National Taiwan University

Unsupervised End-to-End Learning of Discrete Linguistic Units for Voice Conversion

Andy T. Liu, Po-chun Hsu, Hung-yi Lee

National Taiwan University

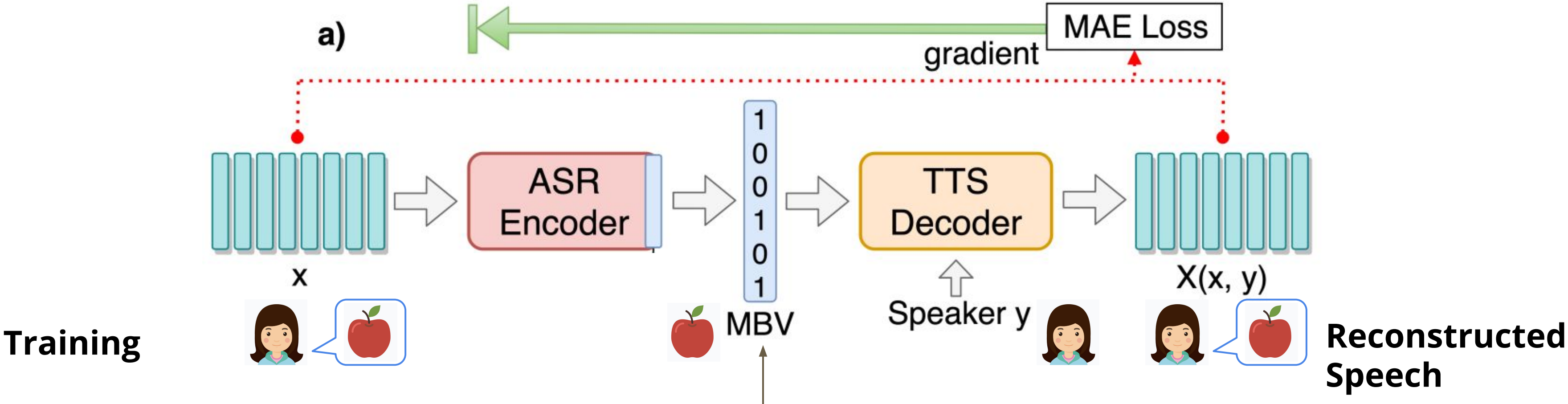
2019/09/17

Highlights

- We present an **unsupervised end-to-end ASR-TTS autoencoder framework**, where we discover discrete subword units from speech without using any labels.
- **Contributions:**
    - Present a discrete encoding method that outperforms continuous encodings.
    - Able to dientangle speech content from speaker style automatically.
    - Achieved many-to-many voice conversion without using any parallel data.

- In our subjective and objective evaluations, we show that **VC quality is improved** when compared to continuous representations (Chao et. el).
- In ZeroSpeech 2019, the proposed method achieved **2nd place** in terms of low bitrate.

Proposed Method (½) - Discrete linguistic units discovery



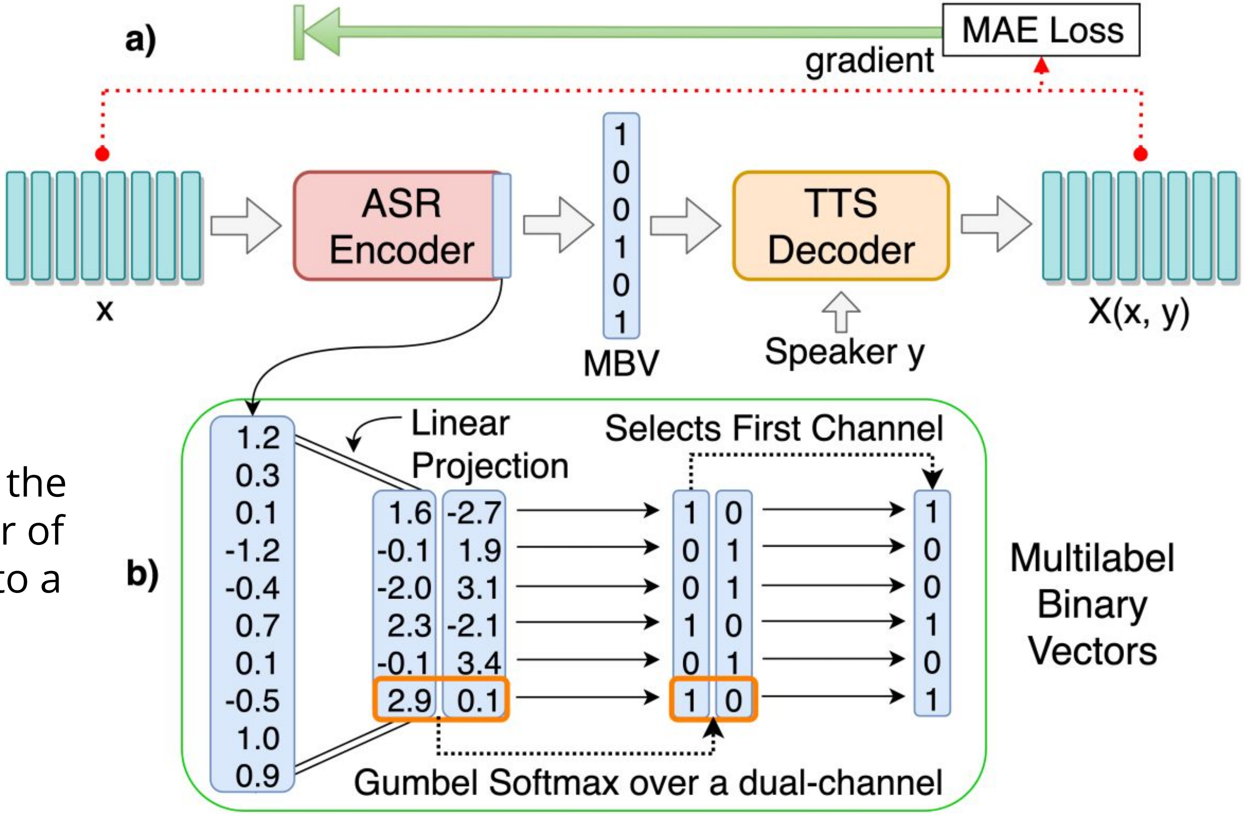
ASR-Encoder is trained to discover a set of common linguistic units given a variety of speakers

