Learning Disentangled Phone and Speaker Representations in a Semi-Supervised VQ-VAE Paradigm

Jennifer Williams, Yi Zhao, Erica Cooper, Junichi Yamagishi

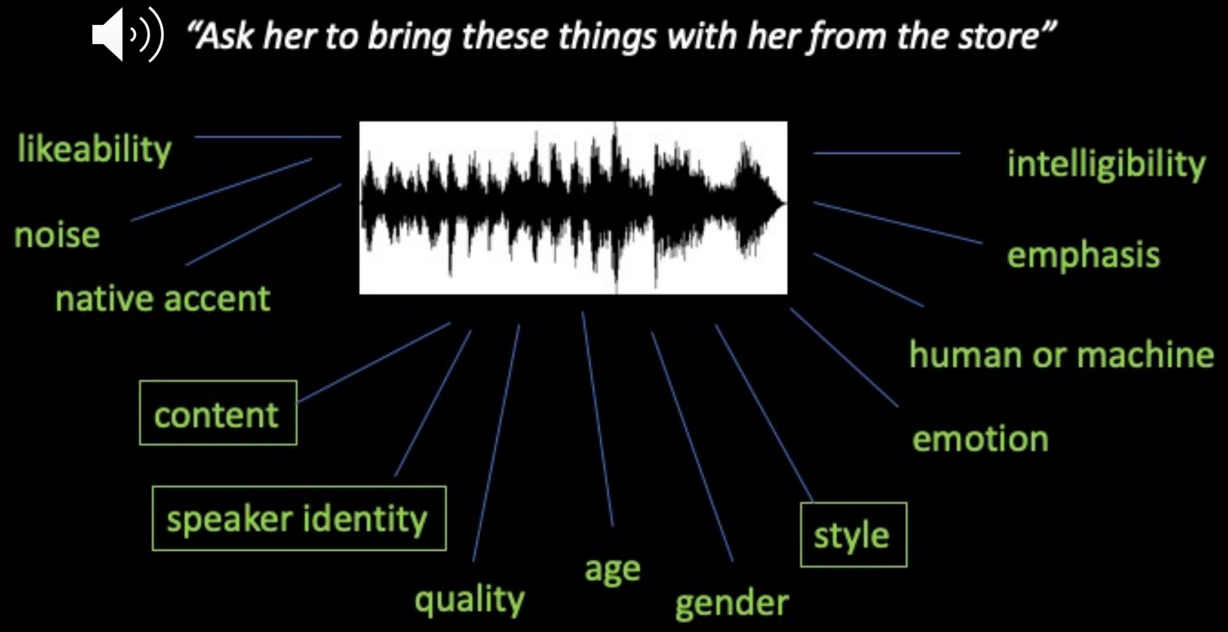
ICASSP 2021

May, 2021

Outline

- Motivation
- Related Work
- VQ-VAE Variants
- Phone/Speaker Disentanglement
- Conclusion & Future Work

Multiple Informational Factors Are Contained in the Speech Signal



- ❑ Traditional representations of speaker identity contain extra information
- ❑ Different kinds of representations are useful for different kinds of speech tasks
- ❑ No end-to-end solutions exist that effectively factorize this information, while also retaining information (and not discard or remove it)

Outline

- Motivation
- Related Work
- VQ-VAE Variants
- Phone/Speaker Disentanglement
- Conclusion & Future Work

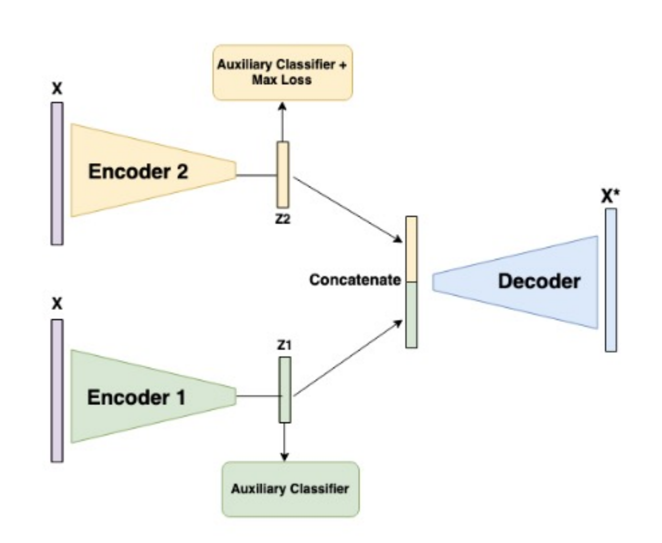
Related Work – Speaker Representations

x-vector

Speaker ID

session ID, gender, speaking rate, transcription, words, phonemes, utterance length, augmentation type

