

Parallel Voice Conversion Based on a Continuous Sinusoidal Model

Mohammed Salah Al-Radhi, Tamás Gábor Csapó, Géza Németh

Department of Telecommunications and Media Informatics
Budapest University of Technology and Economics, Budapest, Hungary

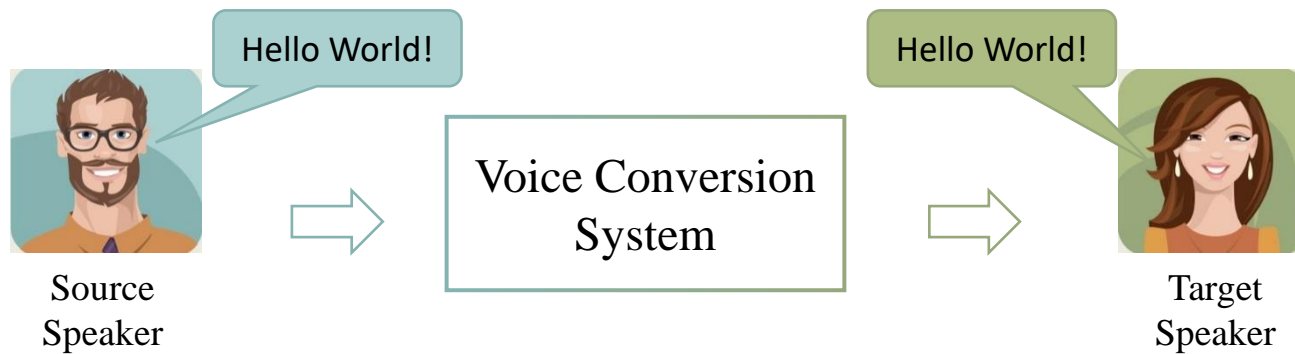
nemeth@tmit.bme.hu

Voice Conversion

What is voice conversion (VC)?

➤ Transformation of the voice characteristics of a (**source**) speaker into those of another (**target**) speaker, while preserving:

- Linguistic content
- Health condition



Application of VC

- ✓ Vocal pathology
- ✓ Voice restoration
- ✓ Speech-to-speech translation
- ✓ Assistive technologies for speech impaired
- ✓ Movies and games dubbing with various persons' voices
- ✓ Security or privacy related usage: hiding the identity of the speaker
- ✓ Computer-aided language learning for personalized perceptual feedback
- ✓ Convert narrow-band speech to wide band speech for telecommunication

VC approaches

Statistical VC main methods:

- Codebook-based mapping
- Non-negative matrix factorization
- Gaussian mixture models
- Maximum likelihood models
- Frequency warping transformations
- Hidden Markov models
- Restricted Boltzmann machines
- Deep neural networks
- Deep belief networks

Problem formulation

- Studies have shown that speech analysis/synthesis solutions play an important role in the overall quality of the converted voice.
 - Source-filter based techniques usually give sound quality and similarity degradation of the converted voice due to:
 - parameterization errors
 - over smoothing} leads to a mismatch in the converted characteristics.
- There is a tradeoff between speaker similarity and computational complexity.

Research Goal

Converting speech into another speaker's voice using continuous sinusoidal model (CSM), which decomposes the source voice into harmonic components to improve VC performance.

The emphasis is on:

- ✓ Good output voice similarity
- ✓ Robustness
- ✓ Computational Complexity

Hypotheses

- Using simple sinusoidal vocoder will
 - obtain higher speaker similarity compared to the conventional methods.
 - whilst at the same time being computationally efficient.
- Using continuous fundamental frequency (contF0) will
 - avoid alignment errors that may happen in voiced and unvoiced segments and can degrade the converted speech, that is important to maintain a high converted speech quality.

