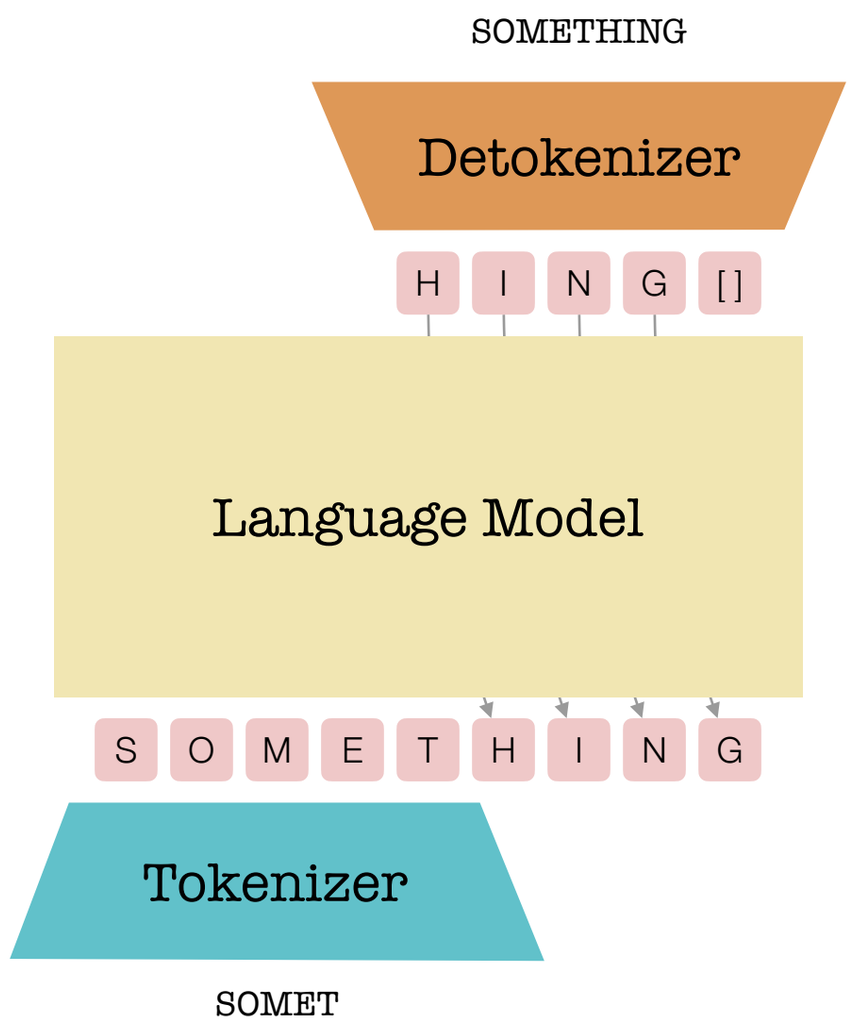
Kanade: A Simple Disentangled Tokenizer for Spoken Language Modeling