Autoencoder Disentanglement



Speaker ID
Speaking Style
Emotion

Not end-to-end
Worked for style not speaker
Requires labeled data
Evaluation by classification

1) Jennifer Williams and Simon King, “Disentangling Style Factors from Speaker Representations,” Proc. Interspeech 2019, pp. 3945–3949, 2019
2) Desh Raj, David Snyder, Daniel Povey, and Sanjeev Khudanpur, “Probing the Information Encoded in X-Vectors,” in2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2019, pp.726–733
3) Raghuveer Peri, Haoqi Li, Krishna Somandepalli, Arindam Jati, and Shrikanth Narayanan, “An Empirical Analysis of Information Encoded in Disentangled Neural Speaker Representations,” in Proc. Odyssey2020 The Speaker and Language Recognition Work-shop, 2020, pp. 194–201
4) Yi Zhao, Haoyu Li, Cheng-I Lai, Jennifer Williams, Er-ica Cooper, and Junichi Yamagishi, “Improved Prosodyfrom Learned F0 Codebook Representations for VQ-VAE Speech Waveform Reconstruction,”Proc. Inter-speech, 2020

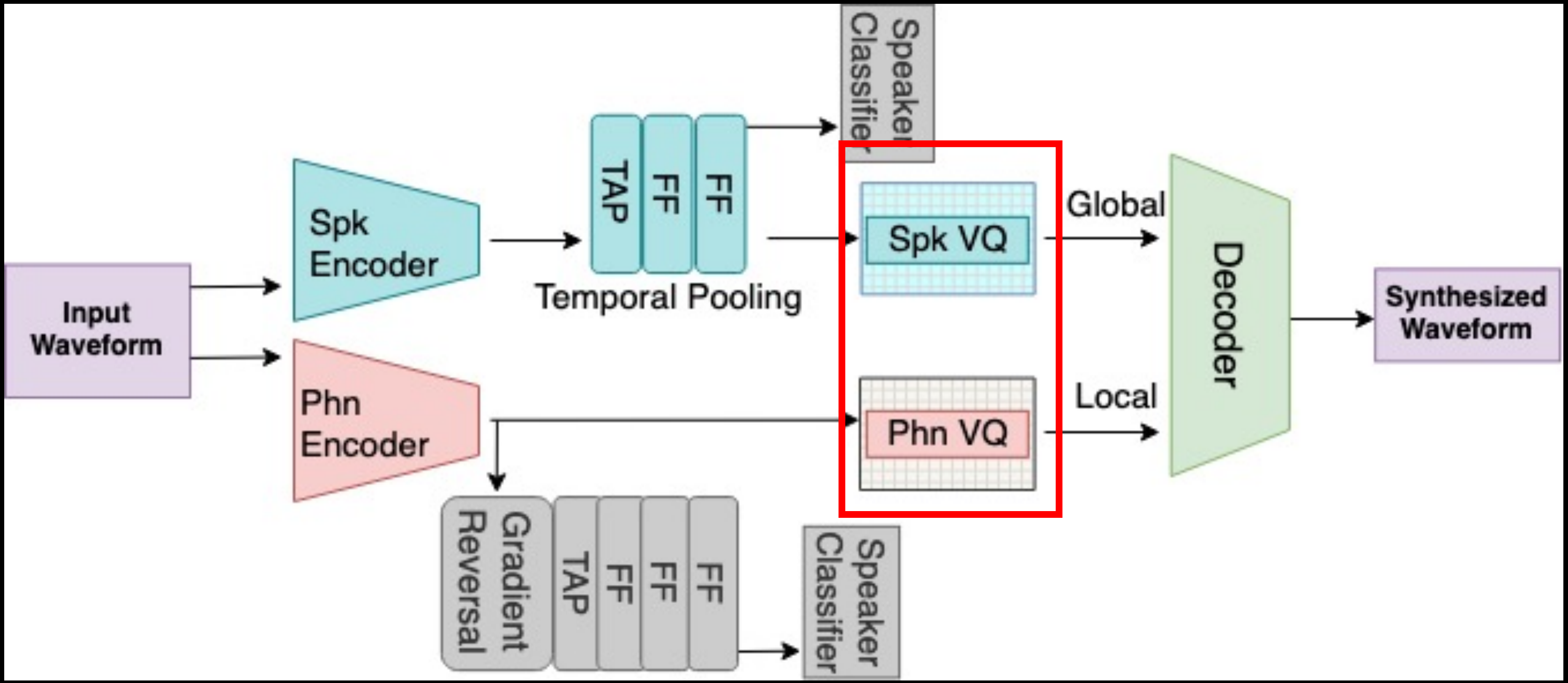
Outline

- Motivation
- Related Work
- VQ-VAE Variants
- Phone/Speaker Disentanglement
- Conclusion & Future Work

