EFFECTS OF F0 ESTIMATION ALGORITHMS ON ULTRASOUND-BASED SILENT SPEECH INTERFACES

Pengyu Dai[1], Mohammed Salah Al-Radhi[1], Tamás Gábor Csapó[1,2]

[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
pengyudaii@gmail.com, {malradhi,csapot}@tmit.bme.hu

ABSTRACT

This paper shows recent progresses on our Silent Speech Interface (SSI) that translates tongue motions into audible speech. In our previous work and also in the current study, the prediction of fundamental frequency (F0) from Ultrasound Tongue Images (UTI) was achieved using articulatory-to-acoustic mapping methods based on deep learning. Here we investigated several traditional discontinuous speech-based F0 estimation algorithms for the target of UTI-based SSI system. Besides, the vocoder parameters (F0, Maximum Voiced Frequency and Mel-Generalized Cepstrum) are predicted using deep neural networks, with UTI as input. We found that during the articulatory-to-acoustic mapping experiments, those discontinuous F0 algorithms are predicted with lower error, and they result in slightly more natural synthesized speech than the Idiap baseline continuous F0 algorithm. Experimental results confirmed that discontinuous algorithms (e.g. Yin) are closest to original speech in objective metrics and subjective listening test.

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

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 Interface" (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 ultrasound tongue imaging (UTI) [2, 3, 4, 5, 6], permanent magnetic articulography (PMA) [8], electromagnetic articulography (EMA) [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, 7, 8, 11]. 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 a support vector machines (SVM) for voiced/unvoiced (V/U) discrimination, and in the second step they applied a Gaussian mixture model (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 result of Nakamura et al. Another two 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].

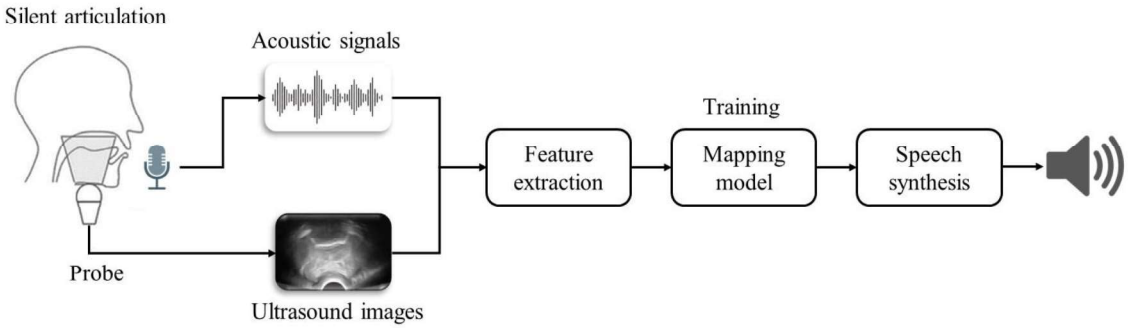

Fig. 1. Workflow of an ultrasound-based silent speech

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 subjective human listening tests [16].

Although there has been some research on articulatory-to-F0 prediction, only two deep learning experiments for estimating the F0 curve from ultrasound tongue images alone are proposed [17, 18]. In a previous study, we presented our results for DNN-based F0 estimation from ultrasound images [18]. In contrast with others worked with EMG signals, our input articulatory representation contains no information directly related to vocal fold vibration. We applied a 2-stage DNN-based approach where one machine learning model seeks to estimate the voicing feature, while another one seeks to predict the F0 value for voiced frames. During the evaluation (synthesis) step, the outputs of the two DNNs are merged. It was achieved by taking the output value of the F0 predictor network where the voicing network decided in favor of voicing, and returning a constant value for frames judged to be unvoiced. In the experiments we attained a correlation rate of 0.74 between the original and the predicted F0 curve. And in subjective listening tests 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. However in the previous experiments only a single F0 estimation algorithm based on Idiap [19] was implemented [17].

Here, we extended our study by investigating different robust F0 estimation techniques: Yaapt [20], Rapt [21], DIO [22] and Yin [23]. In contrast with our recent work where Idiap worked as a continuous pitch algorithm that implemented with a continuous vocoder, the new four algorithms are discontinuous and implemented with a discontinuous vocoder. We discovered in our experiments that all discontinuous algorithms got better values than Idiap (being the baseline of the current paper) in objective and subjective measurements.

2. METHODS

2.1. Data acquisition protocol

Two Hungarian male and two female subjects with normal speaking abilities were recorded while reading sentences aloud (altogether 209 sentences each); and the data of a female speaker was used in our current experiments. The sentences are divided into two distinct sets, 200 were selected for training and validation sets, 9 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 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. In the experiments below, the raw scanline data of the ultrasound was used as input data for the DNNs. The images were reduced to 64×128 pixels (for details see [6]).

