Textless NLP

P. R. Sullivan

UBC NPL-DL, November 18th 2021

Table of Contents

Motivation

Why a textless NLP system?

NLP neglects languages with no standard written form

Purely spoken systems are closer to natural human communication.

What have been main holdups until now?

Metrics

Models

High Level

Textless NLP is
a three part system...

- speech-to-unit (S2u) [ENCODER]
- unit-based-language model (uLM)
- unit-to-spectrogram (u2S) [DECODER]

that performs four tasks...

- Acoustic Unit Discovery [ABX]
- Spoken Language Modeling [Spot-the-word]
- Speech Generation [AUC-of-VERT/PPX]
- Discrete Resynthesis [ASR-PER]

...enabled by new automatic [metrics for evaluation]

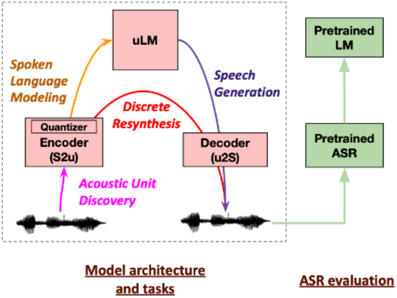


Table of Contents

On Generative Spoken Language Modeling from Raw Audio [4]

First Paper Main idea

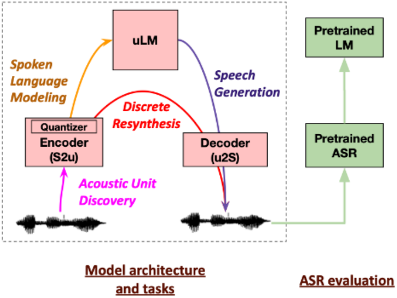
- Overview of entire pipeline
- Introduce two new Metrics (ASR-PER and AUC on Perplexity / VERT)
- Report results with human comparison evaluation

Other two papers focus on LM and Resynthesis in more detail.

Pipeline: S2u

S2u
Models: CPC, wav2vec 2.0, HuBERT all tested.

- Extract frame representation from layer (k) of Model
- k-means clustering discretizian
- "Psuedo-text-units"



Metrics: ABX

Task: Acoustic Unit Discovery

Metric: ABX
"Is X closer to A or B" = ABX
Work on Triphone level ("Bit" vs "Bet")
Can be done within (same speaker) or across speakers (must be normalized).

Standard Evaluation set with tools: LibriLite.

Metrics		S2u	
System	Nb units	ABX with.↓	ABX acr.↓
Toplines			
ASR+LM		-	-
Baselines			
LogMel	50	23.95	35.86
LogMel	100	24.33	37.86
LogMel	200	25.71	39.65
Unsupervised			
CPC	50	5.50	7.20
CPC	100	**5.09**	**6.55**
CPC	200	5.18	6.83
HuBERT-L6	50	7.37	8.61
HuBERT-L6	100	6.00	7.41
HuBERT-L6	200	5.99	7.31
wav2vec-L14	50	22.30	24.56
wav2vec-L14	100	18.16	20.44
wav2vec-L14	200	16.59	18.69

Metrics: ABX cont.

What does this look like?

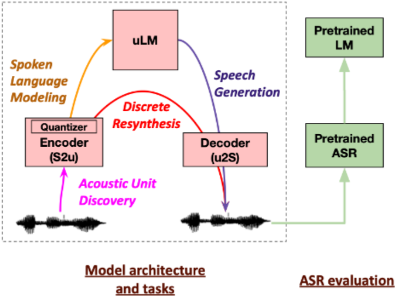
```
#file onset offset #phone prev-phone next-phone speaker
6295-244435-0009 0.2925 0.4725 IH L NG 6295
6295-244435-0009 0.3725 0.5325 NG IH K 6295
6295-244435-0009 0.4325 0.5725 K NG AH 6295
6295-244435-0009 0.4725 0.6125 AH K N 6295
6295-244435-0009 0.5325 0.6925 N AH HH 6295
6295-244435-0009 0.5725 0.7525 HH N AE 6295
6295-244435-0009 0.6125 0.8125 AE HH D 6295
6295-244435-0009 0.6925 0.9125 D AE K 6295
6295-244435-0009 0.7525 1.0125 K D AO 6295
6295-244435-0009 0.8125 1.0725 AO K L 6295
6295-244435-0009 0.9125 1.1125 L AO D 6295
6295-244435-0009 1.0125 1.1925 D L F 6295
6295-244435-0009 1.0725 1.2525 F D ER 6295
6295-244435-0009 1.1125 1.3325 ER F V 6295
```

