F0 Estimation for DNN-Based Ultrasound Silent Speech Interfaces

Tamás Grósz[1], Gábor Gosztolya[1,2], 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 Informatics, Budapest University of Technology and Economics, 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

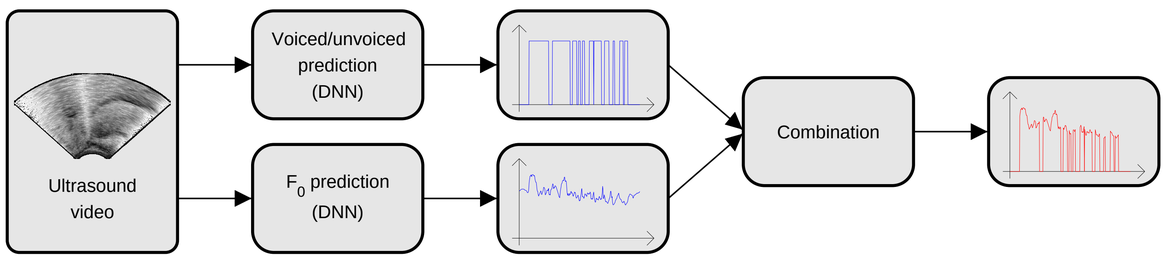
State-of-the-art silent speech interface systems apply vocoders to generate the speech signal directly from articulatory data. Most of these approaches concentrate on estimating just the spectral features of the vocoder, and use the original F0, a constant F0 or white noise as excitation. This solution 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.

Input data: Ultrasound videos

- One Hungarian female speaker, 438 read sentences.
- Tongue ultrasound data was recorded with a Micro system in midsagittal orientation.
- Speech recorded with an Audio-Technica microphone.
- Ultrasound images were reduced to 64×119 pixels.
- A correlation-based feature selection method was also used to further reduce the size of the images.

F0 estimation using DNNs

Separate networks were used to estimate the voicing and the value of F0:



Experimental Setup

DNN for articulatory-to-acoustic mapping

- Fully connected feed-forward network.
- 5 hidden layers with 1000 ReLU neurons.
- 14 linear neurons in the output to estimate the vocoder parameters.
- Binary classification in the case of the U/V network.

Synthesis

For speech synthesis we applied the SPTK vocoder. The synthesizer used the 12 MGC-LSP parameters, the gain and the F0 curve (all estimated by a DNN).

Listening test

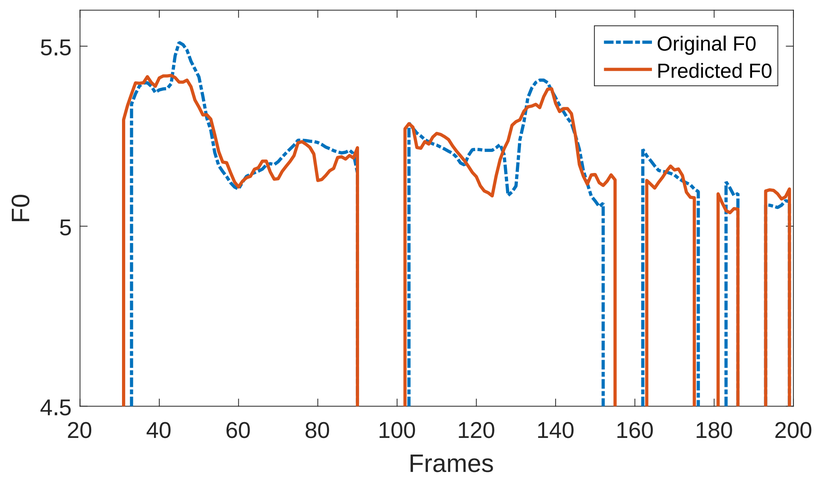
- Online MUSHRA-like listening tests.
- 10 sentences from the test set, 24 listeners (13 females and 11 males).

Conclusions

Here, we described our experiments for performing F0 estimation in ultrasound-based articulatory-to-acoustic mapping. 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 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. At the same time, we think that speaker-dependency is not a drawback, as future SSI systems will inherently be personalized.

