Inter-University Research Institute Corporation /
Research Organization of Information and Systems
National Institute of Informatics

ICASSP 2024
SLP.L20.2

Can Large-scale Vocoded Spoofed Data Improve Speech Spoofing Countermeasure with a Self-supervised Front End?

/wʌn/ /ʃɪn/
Wang Xin, Junichi Yamagishi

National Institute of Informatics

Summary in one slide

- Extension of previous work (Wang 2023)
  - Not use any spoofed training data from text-to-speech or voice conversion
- Method
  - Upstream SSL training, using **vocoded VoxCeleb2**
  - Downstream SSL fine-tuning, using vocoded ASVspoof19
- Our best overall results

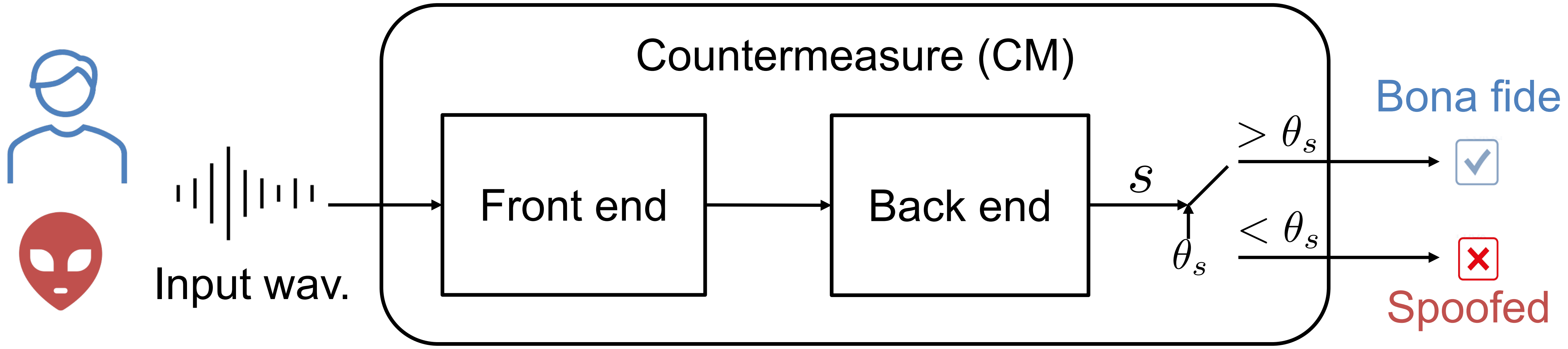
Wang, Xin, and Junichi Yamagishi. “Spoofed Training Data for Speech Spoofing Countermeasure Can Be Efficiently Created Using Neural Vocoders.” In Proc. ICASSP, 1–5. IEEE, 2023.

Introduction

❑ A binary classification task

- Bona fide: human voice
- Spoofed: text-to-speech (TTS) or voice conversion (VC) voice

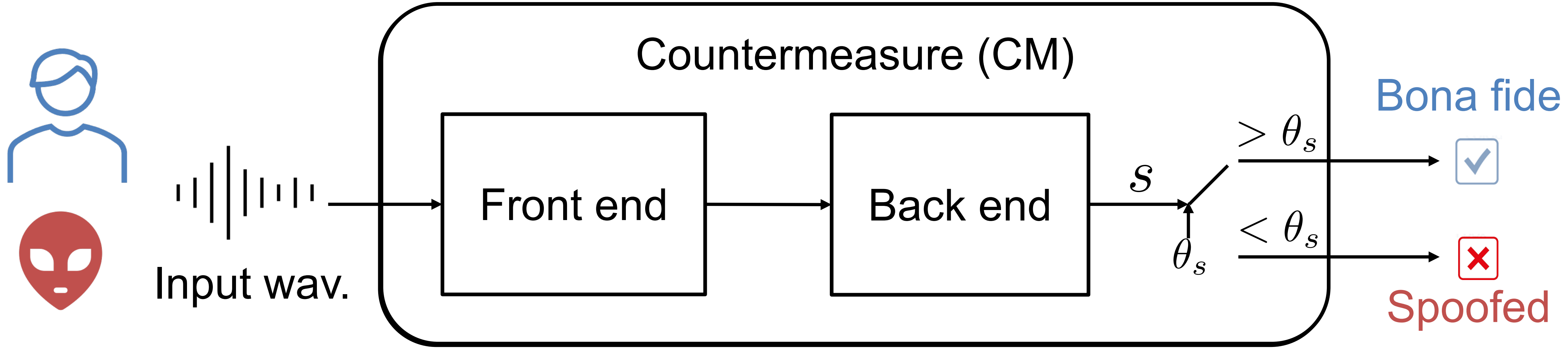
- Metric: equal error rate (EER)



Introduction

❑ CM architecture in our studies (Wang 2022, Tak 2022)

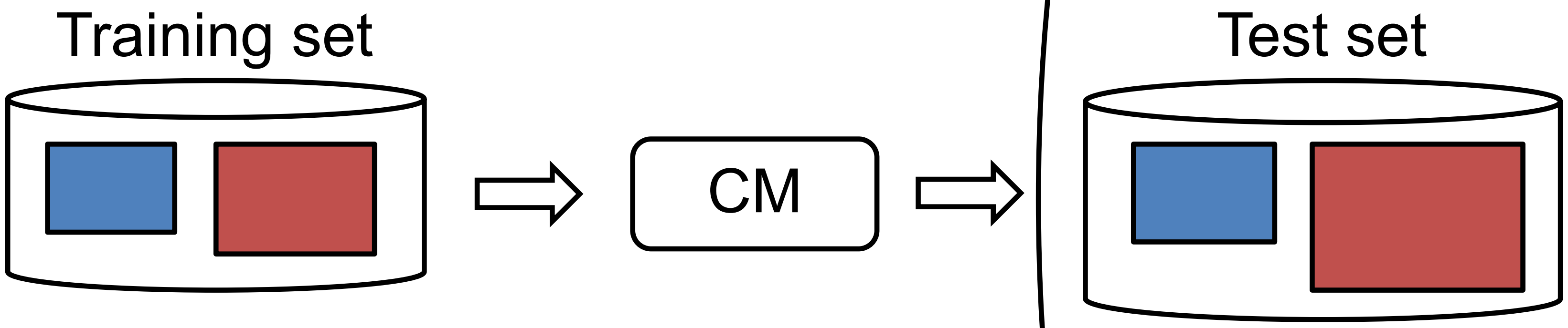
- **Front end**: self-supervissed learning (SSL) model
  - wav2vec 2.0 XLSR-53 (Conneau 2021)
- **Back end**: global average pooling + 4-layer neural network



Conneau, Alexis, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. "Unsupervised Cross-Lingual Representation Learning for Speech Recognition." In *Proc. Interspeech*, 2426–30. ISCA, 2021.
Wang, Xin, and Junichi Yamagishi. "Investigating Self-Supervised Front Ends for Speech Spoofing Countermeasures." In *Proc. Odyssey*, 100–106, 2022..

