Learning Latent Representations for Speech Generation and Transformation

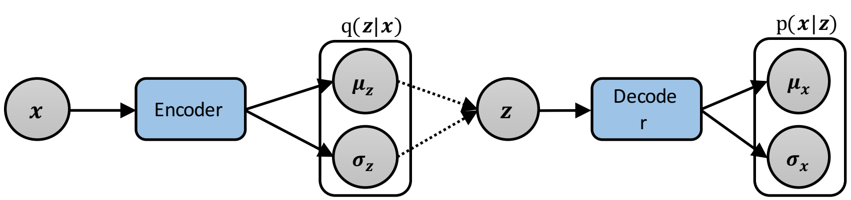
Wei-Ning Hsu, Yu Zhang, James Glass
MIT Computer Science and Artificial Intelligence Laboratory,
Cambridge, MA, USA

Interspeech 2017

CSAIL
MIT COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE LABORATORY

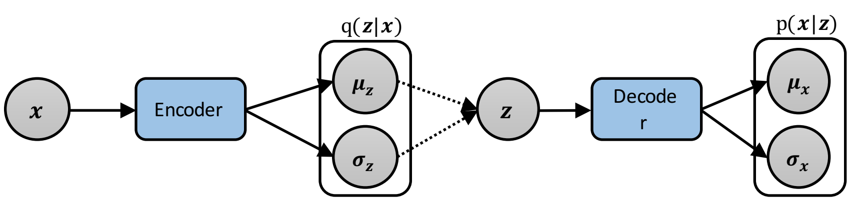
What to Expect in This Talk

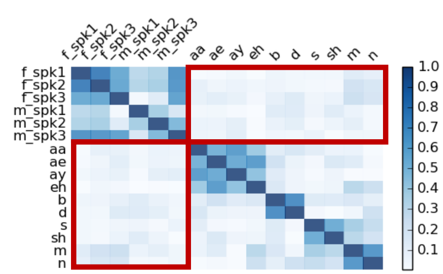
1. A convolutional variational autoencoder framework to model a generative process of speech



What to Expect in This Talk

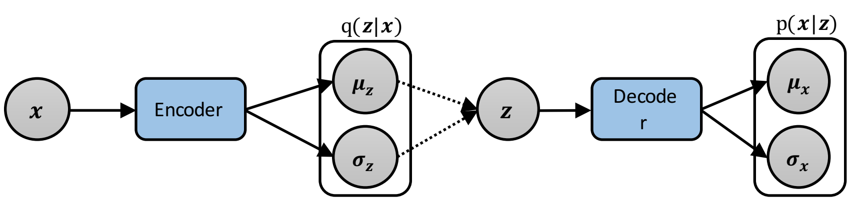
1. A convolutional variational autoencoder framework to model a generative process of speech
2. A method to associate learned latent representations with physical attributes, such as speaker identity and linguistic content

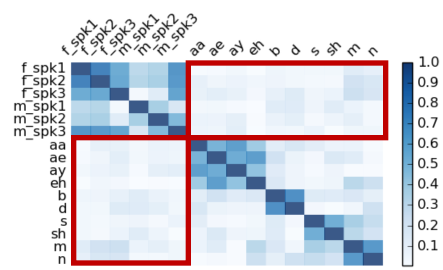


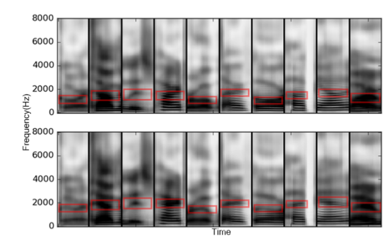


What to Expect in This Talk

1. A convolutional variational autoencoder framework to model a generative process of speech
2. A method to associate learned latent representations with physical attributes, such as speaker identity and linguistic content
3. Simple latent space arithmetic operations to modify speech attributes







Outline

1. Motivation
2. Background and Models
3. Latent Attribute Representations and Operations
4. Experiments
5. Conclusion

Motivation

- We want to learn a generative process of speech
  1. What are the factors that affect speech generation?
  2. How do these factors play a role in speech generation?
  3. How can we infer these factors from observed speech?

Motivation

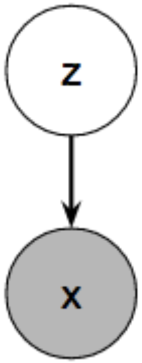
- We want to learn a generative process of speech
  1. What are the factors that affect speech generation?
  2. How do these factors play a role in speech generation?
  3. How can we infer these factors from observed speech?
- Why do we want to learn a generative process?
  - Synthesis (1, 2)
  - Recognition and verification (3)
  - Voice conversion and denoising (1, 2, 3)

Outline

1. Motivations
2. Background and Models
3. Latent Attribute Representations and Operations
4. Experiments
5. Conclusion

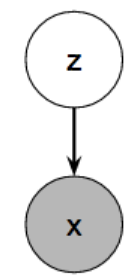
Generative Model Backgrounds

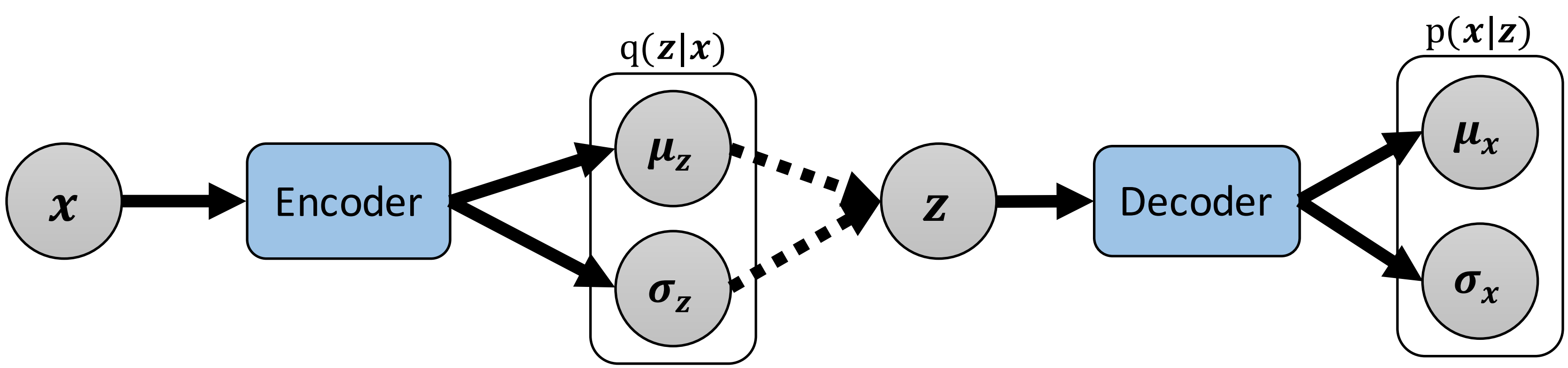
- "Shallow" generative models
  - Hidden Markov model-Gaussian mixture models (HMM-GMMs)
- "Deep" generative models
  - Generative adversarial networks (GANs)
    - model $p(\boldsymbol{x}|\boldsymbol{z})$ and bypass the inference model (generator / discriminator)
  - Auto-regressive models (e.g. WaveNets)
    - model $p(\boldsymbol{x}_t|\boldsymbol{x}_{1:t-1})$ and abstain from using latent variables
  - Variational autoencoders (VAEs)
    - learn an inference model and a generative model jointly



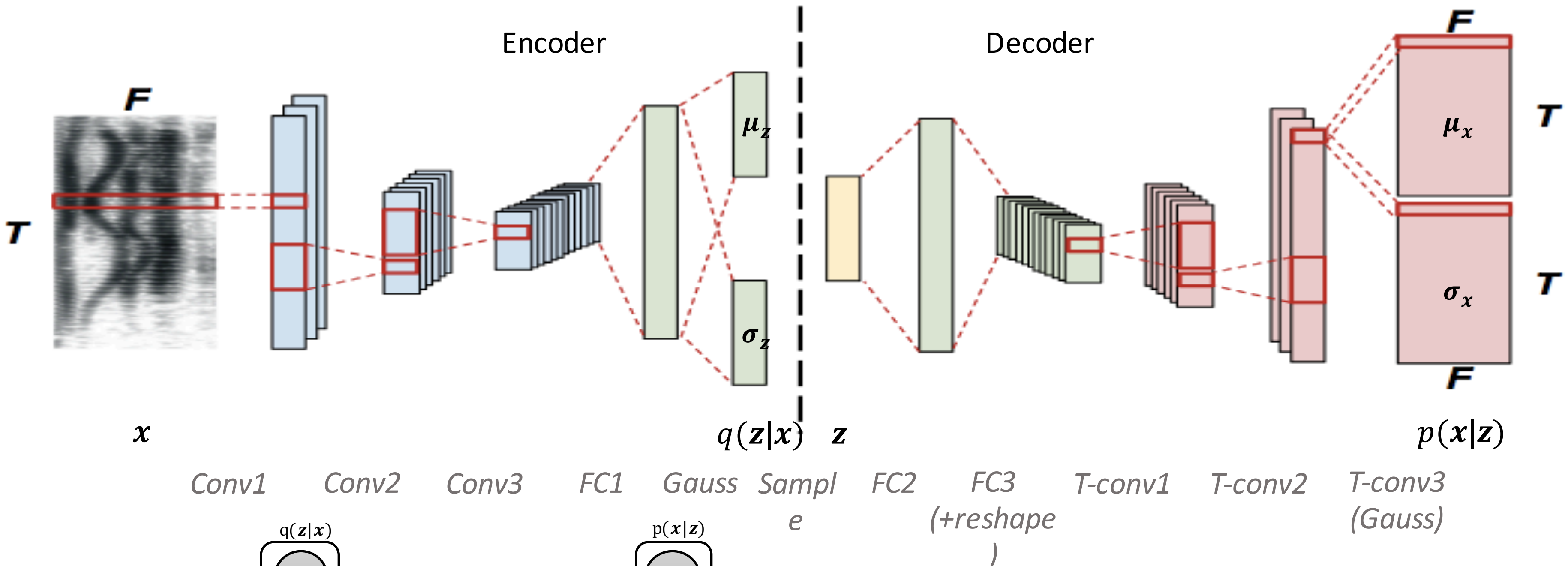
Variational Autoencoders (VAEs)

