Multi-Task Learning of Speech Recognition and Speech Synthesis Parameters for Ultrasound-based Silent Speech Interfaces

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

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

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

Abstract

Silent Speech Interface systems apply two different strategies to solve the articulatory-to-acoustic conversion task. The recognition-and-synthesis approach applies speech recognition techniques to map the articulatory data to a textual transcript, which is then converted to speech by a conventional text-to-speech system. The direct synthesis approach seeks to convert the articulatory information directly to speech synthesis (vocoder) parameters. In both cases, deep neural networks are an evident and popular choice to learn the mapping task. Recognizing that the learning of speech recognition and speech synthesis targets (acoustic model states vs. vocoder parameters) are two closely related tasks over the same ultrasound tongue image input, here we experiment with the multi-task training of deep neural networks, which seeks to solve the two tasks simultaneously. Our results show that the parallel learning of the two types of targets is indeed beneficial for both tasks. Moreover, we obtained further improvements by using multi-task training as a weight initialization step before task-specific training. Overall, we report a relative error rate reduction of about 7% in both the speech recognition and the speech synthesis tasks.

Index Terms: Silent speech interface, silent speech recognition, articulatory-to-acoustic mapping, multi-task, DNN

1. Introduction

Over the last decade, there has been an increased interest in the analysis, recognition and synthesis of silent speech, which is a form of spoken communication where an acoustic signal is not produced, that is, the subject is just silently articulating without producing any sound. Systems which can perform the automatic articulatory-to-acoustic mapping are often referred to as 'Silent Speech Interfaces' (SSI) [1]. Such an SSI can be applied to help the communication of the speaking impaired (e.g. patients after laryngectomy), and in situations where the speech signal itself cannot be recorded (e.g. extremely noisy environments or certain military applications). As the articulatory recording equipment, typically ultrasound tongue imaging (UTI) [2, 3, 4, 5, 6, 7, 8, 9, 10], electromagnetic articulography (EMA) [11, 12, 13, 14], permanent magnetic articulography (PMA) [15, 16] and surface electromyography (sEMG) [17, 18, 19, 20, 21] are used. Of course, the multimodal combination of these methods is also possible [22], and the above methods may also be combined with a simple video recording of the lip movements [4].

There are two distinct ways of SSI solutions, namely 'direct synthesis' and 'recognition-and-synthesis' [23]. In the first case, the speech signal is generated without an intermediate step, directly from the articulatory data, typically using vocoders [2, 5, 6, 7, 8, 13, 16, 19, 20]. In the second case, silent speech recognition (SSR) is applied on the biosignal which extracts the content spoken by the person (i.e., the result is text); this step is then followed by text-to-speech (TTS) synthesis [4, 3, 10, 11, 12, 14, 15, 21]. The drawback of the SSR+TTS approach might be that the errors made by the SSR component inevitably appear as errors in the final TTS output [23], and also that it causes a significant end-to-end delay. Furthermore, any information related to speech prosody is totally lost, while several studies have showed that certain prosodic components may be estimated reasonably well from the articulatory recordings (e.g., energy [7] and pitch [8]). Therefore, state-of-the-art SSI systems mostly prefer the 'direct synthesis' principle.

As deep neural networks (DNNs) became dominant in more and more areas of speech technology, such as speech recognition [24], speech synthesis [25] and language modeling [26], it is natural that the recent studies have attempted to solve the articulatory-to-acoustic conversion problem using deep learning. Diener and his colleagues studied sEMG speech synthesis in combination with a deep neural network [19, 20]. In another study a multimodal Deep AutoEncoder was used to synthesize sung vowels based on ultrasound recordings and a video of the lips [6]. Gonzalez and his colleagues compared GMM, DNN and RNN [16] for PMA-based direct synthesis, while we used DNNs to predict the spectral parameters [7] and F0 [8] of a vocoder using UTI as articulatory input. Liu et al. compared DNN, RNN and LSTM neural networks for the prediction of the V/U flag and voicing [27], while Zhao et al. found that LSTMs perform better than DNNs for articulatory-to-F0 prediction [28].

The multi-task training of DNNs was proposed to improve the generalization ability of the DNN by forcing it to learn two (or more) related tasks at the same time [29]. An application to speech technology was presented by Seltzer and Droppo [30]. They found that besides training the network to recognize the actual phone, the recognition accuracy can be improved by also training the network to identify the phone context as a secondary task. Bell et al. applied the multi-task scheme for the joint training of context-independent and context-dependent phone labels, and they obtained relative improvements of 3-10% in the word error rate compared to conventional train-

ing [31]. Multi-task training was also successfully used for the recognition of reverberant speech [32] and in speech synthesis [33].

