Unsupervised Speech Decomposition Via Triple Information Bottleneck

Kaizhi Qian[1, 2*], Yang Zhang[1*], Shiyu Chang[1],
David Cox[1], Mark Hasegawa-Johnson[2]

1 MIT-IBM Watson AI Lab 2 University of Illinois at Urbana-Champaign

What information does speech contain?

What information does speech contain?

Content	Can be transcribed to text in most languages

What information does speech contain?

Content	Can be transcribed to text in most languages
Timbre	Characterizes the speaker's voice

What information does speech contain?

Content	Can be transcribed to text in most languages
Timbre	Characterizes the speaker's voice
Pitch	Characterizes the intonation

What information does speech contain?

Content	Can be transcribed to text in most languages
Timbre	Characterizes the speaker's voice
Pitch	Characterizes the intonation
Rhythm	Characterizes how fast the speaker speaks

Disentangling Speech Components

Disentangling Speech Components

Disentangling Speech Components

Speech Analysis

Synthesis

Voice Conversion

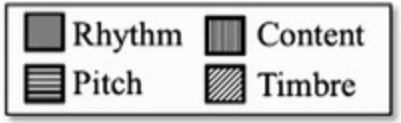
Speech Coding

Voice Conversion

Voice Conversion

Voice Conversion

Voice Conversion

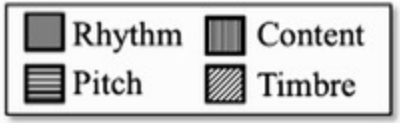


AutoVC: (Qian et al 2017)

Source: Target: Converted:

Qian, Kaizhi, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. "AutoVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss." In *International Conference on Machine Learning*, pp. 5210-5219. 2019.

Voice Conversion



AutoVC: (Qian et al 2017)

Source: Target: Converted:

Only disentangles timbre; other aspects remain intact
Few works can disentangle the remaining aspects

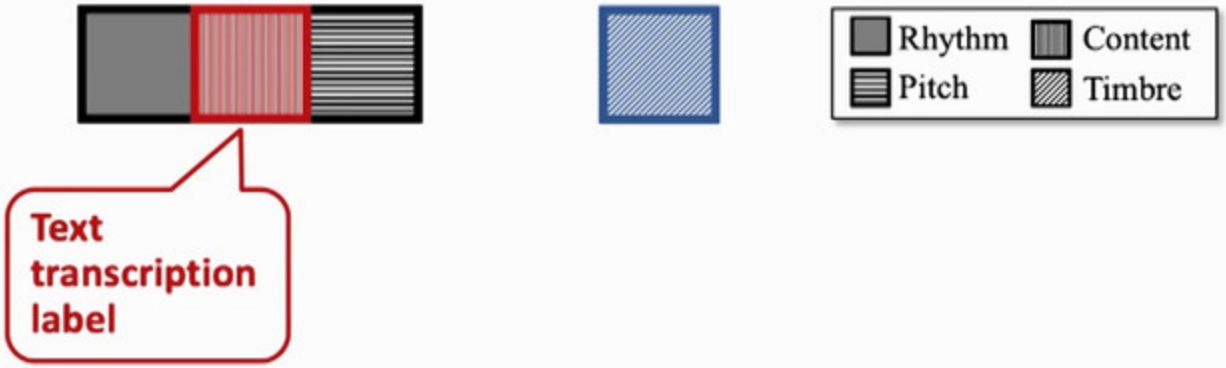
Qian, Kaizhi, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. "AutoVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss." In *International Conference on Machine Learning*, pp. 5210-5219. 2019.

Disentangling Other Aspects

Disentangling Other Aspects

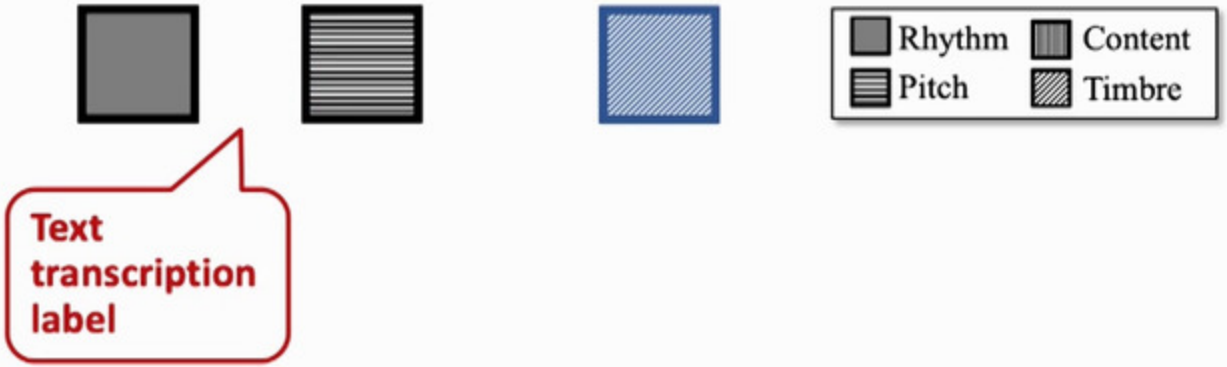
Requires text transcription labels;

Disentangling Other Aspects



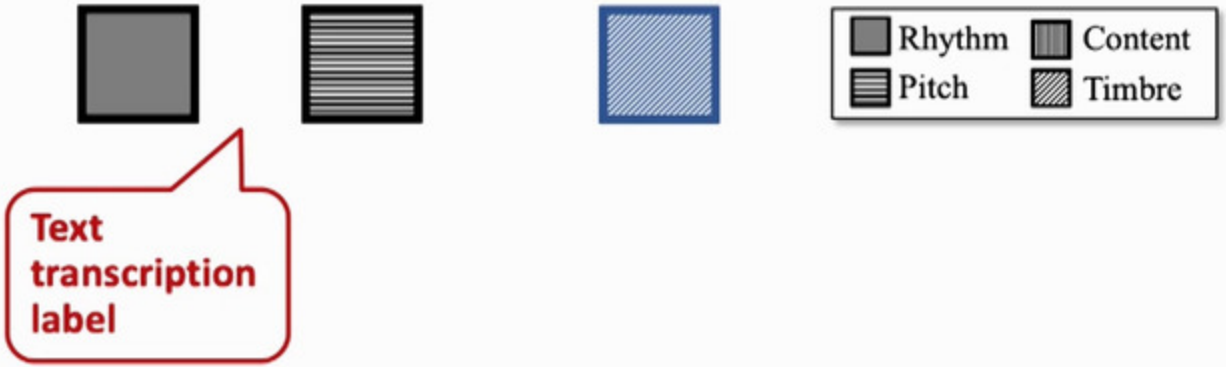
Requires text transcription labels;

Disentangling Other Aspects



Requires text transcription labels;

Disentangling Other Aspects



Requires text transcription labels;
Text transcriptions hard to obtain for low-resource languages.

Is it possible to disentangle speech without the text transcriptions?

Is it possible to disentangle speech without the text transcriptions?

Is it possible to disentangle speech without the text transcriptions?

- Yes!
- We propose **SpeechSplit**, which can decompose speech into content, timbre, pitch and rhythm, without relying on text transcriptions

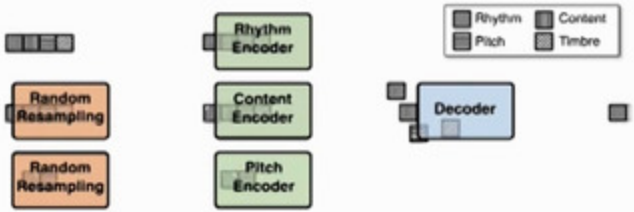
ICML
International Conference
On Machine Learning

Video is Playing

auspicious3000.github.io/SpeechSplit-Demo/

Unsupervised Speech Decomposition Via Triple Information Bottleneck: Audio Demo

This demo webpage provides sound examples for **SpeechSplit**, an autoencoder that can decompose speech into content, timbre, rhythm and pitch. The following GIF illustrates the working mechanism of SpeechSplit.

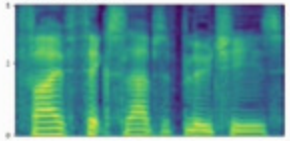


Below are audio clips of aspect-specific conversion, which corresponds to **section 5.2** in the paper. It is an iteractive experience. For each source-target pair, you can:

- **Select the aspect(s) you wish to convert**, and then the corresponding converted speech will automatically load.
- **Flip the source and target speech** by toggling the switch.

Utterance 1

Source Speech



0:02 / 0:02

Target Speech

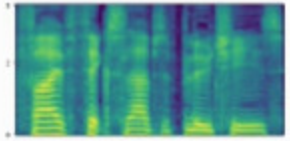
recorder.slideslive.com is sharing your screen. Stop sharing Hide

Below are audio clips of aspect-specific conversion, which corresponds to **section 5.2** in the paper. It is an iteractive experience. For each source-target pair, you can:

- **Select the aspect(s) you wish to convert**, and then the corresponding converted speech will automatically load.
- **Flip the source and target speech** by toggling the switch.

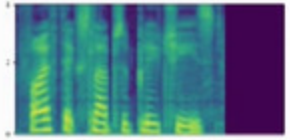
Utterance 1

Source Speech



0:02 / 0:02

Target Speech

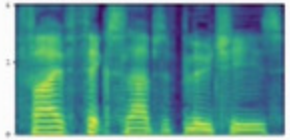


0:01 / 0:01

Choose What To Convert:

☐ Rhythm ☐ Pitch ☐ Timbre Flip source and target

Converted Speech



0:00 / 0:02

Utterance 2

Source Speech

recorder.slideslive.com is sharing your screen. Stop sharing Hide

Is it possible to disentangle speech without the text transcriptions?

