Paper Reading: HIERARCHICALGENERATIVEMODELING FORCONTROLLABLESPEECHSYNTHESIS

Dong Wang

2021/03/08

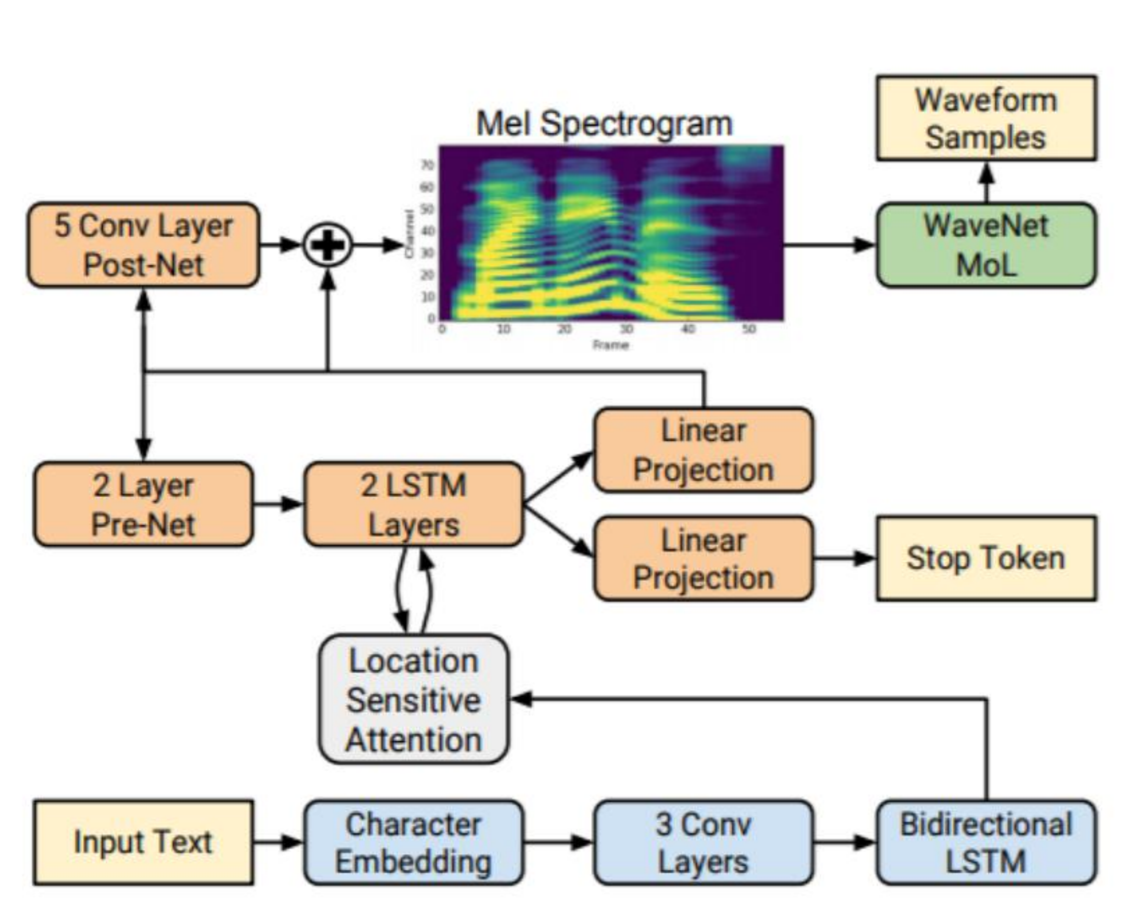
Main question

HIERARCHICAL GENERATIVE MODELING FOR CONTROLLABLE SPEECH SYNTHESIS

Wei-Ning Hsu[1]* Yu Zhang[2] Ron J. Weiss[2] Heiga Zen[2] Yonghui Wu[2] Yuxuan Wang[2]
Yuan Cao[2] Ye Jia[2] Zhifeng Chen[2] Jonathan Shen[2] Patrick Nguyen[2] Ruoming Pang[2]
[1]Massachusetts Institute of Technology [2]Google Inc.
wnhsu@csail.mit.edu, {ngyuzh,ronw}@google.com

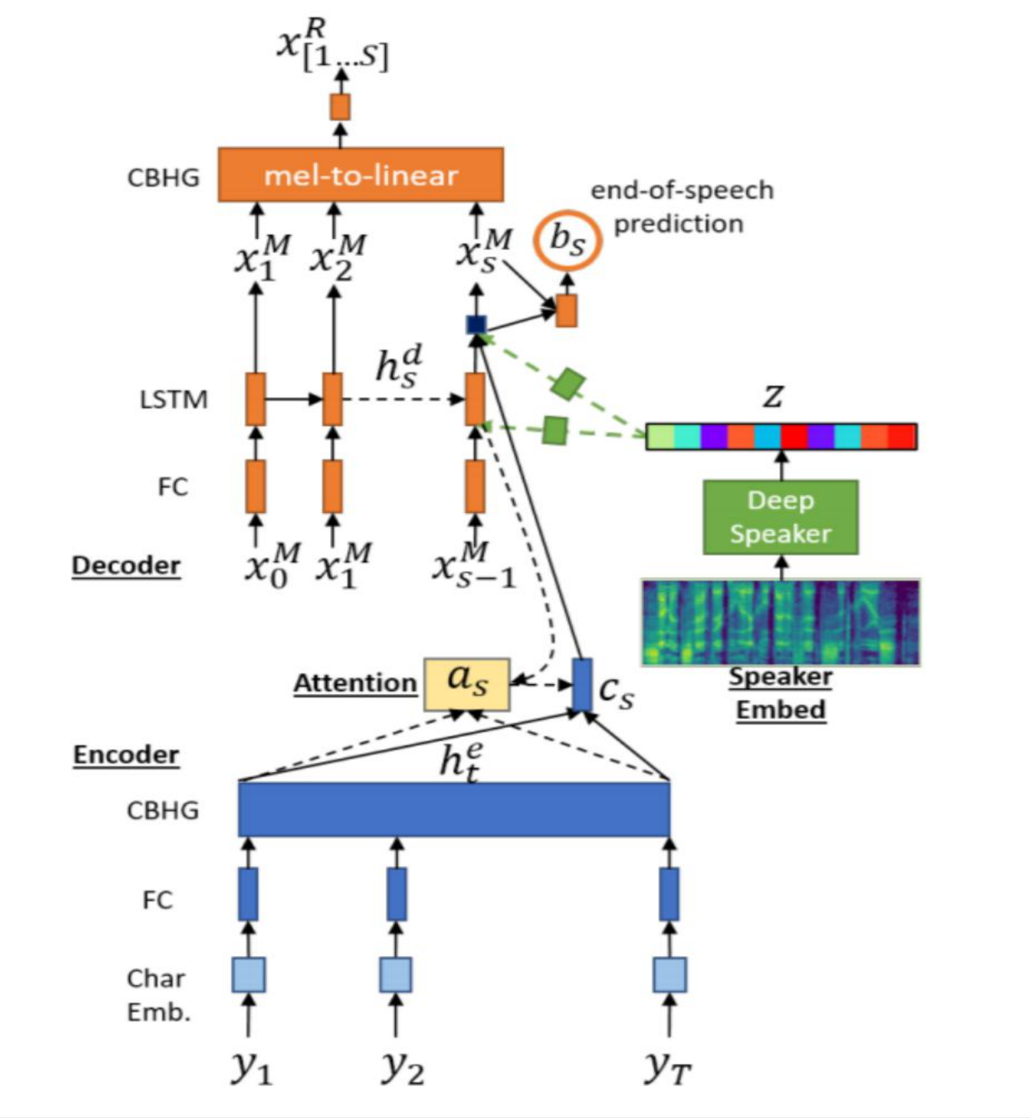
This paper proposes a neural sequence-to-sequence text-to-speech (TTS) model which can control latent attributes in the generated speech that are rarely annotated in the training data, such as speaking style, accent, background noise, and recording conditions. The model is formulated as a conditional generative model based on the variational autoencoder (VAE) framework, with two levels of hierarchical latent variables. The first level is a categorical variable, which represents attribute groups (e.g. clean/noisy) and provides interpretability. The second level, conditioned on the first, is a multivariate Gaussian variable, which characterizes specific attribute configurations (e.g. noise level, speaking rate) and enables disentangled fine-grained control over these attributes. This amounts to using a Gaussian mixture model (GMM) for the latent distribution. Extensive evaluation demonstrates its ability to control the aforementioned attributes. In particular, we train a high-quality controllable TTS model on real found data, which is capable of inferring speaker and style attributes from a noisy utterance and use it to synthesize *clean* speech with controllable speaking style.

