	0.075	0.100	0.143	0.106
Same as test	50	0.501	0.616	0.418	0.512

The first row of Table 2 shows the average correlation values obtained this way. We can see that the DNN predictions are practically worthless, as the average Pearson's correlation values fall between 0.075 and 0.143. (We also confirmed the low quality of these predictions by listening tests, and found the synthesized 'utterances' unintelligible.) In contrast (see the second row), using just 50 sentences for DNN training, but from the same session, we get average correlation scores in the range 0.418-0.616. This huge difference, in our opinion, also demonstrates that ultrasound-based DNN SSI approaches are quite sensitive to misalignments of the ultrasound images, even if these come from the same speaker, and this issue has to be handled if we intend to develop SSI systems for practical use.

4 DNN Adaptation

In the previous section we showed experimentally that DNN models trained on the recordings of one session cannot be utilized to predict speech synthesis parameters in another session, even when both sessions were recorded with the same speaker. Next, we will show that the issue of session-dependency can be handled effectively via the adaptation of the DNN model trained on data from a different session. In practice, adaptation means that we train the DNN further, using recordings taken from the actual session. For the general scheme of the proposed approach, see Fig. 3. Of course, to ease the use of our SSI equipment, this adaptation material has to be as short as possible, hence we simultaneously aim for high-quality spectral parameter estimation while keeping the amount of adaptation data to a minimum. To this end, we performed DNN adaptation experiments using N = 1, 5, 10, 20 and 50 sentences from each session; we used once again the 9 sentences of `The North Wind and the Sun' of the actual session for evaluation purposes.

It is well known (e.g. [8, 11]) that the lower layers of a deep neural network are responsible for low-level feature extraction, while the higher layers perform more abstract and more task-dependent functions. As in our case session dependency appears as a change in the input image, while the task remains the same (i.e. to predict the spectral representation of the speech of the same speaker), it seems

reasonable to expect that it might be sufficient to train just the lower layers of the network instead of adapting all the weights. This way, we might achieve the same level of accuracy with faster training, or obtain better estimates [11]. Since in our experiments we employed DNNs with five hidden layers, we have six choices of which layers to adapt (i.e. only the weights between the input layer and the first hidden layer, adapt the weights among the input layer and the first two hidden layers, etc.). To test this, we also experimented with adapting just the first two and first four layers of the network. Furthermore, as a comparison, we also tried training a DNN from scratch using N = 1, 5, …, 50 sentences on data taken from the same session as our baselines.

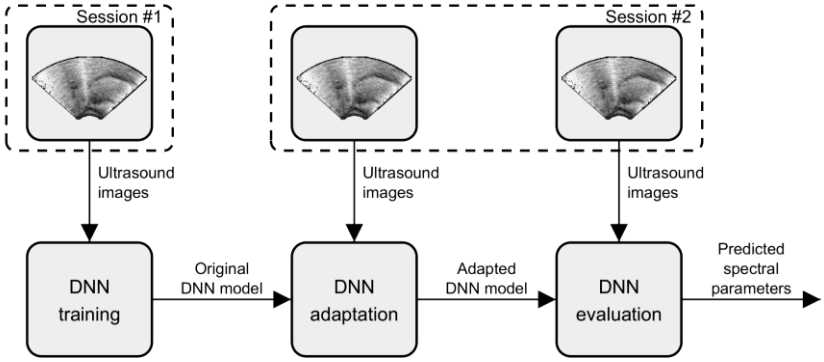

Figure 3
The general workflow of the proposed DNN SSI model adaptation procedure

4.1 DNN Adaptation Results

4.1.1 Correlation Values

Fig. 4 shows the average correlation value**s** measured, as a function of the number of training sentences. The scores are averaged out for the three sessions (i.e. Session #2, #3 and #4); the error bars represent minimal and maximal values. We can see that, in general, if we used more sentences either for DNN training or for adaptation, the accuracy of the predictions improved. It is also quite apparent that when we have only a few sentences taken from the current session, adaptation leads to more accurate predictions than training a randomly initialized DNN. For the N = 20 and N = 50 cases, however, full DNN training resulted only in slightly lower correlation values than adaptation did. Still, even when we have a higher number of sentences, we can state that by using DNN adaptation, fewer sentences are needed to achieve the same performance as with full DNN training. For example, adapting 3 layers with 10 utterances (about 20-25 seconds) of training data from the given session leads to roughly the same averaged correlation score that can be achieved by using 20 sentences and full DNN training.

