F0 ESTIMATION FOR DNN-BASED ULTRASOUND SILENT SPEECH INTERFACES

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

[1]MTA-SZTE Research Group on Artificial Intelligence, Szeged, Hungary
[2]Institute of Informatics, University of Szeged, Hungary
[3]Department of Telecommunications and Media Informatics,
Budapest University of Technology and Economics, Budapest, Hungary
[4]Department of Phonetics, Eötvös Loránd University, Budapest, Hungary
[5]MTA-ELTE Lendület Lingual Articulation Research Group, Budapest, Hungary
{ groszt, ggabor, tothl } @ inf.u-szeged.hu,
csapot@tmit.bme.hu, marko.alexandra@btk.elte.hu

ABSTRACT

State-of-the-art silent speech interface systems apply vocoders to generate the speech signal directly from articulatory data. Most of these approaches concentrate on estimating just the spectral features of the vocoder, and use the original F0, a constant F0 or white noise as excitation. This solution is based on the assumption that the F0 curve is unpredictable from articulatory data that does not contain direct measurements of the vocal fold vibration. Here, we experimented with deep neural networks to perform articulatory-to-acoustic conversion from ultrasound images, with an emphasis on estimating the voicing feature and the F0 curve from the ultrasound input. Contrary to the common belief that F0 is unpredictable, we attained a correlation rate of 0.74 between the original and the predicted F0 curve. What is more, the listening tests revealed that our subjects could not distinguish the sentences synthesized using the DNN-estimated and the original F0 curve, and ranked them as having the same quality.

***Index Terms*—** Silent speech interface, articulatory-to-acoustic mapping, DNN, fundamental frequency

1. INTRODUCTION

During the last decade, there has been a significant interest in articulatory-to-acoustic conversion, which is often referred to as "Silent Speech Interfaces" (SSI) [1]. Here the main idea is to record the soundless articulatory movement (e.g. tongue and/or lips), and automatically generate speech from the movement information, without the subject actually producing any sound. Such SSI systems 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 ultrasound tongue imaging (UTI) [2, 3, 4, 5, 6], electromagnetic articulography (EMA) [7, 8], permanent magnetic articulography (PMA) [9], electromyography (EMG) [10] or multimodal approaches [11] are employed.

State-of-the-art SSI systems use the 'direct synthesis' principle, where the speech signal is generated directly from the articulatory data, using vocoders [3, 4, 5, 8, 9]. Most of these approaches focus on predicting just the spectral features of the vocoder (e.g. Mel-Generalized Cepstrum, MGC). The reason for this is that while there is a direct relation between tongue movement and the spectral content of speech, the F0 value depends on the vocal fold vibration, which has no direct connection with the movement of the tongue and face or the opening of the lips [12]. However, there is some evidence that tongue shapes differ in the case of voiced and unvoiced sounds; for example, the vibration of the vocal folds may slow down during consonant articulation [13]. Along with other factors, these changes correlate with the specific articulatory configuration of the obstruents; that is, the volume of space between the glottis and the obstacle [14]. In spite of these facts, most authors studying SSI systems take the unpredictability of F0 for granted, and use the original F0, a constant F0 or white noise as excitation.

Only a few studies attempted to predict the voicing feature and the F0 curve using articulatory data as input. Nakamura et al. utilized EMG data, and they divided the problem into two steps. First, they used an SVM for voiced/unvoiced (V/U) discrimination, and in the second step they applied a GMM for generating the F0 values. According to their results, EMG-to-F0 estimation achieved a correlation of 0.5, while the V/U decision accuracy was 84% [10]. Hueber et al. experimented with predicting the V/U parameter along with the spectral features of a vocoder, using ultrasound and lip video as input articulatory data. They applied a feed-forward deep neural network (DNN) for the V/U prediction and attained an accuracy score of 82%, which is very similar to the

978-1-5386-4658-8/18/$31.00 ©2018 IEEE

result of Nakamura et al. As the input data contained no direct measurements of vocal fold vibration, they explained this relatively high performance by indirect relationships (e.g. stable vocal tract configurations are likely to correspond to vowels and thus to voiced frames) [3].