Zhijie Huang and Stephen McIntosh
Minematsu-Saito Lab — University of Tokyo

(D)NN-speech Reading Group
May 21, 2026

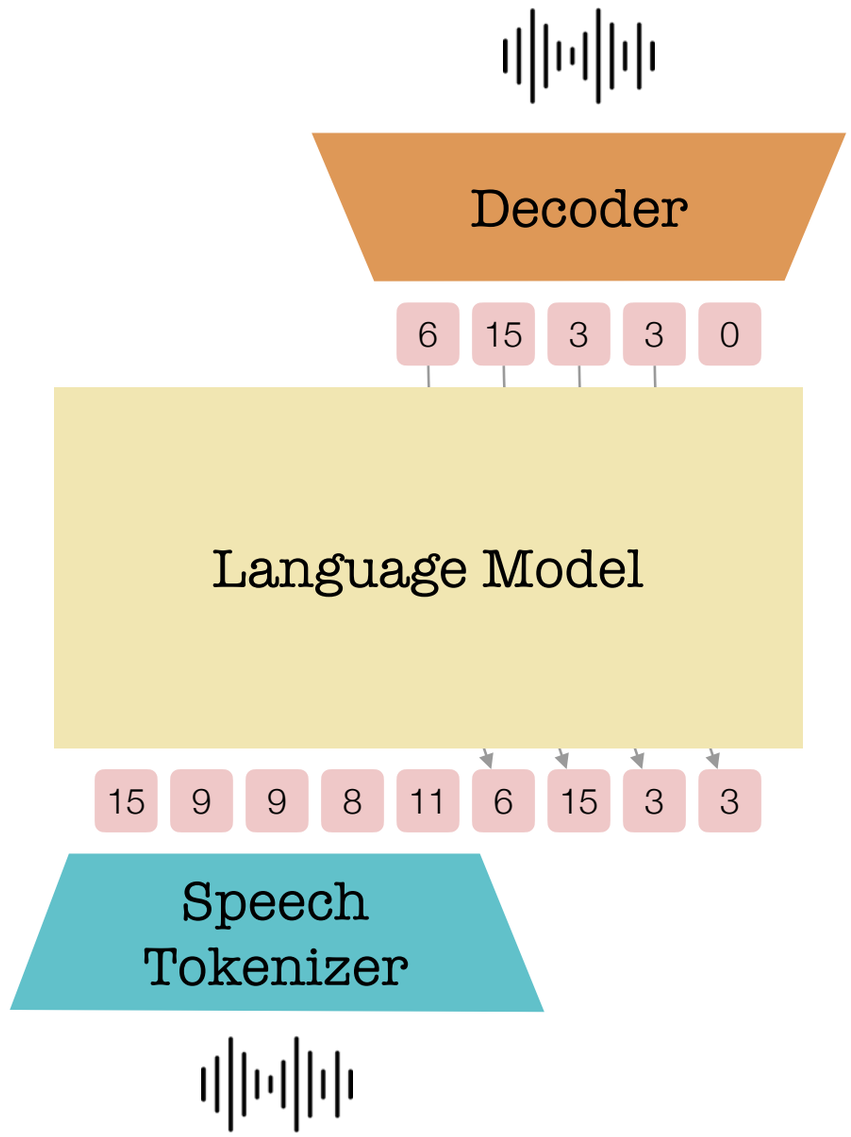
Background

Autoregressive text models



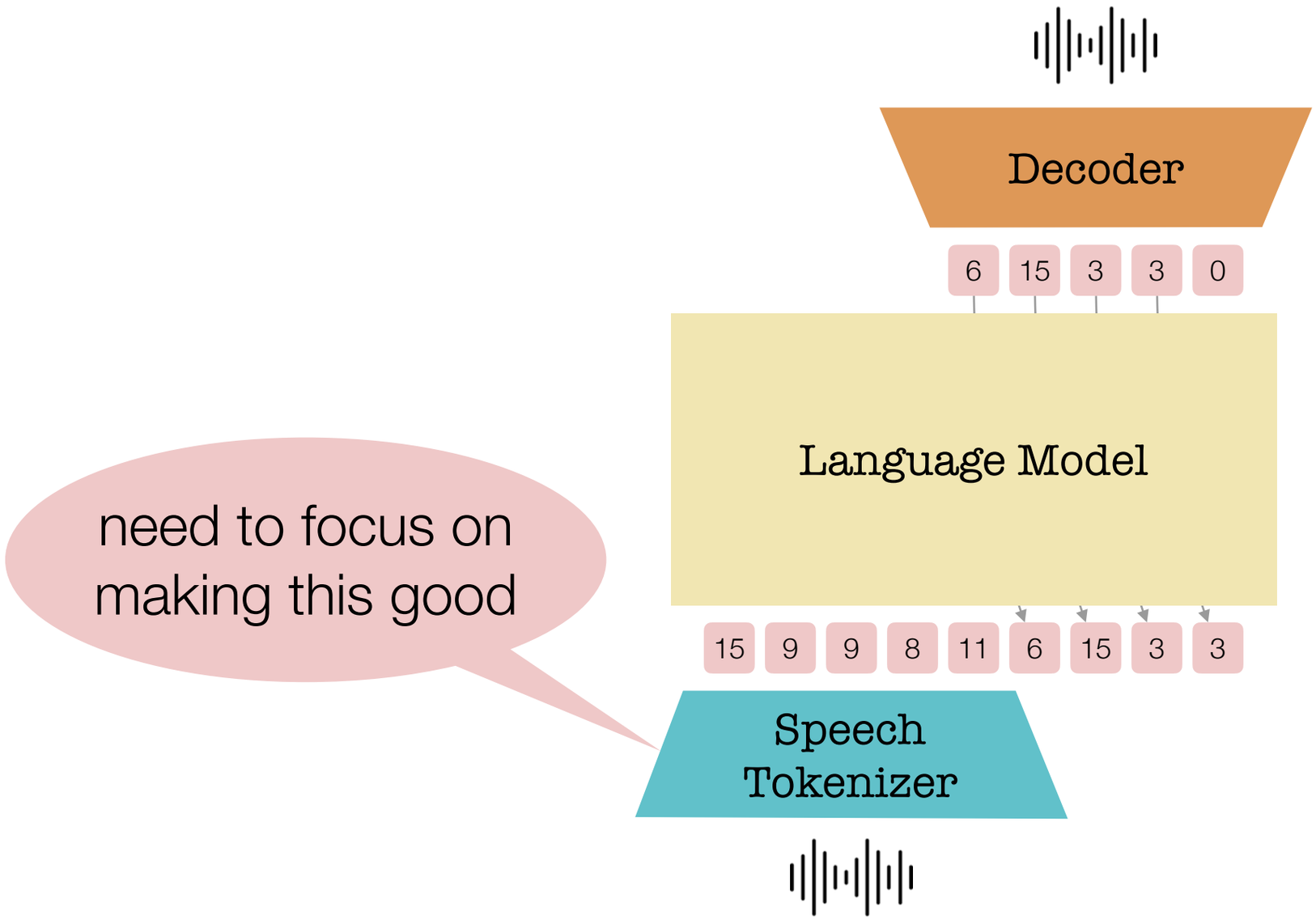
Background

Autoregressive speech models



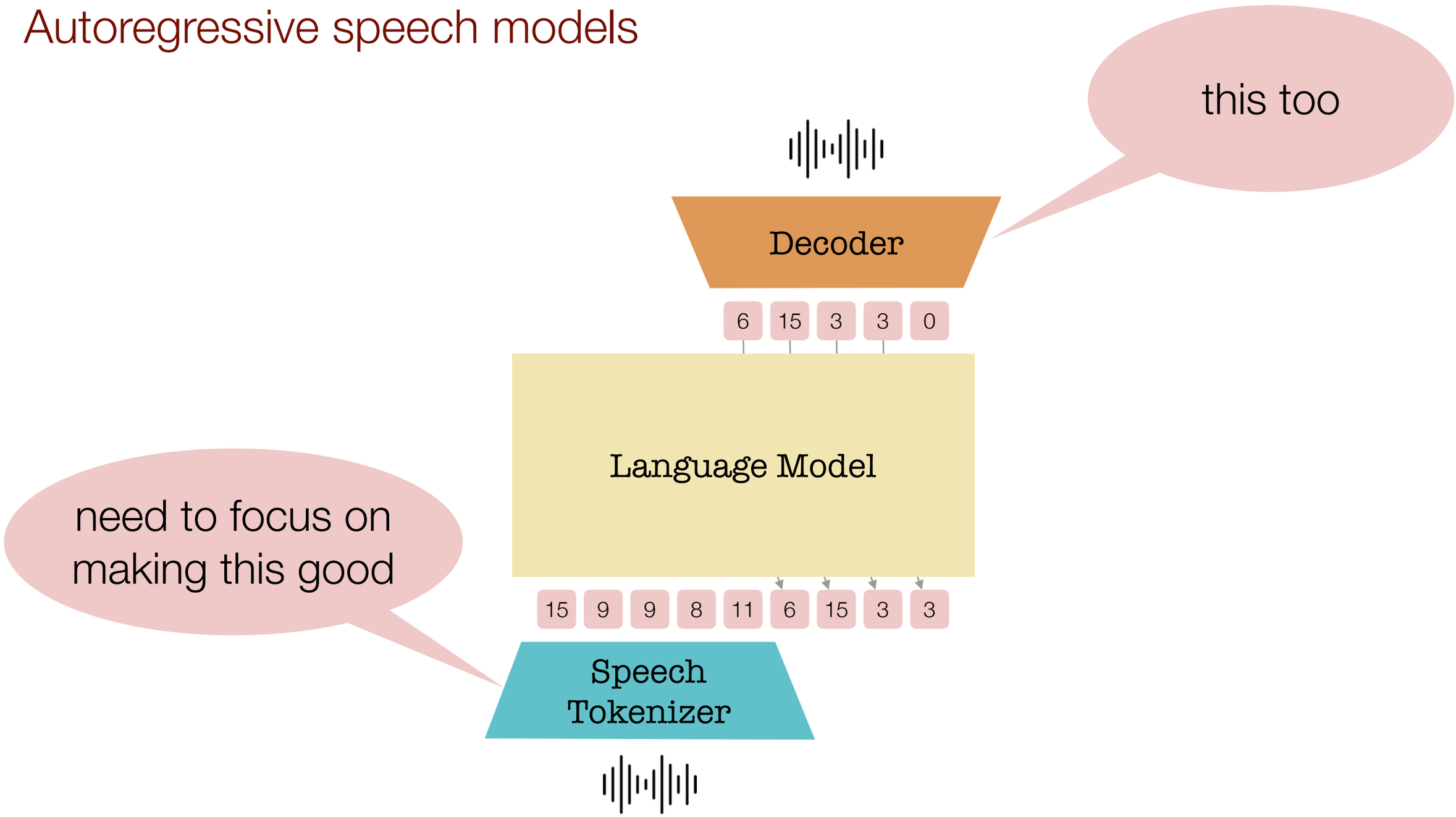
Background

Autoregressive speech models



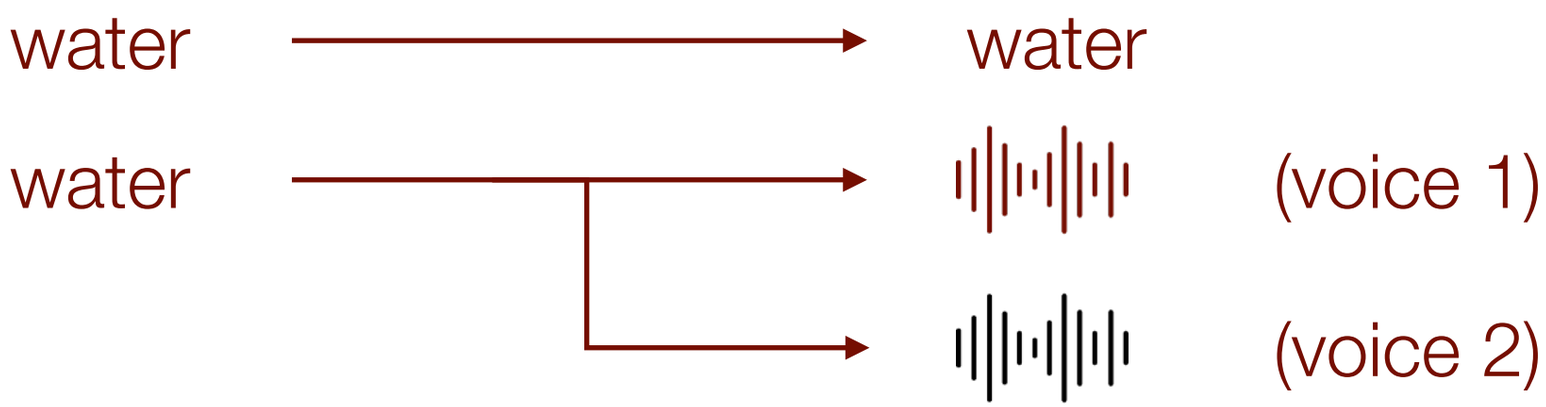
Background

Autoregressive speech models



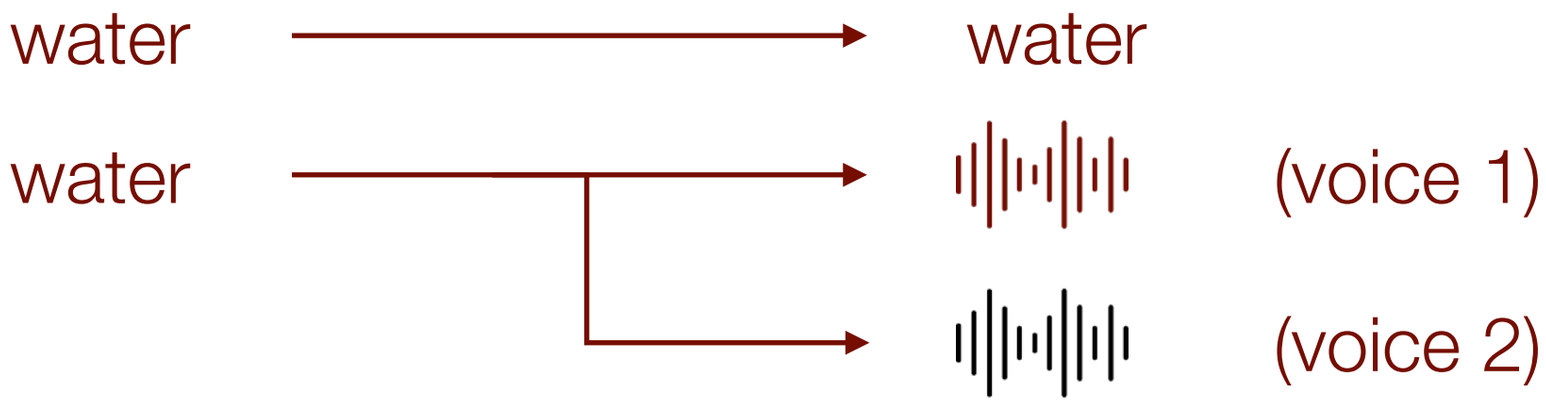
Encoding speech

Speech recordings have much more variability than text

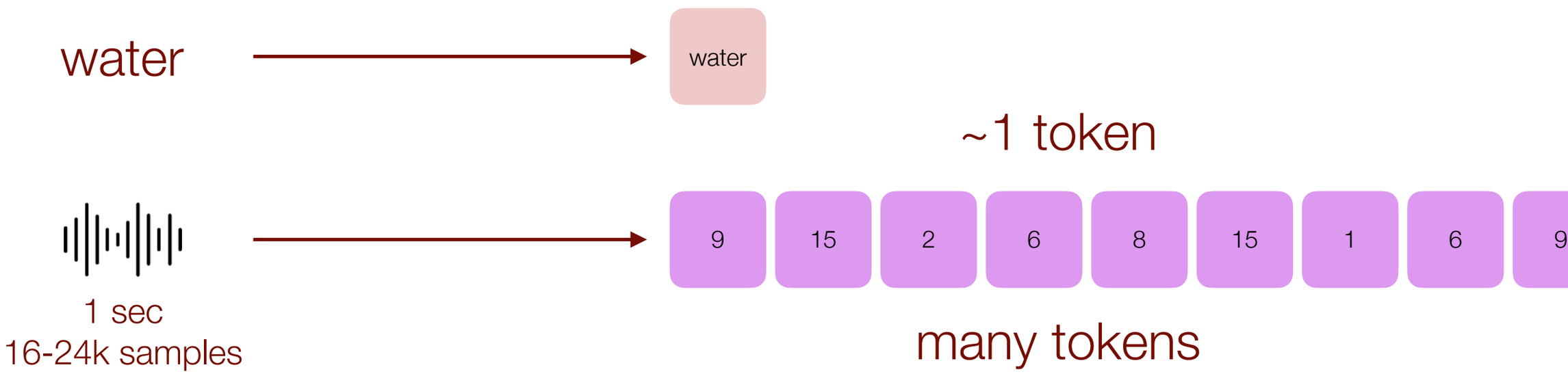


Encoding speech

Speech recordings have much more variability than text



…making speech representations much more verbose



Why are verbose representations a problem?

Downstream models need to model everything at once

- Channel qualities
- Background noise
- Speaker identity
- ...

Why are verbose representations a problem?

Downstream models need to model everything at once

- Channel qualities
- Background noise
- Speaker identity
- ...

→ particularly bad for transformers which prefer short sequences

Why are verbose representations a problem?

Downstream models need to model everything at once

- Channel qualities
- Background noise
- Speaker identity
- ...

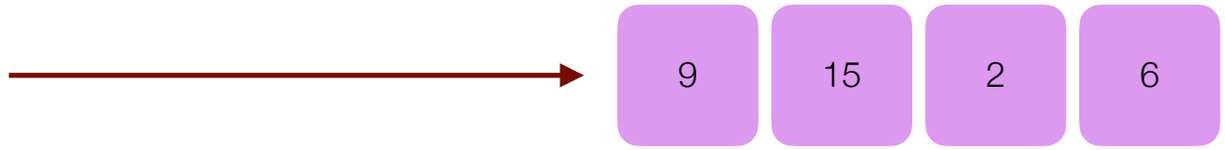
→ particularly bad for transformers which prefer short sequences

→ not impossible, but requires lots of data and a strong model

What can we do about it?

Compress—strip away the information that is not necessary to get outputs we want.

1 sec
16-24k samples

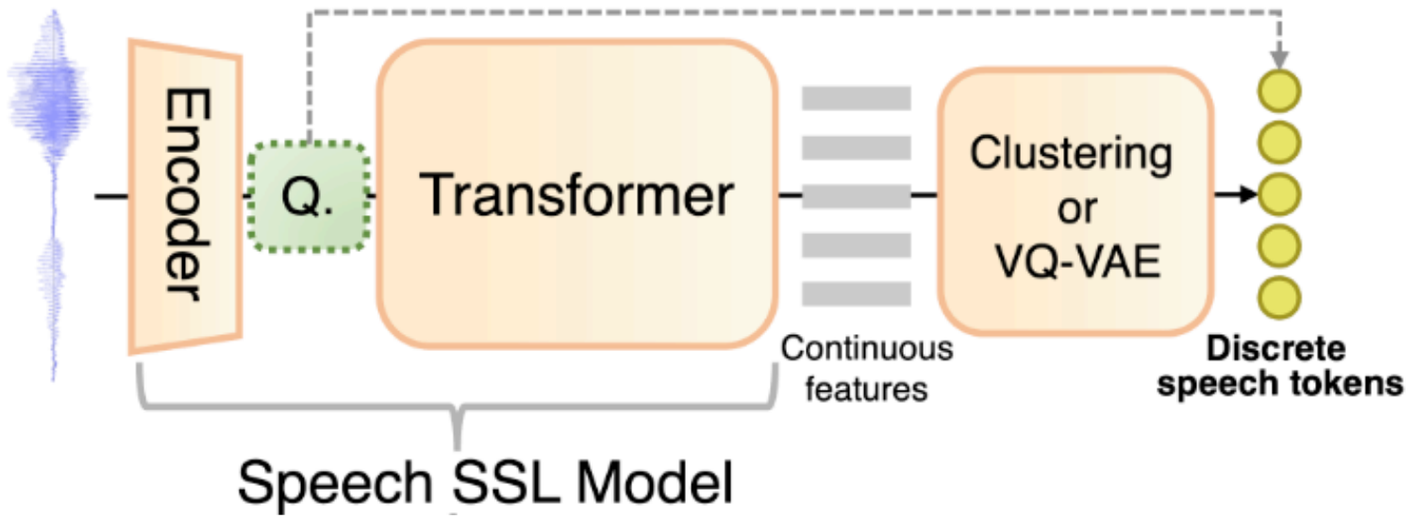


fewer tokens

Option 1: SSL k-means tokens

Use clustered continuous features from a self-supervised model

- Reasonably compressed and rich in phonetic information
- **But**: lack enough information to generate high-quality speech

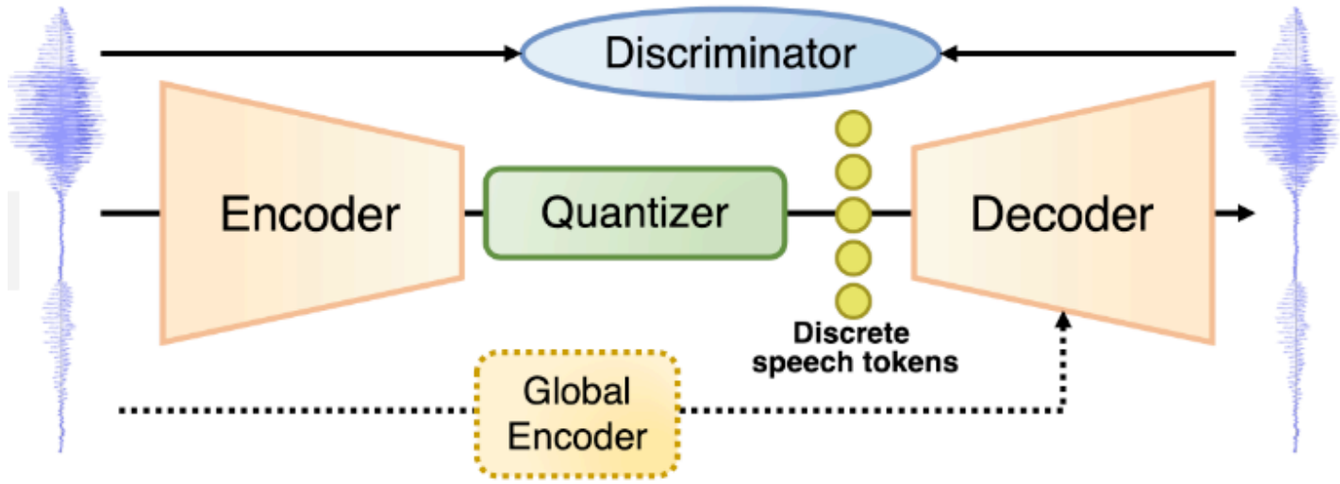


Y. Guo et al., “Recent Advances in Discrete Speech Tokens: A Review,” Feb. 16, 2025, arXiv: arXiv:2502.06490. doi: 10.48550/arXiv.2502.06490.

Option 2: Neural audio codecs

Encode speech aiming to reconstruct it as well as possible

- Have all the information necessary to generate high-quality audio
- **But**: still have a lot of information that is not relevant to language.