Proposed Methodology and Approach

- Stack multiple encoders to learn different representations
- Learn speaker and content
- Use VQ-VAE codebooks (discrete indices, continuous vector-space)
- Use neural vocoder to synthesize speech from discrete codes
- Result is separate disentangled representations
- Explore training methods (self-supervised, semi-supervised)
- Evaluate in meaningful speech tasks (diarization, phone recognition)

Example System Using Stacked VQ-VAE Encoders

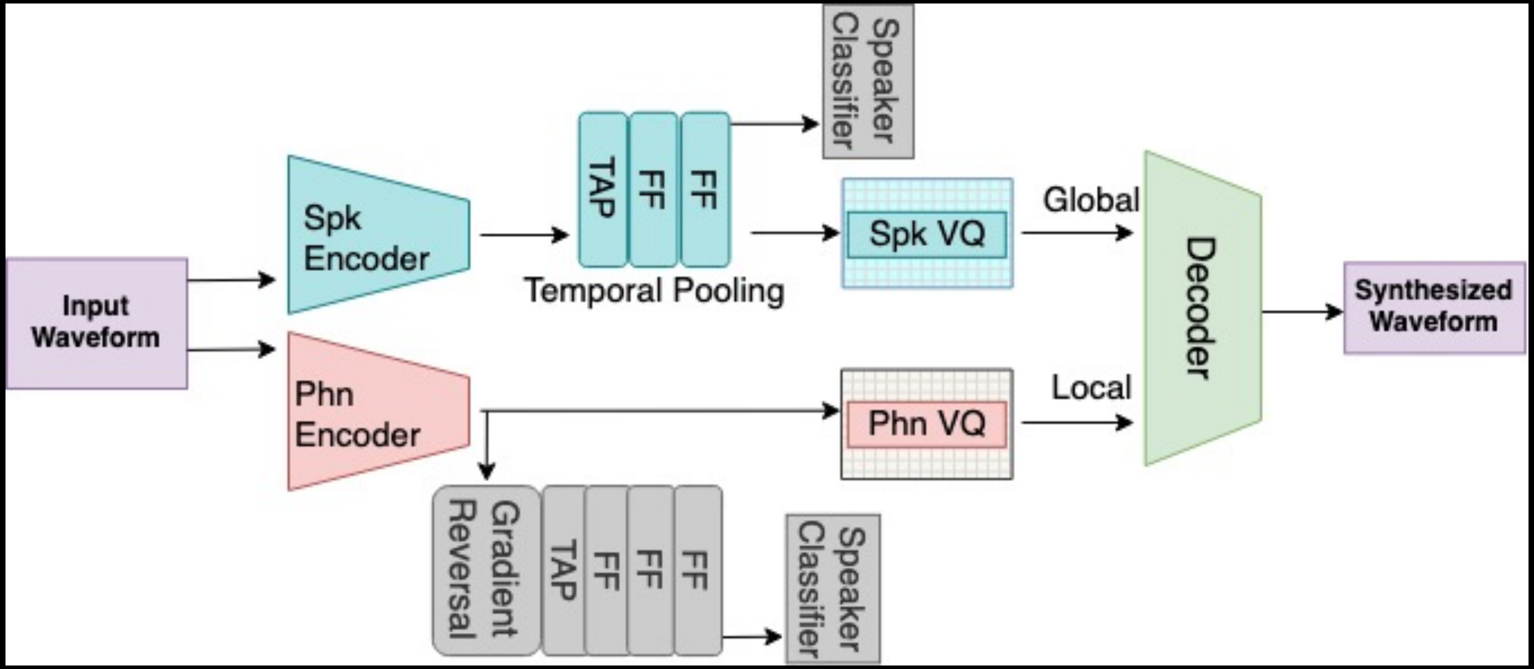


Spk VQ
Speaker Identity

Phn VQ
Speech Content

Speaker VQ learns global conditions with temporal average pooling layer (TAP) and optional speaker classifier. Phone VQ provides local conditions and optional adversarial speaker classifier