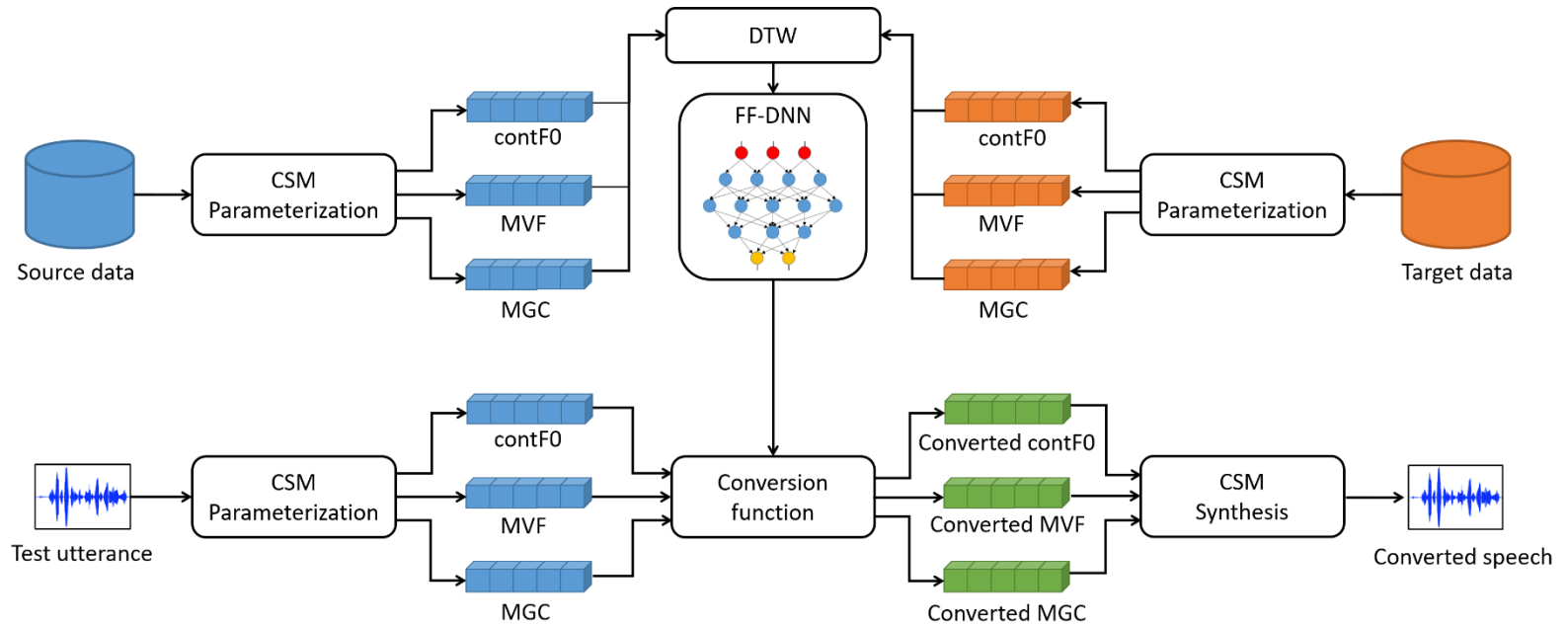
Proposed Methodology

Conversion scheme

Voice Conversion is achieved in 3 stages:

- ❑ **Parameterization** – Analysis the input waveform into acoustic features.
- ❑ **Conversion** – Creating a transformation function between two speakers, using speech segments.
- ❑ **Synthesis** – Synthesizing the converted waveform based on the converted features.

Conversion scheme



Voice conversion process with CSM based waveform generation

Parallel data time alignment

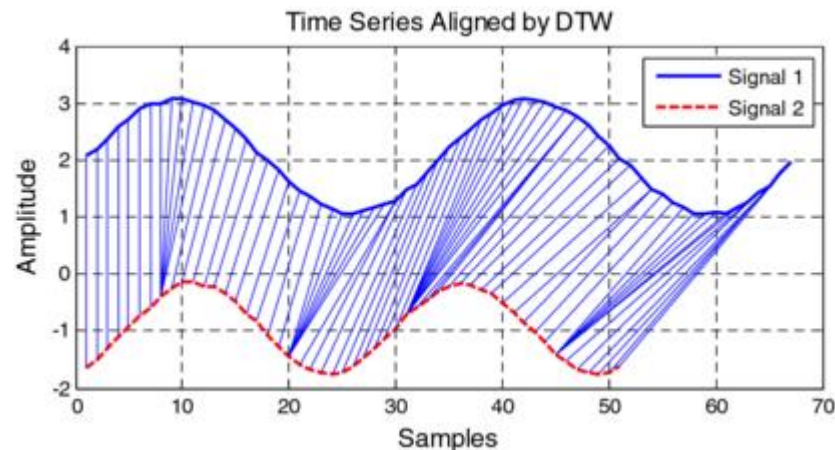
➤ Dynamic Time Warping (DTW)

- is one of the algorithms for measuring similarity between two temporal sequences.
- It allows a non-linear mapping of one signal to another by minimizing the distance between the two.

