

Investigation of F0 estimation algorithms in Ultrasound-to-Speech synthesis

Pengyu Dai, Mohammed Salah Al-Radhi, Tamás Gábor Csapó

pengyudai@gmail.com

Outline

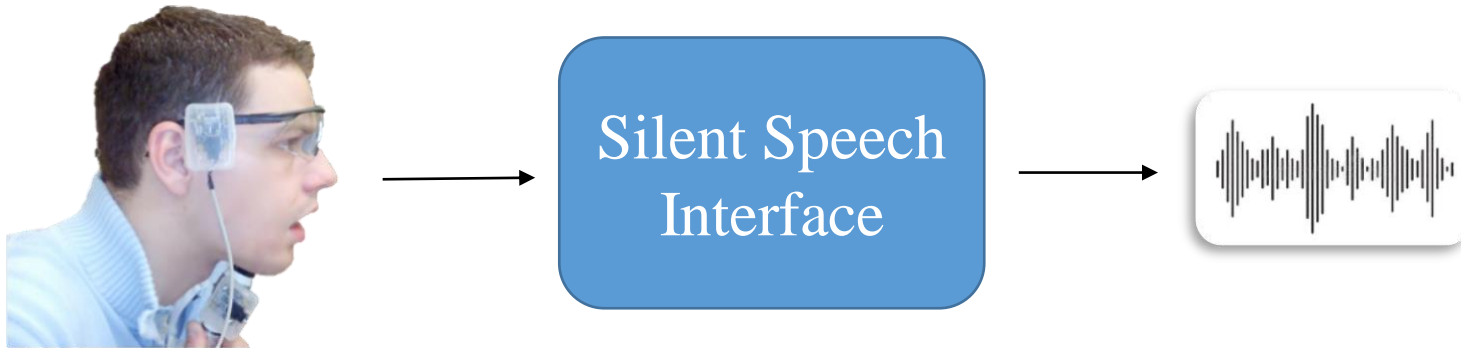
- ✓ Ultrasound tongue imaging (UTI) system
- ✓ F0 prediction based on UTI system
- ✓ Experiment with another 5 F0 estimation algorithms
- ✓ Objective and subjective evaluation

Silent Speech Interface (SSI)

- ✓ The goal: articulatory-to-acoustic conversion

Silent articulation

Synthesized speech



*[Hueber et al. 2016]

- ✓ Useful for:

- speaking impaired people (e.g. after laryngectomy)
- extremely noisy environments
- Silent calls

Typical Technologies

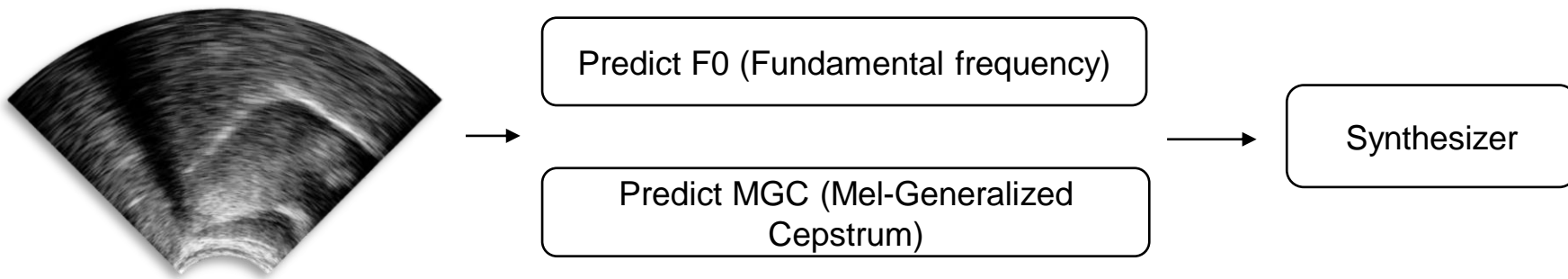
- ✓ ultrasound tongue imaging (UTI)
- ✓ permanent magnetic articulography (PMA)
- ✓ electromagnetic articulography (EMA)
- ✓ electromyography (EMG)
- ✓ multimodal approaches

Feature prediction

✓ Direct synthesis

- Speech generated directly from the articulatory data

✓ Most focus on predicting spectral feature



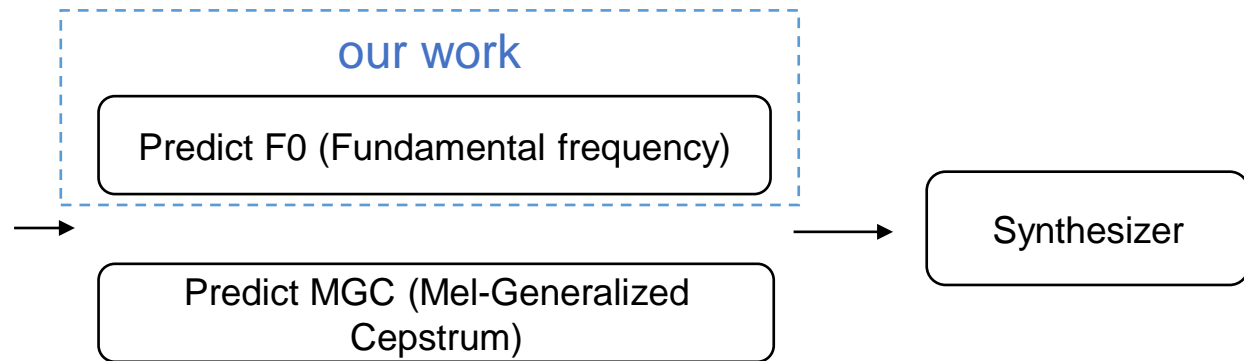
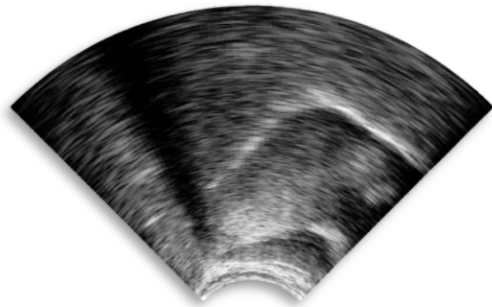
Feature prediction comparison