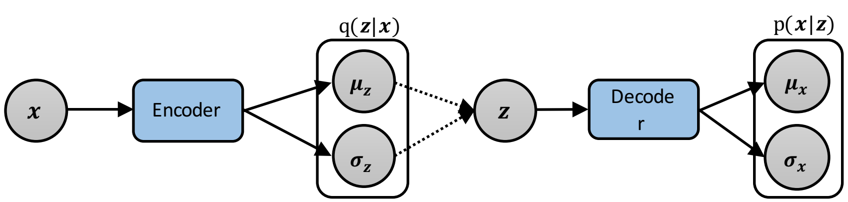
- Define a probabilistic generative process between observation $\boldsymbol{x}$ and latent variable $\boldsymbol{z}$
  - $p(\boldsymbol{z})$, $p(\boldsymbol{x}|\boldsymbol{z})$, and $\mathrm{q}(\boldsymbol{z}|\boldsymbol{x})$ are defined to be in some parametric family
- We define $p(\boldsymbol{x}|\boldsymbol{z})$ (decoder) and $\mathrm{q}(\boldsymbol{z}|\boldsymbol{x})$ (encoder) to be diagonal Gaussians
  - Parameters (mean and variance) are described using some NN
- $p(\boldsymbol{z})$ is defined to be isotropic Gaussian with unit variance





Convolutional Neural Network Architecture

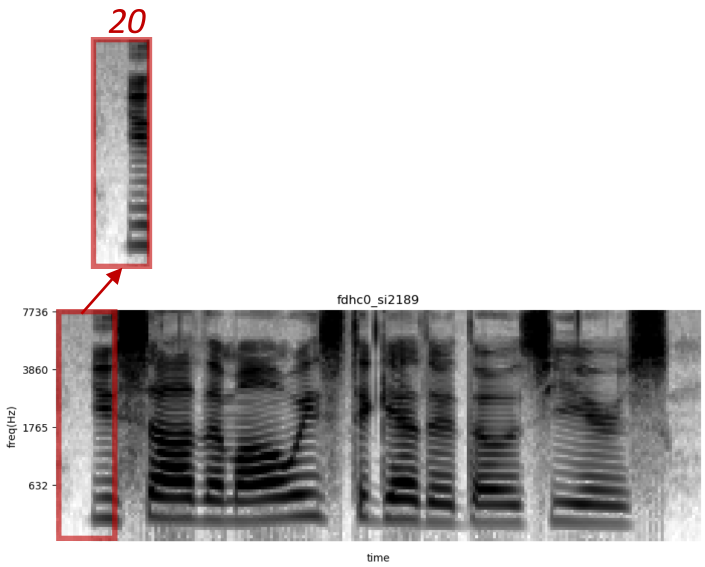




*T-conv stands for transposed convolution

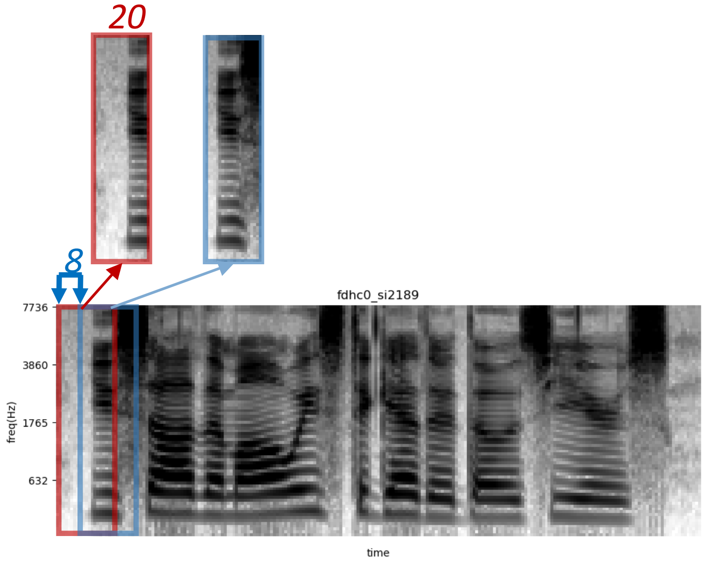
Experiment Setup

- Dataset: TIMIT (5.4hr) (standard 462 speaker sx/si training set)
- Speech Segment Dimension:
  - Unsupervised training (i.e., no use of phonetic transcription)
  - T = 20 frames (with shift of 8 frames)
  - F = 80 (FBank) or 200 (Log Magnitude Spectrogram)
- Training Objective: Variational Lower Bound
- Optimizer: Adam



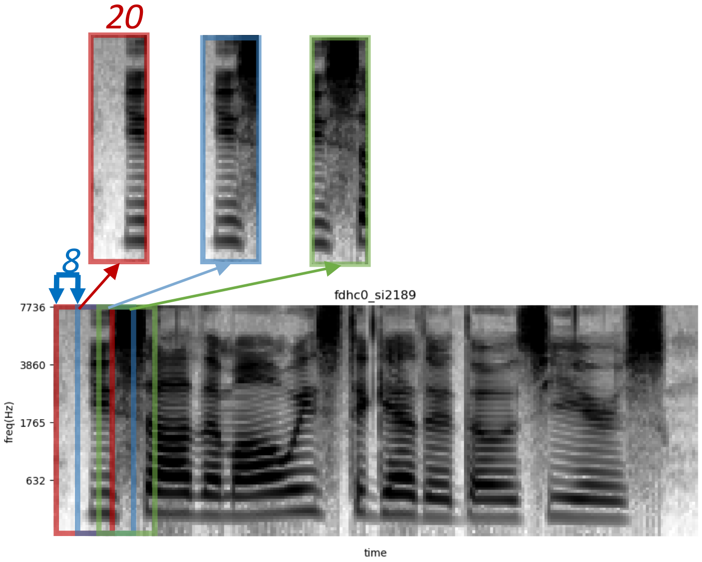
Experiment Setup

- Dataset: TIMIT (5.4hr) (standard 462 speaker sx/si training set)
- Speech Segment Dimension:
  - Unsupervised training (i.e., no use of phonetic transcription)
  - T = 20 frames (with shift of 8 frames)
  - F = 80 (FBank) or 200 (Log Magnitude Spectrogram)
- Training Objective: Variational Lower Bound
- Optimizer: Adam



Experiment Setup

- Dataset: TIMIT (5.4hr) (standard 462 speaker sx/si training set)
- Speech Segment Dimension:
  - Unsupervised training (i.e., no use of phonetic transcription)
  - T = 20 frames (with shift of 8 frames)
  - F = 80 (FBank) or 200 (Log Magnitude Spectrogram)
- Training Objective: Variational Lower Bound
- Optimizer: Adam

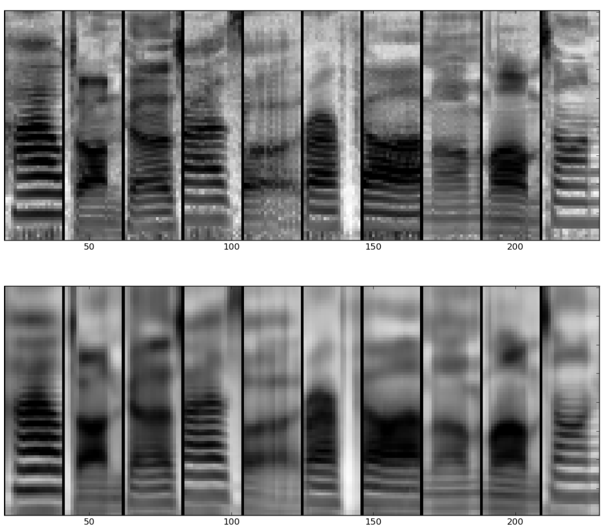


Outline

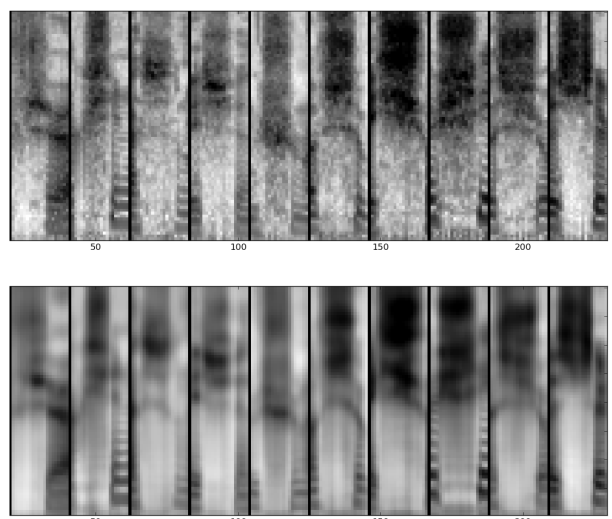
1. Motivations
2. Background and Models
3. Latent Attribute Representations and Operations
4. Experiments
5. Audio Demo
6. Conclusion

Speech Reconstruction Illustration

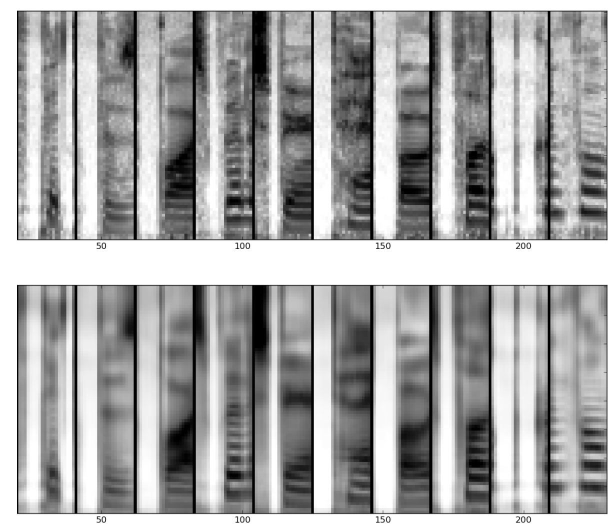
- The trained VAE is able to reconstruct speech segments
- Examples from 10 instances of /aa/, /sh/, and /p/ (sampled at center of segment)



