GENERATIVE MOMENT MATCHING NETWORK-BASED RANDOM MODULATION POST-FILTER FOR DNN-BASED SINGING VOICE SYNTHESIS AND NEURAL DOUBLE-TRACKING

Hiroki Tamaru†, Yuki Saito†, Shinnosuke Takamichi†, Tomoki Koriyama‡, and Hiroshi Saruwatari†

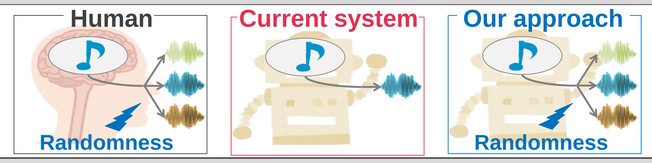
†The University of Tokyo, Japan ‡Tokyo Institute of Technology, Japan

1. SYNOPSIS

We propose a generative moment matching network (GMMN)-based post-filter to

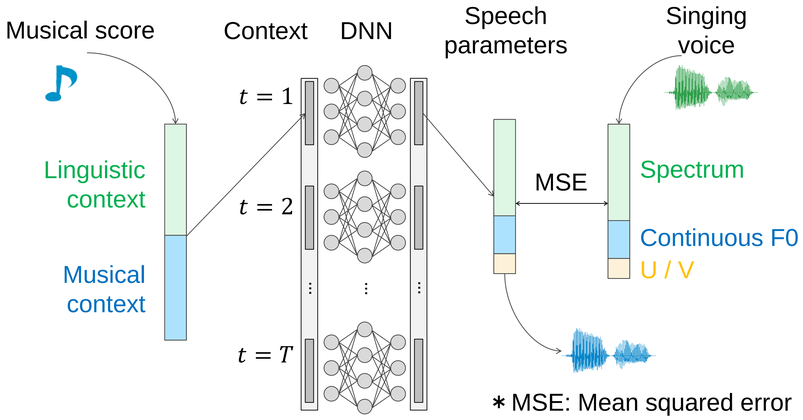
(1) give random *inter-utterance pitch variations* to deterministically synthesized singing voices

(2) give *double-trackedness* to singing voices



2. CONVENTIONAL METHODS

DNN-based singing voice synthesis[1]

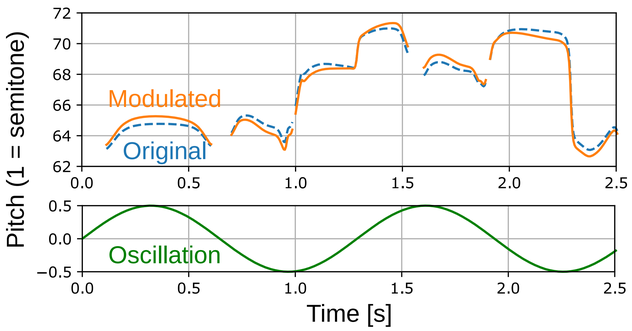


∗ MSE: Mean squared error

Problem: Generated voice has no variation

Artificial double-tracking (ADT)[2]

Mix original voice with its copy modulated using low frequency oscillator to produce double-trackedness



Problem: Unnatural sound artifacts

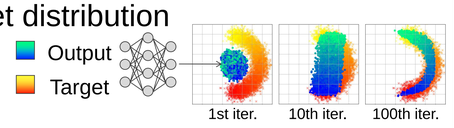
3. GMMN-BASED POST-FILTER AND NEURAL DOUBLE TRACKING

Modulation spectrum (MS)[3]

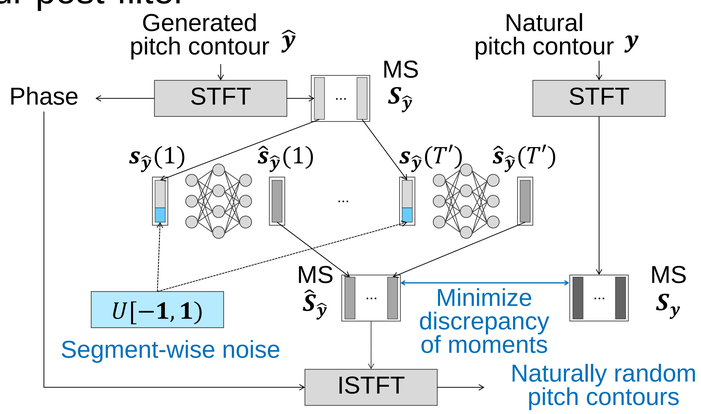
Short-time Fourier transform (STFT) of parameter sequence ⇒ represents envelope of sequence

Conditional GMMN[4]

DNN that minimizes discrepancy of moments between output and target ⇒ enables random sampling from target distribution

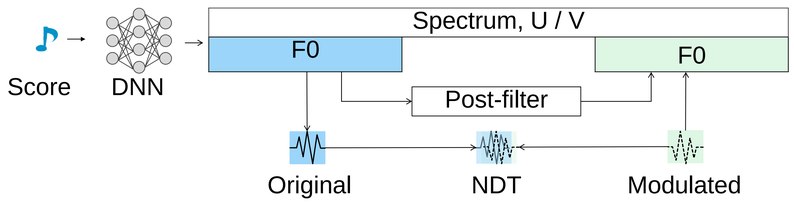


Our post-filter



Neural double-tracking (NDT)

Mix original voice with its copy modulated using our post-filter to produce double-trackedness

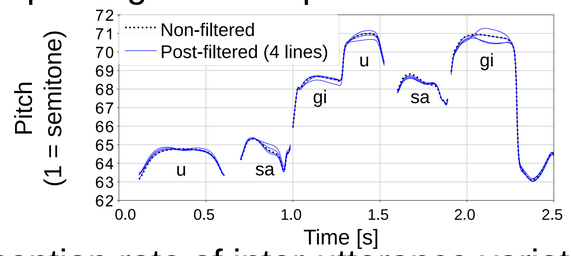


4. EXPERIMENTAL EVALUATION

Experimental conditions

Singing voice corpora	(a) For synthesis model: 58 songs (two female singers) (b) For post-filter: 28 songs (one female singer) (c) For evaluations: 3 songs (one female singer)
DNN architecture	Feed-Forward (see our paper)
Speech params. (including Δ & ΔΔ)	127 dim. (40-dim. mel-cepstral coefficients, Log F0, etc.)
STFT settings for MS	480 ms Hanning window, 240 ms segment shift
Cond. GMMN input	1st order MS and 10 dim. uniform random noise
Evaluation settings	Subjective evaluation, 25 listeners, crowdsourcing

Example of generated pitch contours



Perception rate of inter-utterance variation

Proposed (Two post-filtered versions)	Control (The same version twice)	*p*-Value
0.276	0.176	7.45×10^{-3}

Our post-filter produces perceptible inter-utterance variations!

Double-trackedness score

Sample length	Proposed (NDT)	Conventional (ADT)	*p*-Value
Middle (approx. 4.88 s)	**0.724**	0.276	$< 10^{-10}$
Long (approx. 10.24 s)	**0.736**	0.264	$< 10^{-10}$

Our method achieves more double-trackedness than conventional method!

[References]

[1] Nishimura et al., *Proc. INTERSPEECH*, 2016. [2] Izhaki, *Mixing Audio: Concepts, Practices, and Tools*, 2017. [3] Takamichi et al., *IEEE TRANS. AUDIO SPEECH LANG. PROCESS*, 2016. [4] Ren et al., *Proc. NIPS*, 2016.