✓ Spectral features prediction

- Direct relation between tongue movement and the spectral content of speech (e.g. MGC)

✓ F0 prediction

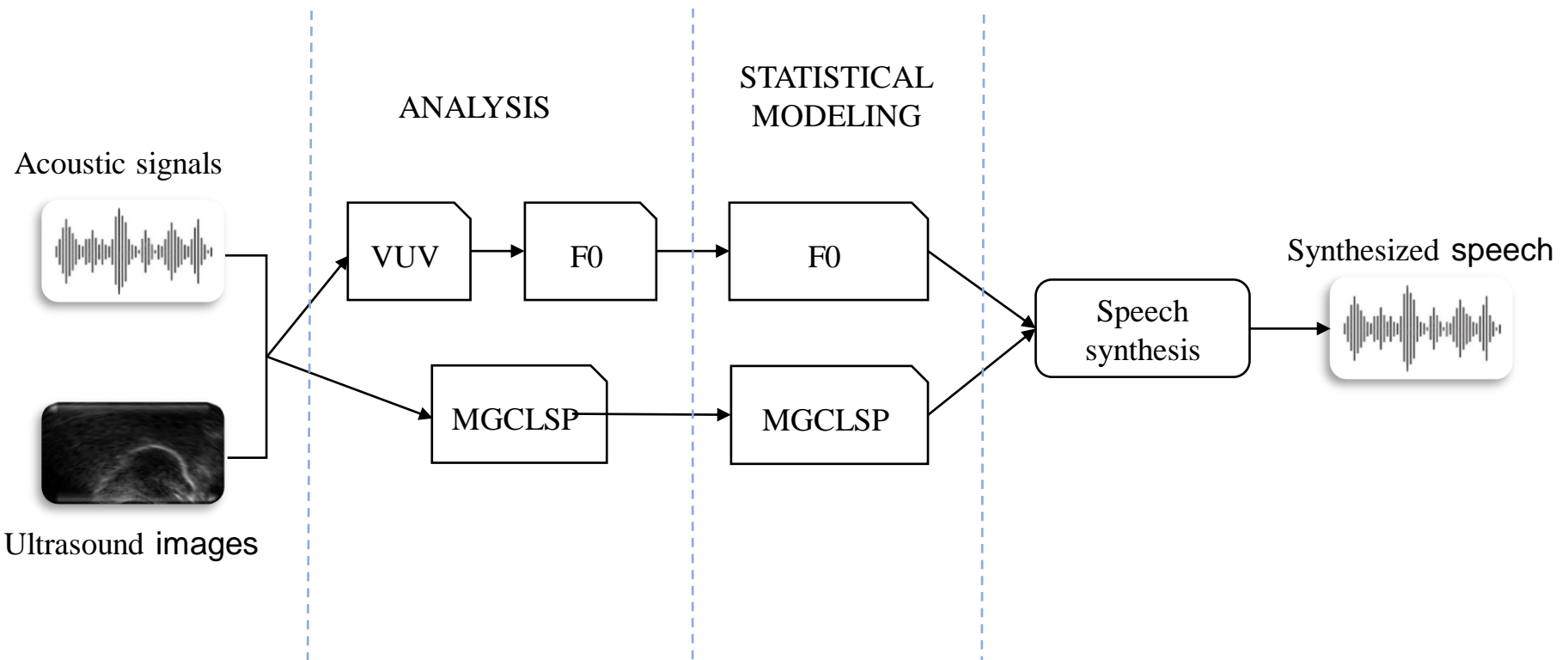
- F0 depends on the vocal fold vibration
- No direct connection with the movement of the tongue/face/lips



DNN-based UTI system

✓ F0 estimate by DNN models

- One model make the voiced/unvoiced decision
- Another one estimate the actual F0 value for voiced frames



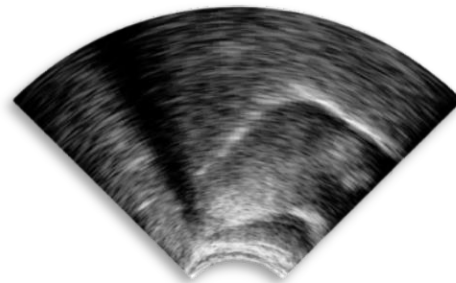
DNN-based UTI system

Data acquisition

- ✓ A female speaker with normal speaking abilities
- ✓ Recorded while reading sentences aloud (altogether 209 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 a Beyerdynamic TG H56c tan omnidirectional condenser microphone