Generalization is desired

Space of all possible bona fide and spoofed data

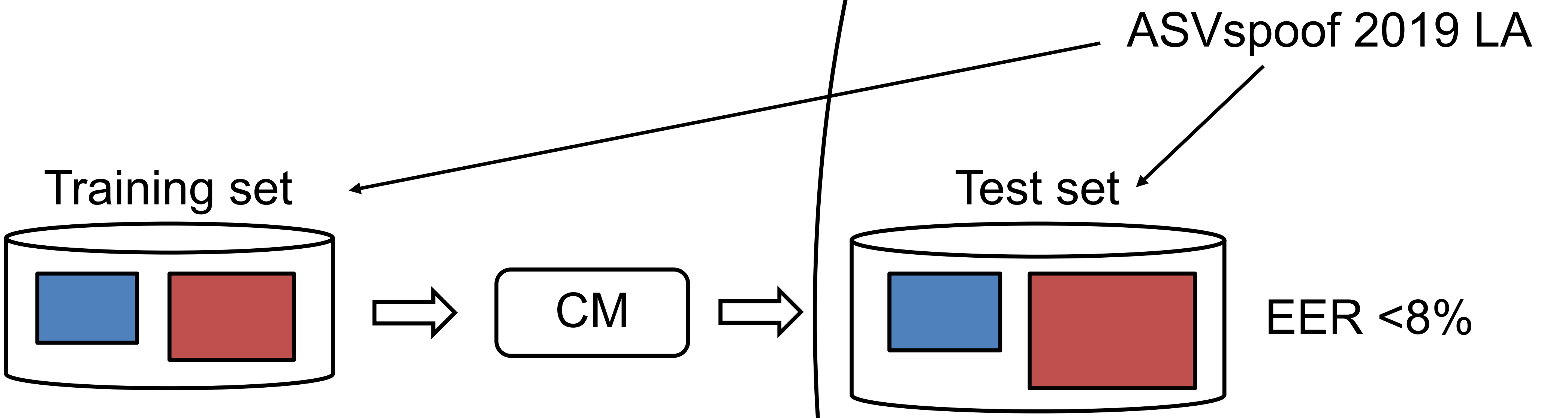


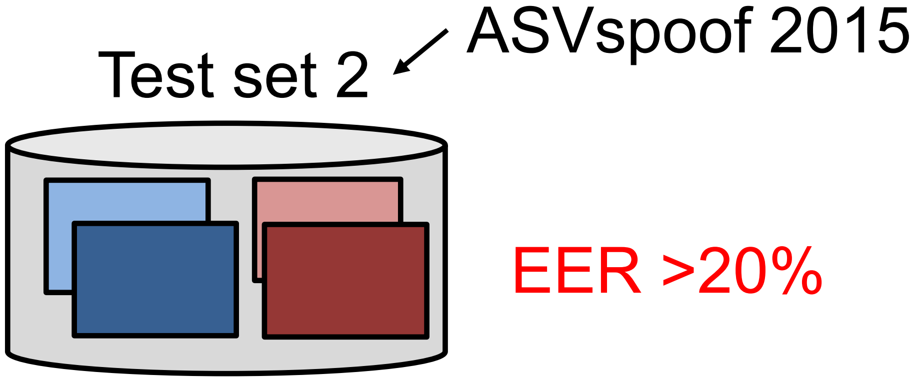
En, Fr, Ch, Jp, …

Wav, mp3, m4a …

New TTS/VC methods

Generalization is challenging

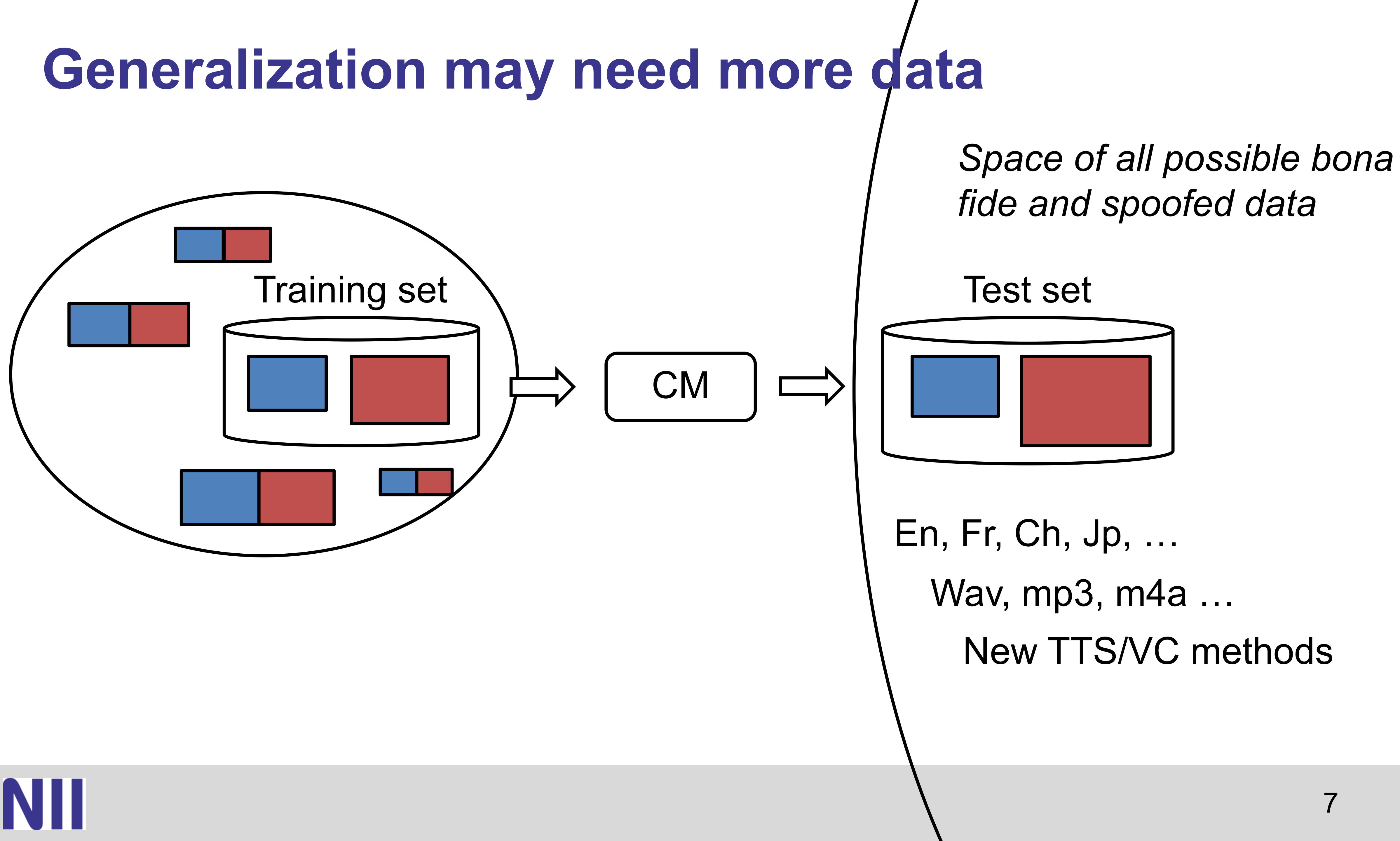




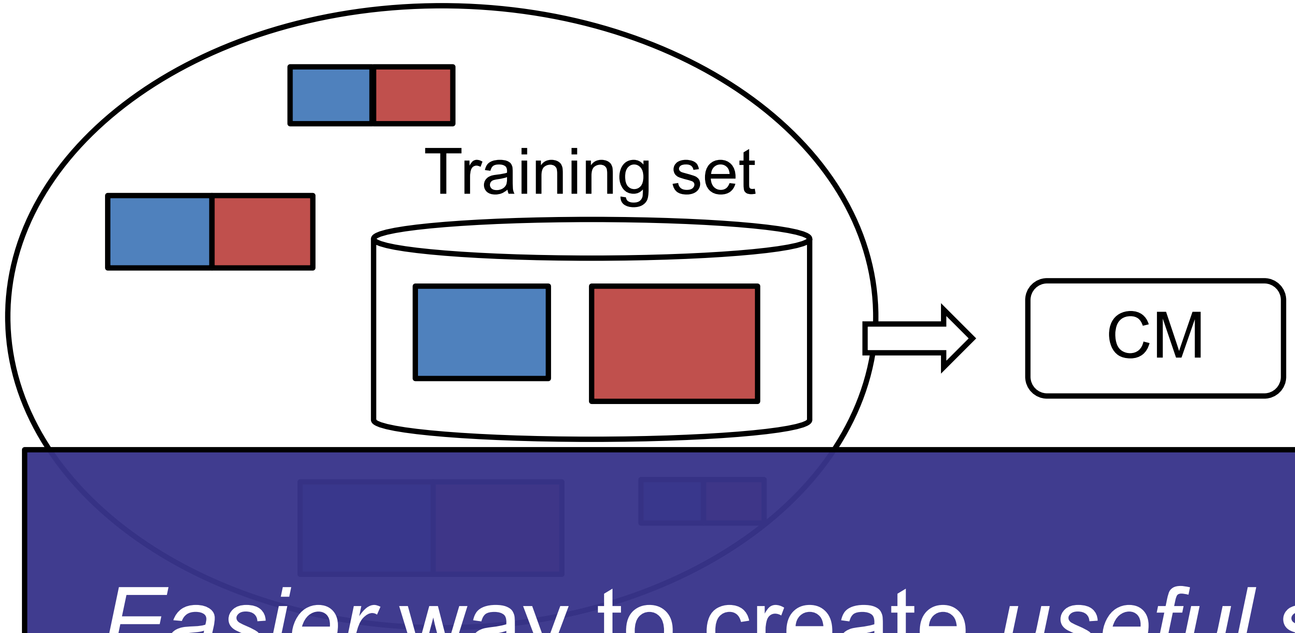
- Poor generalization (Das 2020)
  - Overfit to non-speech (Müller 2021, Liu 2022)
  - Artefacts in high-frequency band (Wang 2022)
  - …

Das, R. K., Yang, J. & Li, H. Assessing the scope of generalized countermeasures for anti-spoofing. in *Proc. ICASSP* 6589–6593 (2020). doi:10.1109/ICASSP40776.2020.9053086* From CQSPIC system in Table 1&2 of (Das 2020)
Nicolas Müller, Franziska Dieckmann, Pavel Czempin, Roman Canals, Konstantin Böttinger, and Jennifer Williams. Speech Is Silver, Silence Is Golden: What Do ASVspoof-Trained Models Really Learn? In *Proc. ASVspoof Challenge Workshop*, 55–60. doi:10.21437/ASVSPOOF.2021-9. 2021.
Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, and Kong Aik Lee. ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild. *ArXiv Preprint ArXiv:2210.02437*. 2022.
Xin Wang, and Junichi Yamagishi. Investigating Self-Supervised Front Ends for Speech Spoofing Countermeasures. In *Proc. Odyssey*, 100–106. doi:10.21437/Odyssey.2022-14. 2022.