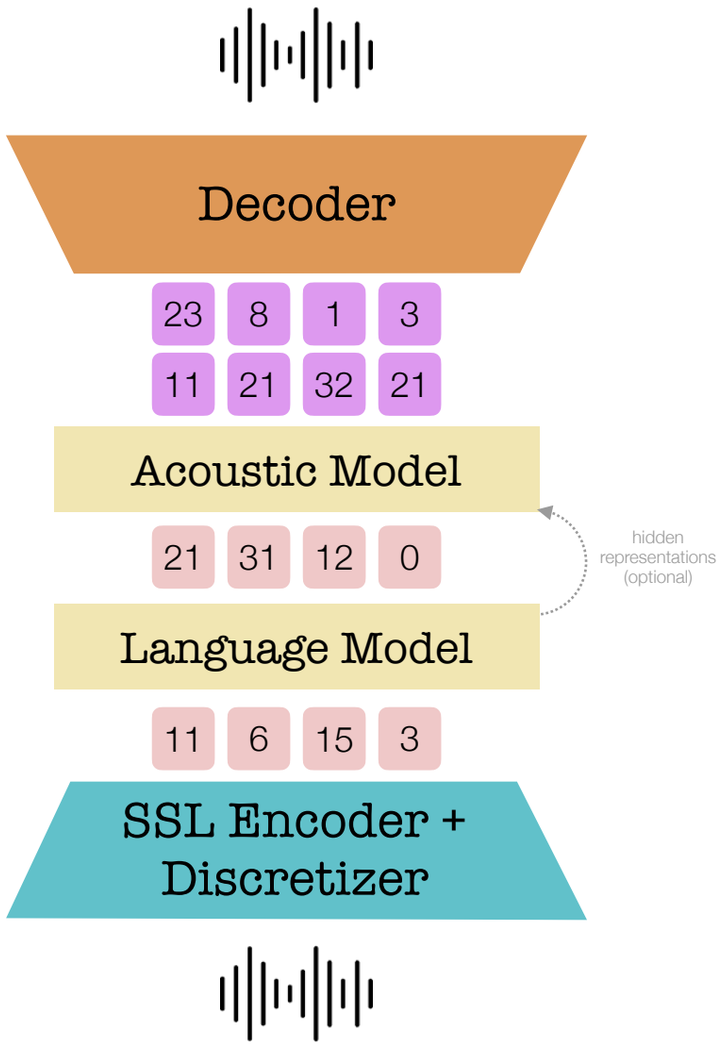


Y. Guo et al., “Recent Advances in Discrete Speech Tokens: A Review,” Feb. 16, 2025, arXiv: arXiv:2502.06490. doi: 10.48550/arXiv.2502.06490.

Option 3: Use both

Use k-means style tokens for language modeling and then generate the remaining features later

- Modeling on only k-means tokens: coherent speech
- Generating acoustic tokens helps improve output quality
- **But**: complicates architecture and requires downstream models to learn acoustic details



What do we actually want?

Downstream models should model speech based on:

- phonetics: the individual sounds
- prosody: rhythm, pitch, intensity, etc.

Semantics, syntax, pragmatics, etc. are higher level constructs built up from these

What do we actually want?

Downstream models should model speech based on:

- phonetics: the individual sounds
- prosody: rhythm, pitch, intensity, etc.

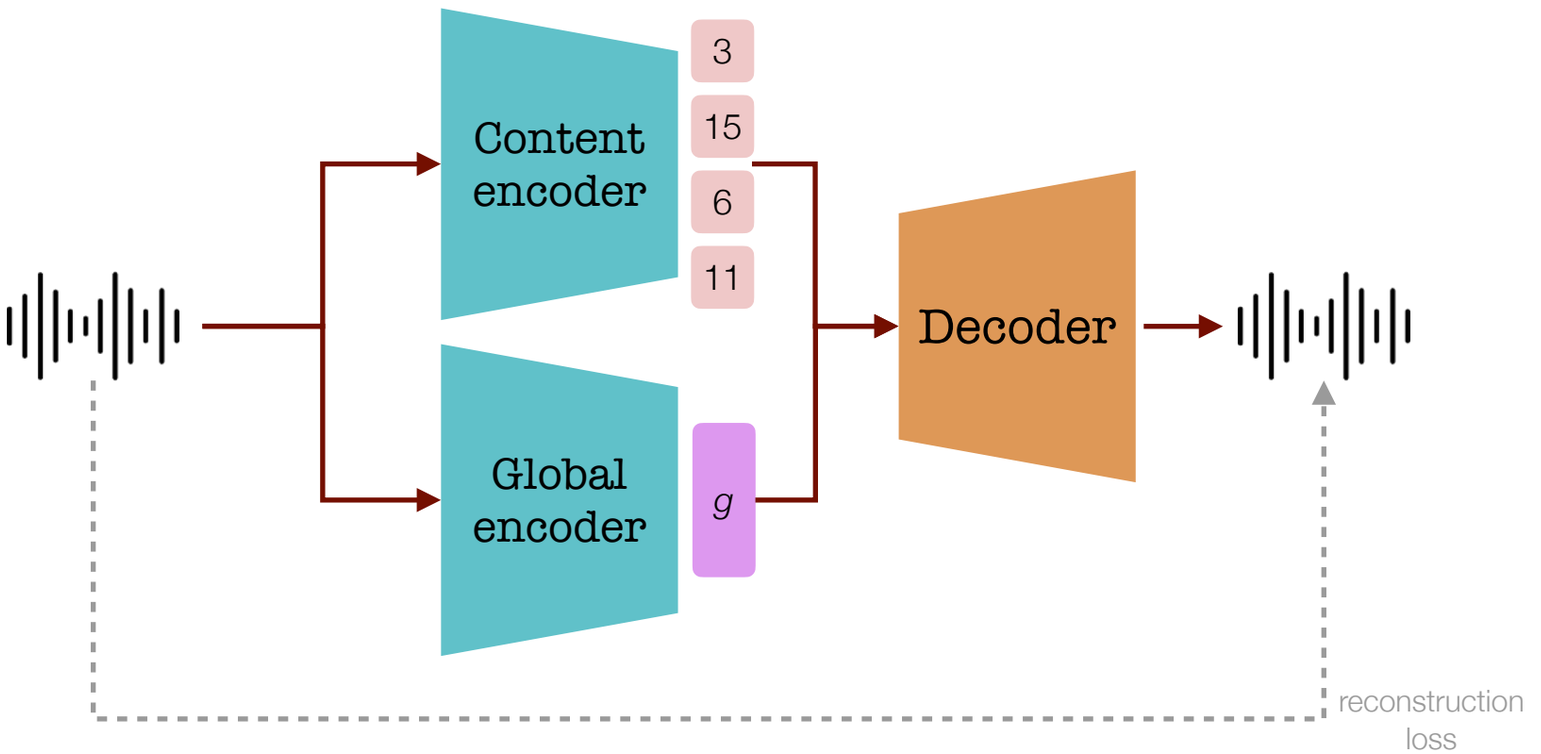
Semantics, syntax, pragmatics, etc. are higher level constructs built up from these

→ **tokens should include phonetics AND prosody**

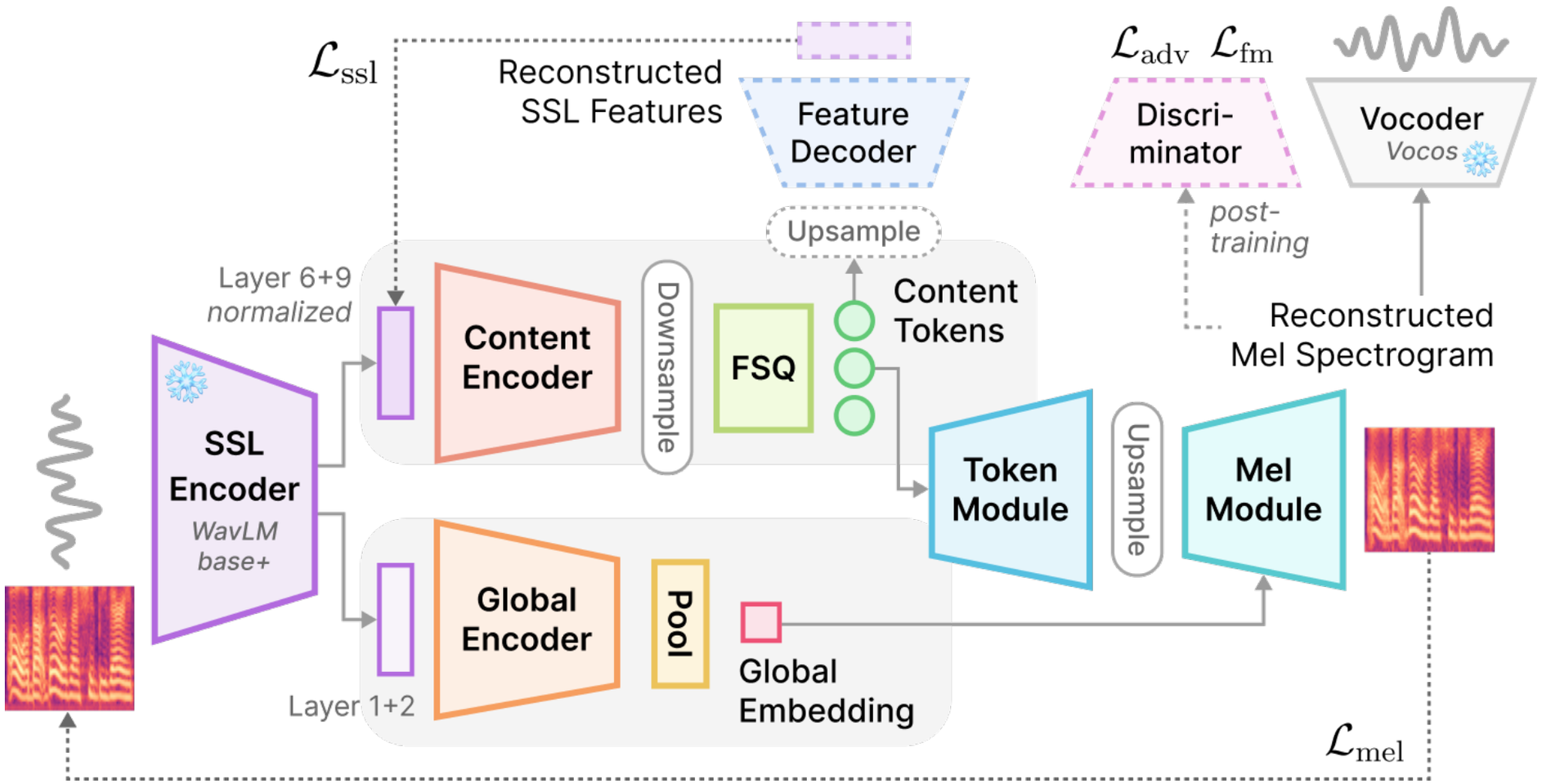
Option 4: Disentangled codec

Uses a reconstruction objective like NACs, but allow constant information to flow through a separate branch.

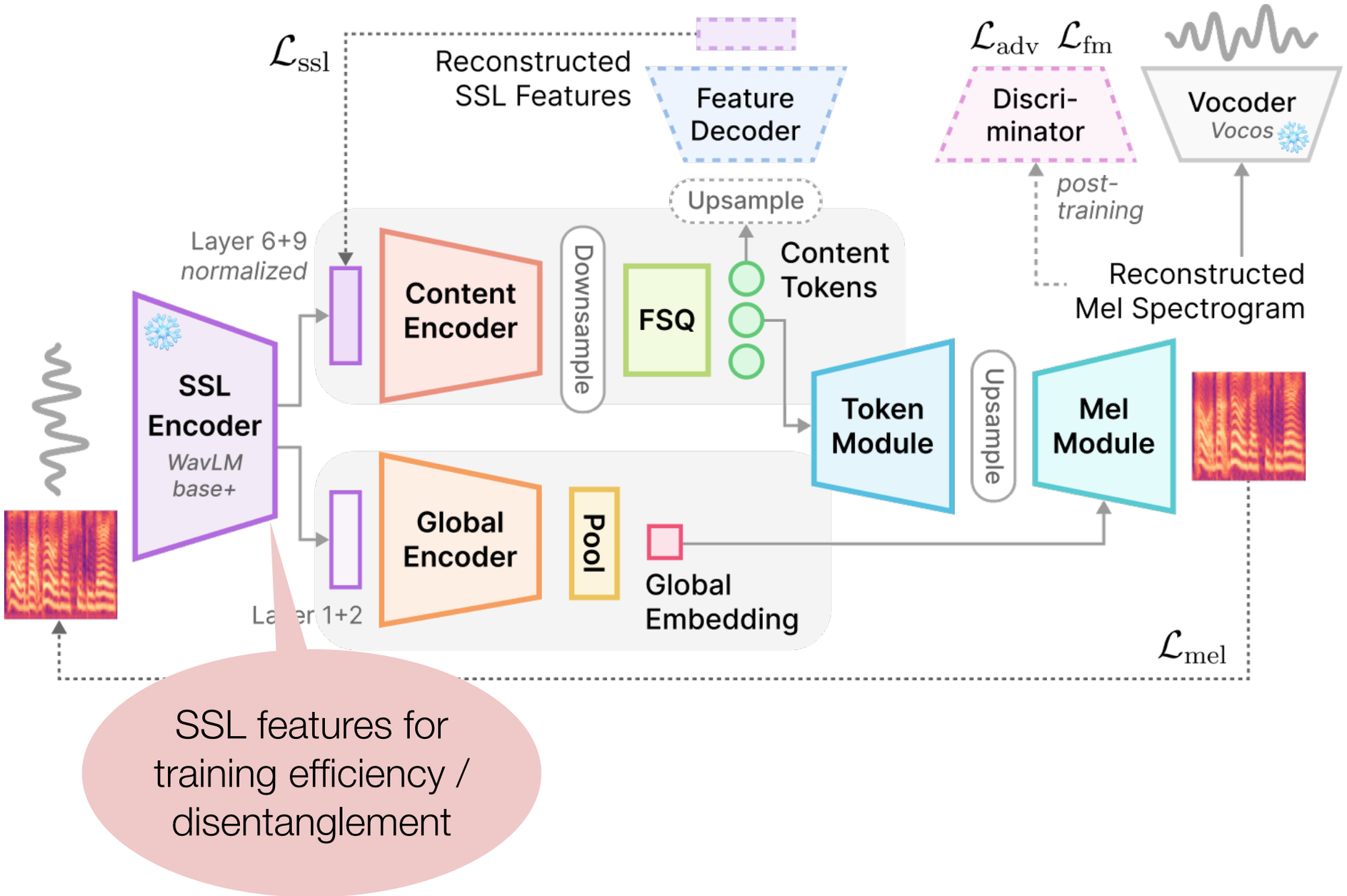
- Can generate coherent speech with appropriate prosody
- Downstream models do not concern themselves with acoustic details
- Does not require complicated downstream architectures