Two recent studies experimented with EMA-to-F0 prediction. Liu et al. compared DNN, RNN and LSTM neural networks for the prediction of the V/U flag and voicing. They found that the strategy of cascaded prediction, namely using the predicted spectral features as auxiliary input increases the accuracy of excitation feature prediction [15]. Zhao et al. found that the velocity and acceleration of EMA movements are effective in articulatory-to-F0 prediction, and that LSTMs perform better than DNNs in this task. Although their objective F0 prediction scores were promising, they did not evaluate their system in human listening tests [16].

The above brief summary of the related work tells us that although there has been some research on articulatory-to-F0 prediction, deep learning experiments for estimating the F0 curve from tongue ultrasound images alone are still lacking. In a previous study, we presented our results for DNN-based articulatory-to-acoustic mapping from ultrasound images [6]. There, similar to other authors, we assumed that F0 cannot be estimated from ultrasound data, so we used the F0 curves extracted from the original recordings. However, it is clear that for a fully operable SSI system the problem of F0 estimation also has to be solved. Here, we extended our experiments with the estimation of the voicing feature and the actual F0 values. Just like Nakamura et al., we applied a 2-stage approach where one machine learning model seeks to estimate the voicing feature, while another one seeks to predict the F0 value for voiced frames [10]. However, in contrast with their study, we applied DNNs for both tasks. Furthermore, while they worked with EMG signals, in our case the input articulatory representation is ultrasound, which contains no information directly related to vocal fold vibration. We evaluate our approach on ultrasound recordings collected from one female subject. In the experiments we attained a correlation rate of 0.74 between the original and the predicted F0 curve. What is more, in subjective listening tests we found that our subjects could not distinguish between the sentences synthesized using the DNN-estimated or the original F0 curve, and ranked them as having the same quality.

2. EXPERIMENTAL SET-UP

The speech of one Hungarian female subject (42 years old) with normal speaking abilities was recorded while reading sentences aloud (altogether 438 sentences). The sentences were divided into three distinct sets, 310 were selected for the training set, 41 for the development set and 87 for the test set. 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. In the experiments below, the raw scanline data of the ultrasound was used as input data for the DNNs. The images were reduced to 64×119 pixels (for details see [6]).

2.1. Preprocesing and Synthesis

We applied the SPTK vocoder for the analysis and synthesis of speech (http://sp-tk.sourceforge.net). The speech signal was lowpass filtered and resampled to 11 050 Hz. The F0 curve was extracted with the SWIPE algorithm [17]. We extracted 12 MGC-LSP features along with the gain, which resulted in a 13-dimensional feature vector. This vector served as the training target during DNN training. In the synthesis phase, we replaced all parameters required by the synthetizer by the estimates produced by the DNN. The vocoder generated an impulse-noise excitation according to the F0 parameter, and applied spectral filtering using the MGC-LSP coefficients and a Mel-Generalized Log Spectral Approximation (MGLSA) filter [18] to reconstruct the speech signal.

2.2. DNNs for the Estimation of Fundamental Frequency

Estimating the fundamental frequency from the ultrasound movies of the tongue movement alone is a difficult task. Similar to the study of Nakamura et al [10], our system consists of two major machine learning components, one dedicated to making the voiced/unvoiced decision, while the role of the second was to estimate the actual F0 value for voiced frames. In contrast to Nakamura et al., we used DNNs for both tasks. The first task, V/U decision for each frame has a binary output, hence we treated it as a classification task. While working on the same input images, the second DNN seeks to learn the F0 curve. As the F0 estimate takes continuous values, we viewed this task as a regression problem, and we trained this second network just using the voiced segments from the training data. During the evaluation (synthesis) step, the outputs of the two DNNs are merged. This is achieved by simply taking the output value of the F0 predictor network where the voicing network decided in favour of voicing, and returning a constant value for frames judged to be unvoiced.

In the past few years Convolutional Neural Networks (CNNs) have become the state-of-the-art architecture in image processing. Yet, in this study we used a simple fully connected network structure for two reasons. First, CNNs have the biggest advantages for images that are rich in local details that build up the picture in a hierarchical manner. Here, however, the input image contains just a few objects (i.e. the tongue, the shadow of the jaw and the hyoid bone), and the resolution is rather low. Second, the pooling step of CNNs loses the precise spatial relationship between these objects, which might worsen the performance of the network in our case. Therefore we opted for a simple fully connected

	Development			Test		
Input	Voiced accuracy	NMSE	Correlation	Voiced accuracy	NMSE	Correlation
Baseline (1 frames)	86.1%	0.562	0.719	85.7%	0.577	0.711
5 frames	88.2%	0.476	0.760	87.2%	0.516	0.742
17 frames + feature selection	87.4%	0.506	0.747	86.9%	0.526	0.736

Table 1. F0 prediction performance of the DNNs for the various feature sets

structure with five hidden layers of 1000 ReLU neurons. We predicted the F0 parameter together with the gain and the 12 LSP parameters, as in pilot tests we got slightly better results this way than training separate networks for each parameter. This DNN contained 14 linear neurons in its output layer. The network trained for the binary U/V decision task had the same structure, but with a binary classification output layer.