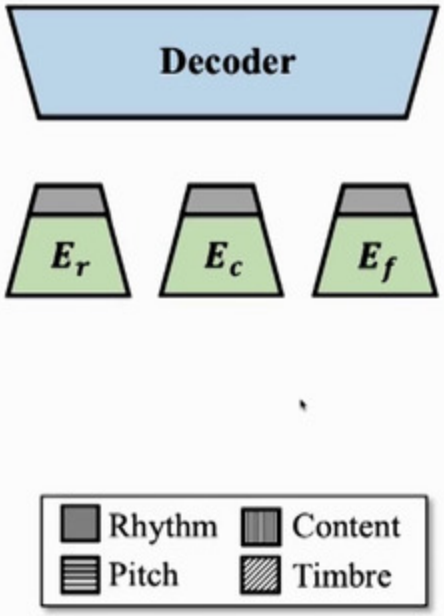
- Yes!
- We propose **SpeechSplit**, which can decompose speech into content, timbre, pitch and rhythm, without relying on text transcriptions

SpeechSplit Demo

https://auspicious3000.github.io/SpeechSplit-Demo/

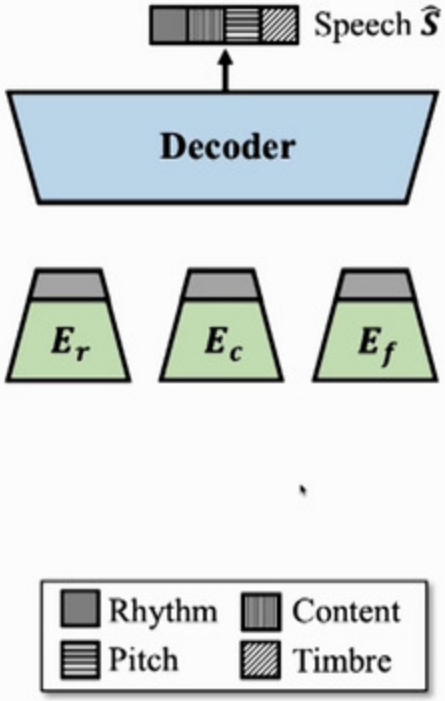
SpeechSplit

- **Autoencoder structure**



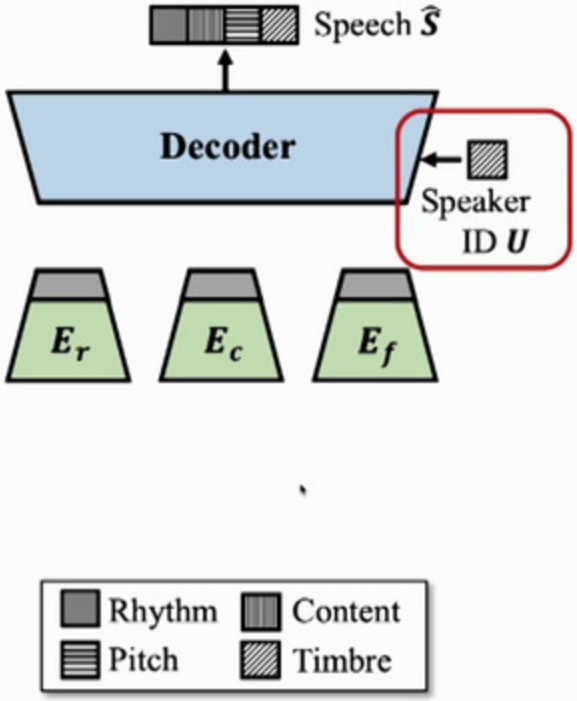
SpeechSplit

- **Autoencoder structure**
- Decoder reconstructs speech; it also takes speaker ID



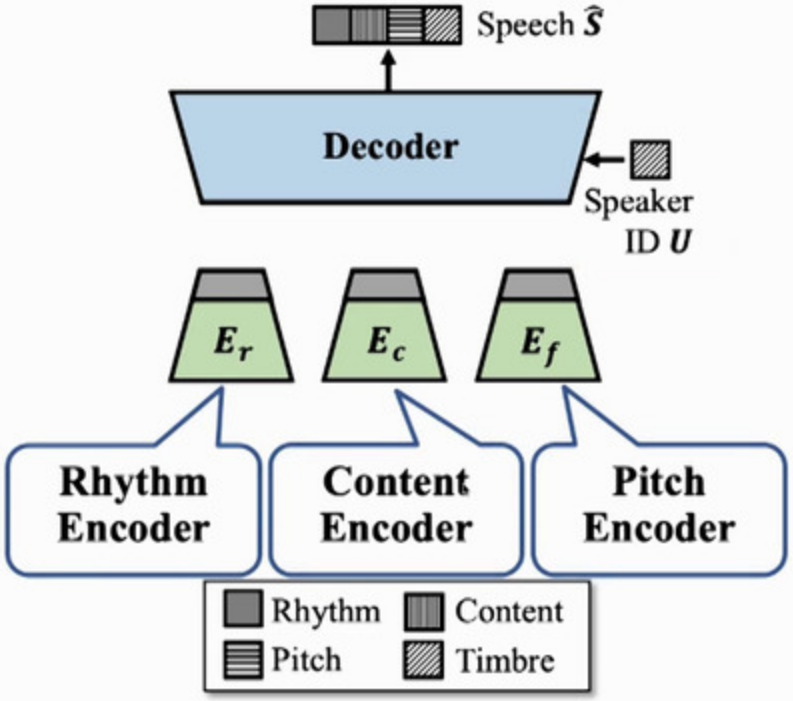
SpeechSplit

- **Autoencoder structure**
- Decoder reconstructs speech; it also takes speaker ID



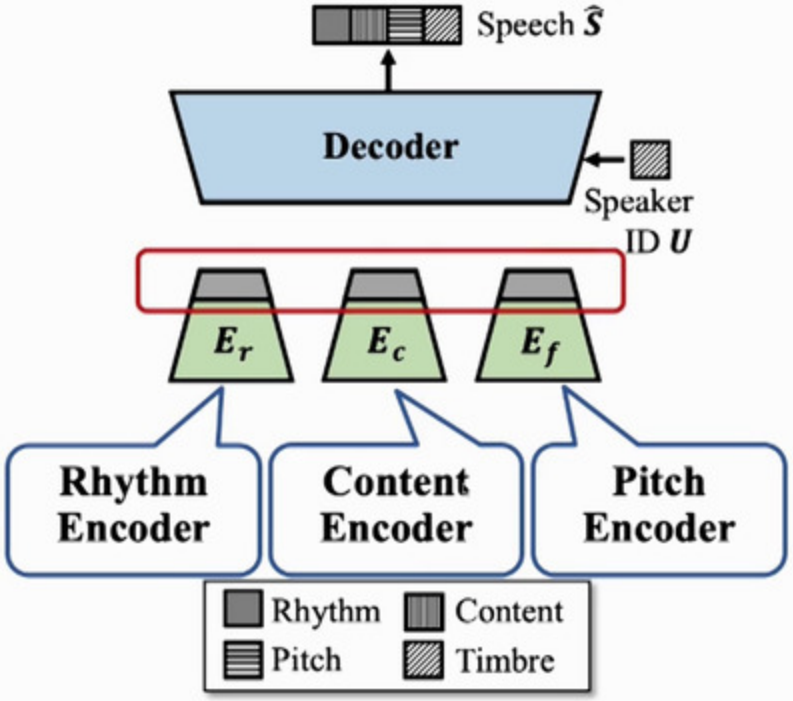
SpeechSplit

- **Autoencoder structure**
- Decoder reconstructs speech; it also takes speaker ID

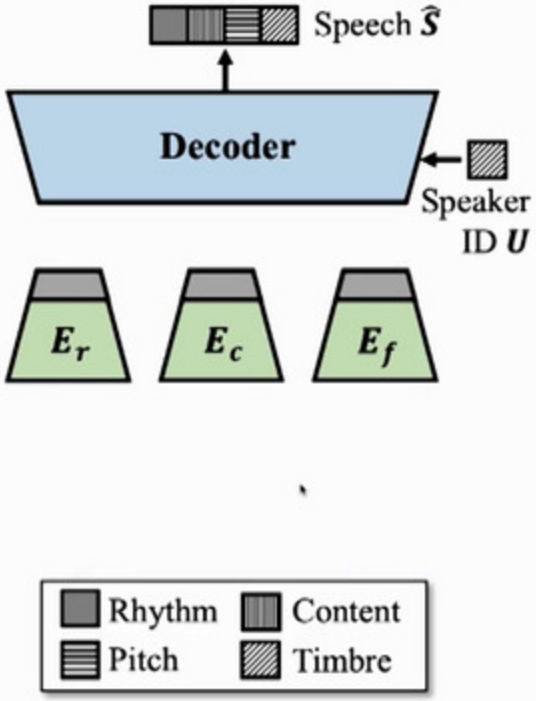


SpeechSplit

- **Autoencoder structure**
- Decoder reconstructs speech; it also takes speaker ID
- Encoder has a stringent physical bottleneck dimension