Overview of Proposed Systems for Experimentation



Original VQ-VAE loss

$$L = L_R + \alpha L_{VQ} + \beta L_C$$

Modified VQ-VAE loss

$$L = L_R + \alpha(L_{VQl} + L_{Cl}) + \beta(L_{VQg} + L_{Cg})$$

System 1: Original VQ-VAE, self-supervised, only phone codebook
System 2: VQ-VAE, self-supervised, global conditioning
System 3: VQ-VAE, semi-supervised w/ speaker labels,
System 4: VQ-VAE, semi-supervised w/ speaker labels + gradient reversal

Data, Training, and Testing Conditions

Data
VCTK v0.92 English (studio-quality)
110 speakers
16 kHz sample rate
Some overlapping text content among speakers

4 Testing Conditions
Condition 1: seen speakers, seen texts (easiest)
Condition 2: seen speakers, unseen texts
Condition 3: unseen speakers, seen texts
Condition 4: unseen speakers, unseen texts

Training
Warm-up model: original VQ-VAE (system1) trained to 800k steps
Dual-encoder models: trained to an additional 800k steps

Fine-tuning on TIMIT
462/168 speakers in train/test split
Freeze all speaker encoder components
Trained to an additional 400k steps

Automatic Assessment of Synthesis Quality

Table 1: Speech synthesis quality estimation for four testing conditions on VCTK data. (S: softmax, AS: angular-softmax).

		Estimated MOS					Speaker Similarity					Intelligibility (WER)				
Method		C1	C2	C3	C4	Avg	C1	C2	C3	C4	Avg	C1	C2	C3	C4	Avg
Natural Speech	–	4.1	3.9	3.8	3.6	3.8	–	–	–	–	–	9.0	10.6	8.4	8.2	9.0
VQ-VAE	–	3.6	3.5	3.7	3.3	3.5	0.94	0.94	0.33	0.46	0.66	40.4	49.1	85.5	87.9	65.6
+ Global VQ	–	2.3	2.4	2.1	2.2	2.2	0.42	0.45	0.54	0.59	0.50	83.8	82.4	87.7	74.5	82.1
+ Speaker label	S	4.1	3.9	3.8	3.8	3.9	**0.89**	**0.88**	**0.91**	**0.91**	**0.89**	25.8	40.2	27.7	30.8	31.1
	AS	3.9	3.8	3.7	3.7	3.7	0.87	0.86	0.87	0.87	0.87	30.4	42.3	30.3	29.2	33.5
+ Adversarial loss	S	**4.1**	**3.9**	**3.9**	**3.9**	**4.0**	**0.89**	**0.89**	**0.89**	**0.90**	**0.89**	**26.3**	**34.9**	**22.5**	**26.8**	**27.6**
	AS	4.1	4.0	4.0	3.8	3.9	0.84	0.82	0.84	0.84	0.84	32.0	39.2	28.3	35.6	33.7

Audio Samples: https://rhoposit.github.io/icassp2021/

1) Jennifer Williams, Joanna Rownicka, Pilar Oplustil, and Simon King," Comparison of Speech Representations for Automatic Quality Estimation in Multi-Speaker Text-to-Speech Synthesis," in Proc. Odyssey2020 The Speaker and Language Recognition Work-shop, 2020, pp. 222–229
2) Andrew Cameron Morris, Viktoria Maier, and Phil Green," From WER and RIL to MER and WIL: Improved Evaluation Measures for Connected Speech Recognition," in Eighth International Conference on Spoken Language Processing, 2004

Outline

- Motivation
- Related Work
- VQ-VAE Variants
- Disentanglement Evaluation
- Conclusion & Future Work