2.3. Input Representation

In our previous study we performed several experiments to seek the optimal input representation. Here, we used only the representation methods that proved the best in that work [6]. Both the V/U decision network and the spectral+F0 parameter estimation network used the same input representation. In the case of the baseline DNN (called *Baseline* later), only one frame of the ultrasound movie 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 images (this model will be called *5-frames*). However, this simple solution significantly increased the size of the input vector, thus slowing down the training and synthesizing process. Hence, we applied a correlation-based feature selection method [6] to reduce each image to 20% of its original size. This allowed us to use a larger left-right context of 8-8 frames, while keeping the network size relatively small. This solution will be called *17-frames+fs*. For more details on the feature selection experiments see our previous study [6].

To evaluate the best F0 predicting system (*5-frames*) via subjective listening tests, we synthesized sentences using three different F0 curves. To have a baseline, we synthesized sentences using a constant F0, where the V/U network predicted the voicing of the actual frame. This system is labeled *F0-const* in the figures. To have an upper glass ceiling, we also synthetized sentences using the original F0 curve (*F0-orig* in the figures). Lastly, the system that applied the DNN-predicted F0 curve is called *F0-DNN* in the following.

3. RESULTS AND DISCUSSION

3.1. Objective Evaluation

Similar to our previous study, we used the Normalized Mean Square Error (NMSE) and the Pearson correlation to objectively measure the quality of our DNN-based estimates [6]. Table 1 lists the results we obtained with the various input representation approaches outlined in the previous section. All the networks achieved quite good V/U accuracy scores, and the best F0 estimation model gave a correlation scores of 0.74. Figure 1 shows an example of the F0 prediction produced by our system, compared to the original F0 curve. It illustrates that our DNNs performed reasonably well in predicting the F0 from ultrasound articulatory data alone.

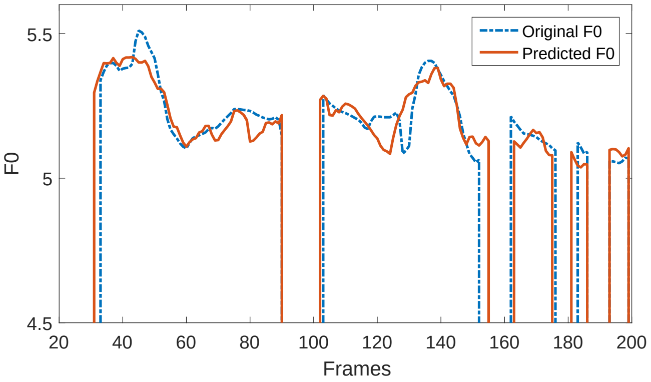

Fig. 1. An example of the F0 prediction produced by our system, compared to the original F0 curve.

3.2. Subjective listening tests

In order to find out which proposed model sounds the closest to natural speech, we conducted an online MUSHRA (MUlti-Stimulus test with Hidden Reference and Anchor) listening test [19]. 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 natural and vocoded sentences with the synthesized sentences using DNN-predicted features. The vocoded reference sentences were synthesized applying impulse-noise excitation using the original F0 and MGC of the signals. We also used an anchor sentence, which had constant F0 and a distorted version of the original MGC features. Ten sentences were selected for the test, which were not included in the training database. Here, the test had two aims. First, we sought to compare the various input representation approaches, and second, we attempted to evaluate the influence of the accuracy of the predicted F0 contour on speech quality. This is why the test was broken down into two sub-tests. In sub-test #1, the *Baseline*, *17-frames+fs*, and *5-frames* input