/aa/



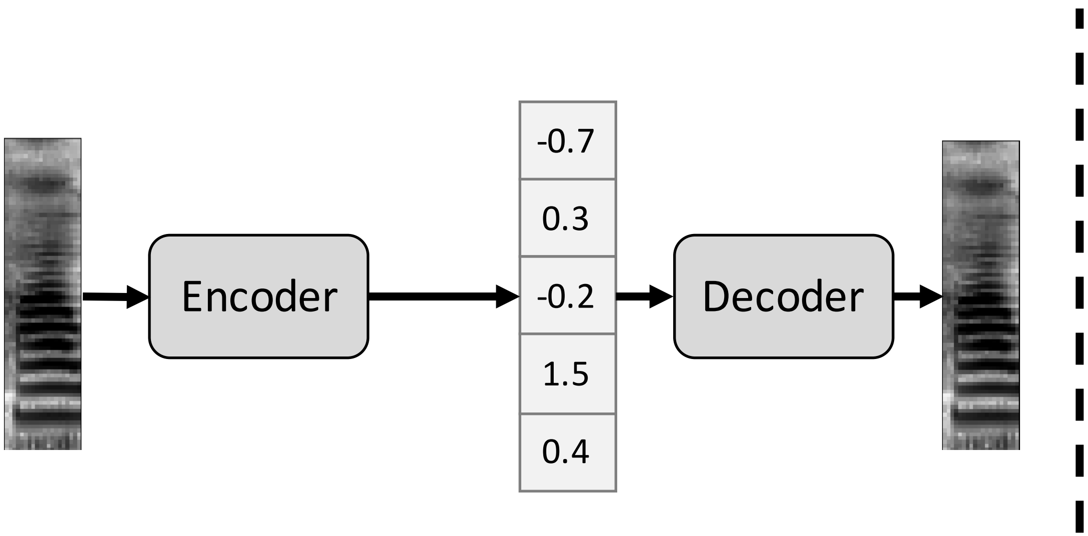
/sh/



/p/

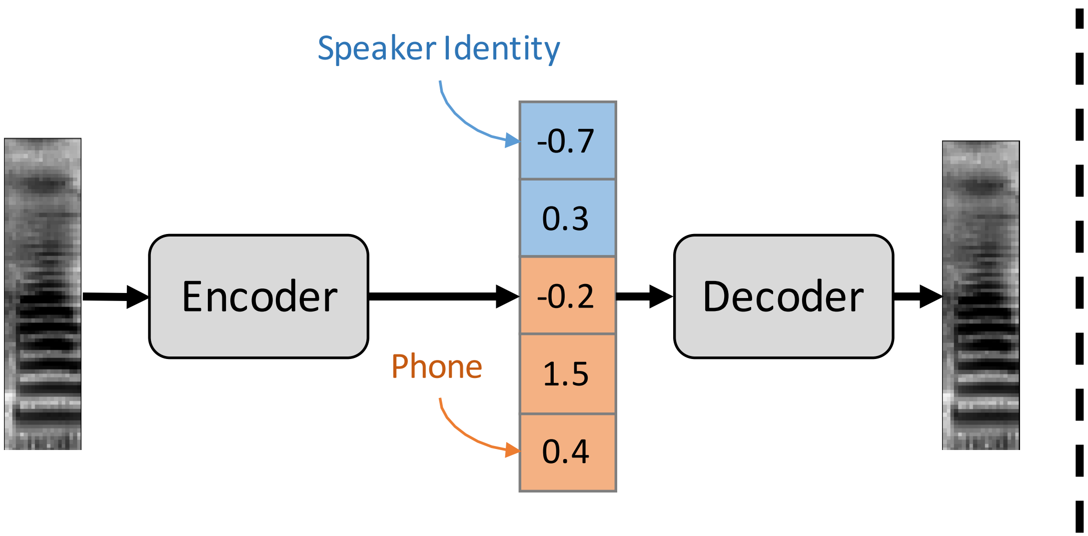
Latent Attribute Representations

- VAE is encouraged to model independent factors using different dimensions
  - Because the prior is assumed to be a diagonal Gaussian



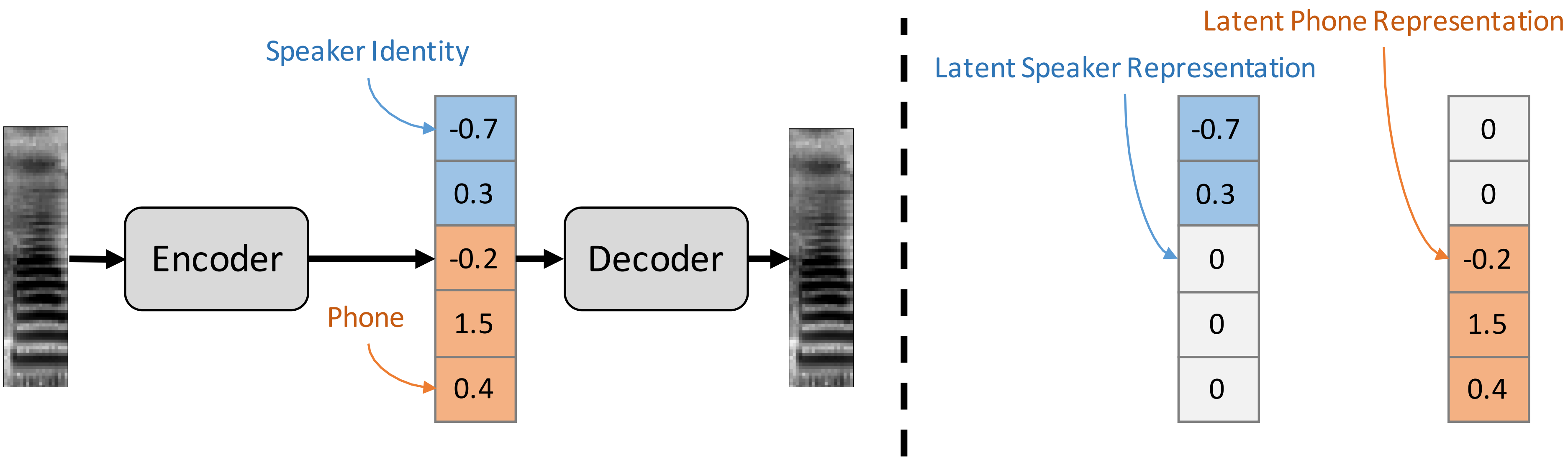
Latent Attribute Representations

- VAE is encouraged to model independent factors using different dimensions
  - Because the prior is assumed to be a diagonal Gaussian
- We want to associate physical attributes with some dimensions



Latent Attribute Representations

- VAE is encouraged to model independent factors using different dimensions
  - Because the prior is assumed to be a diagonal Gaussian
- We want to associate particular dimensions with different physical attributes

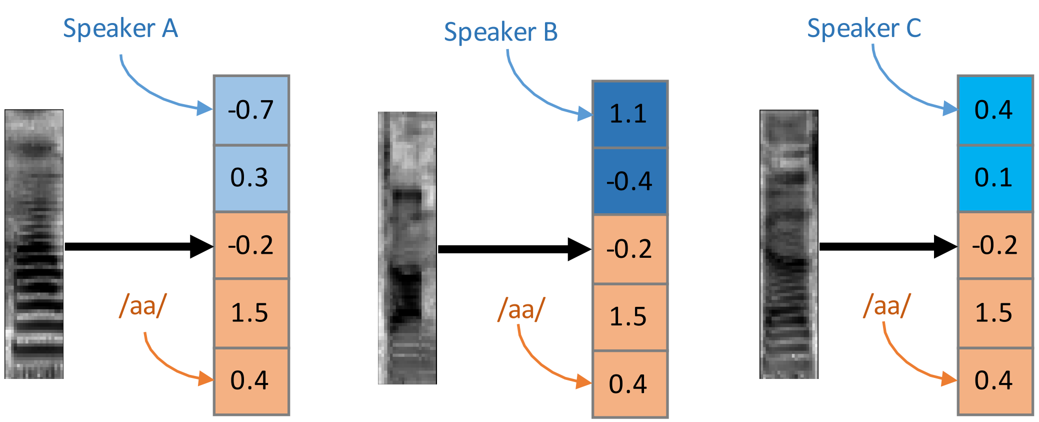


Latent Attribute Representations

- Factors have normal distributions along their associated dimensions

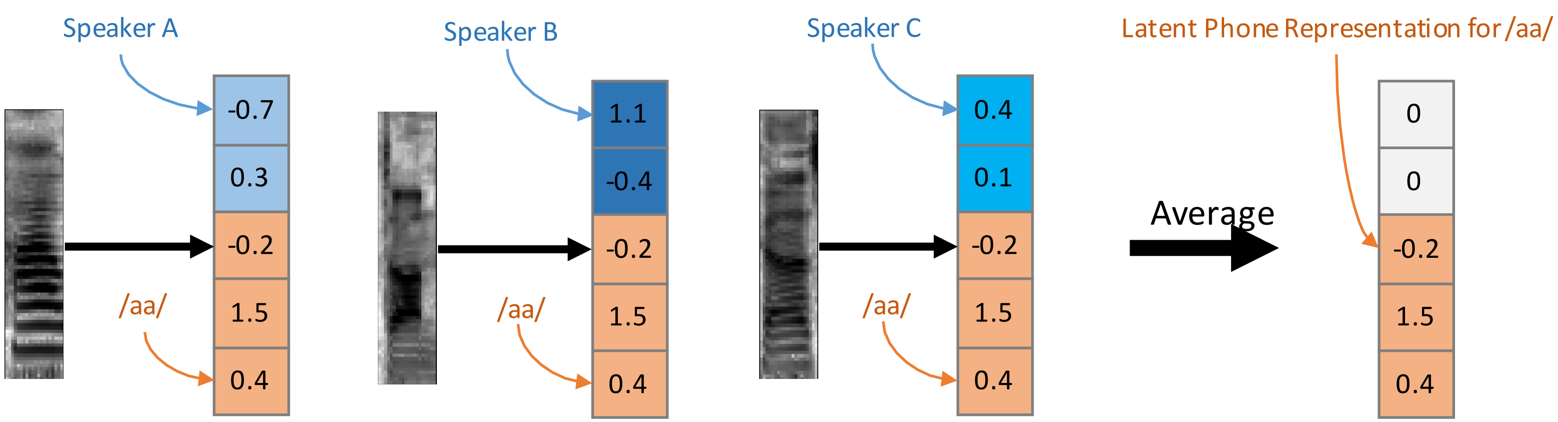
Latent Attribute Representations

- Factors have normal distributions along their associated dimensions
- For example, if we want to estimate the latent phone representation for /aa/:



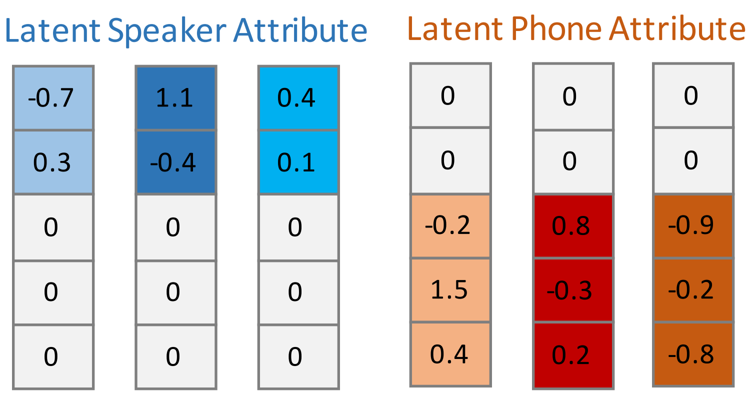
Latent Attribute Representations

- Factors have normal distributions along their associated dimensions
- For example, if we want to estimate the latent phone representation for /aa/:
  - We can estimate latent attribute by **taking the mean latent representations**



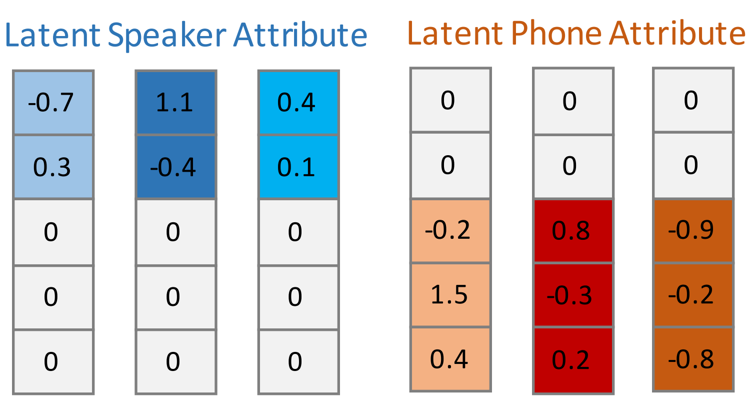
Empirical Study of the Assumptions

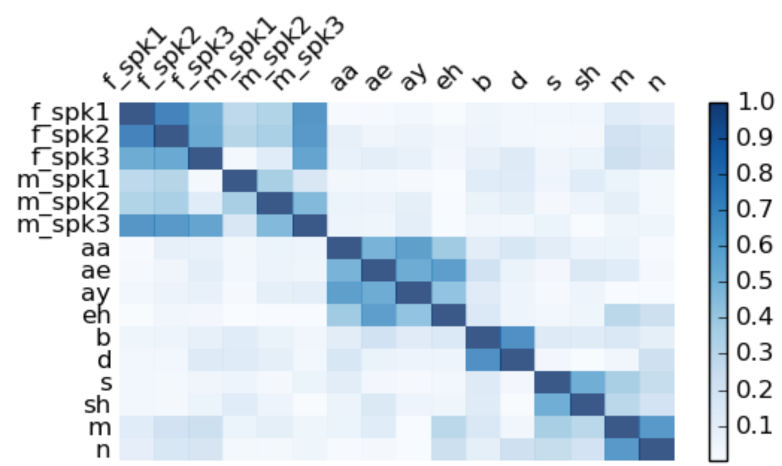
- We compute latent attribute representations of two attributes:



Empirical Study of the Assumptions

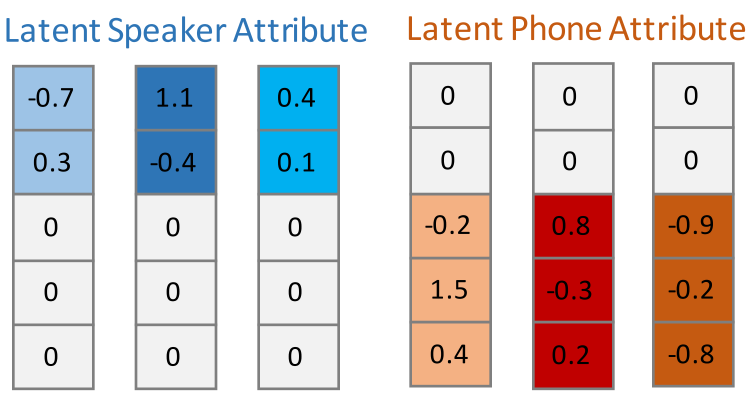
- We compute latent attribute representations of two attributes:
- Compute the absolute cosine similarity between latent attribute representations

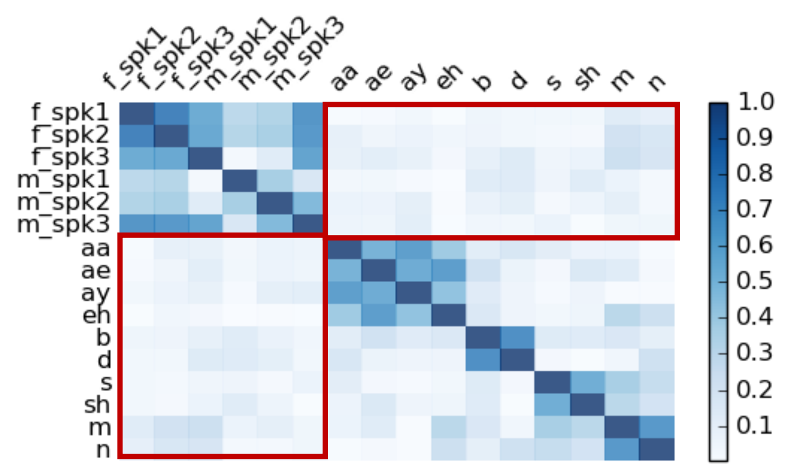




Empirical Study of the Assumptions

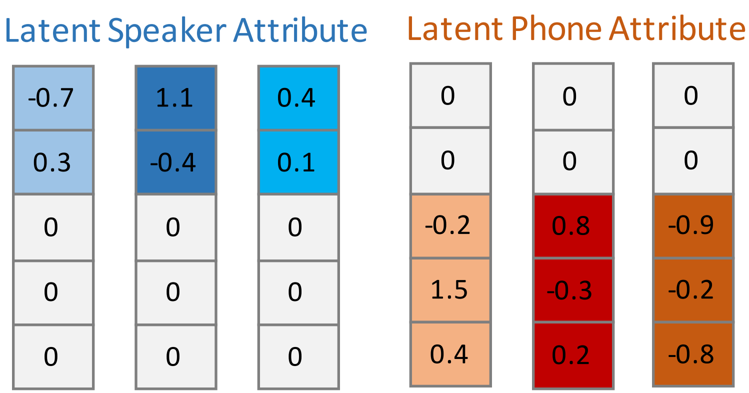
- We compute latent attribute representations of two attributes:
- Compute the absolute cosine similarity between latent attribute representations

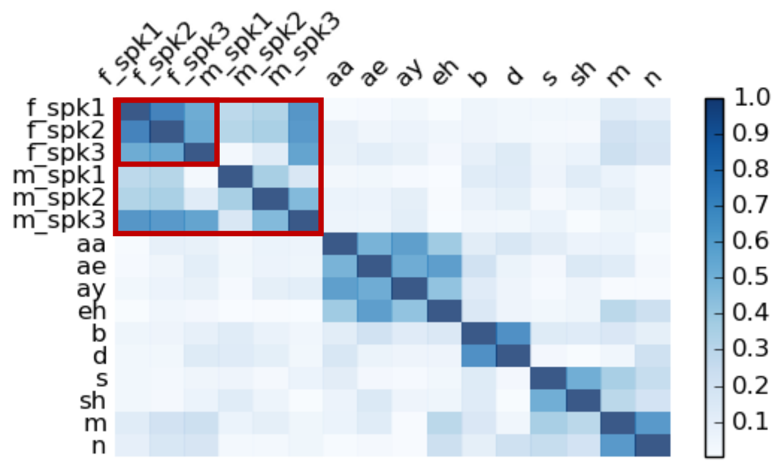




Empirical Study of the Assumptions

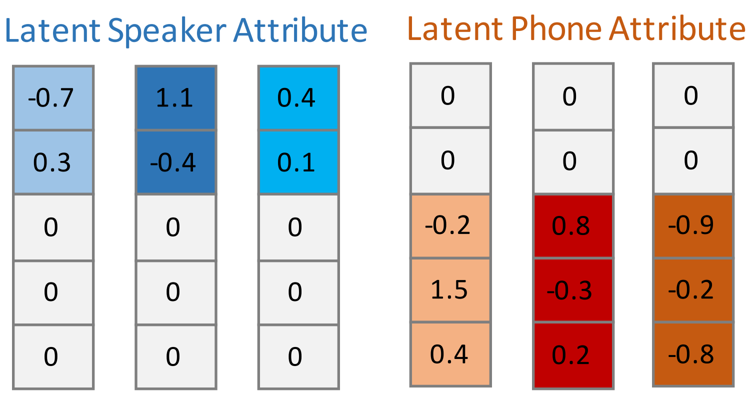
- We compute latent attribute representations of two attributes:
- Compute the absolute cosine similarity between latent attribute representations