Source feature: $X = [x_1, x_2, \dots, x_M]$

Target feature: $y = [y_1, y_2, \dots, y_N]$

DTW alignment $X, Y]$



Continuous Sinusoidal Model (CSM)

➤ Analysis step:

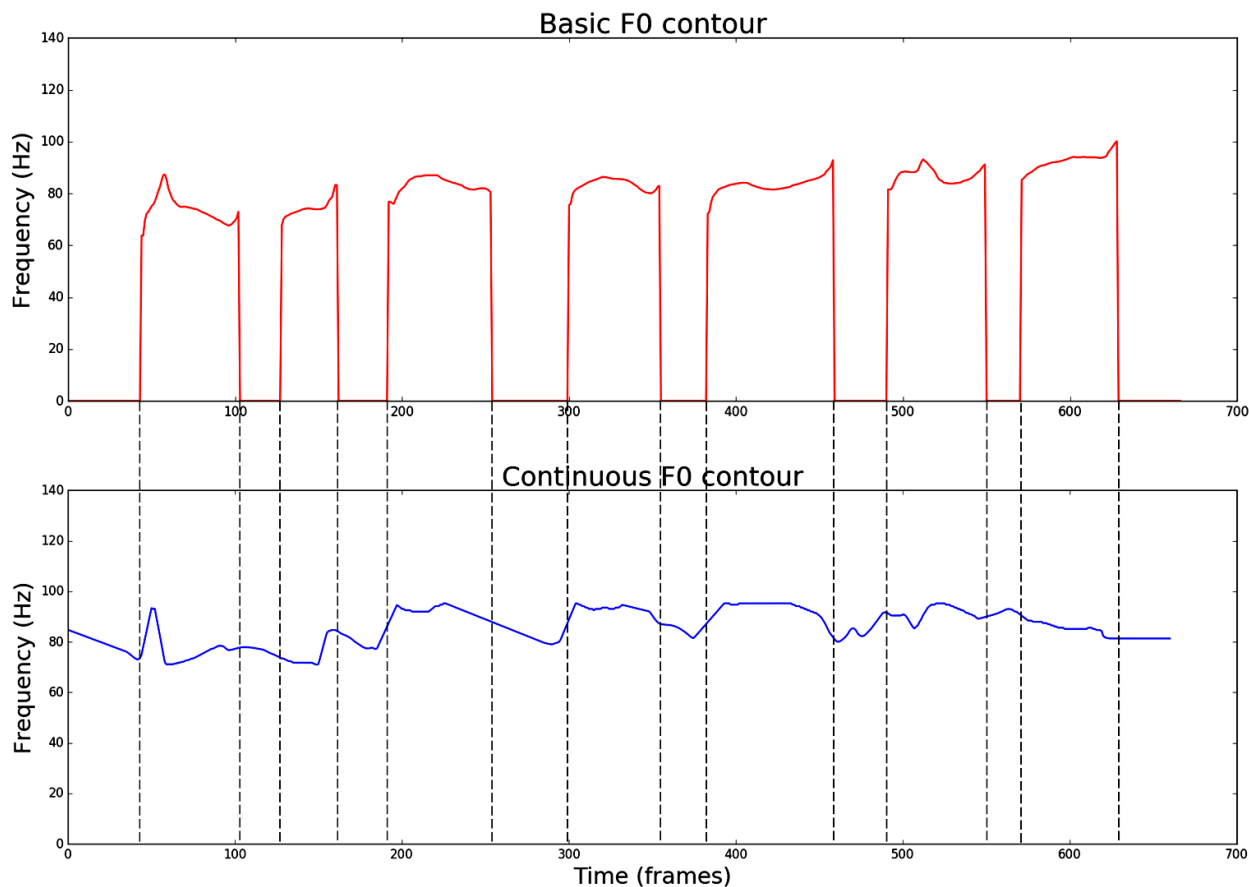
1. **contF0** [Garner et al., 2013]: Continuous fundamental frequency.
2. **MVF** [Drugman and Stylianou, 2014]: Maximum voiced frequency to model the voiced/unvoiced characteristics of sounds.
3. **MGC** [Morise 1994]: Spectral envelope based on Mel-generalized cepstral analysis.

Continuous Sinusoidal Model (CSM)

- Discontinuous F0 model (traditional)
 - continuous ($F_0 > 0$) in voiced regions
 - discontinuous ($F_0 = 0$) in unvoiced regions
 - hard to model boundaries between voiced and unvoiced segments
 - difficult to handle mixed excitation

- Continuous F0 model
 - no voiced/unvoiced decision
 - decrease the disturbing effect of creaky voice
 - easier to handle mixed excitation

Continuous Sinusoidal Model (CSM)