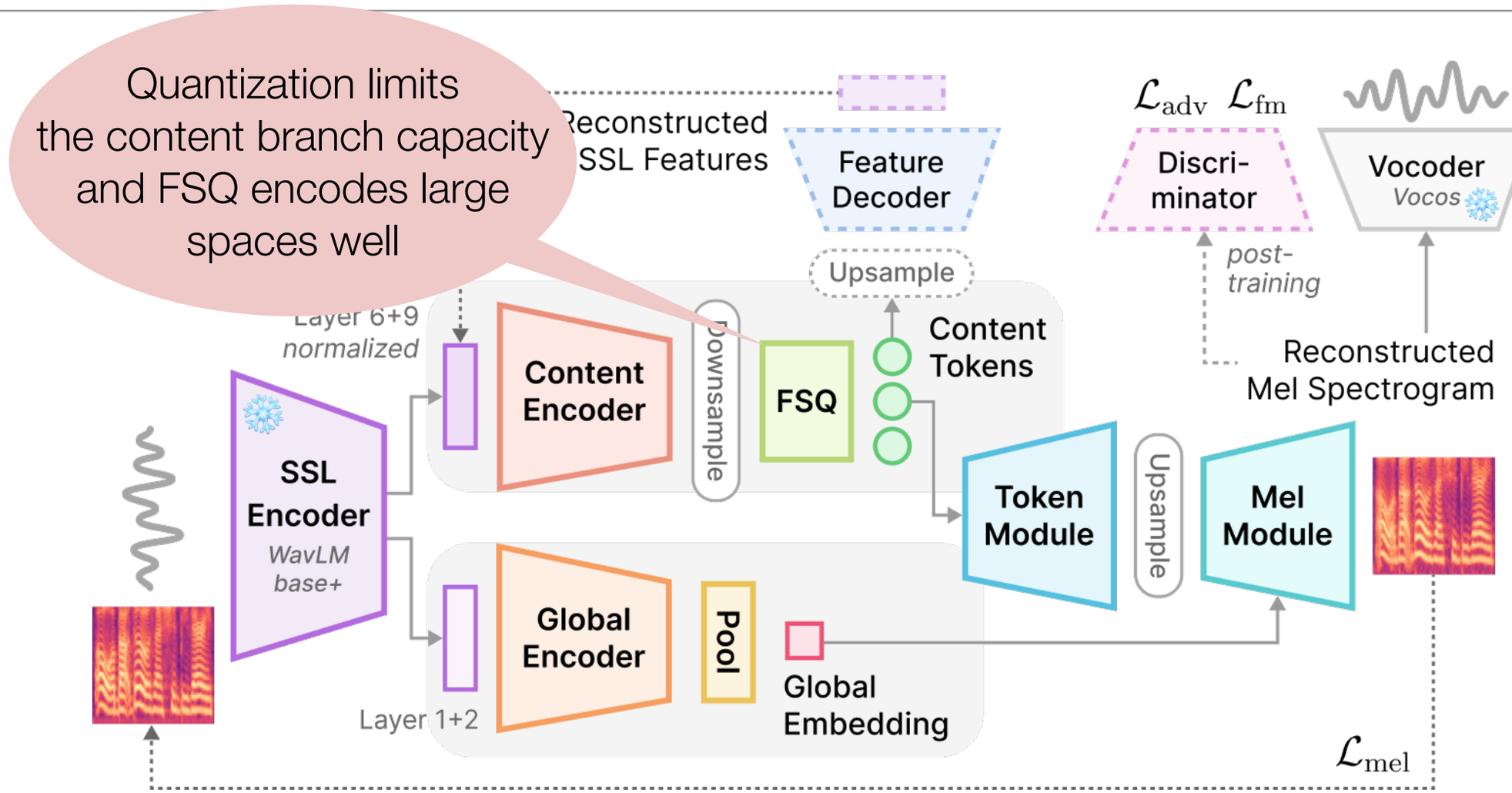
Kanade


$\mathcal{L}_{\text{ssl}}$
Reconstructed SSL Features
Feature Decoder
$\mathcal{L}_{\text{adv}}$ $\mathcal{L}_{\text{fm}}$
Discri-minator
Vocoder
Vocos
post-training
Upsample
Content Tokens
Layer 6+9
normalized
Content Encoder
Downsample
FSQ
Reconstructed Mel Spectrogram
SSL Encoder
WavLM base+
Token Module
Upsample
Mel Module
Global Encoder
Pool
Global Embedding
Layer 1+2
$\mathcal{L}_{\text{mel}}$

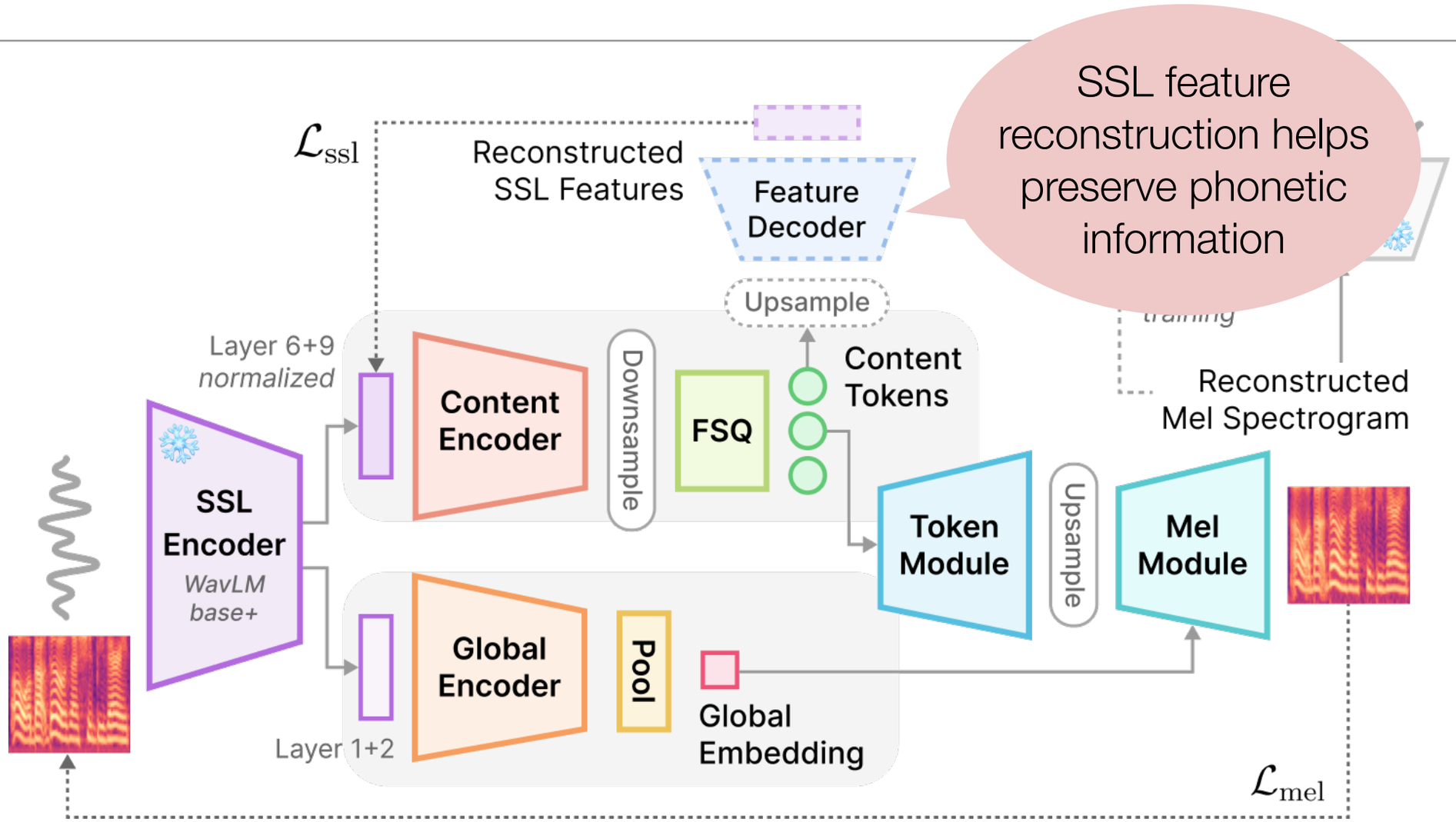
Kanade


$\mathcal{L}_{\text{ssl}}$
Reconstructed SSL Features
Feature Decoder
$\mathcal{L}_{\text{adv}}$ $\mathcal{L}_{\text{fm}}$
Discri-minator
Vocoder
Vocos
post-training
Upsample
Content Tokens
Layer 6+9
normalized
Content Encoder
Downsample
FSQ
Reconstructed Mel Spectrogram
SSL Encoder
WavLM base+
Token Module
Upsample
Mel Module
Global Encoder
Pool
Global Embedding
Layer 1+2
$\mathcal{L}_{\text{mel}}$
SSL features for training efficiency / disentanglement

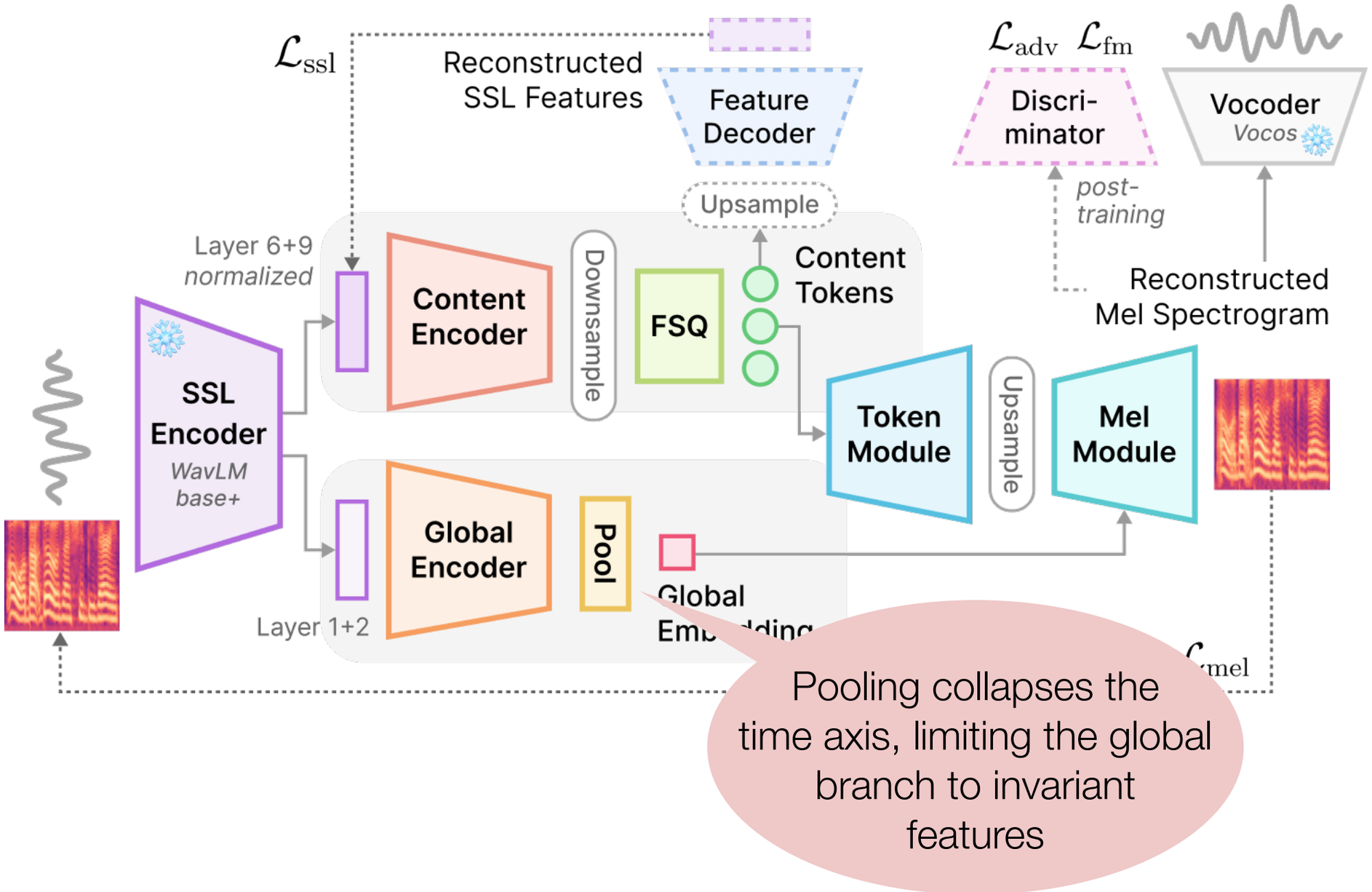
Kanade


Quantization limits the content branch capacity and FSQ encodes large spaces well
Reconstructed SSL Features
Feature Decoder
$\mathcal{L}_{\text{adv}}$ $\mathcal{L}_{\text{fm}}$
Discri-minator
Vocoder
Vocos
post-training
Upsample
Content Tokens
Layer 6+9
normalized
Content Encoder
Downsample
FSQ
Reconstructed Mel Spectrogram
SSL Encoder
WavLM base+
Token Module
Upsample
Mel Module
Global Encoder
Pool
Global Embedding
Layer 1+2
$\mathcal{L}_{\text{mel}}$