Pipeline: uLM

S2u

Models: Transformer LM Big

- Train on pseudo-text-units
- Standard causal LM

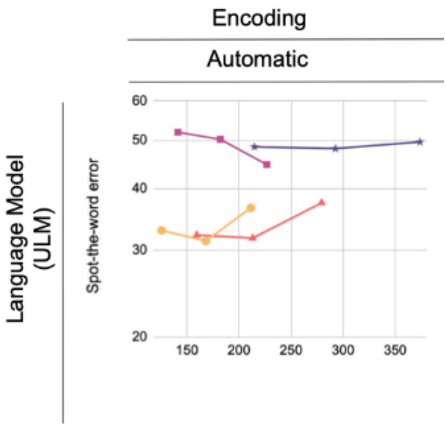


Metrics: Spot-the-Word

Task: Spoken Language Modelling

Metric: Spot-the-Word
Given two one-word wav files:
"p(Real Word) > p(Fake Word)"

Standard Dataset: sWUGGY (ENGLISH)
Found to correlate well with ABX



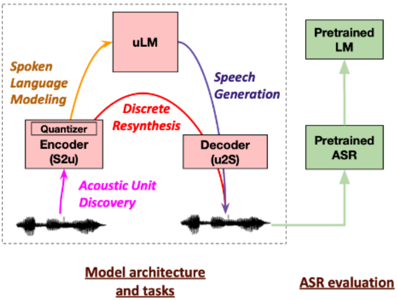
Bottom: Bitrate. Yellow (HuBERT), Red (CPC), Purple (wav2vec), Blue (log-mel)

Pipeline: u2S

u2S

Models: adapted Tacotron-2

- Input psuedo-text-units
- Output log-mel spectrogram
- Trained on LJ Speech
- WaveGlow vocoder



Metrics: ASR-PER

Task: Speech Resynthesis

Metric: ASR-based Phoneme Error Rate
Idea: "Use a standard phone-ASR model as judge"

Note: domain effect between LJ speech and Librispeech (LS) performance.

Systems			End-to-end ASR-based metrics			
S2u architect.	Nb units	Bit-rate	PER↓ (LJ)	PER↓ (LS)	CER↓ (LJ)	CER↓ (LS)
Toplines						
original wav			-	-	-	-
orig text+TTS			7.78	7.92	8.87	5.14
ASR + TTS	27		9.45	8.18	9.48	5.30
Baselines						
LogMel	50	214.8	27.72	49.38	27.73	52.05
LogMel	100	292.7	25.83	45.58	24.88	48.71
LogMel	200	373.8	19.78	45.16	17.86	46.12
Unsupervised						
CPC	50	159.4	10.87	17.16	10.68	12.06
CPC	100	213.1	10.75	15.82	9.84	9.46
CPC	200	279.4	**8.74**	14.23	9.20	8.29
HuBERT-L6	50	125.7	11.45	16.68	11.02	11.85
HuBERT-L6	100	168.1	9.53	13.24	9.31	7.19
HuBERT-L6	200	211.3	8.87	**11.06**	**8.88**	**5.35**
wav2vec-L14	50	141.3	24.95	33.69	25.42	32.91
wav2vec-L14	100	182.1	14.58	22.07	13.72	17.22
wav2vec-L14	200	226.8	10.65	16.34	10.21	10.50