speech

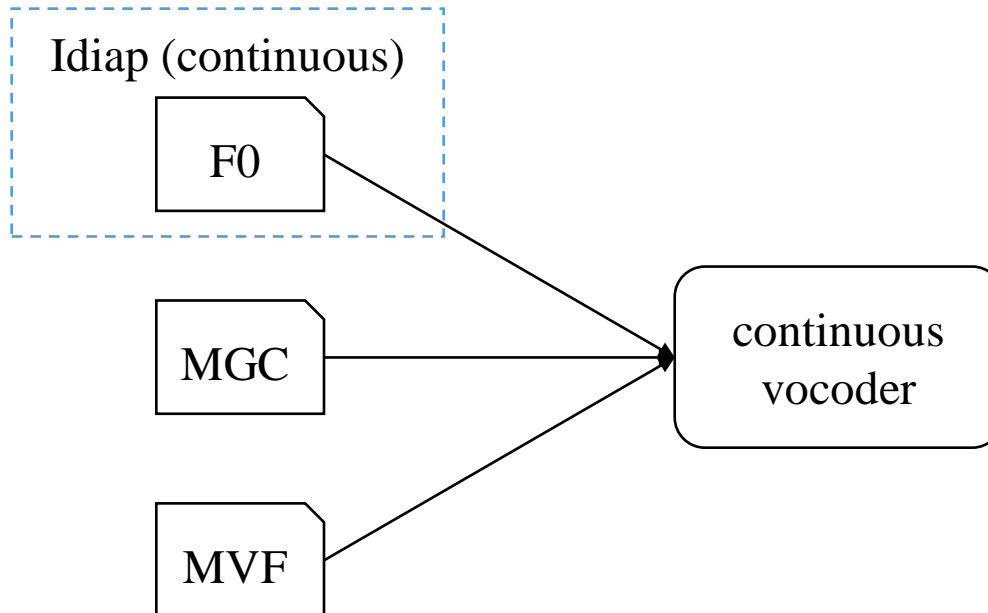


Ultrasound tongue images

DNN-based UTI system

Feature extraction and vocoder

- ✓ 3 features: F0, mgc, mvf
 - F0 extracted by Idiap (a continuous algorithm)
- ✓ A continuous vocoder

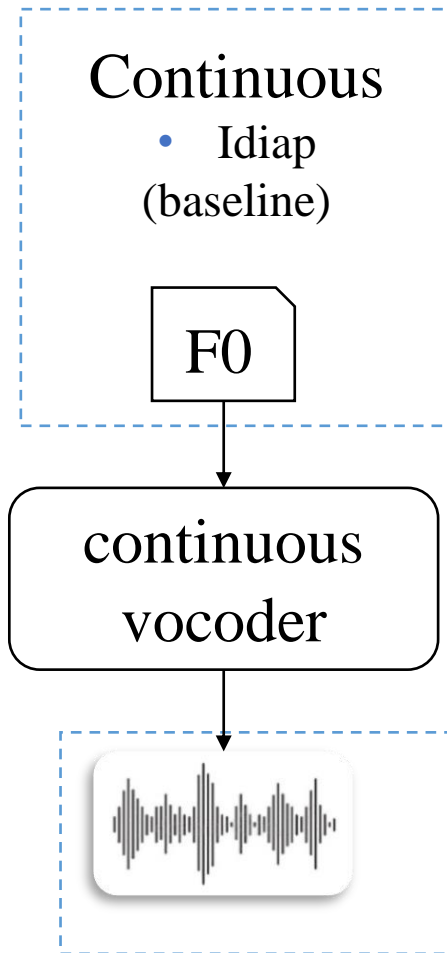


Methodology

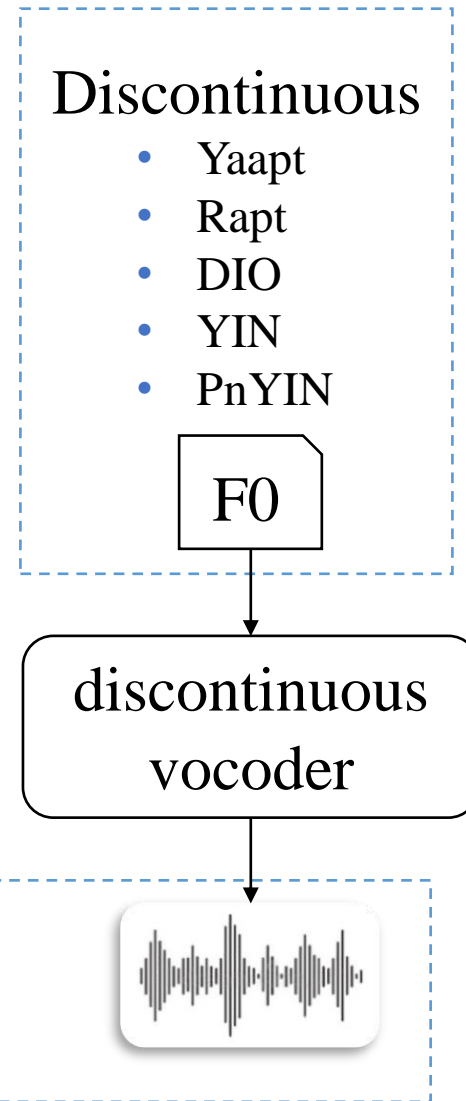
- ✓ Improvement: more F0 estimation algorithms with vocoder could be implemented
- ✓ 5 discontinuous F0 estimation algorithms
 - Rapt
 - Yaapt
 - DIO
 - YIN
 - PnYIN (based on YIN)
- ✓ A discontinuous vocoder
 - Standard SPTK vocoder

Methodology

Baseline model



Experimental model



Objective metrics

- ✓ IS (Itakura–Saito)
 - distance of linear prediction coefficients (LPC) on the original and the predicted speech

- ✓ LLR (log likelihood ratio)
 - linear prediction coefficients (LPC) on the original and the predicted speech

