Global Rhythm Style Transfer

Aakriti
Samiksha Das
Sumit Jain

Contents

Introduction

- Works on non-parallel speech style transfer
- Refers to transferring source speech into speech of target domain
  - In voice style transfer, domains correspond to speaker identities
- Non-parallel style transfer - when the source and target utterances do not need to have the same speech content

Introduction

- Speech has many layers of information
  - Content
  - Prosody
    - Rhythm
    - Pitch
- Prosody is an important aspect
- Prosody must be disentangled from the source utterance to apply the traits of the target utterance
- Disentangling the prosody information is very challenging

Motivation

Generating new voices for TTS (Text-To-Speech) systems

Dubbing in movies and videogames

Speech enhancement

Related Works

Prosody Disentanglement :

- Disentangle prosody from speech content by an auto-encoder based representation
- CHiV explicitly extracts prosodic features and linguistic features for expressive TTS
- Require text transcriptions which limits their applications to high-resource language
- Algorithms that do not rely on text transcriptions
  - Attempts to remove the rhythm information by randomly resampling input speech
  - SPEECHSPLIT relies on fine-grained prosody ground-truth in the target domain
- Prosody conversion not effective

Related Works

Voice Style Transfer :

- Directly learn speaker-independent content representations using a VAE
- ACVAE-VC encourages converted speech to be correctly classified as the target speaker by classifying the output
- Image style transfer approaches like CycleGAN and StarGAN adapted
- AUTOVC disentangles the timbre and content using a simple autoencoder
- Only focus on converting timbre, which is only one of the speech components

Problem to be Solved

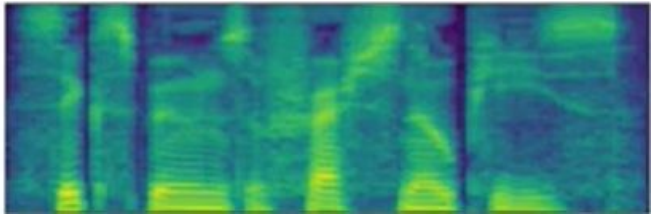
- Most algorithms require text transcriptions to identify content and separate out style
- Cannot be applied to low-resource languages with few text transcriptions
- Some attempts try to disentangle prosody in an unsupervised manner
  - Consists of an auto-encoder with a resampler to corrupt the rhythm
- SPEECHSPLIT : better disentanglement, but needs target ground-truth prosody info
- Prosody style transfer without relying on text transcriptions or local prosody ground truth largely remains unresolved in the research community

Problem to be Solved

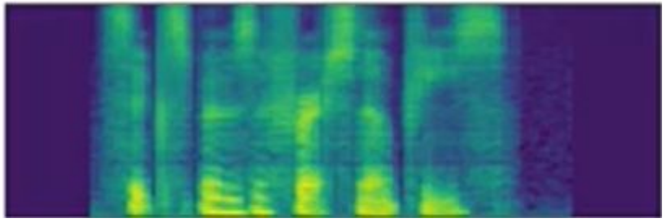
1. The autoVC algorithm doesn't change the rhythm of source speaker it only change the timbre to target source.
2. The two speaker have different speech rate, which is not reflected by autoVC alone.

Example:

Source Speech:



Target Example Speech:



AutoPST

- AUTOPST is an unsupervised speech decomposition algorithm that
    - does not require text annotations
    - can effectively convert prosody style given domain summaries
- Introduces a much more thorough rhythm removal module
- Adopts two-stage training strategy to pass full content without leaking rhythm
- Experiments on different style transfer tasks show that AUTOPST can effectively convert prosody that correctly reflects the styles of the target domains

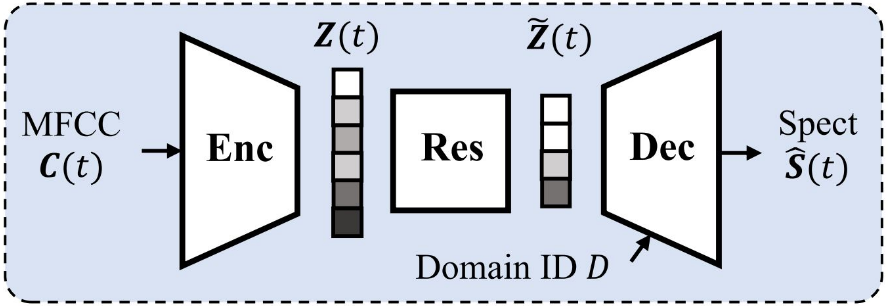
How autoPST changes Speech rate and pauses ?