Modern speech synthesis



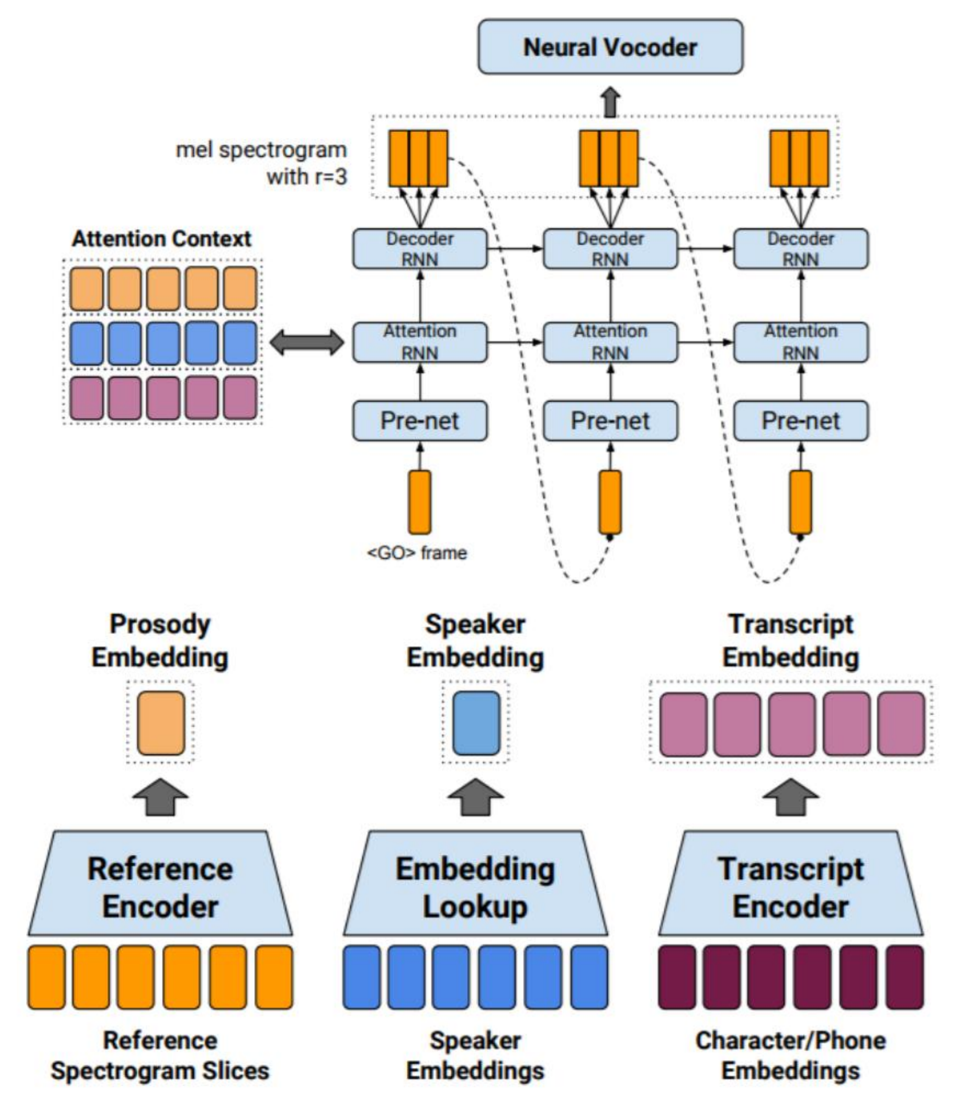
- Yuxuan Wang et al. Tacotron: A fully end-to-end text-to-speech synthesis model. arXivpreprint arXiv:1703.10135, 2017.
- Jonathan Shen et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. arXiv preprint arXiv:1712.05884, 2017.

How to control generation style



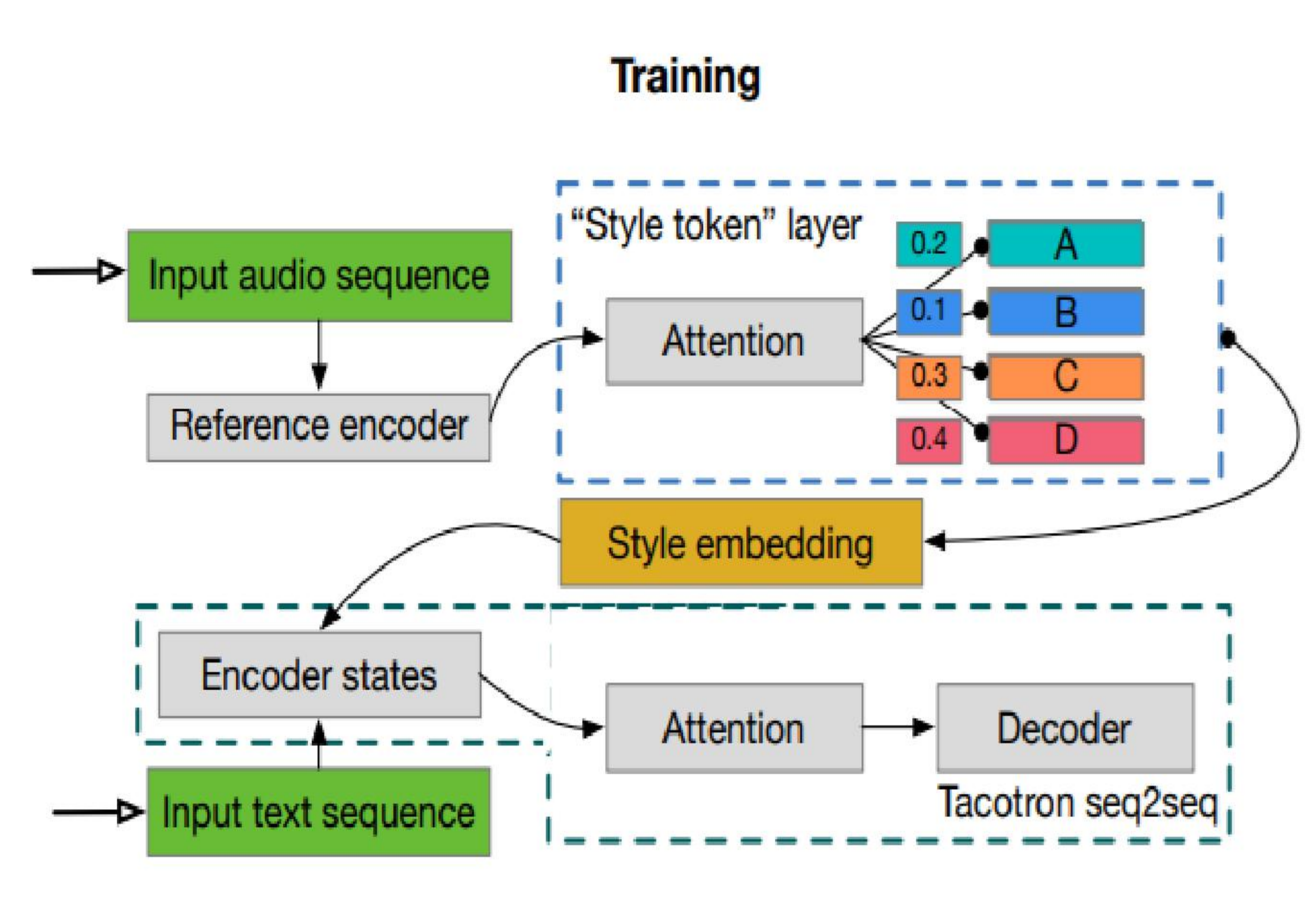
Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. Machine speech chain with one-shot speaker adaptation. arXiv preprint arXiv:1803.10525, 2018.

Control by reference



RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron J Weiss, Rob Clark, and Rif A Saurous. Towards end-to- end prosody transfer for expressive speech synthesis with tacotron. arXiv preprint arXiv:1803.09047, 2018.

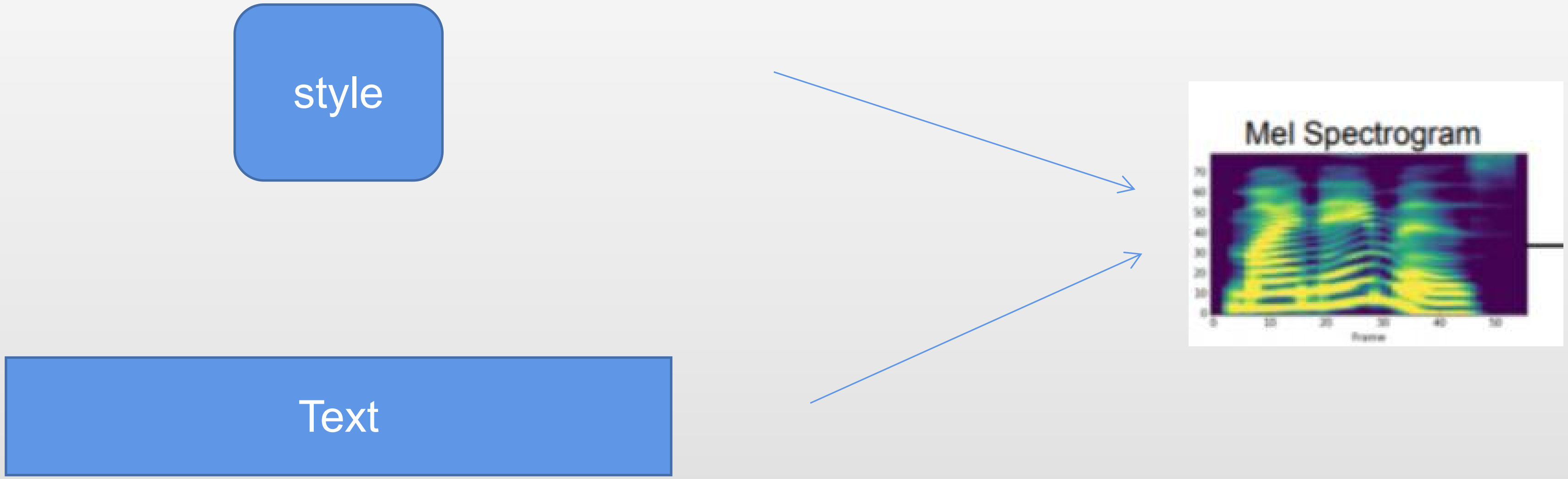
Control by reference



Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ Skerry-Ryan, Eric Battenberg,Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A. Saurous. Style tokens:Unsupervised style modeling, control and transfer in end-to-end speech synthesis. arXiv preprint arXiv:1803.09017, 2018.