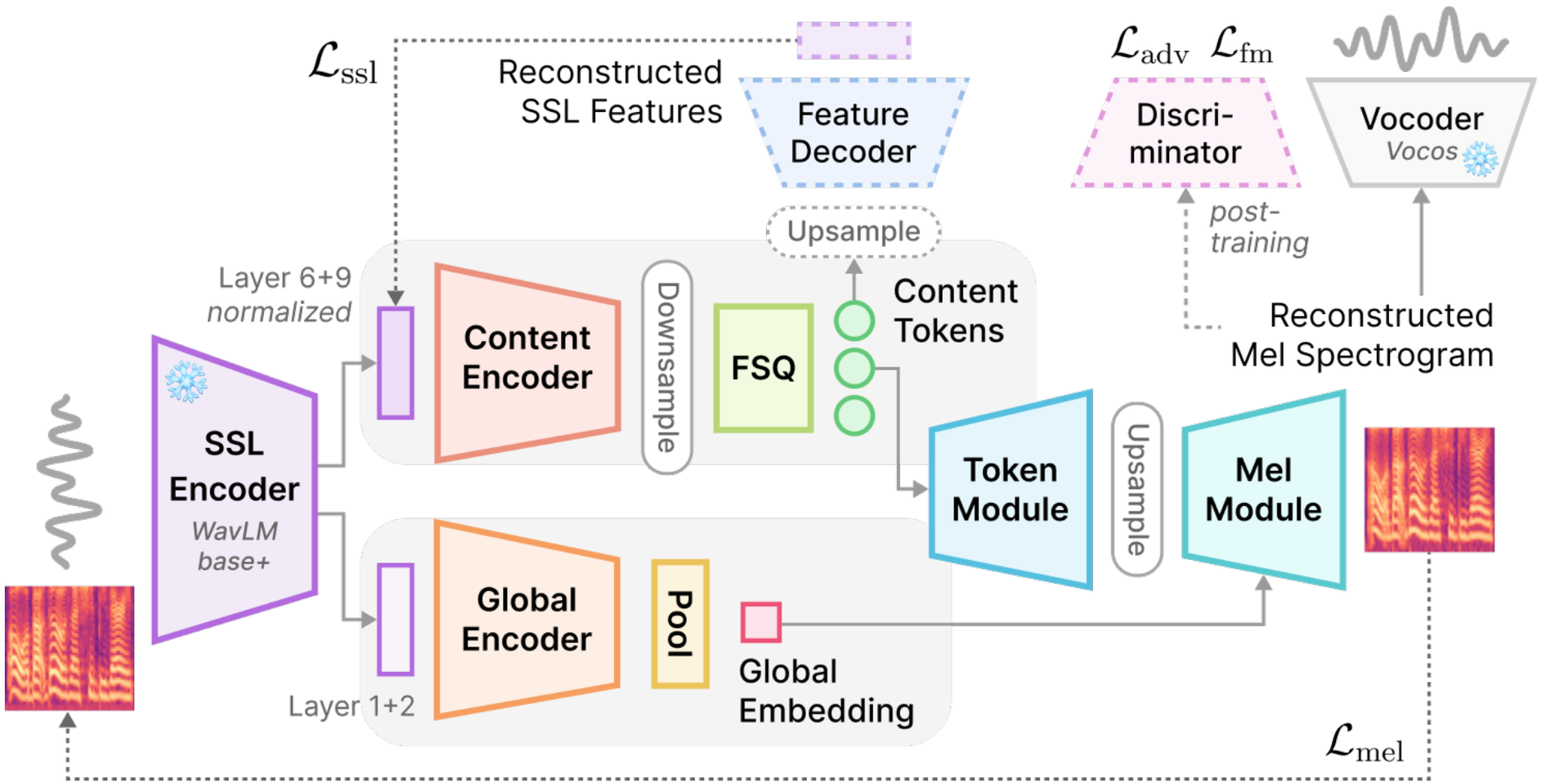
Kanade


$\mathcal{L}_{\text{ssl}}$
Reconstructed SSL Features
Feature Decoder
SSL feature reconstruction helps preserve phonetic information
Upsample
Layer 6+9
normalized
Content Encoder
Downsample
FSQ
Content Tokens
Reconstructed Mel Spectrogram
SSL Encoder
WavLM base+
Token Module
Upsample
Mel Module
Global Encoder
Pool
Global Embedding
Layer 1+2
$\mathcal{L}_{\text{mel}}$

Kanade


$\mathcal{L}_{\text{ssl}}$
Reconstructed SSL Features
Feature Decoder
$\mathcal{L}_{\text{adv}}$ $\mathcal{L}_{\text{fm}}$
Discri-minator
Vocoder
Vocos
post-training
Upsample
Content Tokens
Layer 6+9
normalized
Content Encoder
Downsample
FSQ
Reconstructed Mel Spectrogram
SSL Encoder
WavLM base+
Token Module
Upsample
Mel Module
Global Encoder
Pool
Layer 1+2
Global Embedding
$\mathcal{L}_{\text{mel}}$
Pooling collapses the time axis, limiting the global branch to invariant features

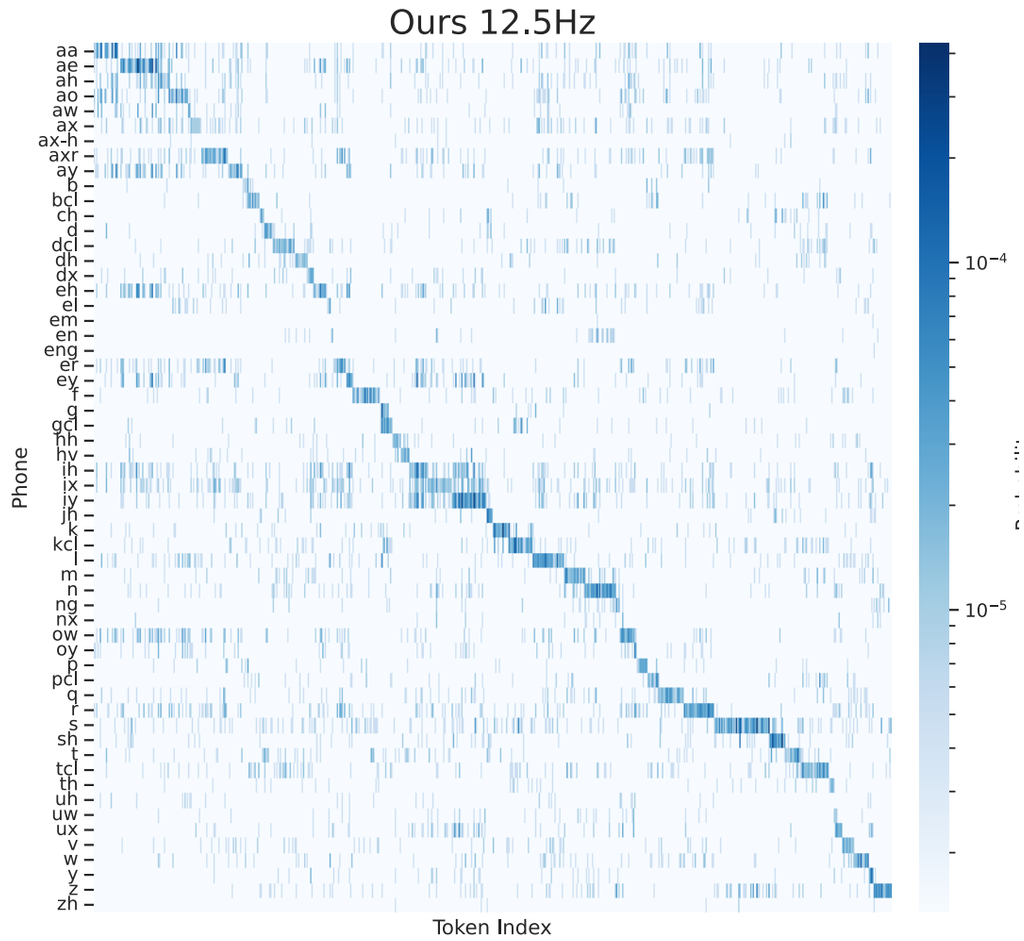
Kanade


$\mathcal{L}_{\text{ssl}}$
Reconstructed SSL Features
Feature Decoder
$\mathcal{L}_{\text{adv}}$ $\mathcal{L}_{\text{fm}}$
Discri-minator
Vocoder
Vocos
post-training
Upsample
Content Tokens
Layer 6+9
normalized
Content Encoder
Downsample
FSQ
Reconstructed Mel Spectrogram
SSL Encoder
WavLM base+
Token Module
Upsample
Mel Module
Global Encoder
Pool
Global Embedding
Layer 1+2
$\mathcal{L}_{\text{mel}}$

What does Kanade learn?

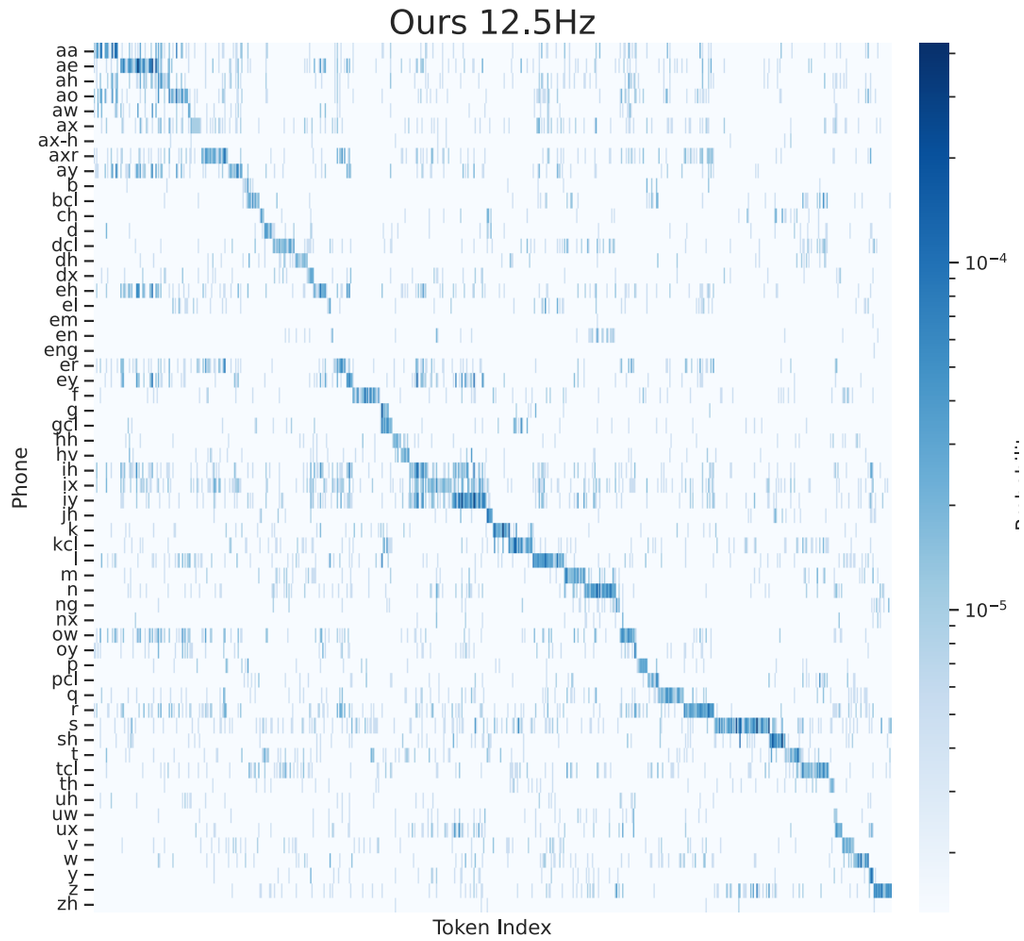
Phonetic information

Model	ABX↓		PNMI↑
	within	across	
KM 12.5Hz	4.4%	5.1%	0.79
KM 25Hz	3.5%	**4.2%**	0.81
PAST	**3.4%**	**4.2%**	**0.87**
ST	3.6%	4.5%	0.69
FACodec	4.4%	5.9%	0.53
Mimi	6.6%	7.8%	0.63
X-Codec 2	15.4%	22.4%	0.44
DualCodec	16.0%	19.1%	0.56
StableCodec	21.9%	25.0%	0.55
TiCodec	22.1%	27.1%	0.18
BiCodec	24.5%	34.3%	0.22
WavTokenizer	25.6%	31.5%	0.17
Kanade 12.5Hz	22.7%	24.3%	0.58
Kanade 25Hz	19.0%	21.6%	0.49