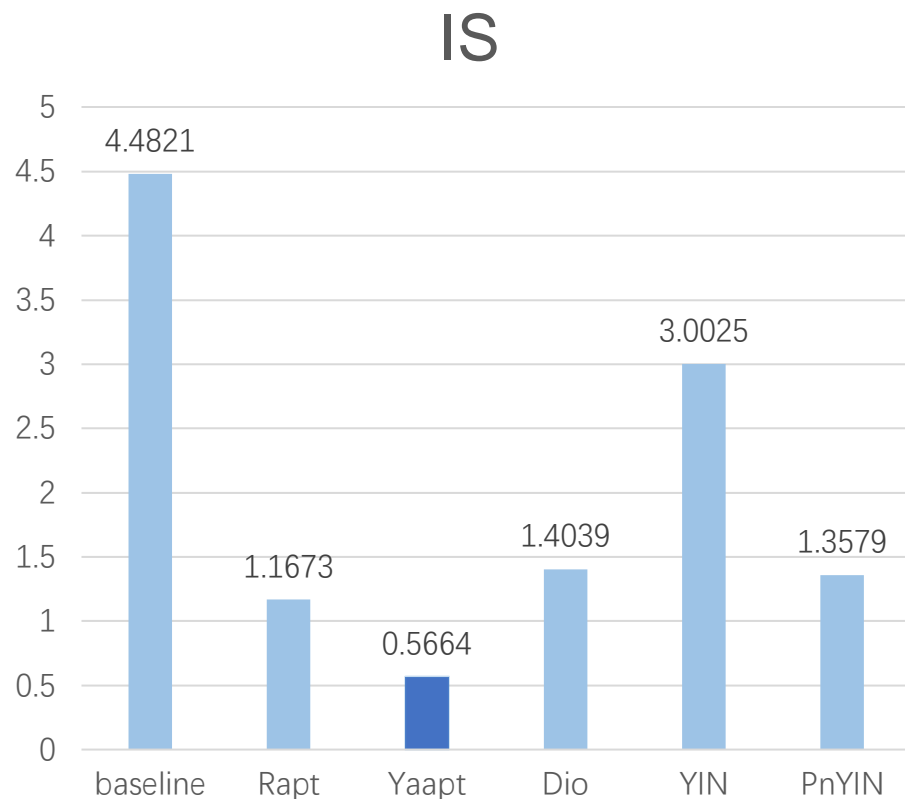
- ✓ CEP (cepstrum distance measures)
 - log spectral distance between two speeches

Objective metrics

- ✓ fwSNRseg (frequency-weighted segmental SNR)
 - for the error criterion

- ✓ ESTOI (Extended ShortTime Objective Intelligibility)
 - the correlation between the temporal envelopes of original and predicted speech.

Objective evaluation



(Smaller value better)

Note:

Original recorded speech

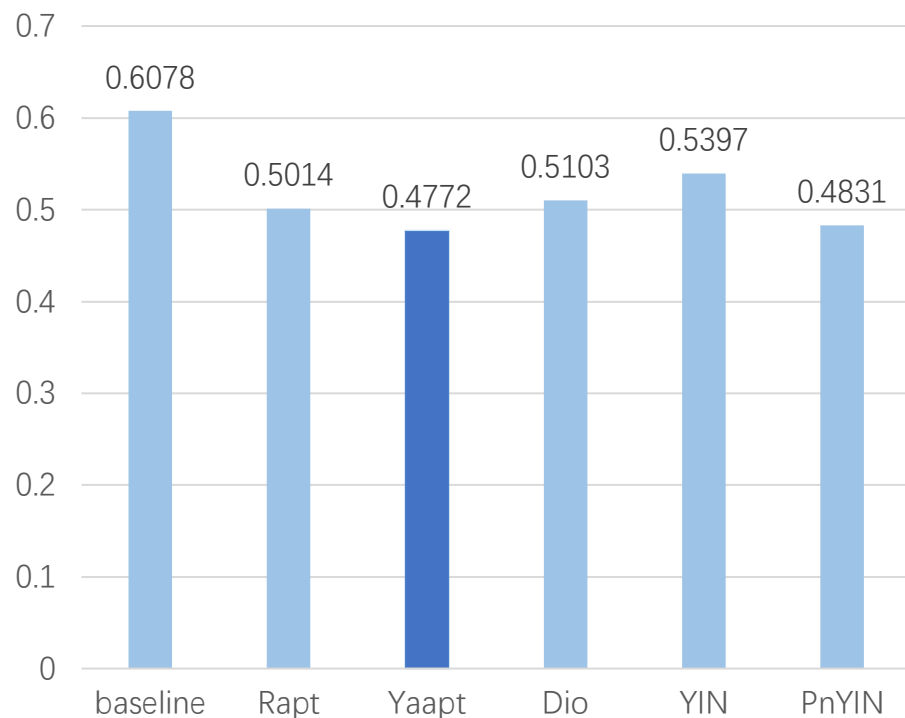
Vs

Synthesized speech using
predicted F0

- Yaapt is the best one
- All discontinuous model better than the baseline

Objective evaluation

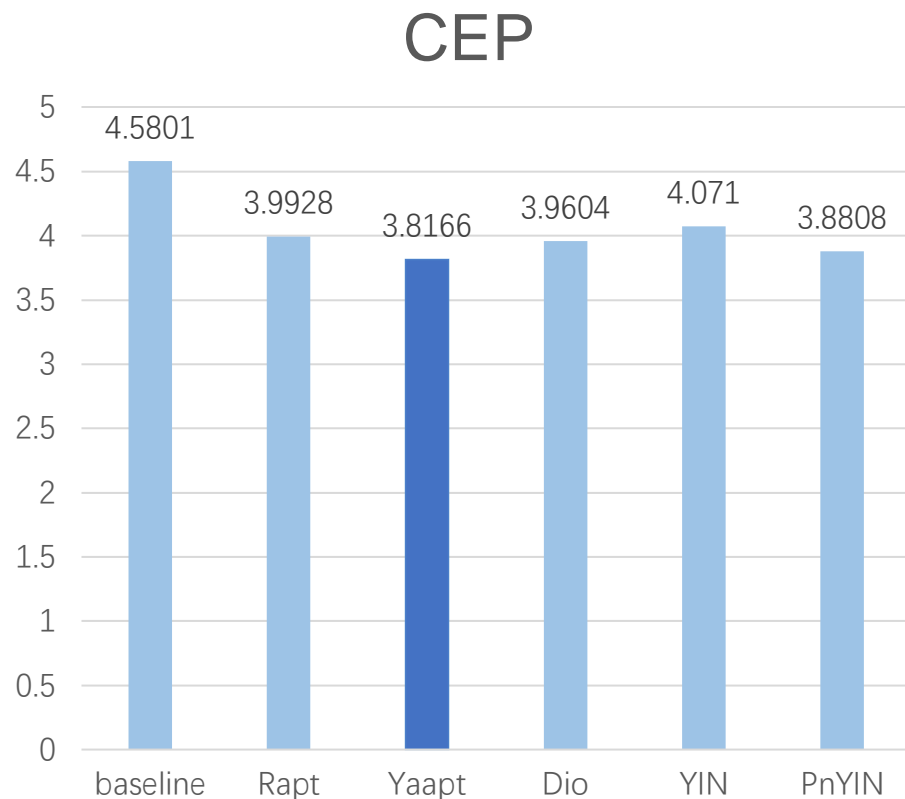
LLR



(Smaller value better)

- Yaapt is the best one
- All discontinuous model better than the baseline

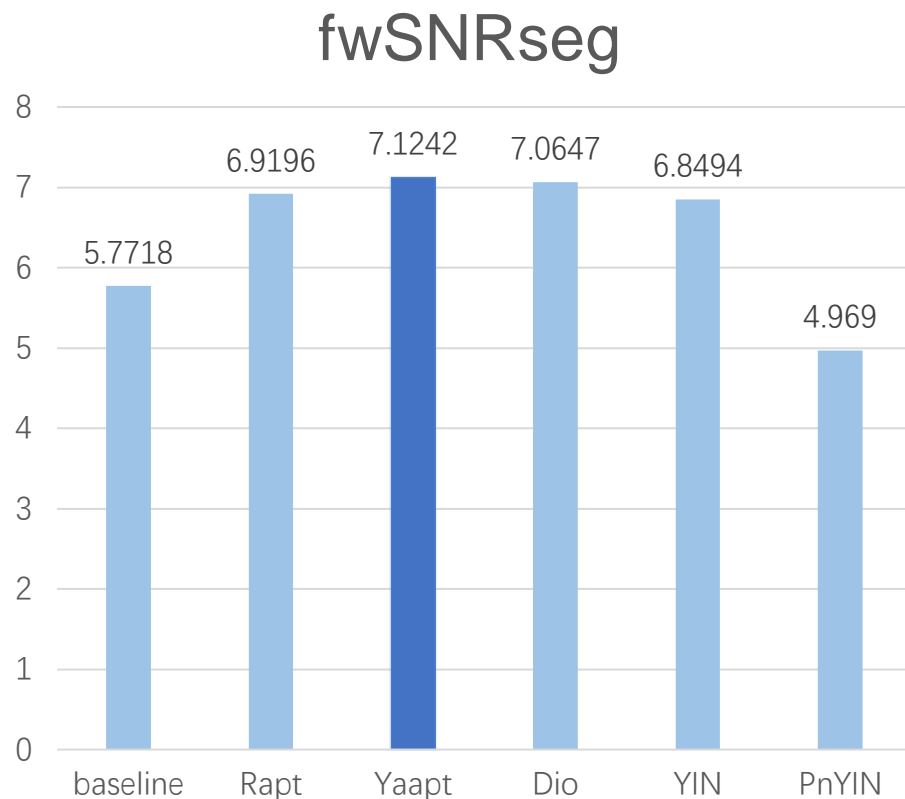
Objective evaluation



(Smaller value better)

