# GLOBAL RHYTHM STYLE TRANSFER WITHOUT TEXT TRANSCRIPTIONS

Kaizhi Qian; Yang Zhang; Shiyu Chang; Jinjun Xiong; Chuang Gan; David Cox; Mark Hasegawa-Johnson.

## AUDIOHEADS

**AUTOPST:** An Autoencoder-based Prosody Style Transfer framework with a thorough rhythm removal module guided by self-expressive representation learning.

It can disentangle global prosody style from speech without relying on any text transcriptions.

## INTRODUCTION

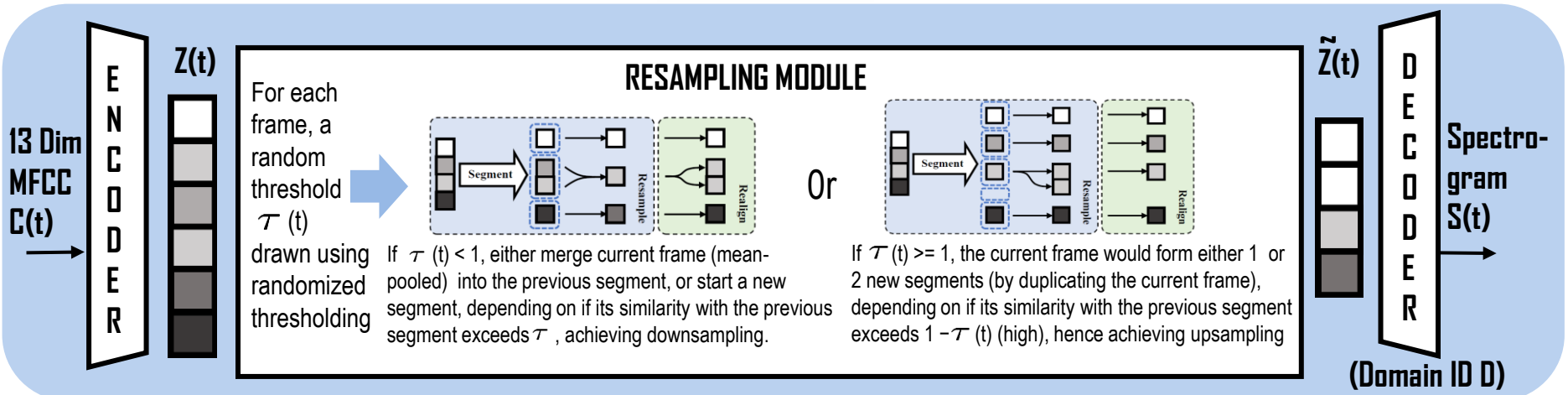
**Speech style transfer:** transferring the source speech into the style of the target domain, while keeping the content unchanged.

**Prosody = Pitch + Rhythm**

- Prosody plays an important role in characterizing the style/emotion of a speaker, but most non-parallel voice or emotion style transfer algorithms do not convert any prosody information.
- **Disentangling the prosody information, particularly the rhythm component, from speech is challenging as it involves breaking the synchrony between the input speech and the disentangled speech representation. (Different phones have different lengths when speech is slowed)**
  - Most existing prosody style transfer algorithms would need to rely on some form of text transcriptions to identify the content information, which confines their application to high-resource languages only.
  - Recently, SPEECHSPLIT (Qian et al., 2020b) has made progress towards unsupervised prosody style transfer, but it is unable to extract high-level global prosody style in an unsupervised manner.

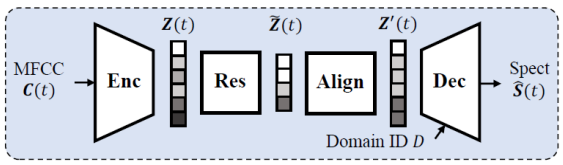
## AUTOPST

### ARCHITECTURE

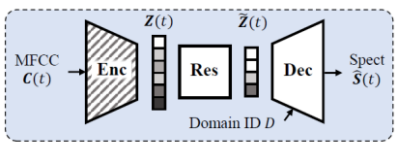



### TRAINING

2-stage training scheme to prevent encoder and decoder from communicating and leaking rhythm information in ways robust to temporal resampling.




(a) Synchronous training.