Phonetic information

Model	ABX↓		PNMI↑
	within	across	
KM 12.5Hz	4.4%	5.1%	0.79
KM 25Hz	3.5%	**4.2%**	0.81
PAST	**3.4%**	**4.2%**	**0.87**
ST	3.6%	4.5%	0.69
FACodec	4.4%	5.9%	0.53
Mimi	6.6%	7.8%	0.63
X-Codec 2	15.4%	22.4%	0.44
DualCodec	16.0%	19.1%	0.56
StableCodec	21.9%	25.0%	0.55
TiCodec	22.1%	27.1%	0.18
BiCodec	24.5%	34.3%	0.22
WavTokenizer	25.6%	31.5%	0.17
Kanade 12.5Hz	22.7%	24.3%	0.58
Kanade 25Hz	19.0%	21.6%	0.49

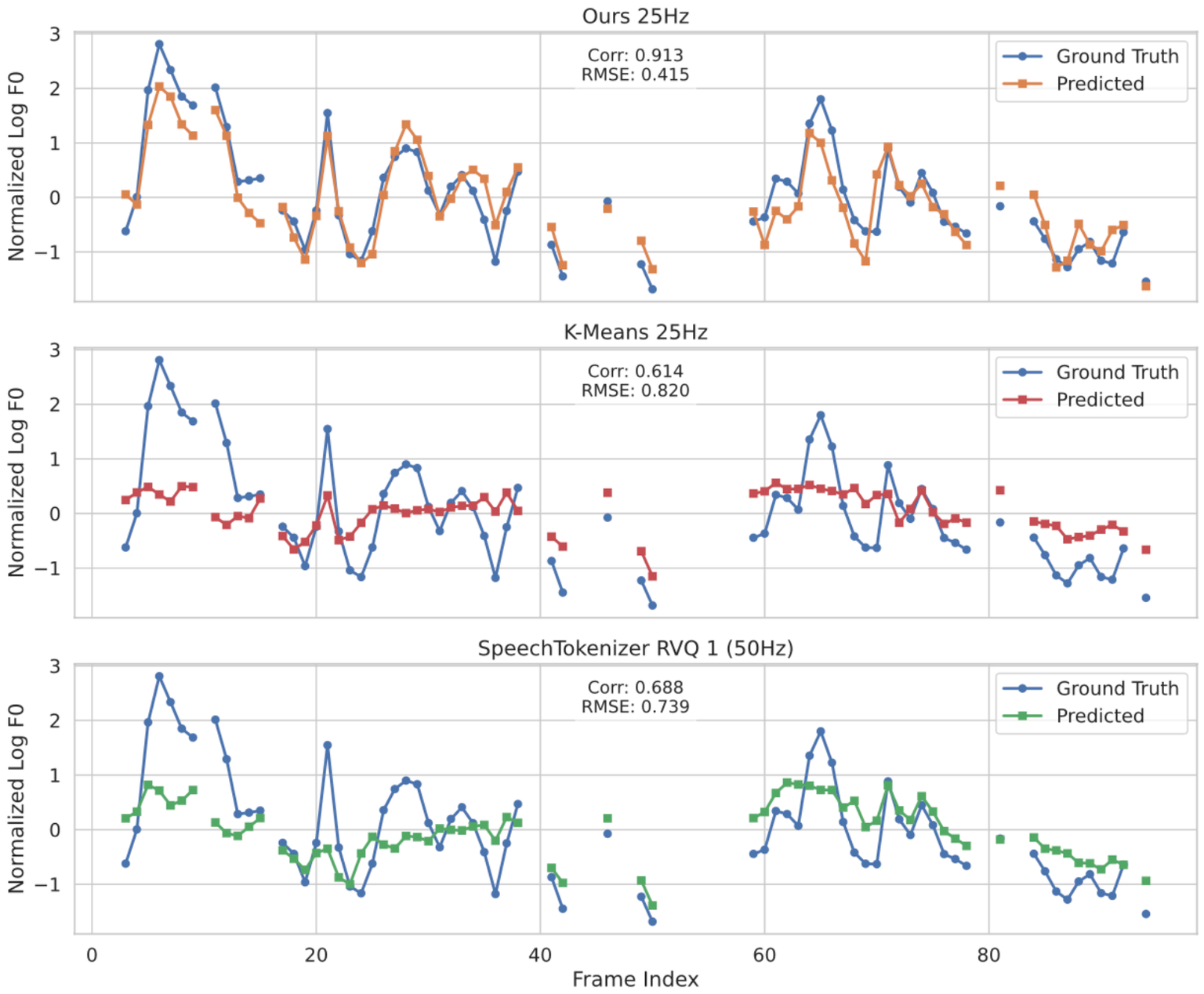


→ not great, possibly due to prosodic information and fat tokens

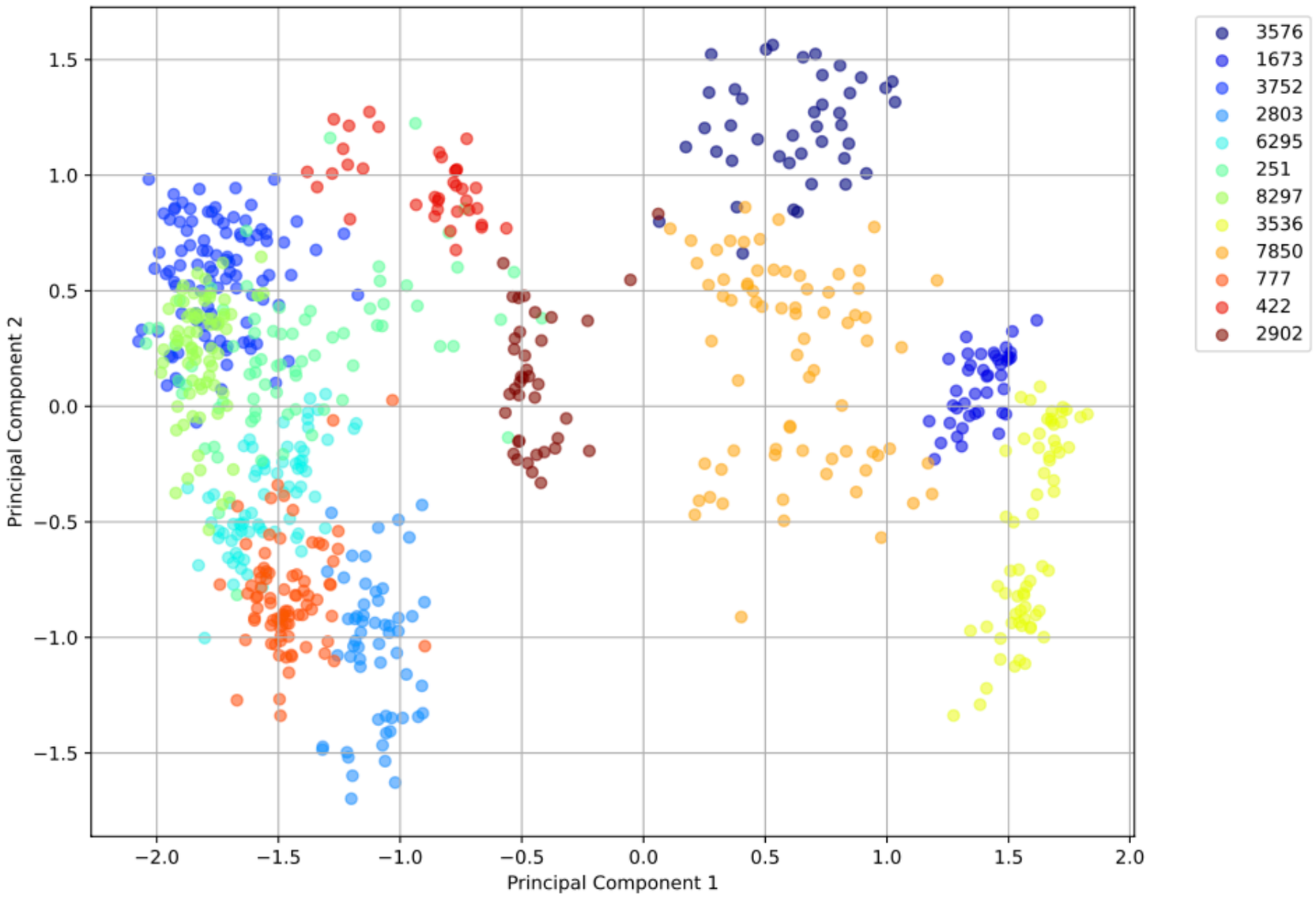
PCA on unquantized content embeddings

large markers are per-phoneme averages

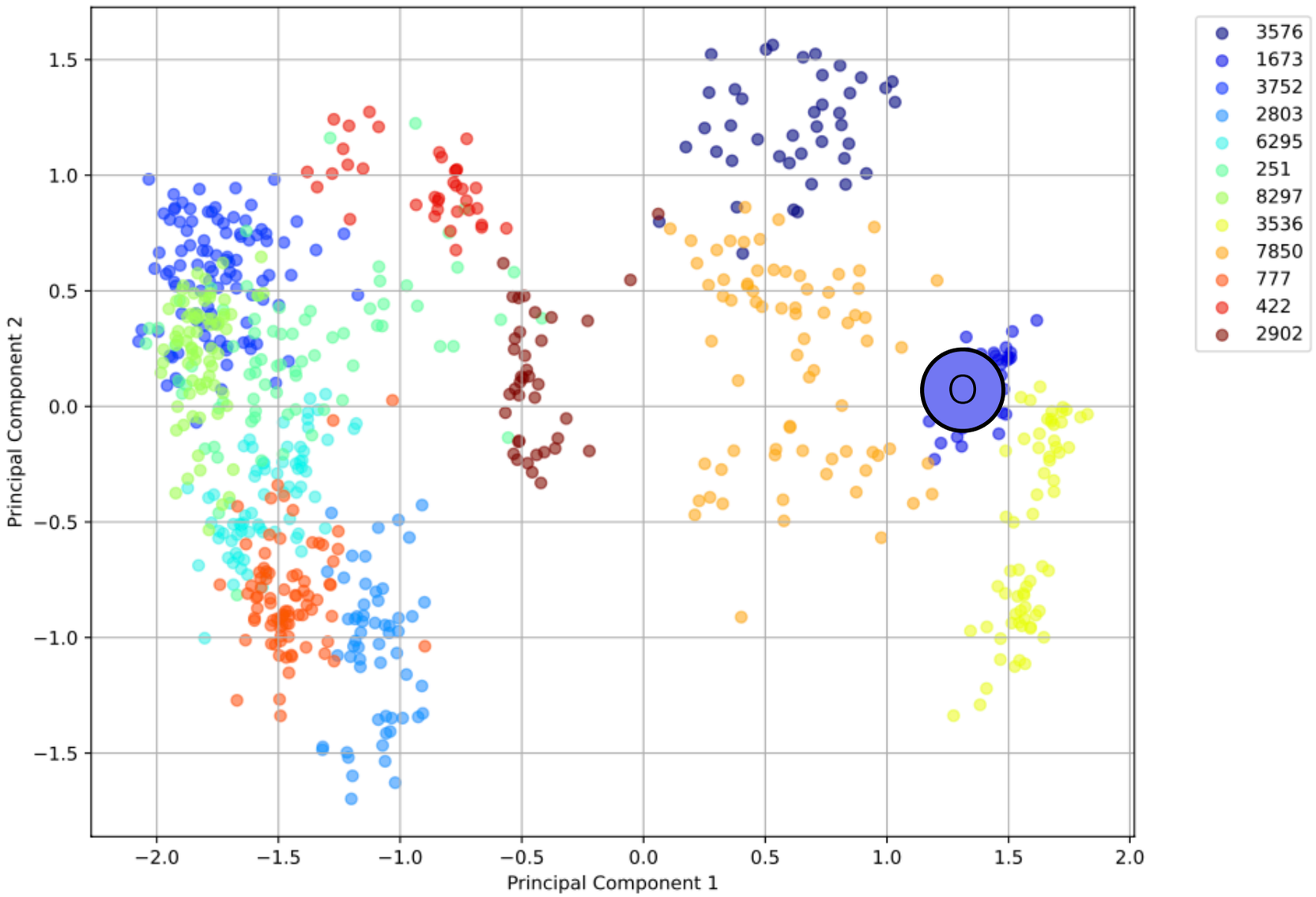
Prosodic information: F0 probing



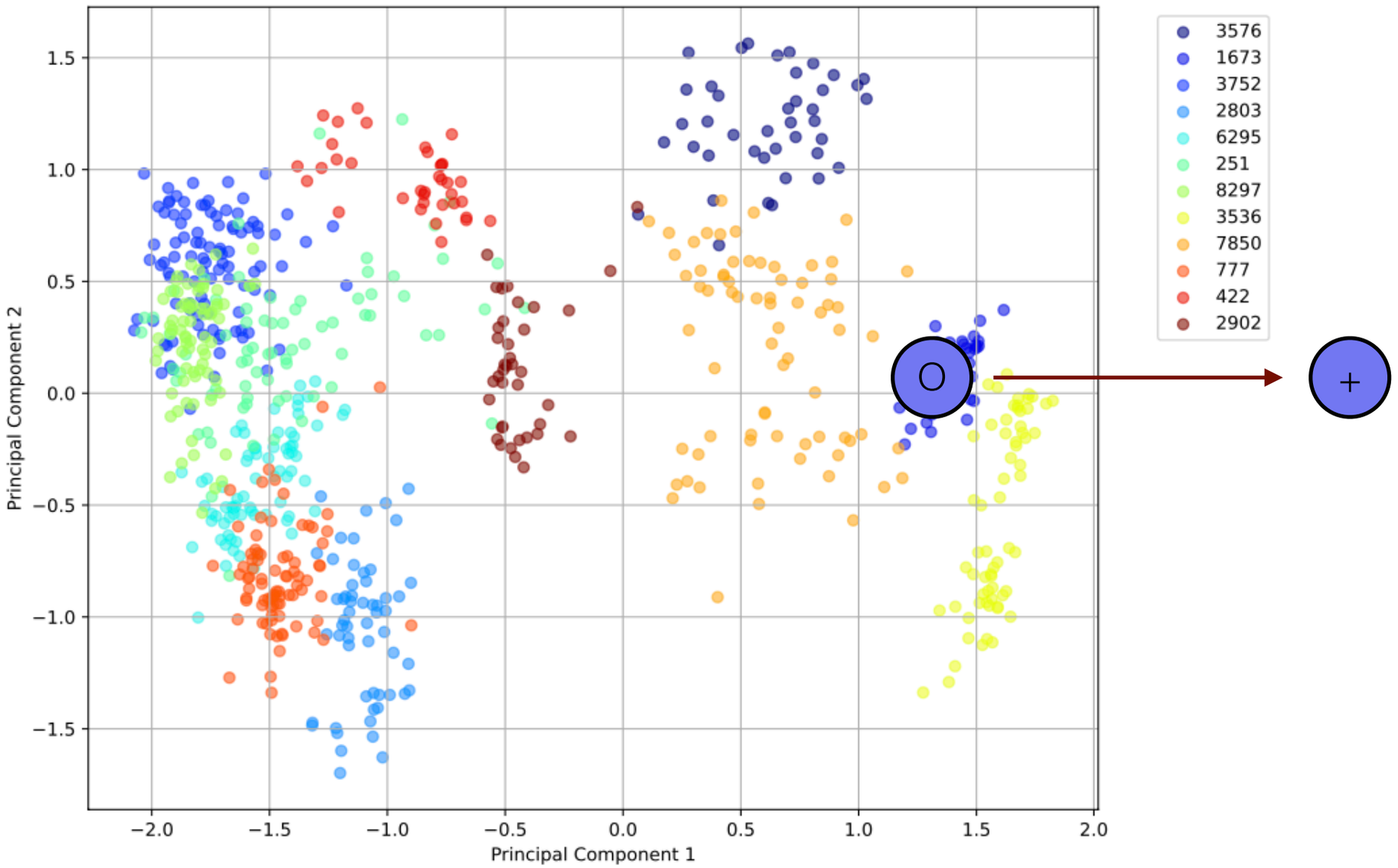
Global embedding in PCA space



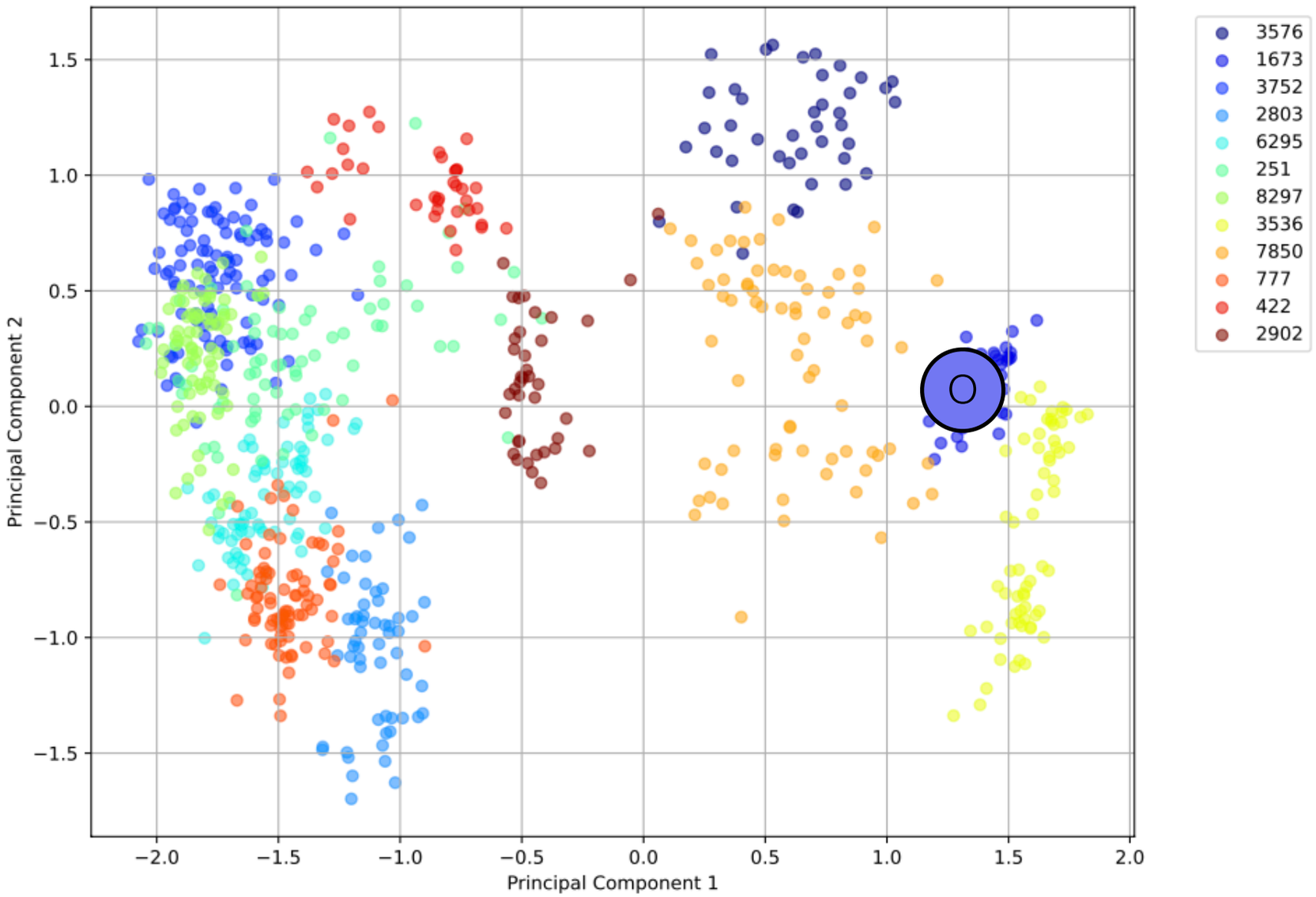
Global embedding in PCA space



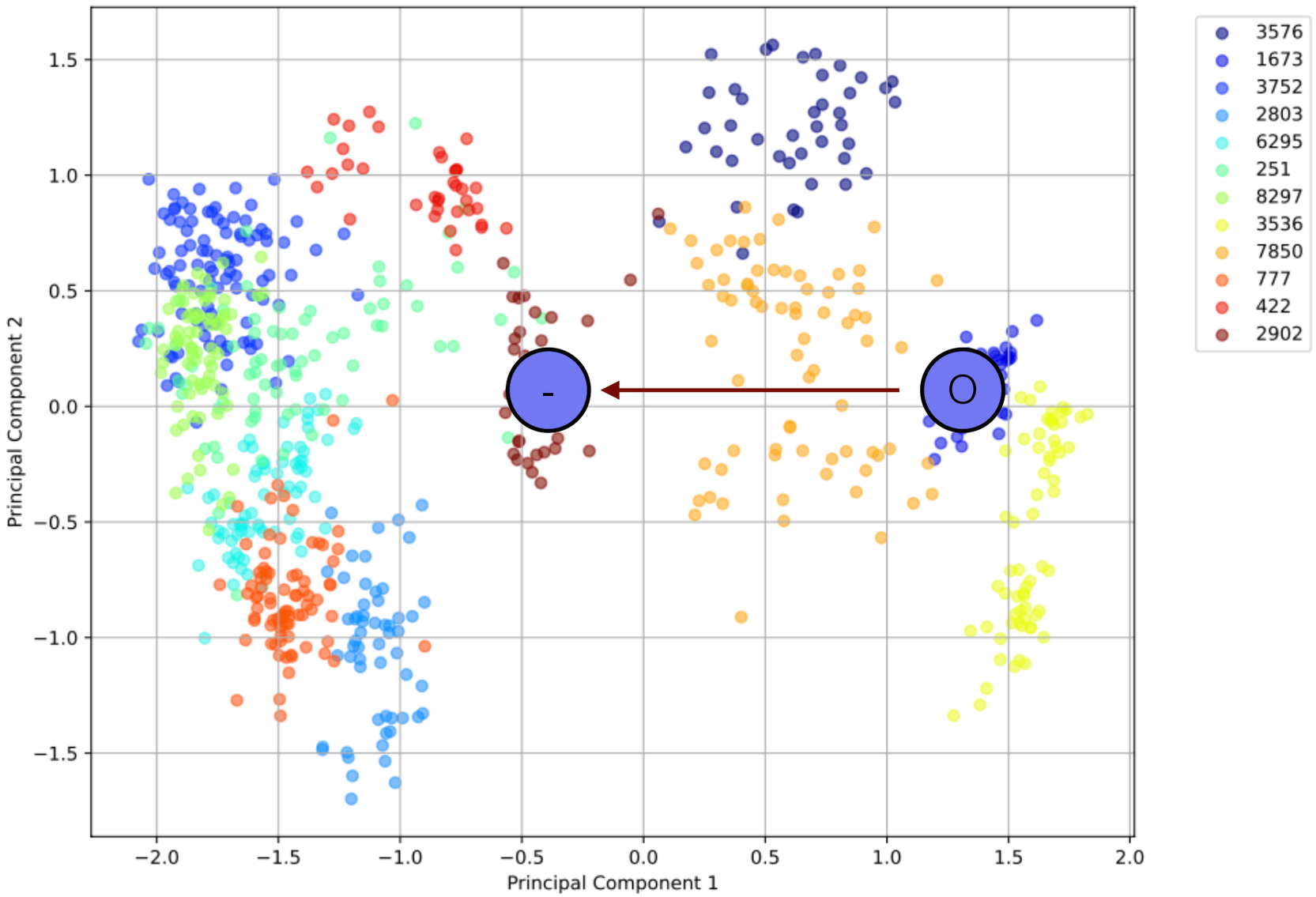
Global embedding in PCA space



Global embedding in PCA space

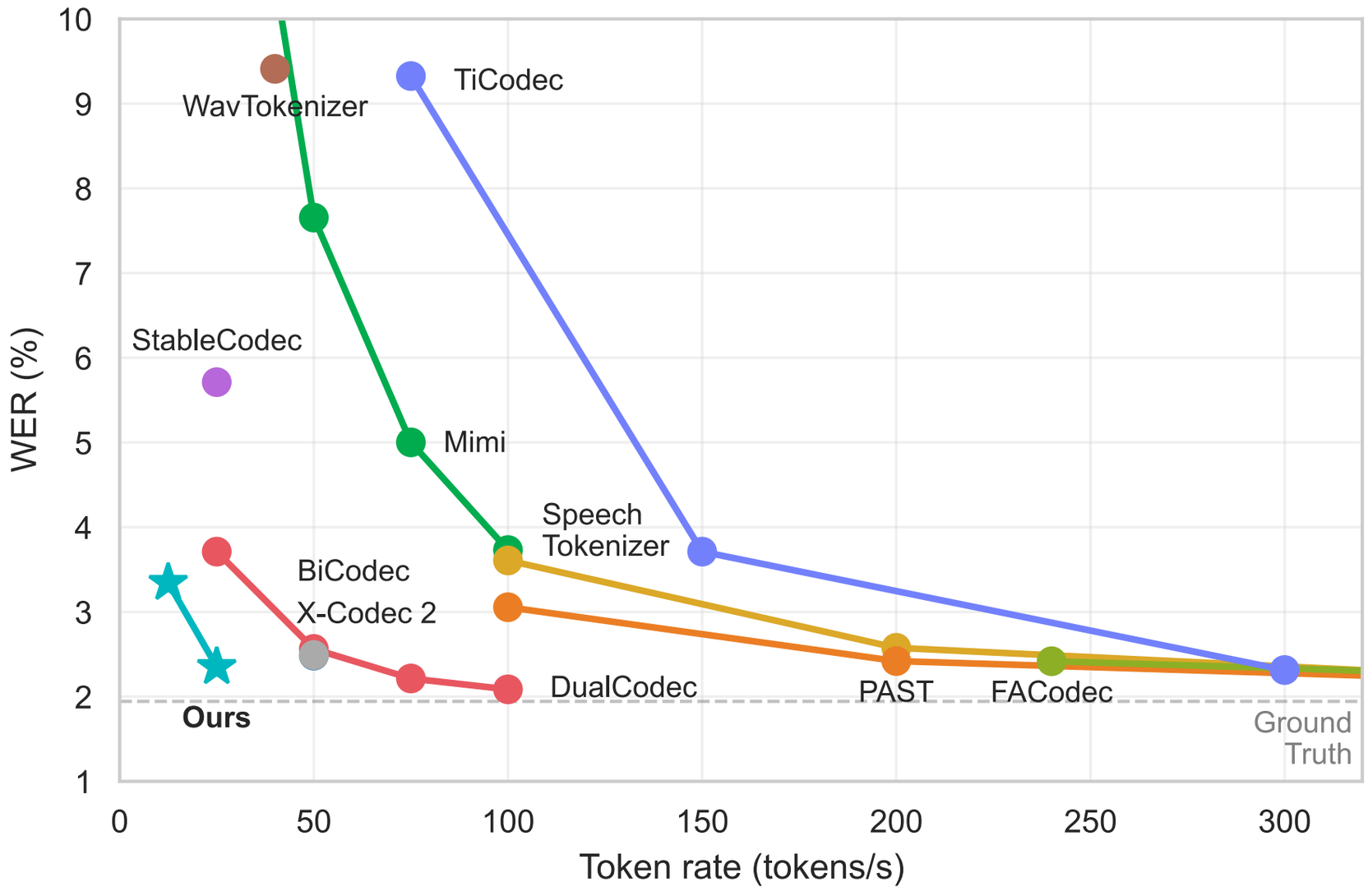


Global embedding in PCA space



Experiments and results

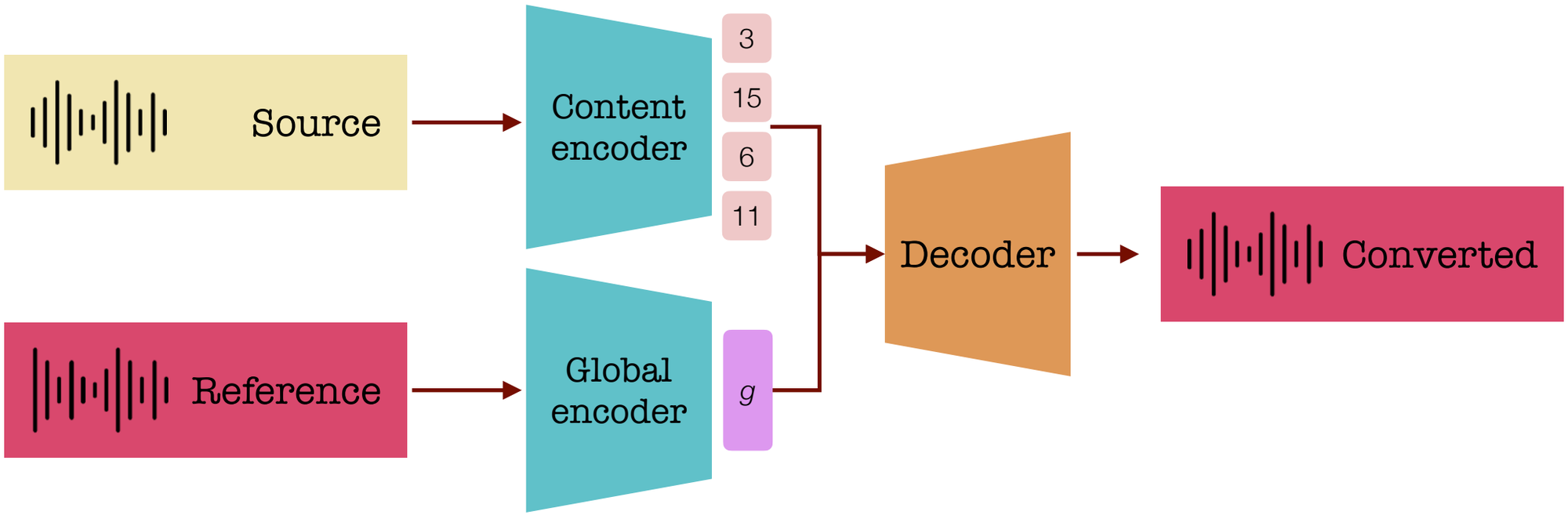
Reconstruction



Reconstruction

Model	Token Rate	Intelligibility		Quality		Spk.	Prosody
		WER↓	CER↓	MUSHRA↑	UTMOS↑	SIM↑	F0Corr↑
Ground Truth	–	1.9	0.6	78.0	4.07	–	–
Cont. 50Hz	–	2.0	0.6	76.7	3.90	98.9	0.94
KM 12.5Hz	12.5	3.0	1.1	72.1	4.04	95.5	0.66
KM 25Hz	25	2.7	1.0	72.4	4.07	95.6	0.67
Multi-layer							
FACodec	480	**2.1**	**0.7**	81.4	4.11	98.3	0.94
PAST	400	**2.1**	**0.7**	**82.4**	**4.18**	**98.8**	0.92
ST	400	**2.1**	**0.7**	76.0	3.90	98.1	0.92
DualCodec	100	**2.1**	**0.7**	75.6	4.12	98.3	**0.95**
Single-layer							
X-Codec 2	50	2.5	0.9	77.0	4.13	**97.8**	0.90
BiCodec	50	2.5	0.9	75.0	4.18	97.6	**0.91**
WavTokenizer	40	9.4	4.7	72.1	3.57	92.4	**0.91**
StableCodec	25	5.7	2.6	**79.3**	**4.31**	93.3	**0.91**
Kanade 12.5Hz	12.5	3.3	1.3	74.6	4.17	96.6	0.85
Kanade 25Hz	25	**2.4**	**0.8**	75.0	4.16	97.2	0.88

Voice conversion



Voice conversion

When the sunlight strikes raindrops in the air, they act as a prism and form a rainbow.

...six spoons of fresh snowpeas, five thick slabs of blue cheese, and maybe a snack for her brother bob