- Yaapt is the best one
- All discontinuous model better than the baseline

Objective evaluation

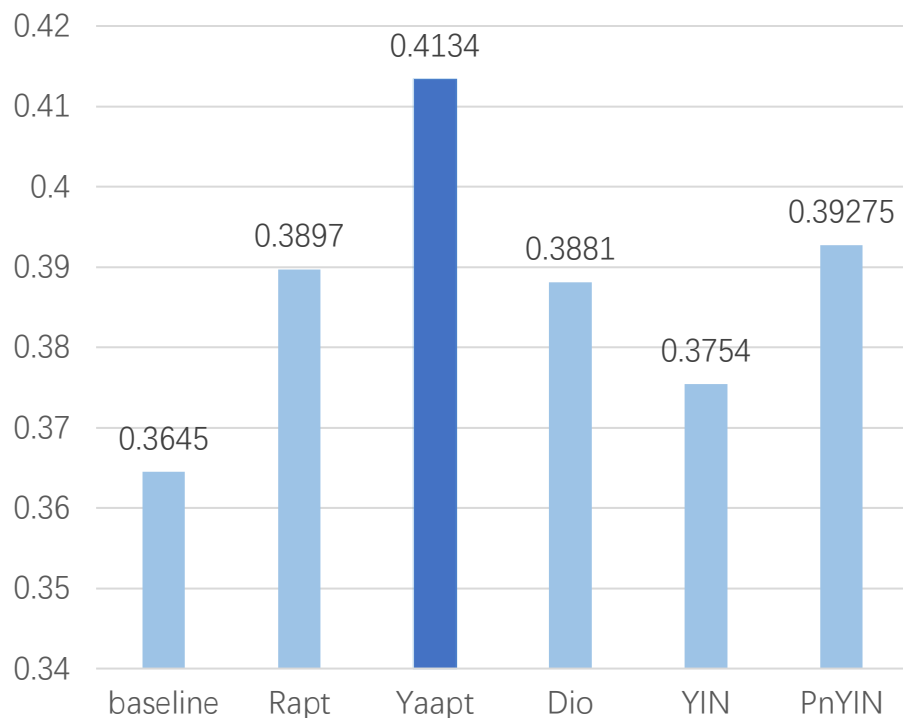


(Bigger value better)

- Yaapt is the best one
- Except PnYin, all discontinuous model better than the baseline

Objective evaluation

ESTOI



(Bigger value better)

- Yaapt is the best one
- All discontinuous model better than the baseline

Objective evaluation

- ✓ F0 predicted by discontinuous algorithms with discontinuous vocoder have better performance than the baseline
- ✓ Yaapt has the best performance followed by PnYIN and Rapt

Method	IS	LLR	CEP	fwSNRseg	ESTOI
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	4.071	6.8494	0.3754
PnYin	1.3579	0.4831	3.8808	4.969	0.3927

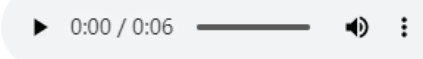


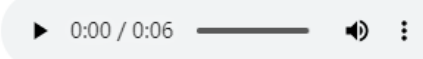

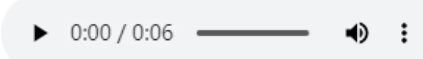

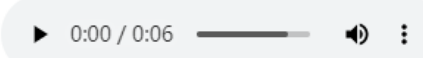

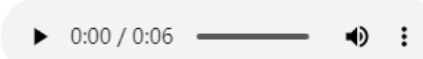



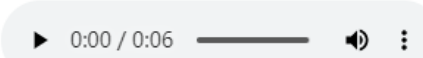



Listening test

- ✓ MUSHRA-like (Multi-Stimulus test with Hidden Reference and Anchor) listening test
 - Five sentences (not included in training data) were selected for the test
 - All sentences appeared in randomized order (different for each listener)
 - Benchmark lower anchor sentence (constant F0)

Listening test

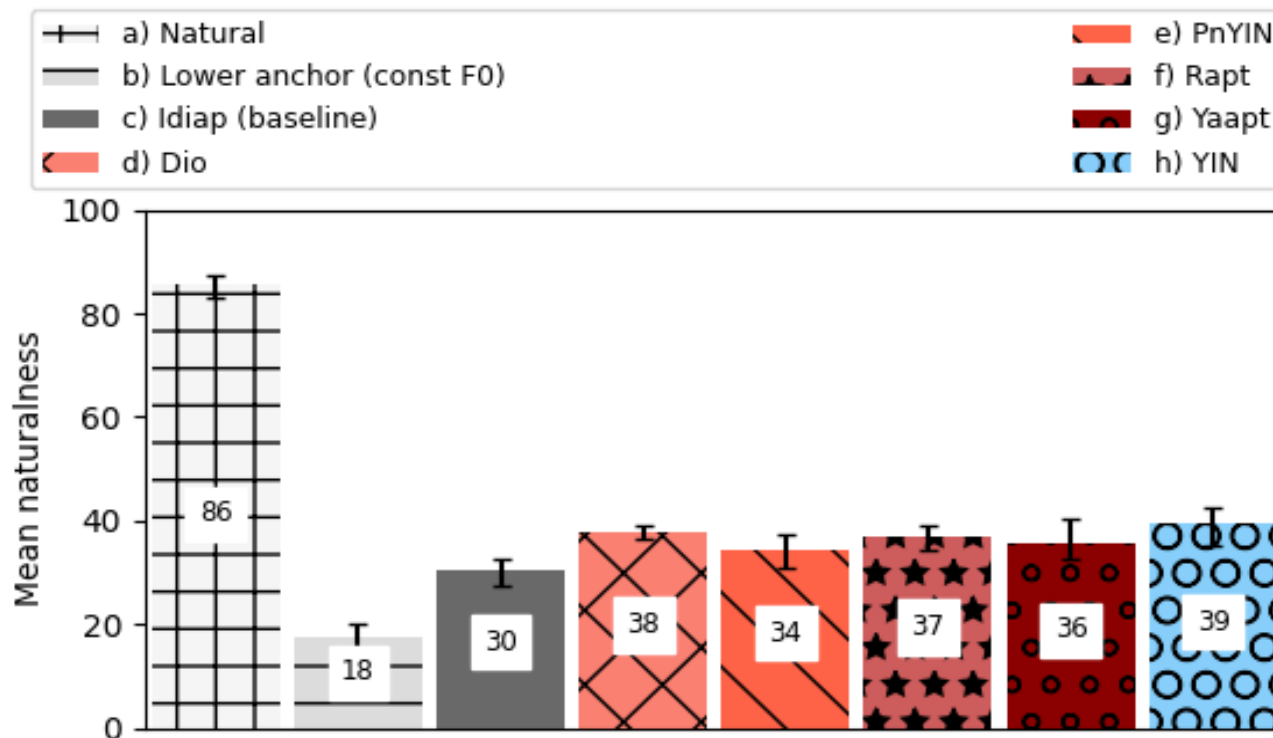
✓ The listeners

- rate the naturalness of each speech in a randomized order
- relative to the reference (the natural sentence), from 0 (very unnatural) to 100 (very natural).