Metrics: AUC PPL/VERT

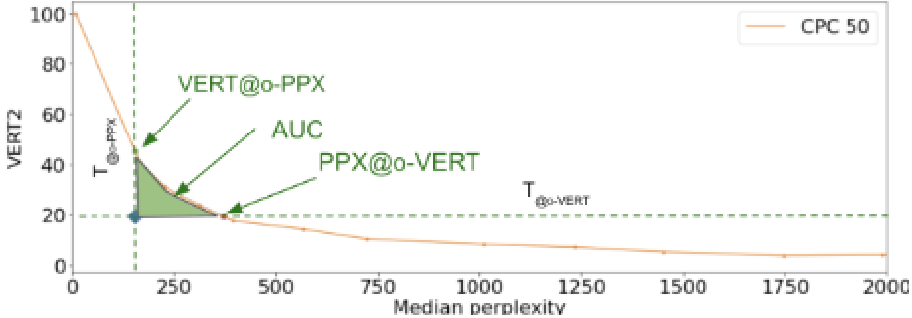
Task: Speech Generation

Metric: AUC on Perplexity and VERT (diVERsiTy)

Pre-trained ASR model to get transcripts

Perplexity: Quality. **VERT**: How diverse

VERT is geometric mean of self-BLEU and (new) auto-BLEU (ratio of repeated N-grams in an utterance)



Metrics: AUC PPL/VERT cont.

Results

Systems		Generation based metrics					
Encoder architect.	Nb units	unconditional PPX↓	unconditional VERT↓	unconditional AUC↓	prompt PPX↓	prompt VERT↓	prompt AUC↓
Controls							
oracle text		154.5	19.43	-	154.5	19.43	-
ASR + LM		178.4	21.31	0.18	162.8	20.49	0.04
Baseline							
LogMel	50	1588.97	-	1083.76	-	-	-
LogMel	100	1500.11	95.50	510.26	-	-	-
LogMel	200	1539.00	-	584.16	-	-	-
Unsupervised							
CPC	50	374.26	46.26	19.68	323.9	39.92	18.44
CPC	100	349.56	41.797	15.74	294.7	42.93	14.06
CPC	200	362.84	40.28	16.46	303.5	43.42	26.67
HuBERT-L6	50	376.33	43.06	19.27	339.8	45.85	21.03
HuBERT-L6	100	**273.86**	**31.36**	**5.54**	**251.2**	**33.67**	**5.88**
HuBERT-L6	200	289.36	33.04	7.49	262.4	34.30	6.13
wav2vec-L14	50	936.97	-	307.91	1106.3	-	330.8
wav2vec-L14	100	948.96	79.51	208.38	775.1	-	205.7
wav2vec-L14	200	538.56	61.06	61.48	585.8	-	91.07

First Paper Wrap Up

- Complete pipeline Spoken Language Generation
- Test two new metrics (AUC PPL/VERT and ASR-PER), both of which correlate well to human judgement.
- Human preference for large number of k-means units.
- CPC and HuBERT both perform well as encoders
- GSLM with off the shelf parts

Table of Contents

Speech Resynthesis from Discrete Disentangled Self-Supervised Representations [6]

Second Paper Main idea

- Replaces u2S component with multistream model
- Three streams: pseudo-units, prosody, speaker embedding
- Eliminate log-Mel spectrogram
- Results in high efficiency codec

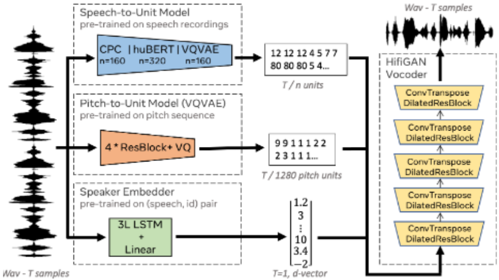
Modified Resynthesis

New u2S system

Multi-stream Cascade model:

- (From before) S2u model CPC/Hubert, (New) VQVAE
- (New) Pitch-to-Unit Model (VQVAE)
- (New) Speaker embedding
- (New) Vocoder directly integrated (HiFiGAN)