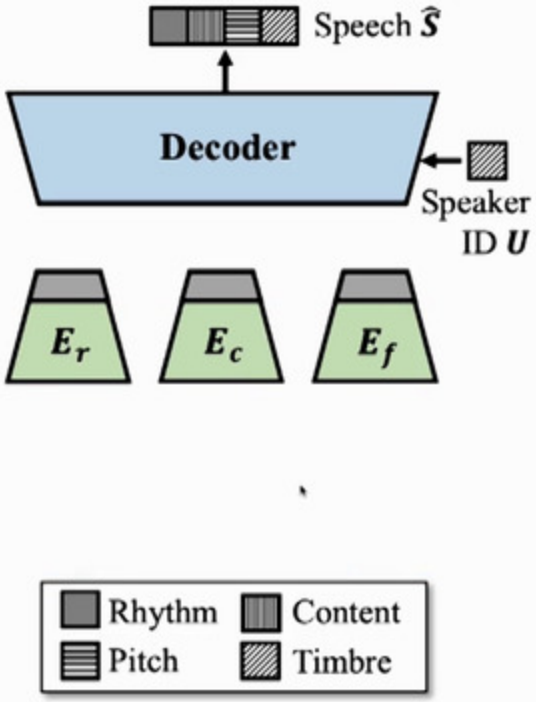


SpeechSplit



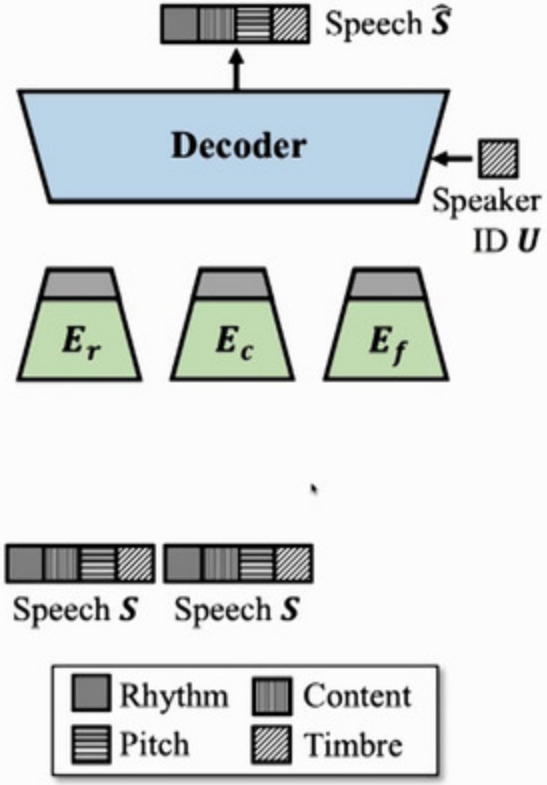
SpeechSplit

- **Two differences among encoders**



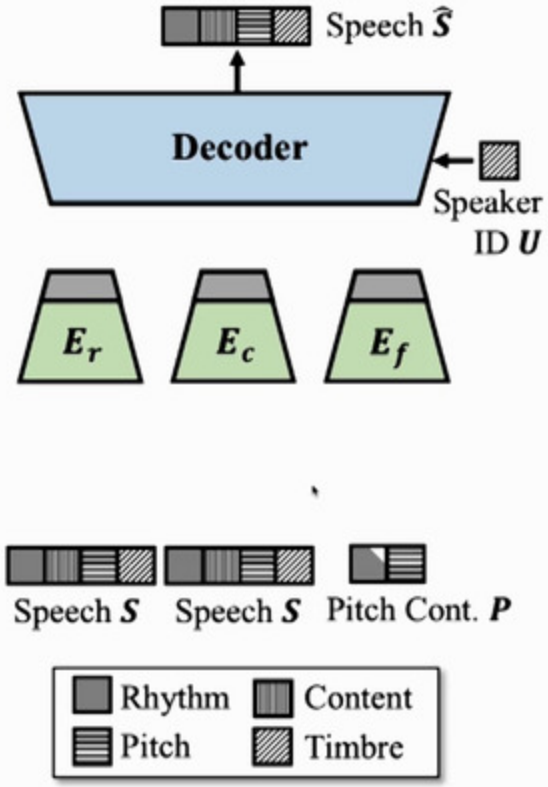
SpeechSplit

- **Two differences among encoders**
  - Input



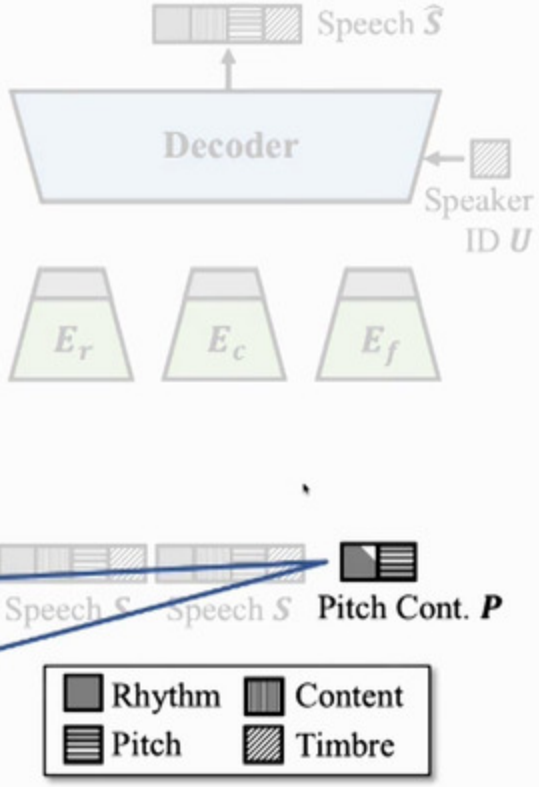
SpeechSplit

- **Two differences among encoders**
  - Input



SpeechSplit

- Two differences among encoders
  - Input

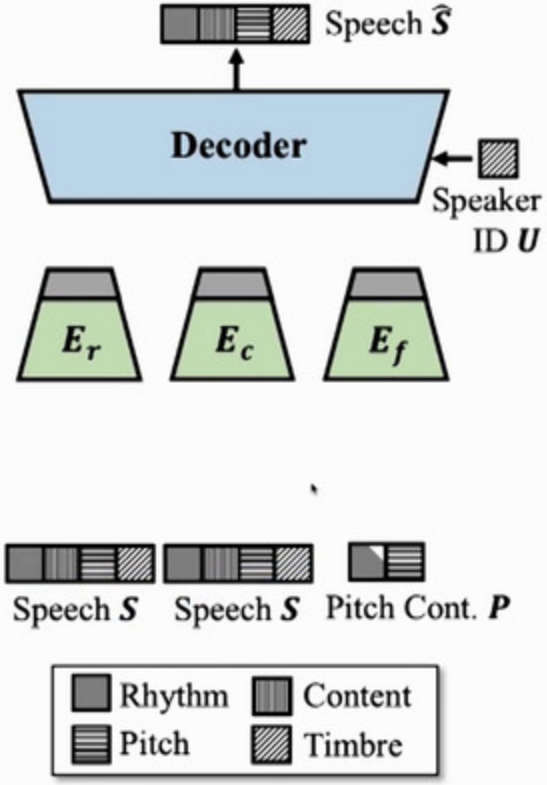


Pitch Contour:

- Pitch frequency against time
- Contains pitch information and parts of the rhythm information

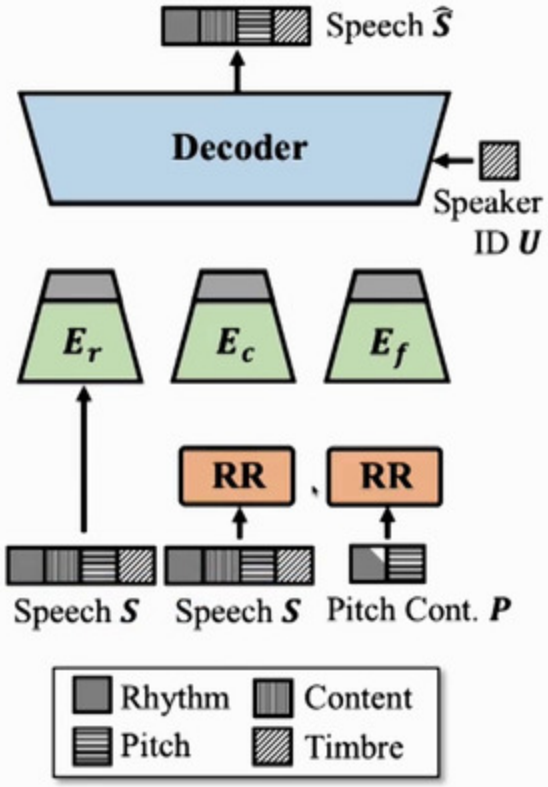
SpeechSplit

- **Two differences among encoders**
  - Input



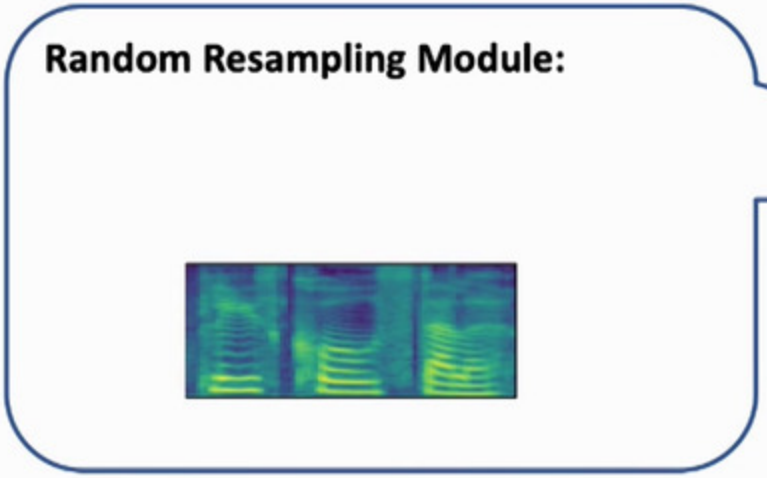
SpeechSplit

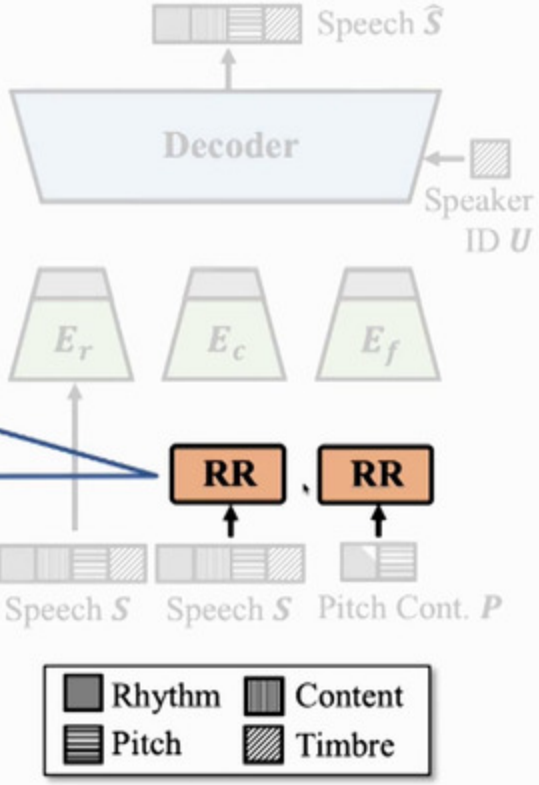
- **Two differences among encoders**
  - Input
  - Random Resampling Module