TTS-Decoder is trained to project the discovered units back to the designated speech

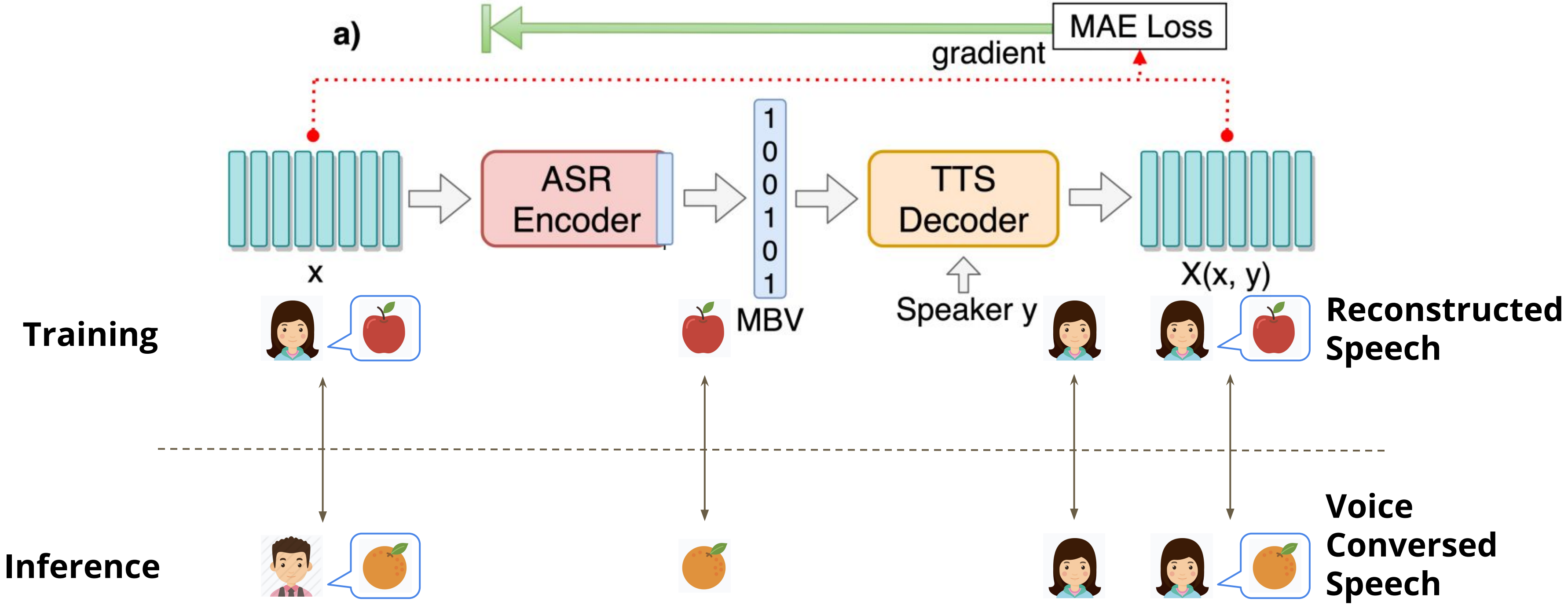
In this **unsupervised end-to-end** manner, **discrete linguistic units** are learned and **represented as multilabel binary vectors (MBVs)**.

Proposed Method (½) - MBV: discrete vectors of zeros one ones

Linearly project the last hidden layer of ASR-Encoder into a $\mathbb{R}^{n\times 2}$ space.



Proposed Method (½) - VC using the ASR-TTS autoencoder

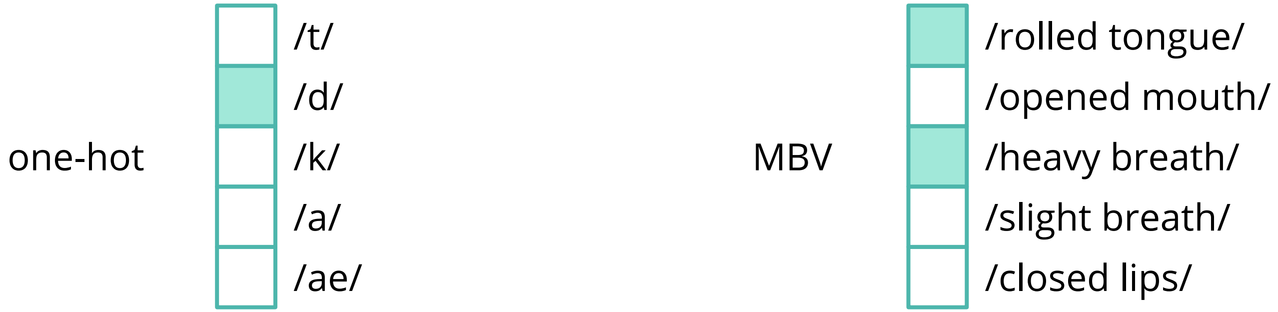


The voice converted speech would sound like uttering 's content.

Why Multilabel-Binary Vectors ? (MBV)

Both one-hot and MBV are discrete, however

each dimension of **one-hot** vector corresponds to a **linguistic unit** (phoneme), while each dimension of **MBV** may corresponds to a **pronunciation attribute**.



This makes MBV more data efficient than one-hot vectors as a linguistic unit,
We also verified that one-hot vectors is incapable for this task.