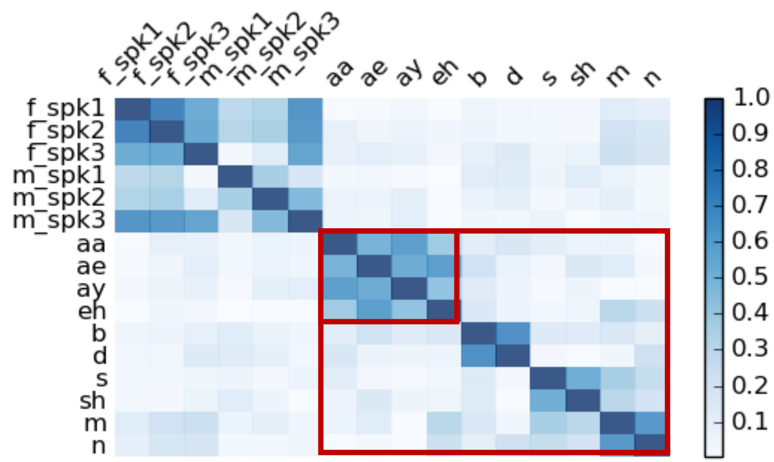




Empirical Study of the Assumptions

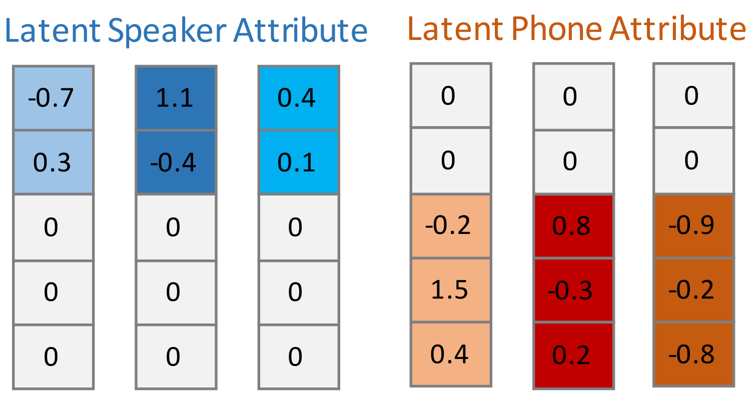
- We compute latent attribute representations of two attributes:
- Compute the absolute cosine similarity between latent attribute representations

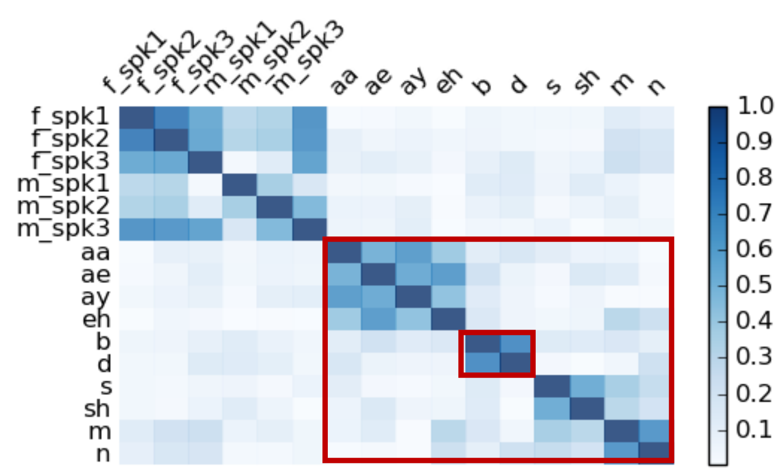




Empirical Study of the Assumptions

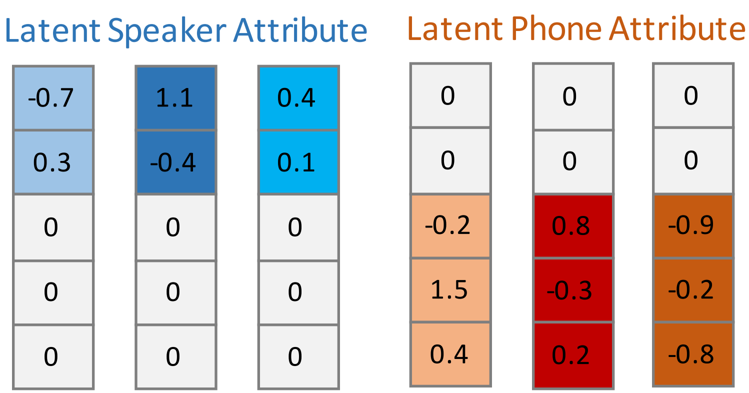
- We compute latent attribute representations of two attributes:
- Compute the absolute cosine similarity between latent attribute representations

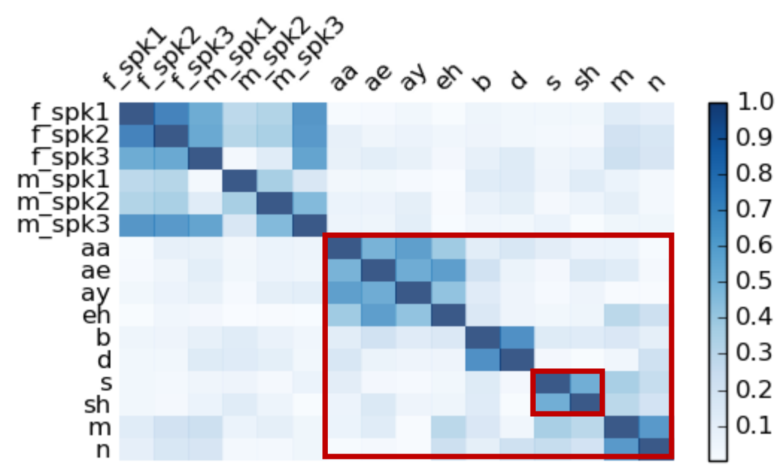




Empirical Study of the Assumptions

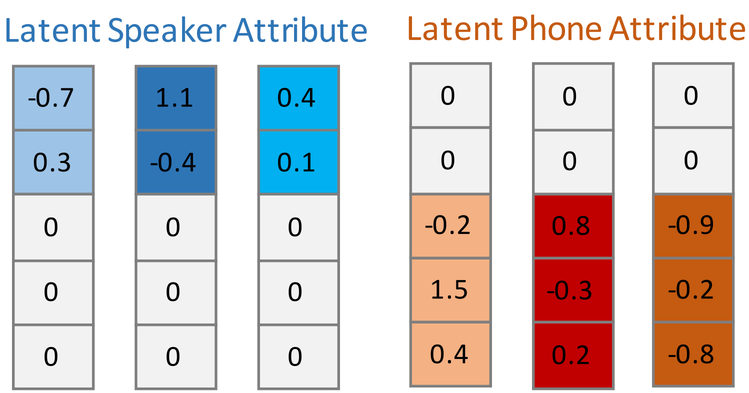
- We compute latent attribute representations of two attributes:
- Compute the absolute cosine similarity between latent attribute representations

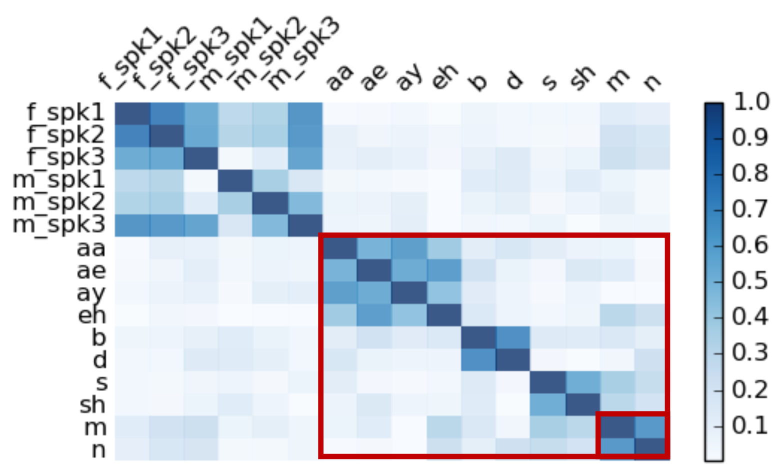




Empirical Study of the Assumptions

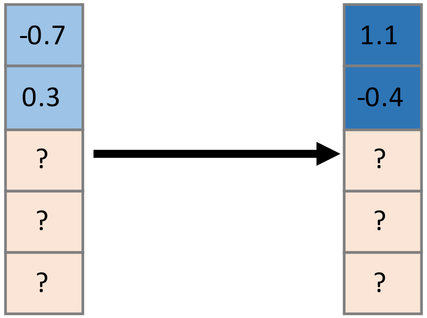
- We compute latent attribute representations of two attributes:
- Compute the absolute cosine similarity between latent attribute representations





Arithmetic Operations to Modify Attributes

- The result suggests that we can modify a specific attribute without altering the others
  - Suppose we want to convert the voice from speaker A (light blue) to speaker B (dark blue)
  - We can do the following operations:

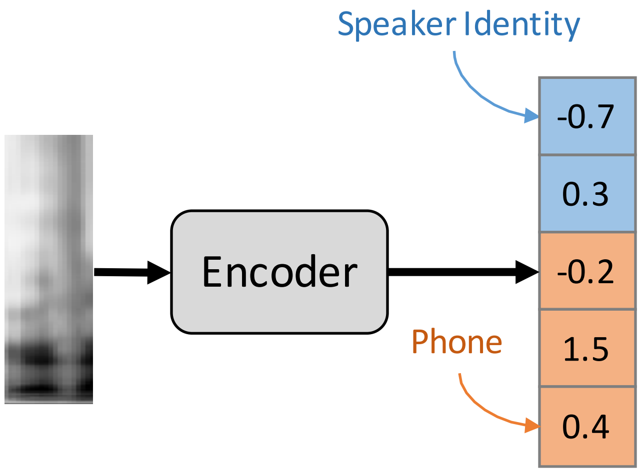


Arithmetic Operations to Modify Attributes

- The result suggests that we can modify a specific attribute without altering the others
  - Suppose we want to convert the voice from speaker A (light blue) to speaker B (dark blue)
  - We can do the following operations:

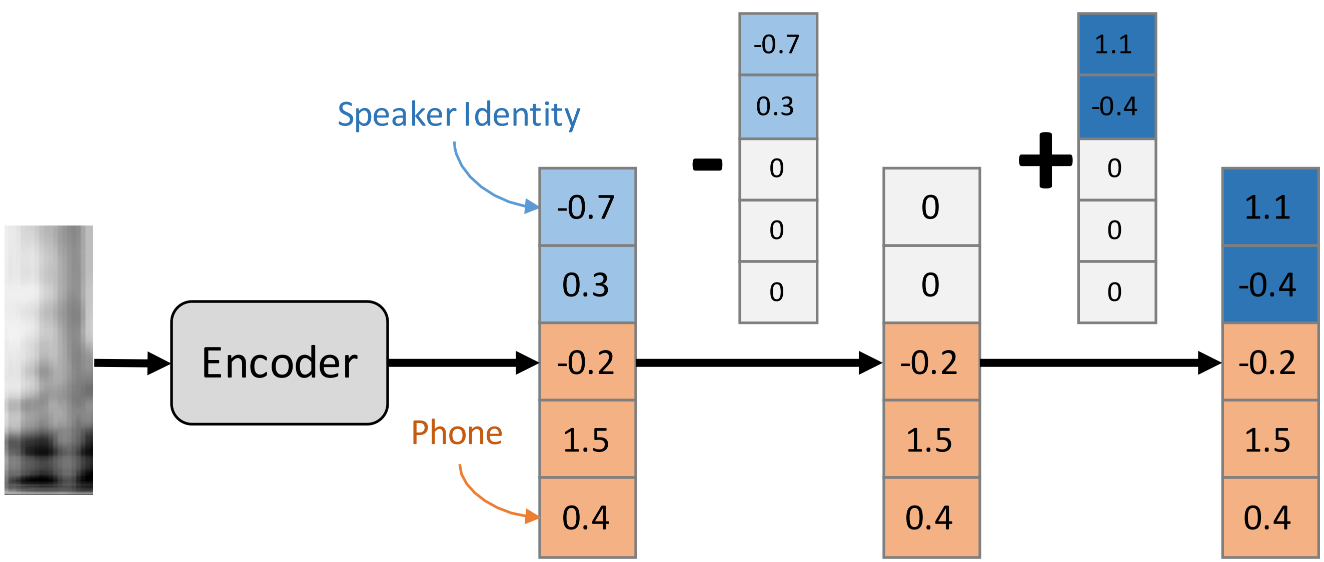
Arithmetic Operations to Modify Attributes

- The result suggests that we can modify a specific attribute without altering the others
  - Suppose we want to convert the voice from speaker A (light blue) to speaker B (dark blue)
  - We can do the following operations:



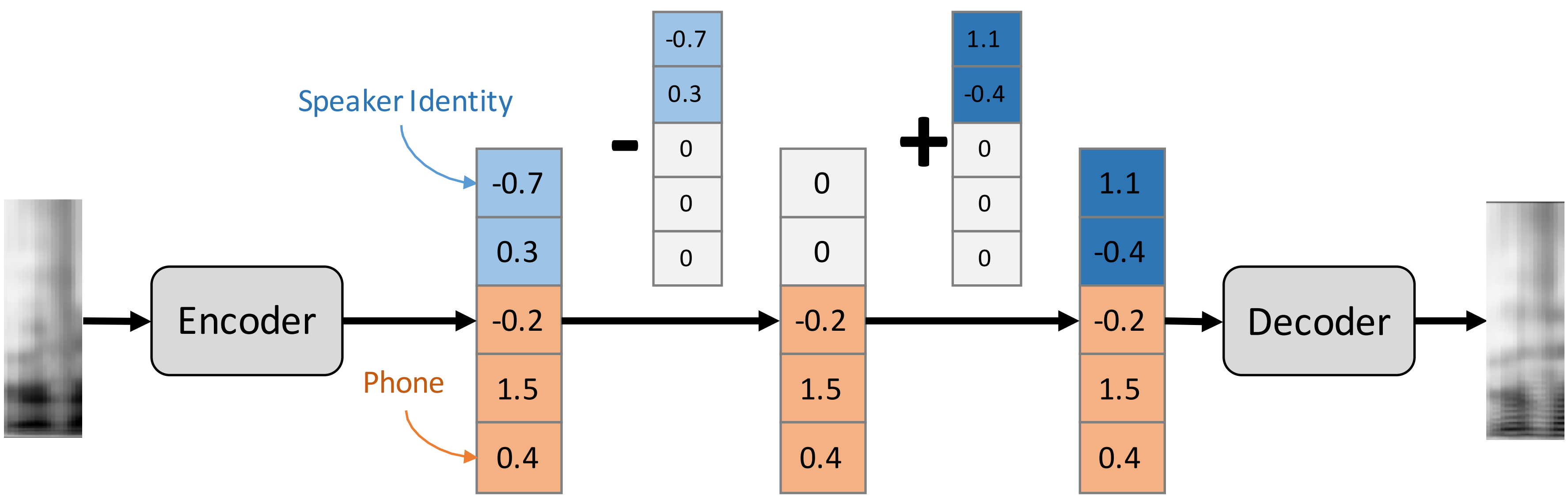
Arithmetic Operations to Modify Attributes

- The result suggests that we can modify a specific attribute without altering the others
  - Suppose we want to convert the voice from speaker A (light blue) to speaker B (dark blue)
  - We can do the following operations:



Arithmetic Operations to Modify Attributes

- The result suggests that we can modify a specific attribute without altering the others
  - Suppose we want to convert the voice from speaker A (light blue) to speaker B (dark blue)
  - We can do the following operations:



Outline

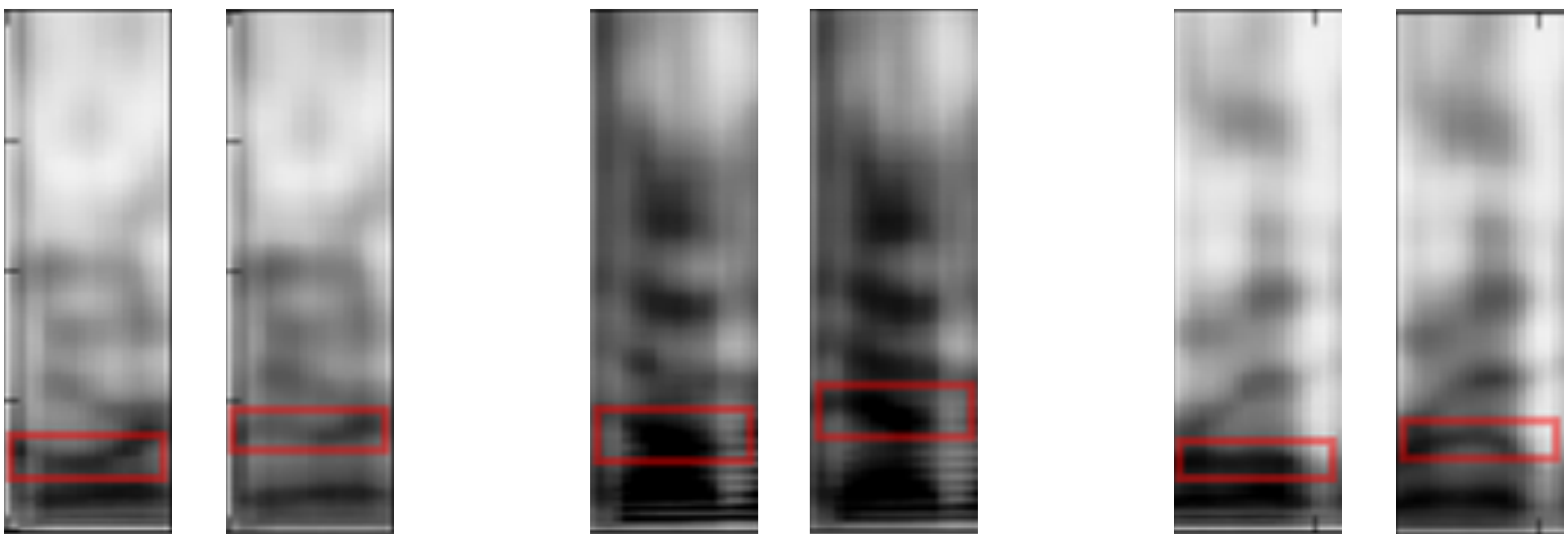
1. Motivations
2. Background and Models
3. Latent Attribute Representations and Operations
4. Experiments
5. Conclusion

Magnitude Spectrogram Reconstruction

- **Griffin and Lim algorithm** is used for waveform reconstruction
  - Iteratively estimate phase

Modify the Phoneme

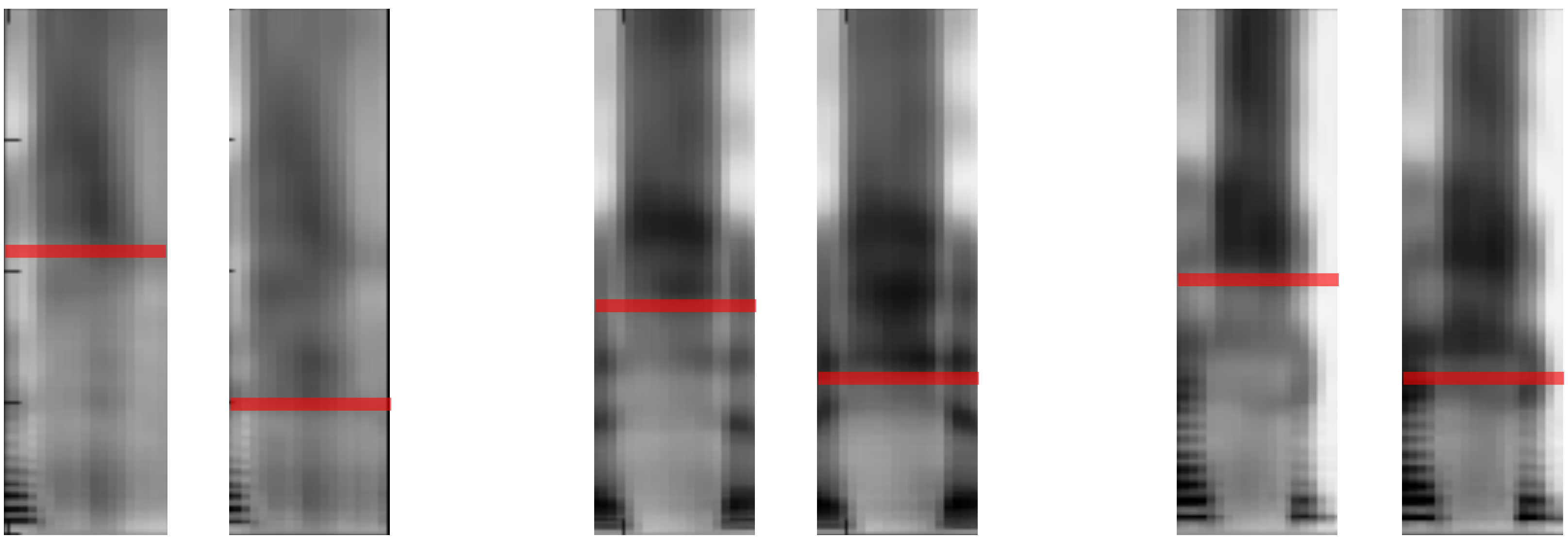
- Modify /aa/ to /ae/, F2 goes up (back vowel -> front vowel)