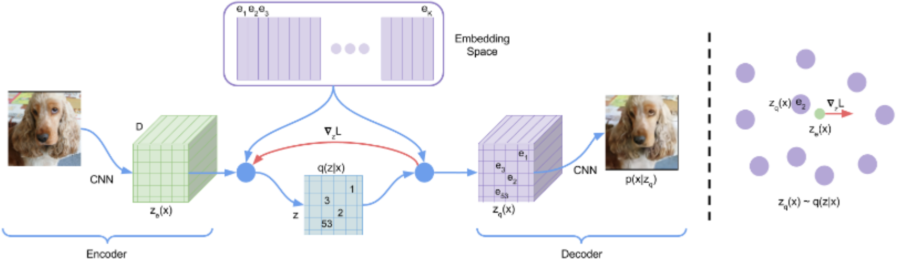
True "unit to speech" instead of "unit to spectrogram"



P2u

Pitch-to-unit

- train on F_0 information extracted from wav
- shift pitch information T_{n-1}



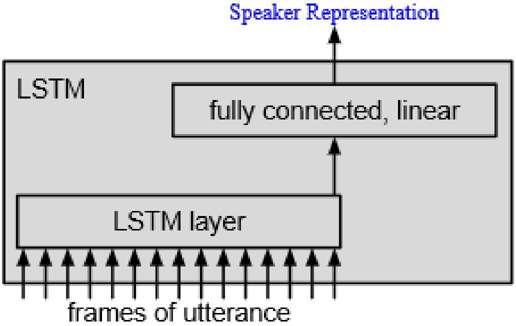
1

[1]Image credits: [5]

Speaker Embedding

Speaker Embedding

- Multi-layer LSTM classifier
- d-vector is just the final output
- LSTM uses crossentropy loss



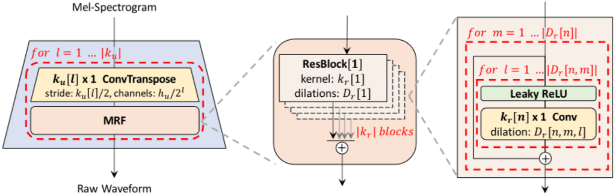
2

[2]Image credits: [1]

Connected Vocoder

Modified HiFiGAN

- Original: GAN upscale log-Mel spectrogram to wav.
- Modified: units, pitch, embedding instead of spectrogram
- Added: Reconstruction and Feature Matching loss



[3]

[3] Image credits: [3]