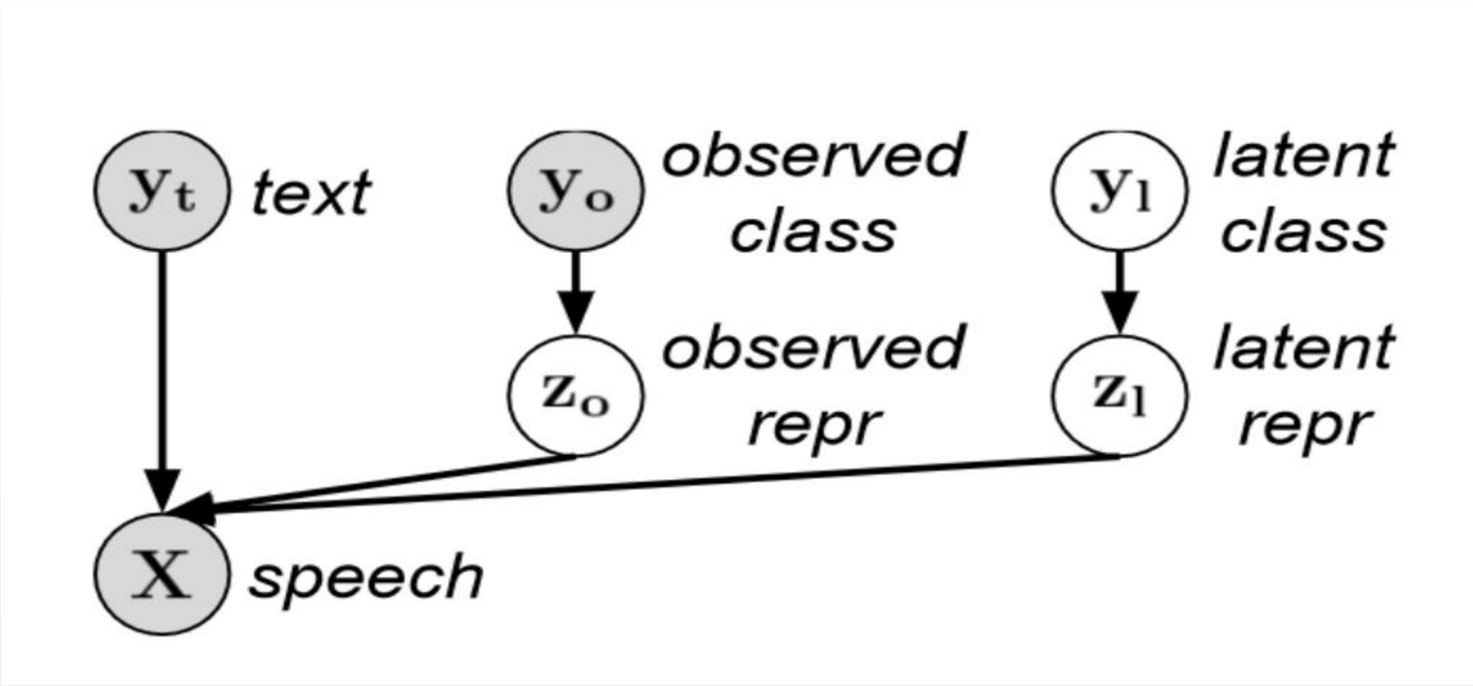
A generative perspective

- Only the text cannot recover the speech signal. A style distribution is required to improve the model.
- The distribution should reflect the true hidden factors.
- Important when using complex datasets



Explict generative modeling

- Treat text as the main input
- Involve speaker embedding as the observed condition
- Model unseen variation as a mixture Gaussian
- Diagonal covariance to encourage disentanglement



Likelihood function

speech

text

$$p(\mathbf{X}, \mathbf{y}_l, \mathbf{z}_l \mid \mathbf{Y}_t, \mathbf{y}_o) = p(\mathbf{X} \mid \mathbf{Y}_t, \mathbf{y}_o, \mathbf{z}_l)\, p(\mathbf{z}_l \mid \mathbf{y}_l)\, p(\mathbf{y}_l).$$

hidden class

hidden continuous variable

speaker

Maximum likelihood by VAE

- Likelihood is not tractable (by marginalizing the lagent variable y_l and z_l), due to the complex decoder
- Using variational approach to approximate the posterior

$$p(\mathbf{y}_l, \mathbf{z}_l \mid \mathbf{X}, \mathbf{Y}_t, \mathbf{y}_o) \approx q(\mathbf{y}_l \mid \mathbf{X})\, q(\mathbf{z}_l \mid \mathbf{X})$$

- $q(y_l|X)$ and $q(z_l|x)$ can be approximated by a Gaussian, using a nueral net encoder

Maximum likelihood by VAE

- Since the y_l and z_l form a Gaussian mixture, it is possible to infer $p(y_l|z_l)$ with known z_l. This makes $q(y_l|X)$ not necessary if we have known $q(z_l|X)$.

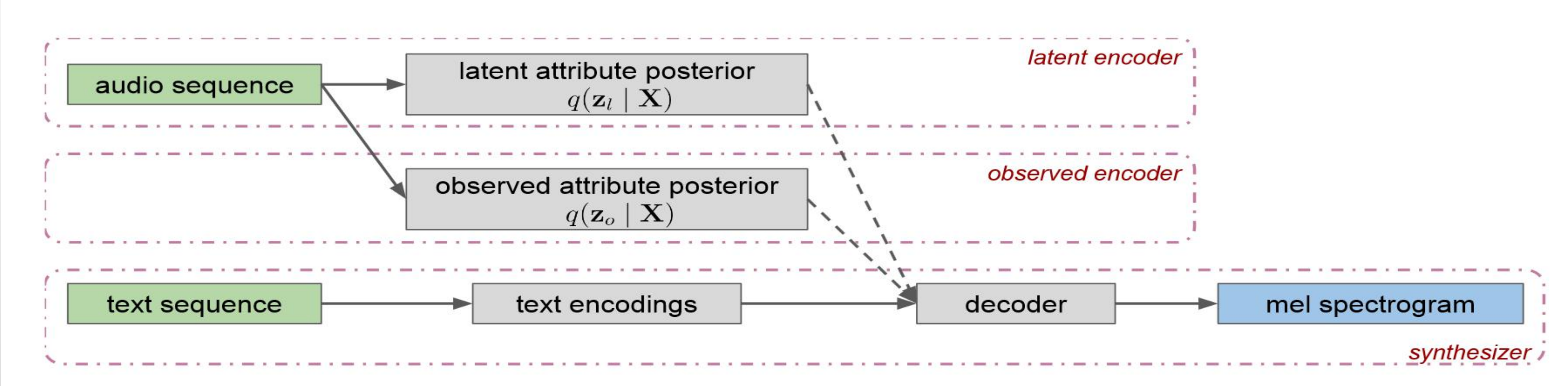
$$p(\mathbf{y}_l|\mathbf{X}) = \int_{\mathbf{z}_l} p(\mathbf{y}_l \mid \mathbf{z}_l)\, p(\mathbf{z}_l|\mathbf{X})\, d\mathbf{z}_l = \mathbb{E}_{p(\mathbf{z}_l|\mathbf{X})}\left[p(\mathbf{y}_l \mid \mathbf{z}_l)\right] \approx \mathbb{E}_{q(\mathbf{z}_l|\mathbf{X})}\left[p(\mathbf{y}_l \mid \mathbf{z}_l)\right] := q(\mathbf{y}_l|\mathbf{X})$$