Generalization may need more data



Generalization may need more data




ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech

Xin Wang[a,*], Junichi Yamagishi[a,b,**], Massimiliano Todisco[c,**], Héctor Delgado[c,**], Andreas Nautsch[c,**], Nicholas Evans[c,**], Md Sahidullah[d,**], Ville Vestman[e,**], Tomi Kinnunen[e,**], Kong Aik Lee[f,**], Lauri Juvela[g], Paavo Alku[g], Yu-Huai Peng[h], Hsin-Te Hwang[h], Yu Tsao[h], Hsin-Min Wang[h], Sébastien Le Maguer[i], Markus Becker[j], Fergus Henderson[j], Rob Clark[j], Yu Zhang[j], Quan Wang[j], Ye Jia[j], Kai Onuma[k], Koji Mushika[k], Takashi Kaneda[k], Yuan Jiang[l], Li-Juan Liu[l], Yi-Chiao Wu[m], Wen-Chin Huang[m], Tomoki Toda[m], Kou Tanaka[n], Hirokazu Kameoka[n], Ingmar Steiner[o], Driss Matrouf[p], Jean-François Bonastre[p], Avashna Govender[b], Srikanth Ronanki[q], Jing-Xuan Zhang[r], Zhen-Hua Ling[r]

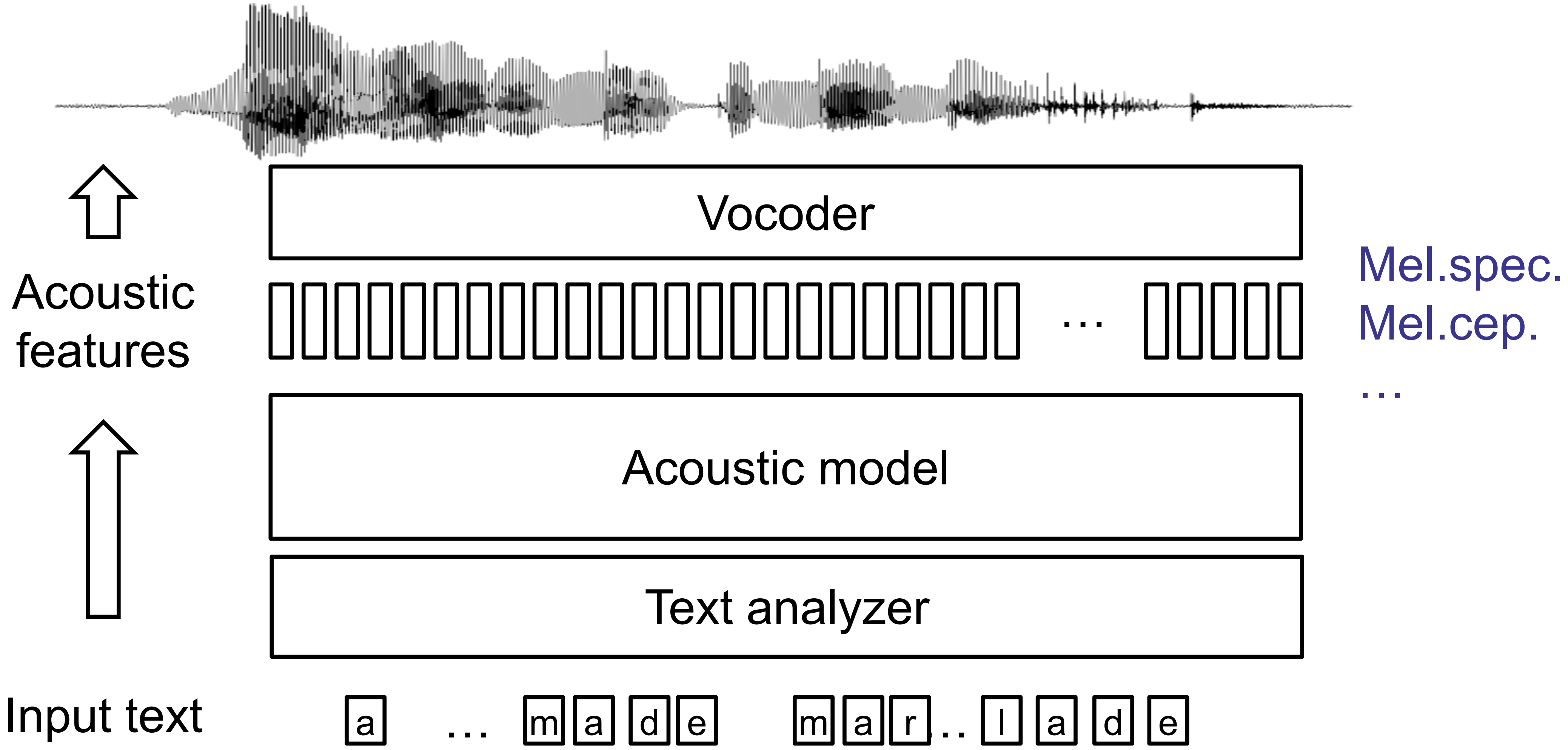
[a]National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan
[b]Centre for Speech Technology Research, University of Edinburgh, UK
[c]EURECOM, Campus SophiaTech, 450 Route des Chappes, 06410 Biot, France
[d]Université de Lorraine, CNRS, Inria, LORIA, F-54000, Nancy, France
[e]University of Eastern Finland, Joensuu campus, Länsikatu 15, FI-80110 Joensuu, Finland
[f]NEC Corp., 7-1, Shiba 5-chome Minato-ku, Tokyo 108-8001, Japan
[g]Aalto University, Rakentajanaukio 2 C, 00076 Aalto, Finland
[h]Academia Sinica, 128, Sec. 2, Academia Road, Nankang, Taipei, Taiwan

~6 months of work

Easier way to create *useful* spoofed training data?

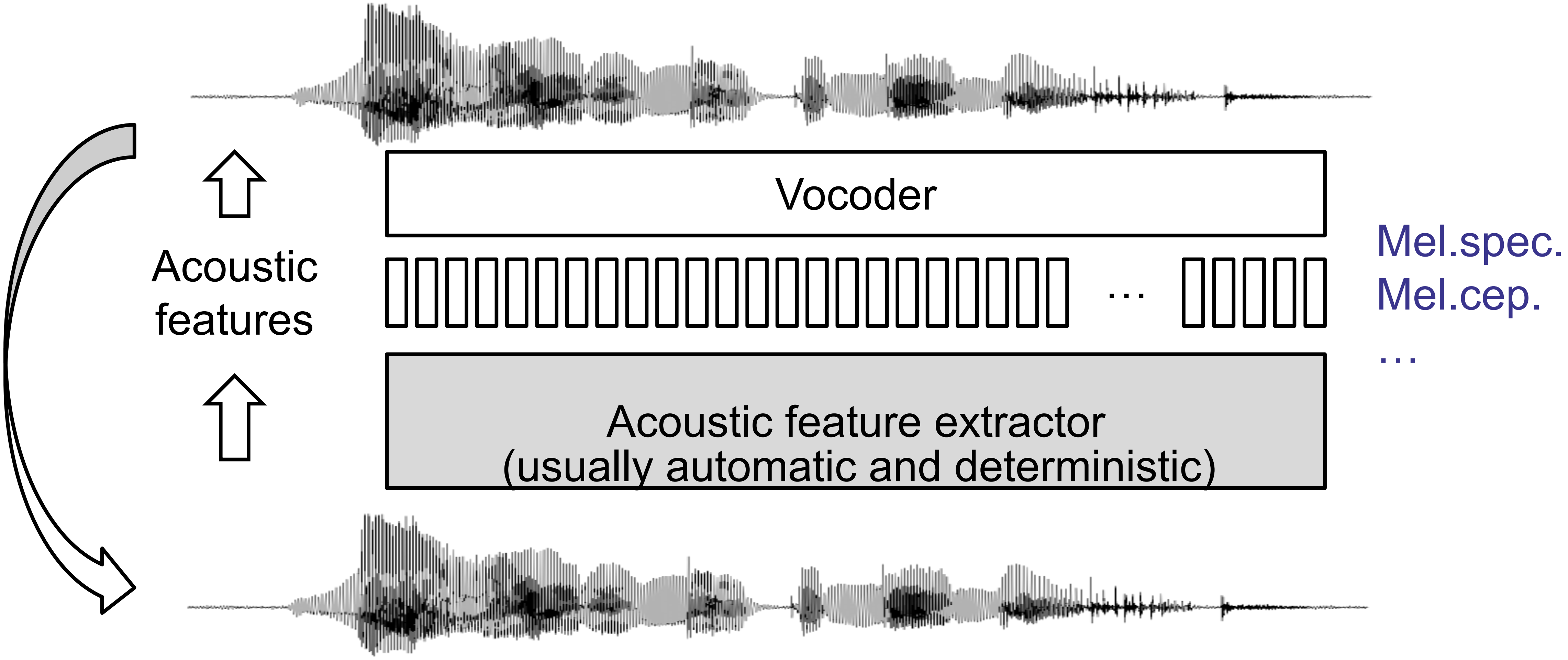
- However, building diverse TTS and VC systems is **time consuming**