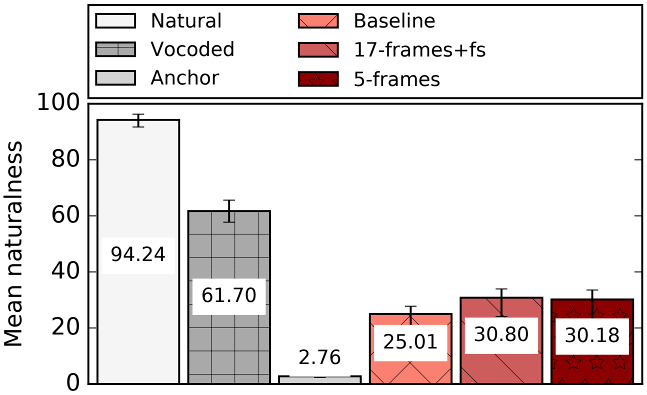

Fig. 2. Results of the listening test for the comparison of the input representation approaches.

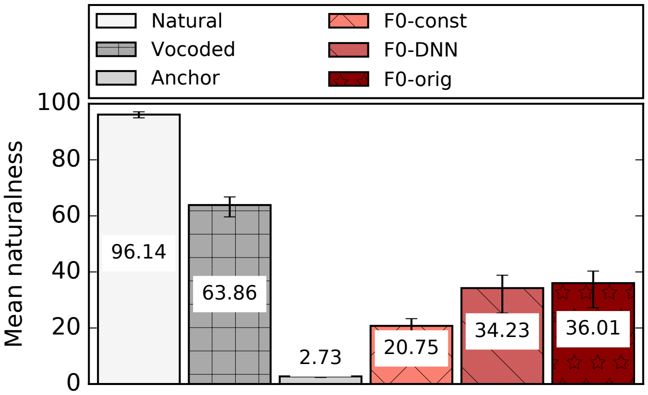

Fig. 3. Results of the listening test for the comparison of the F0 prediction methods.

representation strategies were compared, whereas with sub-test #2 we compared the *F0-const*, *F0-DNN*, and *F0-orig* variants of F0 prediction. All of the ten sentences, and all of the sentence variants (6 variants in each case) appeared in randomized order (different for each listener). In the MUSHRA test the listeners had to rate the naturalness of each stimulus relative to the reference (which was the natural sentence), on a scale from 0 (highly unnatural) to 100 (highly natural).

3.2.1. Results of the Listening Test

Altogether 24 listeners participated in the main test (13 females, 11 males). All of them were native speakers of Hungarian with no known hearing loss, and four of them were speech experts. The subjects were between 18–74 years (mean: 28 years). On the average, the whole test took 18 minutes to complete. The MUSHRA scores of the listening test are presented in Fig. 2 for sub-test #1 and in Fig. 3 for sub-test #2 (the errorbars showing the 95% confidence intervals). In general, the 'natural' sentences attained roughly 100% on the naturalness scale. The 'vocoded' references achieved naturalness scores around 60%, and the anchor sentences were rated very low by the listeners. The three utterance types where the features were predicted using DNNs were ranked between 20–36% by the subjects, indicating that they are roughly half-way between the vocoded references and the anchor in terms of naturalness.

The rankings by the listeners were compared by using Mann-Whitney-Wilcoxon ranksum tests as well, with a 95% confidence level. In both sub-tests, the DNN-based synthetic signals significantly differed from the natural, anchor, and vocoded references. In sub-test #1, the *17-frames+fs* and *5-frames* strategies proved both significantly different from *Baseline*, while their naturalness was not judged to be significantly different from each other. In sub-test #2, the *F0-const* model ranked significantly lower than *F0-DNN* and *F0-orig* and, most importantly, the scores of the latter two were not significantly different. This result tells us that the listeners could not differentiate the synthetized sentences with DNN-predicted F0 from those using the original F0 curve.

4. CONCLUSIONS

Here, we described our experiments for performing F0 estimation in ultrasound-based articulatory-to-acoustic mapping. We used two separate fully-connected DNNs to estimate the voicing of the actual frame, and to predict the F0 for voiced frames. Using objective evaluation metrics, we attained a correlation rate of 0.74. According to our subjective listening tests, the listeners ranked the synthesized sentences with the original and the DNN-predicted F0 curves as being equally natural. These findings justify that articulatory-to-F0 prediction is promising, even if the input features (ultrasound data in our case) do not contain direct measurements of the vocal cord vibration. We worked with the voice of only one person and we assume that this fact significantly contributed to our good results. While we think that speaker-dependency is not a drawback, as future SSI systems will inherently be personalized, in the future we plan to repeat our experiments with more speakers (both male and female) in the training database. Also, while SSI systems should be able to cope without the vibration of the vocal cords, here our subject was a healthy person producing normal speech. Hence, it is also a future task to evaluate the system with real silent speech.