- ELBO is given:

$$\begin{aligned}\mathcal{L}(p, q; \mathbf{X}, \mathbf{Y}_t, \mathbf{y}_o) = {} & \mathbb{E}_{q(\mathbf{z}_l|\mathbf{X})}[\log p(\mathbf{X} \mid \mathbf{Y}_t, \mathbf{y}_o, \mathbf{z}_l)] \\ & - \mathbb{E}_{q(\mathbf{y}_l|\mathbf{X})}[D_{KL}(q(\mathbf{z}_l \mid \mathbf{X}) \,||\, p(\mathbf{z}_l \mid \mathbf{y}_l))] - D_{KL}(q(\mathbf{y}_l \mid \mathbf{X}) \,||\, p(\mathbf{y}_l))\end{aligned}$$

Involing latent variables related to speaker

$$\mathcal{L}_o(p,q;\mathbf{X},\mathbf{Y}_t,\mathbf{y}_o) = \mathbb{E}_{\boxed{q(\mathbf{z}_o|\mathbf{X})}q(\mathbf{z}_l|\mathbf{X})}[\log p(\mathbf{X} \mid \mathbf{Y}_t, \boxed{\mathbf{z}_o}, \mathbf{z}_l)] - \boxed{D_{KL}(q(\mathbf{z}_o \mid \mathbf{X}) \,||\, p(\mathbf{z}_o \mid \mathbf{y}_o))}$$
$$- \mathbb{E}_{q(\mathbf{y}_l|\mathbf{X})}[D_{KL}(q(\mathbf{z}_l \mid \mathbf{X}) \,||\, p(\mathbf{z}_l \mid \mathbf{y}_l))] - D_{KL}(q(\mathbf{y}_l \mid \mathbf{X}) \,||\, p(\mathbf{y}_l)).$$

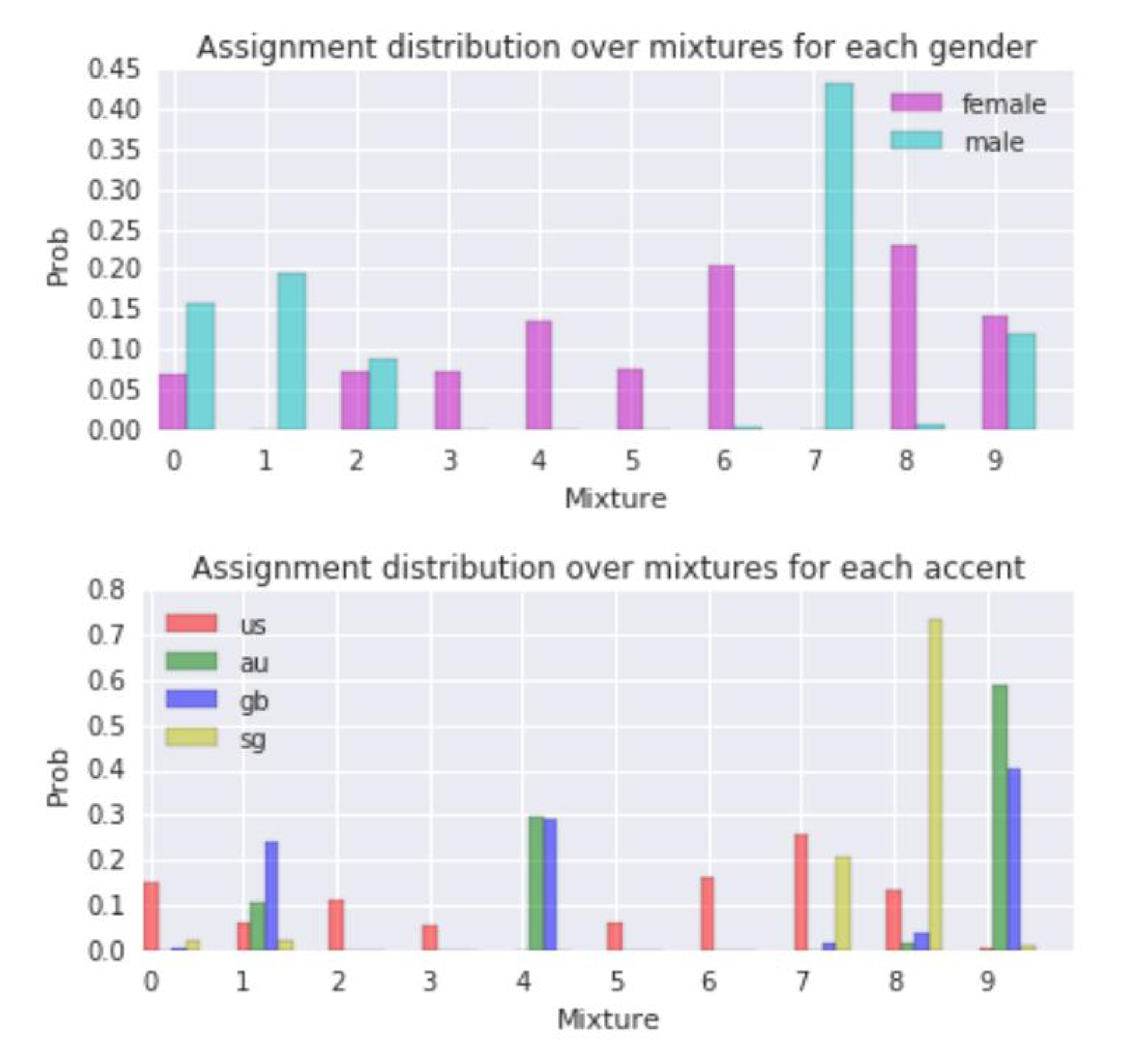


Experiment setting

	multi-spk (Section 4.1)	noisy-multi-spk (Section 4.2)	audiobooks (Section 4.3)	crowd-sourced (Section 4.4)
$\dim(\mathbf{y}_l)$	10	10	10	10
$\dim(\mathbf{z}_l)$	16	16	16	16
initial $\boldsymbol{\sigma}_l$	e^0	e^{-1}	e^{-1}	e^{-1}
minimum $\boldsymbol{\sigma}_l$	e^{-1}	e^{-2}	e^{-2}	e^{-2}
$\dim(\mathbf{y}_o)$	N/A	84	N/A	1,172
$\dim(\mathbf{z}_o)$	N/A	N/A	N/A	16
initial $\boldsymbol{\sigma}_o$	N/A	N/A	N/A	e^{-2}
minimum $\boldsymbol{\sigma}_o$	N/A	N/A	N/A	e^{-4}

