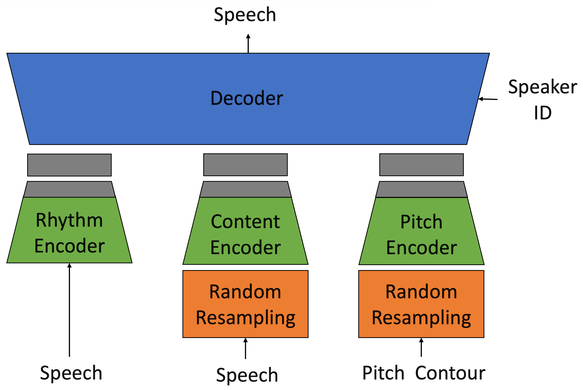
Disentangle speech for multi-aspect voice conversion by combining autoencoder and signal processing methods

Chak Ho Chan[1], Kaizhi Qian[2], Yang Zhang[2], Mark Hasegawa-Johnson[1]
[1]University of Illinois at Urbana-Champaign
[2]MIT-IBM Watson AI Lab

Motivation

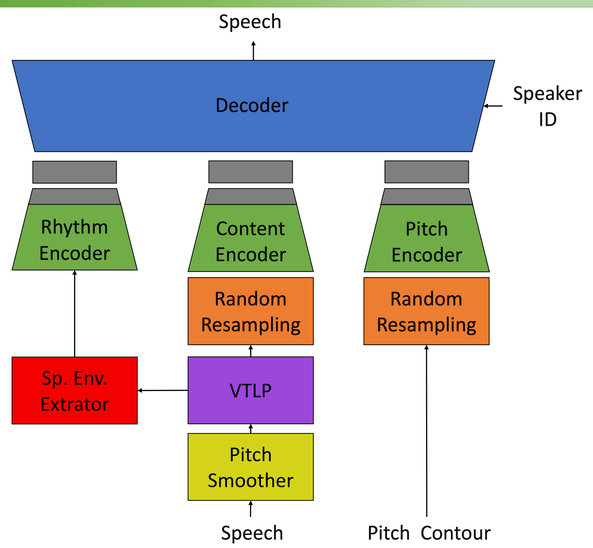
- Speech can be decomposed into rhythm, content, pitch, and timbre
- Existing voice conversion systems:
  - Focus on timbre-only conversion
  - Converting other aspects is under-explored

SpeechSplit



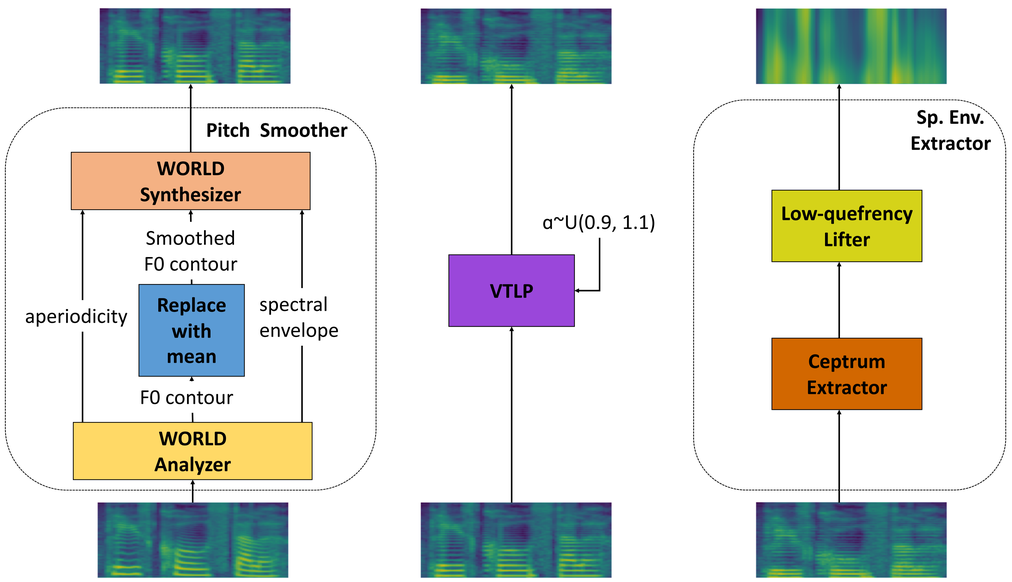
- Two differences among encoders
  - Inputs to the encoders
  - Random Resampling: corrupt rhythm
- Why does it work?
  - Information is corrupted in different inputs
  - Only encode one aspect if bottleneck is binding
- Limitation: exhaustive bottleneck tuning

SpeechSplit 2.0



- **Basic idea**: corrupt information in the inputs so that each encoder can only access full information of one aspect
- Corrupt pitch with Pitch Smoother
  - Extract spectral envelope, F0 contour, and aperiodicy with WORLD
  - Resynthesize speech with the smoothed F0 contour
- Corrupt timbre with VTLP
  - Change timbre by warping the frequency
- Corrupt content with Spectral Envelope Extractor
  - Discard fine-grained details
  - Preserve unique patterns for different phonemes

Code

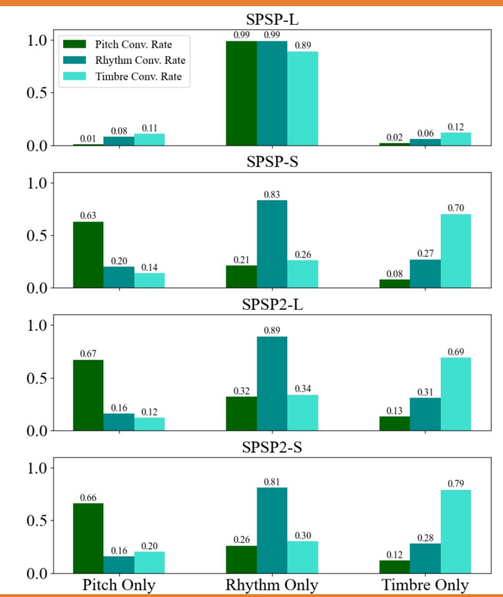


Evaluation

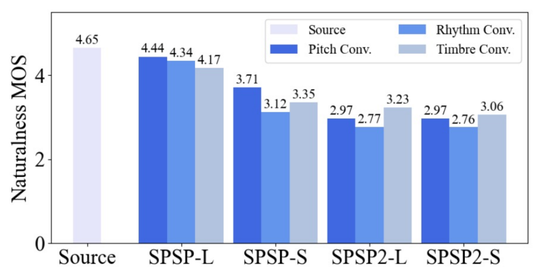
	Just Right	Very Large
SpeechSplit	SPSP-S	SPSP-L
SpeechSplit2.0	SPSP2-S	SPSP-L

Results

Conversion Capability



Speech Quality



Model	Pitch Conv.	Rhythm Conv.	Timbre Conv.
SPSP-L	12.9%	14.4%	9.8%
SPSP-S	30.8%	46.3%	34.0%
SPSP2-L	37.8%	54.5%	39.2%
SPSP2-S	54.4%	62.6%	43.5%