As the 'recognition-and-synthesis' and the 'direct synthesis' approaches represent the problem by different, yet closely related machine learning tasks over the same input, multi-task training seems directly applicable here. In this study we attempt to jointly estimate the speech recognition targets and the speech synthesis targets using a DNN adjusted to multi-task training. While the input for both tasks consists of the same ultrasound recording, the training targets for the speech recognition task are the HMM states of the acoustic model, and in the case of the speech synthesis task they correspond to the vocoder parameters. The joint DNN model contains several shared layers, but it has two dedicated output layers for the two tasks. The shared layers are forced to focus on both tasks simultaneously, and as the tasks are related but different, it helps the network gain extra knowledge and thus attain a better local optimum.

In our experiments, ultrasound recordings of about a half an hour from a female speaker served as the input for the multi-task DNN. We created several variants of the multi-task network by varying the number of shared and task-specific layers. As the last step, after multi-task training we converted the multi-task DNN into two separate networks and continued their training by task-specific training steps. Overall, we obtained relative error rate reductions of about 7% for both the HMM state probability estimation and the vocoder parameter estimation tasks, which justifies the viability of our approach.

2. Multi-task modeling

Fig. 1 shows the architecture of the DNN used in our experiments. The input of the network is a (series of) ultrasound images, for which the corresponding feature extraction process will be presented in detail later on. The network consisted of 5 hidden layers for both tasks, but only the upper layers were task-specific, while the lower ones were shared between the two tasks. Naturally, the network contained two output layers, one dedicated to each task. In the case of the speech recognition task the training targets were the acoustic state labels of the HMM/DNN recognizer. As regards the speech synthesis branch of the network, the training targets were the vocoder parameters of the given speech frame. Notice that the former is a classification task, while the latter is a regression task; so they require a different type of output layer and a different cost function. Hence, joint learning by simply concatenating the target vectors and feeding them into a standard DNN would not be straightforward.

The motivation behind multi-task learning is the assumption that forcing the network to create a shared representation for the two tasks might be beneficial for both of them. We can reasonably expect this only when the two tasks are closely related. In our case the training targets are the phone state labels and the vocoder (spectral) parameters. While these are clearly related, the connection between them is far from trivial (mapping the spectrum into phones is basically what speech recognition is about). Hence, besides forcing the network to create a shared representation, we must also leave room for it to learn the actual task. This was our motivation for varying the number of shared and task-specific layers (while keeping the overall depth fixed at 5 layers). Fig. 1 shows the configuration with 3 shared and 2 task-specific layers. We note that, due to layer sharing, the multi-task configuration always had fewer parameters than the baseline consisting of two separate networks.

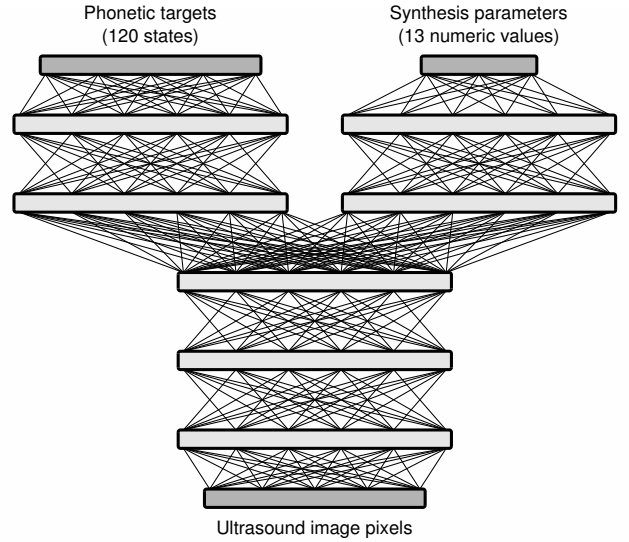

Figure 1: *The structure of the multi-task DNN. The configuration shown is for the case of 3 shared and 2 task-specific layers.*

3. Experimental set-up

3.1. Data acquisition

A Hungarian female subject with normal speaking abilities was recorded while reading sentences aloud (altogether 438 sentences). The 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 an Audio-Technica - ATR 3350 omnidirectional condenser microphone. The ultrasound data and the audio signals were synchronized using the tools provided by Articulate Instruments Ltd. More details about the recording set-up can be found in [7]. In the experiments below, the scanline data of the ultrasound was used. The original raw ultrasound images of 64×946 pixels were resized to 64×119 in the same way as in [7], which resulted in 7616 features per time frame. The overall duration of the recordings was about half an hour, which was partitioned into training, development and test sets in a 70-10-20 ratio.