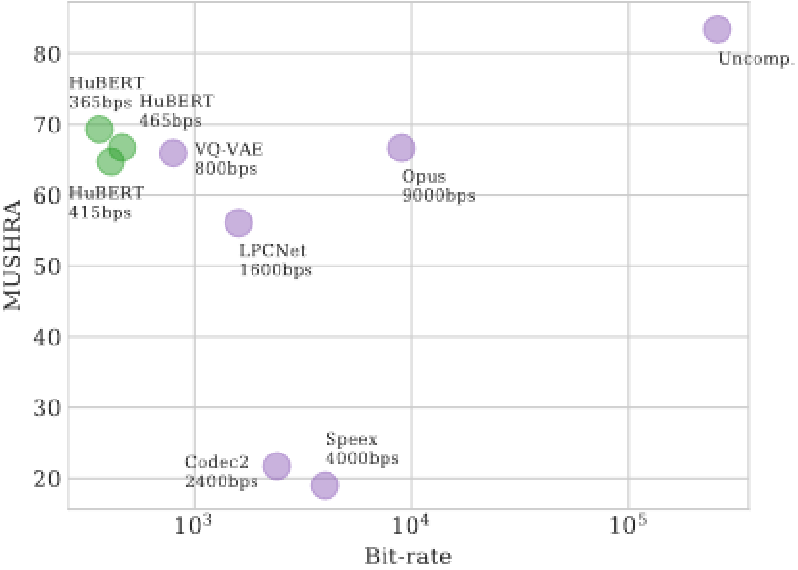
Results

Dataset	Method	Content		F0		Speaker	Overall Quality
		PER ↓	WER ↓	VDE ↓	FFE ↓	EER ↓	MOS ↑
LJ	GT	6.93	5.60	–	–	–	4.33±0.20
	CPC	9.66	8.51	13.48	15.19	–	3.31±0.33
	HuBERT	**9.52**	**6.96**	13.09	15.00	–	**3.66±0.33**
	VQ-VAE	12.77	8.85	**7.19**	**8.54**	–	**3.66±0.31**
VCTK	GT	17.16	4.32	–	–	3.25	4.08±0.66
	CPC	23.01	14.49	10.56	11.13	**4.25**	3.33±0.61
	HuBERT	**19.66**	**11.44**	9.77	10.43	5.79	**3.41±0.66**
	VQ-VAE	31.97	19.80	**5.20**	**5.59**	4.28	3.39±0.58

Results cont.

Dataset	Method	Voice Conversion				F0 Manipulation	
		PER ↓	WER ↓	EER ↓	MOS ↑	VDE ↑	FFE ↑
VCTK	GT	17.16	4.32	3.25	4.11±0.29	–	–
LJ	CPC	22.22	16.11	0.46	3.57±0.15	**46.68**	**48.71**
	HuBERT	**19.09**	**12.23**	**0.31**	**3.71±0.24**	39.20	48.42
	VQ-VAE	40.88	36.96	9.65	2.90±0.17	10.54	12.08
VCTK	CPC	23.58	15.98	**4.83**	3.42 ± 0.24	**25.29**	**26.97**
	HuBERT	**20.85**	**12.72**	6.01	**3.58 ± 0.28**	23.46	26.67
	VQ-VAE	36.88	29.44	11.56	3.08 ± 0.34	7.03	7.80

Results cont.



Second Paper Wrap Up

- Multistream!
- Elimination of log-mel step
- Evaluate Codec performance

Table of Contents

Text-Free Prosody-Aware Generative Spoken Language Modeling [2]

Final Paper Main idea
LMing now with Prosody!

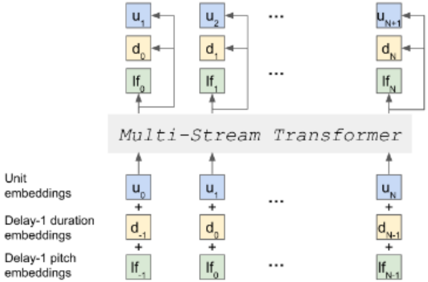
- Replaces uLM with multistream transformer
- Three streams: pseudo-units, prosody, speaker embedding
- New Prosody quantization *speaker-mean normalized log F0*

Modified uLM

New uLM system

Multi-stream Transformer model

Note: Time delay of prosody and embedding



Prosody Quantization

Speaker-mean normalized log F0
"ratio to the mean pitch in the log space"

References I

[1] Heigold, G., Moreno, I., Bengio, S., and Shazeer, N.
End-to-end text-dependent speaker verification.
In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2016), IEEE, pp. 5115–5119.

[2] Kharitonov, E., Lee, A., Polyak, A., Adi, Y., Copet, J., Lakhotia, K., Nguyen, T.-A., Rivière, M., Mohamed, A., Dupoux, E., et al.
Text-free prosody-aware generative spoken language modeling.
arXiv preprint arXiv:2109.03264 (2021).

[3] Kong, J., Kim, J., and Bae, J.
Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis.
arXiv preprint arXiv:2010.05646 (2020).

References II

[4] Lakhotia, K., Kharitonov, E., Hsu, W.-N., Adi, Y., Polyak, A., Bolte, B., Nguyen, T.-A., Copet, J., Baevski, A., Mohamed, A., et al.
Generative spoken language modeling from raw audio.
arXiv preprint arXiv:2102.01192 (2021).

[5] Oord, A. v. d., Vinyals, O., and Kavukcuoglu, K.
Neural discrete representation learning.
arXiv preprint arXiv:1711.00937 (2017).

[6] Polyak, A., Adi, Y., Copet, J., Kharitonov, E., Lakhotia, K., Hsu, W.-N., Mohamed, A., and Dupoux, E.
Speech resynthesis from discrete disentangled self-supervised representations.
arXiv preprint arXiv:2104.00355 (2021).