2.2. Feature extraction and speech synthesis

The general workflow of ultrasound-based silent speech interface is show in Fig. 1. 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 22 050 Hz. The F0 curve was extracted by Idiap, Yaapt, Rapt, Dio and Yin respectively. We extracted 12 Mel-Generalized Cepstrum-based Line Spectral Pair (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 synthesizer 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 [24] to reconstruct the speech signal.

Table 1. Average objective scores based on synthesized speech signals. Bold value denotes the best results.

Method	Evaluation Metric				
	IS	LLR	CEP	fwSNRseg	ESTOI
Idiap (baseline)	4.4821	0.6078	4.5801	5.7718	0.3645
Rapt	1.1673	0.5014	3.9928	6.9196	0.3897
Yaapt	**0.5664**	**0.4772**	**3.8166**	**7.1242**	**0.4134**
DIO	1.4039	0.5103	3.9604	7.0647	0.3881
Yin	3.0025	0.5397	4.0710	6.8494	0.3754

2.3. DNN-based Fundamental Frequency Estimation

DNNs were used in 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. The first tasks, since V/U decision for each frame has a binary output, we treated it as a classification task. While working on the same input images, the second DNN seeks to learn the F0 curve. This second task was viewed as a regression problem, and it was trained with the voiced segments from the training data. The outputs of the two DNNs were merged during the evaluation (synthesis) step. For Idiap, this is achieved by taking the output value of the F0 predictor network where the voicing network decided in favor of voicing, and returning a constant value for frames judged to be unvoiced. For Yaapt and another three algorithms, only those predicted F0 values from voiced frames are used.

We trained DNNs with 5 hidden layers of 1000 ReLU neurons. The F0 parameter was predicted together with the gain and the 12 LSP parameters. 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.

To evaluate the best F0 predicting algorithm via subjective listening test, we synthesized 2 reference sentences. To have an upper glass ceiling, we synthesized sentences using the original F0 curve (*natural* in Fig. 2). To have a benchmark / lower anchor version, we synthesized sentences using a constant F0 (*const F0* in Fig. 2), where the V/U network predicted the voicing of the actual ultrasound images.

3. RESULTS AND DISCUSSION

3.1. Objective Evaluation

Performance of F0 detection algorithms are evaluated by comparing their synthesized speech and original speech. 5 metrics are used: IS (Itakura–Saito) [25], LLR (log likelihood ratio) [25], CEP (cepstrum distance measures) [26], fwSNRseg (frequency-weighted segmental SNR) [27] and ESTOI (Extended ShortTime Objective Intelligibility) [28]. IS and LLR directly calculate the distance between two sets of linear prediction coefficients (LPC) on the original and the predicted speech, while CEP distance provides an estimate of the log spectral distance between two speeches. fwSNRseg was adopted in time domain for the error criterion. ESTOI calculates the correlation between the temporal envelopes of original and predicted speech. This objective evaluation was done on test data (9 sentences)

Table 1 list the results of various measurement methods (note that our goal is to minimize IS, LLR and CEP, while maximize fwSNRseg and ESTOI). Comparing the baseline with others, we can observe that discontinuous algorithms get better score than the baseline in every metrics. It shows that speech signal synthesized by predicted discontinuous F0 curve are much closer to original speech signal. It is clear that F0 predicted by discontinuous algorithms with discontinuous vocoder have better performance than the baseline.

3.2. Subjective Evaluation

In order to find out which investigated model is closer to natural speech, we conducted an online MUSHRA-like (MultiStimulus test with Hidden Reference and Anchor) listening test [30]. 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 synthesized baseline sentences with the synthesized sentences using another four discontinuous F0 extraction algorithms. We used a benchmark/ lower anchor sentence, which had constant F0 and a distorted version of the original MGC features. Five sentences were selected for the test, which are not included in the training database. All sentences appeared in randomized order (different for each listener). In the MUSHRA 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 (highly unnatural) to 100 (high natural).

Altogether 16 listeners participated in the main test (6 females, 10 males). None of them indicated any hearing loss. The subjects were between 21-47 years (mean 24 years). On the average, the whole test took 12 minutes to complete. Fig. 2 shows the average naturalness score for these experimented algorithms. The benchmark version (*const F0*) achieved the lowest score, while the natural sentences (*natural*) were rated the highest, as expected. Comparing with other discontinuous algorithms, the baseline Idiap get the lowest score, which means all discontinuous algorithms based predicted sentences sound more natural than baseline. We also noticed that the score of four discontinuous algorithms are very similar. The

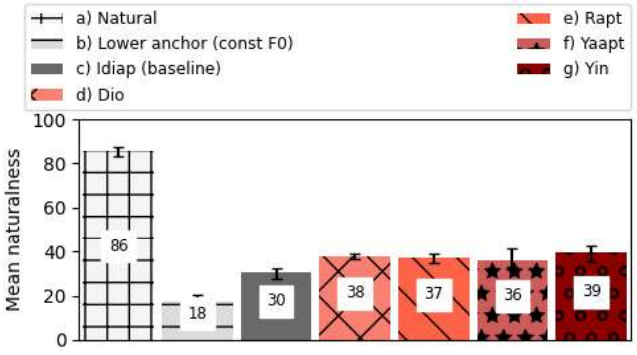

Fig. 2. Results of the subjective listening test. The error bars show the 95% confidence intervals.

reason might be their synthesized sentences are relatively close and it is hard for human being to distinguish their subtle differences. To check the statistical significance of the differences, we conducted Mann-Whitney-Wilcoxon rank-sum tests with a 95% confidence level, showing that the result of the Yin algorithm was significantly different from the baseline, while the other differences are not significant.