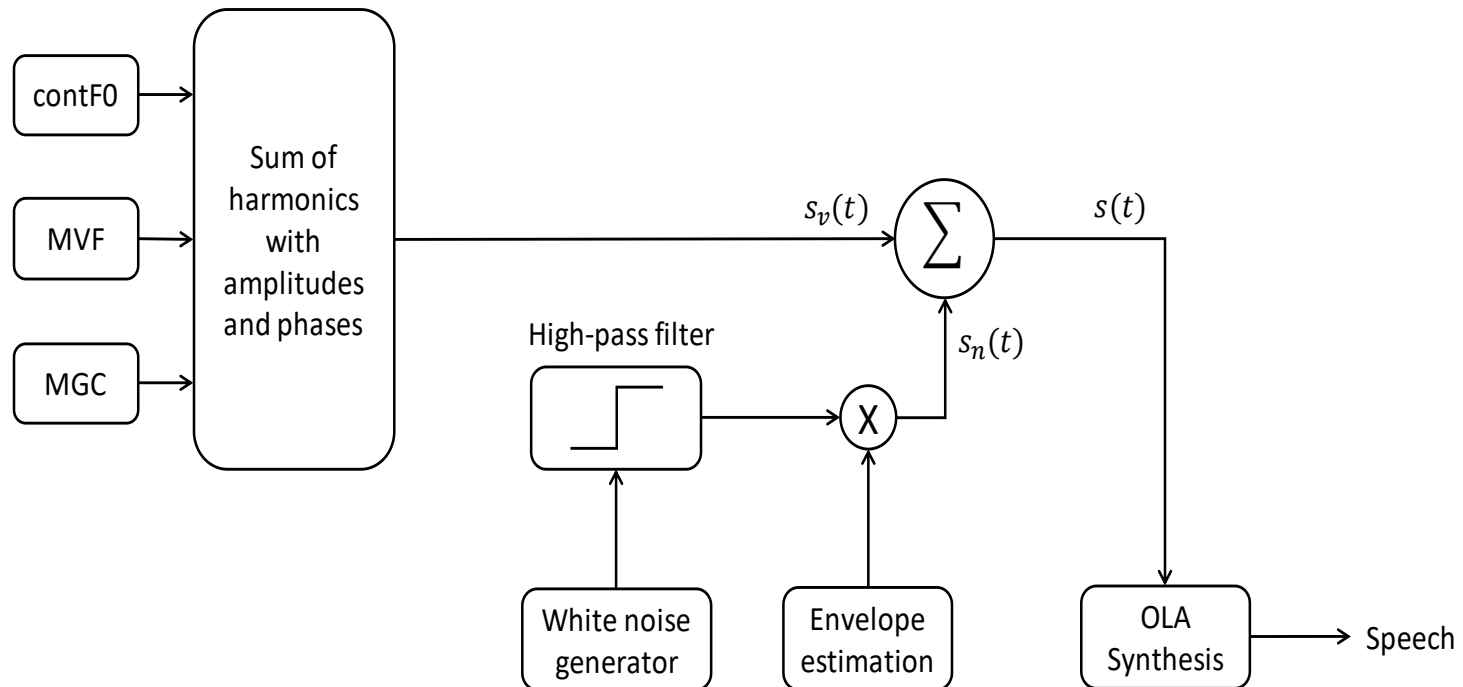


“The girl faced him, her eyes shining with sudden fear.”

Continuous Sinusoidal Model (CSM)

➤ Sinusoidal synthesis:

Decompose the speech frames into a harmonic/voiced component lower band and a stochastic/noise component upper band based on MVF values.



Mathematical representation of CSM

$$s(t) = s_v(t) + s_n(t)$$

$$s_v^i(t) = \sum_{k=1}^{K^i} A_k^i(t) \cos(w_k^i t + \phi_k^i(t)) \quad , \quad w_k^i = 2\pi k(\text{contF0})^i$$

where $A_k(t)$ and $\phi_k(t)$ are the amplitude and phase at frame i , $t = 0, 1, \dots, N$ and N is the length of the synthesis frame. K is the time-varying number of harmonics that depends on the contF0 and MVF :

$$K^i = \begin{cases} \text{round}\left(\frac{\text{MVF}^i}{\text{contF0}^i}\right) - 1, & \text{voiced frames} \\ 0, & \text{unvoiced frames} \end{cases}$$

If the current frame is voiced, the synthesized noise part $n(t)$ is filtered by a high-pass filter $f_h(t)$ with cutoff frequency equal to the local MVF , and then modulated by its time-domain envelope $e(t)$. For unvoiced frames, the harmonic part is obviously zero and the synthetic frame is typically equal to the generated noise.

$$s_n^i(t) = e^i(t) [f_h^i(t) * n^i(t)]$$

Feedforward deep neural network (FF-DNN)