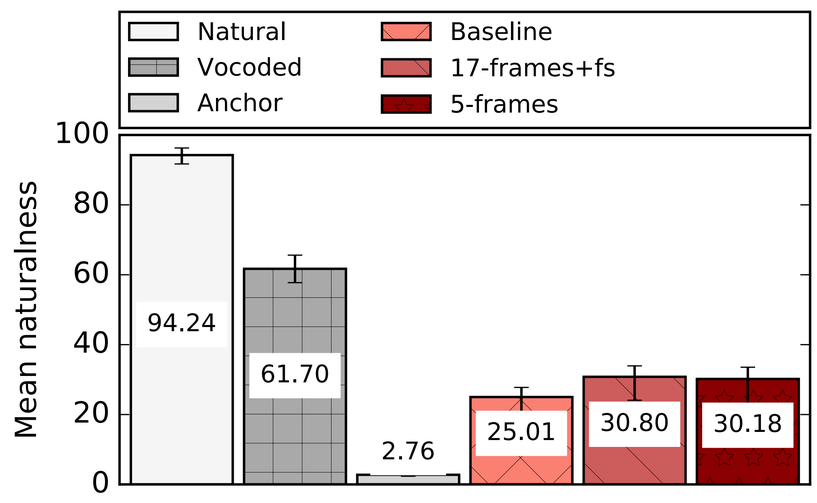
Objective Results

An example F0 prediction produced by our system:



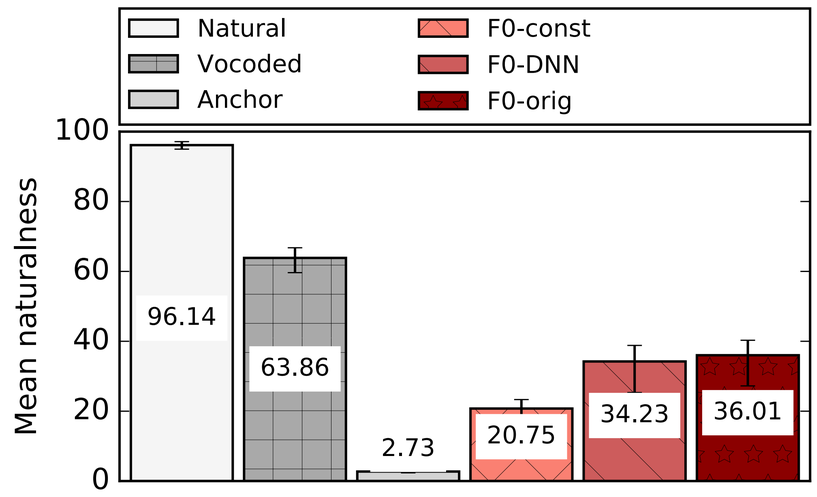
	Development			Test		
Input	Voiced accuracy	NMSE	Corr.	Voiced accuracy	NMSE	Corr.
Baseline (1 frame)	86.1%	0.562	0.719	85.7%	0.577	0.711
5 frames	88.2%	0.476	0.760	87.2%	0.516	0.742
17 frames + f.s.	87.4%	0.506	0.747	86.9%	0.526	0.736

Listening Tests (Input representation)



The *17 frames+f.s.* and *5 frames* strategies both proved significantly better than the *Baseline*, their naturalness was not judged to be significantly different from each other.

Listening Tests (F0 tests)



As can be seen, the listeners could not differentiate the synthetized sentences with DNN-predicted F0 from those using the original F0 curve.

References

[1] Tamás Gábor Csapó et al. "DNN-Based Ultrasound-to-Speech Conversion for a Silent Speech Interface". In: *Proc. Interspeech*. Stockholm, Sweden, 2017, pp. 3672–3676.

[2] Keigo Nakamura et al. "Estimation of fundamental frequency from surface electromyographic data: EMG-to-F0". In: *Proc. ICASSP*. Prague, Czech Republic, 2011, pp. 573–576.