SpeechSplit

- Two differences among encoders
  - Input
  - Random Resampling Module



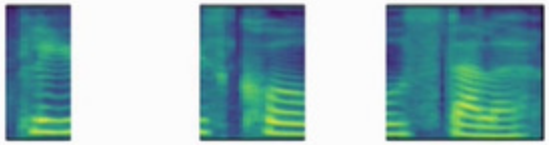


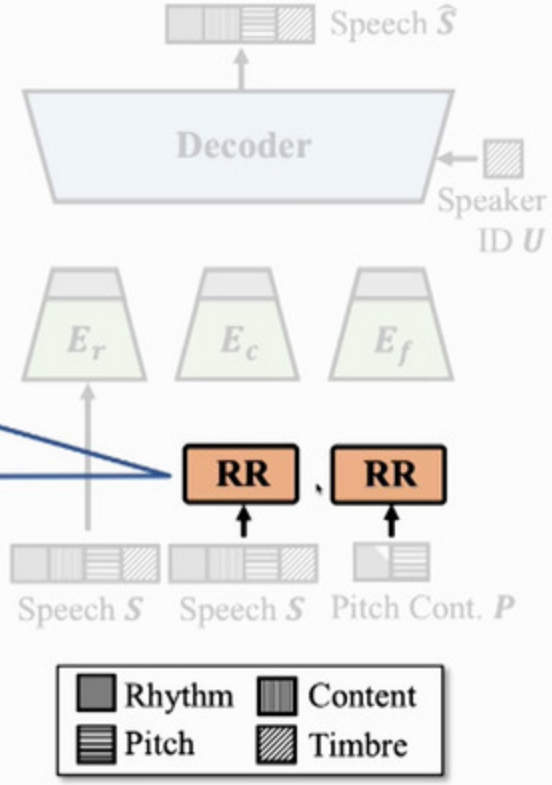
SpeechSplit

- Two differences among encoders
  - Input
  - Random Resampling Module

Random Resampling Module:

- Randomly split into segments



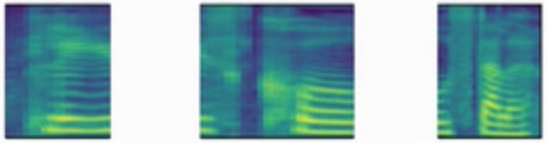


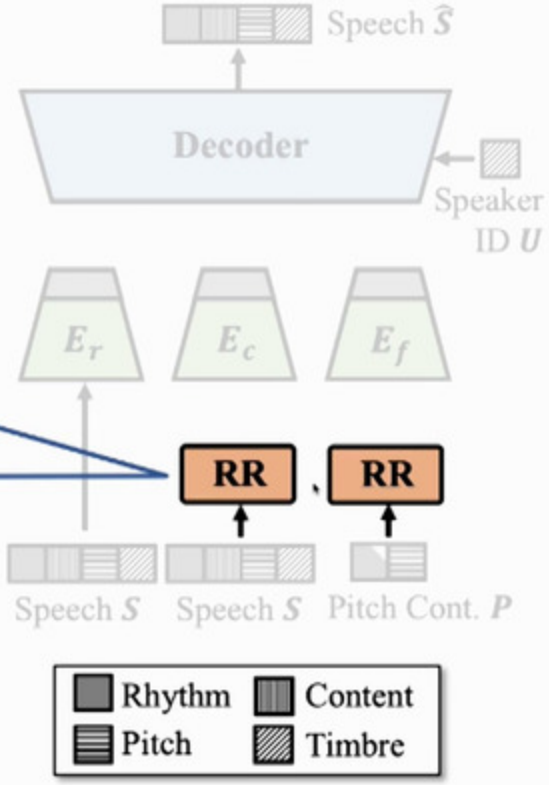
SpeechSplit

- Two differences among encoders
  - Input
  - Random Resampling Module

Random Resampling Module:

- Randomly split into segments
- Randomly stretch or squeeze



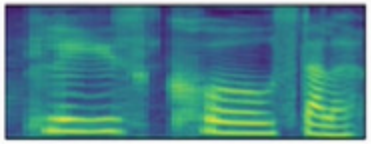


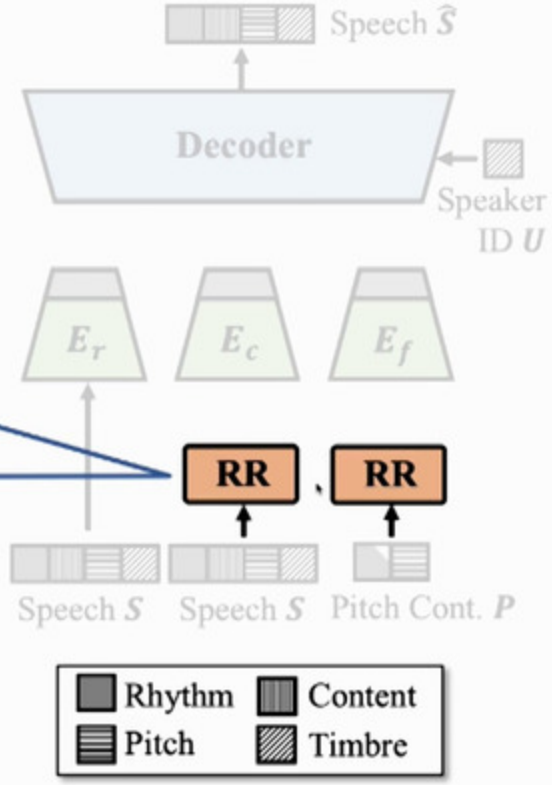
SpeechSplit

- Two differences among encoders
  - Input
  - Random Resampling Module

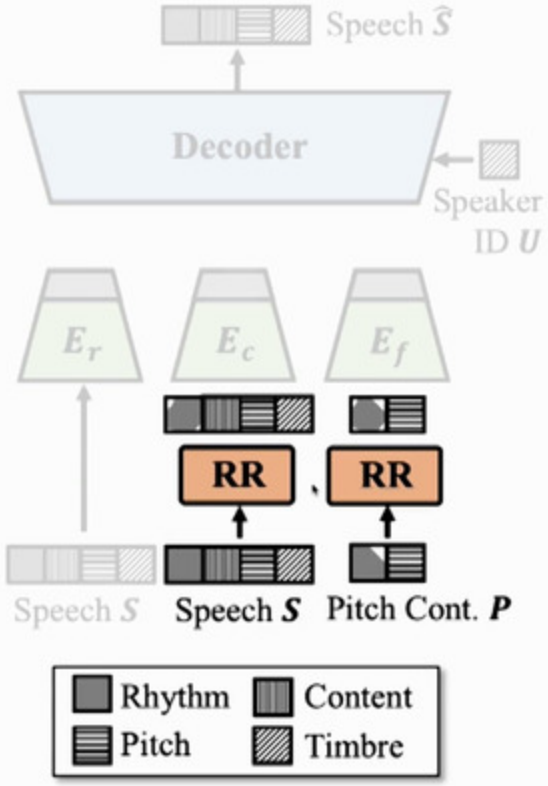
Random Resampling Module:

- Randomly split into segments
- Randomly stretch or squeeze





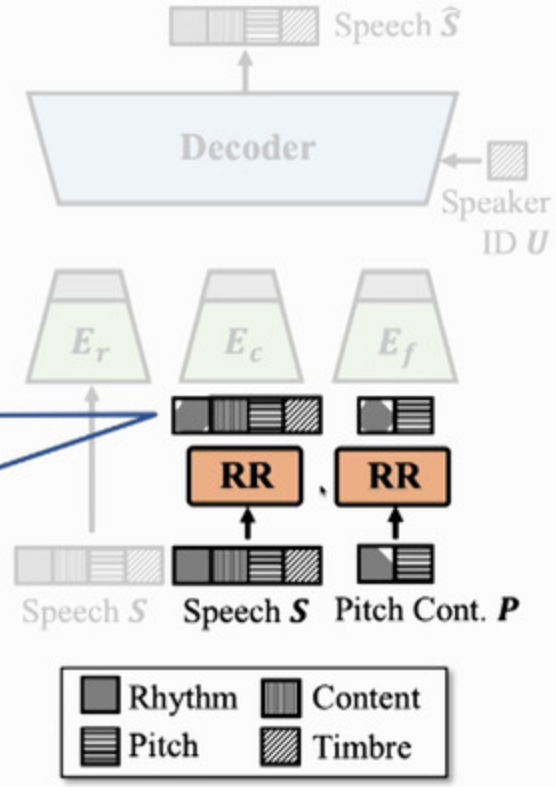
SpeechSplit



SpeechSplit

Random Resampling:

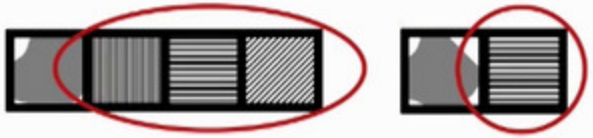
- Partially & randomly contaminate rhythm information

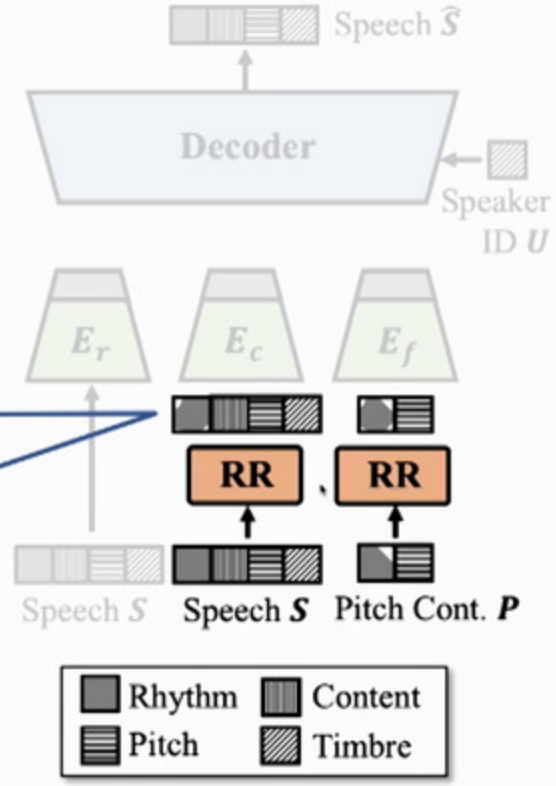


SpeechSplit

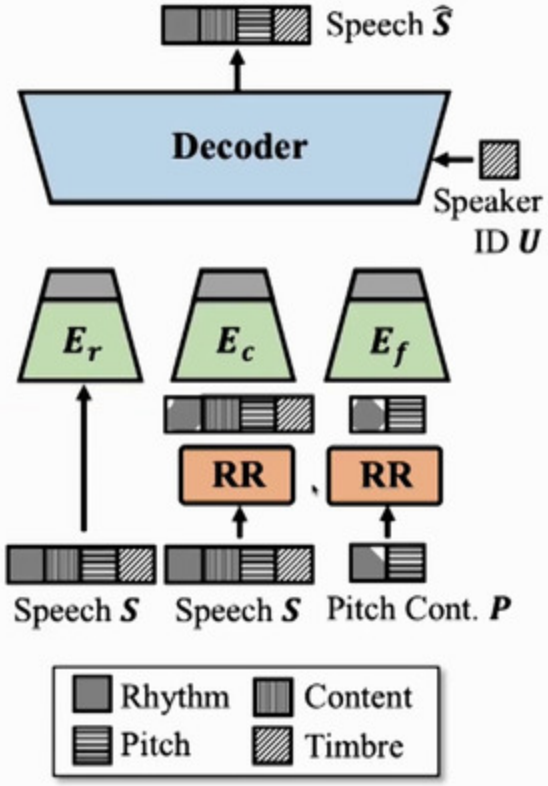
Random Resampling:

- Partially & randomly contaminate rhythm information
- Other information intact



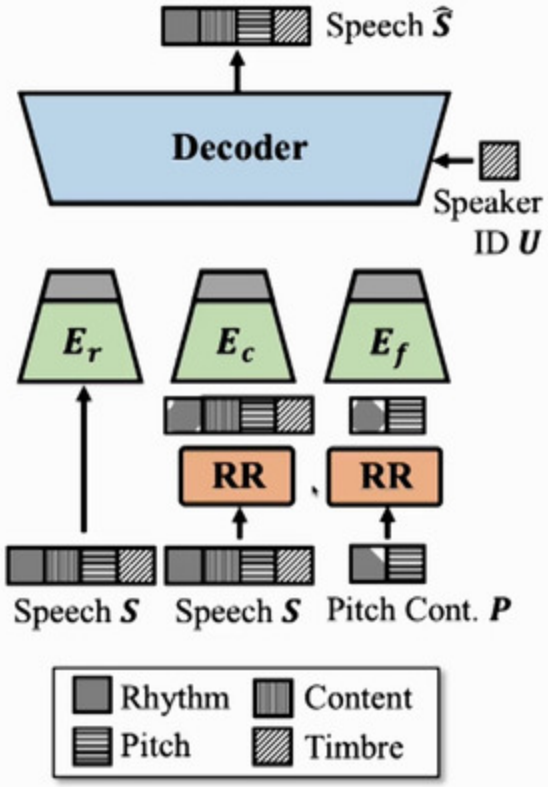


SpeechSplit – Why does it work?



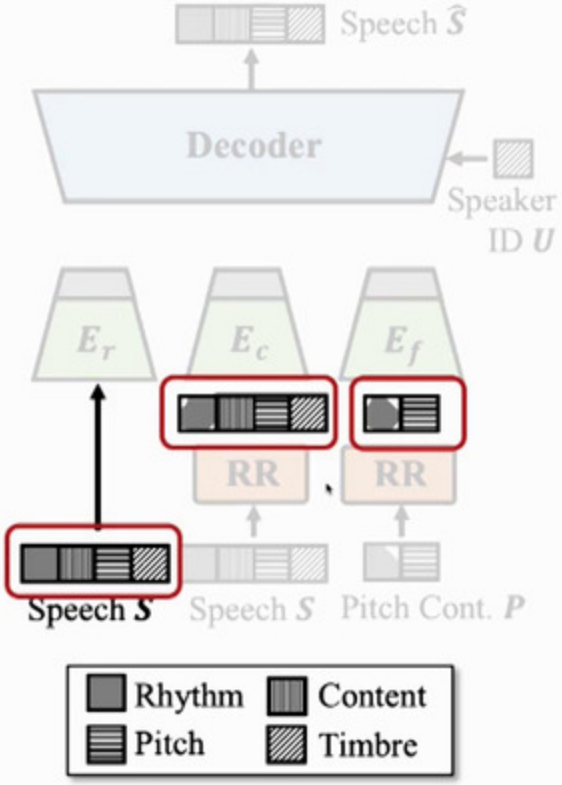
SpeechSplit – Why does it work?

Basic idea: The encoders need to pass all the blocks for perfect reconstruction, subject to the bottleneck constraints



SpeechSplit – Why does it work?

Basic idea: The encoders need to pass all the blocks for perfect reconstruction, subject to the bottleneck constraints

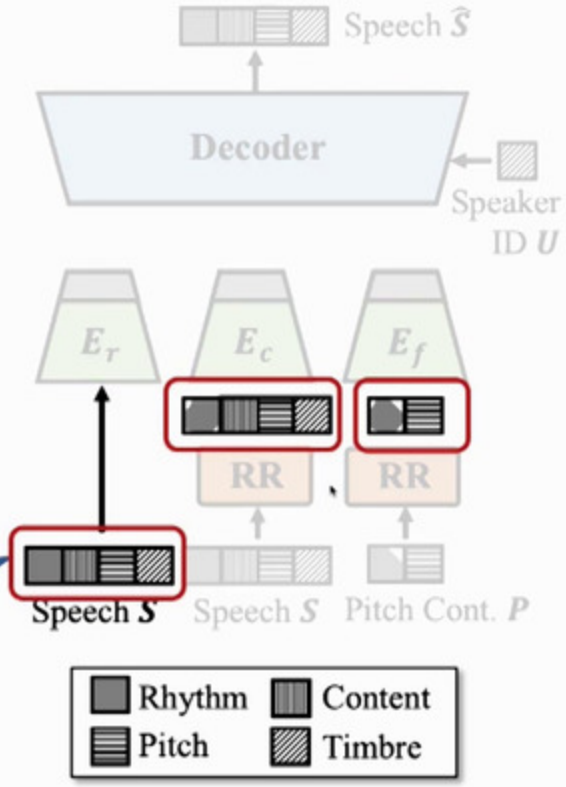


SpeechSplit – Why does it work?

Basic idea: The encoders need to pass all the blocks for perfect reconstruction, subject to the bottleneck constraints

Rhythm Information:

Only the rhythm encoder has access to full rhythm information



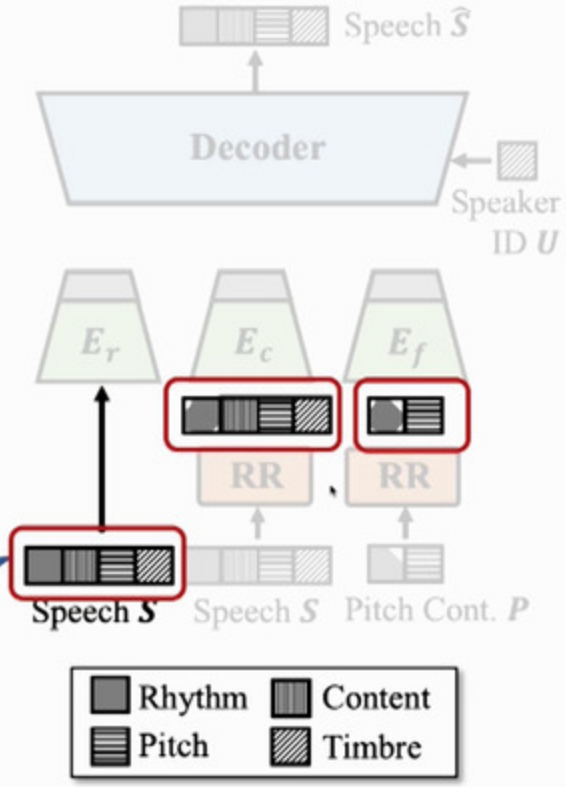
SpeechSplit – Why does it work?

Basic idea: The encoders need to pass all the blocks for perfect reconstruction, subject to the bottleneck constraints

Rhythm Information:

Only the rhythm encoder has access to full rhythm information

The other information can be supplied elsewhere



SpeechSplit – Why does it work?

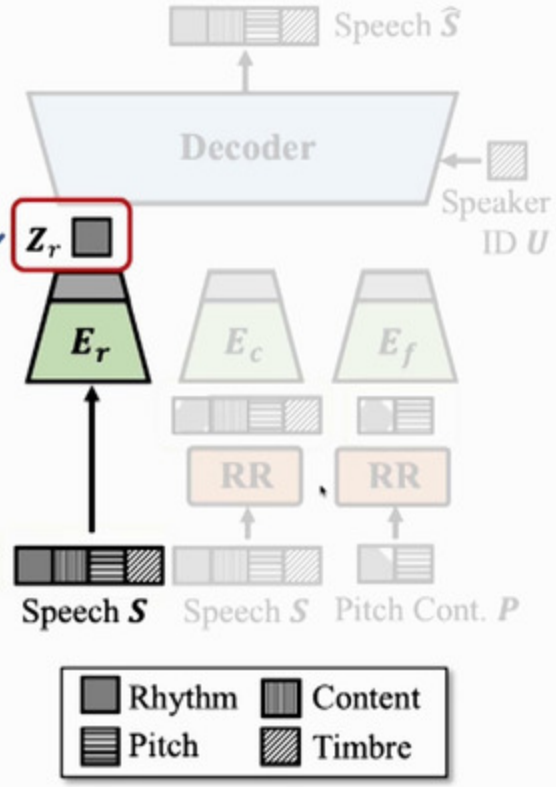
Basic idea: The encoders need to pass all the blocks for perfect reconstruction, subject to the bottleneck constraints

Rhythm Information:

Only the rhythm encoder has access to full rhythm information

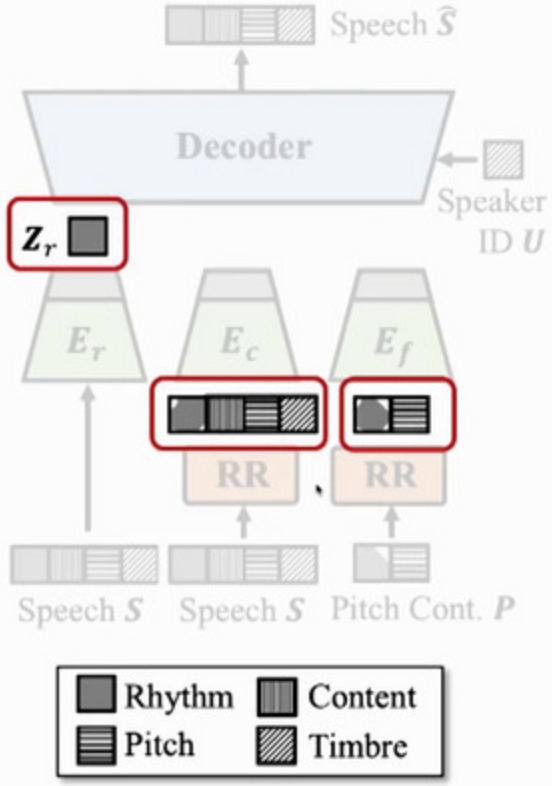
The other information can be supplied elsewhere

The rhythm encoder will prioritize passing the rhythm information; and remove others if bottleneck is binding



SpeechSplit – Why does it work?