/aa/ /ae/ /aa/ /ae/ /aa/ /ae/

Modify the Phoneme

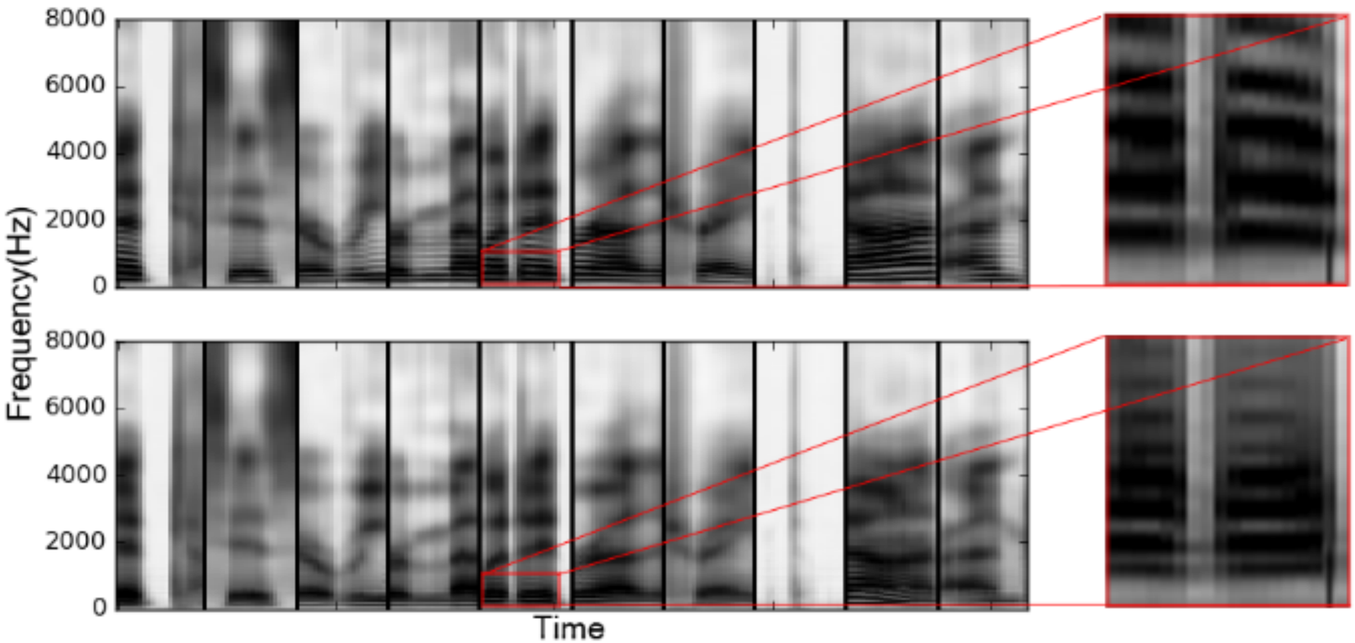
- Modify /s/ to /sh/, cutoff goes down (alveolar -> palatal strident)



/s/ /sh/ /s/ /sh/ /s/ /sh/

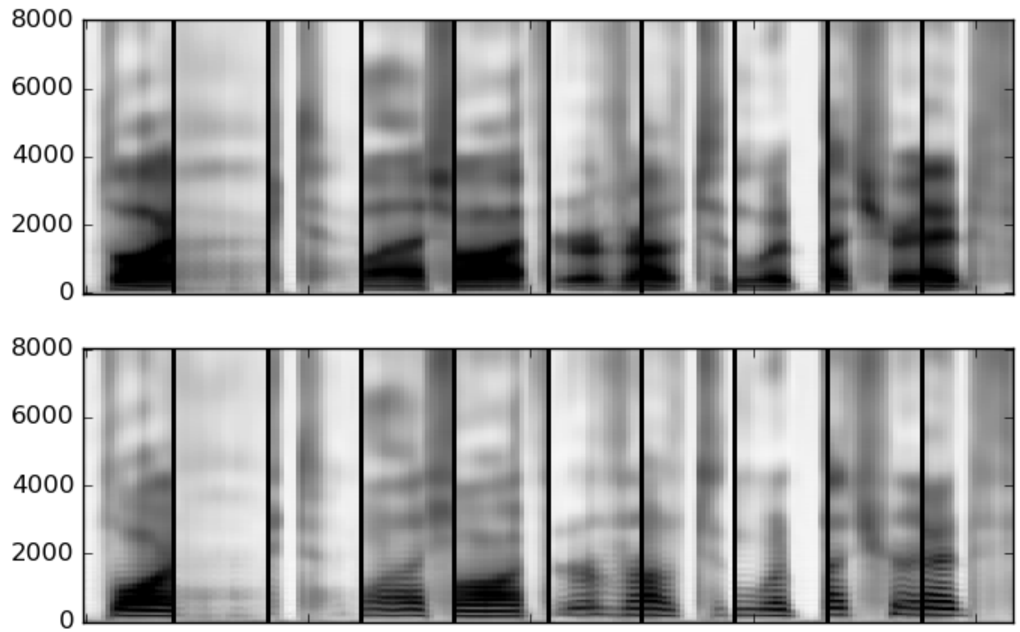
Modify the Speaker

- Modify a female to a male, pitch decreases



Modify the Speaker

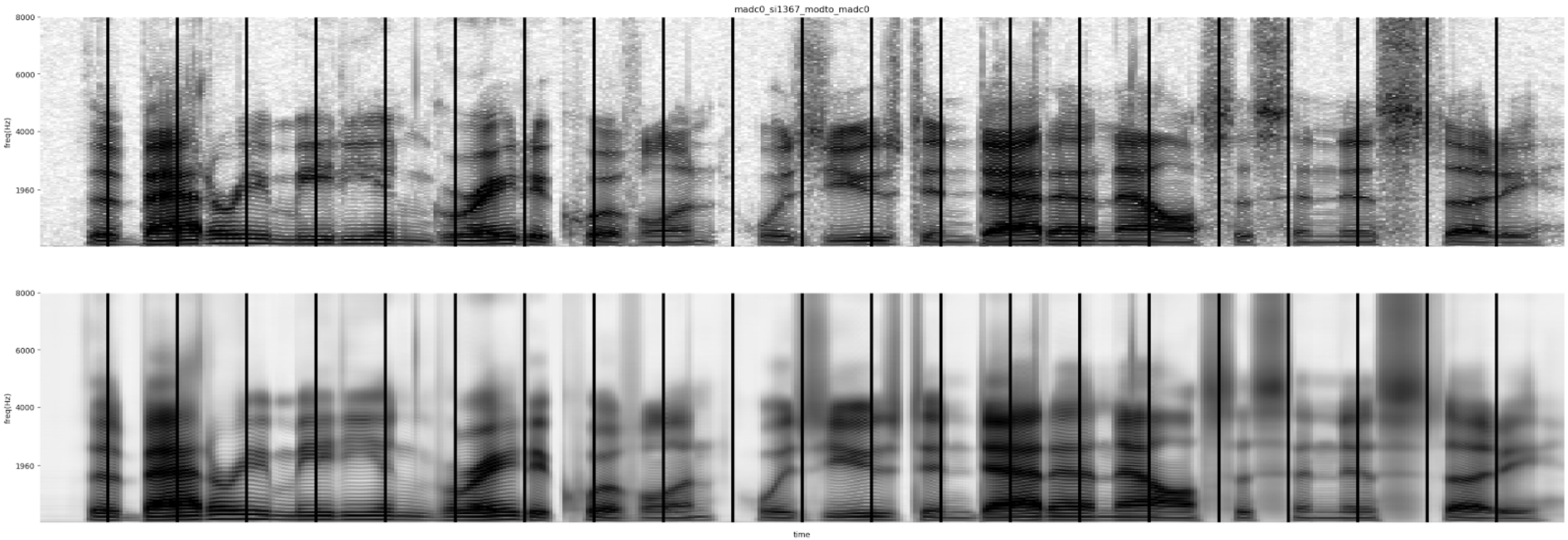
- Modify a male to a female, pitch increases



Modify the Speaker for An Entire Utterance

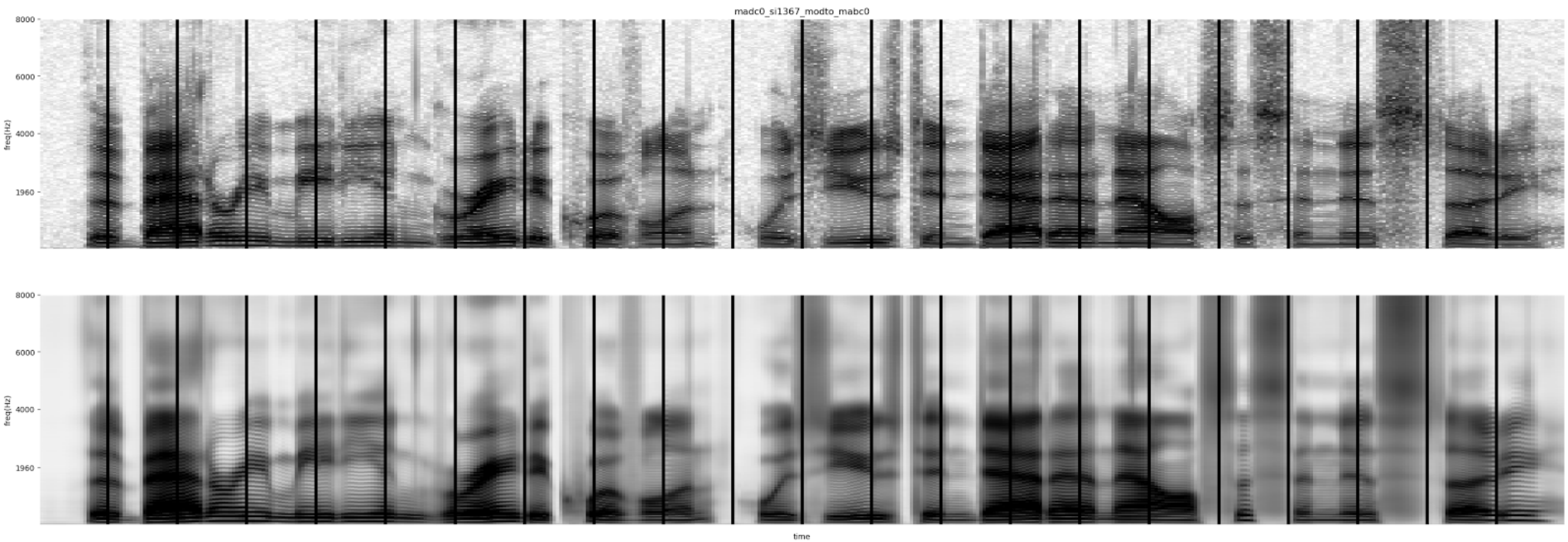
- We choose an utterance from a male speaker (madc0)
  - Modify to another male speaker (mabc0), and a female speaker (fajw0)
- Each speaker has only 8 utterances in the set
  - ~4s/utterances
- Estimate the latent speaker representation using only 30s of speech