Based on the speech recordings and the corresponding transcripts, the phonetic labels and boundaries were obtained by using a Hungarian speech recognizer [34] in forced alignment mode. The aligned acoustic model states served as the training targets of our DNN in the speech recognition experiments. We worked with tri-state monophone models, as the amount of training data was quite limited.

To create the speech synthesis targets, the speech recordings (resampled at 11 050 Hz) were analyzed using an MGLSA vocoder [35] at a frame shift of 1 / (82 fps), which resulted in F0, energy and 12-order spectral (MGC-LSP) features [36]. The vocoder spectral parameters (excluding F0) served as the training targets of the DNN in our speech synthesis experiments.

3.2. Feature extraction

We experimented with three input representation methods. In the simplest case only one frame of the ultrasound video was used as input to the neural network. Then, to improve the performance, we also extended the input vector of the DNN to contain 5 consecutive data vectors (this model will be labeled as "5 neigh." in the figures). However, this simple solution increased

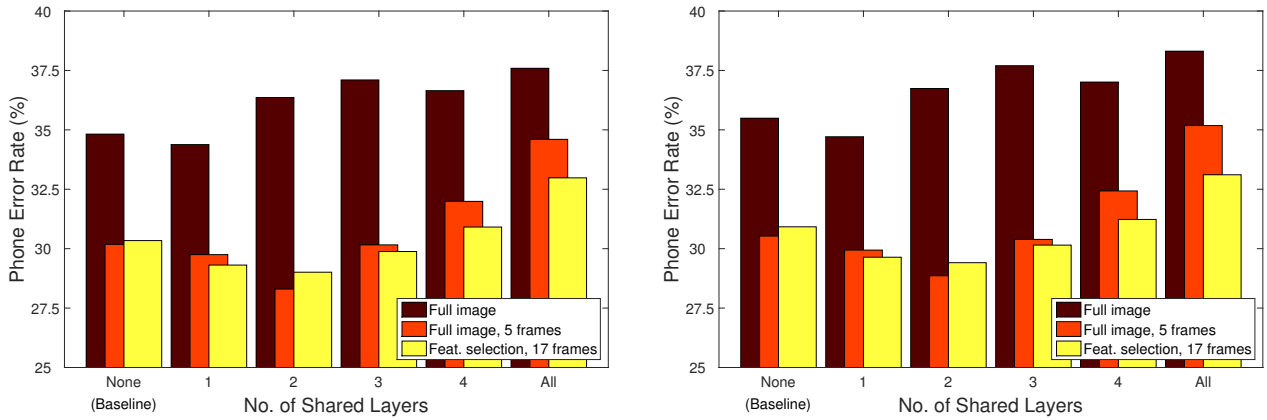

Figure 2: *Phone error rates attained with the speech recognition branch of the multi-task DNN as a function of the number of shared hidden layers, for the various feature sets, for the development set (left) and the test set (right).*

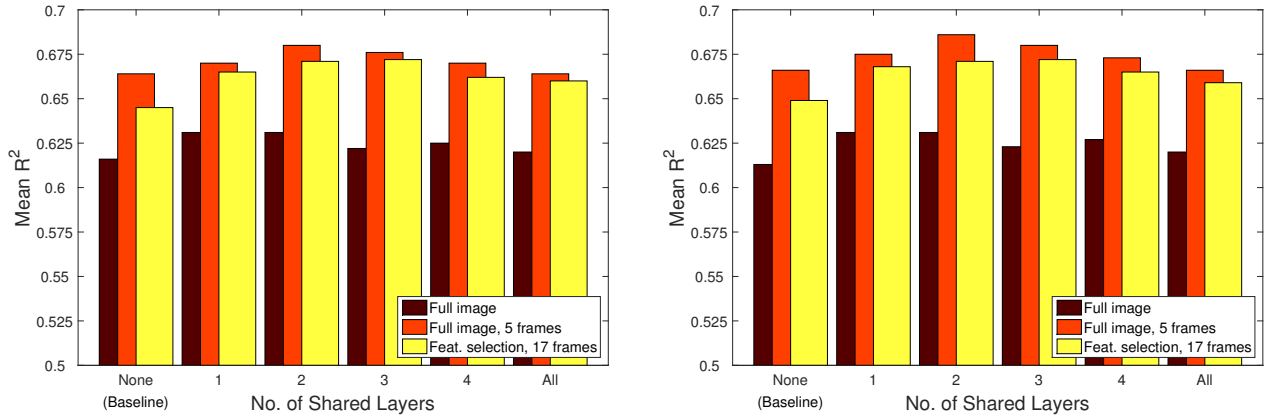

Figure 3: *Mean R^2 values attained with the speech synthesis branch of the multi-task DNN as a function of the number of shared hidden layers, for the various feature sets, for the development set (left) and the test set (right).*