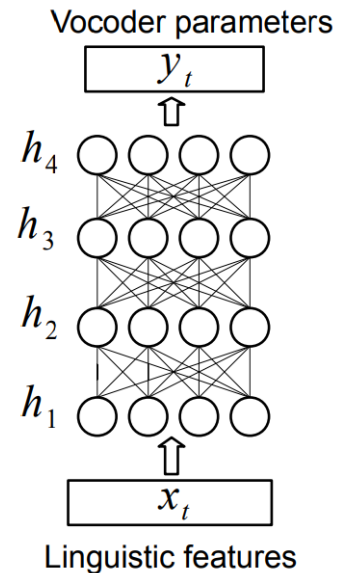
- The input is used to predict the output via several layers of hidden units, each of which performs a nonlinear function.
- For a given
 - input vector sequence $x = (x_1, \dots, x_T)$,
 - hidden state vector sequence $h = (h_1, \dots, h_T)$
 - outputs vector sequence $y = (y_1, \dots, y_T)$

$$h_t = \mathcal{H}(W^{xh}x_t + b^h)$$

$$y_t = W^{hy}h_t + b^y$$

Where W^{xh} and W^{hy} are the weight matrices, b^h and b^y are bias vectors, $W^{hy}h_t$ is a linear regression to predict target features from the activations in the preceding hidden layer, and $\mathcal{H}(\cdot)$ is a nonlinear activation function in a hidden layer, defined as

$$f(x) = \begin{cases} \frac{e^{2x} - 1}{e^{2x} + 1}, & \text{in the hidden layer} \\ x, & \text{in the output layer} \end{cases}$$



Results & Evaluation

Experimental Conditions

- English speaker from CMU-ARCTIC database [Kominek and Black, 2003]
 - BDL (American English, male)
 - JMK (Canadian English, male)
 - SLT (American English, female)
 - CLB (US English, female)
- Each one consisting of 1132 sentences.
- Waveform sampling rate of the database is 16 kHz.
- We conducted intra-gender and cross-gender pairs.
- Training procedures were conducted on an NVidia Titan X GPU.

Experimental Conditions

➤ FF-DNN Setting

- 6 feed-forward hidden layers; each one has 1024 hyperbolic tangent units.
- 90% of these utterances were used for training and the rest were used for testing.
- High performance NVidia Titan X GPU
- Merlin: Open source neural network toolkit [Wu et al. 2016]

➤ Reference baseline systems:

- WORLD
- MagPhase
- Sprocket

Error metrics

1. frequency-weighted segmental SNR

$$\text{fwSNR}_{\text{seg}} = \frac{1}{N} \sum_{j=1}^N \left(\frac{\sum_{i=1}^K W_{i,j} \cdot \log \frac{X_{i,j}^2}{X_{i,j}^2 - Y_{i,j}^2}}{\sum_{i=1}^K W_{i,j}} \right)$$

where $X_{i,j}^2$, $Y_{i,j}^2$ are critical-band magnitude spectra in the j^{th} frequency band of the target and converted frame signals respectively, K is the number of bands, and W is a weight vector.

2. Log-Likelihood Ratio (LLR): It is used to evaluate the distance between the converted and target speech from their linear prediction coefficients.

$$\text{LLR} = \frac{1}{N} \sum_{i=1}^N \log \left(\frac{\mathbf{a}_{y,i}^T \mathbf{R}_{x,i} \mathbf{a}_{y,i}}{\mathbf{a}_{x,i}^T \mathbf{R}_{x,i} \mathbf{a}_{x,i}} \right)$$

where \mathbf{a}_x , \mathbf{a}_y , and \mathbf{R}_x are the LPC vector of the target signal frame, converted signal frame, and the autocorrelation matrix of the target speech signal, respectively.