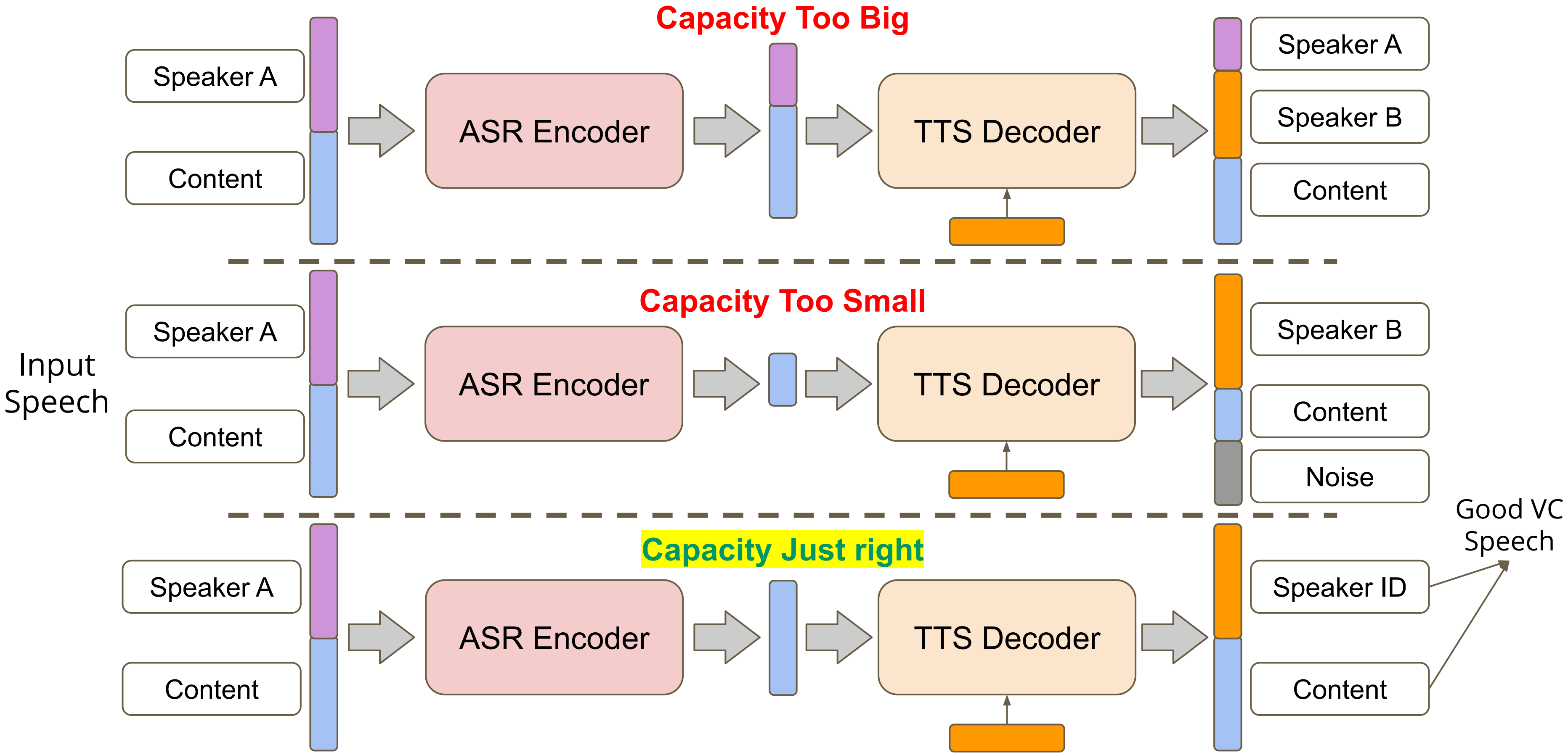
Why does it work?

How can the model automatically learn
how to disentangle speech content form speaker identity?

**If the bottleneck dimension of the ASR-Encoder is set just right,
so that there is just enough capacity to encode the content,
ideal VC output can be achieved.**

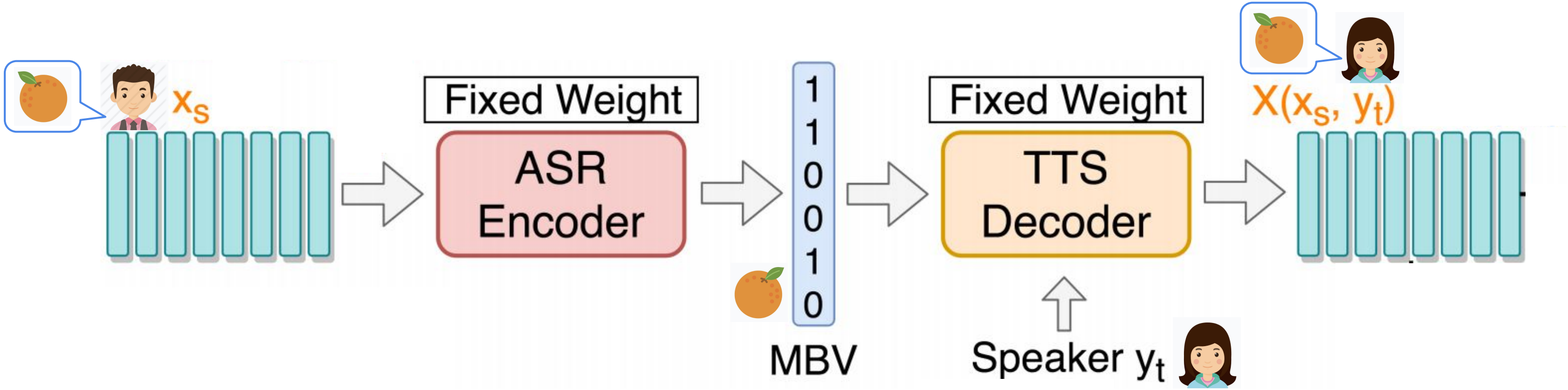
This is later formally proved by Qian et. al (ICML 2019),
where they used RNN + downsampling to form the bottleneck,
in comparison we use discrete encodings to form this bottleneck.