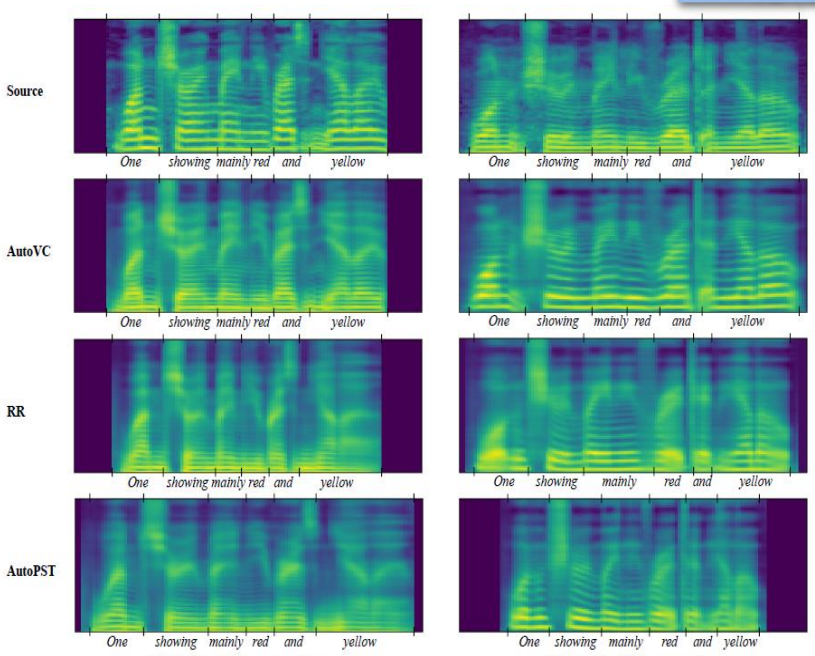

(b) Asynchronous training.

Downsampling: Copy each Z̃(m) to match the length of the original segment.
Upsampling: Delete the newly inserted Z̃ (m).
The network is then trained end-to-end to reconstruct the input with the realignment module. Since decoder has full access to the rhythm information, encoder will be trained to pass the content and not the rhythm information.

Remove Realignment module, fix encoder and only update decoder. We want to scramble rhythm information

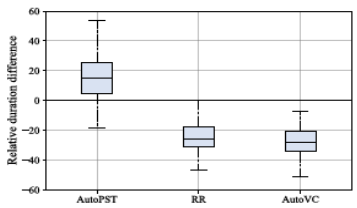
## EXPERIMENTS

### FAST TO SLOW AND SLOW TO FAST CONVERSION

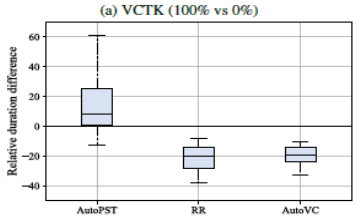



(a) Fast-to-slow conversion. (b) Slow-to-fast conversion.

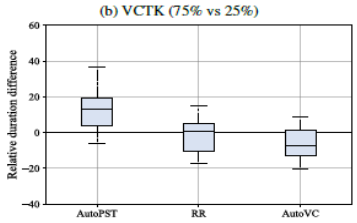
Mel spectrograms: AUTOPST is the only algorithm that can significantly change the rhythm to match the target speakers' styles. (RR uses SPEECHSPLIT for resampling)(Qian et. al. 2020)




(a) VCTK (100% vs 0%)



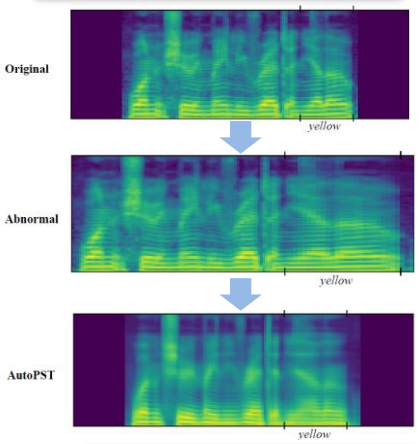

(b) VCTK (75% vs 25%)



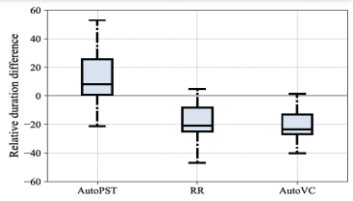

(c) VCTK (60% vs 40%)

Positive duration differences indicate more sufficient rhythm disentanglement. Percentiles indicate speech speed (fast to slow). Dataset VCTK (Veaux et.al. 2016)

### RESTORING ABNORMAL RHYTHM PATTERNS




### EMOTION STYLE TRANSFER (EMO V-DB)




Tested for amused, sad, sleepy, neutral, and angry emotions