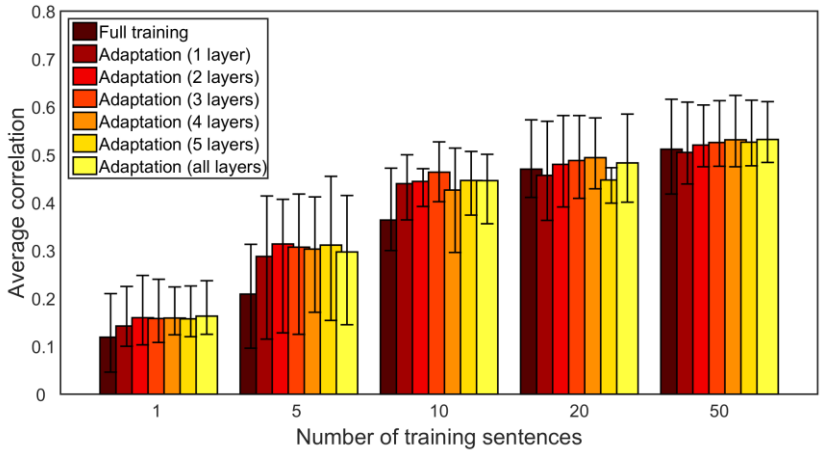

Figure 4
Average correlation scores via full DNN training and DNN adaptation as a function of the number of sentences used

Regarding the number of layers adapted, there are only slight differences in DNN performance. Although adapting only one layer (i.e. the weights between the input and the first hidden layer) led to the lowest correlation value in each case, the remaining five variations proved to be quite similar, and usually adapting the first four layers (for $N = 10$, three layers), proved to be optimal.

Inspecting the minimal and maximal correlation scores for each configuration, these values usually behaved just like the mean correlation scores did: adapting only one layer resulted in a suboptimal performance, but when we adapted at least two layers, there were no large differences. However, it is quite apparent that for the case $N = 50$ and adapting at least two layers, the minimal correlation value greatly exceeded that of full training, while the maximal scores appeared to be roughly the same. For an SSI system used in everyday practice, where we have no guarantee of the precision of the current equipment positioning, the minimal performance of the (adapted or newly trained) DNN model might be just as important as the average one; and in this respect, DNN adaptation performed much better than full DNN training did.

Table 3 lists the notable correlation scores for all three sessions and their average. These numeric values confirm our previous findings; namely, the average performance of full DNN training always falls closer to the best correlation score of DNN adaptation using fewer sentences than using the same amount of training data. Furthermore, for the case $N = 50$, full DNN training led to a correlation value of 0.418 as the worst score, while for adaptation it is never lower than 0.475.

Table 3
Average correlation scores obtained for `The North Wind and the Sun' depending on the amount of DNN adaptation data

No. of Train Sentences	Adapted Layers	Average correlation for sessions			
		#2	#3	#4	Avg.
10	Full training	0.300	0.472	0.319	0.364
	Input to 2nd	0.471	0.470	0.392	0.444
	Input to 3rd	0.462	0.527	0.402	0.464
	All layers	0.481	0.501	0.356	0.446
20	Full training	0.426	0.573	0.411	0.470
	Input to 2nd	0.467	0.582	0.391	0.480
	Input to 3rd	0.476	0.577	0.429	0.494
	All layers	0.463	0.585	0.401	0.483
50	Full training	0.501	0.616	0.418	0.512
	Input to 2nd	0.475	0.604	0.482	0.520
	Input to 3rd	0.475	0.624	0.495	0.531
	All layers	0.484	0.611	0.501	0.532

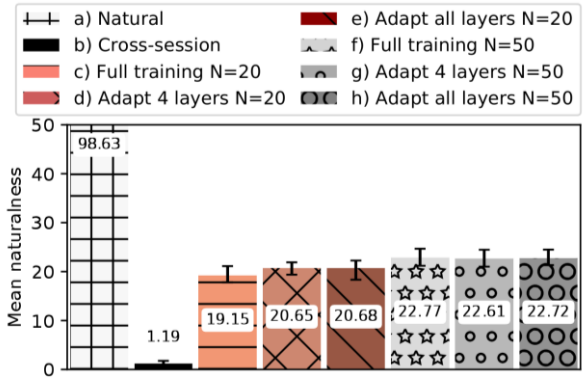

Figure 5
Mean naturalness scores of the MUSHRA listening test; error bars show the 95% confidence intervals