When the sunlight strikes raindrops in the air, they act as a prism and form a rainbow.

Voice conversion

When the sunlight strikes raindrops in the air, they act as a prism and form a rainbow.

...six spoons of fresh snowpeas, five thick slabs of blue cheese, and maybe a snack for her brother bob

When the sunlight strikes raindrops in the air, they act as a prism and form a rainbow.

Voice conversion

When the sunlight strikes raindrops in the air, they act as a prism and form a rainbow.

...six spoons of fresh snowpeas, five thick slabs of blue cheese, and maybe a snack for her brother bob

When the sunlight strikes raindrops in the air, they act as a prism and form a rainbow.

Voice conversion

When the sunlight strikes raindrops in the air, they act as a prism and form a rainbow.

...six spoons of fresh snowpeas, five thick slabs of blue cheese, and maybe a snack for her brother bob

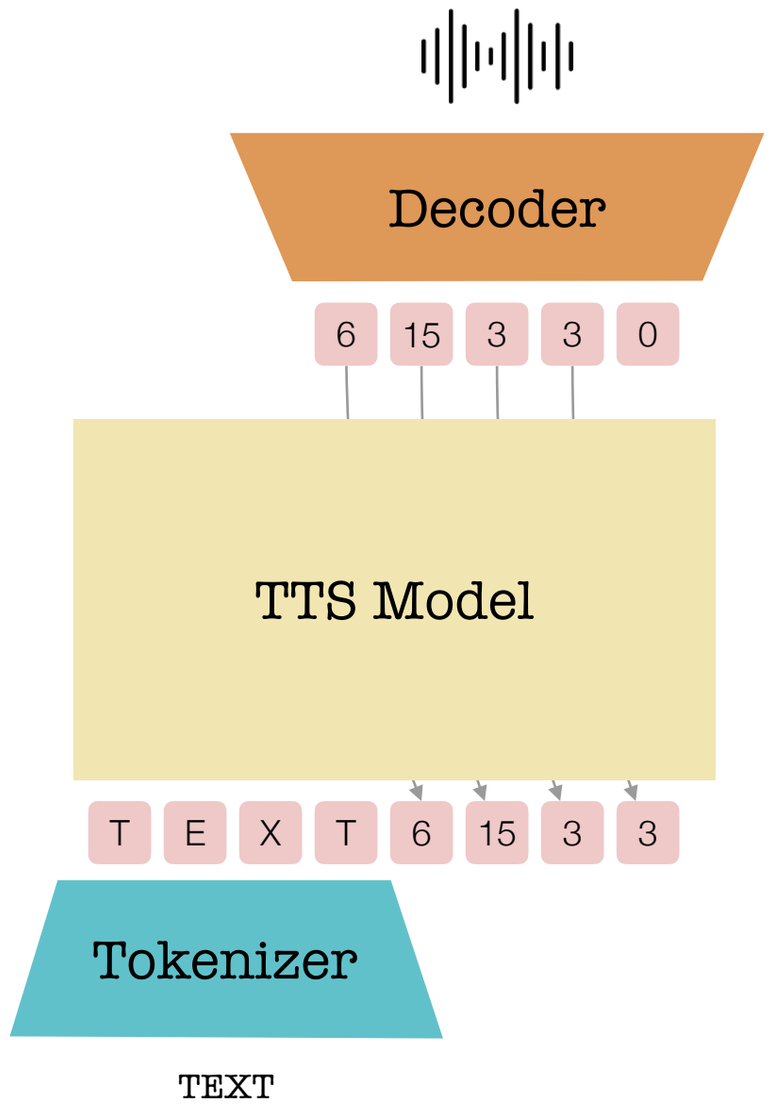
When the sunlight strikes raindrops in the air, they act as a prism and form a rainbow.

VC results

Model	Lexical		Quality	Speaker Timbre		Prosody
	WER↓	CER↓	UTMOS↑	EER↑	Similarity↑	F0Corr↑
Ground Truth	0.0	0.0	4.08	–	–	–
LinearVC	0.6	0.2	3.94	29.7	73.4	0.62
FreeVC	0.6	0.3	3.99	29.0	74.5	0.67
CosyVoice 2	1.1	0.5	4.11	31.0	76.0	0.64
TiCodec ↛	**0.5**	**0.2**	3.32	5.4	68.0	0.77
FACodec ↛	0.7	0.4	3.77	15.2	62.6	**0.79**
BiCodec ↛	1.2	0.6	3.84	18.5	71.4	0.61
DualCodec ↛≠	21.5	12.9	2.51	6.8	52.0	0.54
PAST ↛≠	22.9	15.1	1.84	8.2	23.3	0.20
ST ↛≠	74.7	61.7	1.54	10.6	35.0	0.20
Mimi ≠	120.3	86.8	3.09	**38.5**	**81.7**	0.21
Kanade 12.5Hz	1.6	0.7	**4.17**	32.0	77.6	0.64
Kanade 25Hz	0.7	0.3	4.16	30.7	77.1	0.71

↛ and ≠ denote models exhibiting poor timbre transfer and serious content degradation, respectively.

Text-to-speech



Text-to-speech results

Model	LibriTTS test-clean					Seed-TTS-eval	
	WER↓	SIM↑	UTMOS↑	Quality↑	Prosody↑	WER↓	SIM↑
Ground Truth	2.3	–	4.13	74.9	80.9	1.9	–
CosyVoice 2	1.8	95.5	4.42	77.1	83.0	2.1	65.6
KM 12.5Hz	4.6	94.7	3.96	72.0	67.0	5.4	41.9
KM 25Hz	4.3	94.9	4.05	74.9	75.9	4.9	42.0
WavTokenizer	13.9	92.0	3.76	74.5	77.0	15.6	27.5
TiCodec	11.5	93.9	3.86	73.8	72.9	12.9	32.5
DualCodec	10.0	**95.7**	3.68	73.0	80.0	5.5	34.3
ST	9.7	95.3	3.95	75.0	79.0	11.2	35.2