Idea: creating spoofed training data by vocoding



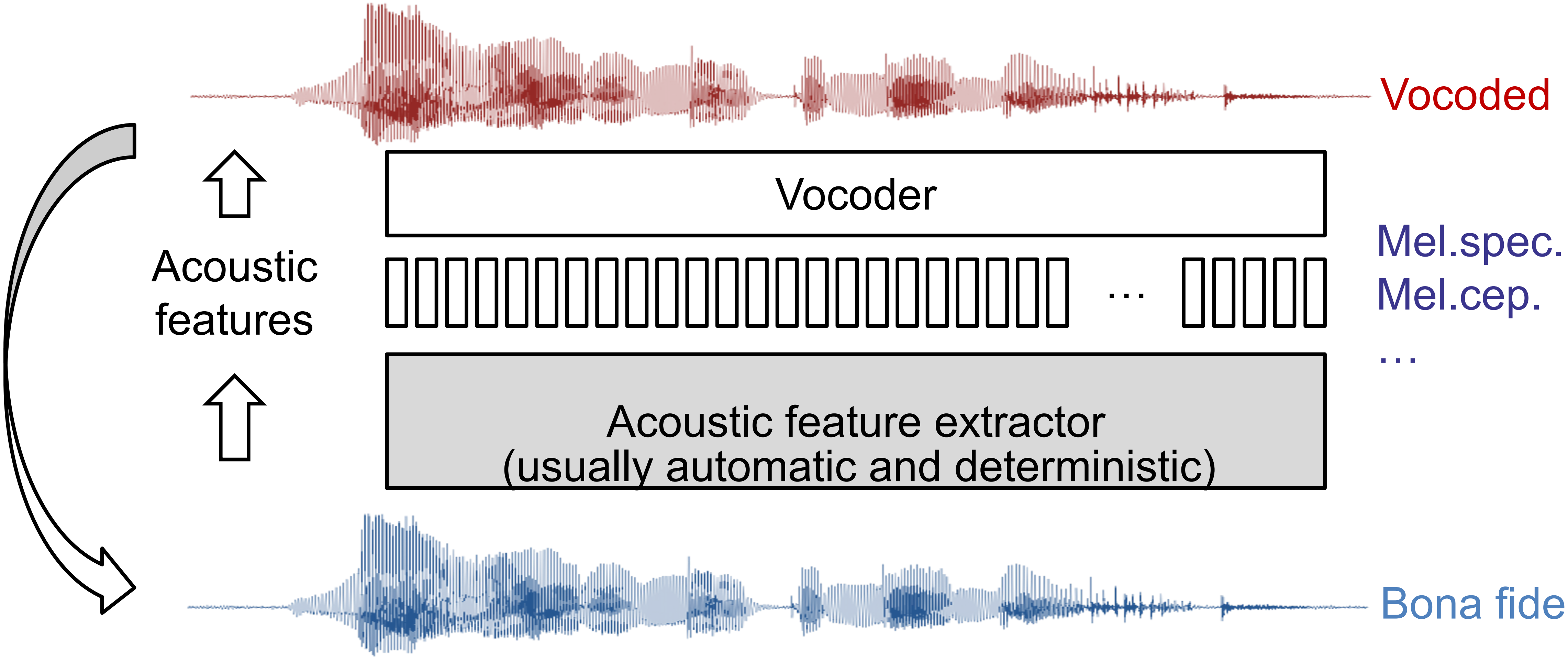
Example of full-fledged TTS

Idea: creating spoofed training data by vocoding



Vocoding is TTS using a perfect acoustic model

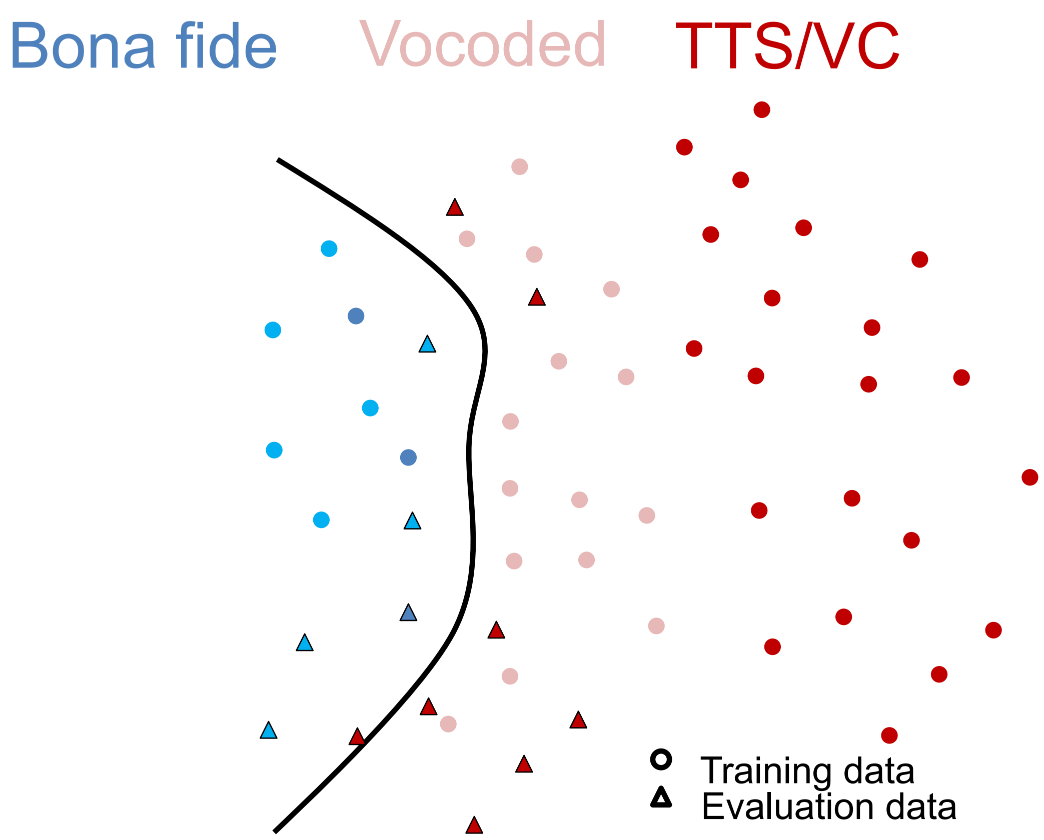
Idea: creating spoofed training data by vocoding