Bottleneck Visualization



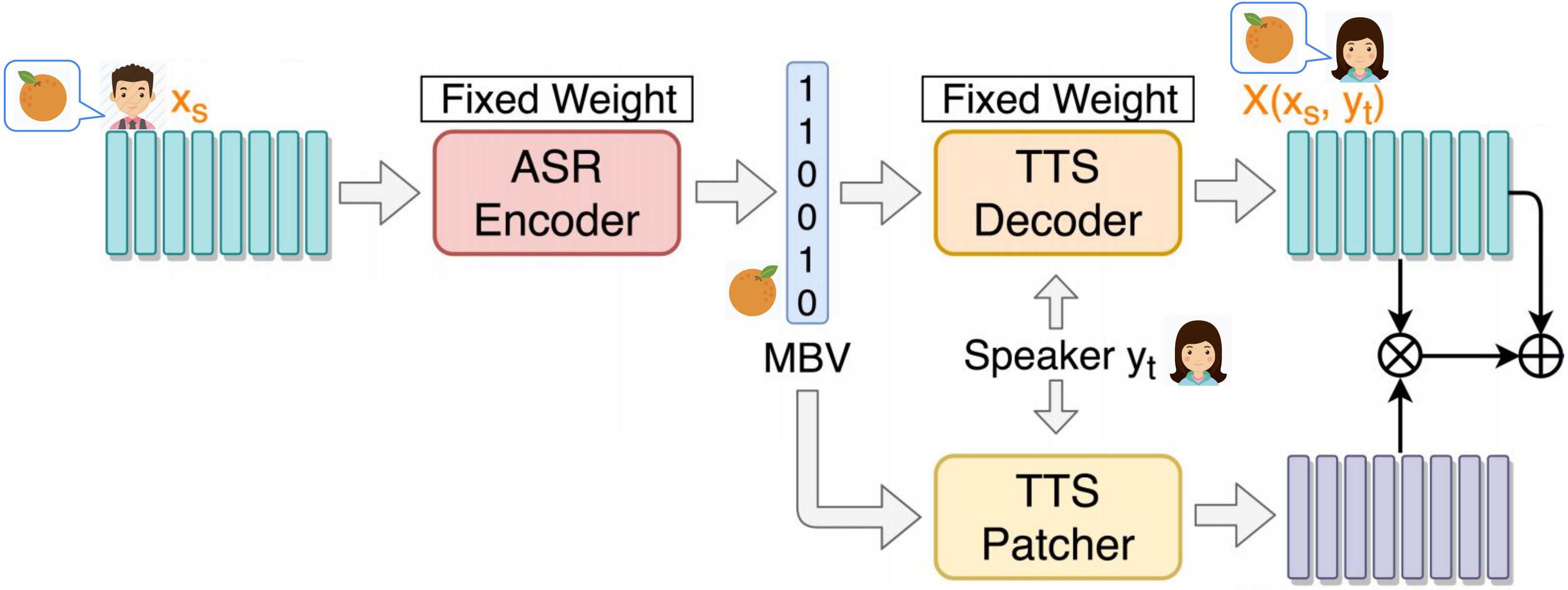
Proposed Method (2/2) - Target guided adversarial learning

Step 0: Given the trained ASR-TTS autoencoder framework shown previously



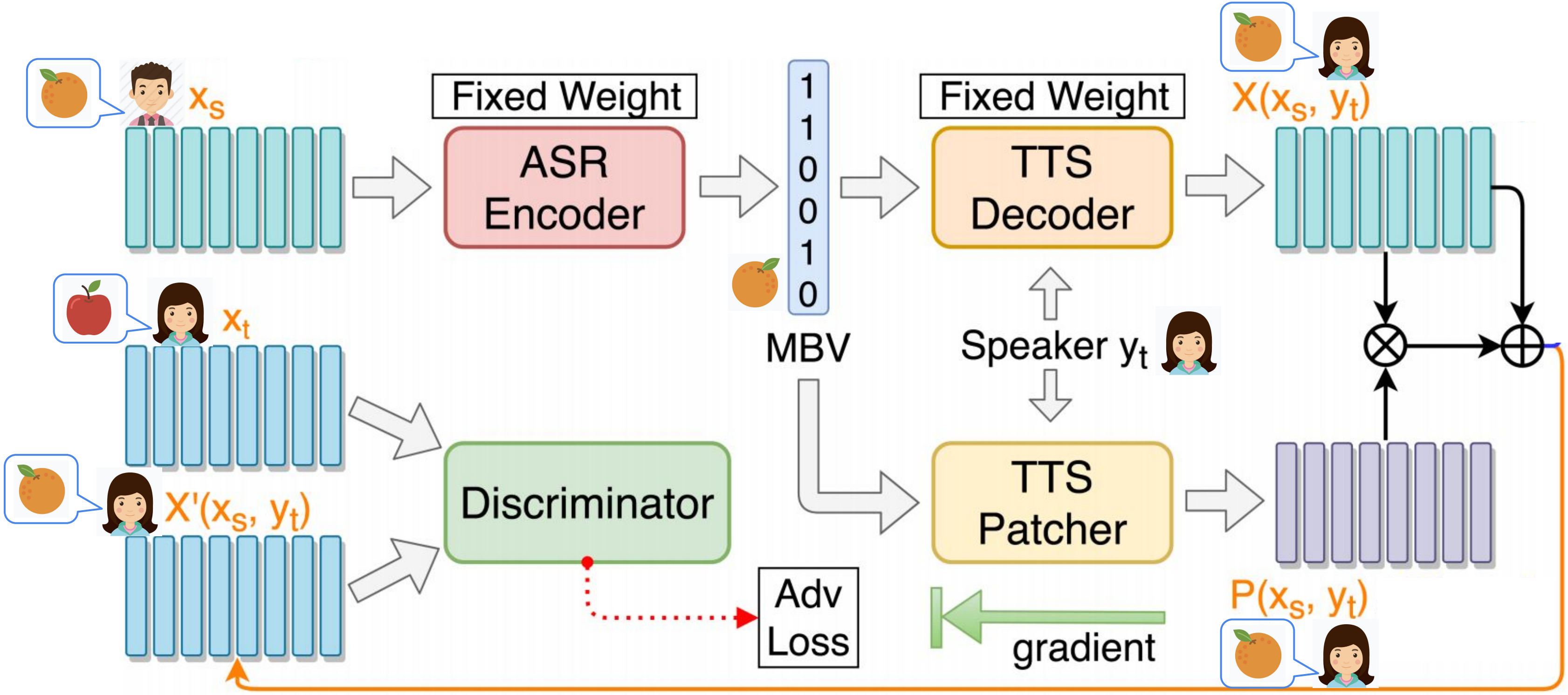
Proposed Method (2/2) - Target guided adversarial learning