Basic idea: The encoders need to pass all the blocks for perfect reconstruction, subject to the bottleneck constraints



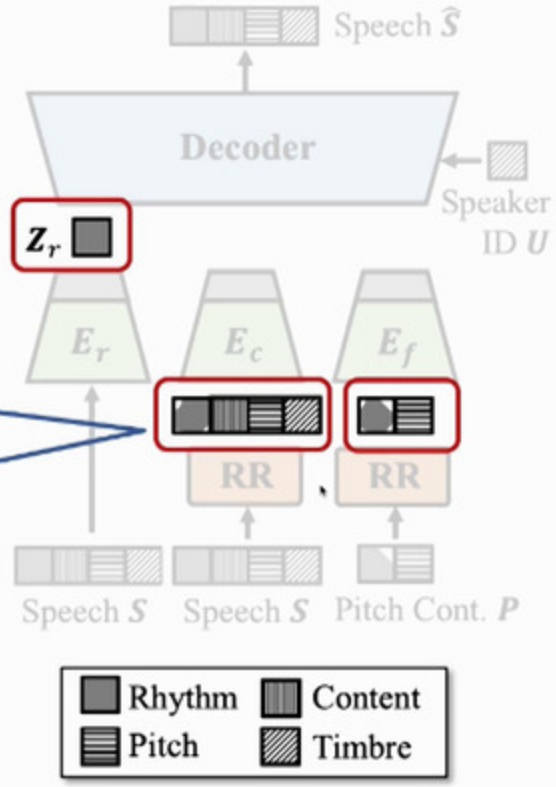
SpeechSplit – Why does it work?

Basic idea: The encoders need to pass all the blocks for perfect reconstruction, subject to the bottleneck constraints

Content Information:

Only the content encoder has access to full content information

The other information can be supplied elsewhere



SpeechSplit – Why does it work?

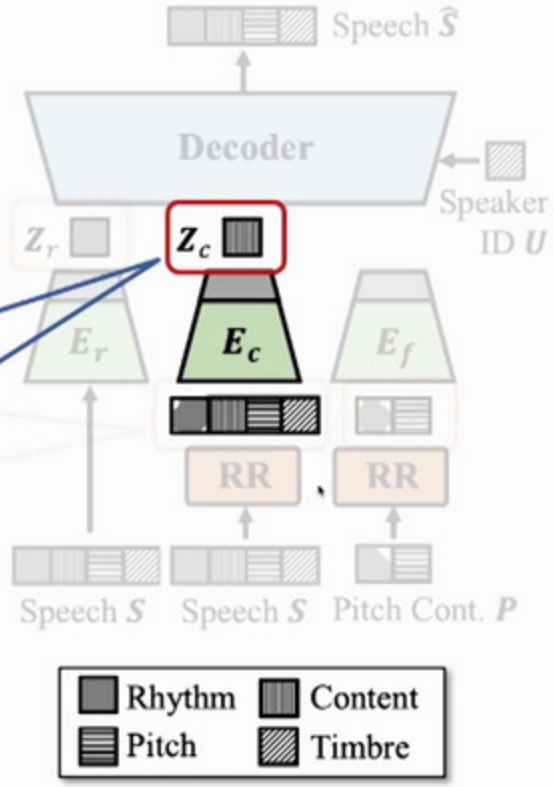
Basic idea: The encoders need to pass all the blocks for perfect reconstruction, subject to the bottleneck constraints

Content Information:

Only the content encoder has access to full content information

The other information can be supplied elsewhere

The content encoder will prioritize passing the content information; and remove others if bottleneck is binding



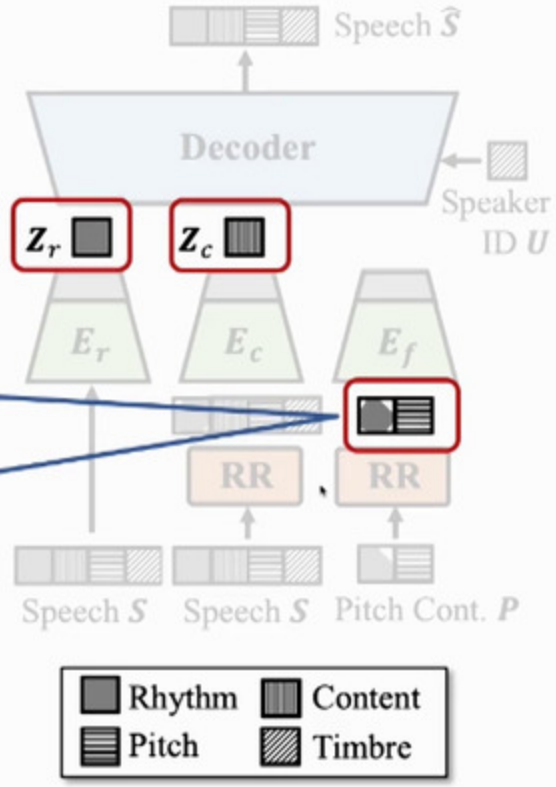
SpeechSplit – Why does it work?

Basic idea: The encoders need to pass all the blocks for perfect reconstruction, subject to the bottleneck constraints

Pitch Information:

Only the pitch encoder has access to full pitch information

The other information can be supplied elsewhere



SpeechSplit – Why does it work?

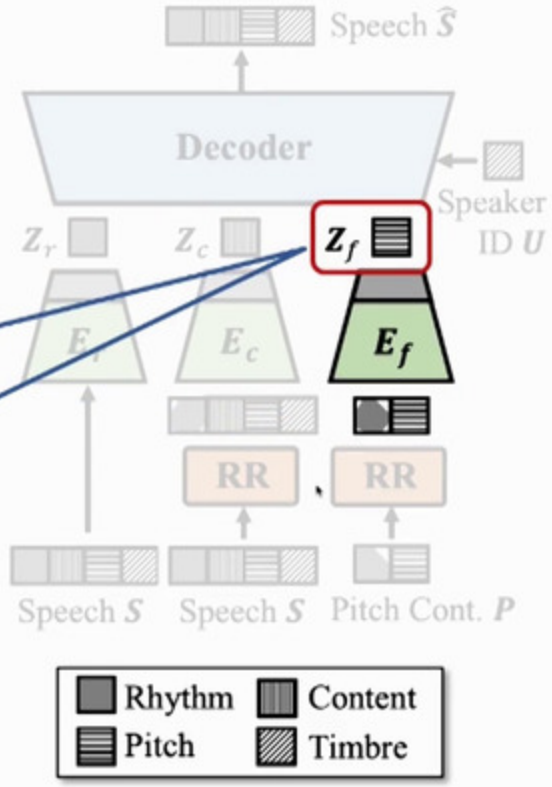
Basic idea: The encoders need to pass all the blocks for perfect reconstruction, subject to the bottleneck constraints

Pitch Information:

Only the pitch encoder has access to full pitch information

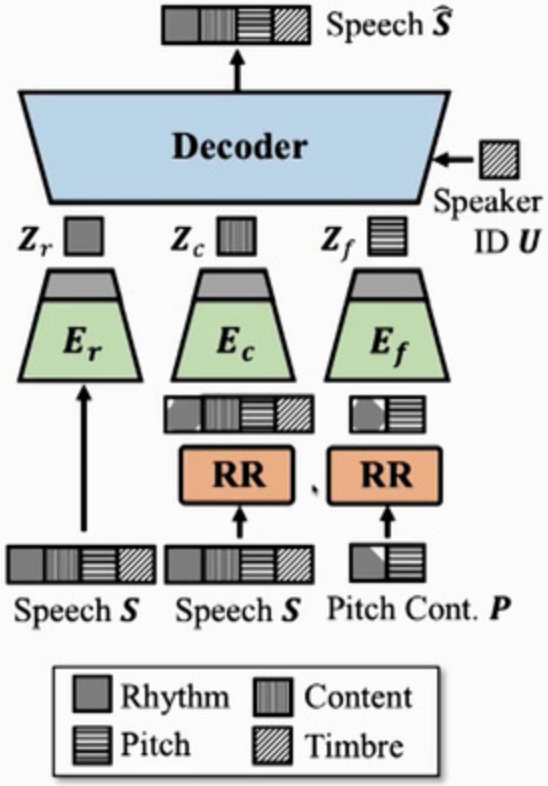
The other information can be supplied elsewhere

The pitch encoder will prioritize passing the pitch information; and remove others if bottleneck is binding



SpeechSplit – Why does it work?

Summary: Given the appropriate bottleneck constraints, this arrangement is the only way to achieve perfect reconstruction



Does SpeechSplit Work As Expected?

Does SpeechSplit Work As Expected?

Exper 1: Aspect-specific Conversion

- For a parallel speech pair, replace rhythm/pitch/timbre embedding of the source with that of the target.

Does SpeechSplit Work As Expected?

Exper 1: Aspect-specific Conversion

- For a parallel speech pair, replace rhythm/pitch/timbre embedding of the source with that of the target.

Hearing is Believing

https://auspicious3000.github.io/SpeechSplit-Demo/

Does SpeechSplit Work As Expected?