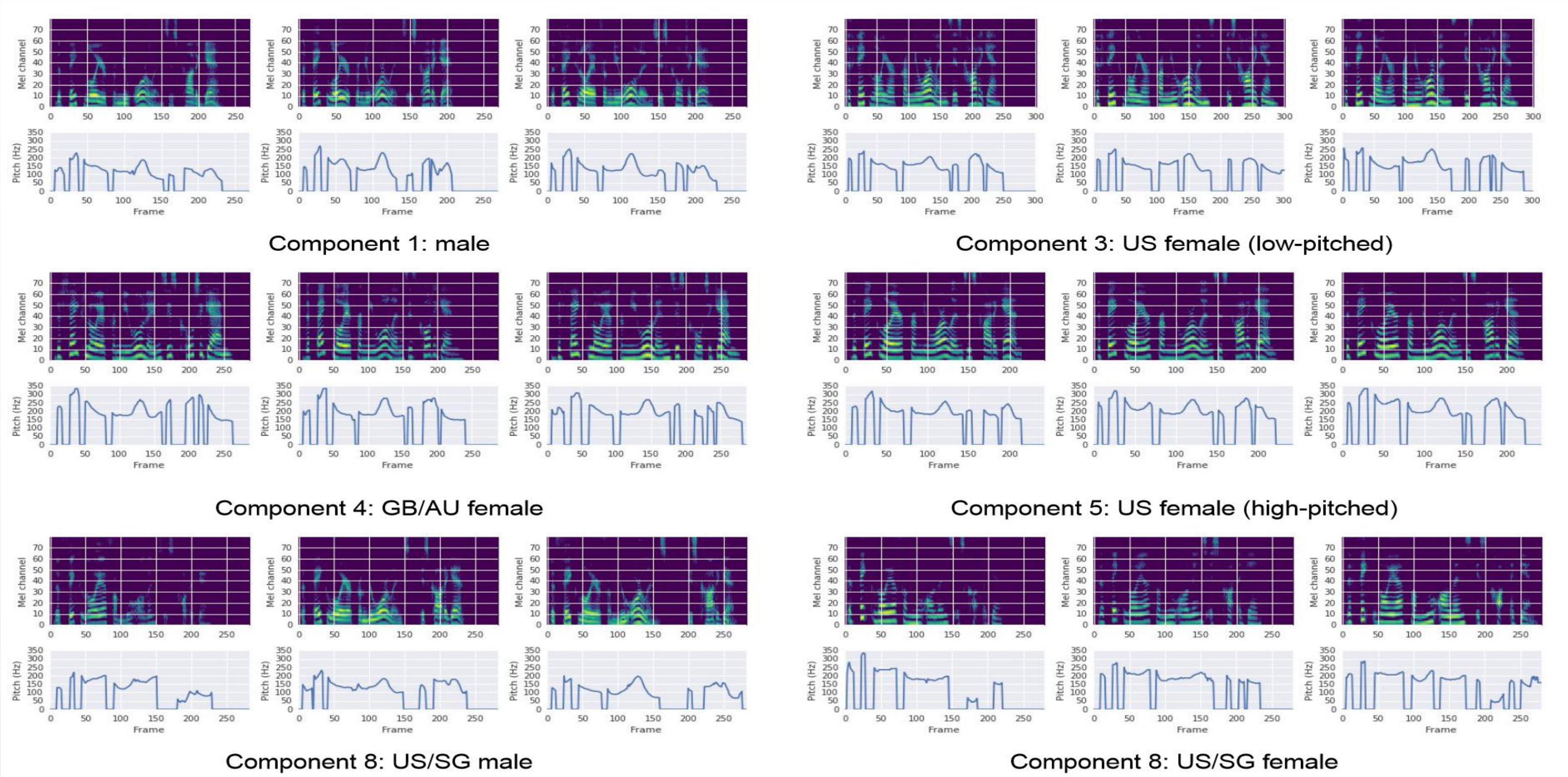
Experiments 1

- 84 English speakers with different accents
- Assign to the largest $q(y_l|X)$.
- Look at the distribution of gender and accent within each mixture
- Most mixtures represent one gender, and a few accents



Experiments 1

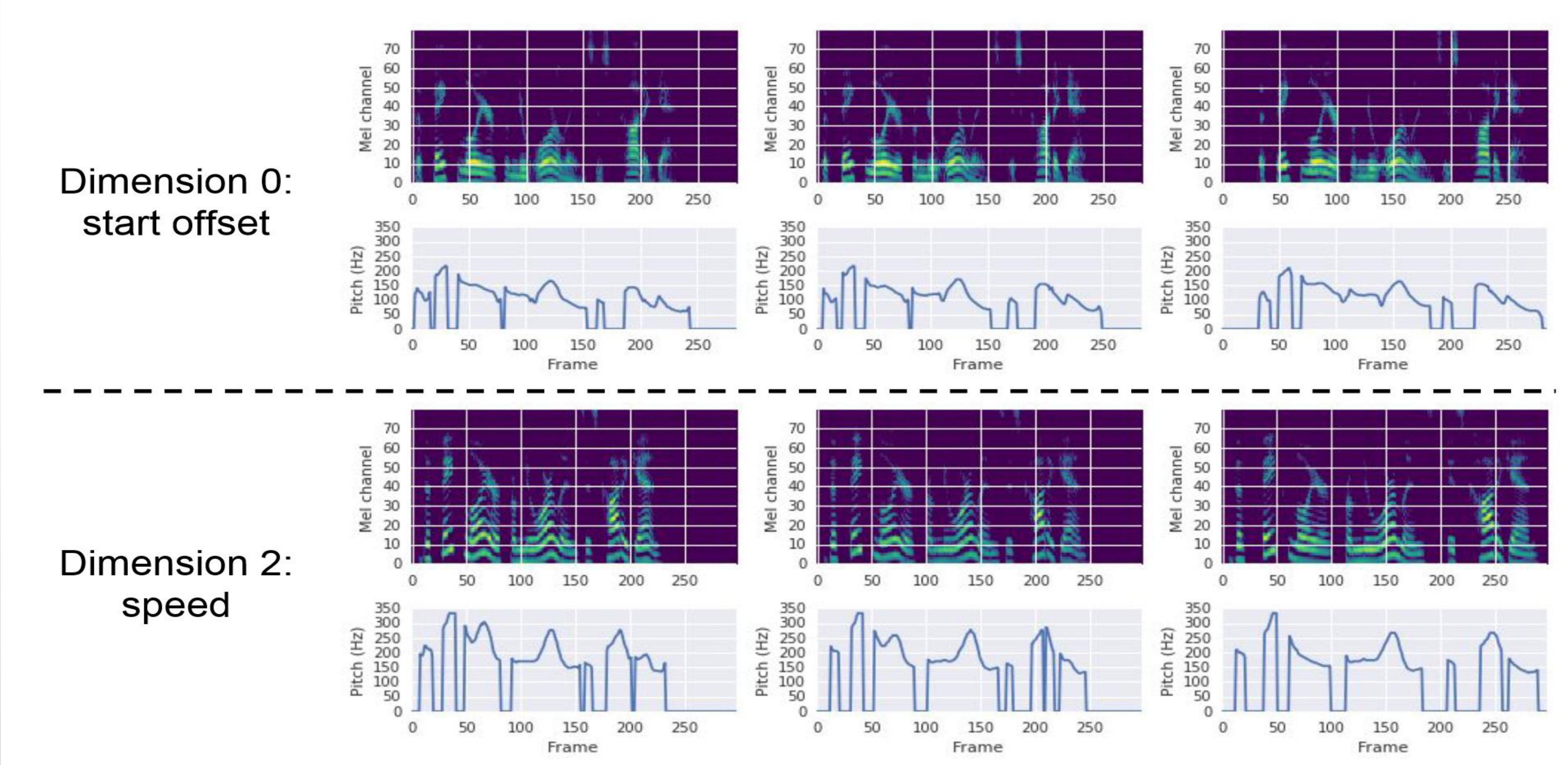
- Random samples by mixture components



https://google.github.io/tacotron/publications/gmvae_controllable_tts

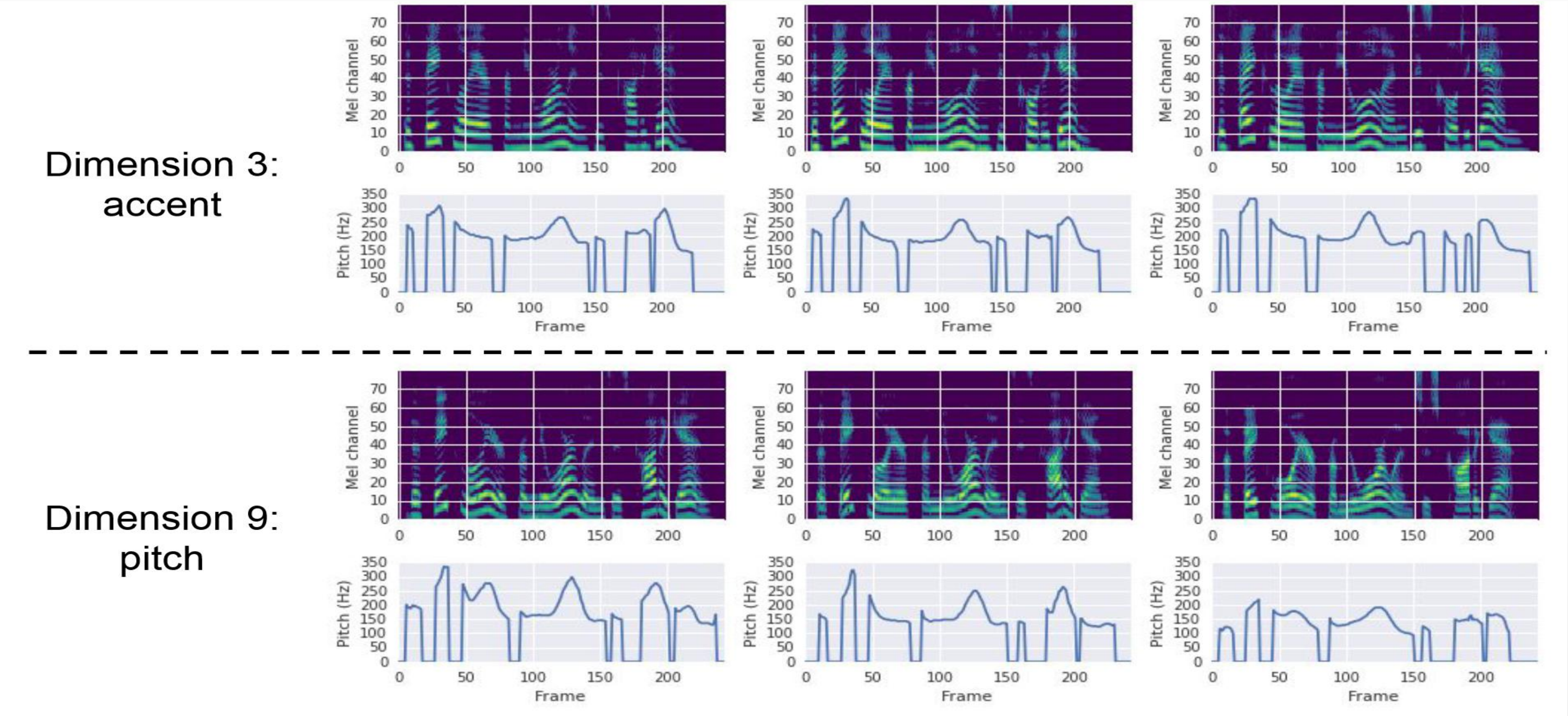
Experiment 1

- Different dimensions control different characters



Experiment 1

- Different dimensions control different characters



Experiment 1

- Classification using the latent variable

	Gender	Accent	Speaker Identity
Train	100.00	98.76	97.66
Eval	98.72	98.72	95.39

Experiment 2

- Using noisy data to train model, and then generate clean speech.
- Design 8 clusters clusters (speakers are known), some clusters will represent clean and others represent noisy.