4.1.2 MUSHRA Listening Tests

Fig. 5 shows the results obtained from the MUSHRA listening tests. (The samples used in the test can be found at http://smartlab.tmit.bme.hu/actapol2019_ssi_session.) The naturalness of the synthesized utterances turned out to be somewhat low in each case, probably due to the small size of the training data (i.e. 20 or 50 sentences overall, equivalent to about 90 seconds and less than 4 minutes of duration, respectively). Still, the effect of the number of sentences used for training or adaptation is clearly visible: using no adaptation led to unintelligible speech (a mean naturalness score of only 1.19), while using 20

sentences resulted in naturalness scores between 19.15 and 20.68, which increased to 22.61-22.77 for the case $N = 50$. The listening tests also reinforced our previous findings that for $N = 20$, DNN adaptation is a better approach, while for $N = 50$ there is no observable difference among the output of the full DNN training and the DNN adaptation techniques. According to the Mann-Whitney-Wilcoxon ranksum test with a 95% confidence level, differences between variants c) to h) (i.e. the tested models with $N = 20$ and $N = 50$) were not statistically significant.

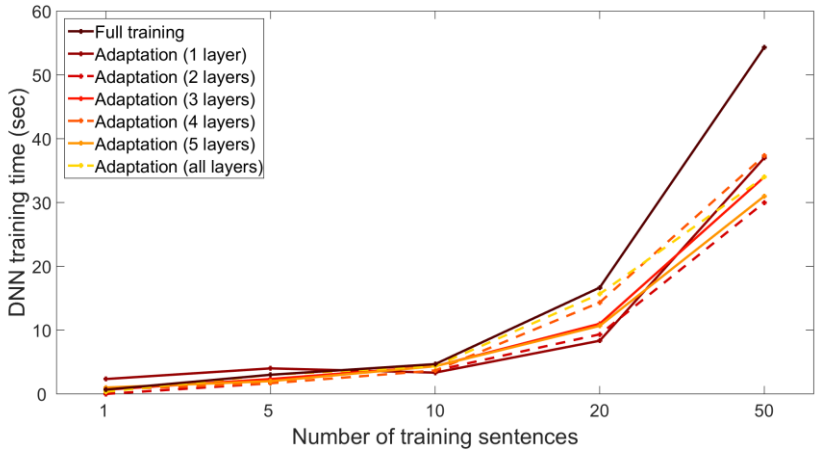

Figure 6
Average wall clock training times as a function of the number of sentences used for training

4.1.3 DNN Training Times

Fig. 6 shows the (wall clock) DNN training and DNN adaptation times expressed in seconds (averaged out for the three sessions), measured on an Intel i7 4.2 GHz PC with 32 GB RAM and an NVidia Titan X video card. From these values, it is clear that the DNN adaptation time is primarily affected by the size of the adaptation data: for $N = 10$, the average values fell between 3 and 5 seconds, which increased to 8-15 seconds for $N = 20$ and to 30-37 seconds for $N = 50$. In contrast, full DNN training took 17 seconds for $N = 20$ and 54 seconds for $N = 50$. From these values, however, we cannot confirm that adapting fewer layers leads to lower execution times; in our experience, DNN adaptation time is primarily affected by the size of the adaptation data. Full DNN training led to by far the highest training time in the $N = 50$ case, while for $N = 20$ its training time is much higher than those of most adaptation configurations. This indicates that DNN adaptation has a further advantage: it allows quicker convergence than training a DNN with random initial weights. Specifically, for the case $N = 50$, DNN adaptation required about two-thirds the time compared to DNN training from scratch did; and adapting a DNN with 20 sentences needed far less training time (17-29%) to achieve the same performance as full DNN training did with $N = 50$.

Overall, from these results, DNN adaptation with 20 sentences seems to be the best approach, since it requires significantly less training material than full DNN training in the case $N = 50$, and it was also much quicker to train. Furthermore, it led to a higher minimal correlation value, while the average correlation and MUSHRA naturalness scores appeared to be quite similar, and the difference was not statistically significant.