	Recording identifier & sample	Highly unnatural	Unnatural	Intermediate	Natural	Highly natural
reference	 0:00 / 0:06					
a	 0:00 / 0:06					
b	 0:00 / 0:06					
c	 0:00 / 0:06					
d	 0:00 / 0:06					
e	 0:00 / 0:06					
f	 0:00 / 0:06					
g	 0:00 / 0:06					
h	 0:00 / 0:06					

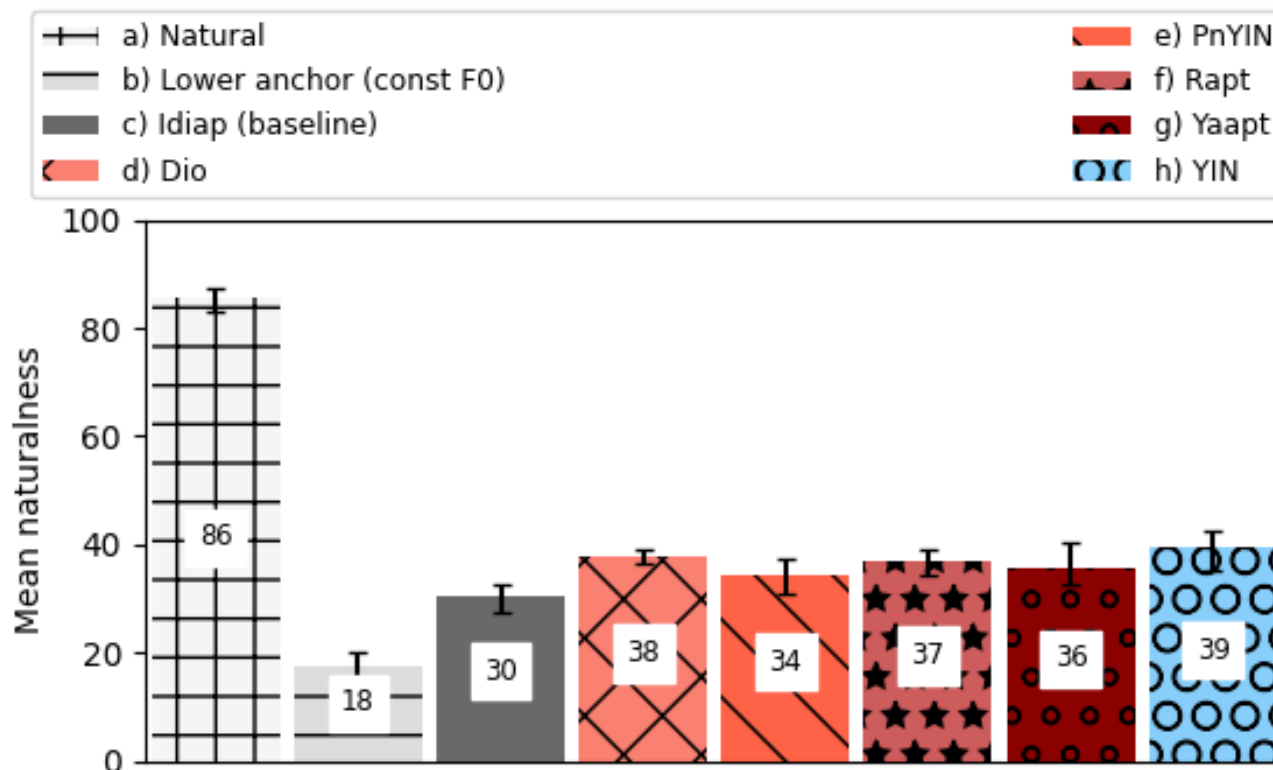
Listening test results

- ✓ 20 listeners
- ✓ benchmark version (18); natural sentences (86)
- ✓ baseline get lowest score (30)
 - all discontinuous algorithms based predicted sentences sound more natural than baseline