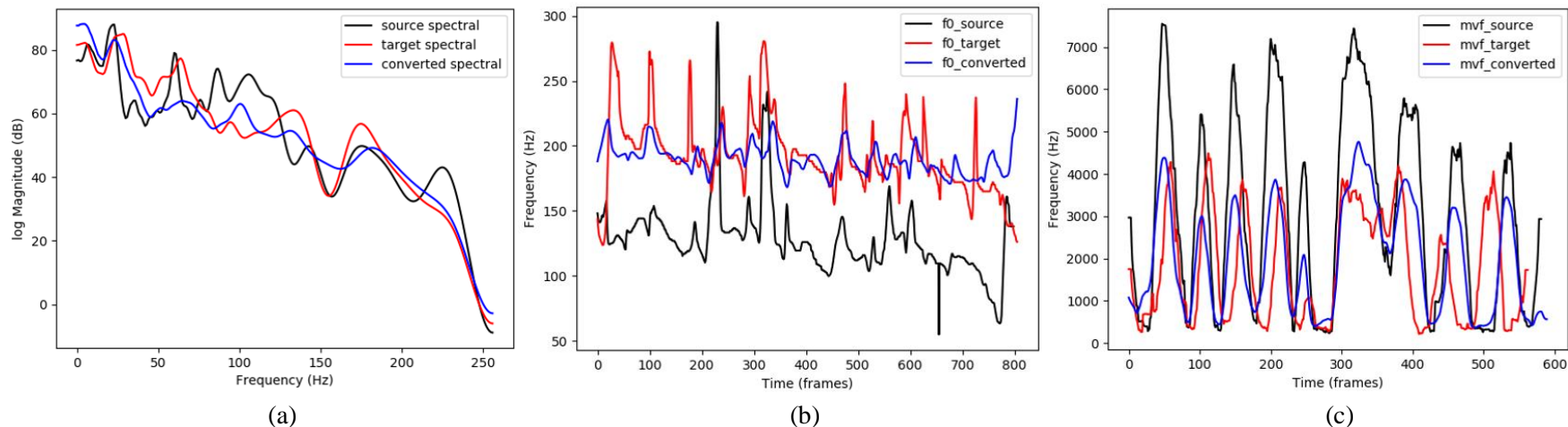
A) Objective evaluation

TABLE I. AVERAGE SCORES ON CONVERTED SPEECH SIGNAL PER EACH OF THE SPEAKER PAIRS CONVERSION

Model	WORLD		MagPhase		Sprocket		Proposed	
	fwSNRseg	LLR	fwSNRseg	LLR	fwSNRseg	LLR	fwSNRseg	LLR
BDL → JMK	2.19	1.57	3.21	1.37	2.20	1.48	2.47	1.50
BDL → SLT	1.12	1.72	1.25	1.69	1.04	1.49	2.33	1.57
BDL → CLB	0.79	1.83	1.65	1.72	0.37	1.69	1.66	1.74
JMK → BDL	1.31	1.76	2.49	1.56	1.73	1.63	2.15	1.57
JMK → SLT	0.55	1.74	1.93	1.56	0.11	1.64	1.54	1.65
JMK → CLB	1.45	1.74	1.75	1.66	0.69	1.60	1.81	1.67
SLT → BDL	1.65	1.71	1.60	1.70	1.80	1.51	2.95	1.49
SLT → JMK	2.16	1.61	2.71	1.42	0.713	1.56	2.59	1.39
SLT → CLB	1.51	1.75	2.89	1.59	2.32	1.56	2.51	1.50
CLB → BDL	0.97	1.81	1.60	1.70	0.95	1.72	1.92	1.60
CLB → JMK	2.50	1.49	2.74	1.40	0.98	1.46	3.00	1.30
CLB → SLT	0.98	1.70	2.17	1.53	1.96	1.54	2.12	1.47

- The proposed approach based sinusoidal model succeeded in the voice conversion.
- the CSM can convert voice characteristics more accurately than other methods when a female is a source speaker.

A) Objective evaluation



Example of the natural source (black), target (red), and converted (blue) spectral envelope, contF0, and MVF trajectories using the proposed method.

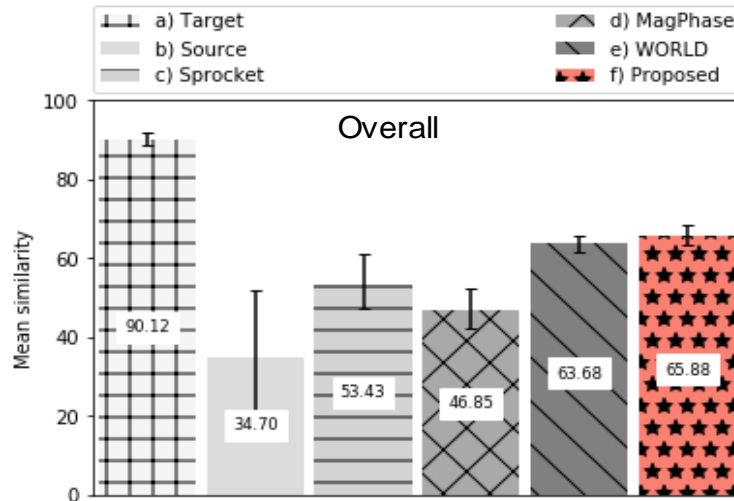
- The converted spectral envelope (a) is more similar in general to the target one than the source one.
- The converted contF0 trajectory (b) generated from the proposed method follow the same shape of the target confirming the similarity between them and can provide better F0 predictions.
- The proposed framework produces converted speech with MVF more similar to the target trajectories rather than to the source ones.