Modify the Speaker for An Entire Utterance



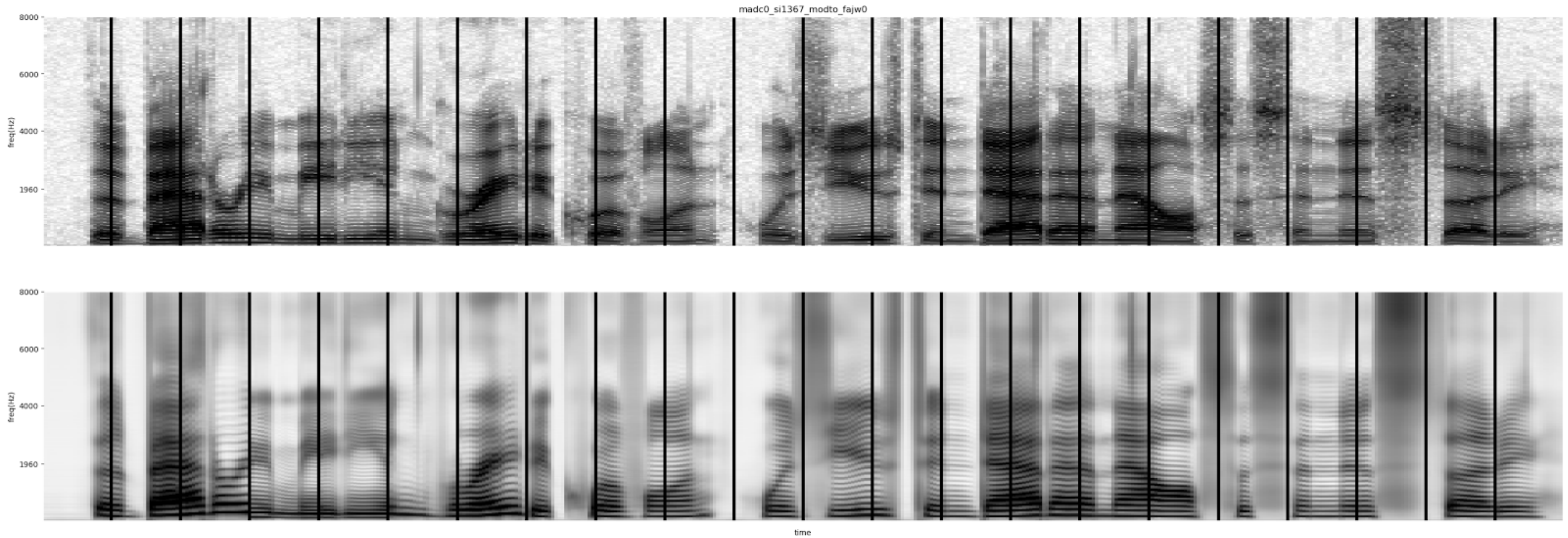
Original Speaker
(top) original spectrogram, (bottom) reconstructed spectrogram

Modify the Speaker for An Entire Utterance



Convert to Speaker mabc0
(top) original spectrogram, (bottom) modified spectrogram

Modify the Speaker for An Entire Utterance



Convert to Speaker fajw0
(top) original spectrogram, (bottom) modified spectrogram

Quantitative Evaluation

- We train discriminators for phone classification and speaker classification
- Posteriors as the quantitative metric
  - Discriminators' mean opinion score on the two attributes
  - Posterior of target attribute increases; posterior of source attribute decreases
  - Posteriors of irrelevant attributes unchanged

Quantitative Evaluation

- We train discriminators for phone classification and speaker classification
- Posteriors as the quantitative metric
  - Discriminators' mean opinion score on the two attributes
  - Posterior of target attribute increases; posterior of source attribute decreases
  - Posteriors of irrelevant attributes unchanged

		/aa/	/ae/	ori. spk.
Modify Phone	before	34.06%	0.45%	50.78%
	after	0.24%	29.73%	41.66%

Quantitative Evaluation

- We train discriminators for phone classification and speaker classification
- Posteriors as the quantitative metric
  - Discriminators' mean opinion score on the two attributes
  - **Posterior of target attribute increases**; posterior of source attribute decreases
  - Posteriors of irrelevant attributes unchanged

		/aa/	/ae/	ori. spk.
Modify Phone	before	34.06%	0.45%	50.78%
	after	0.24%	29.73%	41.66%

Quantitative Evaluation

- We train discriminators for phone classification and speaker classification
- Posteriors as the quantitative metric
  - Discriminators' mean opinion score on the two attributes
  - Posterior of target attribute increases; **posterior of source attribute decreases**
  - Posteriors of irrelevant attributes unchanged

		/aa/	/ae/	ori. spk.
Modify Phone	before	34.06%	0.45%	50.78%
	after	0.24%	29.73%	41.66%

Quantitative Evaluation

- We train discriminators for phone classification and speaker classification
- Posteriors as the quantitative metric
  - Discriminators' mean opinion score on the two attributes
  - Posterior of target attribute increases; posterior of source attribute decreases
  - **Posteriors of irrelevant attributes unchanged**

		/aa/	/ae/	ori. spk.
Modify Phone	before	34.06%	0.45%	50.78%
	after	0.24%	29.73%	41.66%

Quantitative Evaluation

- We train discriminators for phone classification and speaker classification
- Posteriors as the quantitative metric
  - Discriminators' mean opinion score on the two attributes
  - Posterior of target attribute increases; posterior of source attribute decreases
  - Posteriors of irrelevant attributes unchanged

Modify Phone		/aa/	/ae/	ori. spk.
	before	34.06%	0.45%	50.78%
	after	0.24%	29.73%	41.66%
Modify Speaker		falk0	madc0	ori. phone
	before	44.48%	0.02%	54.61%
	after	3.11%	28.71%	48.71%

Quantitative Evaluation

- We train discriminators for phone classification and speaker classification
- Posteriors as the quantitative metric
  - Discriminators' mean opinion score on the two attributes
  - **Posterior of target attribute increases;** posterior of source attribute decreases
  - Posteriors of irrelevant attributes unchanged

Modify Phone		/aa/	/ae/	ori. spk.
	before	34.06%	0.45%	50.78%
	after	0.24%	29.73%	41.66%
Modify Speaker		falk0	madc0	ori. phone
	before	44.48%	0.02%	54.61%
	after	3.11%	28.71%	48.71%

Quantitative Evaluation

- We train discriminators for phone classification and speaker classification
- Posteriors as the quantitative metric
  - Discriminators' mean opinion score on the two attributes
  - Posterior of target attribute increases; **posterior of source attribute decreases**
  - Posteriors of irrelevant attributes unchanged

Modify Phone		/aa/	/ae/	ori. spk.
	before	34.06%	0.45%	50.78%
	after	0.24%	29.73%	41.66%
Modify Speaker		falk0	madc0	ori. phone
	before	44.48%	0.02%	54.61%
	after	3.11%	28.71%	48.71%

Quantitative Evaluation

- We train discriminators for phone classification and speaker classification
- Posteriors as the quantitative metric
  - Discriminators' mean opinion score on the two attributes
  - Posterior of target attribute increases; posterior of source attribute decreases
  - **Posteriors of irrelevant attributes unchanged**

		/aa/	/ae/	ori. spk.
Modify Phone	before	34.06%	0.45%	50.78%
	after	0.24%	29.73%	41.66%
		falk0	madc0	ori. phone
Modify Speaker	before	44.48%	0.02%	54.61%
	after	3.11%	28.71%	48.71%

Outline

1. Motivations
2. Background and Models
3. Latent Attribute Representations and Operations
4. Experiments
5. Conclusion

Conclusion and Future Work

- We present a CNN-VAE to model generation process of speech segments

Conclusion and Future Work

- We present a CNN-VAE to model generation process of speech segments
- The framework leverages vast quantities of unannotated data to learn a general speech analyzer and a general speech synthesizer.

Conclusion and Future Work

- We present a CNN-VAE to model generation process of speech segments
- The framework leverages vast quantities of unannotated data to learn a general speech analyzer and a general speech synthesizer.
- We demonstrate qualitatively and quantitatively the ability to modify speech attributes.

Conclusion and Future Work

- We present a CNN-VAE to model generation process of speech segments
- The framework leverages vast quantities of unannotated data to learn a general speech analyzer and a general speech synthesizer.
- We demonstrate qualitatively and quantitatively the ability to modify speech attributes.
- We have applied the modification operation to data augmentation for ASR and achieved significant improvement for domain adaptation. (submitted to ASRU)

Conclusion and Future Work

- We present a CNN-VAE to model generation process of speech segments
- The framework leverages vast quantities of unannotated data to learn a general speech analyzer and a general speech synthesizer.
- We demonstrate qualitatively and quantitatively the ability to modify speech attributes.
- We have applied the modification operation to data augmentation for ASR and achieved significant improvement for domain adaptation. (submitted to ASRU)
- For future work, we plan to investigate the use of VAE on voice conversion and speech de-noising under the setting of no parallel training data.

Thanks for Listening.
Q&A?

Paper, slides, samples and follow-up works can be found on

http://people.csail.mit.edu/wnhsu/