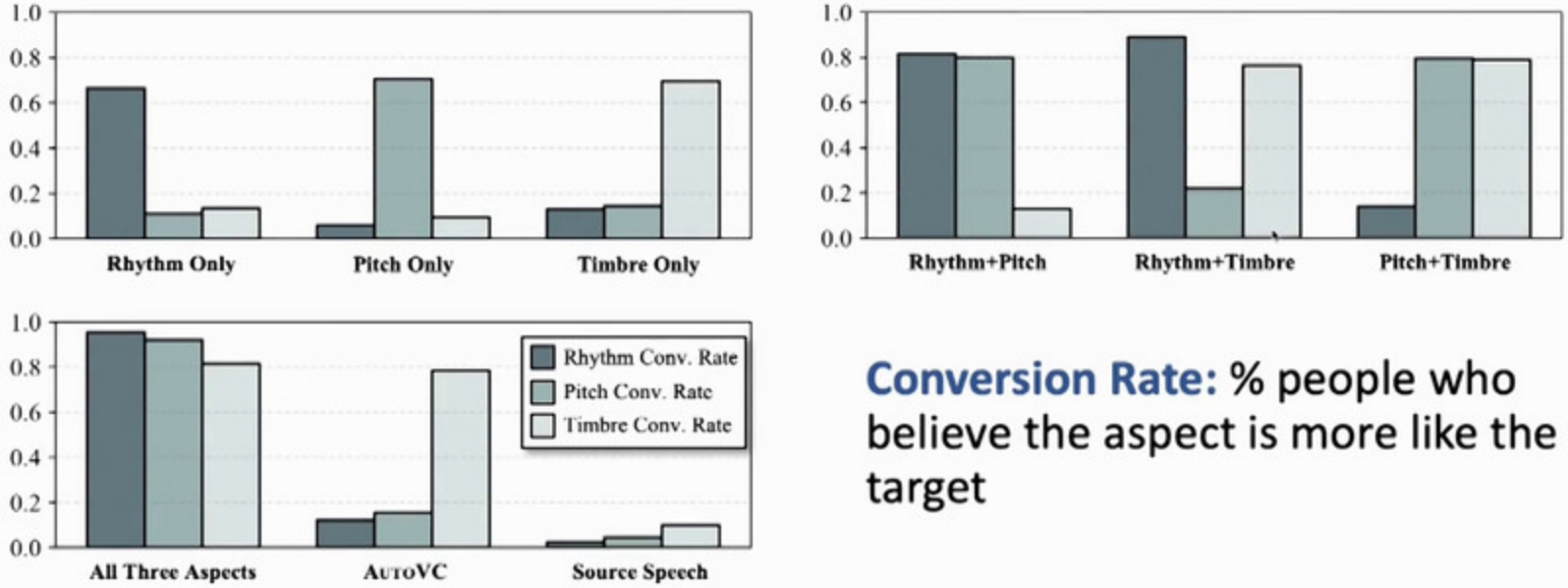
Exper 1: Aspect-specific Conversion

- AMT Study: Subjects choose whether rhythm/pitch/timbre of the conversion speech is more like the source or the target (randomized)

Does SpeechSplit Work As Expected?

Exper 1: Aspect-specific Conversion

- AMT Study: Subjects choose whether rhythm/pitch/timbre of the conversion speech is more like the source or the target (randomized)

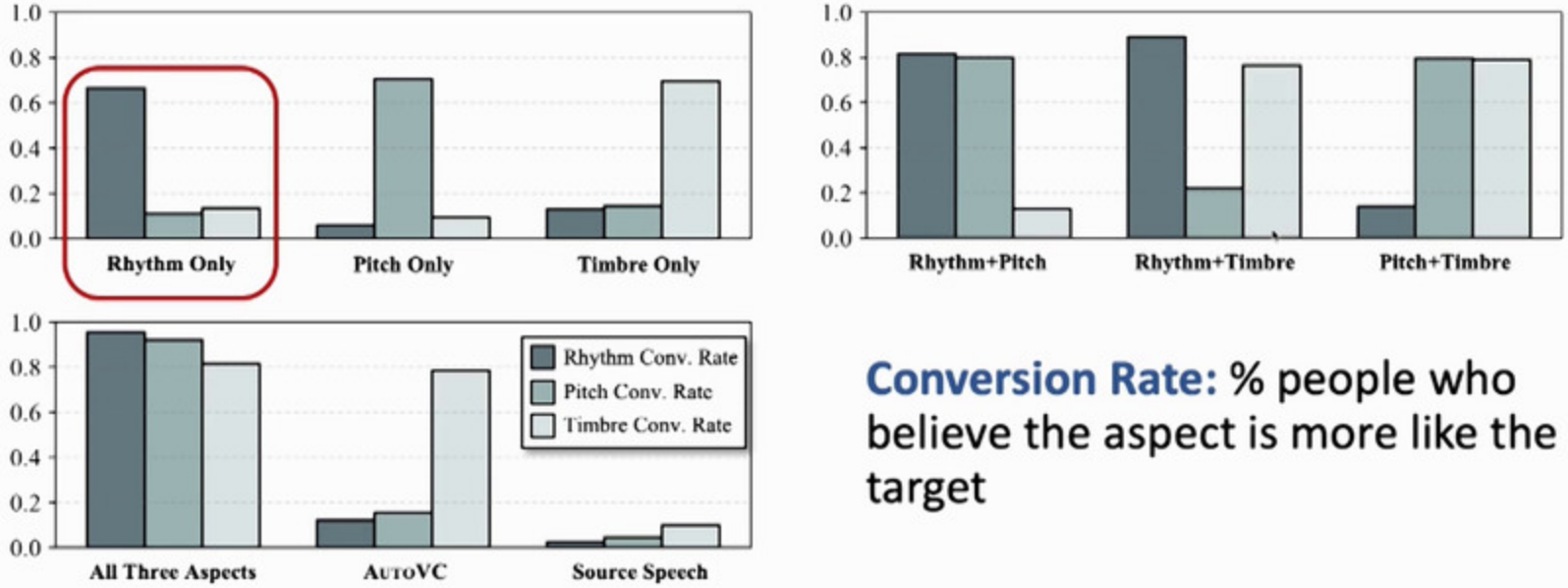


Conversion Rate: % people who believe the aspect is more like the target

Does SpeechSplit Work As Expected?

Exper 1: Aspect-specific Conversion

- AMT Study: Subjects choose whether rhythm/pitch/timbre of the conversion speech is more like the source or the target (randomized)



Conversion Rate: % people who believe the aspect is more like the target

Does SpeechSplit Work As Expected?

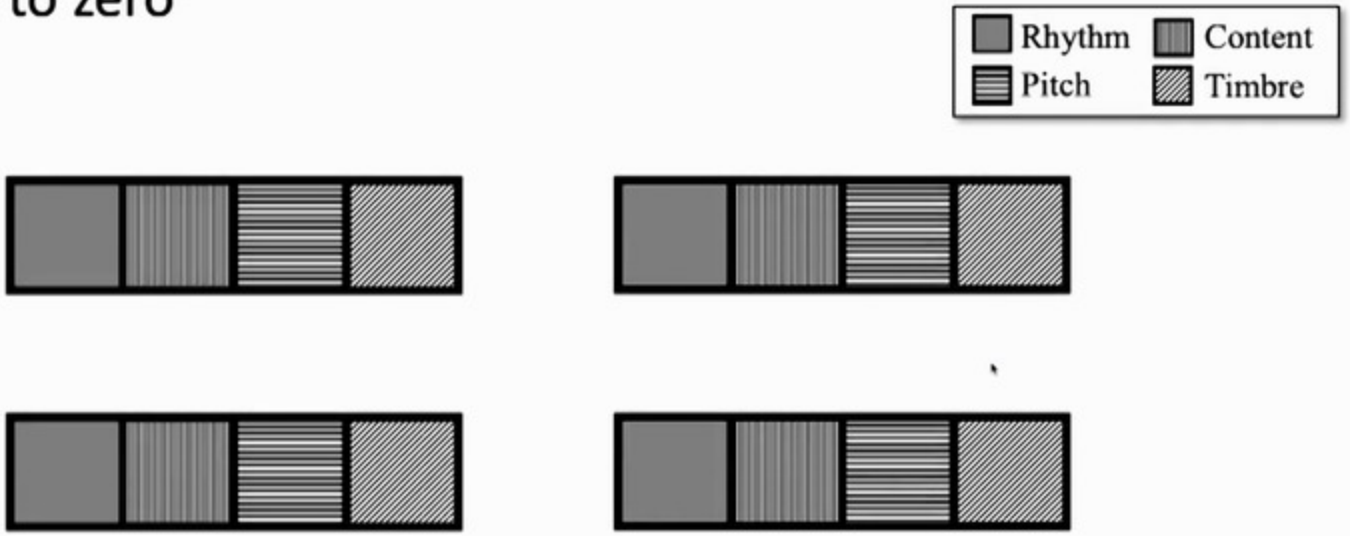
Exper 2: Removing Speech Components

- Set a specific to zero

Does SpeechSplit Work As Expected?

Exper 2: Removing Speech Components

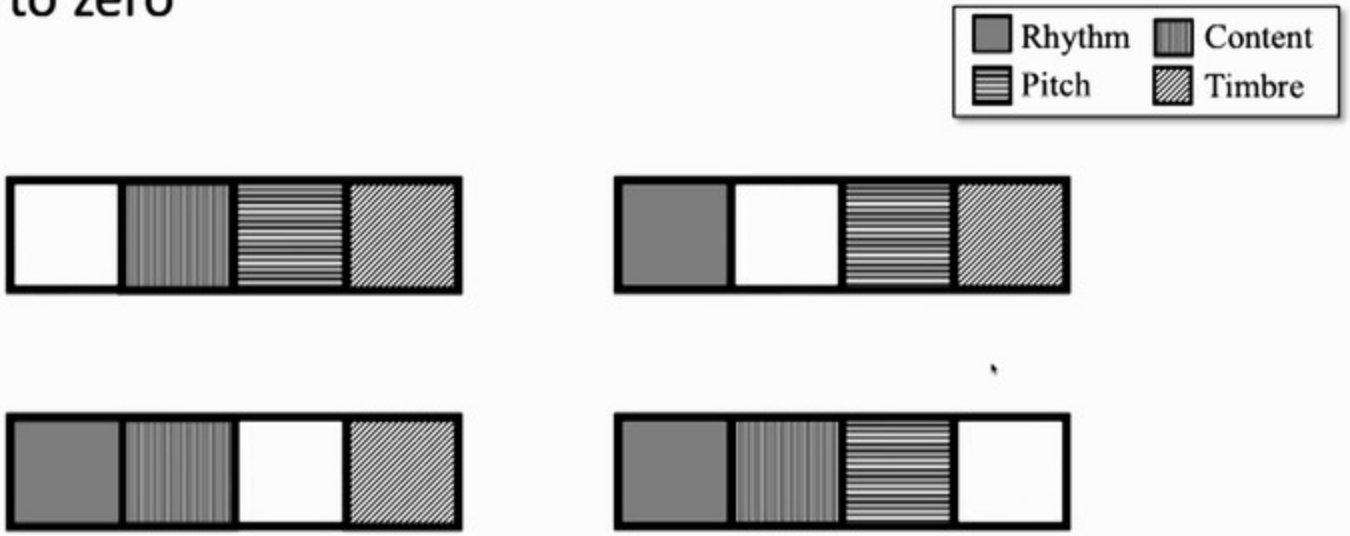
- Set a specific to zero



Does SpeechSplit Work As Expected?