B) Subjective evaluation

- MUSHRA: enables evaluation of multiple samples in a single trial without breaking the task into many pairwise comparisons.
- reference: natural target speech
- anchor: pulse-noise excitation
- 72 utterances were included in the test (6 types x 12 sentences)
- 20 participants (11 males, 9 females).
- The test took 17 minutes to fill

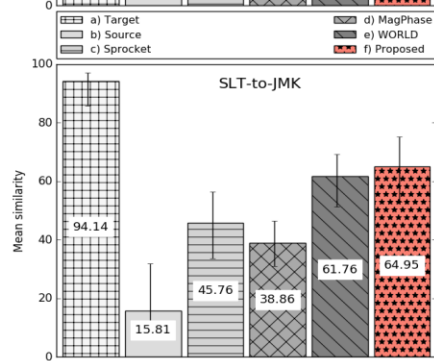
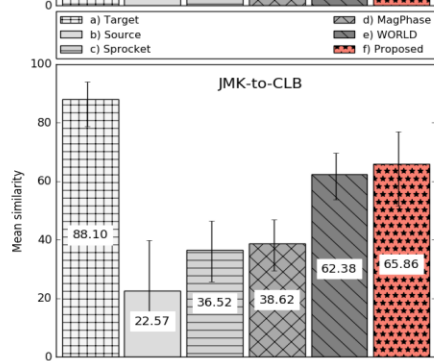
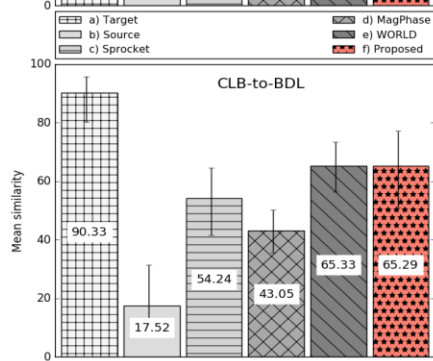
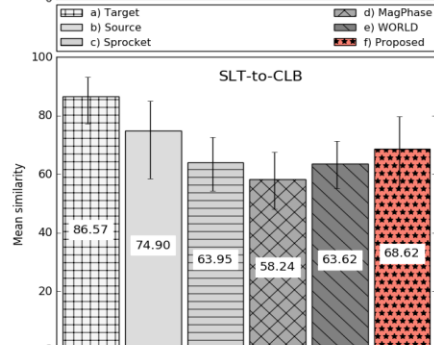
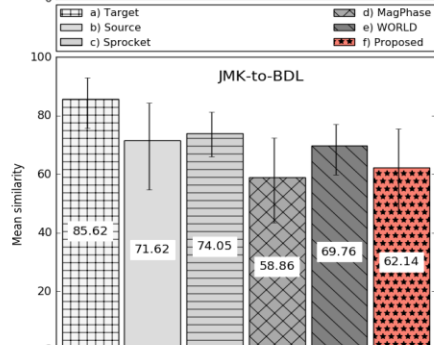
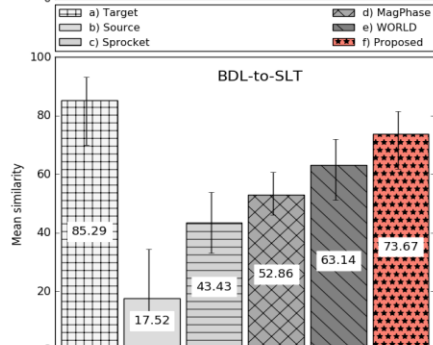
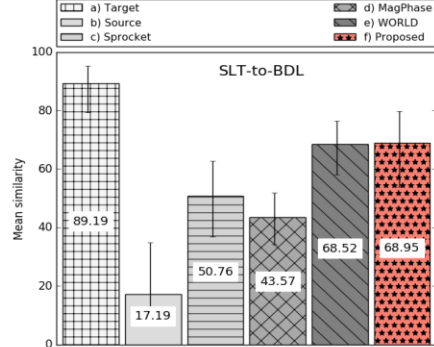
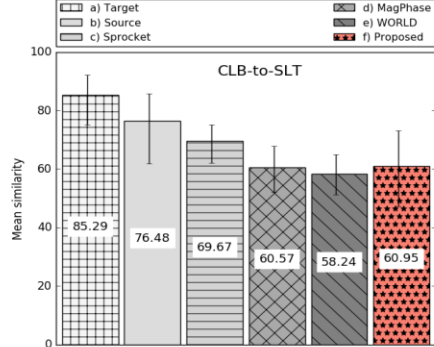
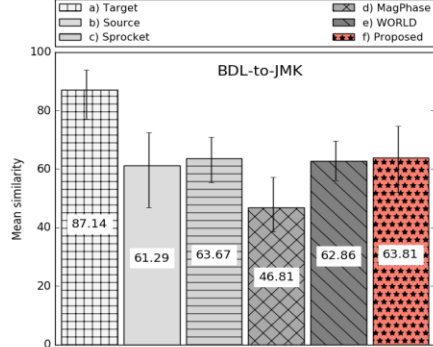
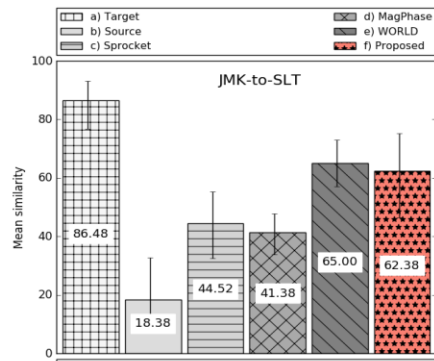
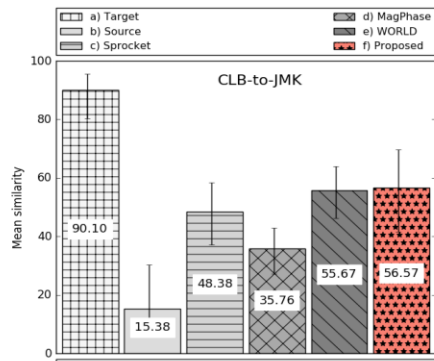
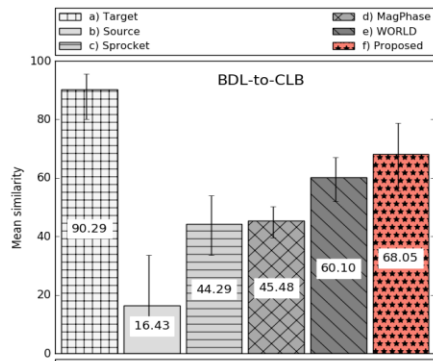
Online test samples: http://smartlab.tmit.bme.hu/sped2019_vc