Conclusions

In this study, we focused on the session dependency of the ultrasound-based direct speech synthesis process, during articulatory-to-acoustic mapping. Similarly to studies using sEMG [16, 22] and EMA [19, 20], we investigated how the reattachment of the articulatory equipment affects the final output. For the first time in the scientific community, we used ultrasound tongue imaging for this purpose, building on our earlier single-session studies [4, 12, 28]. We expected that reattaching the probe would greatly diminish the accuracy of a previously trained system.

We found that our hypothesis was supported by the following results:

1) The synthesized speech was unintelligible if the network was trained on one session and evaluated on another session as-is (without the adaptation of the network weights)
2) We found large differences even among the performance of DNN models used within the same session, depending on the actual session
3) To create a DNN model for the actual session, DNN adaptation performed better than full DNN training did during UTI-to-spectral feature conversion

Furthermore, DNN adaptation had the advantage of allowing quicker convergence than random DNN weight initialization did.

The findings of our experiments are an important step within the articulatory-to-acoustic research area, as the simple-yet-effective adaptation method proposed herein, should contribute to the development of practical and efficient Silent Speech Interfaces. For example, a DNN adaptation with 20 sentences takes roughly 15 seconds on a current computer (such as the Intel i7 4.2 GHz PC used in our experiments), after which, speech can be synthesized directly from ultrasound-based articulatory data. However, the current study was conducted on regular speech and it is a future task to experiment with real silent (mouthed) speech. In the future we also plan to investigate the speaker-dependency of the ultrasound tongue imaging.

Acknowledgement

The authors were partially funded by the Ministry of Human Capacities, Hungary (grant no. 20391-3/2018/FEKUSTRAT), the National Research, Development and Innovation Office of Hungary (grants no. FK 124584 and PD 127915) and by the

MTA “Lendület” grant. Gábor Gosztolya and László Tóth were supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences and by the ÚNKP-19-4 New National Excellence Program of the Ministry of Human Capacities.

References

[1] H. Akbari, H. Arora, L. Cao, and N. Mesgarani, “LIP2AUDSPEC : Speech reconstruction from silent lip movements video,” in Proceedings of ICASSP, Calgary, Canada, 2018, pp. 2516-2520

[2] M. S. Al-Radhi, T. G. Csapó, and G. Németh, “Deep Recurrent Neural Networks in speech synthesis using a continuous vocoder,” in Proceedings of SPECOM, Hatfield, Hertfordshire, UK, 2017, pp. 282-291

[3] B. Cao, M. Kim, J. R. Wang, G. Van Santen, T. Mau, and J. Wang, “Articulation-to-speech synthesis using articulatory flesh point sensors' orientation information,” in Proceedings of Interspeech, Hyderabad, India, 2018, pp. 3152-3156

[4] T. G. Csapó, T. Grósz, G. Gosztolya, L. Tóth, and A. Markó, “DNN-based ultrasound-to-speech conversion for a silent speech interface,” in Proceedings of Interspeech, Stockholm, Sweden, 2017, pp. 3672-3676

[5] 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

[6] L. Diener, and T. Schultz, “Investigating objective intelligibility in real-time EMG-to-Speech Conversion,” in Proceedings of Interspeech, Hyderabad, India, 2018, pp. 3162-3166

[7] A. Ephrat and S. Peleg, “Vid2speech: Speech reconstruction from silent video,” in Proceedings of ICASSP, New Orleans, LA, USA, 2017, pp. 5095-5099

[8] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, “Joint training of frontend and back-end Deep Neural Networks for robust speech recognition,” in Proceedings of ICASSP, Brisbane, Australia, 2015, pp. 4375-4379

[9] X. Glorot, A. Bordes, and Y. Bengio, “Deep sparse rectifier networks,” in Proceedings of AISTATS, Fort Lauderdale, FL, USA, 2011, pp. 315-323

[10] J. A. Gonzalez, L. A. Cheah, A. M. Gomez, P. D. Green, J. M. Gilbert, S. R. Ell, R. K. Moore, and E. Holdsworth, “Direct speech reconstruction from articulatory sensor data by machine learning,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 25, No. 12, pp. 2362-2374, 2017