StableCodec	9.0	90.7	3.78	71.0	66.0	10.9	22.6
PAST	8.0	95.4	4.14	74.9	78.4	9.0	35.4
BiCodec	7.8	95.3	4.12	73.8	78.9	7.5	45.6
Mimi	6.6	94.6	3.48	74.9	73.9	6.0	31.6
X-Codec 2	6.5	95.2	**4.21**	72.0	78.0	7.2	34.7
Kanade 12.5Hz	5.9	95.2	4.13	**77.1**	77.9	5.7	46.8
Kanade 25Hz	**4.2**	95.4	4.18	73.0	**81.0**	**4.0**	**48.0**

Unconditional generation (pure SLM) results

Model	Tok. Rate	Vocab.	sWUGGY↑	sBLIMP↑	sSC↑	tSC↑
KM 12.5Hz	12.5	12800	75.8	**57.5**	51.8	66.7
KM 25Hz	25	12800	68.1	53.5	51.1	63.5
ST	50	1024	75.8	54.9	52.0	64.4
PAST	50	1024	76.8	53.6	51.8	59.5
Mimi	12.5	2048	**77.6**	56.1	52.0	**67.8**
Kanade 12.5Hz	12.5	12800	76.6	55.2	**52.1**	65.3
Kanade 25Hz	25	12800	69.7	52.4	51.3	60.0

Unconditional generation (pure SLM) results

Model	Tok. Rate	Vocab.	sWUGGY↑	sBLIMP↑	sSC↑	tSC↑
KM 12.5Hz	12.5	12800	75.8	**57.5**	51.8	66.7
KM 25Hz	25	12800	68.1	53.5	51.1	63.5
ST	50	1024	75.8	54.9	52.0	64.4
PAST	50	1024	76.8	53.6	51.8	59.5
Mimi	12.5	2048	**77.6**	56.1	52.0	**67.8**
Kanade 12.5Hz	12.5	12800	76.6	55.2	**52.1**	65.3
Kanade 25Hz	25	12800	69.7	52.4	51.3	60.0

→ note: these are all text-based metrics

Conclusion

- Spoken language models need good speech tokenizers that
  - Include phonetic and prosodic information
  - Suppress non-linguistic information
  - Enable high-quality synthesis
- Kanade fulfills these requirements using a simple method for disentanglement

Future work

- Speech representation
  - Variable-length tokenization (e.g., Sylber, ZeroSyl)
  - What to do with
    - multiple, possibly overlapping speakers?
    - dynamic background noise?
  - Disentanglement of prosody
  - Quantization-free methods
- Objective metrics of speech quality beyond WER and audio quality

Future work

- Speech representation
  - Variable-length tokenization (e.g., Sylber, ZeroSyl)
  - What to do with
    - multiple, possibly overlapping speakers?
    - dynamic background noise?
  - Disentanglement of prosody
  - Quantization-free methods
- Objective metrics of speech quality beyond WER and audio quality

→ if you're interested in collaborating, contact me!

Thank you!

Paper https://arxiv.org/abs/2602.00594
Demo https://frothywater.github.io/kanade-tokenizer/
Code https://github.com/frothywater/kanade-tokenizer