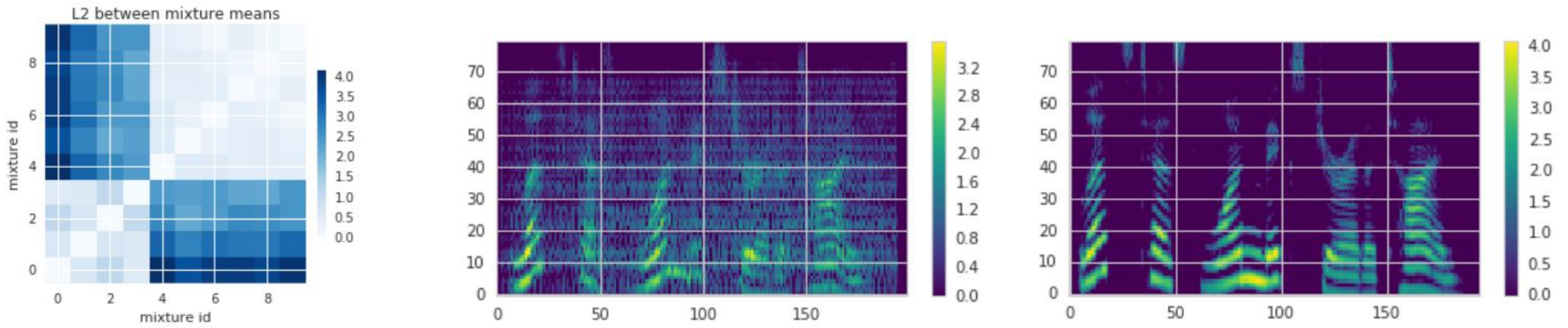


Figure 4: Left: Euclidean distance between the means of each mixture component pair. Right: Decoding the same text conditioned on the mean of a noisy (center) and a clean component (right).

Experiment 2

- Using LDA to find the noise-related dimension and the use the dimension to control the noise level.

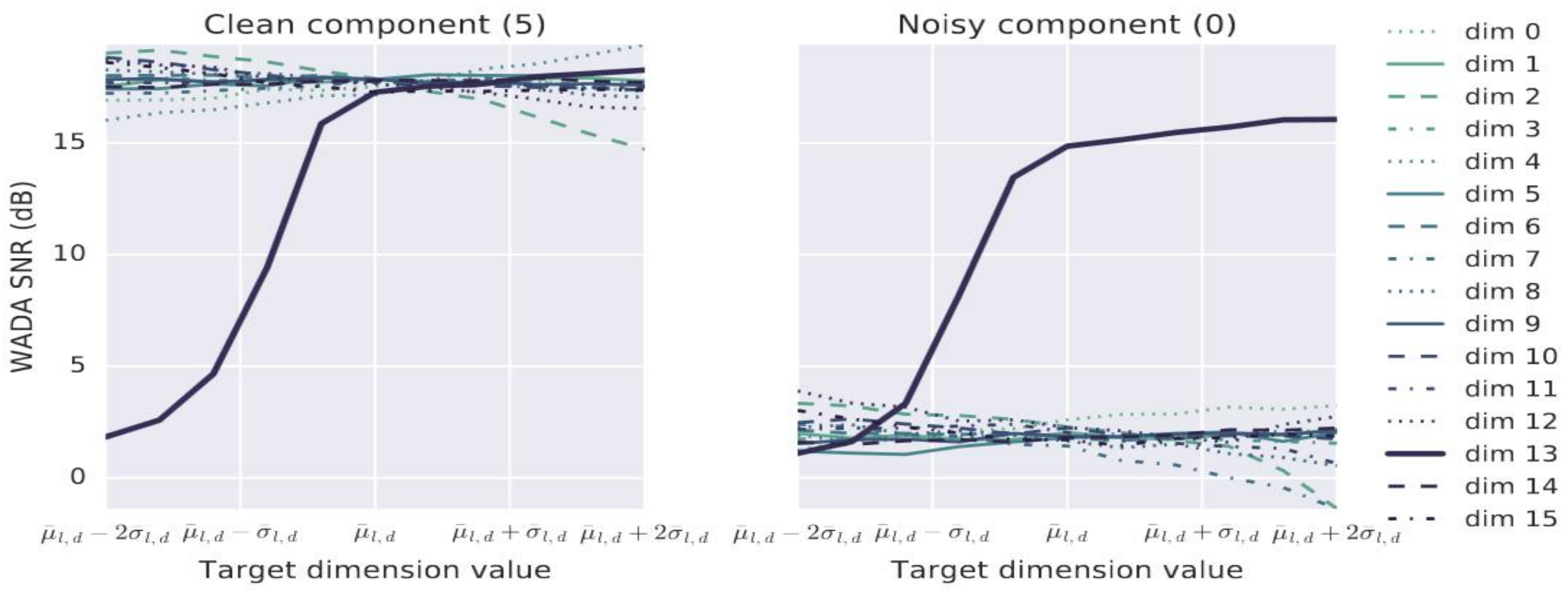
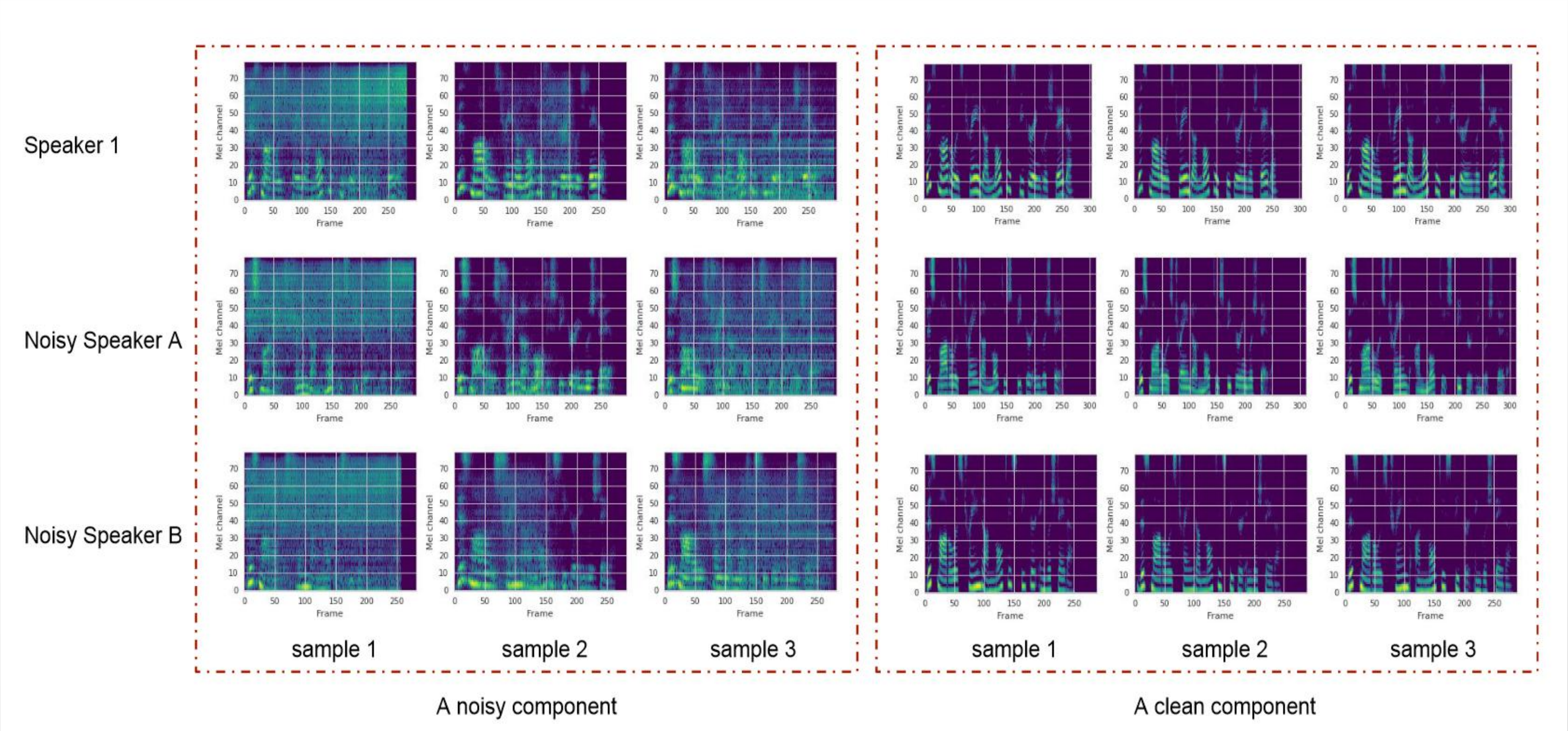
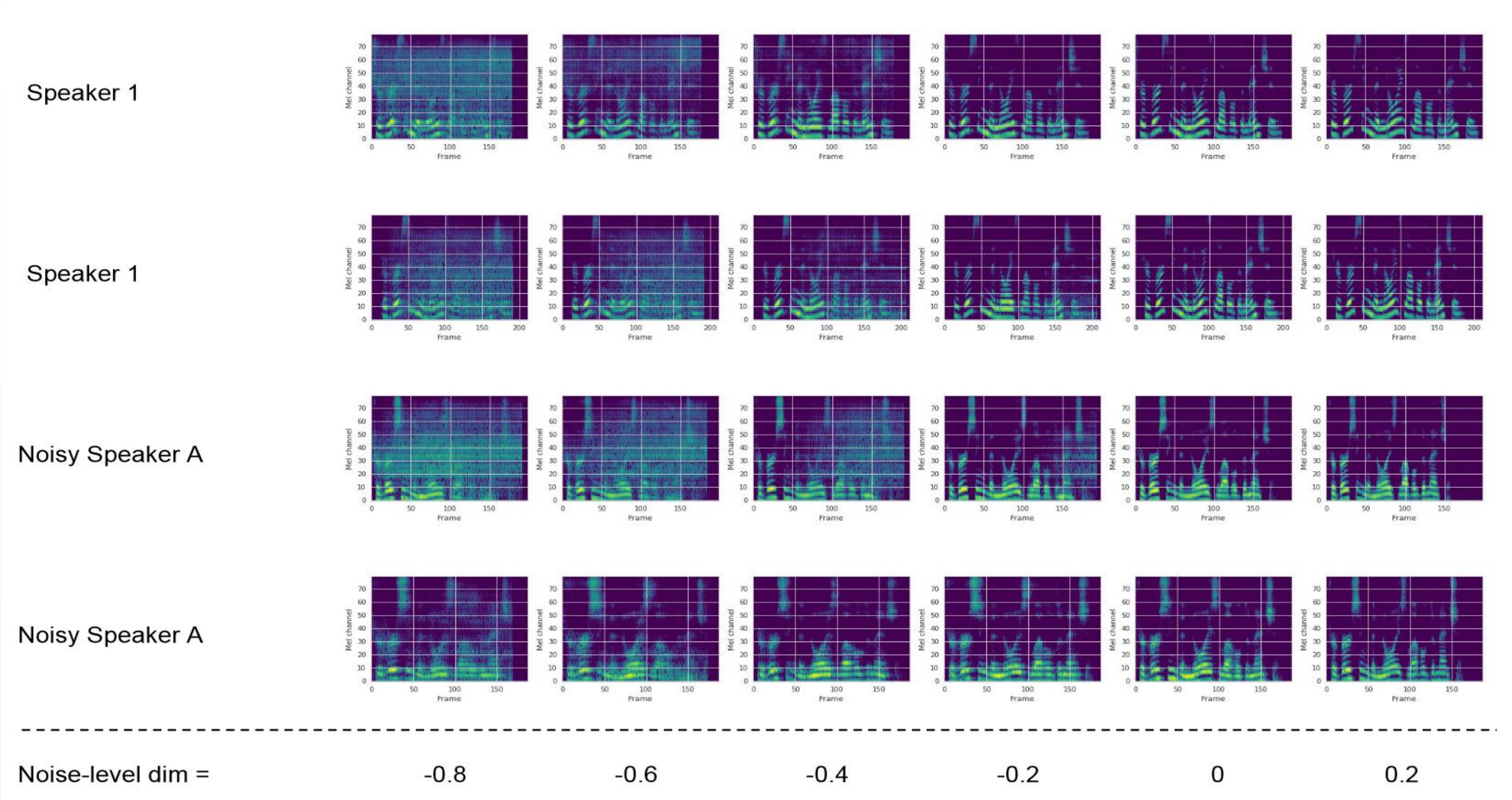