Step 1: **Add a TTS-Patcher on top of it to improve VC quality**

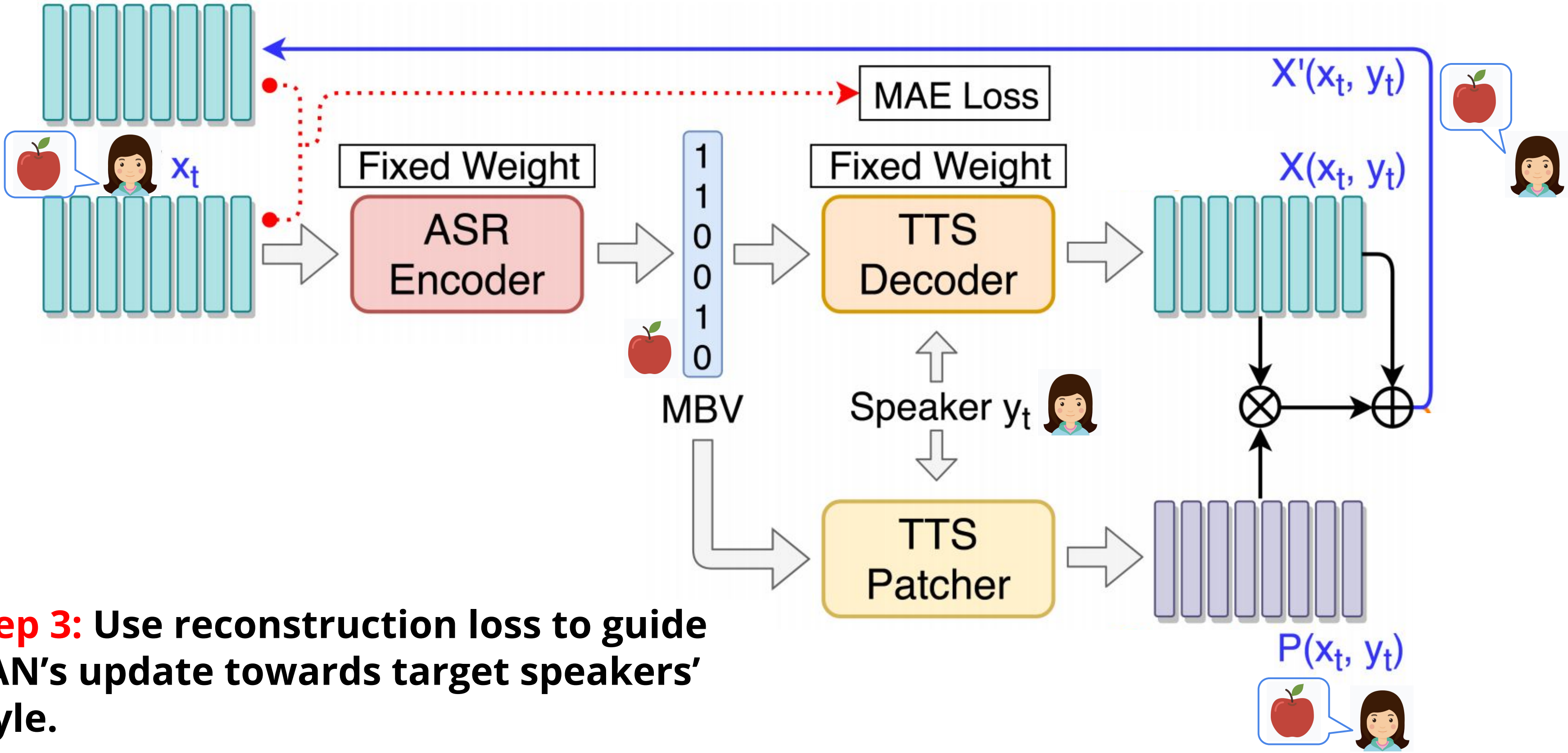


Proposed Method (2/2) - Target guided adversarial learning

Step 2: **Train the TTS-Patcher (Generator) in the framework of GAN**

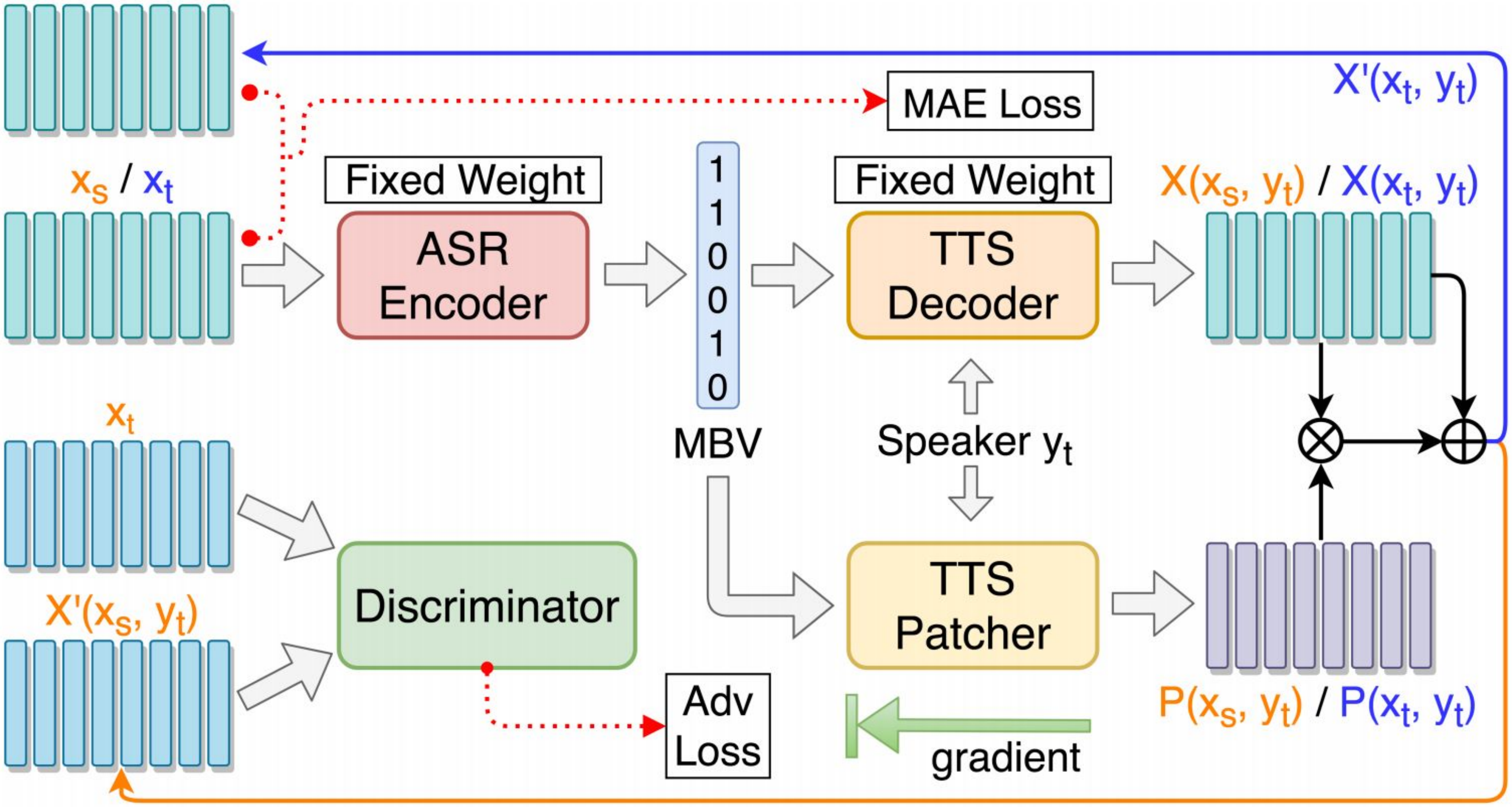


Proposed Method (2/2) - Target guided adversarial learning



Step 3: **Use reconstruction loss to guide GAN's update towards target speakers' style.**

Proposed Method (2/2) - Target guided adversarial learning



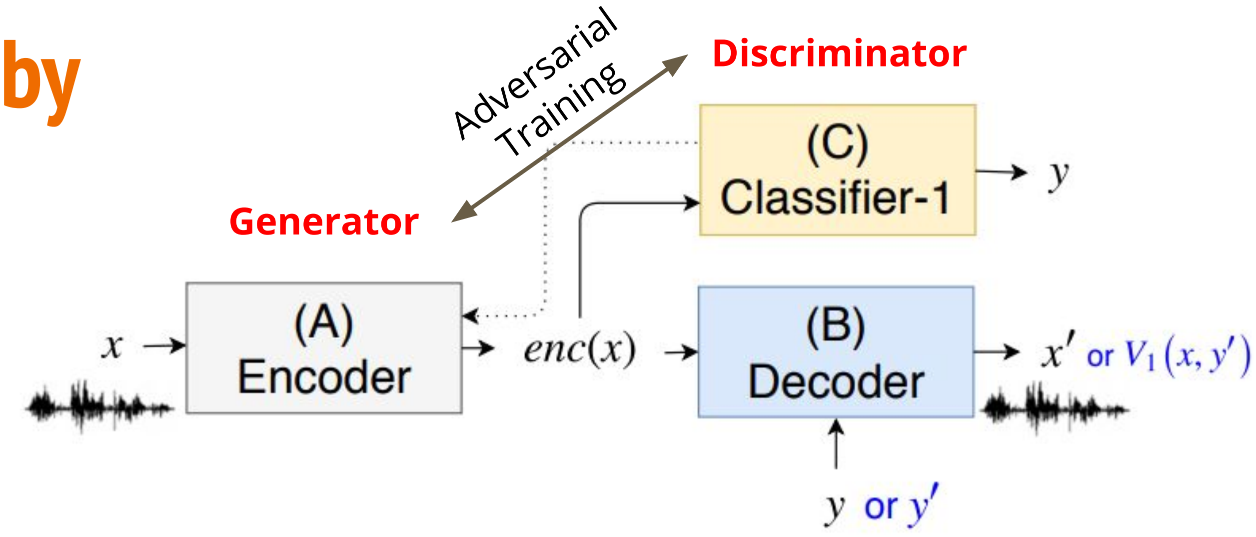
Experiment - Setups

- **The ZeroSpeech 2019 Challenge provides two datasets:**
  - Development English Set
  - Testing Suprise Set

- **For our experiments:**
  We use only the English set (Voice / Unit set, no parallel data are used) for training, and evaluate our model on the English Test set.

- **For ZeroSpeech 2019 Challenge:**
  Tune our model's hyperparameters with the Development English set, and use those hyperparameters to train our Surprise language model.

Previous Work by Chao et al.
(Voice Conversion)



Chou et al. (2018) introduce a **classifier** on the latent code, which needs to be trained in a **GAN** setting.

However, **we did not use additional supervision on the latent code** and achieved improved results.