[11] G. Gosztolya and T. Grósz, “Domain adaptation of Deep Neural Networks for Automatic Speech Recognition via wireless sensors,” Journal of Electrical Engineering, Vol. 67, No. 2, pp. 124-130, 2016

[12] T. Grósz, G. Gosztolya, L. Tóth, T. G. Csapó, and A. Markó, "F0 estimation for DNN-based ultrasound silent speech interfaces," in Proceedings of ICASSP, Calgary, Alberta, Canada, 2018, pp. 291-295

[13] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," IEEE Signal Processing Magazine, Vol. 29, No. 6, pp. 82-97, 2012

[14] T. Hueber, E.-l. Benaroya, B. Denby, and G. Chollet, "Statistical mapping between articulatory and acoustic data for an ultrasoundbased silent speech interface," in Proceedings of Interspeech, Florence, Italy, 2011, pp. 593-596

[15] 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

[16] M. Janke, M. Wand, K. Nakamura, and T. Schultz, "Further investigations on EMG-to-speech conversion," in Proceedings of ICASSP, Kyoto, Japan, 2012, pp. 365-368

[17] M. Janke and L. Diener, "EMG-to-speech: Direct generation of speech from facial electromyographic signals," IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 25, No. 12, pp. 2375-2385, 2017

[18] 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 Proceedings of Interspeech, San Francisco, CA, USA, 2016, pp. 1467-1471

[19] M. Kim, B. Cao, T. Mau, and J. Wang, "Multiview representation learning via deep CCA for silent speech recognition," in Proceedings of Interspeech, Stockholm, Sweden, 2017, pp. 2769-2773

[20] M. Kim, B. Cao, T. Mau, and J. Wang, "Speaker-Independent Silent Speech Recognition From Flesh-Point Articulatory Movements Using an LSTMNeural Network," IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 25, No. 12, pp. 2323-2336, 2017

[21] H.-Z. Ling, S.-Y. Kang, H. Zen, A. Senior, M. Schuster, X.-J. Qian, H. Meng, and L. Deng, "Deep learning for acoustic modeling in parametric speech generation," IEEE Signal Processing Magazine, Vol. 32, No. 3, pp. 35-52, 2015

[22] L. Maier-Hein, F. Metze, T. Schultz, and A. Waibel, "Session independent non-audible speech recognition using surface electromyography," in Proceedings of ASRU, San Juan, Puerto Rico, 2005, pp. 331-336

[23] G. Melis, C. Dyer, and P. Blumsom, “On the state of the art of evaluation in neural language models,” Proceedings of ICLR, Vancouver, BC, Canada, 2018

[24] T. Mikolov, M. Karafiát, L. Burget, J. Černocký, and Sanjeev Khudanpur, “Recurrent neural network based language model,” in Proceedings of Interspeech, Makuhari, Japan, 2010, pp. 1045-1048

[25] T. Schultz, M. Wand, T. Hueber, D. J. Krusienski, C. Herff, and J. S. Brumberg, “Biosignal-based spoken communication: A survey,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 25, No. 12, pp. 2257-2271, 2017

[26] F. Seide, G. Li, X. Chen, and D. Yu, “Feature engineering in context-dependent deep neural networks for conversational speech transcription,” in Proceedings of ASRU, Big Island, HI, USA, 2011, pp. 24-29

[27] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, “Melgeneralized cepstral analysis – a unified approach to speech spectral estimation,” in Proceedings of ICSLP, Yokohama, Japan, 1994, pp. 1043-1046

[28] L. Tóth, G. Gosztolya, T. Grósz, A. Markó, and T. G. Csapó, “Multi-task learning of phonetic labels and speech synthesis parameters for ultrasound-based silent speech interfaces,” in Proceedings of Interspeech, Hyderabad, India, 2018, pp. 3172-3176

[29] Z. Tüske, R. Schlüter, and H. Ney, "Investigation on LSTM Recurrent N-gram Language Models for Speech Recognition," in Proceedings of Interspeech, Hyderabad, India, pp. 3358-3362, 2018

[30] M. Wand, T. Schultz, and J. Schmidhuber, “Domain-adversarial training for session independent EMG-based speechrecognition,” in Proceedings of Interspeech, Hyderabad, India, 2018, pp. 3167-3171

[31] “ITU-R recommendation BS.1534: Method for the subjective assessment of intermediate audio quality,” 2001