B) Subjective evaluation



























Overall MUSHRA scores for the similarity question. Higher value means better overall quality. Errorbars show the bootstrapped 95% confidence intervals.

- Our proposed model has successfully converted the source voice to the target voice on the same-gender and cross-gender cases.
- The advantage of the CSM is that it gives the closest results to the target speaker in both objective and similarity tests compared to other approaches.



MUSHRA scores for the similarity question per each conversion process.

Sound samples

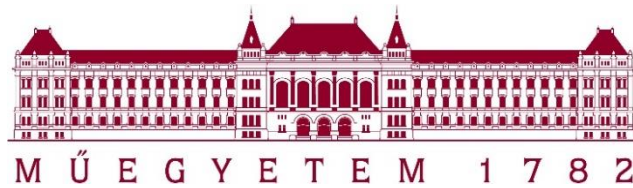
Conversion	Source	Target	Sprocket	MagPhase	WORLD	Proposed
Male1 – to – Male2 (bdl2jmk)						
Male1 – to – Female2 (bdl2slt)						
Female1 – to – Female2 (clb2slt)						
Female2 – to – Male1 (slt2bdl)						

Summary and Future plans

- ✓ The proposed method obtained higher speaker similarity compared to the conventional methods.
- ✓ Continuous sinusoidal vocoder has fewer parameters
 - computationally feasible
 - suitable for real-time operation
- ✓ Future works will aim at improving the quality scores through the use of bidirectional recurrent neural networks, in which many-to-one and one-to-many voice conversion can be achieved.

Key reference

- ❑ Garner, P. N., Cernak, M., and Motlicek, P., "A simple continuous pitch estimation algorithm," *IEEE Signal Processing Letters*, vol. 20, no. 1, pp. 102-105, 2013.
- ❑ Drugman, T., and Stylianou, Y., "Maximum Voiced Frequency Estimation : Exploiting Amplitude and Phase Spectra," *IEEE Signal Processing Letters*, vol. 21, no. 10, p. pp. 1230–1234, 2014.
- ❑ Morise, M., "CheapTrick, a spectral envelope estimator for high-quality speech synthesis," *Speech communication*, 67, pp. 1-7, 2015.
- ❑ Kominek, J., and Black, A.W., "CMU ARCTIC databases for speech synthesis," Carnegie Mellon University, 2003.
- ❑ Wu, Z., Watts, O., King, S., "Merlin: An Open Source Neural Network Speech Synthesis System" in *Proc. 9th ISCA Speech Synthesis Workshop (SSW9)*, September 2016, Sunnyvale, CA, USA.



Thank you for your attention !

nemeth@tmit.bme.hu