the size of the input layer to 38080, thus slowing down the training process and increasing the risk of overfitting. Hence, we applied a correlation-based feature selection method to reduce each image to 20% of its original size [7]. This allowed us to use a larger left-right context of 8-8 neighboring frames, resulting in an input size of 25891 features. This solution will be called "feat. selection, 17 neigh." in the tables.

3.3. DNN training

All our DNNs were simple fully-connected feed-forward networks, with each hidden layer consisting of 1000 ReLU neurons. The output layer dedicated for the speech recognition task was a softmax layer containing 120 output units (corresponding to the states in our HMM/DNN recognizer). We applied the frame-level cross-entropy (CE) cost function to calculate the error of this branch of the network. The output layer for the speech synthesis task contained 13 linear neurons (the vocoder spectral parameters), and the cost function applied for this regression task was the mean-squared error (MSE). We trained the network using backpropagation, where the learning rate decay was controlled by the error rate of the phone state classifier output layer on the validation set. We also tried to use the other output layer for this purpose, but we obtained no significant difference in the overall number of the training epochs or when the halving of the learning rate commenced.

3.4. Evaluation

To evaluate the speech recognition branch of the DNN, the estimated phone states were turned into a phone sequence by a simple Viterbi decoder. This decoding step applied a phone bigram language model, with its weight tuned on the development set. In Section 4, we will report phone error rate results obtained this way. While we could have achieved better results by applying a more sophisticated word-level language model, our main goal here was just to demonstrate the feasibility of the multi-task training approach.

As regards the speech synthesis branch, we might have simply reported the MSE values attained. However, as the mean of the Pearson correlation (R^2) is a more common error metric in the field of SSI,, and we also reported R^2 values in our earlier study [7], here we will again present Pearson correlation values. While a thorough evaluation would also require subjective listening tests, here we skipped this tedious step, as our main goal was just to justify the applicability of multi-task training.

4. Results and discussion

Fig. 2 shows the phone error rates obtained by using the speech recognition output of the multi-task network. As the first attempt, we shared all the hidden layers among the tasks, so that only the output layer was task-specific, as this is the most

Feature set	Phone Error Rate (dev/test)		
	Baseline	Multi-task	Adapted Multi-task
feat. selection + 17 neigh.	30.3% / 30.9%	29.0% / 29.4%	28.3% / 28.7%
Full image, 5 neigh.	30.2% / 30.5%	28.3% / 28.9%	28.2% / 28.4%

Table 1: *Phone error rates of the baseline, the 2-shared-layer multi-task, and the task-adapted DNNs for the two best feature sets.*

Feature set	Mean R^2 (dev/test)		
	Baseline	Multi-task	Adapted Multi-task
feat. selection + 17 neigh.	0.645 / 0.649	0.671 / 0.671	0.672 / 0.679
Full image, 5 neigh.	0.664 / 0.666	0.680 / 0.686	0.685 / 0.689

Table 2: *Mean R^2 values of the baseline, the 2-shared-layer multi-task, and the task-adapted DNNs for the two best feature sets.*

widespread solution in the literature [30, 31]. Unfortunately, we got worse results than those with the baseline model (which was a simple task-specific DNN) for all feature sets (cf. the rightmost column in the figures). We conjectured that the two tasks are simply too different, so just one task-specific layer is not enough to learn them. Hence, we repeated the experiment, but gradually decreased the number of shared layers, while increasing the number of task-specific ones. As Fig. 2 shows, the multi-task system outperformed the baseline when sharing 1-3 layers, and the optimum was achieved with two shared layers. A similar behavior was observed with both feature sets that involved several neighboring frames, while the simplest feature set that consisted of only one frame of data produced far worse and randomly fluctuating results in all cases. This reinforced our previous finding that involving neighboring frames in the feature set is vital for a good performance [7, 8]. While the feature set that consisted of a 'compressed' version of 17 neighboring frames slightly outperformed the set that contained 5 subsequent full images in most cases, the latter set was superior just in the case of the baseline and the best multi-task configurations.