Figure 5: SNR as a function of the value in each latent dimension, comparing clean (left) and noisy (right) components.

Experiment 2



Experiment 2

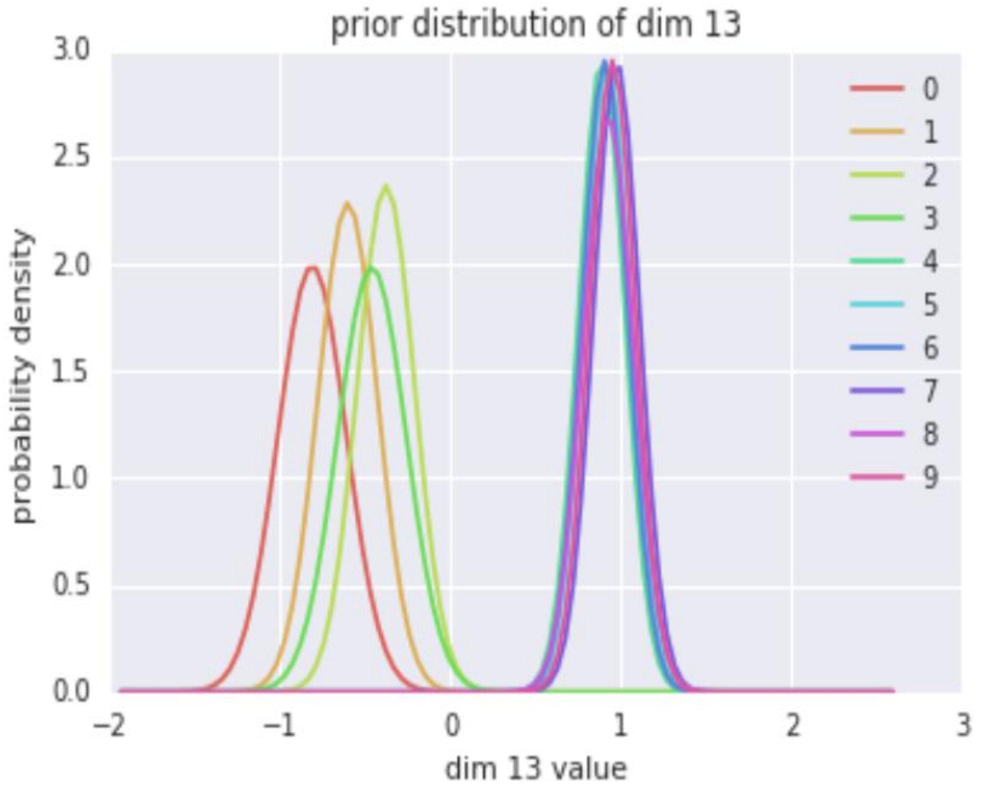


Experiment 2

- Synthesizing for speaker with only noisy training data

Table 1: MOS and SNR comparison among clean original audio, baseline, GST, VAE, and GMVAE models.

Model	MOS	SNR
Original	4.48 ± 0.04	17.71
Baseline	2.87 ± 0.25	11.56
GST	3.32 ± 0.13	14.43
VAE	3.55 ± 0.17	12.91
GMVAE	$\mathbf{4.25 \pm 0.13}$	**17.20**



Experiment 3

- A single speaker US English audiobook dataset of 147 hours, recorded by professional speaker

Table 2: MOS comparison of the original audio, baseline and GMVAE.

Model	MOS
Original	4.67 ± 0.04
Baseline	4.29 ± 0.11
Proposed	$\mathbf{4.67 \pm 0.07}$

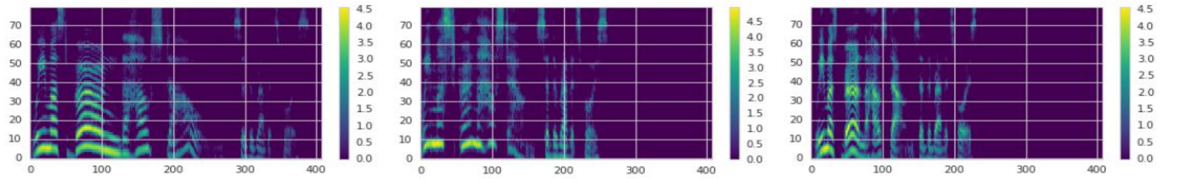


Figure 6: Mel-spectrograms of three samples with the same text, *"We must burn the house down! said the Rabbit's voice."* drawn from the proposed model, showing variation in speed, F_0, and pause duration.