Experiment - Degree of Disentanglement

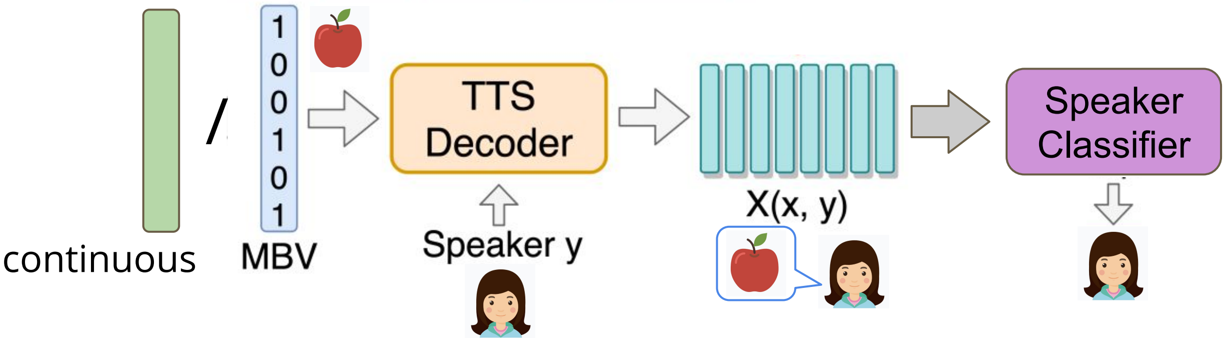


Table 2: *Comparison of different latent representations.*

Incapable for this task

Types of encodings	**Dim**	**Acc**
One-hot	1024	43.3%
continuous	1024	84.1%
continuous	128	79.9%
continuous (with add'l loss)	1024	78%
continuous (with add'l loss)	128	81.3%
Ours (MBV)	1024	**92.3%**
Ours (MBV)	128	**93.9%**

A disentangled representation should produce voice similar to the target speakers and leads to higher classification accuracy.

Experiment - Subjective Evaluation

- Human participants are required to grade each method on a 1 to 5 scale under two measures:
  - **Naturalness**
    Whether the converted speech is human-like.
  - **Similarity**
    Whether the converted speech's has similar speaker characteristics to the target speaker.

Table 3: *Results of subjective human evaluation. All methods used an encoding dimension of 1024 if not specified otherwise.*

Types of encodings	**naturalness**	**similarity**
continuous	3.80	2.14
continuous (with add'l loss)	3.21	2.58
Ours (MBV with dim 6)	1.61	1.51
Ours (MBV)	3.36	**3.06**
Ours (with adv. training)	2.57	**3.15**

Experiment - Subjective Evaluation

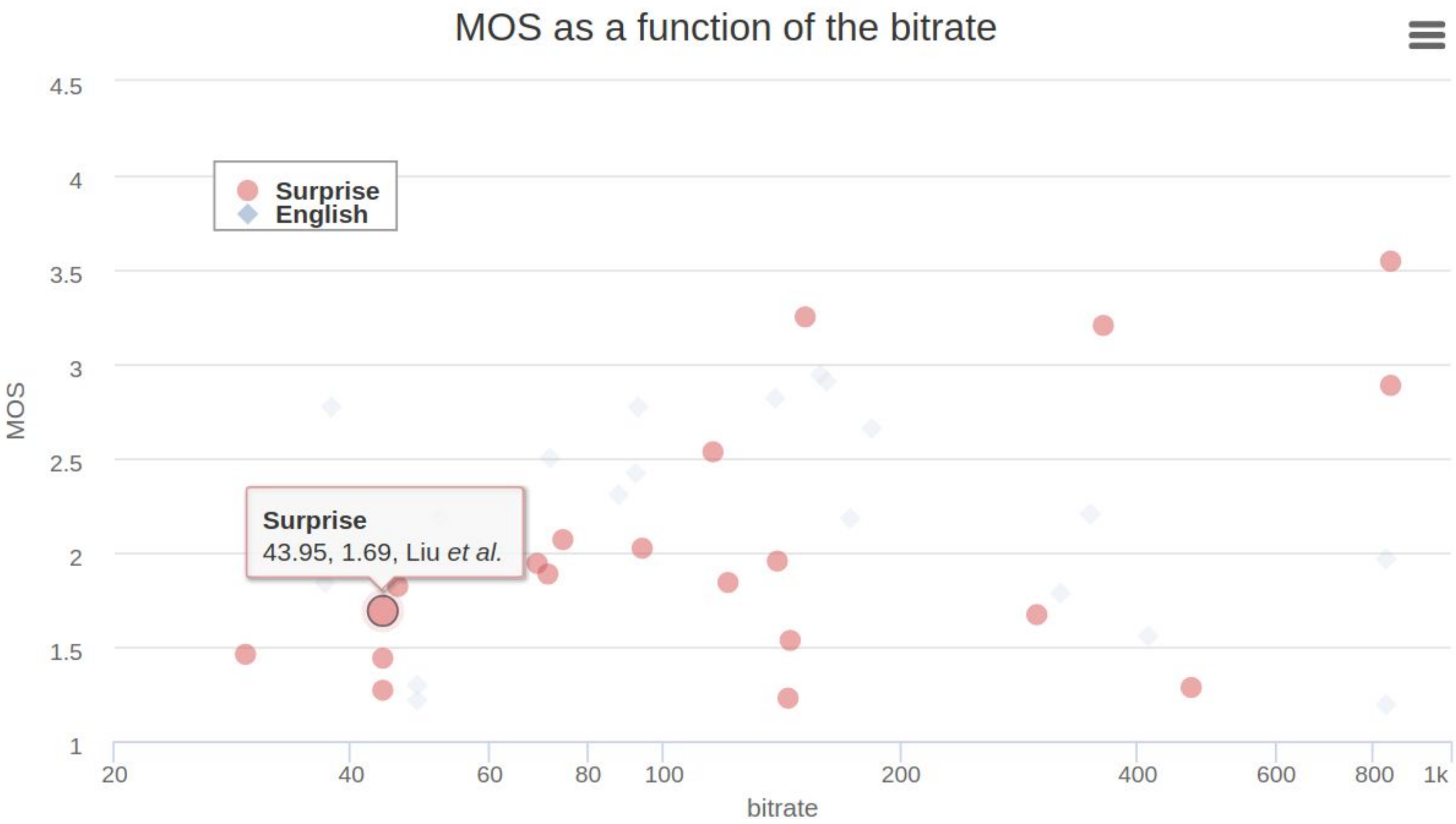
The setting we used to compete in the ZeroSpeech Challenge

- Human participants are required to grade each method on a 1 to 5 scale under two measures:
    - **Naturalness**
      Whether the converted speech is human-like.
    - **Similarity**
      Whether the converted speech's has similar speaker characteristics to the target speaker.