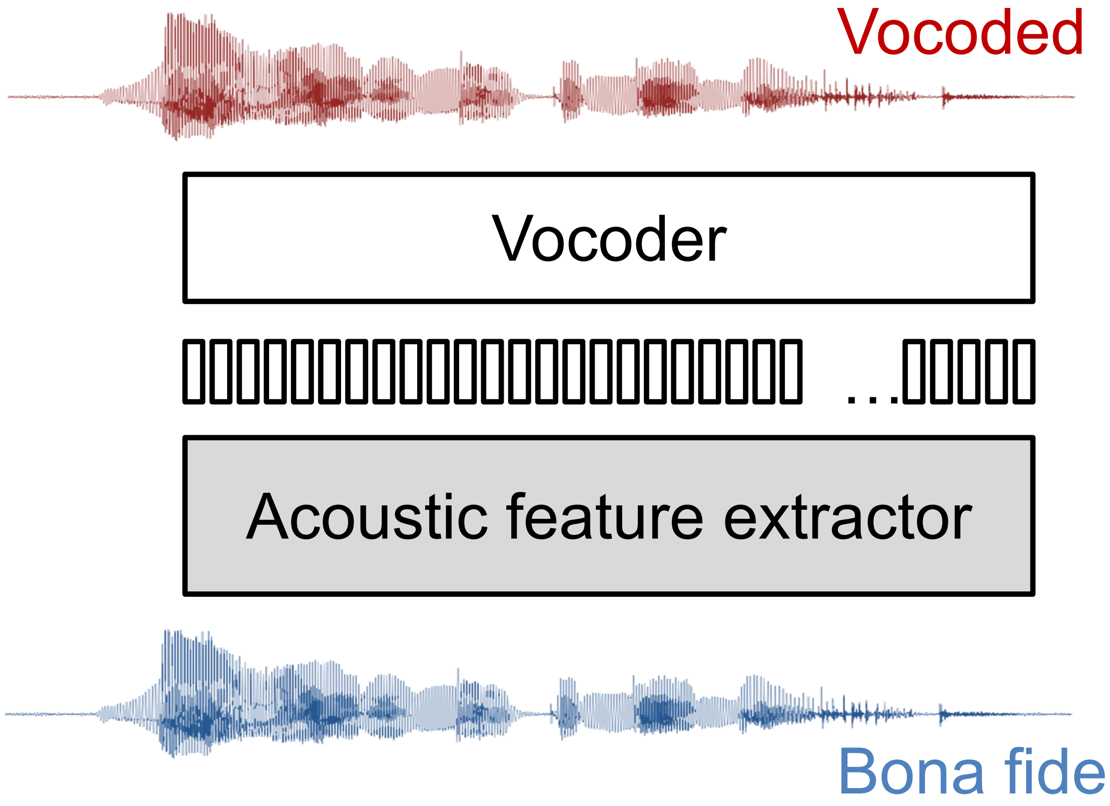


Vocoding is TTS using a perfect acoustic model

Idea: creating spoofed training data by vocoding

❑ Assumption (ideally)





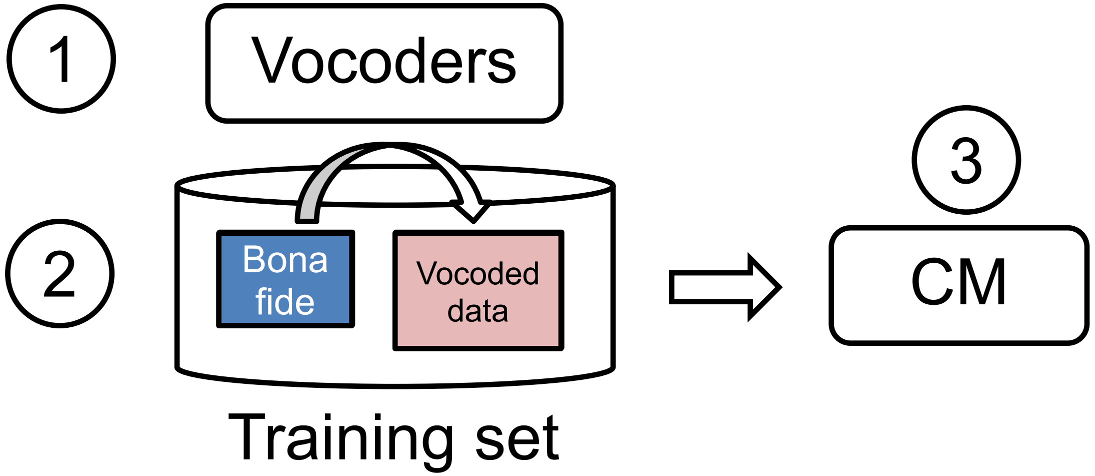
Vocoding is TTS using a perfect acoustic model

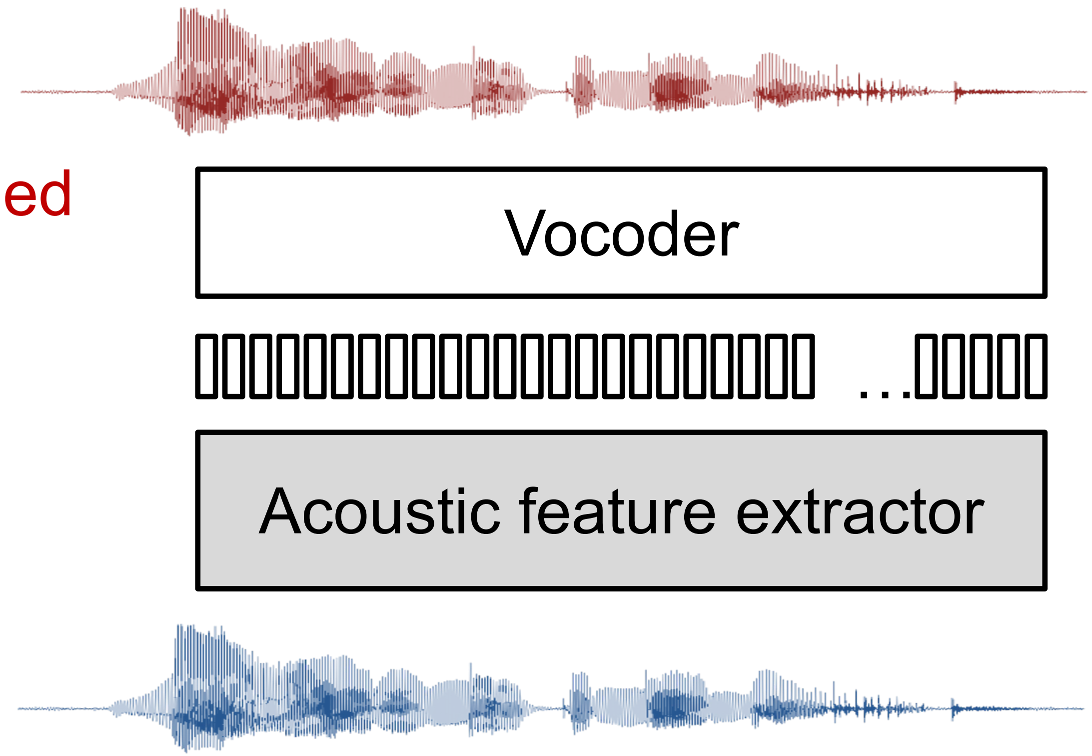
See actual embedding space in appendix

Idea: creating spoofed training data by vocoding

❑ Procedure

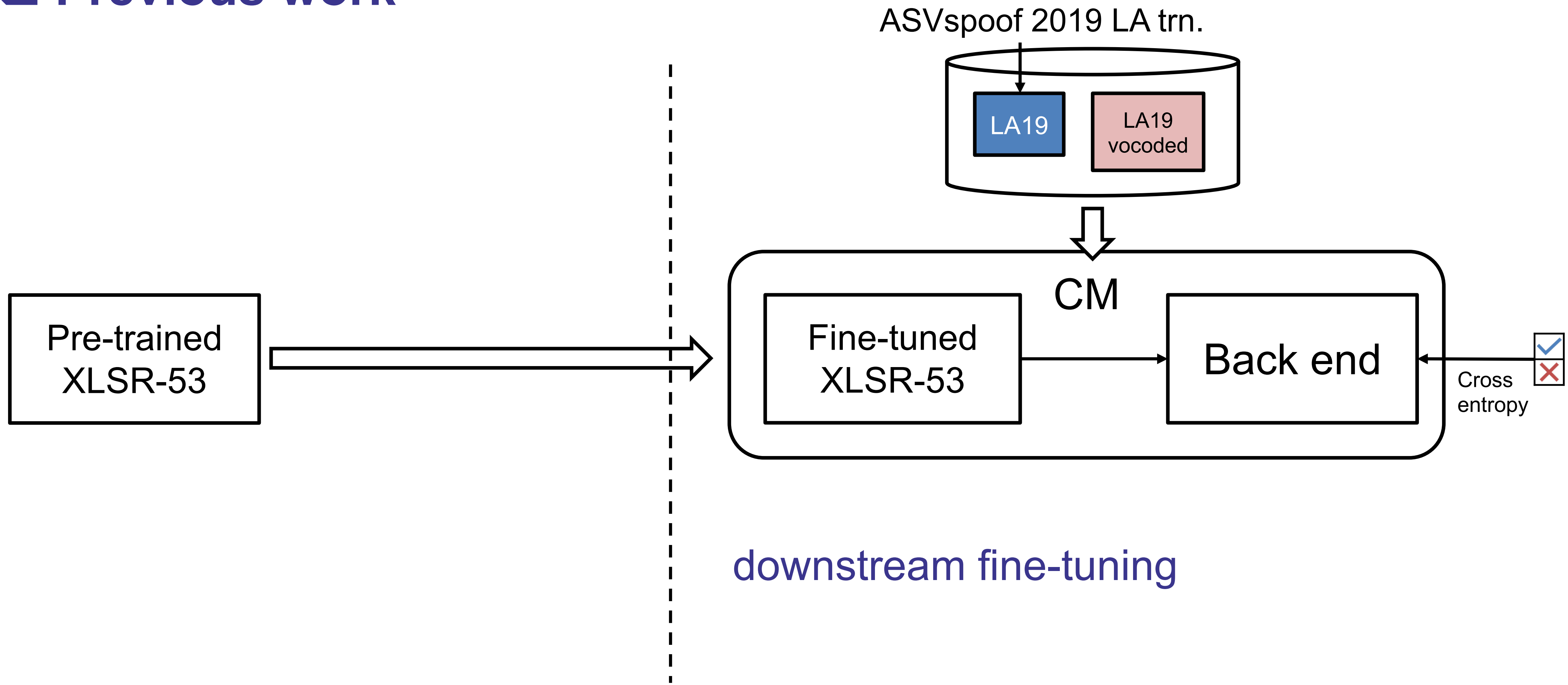
1. Prepare (or download) vocoder***s***
2. Vocode the bona fide data
3. Train CM using bona fide and vocoded data