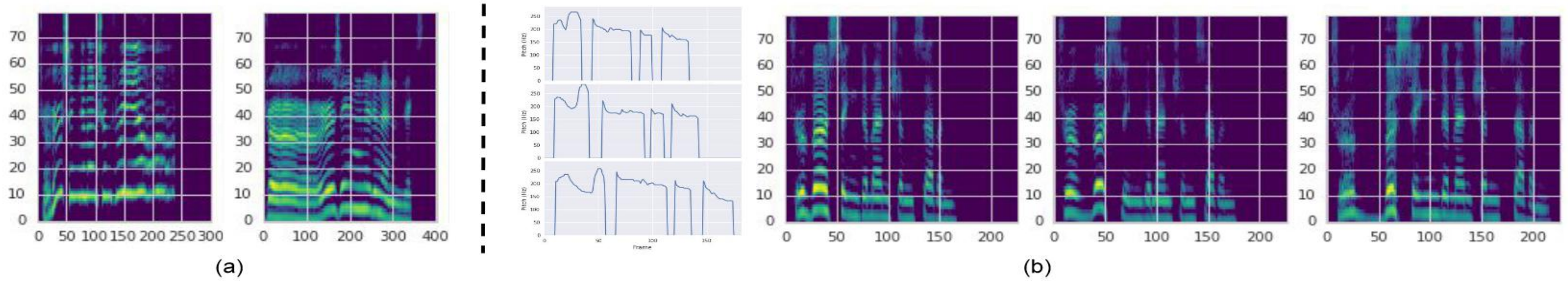


Figure 7: (a) Mel-spectrograms of two unnatural GST samples when setting the weight for one token -0.1: first with tremolo at the end, and second with abnormally long duration for the first syllable. (b) F_0 tracks and spectrograms from GMVAE-Tacotron using different values for the "speed" dimension.

Experiment 4

- Audioset dataset, with thounsands of speakers
- Using cluster mean or dimension to perform clean speech synthesis

Table 3: SNR of original audio, baseline, and the proposed models with different conditioned $\mathbf{z}_l$, on different speakers.

Set	Original	Baseline	Proposed mean	Proposed latent	Proposed latent-dn
SC	18.61	14.33	15.90	16.28	**17.94**
SN	11.80	9.69	15.82	6.78	**18.94**
UC	20.39	N/A	15.70	16.40	**18.83**
UN	10.92	N/A	15.27	4.81	**16.89**

Table 4: Subjective preference (%) between baseline and proposed model with denoised $\mathbf{z}_l$ on the set of "seen noisy" (SN) speakers.

Baseline	Neutral	Proposed
4.0	10.5	**85.5**

Experiment 4

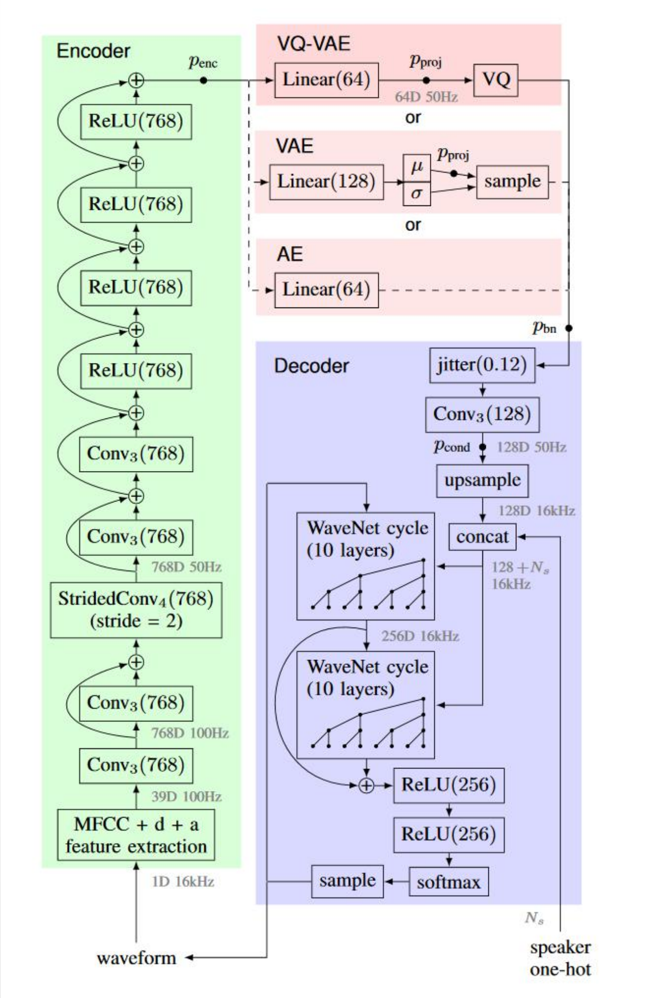
Table 5: Naturalness MOS of original audio, baseline, and proposed model with the clean component mean.

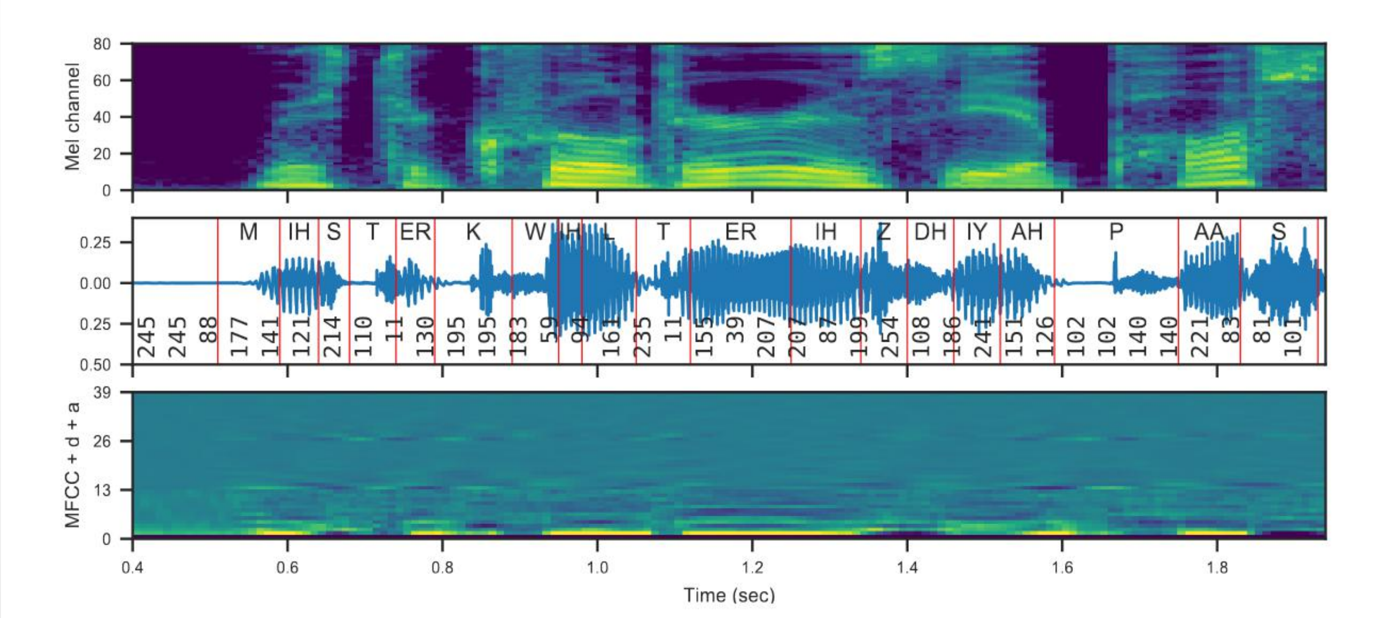
Set	Model	MOS
SC	Original	4.60 ± 0.07
	Baseline	4.17 ± 0.07
	Proposed	4.18 ± 0.06
SN	Original	4.45 ± 0.08
	Baseline	3.64 ± 0.10
	+ denoise	3.84 ± 0.10
	Proposed	$\mathbf{4.09} \pm \mathbf{0.08}$
UC	Original	4.54 ± 0.08
	d-vector	4.10 ± 0.06
	Proposed	$\mathbf{4.26} \pm \mathbf{0.05}$
UN	Original	4.34 ± 0.07
	d-vector	3.76 ± 0.12
	Proposed	$\mathbf{4.20} \pm \mathbf{0.08}$

Table 6: Speaker similarity MOS.

Set	Model	MOS
SC	Baseline	3.54 ± 0.09
	Proposed	3.60 ± 0.09
SN	Original (different channels)	3.30 ± 0.27
	Baseline	$\mathbf{3.83} \pm \mathbf{0.08}$
	Baseline + denoise	3.23 ± 0.20
	Proposed	3.11 ± 0.08
UC	*d*-vector	2.23 ± 0.08
	d-vector (large)	$\mathbf{3.03} \pm \mathbf{0.09}$
	Proposed	2.79 ± 0.08

Additional: Learning short-time feature





Chorowski J, Weiss R J, Bengio S, et al. Unsupervised speech representation learning using wavenet autoencoders[J]. IEEE/ACM transactions on audio, speech, and language processing, 2019, 27(12): 2041-2053.

Conclusions

- It is possible to design a generative model and train it following the ML property.
- An VAE architecture can be used to perform the ML training and infer the latent variables.
- Defining latent distribution by GMM seems a good choice.
- An interesting trend that merges speech recognition, speaker recognition and speech synthesis.
- An interesting way of dealing with data explosion.
- An interesting way of dealing with problems like speech enhancement.