Turning our attention to the test set, we see that the scores follow the same pattern as that for the development set (apart from the simplest feature set that gave worse and inconsistent results). Consistent with the development set, the best results were attained with 2 shared and 3 task-specific layers. For this case, the numeric results are listed in the middle column of Table 1 for the sake of comparison. The best system, namely the one that used 5 neighboring full images as input achieved a relative error rate reduction of 5% over the baseline.

We repeated the same evaluation for the other branch of the multi-task DNN, which estimated the speech synthesis parameters. The R^2 values attained are shown in Fig. 3, both for the development and the test sets. Apart from the fact that this is a maximization and not a minimization task, the trends of the scores are very similar to those for the recognition task. Once again, the largest improvements were attained with 2 shared and 3 task-specific layers. Here, the feature set that contained 5 frames of full images outperformed the one with 17 reduced images in each case. This is different from what we saw for the phone recognition task, and one may suppose that estimating the actual articulatory positions is a more subtle task than just identifying the phone label, and hence it requires the full images. However, justifying this would require more experiments.

Again, apart from the simplest and significantly worse feature set, the improvements on the development set are consistently present on the test set. The scores for the case of 2 shared layers are listed in the middle column of Table 2. For the best system the relative improvement over the baseline was 6%.

4.1. Task-specific training

The fact that sharing all the layers between the two tasks resulted in worse results made us think that the joint learning of the two tasks was more difficult than expected. This gave us the idea of applying the multi-task training only in the first phase of training. After multi-task training, we converted the multi-task network into two separate networks, and continued their training in a task-specific manner. This process can be interpreted as initializing the weights of the two networks with a joint, multi-task training phase. Another interpretation is that with the additional training steps we adapt the multi-task network to the given task. Based on this, we will simply refer to these models as the 'adapted' models.

We applied this task-specific training step only to the two best models that contained two shared layers. The results are shown in the rightmost column of Table 1 for the phone recognition task, and in the rightmost column of Table 2 for the speech synthesis task. The task-specific adaptation of the multi-task DNN yielded a further relative improvement of 1-2% for both tasks. Overall, compared to the baseline, the best models attained a relative error rate reduction of 7% in the case of both the recognition task and the synthesis task.

5. Conclusions

The articulatory-to-acoustic mapping problem of SSI systems can be approached either as a speech recognition and synthesis or as a direct speech synthesis task. While DNNs have already been applied to both tasks, here we realized that the input is the same biosignal in both cases, so we proposed the application of the multi-task training scheme. We experimented with the multi-task training of a DNN, where the parallel training targets were the states of a HMM/DNN speech recognizer and the spectral parameters of a vocoder. We attained relative error rate reductions of about 7% for both tasks, compared to training two separate, task-specific models. These results justify our initial assumption that the joint learning is beneficial for finding a good approximation for both mapping problems.

6. Acknowledgements

The authors were partially funded by the National Research, Development and Innovation Office of Hungary (FK 124584) and by the MTA Lendület program. Tamás Grósz was supported by the UNKP-17-3 New National Excellence Program of the Ministry of Human Capacities. László Tóth was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

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

[3] B. Denby, J. Cai, T. Hueber, P. Roussel, G. Dreyfus, L. Crevier-Buchman, C. Pillot-Loiseau, G. Chollet, S. Manitsaris, and M. Stone, "Towards a Practical Silent Speech Interface Based on Vocal Tract Imaging," in *Proc. ISSP*, 2011, pp. 89–94.

[4] 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.

[5] 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*, 2011, pp. 593–596.

[6] 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.

[7] T. G. Csapó, T. Grósz, G. Gosztolya, L. Tóth, and A. Markó, "DNN-Based Ultrasound-to-Speech Conversion for a Silent Speech Interface," in *Proc. Interspeech*, 2017, pp. 3672–3676.

[8] T. Grósz, G. Gosztolya, L. Tóth, T. G. Csapó, and A. Markó, "F0 Estimation for DNN-Based Ultrasound Silent Speech Interfaces," in *Proc. ICASSP*, 2018.

[9] K. Xu, P. Roussel, T. G. Csapó, and B. Denby, "Convolutional neural network-based automatic classification of midsagittal tongue gestural targets using B-mode ultrasound images," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. EL531–EL537, 2017.