5. ACKNOWLEDGEMENTS

The authors were partially funded by the National Research, Development and Innovation Office of Hungary (FK 124584). 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.

6. REFERENCES

[1] Bruce Denby, Tanja Schultz, Kiyoshi Honda, Thomas Hueber, James M. Gilbert, and Jonathan S. Brumberg, "Silent speech interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270–287, 2010.

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

[3] Thomas Hueber, Elie-laurent Benaroya, Bruce Denby, and Gérard Chollet, "Statistical Mapping Between Articulatory and Acoustic Data for an Ultrasound-Based Silent Speech Interface," in *Proc. Interspeech*, Florence, Italy, 2011, pp. 593–596.

[4] Thomas Hueber, Gérard Bailly, and Bruce 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.

[5] Aurore Jaumard-Hakoun, Kele Xu, Clémence Leboullenger, Pierre Roussel-Ragot, and Bruce Denby, "An Articulatory-Based Singing Voice Synthesis Using Tongue and Lips Imaging," in *Proc. Interspeech*, 2016, pp. 1467–1471.

[6] Tamás Gábor Csapó, Tamás Grósz, Gábor Gosztolya, László Tóth, and Alexandra Markó, "DNN-Based Ultrasound-to-Speech Conversion for a Silent Speech Interface," in *Proc. Interspeech*, Stockholm, Sweden, 2017, pp. 3672–3676.

[7] Jun Wang, Ashok Samal, and Jordan Green, "Preliminary Test of a Real-Time, Interactive Silent Speech Interface Based on Electromagnetic Articulograph," in *Proc. SLPAT*, 2014, pp. 38–45.

[8] Florent Bocquelet, Thomas Hueber, Laurent Girin, Christophe Savariaux, and Blaise Yvert, "Real-Time Control of an Articulatory-Based Speech Synthesizer for Brain Computer Interfaces," *PLOS Computational Biology*, vol. 12, no. 11, pp. e1005119, nov 2016.

[9] Jose A. Gonzalez, Lam A. Cheah, Phil D. Green, James M. Gilbert, Stephen R. Ell, Roger K. Moore, and Ed Holdsworth, "Evaluation of a silent speech interface based on magnetic sensing and deep learning for a phonetically rich vocabulary," Stockholm, Sweden, 2017, pp. 3986–3990.

[10] Keigo Nakamura, Matthias Janke, Michael Wand, and Tanja Schultz, "Estimation of fundamental frequency from surface electromyographic data: EMG-to-F0," in *Proc. ICASSP*, Prague, Czech Republic, 2011, pp. 573–576.

[11] João Freitas, Artur J Ferreira, Mário A T Figueiredo, António J S Teixeira, and Miguel Sales Dias, "Enhancing multimodal silent speech interfaces with feature selection," in *Proc. Interspeech*, Singapore, Singapore, 2014, pp. 1169–1173.

[12] Jintao Jiang, Abeer Alwan, Lynne E. Bernstein, Patricia Keating, and Ed Auer Jr., "On the correlation between facial movements, tongue movements, and speech acoustics," *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 11, pp. 174–1188, 2002.

[13] Corine A. Bickley and Kenneth N Stevens, "Effects of a vocal tract constriction on the glottal source: experimental and modeling studies," *Journal of Phonetics*, vol. 14, pp. 373–382, 1986.

[14] John R. Westbury and Patricia A. Keating, "On the naturalness of stop consonant voicing," *Journal of Linguistics*, vol. 22, pp. 145–166, 1986.

[15] Zheng-Chen Liu, Zhen-Hua Ling, and Li-Rong Dai, "Articulatory-to-acoustic conversion with cascaded prediction of spectral and excitation features using neural networks," in *Proc. Interspeech*, San Francisco, CA, USA, 2016, pp. 1502–1506.

[16] Cenxi Zhao, Longbiao Wang, Jianwu Dang, and Ruiguo Yu, "Prediction of F0 based on articulatory features using DNN," in *Proc. ISSP*, Tienjin, China, 2017.

[17] Arturo Camacho and John 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] Satoshi Imai, Kazuo Sumita, and Chieko 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.

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