Method: CM training using bona fide & vocoded data

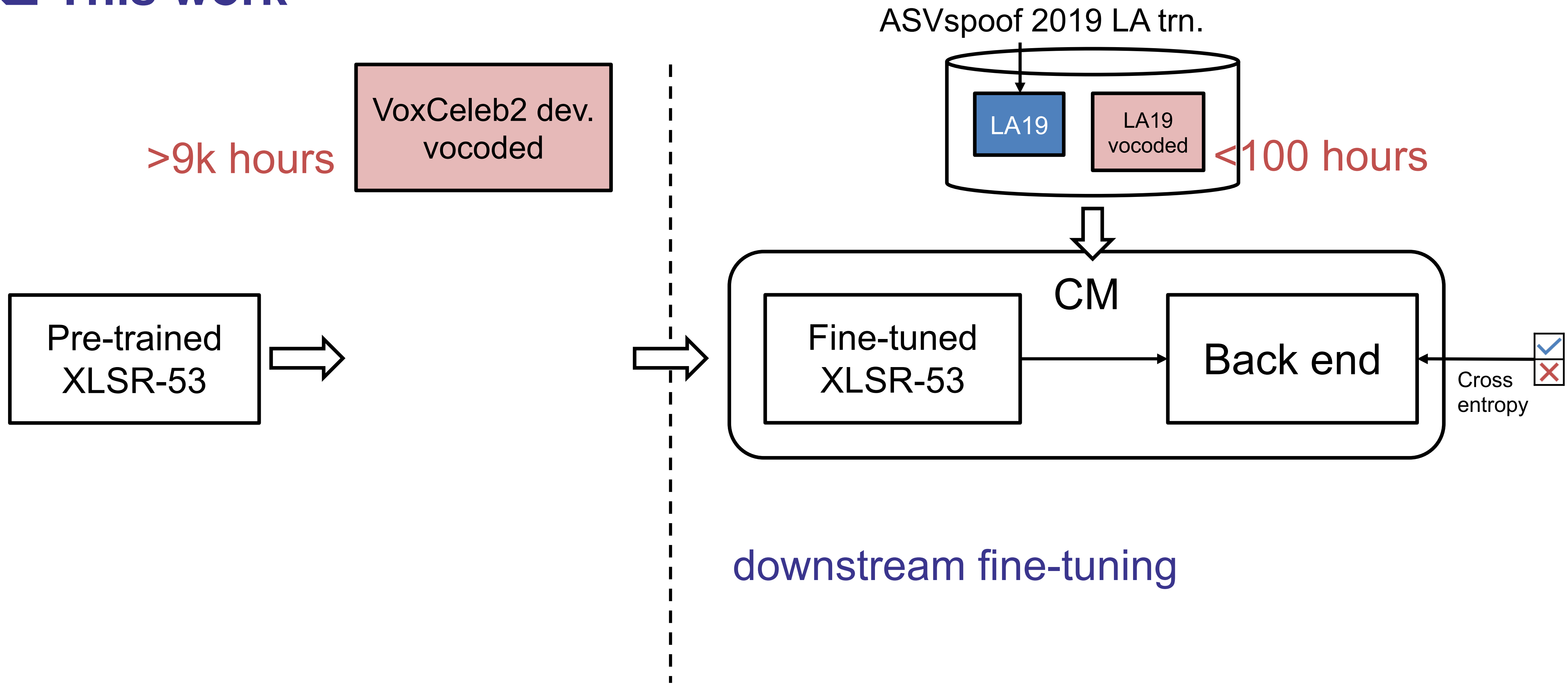
- ❑ Previous work (Wang 2023)



Contrastive feature loss is not plotted.
See slide appendix for more.

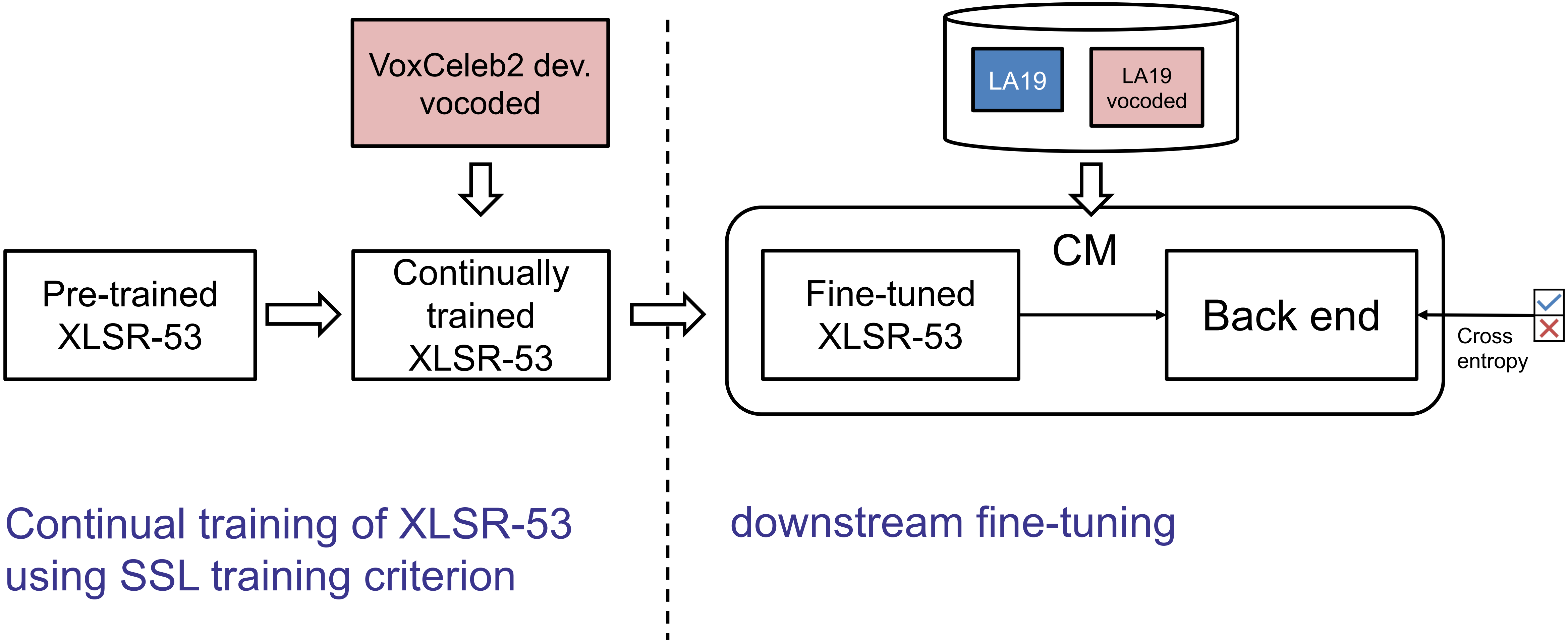
Method: CM training using bona fide & vocoded data