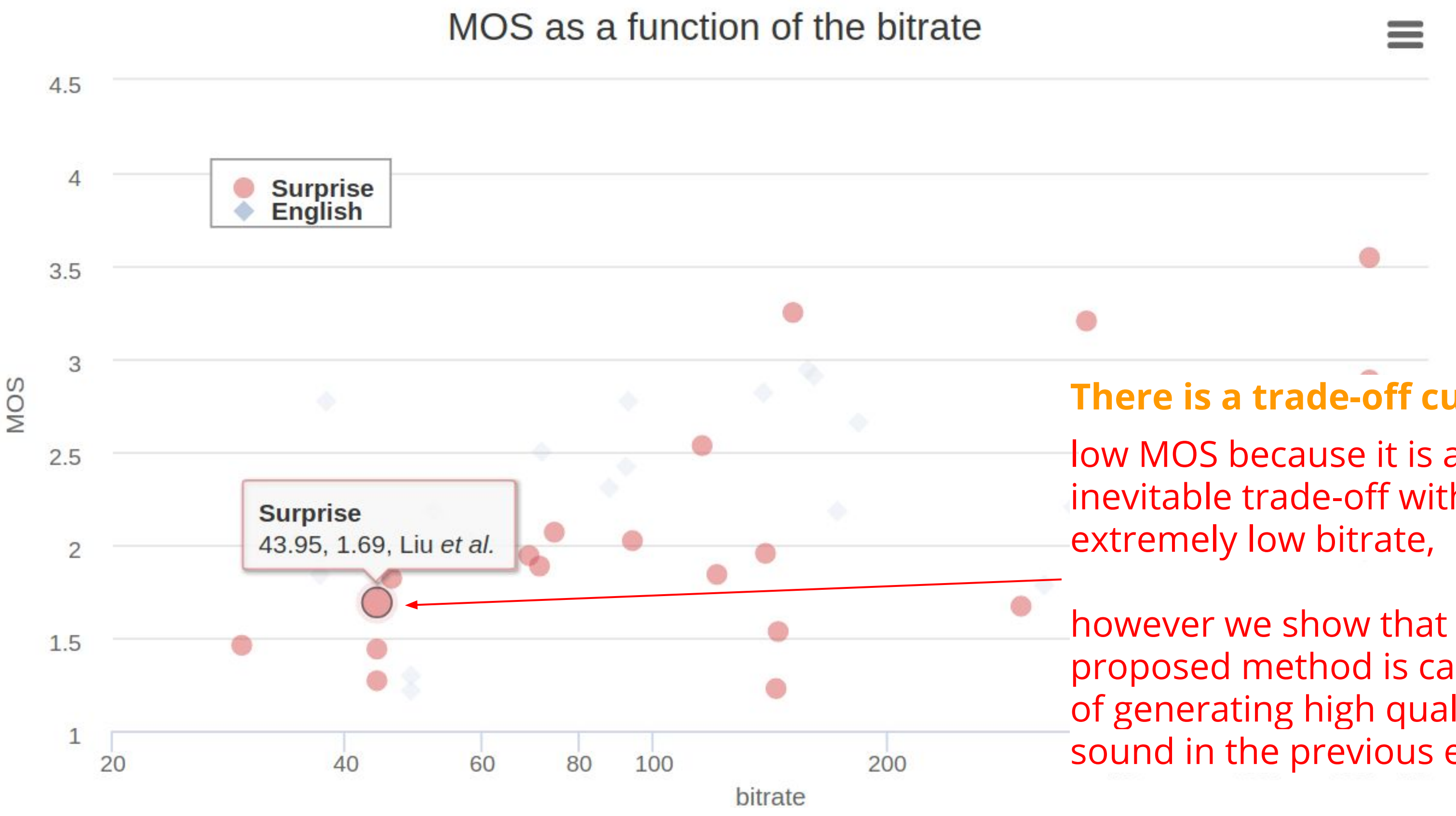
Table 3: *Results of subjective human evaluation. All methods used an encoding dimension of 1024 if not specified otherwise.*

Types of encodings	**naturalness**	**similarity**
continuous	3.80	2.14
continuous (with add'l loss)	3.21	2.58
Ours (MBV with dim 6)	1.61	1.51
Ours (MBV)	3.36	**3.06**
Ours (with adv. training)	2.57	**3.15**

Surprise Set Leaderboard - ZeroSpeech 2019 Challenge



Surprise Set Leaderboard - ZeroSpeech 2019 Challenge



There is a trade-off curve

low MOS because it is an inevitable trade-off with extremely low bitrate,

however we show that the proposed method is capable of generating high quality sound in the previous exp.

Experiment - Encoding Dimension Analysis

The proposed method achieves **lower "bit rate" and "distinct units" with comparable ABX scores.**

More analysis can be found in our paper.

Table 4: *Performance of different encoding dimensions.*

Method	Dim	CER	BR	ABX	distinct
Baseline	200	1.000	71.98	35.90	65
Cont.	1024	0.036	138.45	31.83	16849
	128	0.040	138.45	33.96	16849
Ours	1024	0.196	138.45	32.02	16849
	512	0.313	138.45	32.82	16849
	256	0.430	138.45	32.52	16849
	128	0.629	138.45	31.58	16849
	64	0.717	138.35	32.57	16772
	32	0.797	134.80	31.82	14591
	16	0.887	105.96	35.62	3723
	8	0.998	61.79	38.10	146
	7	0.998	55.97	37.71	94
	6	1.000	**48.78**	39.60	**51**
	5	1.000	**41.32**	41.79	**28**

Conclusion

- The proposed encoding method MBV offers a strong bottleneck for content extraction in VC.
- As a result strong VC performance is achieved as speaker identity is eliminated from extracted encodings, while speech content is preserved.
- In the ZeroSpeech 2019 Challenge Surprise Dataset Leaderboard, the proposed method achieved outstanding results in terms of low bitrate.

Thank You

Q&A