Disentanglement Evaluation: Speaker Diarization Task

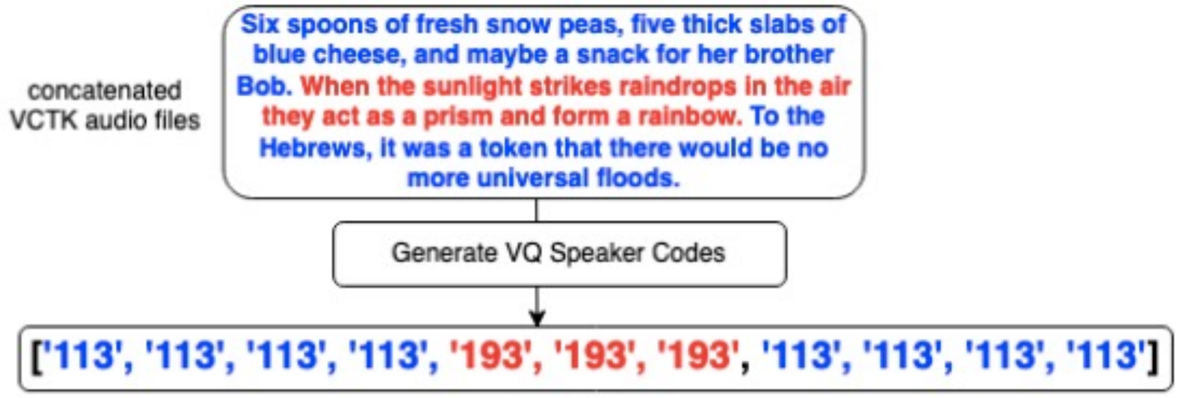


Fig. 2: VQ speaker codes are generated at each sliding window, for a given audio file that contains two different speakers. The codes determine which regions of speech belong to speaker A (113) versus speaker B (193).

Concatenate audio: 2 speakers 3 turns
2s sliding window (250s overlap)

Baseline:
DIHARD 2019 Track 1 x-vector
PLDA + agglomerative clustering
Trained on LDC development data

VQ Method:
Obtain codes from single-speaker audio
Obtain speaker codes from 2 speaker audio
Look-up codes using single-speaker reference

Disentanglement Evaluation: Speaker Diarization Task

Table 2: Speaker diarization error (DER) scores on concatenated VCTK audio. (S: softmax, AS: angular-softmax).

Method		C1	C2	C3	C4	Avg
x-vector		24.3	44.6	27.4	46.7	35.8
VQ-VAE		–	–	–	–	–
+ Global VQ		44.4	39.1	44.7	39.6	42.0
+ Speaker label	S	32.4	32.2	31.0	33.1	32.2
	AS	34.6	35.9	36.4	35.9	35.7
+ Adversarial loss	S	32.2	32.3	30.5	32.9	31.9
	AS	37.2	35.6	36.1	35.2	36.0

+Speaker label / +Adversarial loss systems:

- performed better than x-vector baseline (on avg)
- Significantly better than +GlobalVQ
- +GlobalVQ did not learn a diverse speaker space

DER (diarization error rate)
Speaker error
False alarm speech
Missed speech

Disentanglement Evaluation: Phone Recognition Task

Table 3: Phone error rate (% PER) on TIMIT from sub-phone VQ codes or audio. (S: softmax, AS: angular-softmax).

Method		# VQ Codes	% PER Sub	Ins	Del	Total
Audio Baseline		–	13.8	9.4	7.4	30.6
VQ-VAE		140	**26.6**	**8.3**	**6.0**	**40.9**
+ Global VQ		119	28.0	7.7	6.4	42.1
+ Speaker label	S	139	28.1	9.6	5.8	43.4
	AS	138	**27.6**	**8.0**	**6.3**	**41.9**
+ Adversarial loss	S	176	28.0	8.6	6.5	43.1
	AS	154	30.4	9.6	6.5	46.5

TIMIT data
ESPNet / Kaldi CMU AN4 recipe
LSTM encoder-decoder model
64 units, 100 epochs, CTC loss, no attention
Decoder beam size = 20
TIMIT has 63 unique phone types

Baseline: audio features

Experiments: string of code indexes

Adding a speaker component to VQ-VAE does not sacrifice phone quality
+Speaker label AS performed better than +Global VQ
VQ-VAE systems make similar proportion of error types (high substitution, low deletion)

Outline

- Motivation
- Related Work
- VQ-VAE Variants
- Disentanglement Evaluation
- Conclusion & Future Work

Findings

- Adding a speaker VQ codebook does not cause problems for phone codebook
- (new) Speaker codes are meaningful in diarization task
- Speaker VQ codebook helps system generalize to unseen conditions
- Semi-supervised w/ adversarial loss is the best system variant
- None of the system variants utilized the full phone or speaker codebook space

Ongoing and Future Work

- Building and testing triple-encoder systems (F0, speaker, phones)
- Learning multi-lingual speech synthesis (French, German, Italian, English)
- Learning VQ sub-phone space for code-switching speech
- Building and testing voice conversion using learned speaker dictionary
- Translating text/phones into VQ codes for text-to-speech synthesis

It is not yet known if certain types of information ***should*** remain entangled or not
e.g. should speaker identity include gender, age, and accent?

Thank You