Exper 2: Removing Speech Components

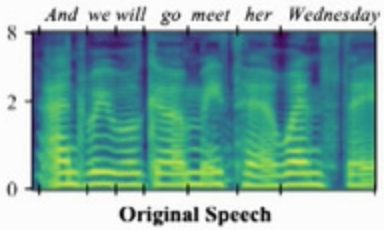
- Set a specific to zero



Does SpeechSplit Work As Expected?

Exper 2: Removing Speech Components

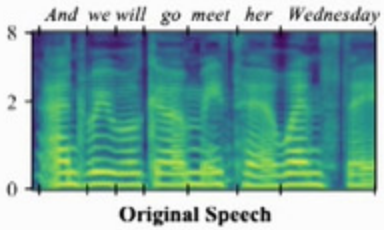
- Set a specific to zero

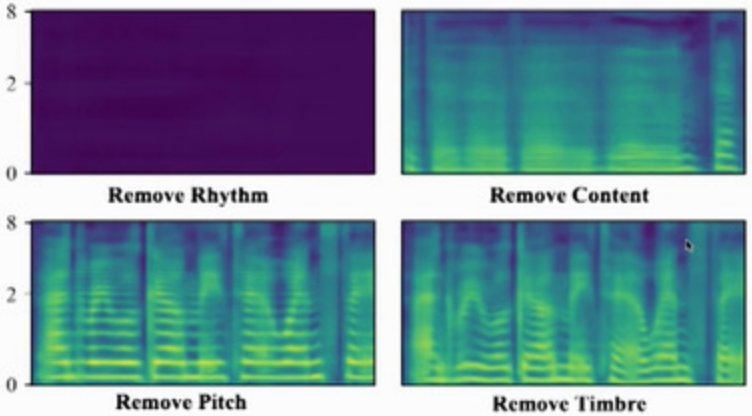


Does SpeechSplit Work As Expected?

Exper 2: Removing Speech Components

- Set a specific to zero

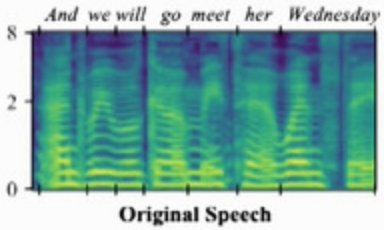


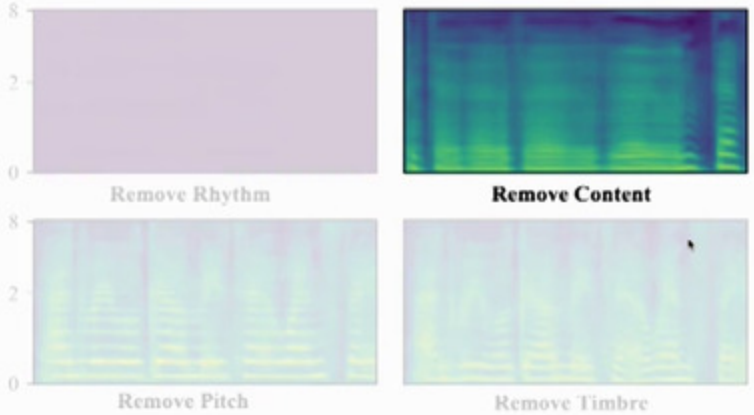


Does SpeechSplit Work As Expected?

Exper 2: Removing Speech Components

- Set a specific to zero

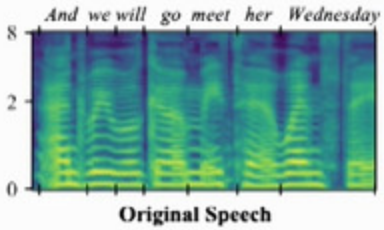


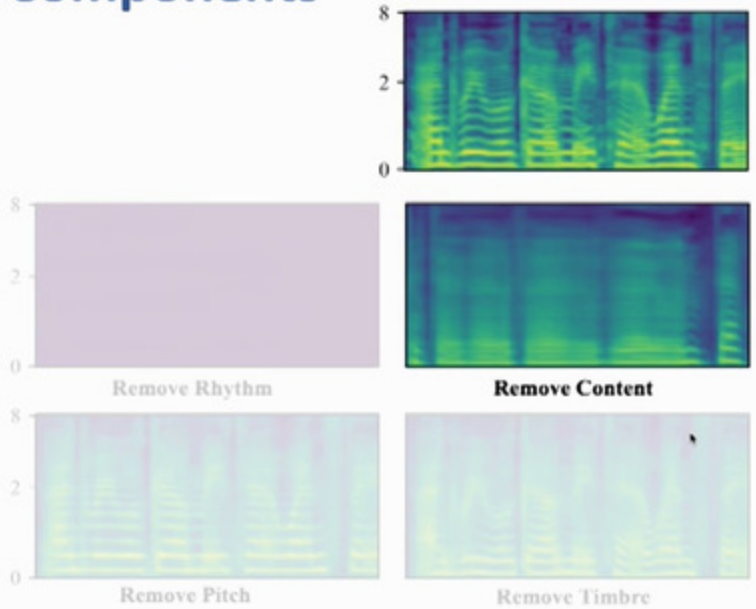


Does SpeechSplit Work As Expected?

Exper 2: Removing Speech Components

- Set a specific to zero

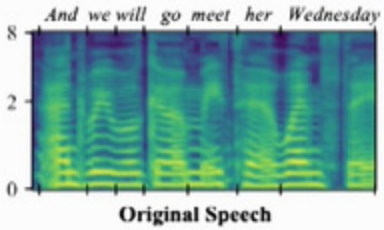


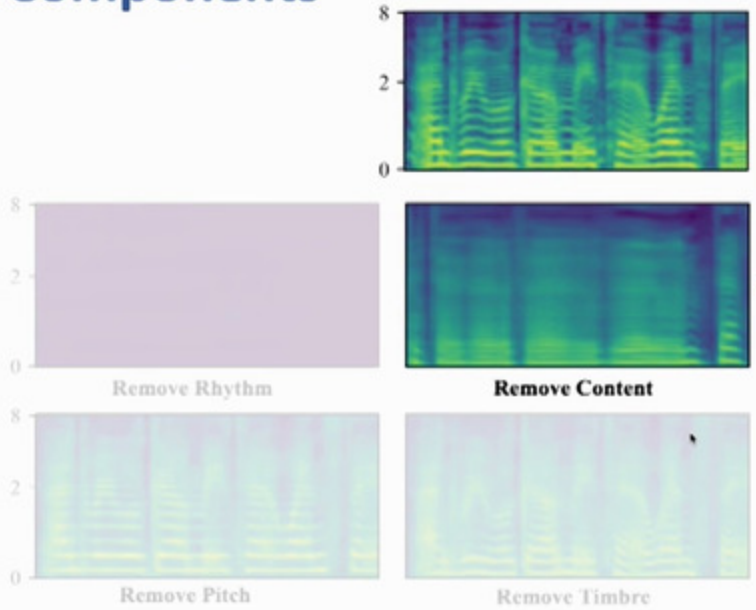


Does SpeechSplit Work As Expected?

Exper 2: Removing Speech Components

- Set a specific to zero



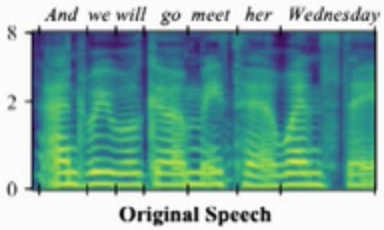


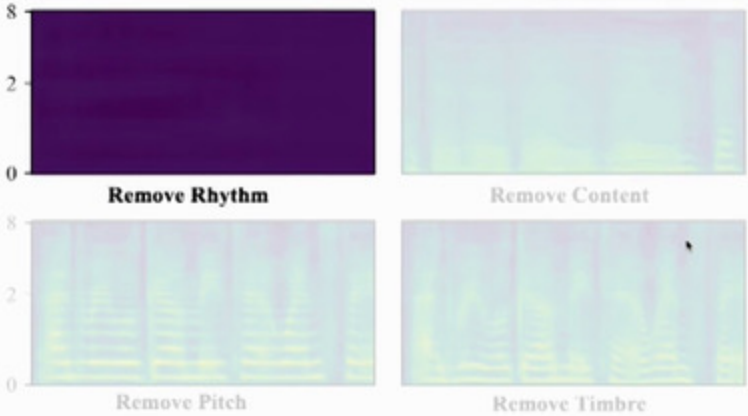
Fill in the blank hypothesis

Does SpeechSplit Work As Expected?

Exper 2: Removing Speech Components

- Set a specific to zero

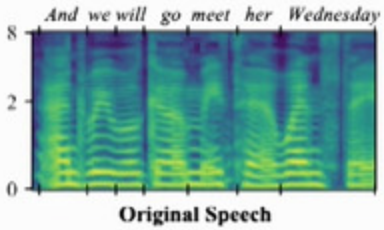


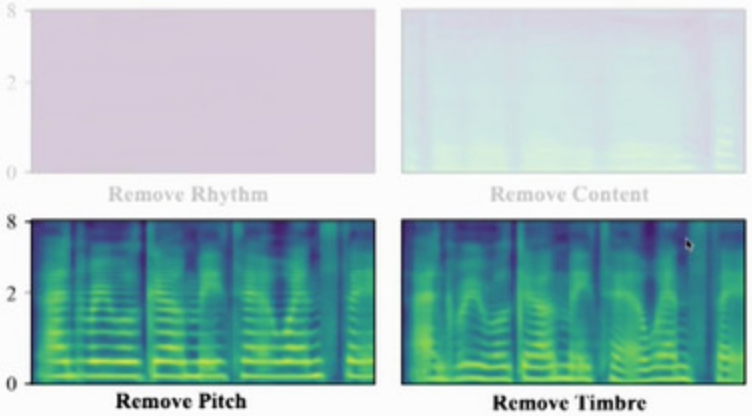


Does SpeechSplit Work As Expected?

Exper 2: Removing Speech Components

- Set a specific to zero





Summary

Summary

- Disentanglement without text transcription is a challenging task

Summary

- Disentanglement without text transcription is a challenging task
- SpeechSplit can achieve so by carefully-designed bottleneck

Summary

- Disentanglement without text transcription is a challenging task
- SpeechSplit can achieve so by carefully-designed bottleneck
- SpeechSplit can perform aspect-specific conversions

Thank you!

Demo

https://auspicious3000.github.io/SpeechSplit-Demo/

Paper

https://arxiv.org/pdf/2004.11284.pdf