1. Our goal is to retain phonetic sequence and obscure repetition information
2. The autoPST contains a hidden module called resampling module which obscure rhythm by resampling so that the decoder couldn’t guess original repetition of sequence

Framework Overview

- AUTOPST adopts an autoencoder based structure
- 13-dimension MFCC is taken by encoder (ENC) having very little pitch information
- Novel resampling module (downsampling/upsampling) to disentangle rhythm from source
- Decoder aims to reconstruct speech based on random resampling module output and the domain identifiers (pitch and rhythm of domain)

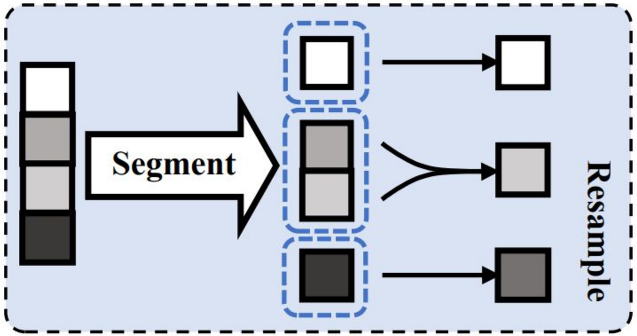
$$\boldsymbol{Z}(t) = \mathrm{Enc}(\boldsymbol{C}(t)),\ \tilde{\boldsymbol{Z}}(t) = \mathrm{Res}(\boldsymbol{Z}(t))$$
$$\hat{\boldsymbol{X}}(t) = \mathrm{Dec}(\tilde{\boldsymbol{Z}}(t), D) \longleftrightarrow \boldsymbol{X}(t)$$



Similarity based downsampling

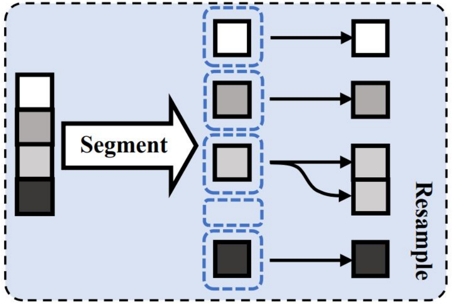
1. Based on observation that relatively steady segments in speech have more flexible durations
2. Uses a self-expressive autoencoder to derive frame-level representations with high cosine similarity between similar frames
3. Similarity threshold **τ** - if similarity between two frames is less than τ, segment boundary is added
4. Later all frames within two segment boundaries are merged to one code with mean pooling

$$\mathsf{G}(t, t') = \frac{\boldsymbol{A}^T(t)\boldsymbol{A}(t')}{\|\boldsymbol{A}(t)\|_2 \|\boldsymbol{A}(t')\|_2}.$$



Similarity based upsampling

1. If τ < 1, we perform the aforementioned downsampling
2. If 1 ≤ τ < 2, we create a boundary
3. If similarity between two frames is high enough we insert code of previous frame
4. As a result, some part of sequence is upsampled
5. At the end of resampling, frames with the most similarity between them are stretched or collapsed the most - rhythm is scrambled



Method: Thresholding ($\boldsymbol{\tau}$)

- Trade-off between rhythm disentanglement and information loss

- Randomized thresholding - to keep all content and forget all rhythm

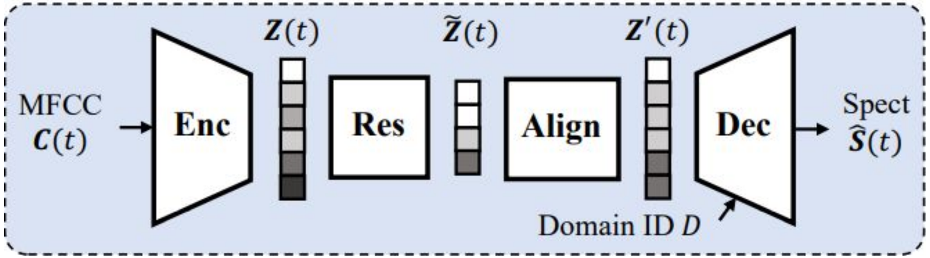
- Double randomized thresholding:
    - Randomly draw global variable G from U[u_l, u_r] that is shared across the whole utterance
    - Local variable L(t) from U[G - 0.05, G + 0.05]