- This work



Method: CM training using bona fide & vocoded data

❑ This work – method 1



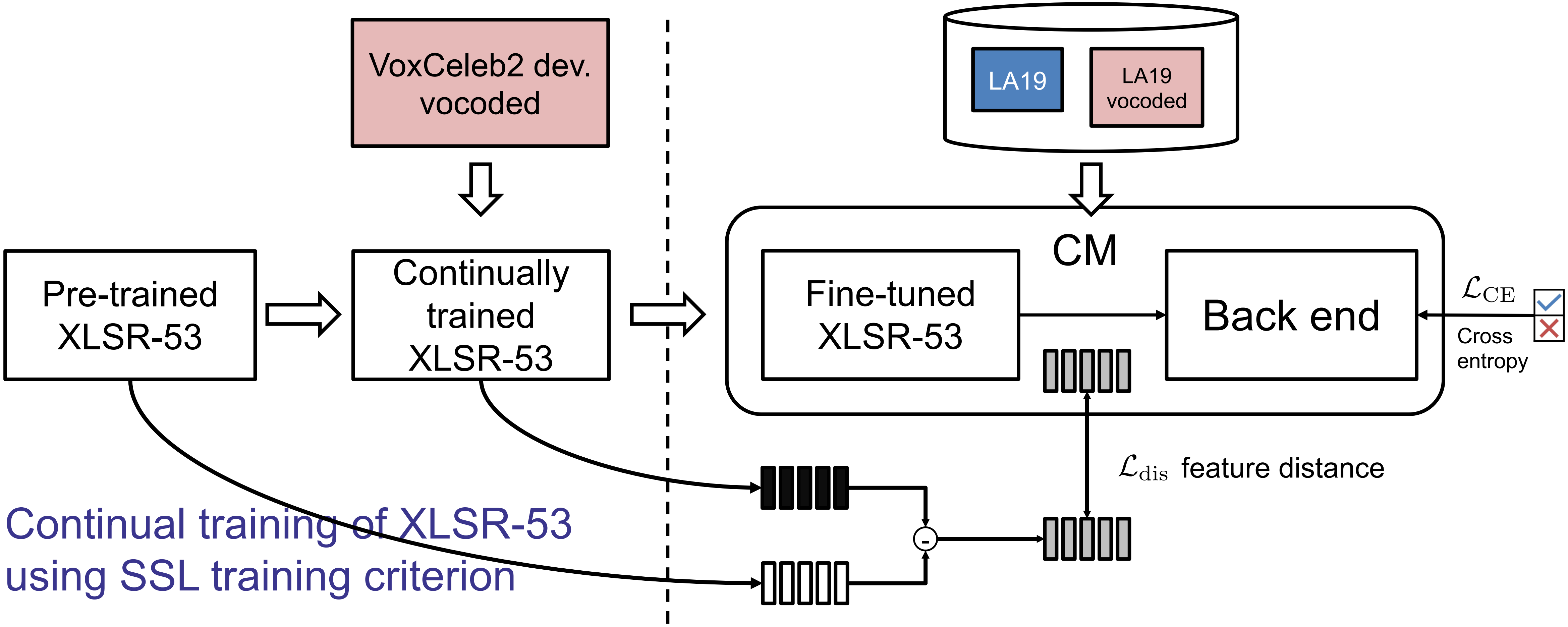
Continual training of XLSR-53 using SSL training criterion

downstream fine-tuning

Contrastive feature loss is not plotted.
See slide appendix for more.

Method: CM training using bona fide & vocoded data

❑ This work – method 2



Continual training of XLSR-53 using SSL training criterion

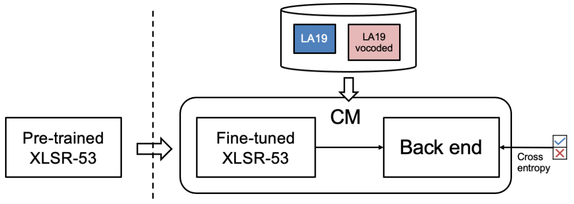
downstream fine-tuning + model distillation

Contrastive feature loss is not plotted.
See slide appendix for more.

Method: CM training using bona fide & vocoded data

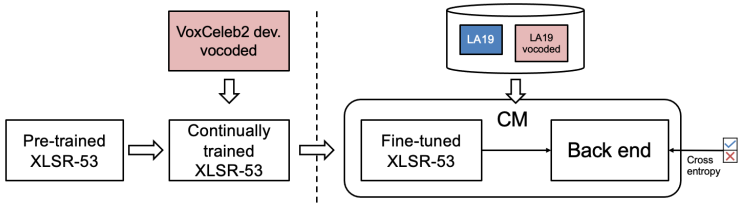
❑ Previous work (Wang 2023)

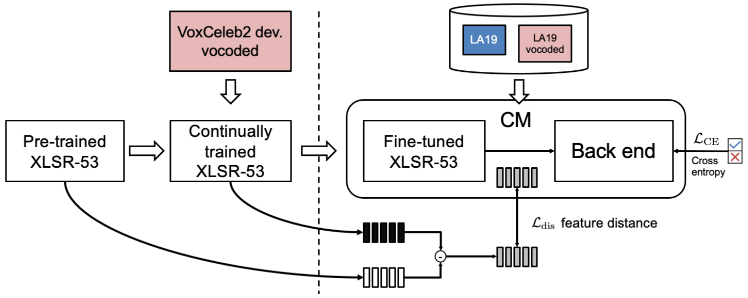
- Vocoded ASVspoof 2019 LA trn.
- downstream fine-tuning of SSL model



❑ This work

- Vocoded VoxCeleb2 dev.
- Upstream training of SSL
- Downstream training + dis***

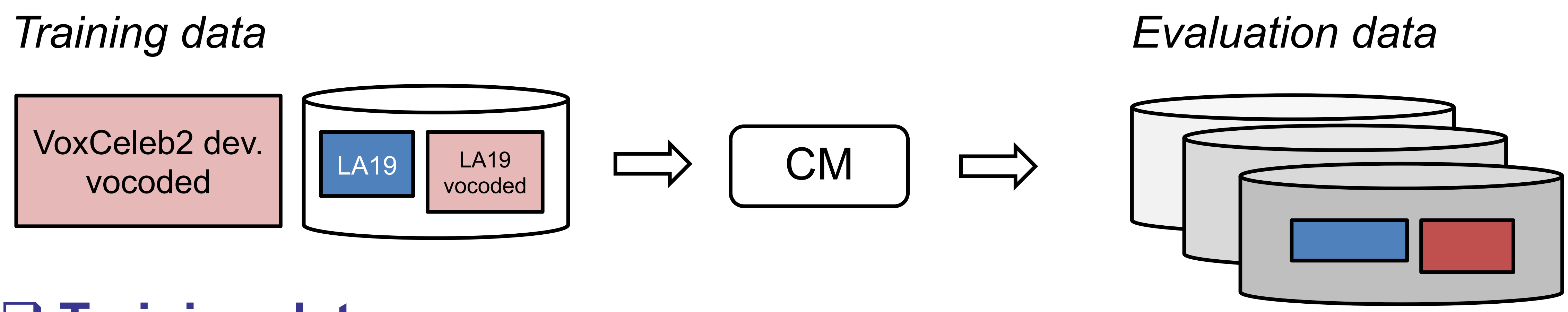




❑ Related studies using DSP-based vocoders (Wu 2013, Khoury 2014, Sizov 2015, Saratxaga 2016, Pal 2018)

See related papers in slide appendix

Experiment



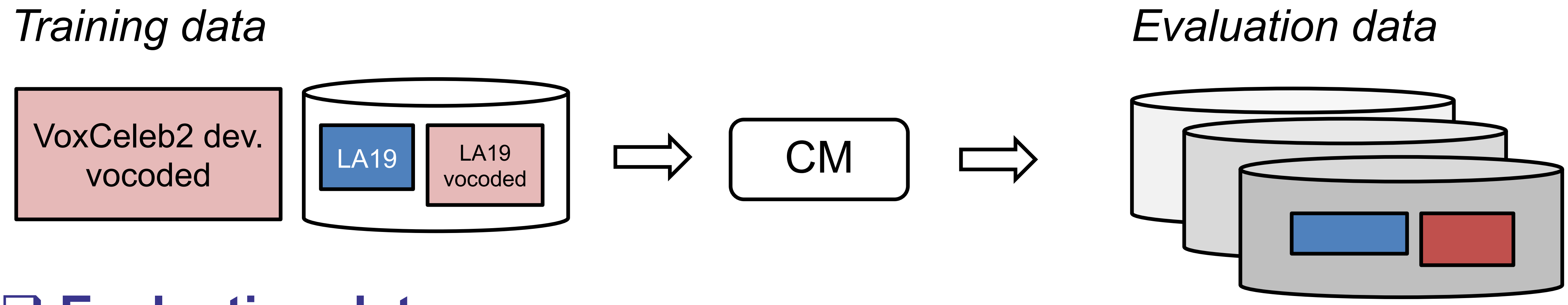
❑ Training data

- SSL upstream training: vocoded VoxCeleb2 dev.
- Downstream fine-tuning: bonafide + vocoded ASVspoof 2019 LA trn.

- Vocoders: Hifi-GAN (Kong 2020), NSF (Wang 2019), NSF-GAN, WaveGlow (Prenger 2019)

Kong, Jungil, Jaehyeon Kim, and Jaekyoung Bae. "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis." In *Proc. NIPS*, 33:17022–33, 2020.
Wang, Xin, Shinji Takaki, and Junichi Yamagishi. "Neural Source-Filter-Based Waveform Model for Statistical Parametric Speech Synthesis." In *Proc. ICASSP*, 5916–20. IEEE, 2019.
Prenger, Ryan, Rafael Valle, and Bryan Catanzaro. "WaveGlow: A Flow-Based Generative Network for Speech Synthesis." In *Proc. ICASSP*, 3617–21, 2019.