4. CONCLUSIONS

Here we described our experiments for comparing several discontinuous F0 estimation algorithms with a continuous baseline one in ultrasound-based articulatory-to-acoustic mapping. We used four accurate discontinuous F0 estimation algorithms to predict the F0 value of voiced frames. The results of objective and subjective evaluation demonstrated that F0 predicted by discontinuous algorithms and the synthesized sentences outperform the one based on continuous F0 (baseline). The experiments were run on the voice of only one Hungarian female speaker. In the future we plan to repeat our experiments with more speakers (both male and female) and also with English data. Besides, it will be worth to apply recurrent neural networks to take into account the sequential nature of articulatory and speech data. For a practical Silent Speech Interface, it will be necessary to apply speaker adaptation techniques, i.e. in the future we plan to test how the UTI-to-F0 algorithms trained on one speaker work with other speakers or with real silent articulation.

5. ACKNOWLEDGEMENTS

The authors were partially funded by the National Research, Development and Innovation Office of Hungary (FK 124584, PD 127915 grants). The Titan X GPU for the deep learning experiments was donated by the NVIDIA Corporation. We would like to thank Gábor Gosztolya for his comments on this manuscript. We thank the listeners for participating in the subjective test.

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, pp. 685–688, 2004.

[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, pp. 593–596, 2011.

[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, pp. 723–726, 2012.

[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*, pp. 1467–1471, 2016.

[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, pp. 3672–3676, 2017.

[7] 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," In *Proc. ISCA, Interspeech*, Stockholm, Sweden, pp. 3986–3990, 2017.

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

[9] 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.

[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, pp. 573–576, 2011.

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

[12] Jintao Jiang, Abeer Alwan, Lynne E. Bernstein, Patricia Keating, and Ed Auer, "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, pp. 1502–1506, 2016.

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

[17] Tamás Gábor Csapó, Mohammed Salah Al-Radhi, Géza Németh, Gábor Gosztolya, Tamás Grósz, László Tóth, Alexandra Markó, "Ultrasound-based Silent Speech Interface Built on a Continuous Vocoder", in *Interspeech*, pp. 894-898, 2019.

[18] Tamás Grósz, Gábor Gosztolya, László Tóth, Tamás Gábor Csapó, and Alexandra Markó, "F0 Estimation for DNN-Based Ultrasound Silent Speech Interfaces," in *Proc. ICASSP*, pp. 291-295, 2018.

[19] Philip N. Garner, Milos Cernak, Petr Motlicek, "A simple continuous pitch estimation algorithm," *IEEE Signal Processing Letters*, vol. 20, no. 1, pp. 102–105, 2013.

[20] Stephen A. Zahorian, and Hongbing Hu, "A spectral/temporal method for robust fundamental frequency tracking." *The Journal of the Acoustical Society of America*, vol. 123, no .6, pp. 4559-4571, 2008.

[21] Talkin, David, and W. Bastiaan Kleijn. "A robust algorithm for pitch tracking (RAPT)." *Speech coding and synthesis*, pp. 495-518, 1995.

[22] M. Morise, H. Kawahara, and T. Nishiura, "Rapid f0 estimation for high-snr speech based on fundamental component extraction," *IEICE Transactions on Information and Systems*, (Japanese Edition), vol. J93-D, no.2, pp.109-117, 2010.

[23] De Cheveigné, Alain, and Hideki Kawahara. "YIN, a fundamental frequency estimator for speech and music." *The Journal of the Acoustical Society of America* vol. 111, no. 4, pp. 1917-1930, 2002.

[24] 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.

[25] Schuyler R. Quackenbush, Thomas Pinkney Barnwell and Mark A. Clements, *Objective Measures of Speech Quality*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1988.

[26] Kitawaki Nobuhiko, Nagabuchi Hiromi, and Itoh Kenzo, "Objective quality evaluation for low bit-rate speech coding systems", *IEEE Journal on Selected Areas in Communications*, vol. 6, pp. 262–273, 1988.

[27] J. M. Tribolet, P. Noll, B. J. McDermott and R. E. Crochiere, "A study of complexity and quality of speech waveform coders," in *Proc. ICASSP*, Oklahoma, USA, pp.586-590, 1978.

[28] Jesper Jensen and Cees H. Taal, "An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[29] Jianfen Ma, Yi Hu and Philipos C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions", *Journal of the Acoustical Society of America*, vol. 125, no. 5, pp. 3387-3405, 2009.

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