$$\tau(t) = L(t)\text{-quantile}\Big[\mathsf{G}(t_m, t_m - b : t_m + b)\Big]$$

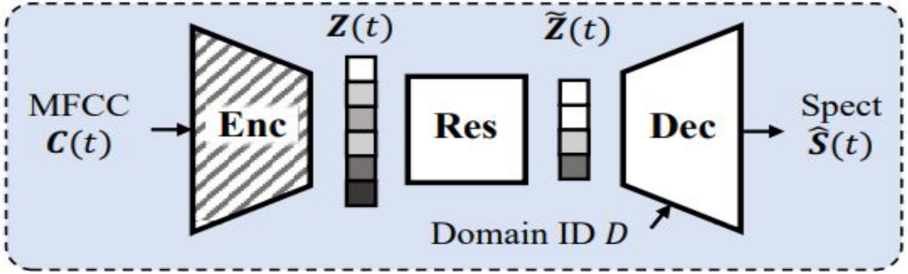
- Quantile : q-quantile, b: sliding window in which threshold is computed, G: similarity function

Training strategy

- Introduce a two-stage training scheme to prevent rhythm information leaking
- Stage 1 : Synchronous Training



- Stage 2 : Asynchronous Training



Architecture and Results

Architecture:

- Encoder
  - 1*5 conv layers with group normalization
  - Output dimension - 4
- Decoder
  - Transformer with 4 encoder and 4 decoder layers
- Spectrogram Conversion to Wavelength
  - WaveNet Vocoder:

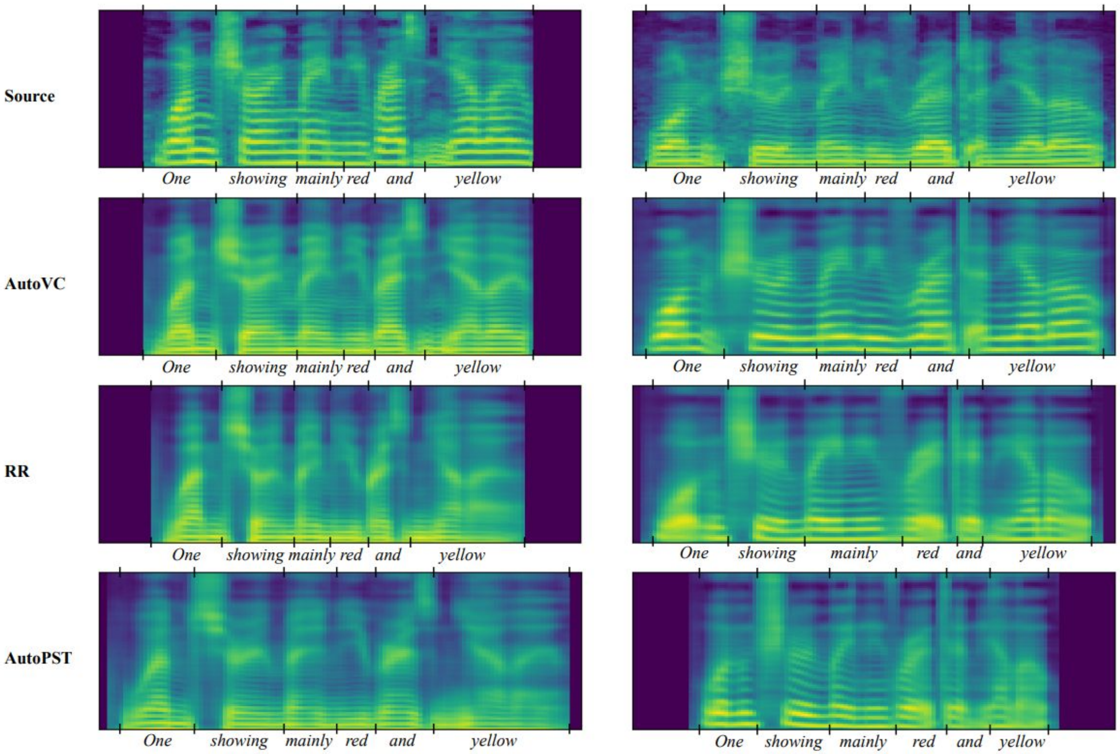
Dataset (VCTK):

- 44 hours of speech, 109 speakers

Baselines:

- RR
- AutoVC

Spectrogram Visualisation:

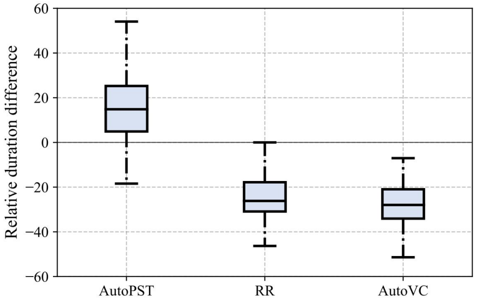


AutoVC and AutoPST: Result Comparison

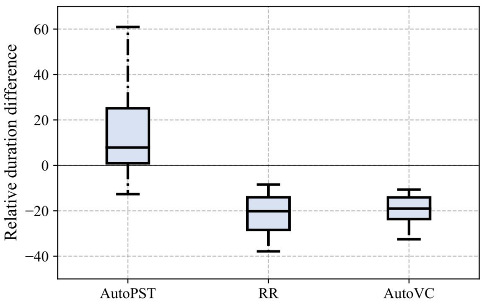
Source speech	
Target speech	
AutoVC output	
AutoPST	

More Experiments

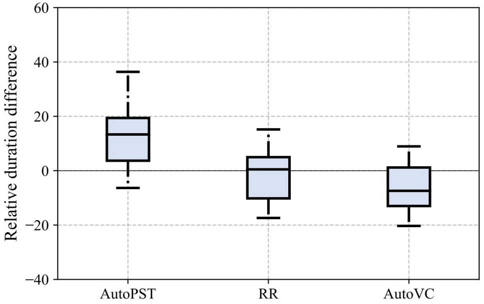
Relative Duration Difference = $(L_{F2S} - L_{S2F})/L_{S2F}$



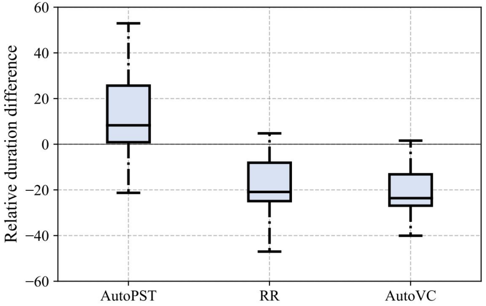
(a) VCTK (100% vs 0%)



(b) VCTK (75% vs 25%)



(c) VCTK (60% vs 40%)



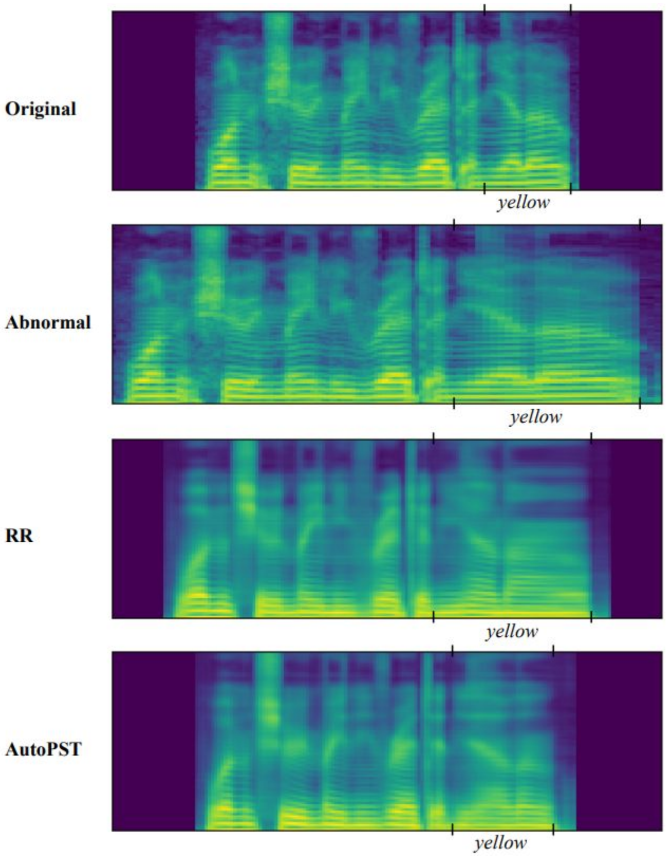
(d) Emo-VDB

More Experiments

- Subjective Evaluation:

	AUTOPST	RR	AUTOVC
Timbre	**4.29 ± 0.032**	4.07 ± 0.037	4.26 ± 0.034
Prosody	**3.61 ± 0.053**	2.97 ± 0.063	2.64 ± 0.066
Overall	**3.99 ± 0.036**	3.63 ± 0.045	3.49 ± 0.052

- Can AutoPST restore abnormal localised rhythm patterns? (right)



Conclusion

- AutoPST performs non-parallel voice style transfer and succeeds at transferring prosody characteristics
- Successfully transfers the rhythm aspect of prosody

Limitations:

- Severe limitations on the dimensions of the hidden representation, compromising the quality of the converted speech
- Performs poorly on in-the-wild examples