[10] E. Tatulli and T. Hueber, "Feature extraction using multimodal convolutional neural networks for visual speech recognition," in *Proc. ICASSP*, 2017, pp. 2971–2975.

[11] J. Wang, A. Samal, J. R. Green, and F. Rudzicz, "Sentence Recognition from Articulatory Movements for Silent Speech Interfaces," in *Proc. ICASSP*, 2012, pp. 4985–4988.

[12] 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.

[13] 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.

[14] M. Kim, B. Cao, T. Mau, and J. Wang, "Speaker-Independent Silent Speech Recognition From Flesh-Point Articulatory Movements Using an LSTM Neural Network," *IEEE/ACM Trans. ASLP*, vol. 25, no. 12, pp. 2323–2336, 2017.

[15] M. J. Fagan, S. R. Ell, J. M. Gilbert, E. Sarrazin, and P. M. Chapman, "Development of a (silent) speech recognition system for patients following laryngectomy," *Medical Engineering and Physics*, vol. 30, no. 4, pp. 419–425, 2008.

[16] 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 Trans. ASLP*, vol. 25, no. 12, pp. 2362–2374, 2017.

[17] K. Nakamura, M. Janke, M. Wand, and T. Schultz, "Estimation of fundamental frequency from surface electromyographic data: EMG-to-F0," in *Proc. ICASSP*, 2011, pp. 573–576.

[18] Y. Deng, J. T. Heaton, and G. S. Meltzner, "Towards a practical silent speech recognition system," in *Proc. Interspeech*, 2014, pp. 1164–1168.

[19] L. Diener, M. Janke, and T. Schultz, "Direct conversion from facial myoelectric signals to speech using Deep Neural Networks," in *Proc. IJCNN*, 2015, pp. 1–7.

[20] M. Janke and L. Diener, "EMG-to-Speech: Direct Generation of Speech From Facial Electromyographic Signals," *IEEE/ACM Trans. ASLP*, vol. 25, no. 12, pp. 2375–2385, dec 2017.

[21] G. S. Meltzner, J. T. Heaton, Y. Deng, G. De Luca, S. H. Roy, and J. C. Kline, "Silent speech recognition as an alternative communication device for persons with laryngectomy," *IEEE/ACM Trans. ASLP*, vol. 25, no. 12, pp. 2386–2398, 2017.

[22] 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*, 2014, pp. 1169–1173.

[23] T. Schultz, M. Wand, T. Hueber, D. J. Krusienski, C. Herff, and J. S. Brumberg, "Biosignal-Based Spoken Communication: A Survey," *IEEE/ACM Trans. ASLP*, vol. 25, no. 12, pp. 2257–2271, 2017.

[24] 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, nov 2012.

[25] Z.-H. 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.

[26] E. Arisoy, T. N. Sainath, B. Kingsbury, and B. Ramabhadran, "Deep neural network language models," in *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*, ser. WLM '12, 2012, pp. 20–28.

[27] Z.-C. Liu, Z.-H. Ling, and L.-R. Dai, "Articulatory-to-acoustic conversion with cascaded prediction of spectral and excitation features using neural networks," in *Proc. Interspeech*, 2016, pp. 1502–1506.

[28] C. Zhao, L. Wang, J. Dang, and R. Yu, "Prediction of F0 based on articulatory features using DNN," in *Proc. ISSP*, Tienjin, China, 2017.

[29] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, Jul 1997.

[30] M. Seltzer and J. Droppo, "Multi-task learning in deep neural networks for improved phoneme recognition," in *Proc. ICASSP*, 2013, pp. 6965–6969.

[31] P. Bell and S. Renals, "Regularization of deep neural networks with context-independent multi-task training," in *Proc. ICASSP*, 2015, pp. 4290–4294.

[32] R. Giri, M. Seltzer, J. Droppo, and D. Yu, "Improving speech recognition in reverberation using a room-aware deep neural network and multi-task learning," in *Proc. ICASSP*, 2015, pp. 5014–5018.

[33] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King, "Deep neural networks employing Multi-Task Learning and stacked bottleneck features for speech synthesis," in *Proc. ICASSP*. IEEE, apr 2015, pp. 4460–4464.

[34] P. Mihajlik, Z. Tüske, B. Tarján, B. Németh, and T. Fegyó, "Improved Recognition of Spontaneous Hungarian Speech – Morphological and Acoustic Modeling Techniques for a Less Resourced Task," *IEEE Trans. ASLP*, vol. 18, no. 6, pp. 1588–1600, aug 2010.

[35] 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.

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