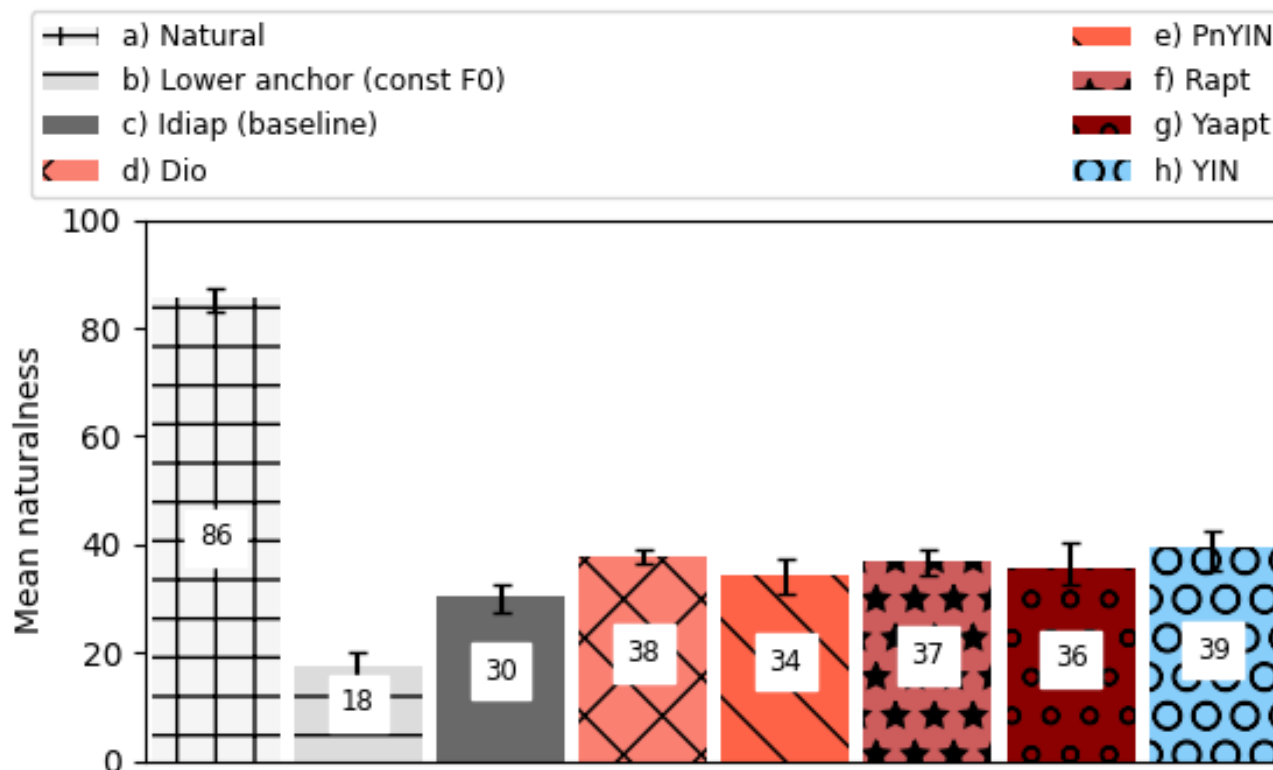
Listening test results

- ✓ Score of discontinuous algorithms are very similar
 - their synthesized sentences are relatively close
 - Hard for human being to distinguish their subtle difference



Listening test results

- ✓ Mann-Whitney-Wilcoxon rank-sum tests (95% confidence level)
 - the result of the YIN algorithm was significantly different from the baseline
 - the differences of other algorithms are not significant.



Samples

Natural



Baseline
(Idiap)



Continuous F0 algorithm
&
Continuous vocoder

DIO



YIN



PnYIN



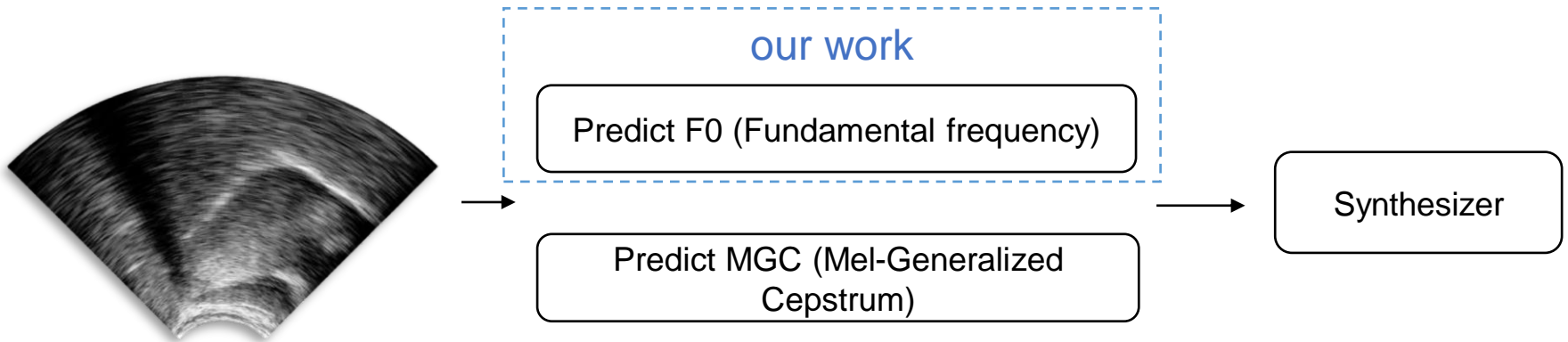
Rapt



Discontinuous F0 algorithm
&
Discontinuous vocoder

Summary and Future work

- ✓ Discontinuous algorithms with discontinuous vocoder have better performance than continuous algorithms with continuous vocoder
- ✓ Yaapt and YIN are slightly better than others
- ✓ The experiment only run on one female Hungarian speaker
- ✓ We plan to experiment with both male and female speaker, and also with English data



Thanks for your attention!

pengyudaii@gmail.com

Key references

- Grósz, T., Gosztolya, G., Tóth, L., Csapó, T. G., & Markó, A., “F0 Estimation for DNN-Based Ultrasound Silent Speech Interfaces,” in Proc. ICASSP, pp. 291-295, 2018.
- Csapó, T. G., Al-Radhi, M. S., Németh, G., Gosztolya, G., Grósz, T., Tóth, L., and Markó, A., “Ultrasound-based Silent Speech Interface Built on a Continuous Vocoder”, in Interspeech, pp. 894-898, 2019.
- Csapó, T. G., Grósz, T., Gosztolya, G., Tóth, L., & Markó, A., “DNN-Based Ultrasound-to-Speech Conversion for a Silent Speech Interface,” in Proc. Interspeech, Stockholm, Sweden, pp. 3672–3676, 2017