Experiment



❑ Evaluation data

- ASVspoof 2019 LA test set, 2021 LA & DF eval sets
- ASVspoof 2019 LA test set w/o non-speech, 2021 LA & DF hidden track
- WaveFake (Frank 2021), In-the-Wild (Müller 2022)

More challenging due to domain mismatch

- three independent training-evaluation rounds
- averaged EERs

Joel Frank, and Lea Schönherr. WaveFake: A Data Set to Facilitate Audio DeepFake Detection. In *Proc. NeurIPS Datasets and Benchmarks 2021*. 2021.
Nicolas M Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger. Does Audio Deepfake Detection Generalize? In *Proc. Interspeech*, 2783–2787. 2022.

Experiment results

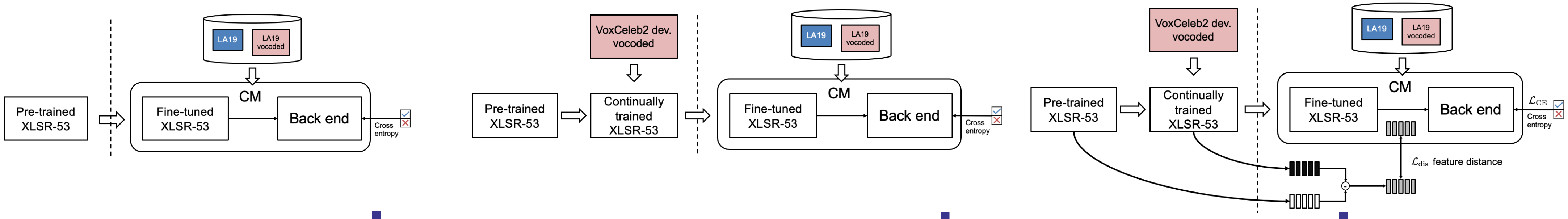
☺ Low EER ☹ High EER

Systems using different training configurations

	ID	B1	B2	B3	P1	P2	P3
CM	Front end SSL(s)	xlsr	xlsr, w2v		v.vox	xlsr, v.vox	xlsr, v.vox
	SSL distilling	-	×	✓	-	×	✓
Data for fine-tune CM		voc.LA			voc.LA		
EER on each test set	LA19eval	3.45	1.97	1.26	2.09	2.01	1.91
	LA21eval	17.59	13.94	21.09	16.88	14.94	15.92
	DF21eval	6.53	4.04	14.72	4.34	5.28	5.67
	LA19etrim	2.69	2.80	3.74	3.33	2.79	3.28
	LA21hid	13.93	14.05	20.03	16.02	13.95	14.97
	DF21hid	8.89	9.10	15.27	7.71	8.40	8.84
	WaveFake	7.33	1.48	5.88	1.94	0.89	1.30
	InWild	6.78	4.25	13.20	5.84	4.07	6.10
Pooled EER single threshold	Pooled	11.13	12.95	14.06	10.54	9.07	9.98

Note: row B3 SSL is "xlsr, w2v"; B2 SSL is "xlsr, w2v".

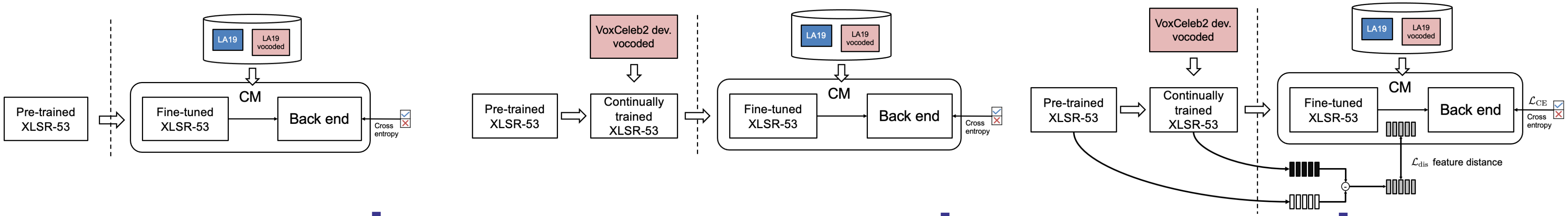
Experiment results



Test sets ID	B1	B2	B3	P1	P2	P3
LA19eval	3.45	1.97	1.26	2.09	2.01	1.91
LA21eval	17.59	13.94	21.09	16.88	14.94	15.92
DF21eval	6.53	4.04	14.72	4.34	5.28	5.67
LA19etrim	2.69	2.80	3.74	3.33	2.79	3.28
LA21hid	13.93	[illegible]	20.03	16.02	13.95	14.97
DF21hid	8.89	9.10	15.27	7.71	8.40	8.84
WaveFake	7.33	1.48	5.88	1.94	0.89	1.30
InWild	6.78	4.25	13.20	5.84	4.07	6.10
Pooled	11.13	[illegible]	[illegible]	10.54	9.07	9.98

B1 vs P1: continually trained SSL is *not* useless

Experiment results

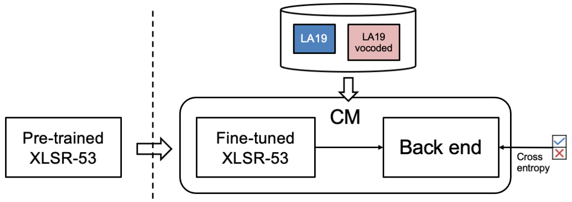


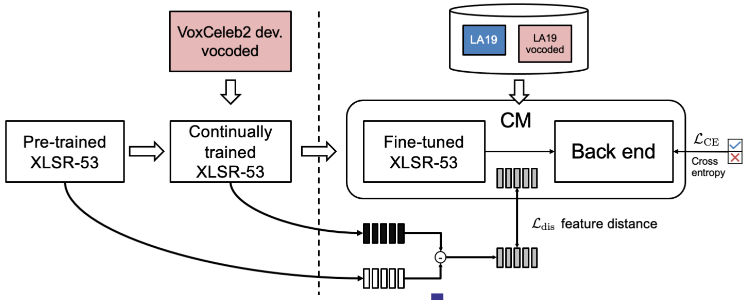
Test sets	ID	B1	B2	B3	P1	P2	P3
	LA19eval	3.45	1.97	1.26	2.09	2.01	1.91
	LA21eval	17.59	13.94	21.09	16.88	14.94	15.92
	DF21eval	6.53	4.04	14.72	4.34	5.28	5.67
	LA19etrim	2.69	2.80	3.74	3.33	[illegible]	3.28
	LA21hid	13.93	[illegible]	20.03	16.02	13.95	14.97
	DF21hid	8.89	[illegible]	[illegible]	7.71	8.40	8.84
	WaveFake	7.33	1.48	5.88	1.94	0.89	1.30
	InWild	6.78	4.25	13.20	5.84	4.07	6.10
	Pooled	11.13	[illegible]	[illegible]	10.54	[illegible]	9.98

B1 vs P1: continually trained SSL is *not* useless

Merging two SSLs is helpful

Experiment results





	ID	B1	P3
Test sets	LA19eval	3.45	1.91
	LA21eval	17.59	15.92
	DF21eval	6.53	5.67
	LA19etrim	2.69	3.28
	LA21hid	13.93	14.97
	DF21hid	8.89	8.84
	WaveFake	7.33	1.30
	InWild	6.78	6.10
	Pooled	11.13	9.98

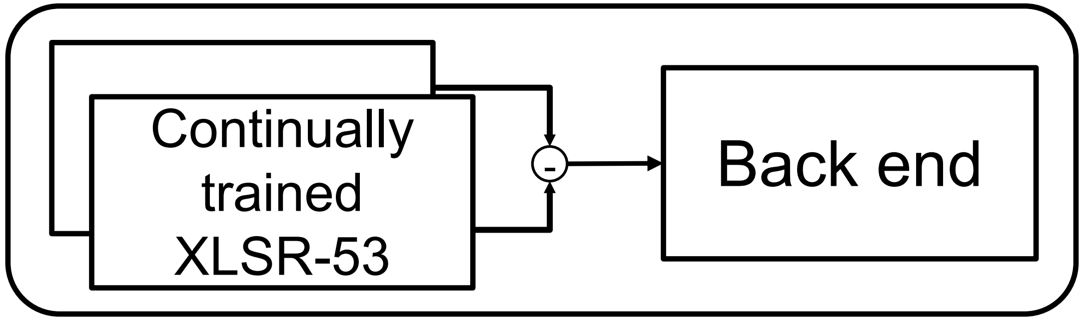
B1 is the best model in our previous work (Wang 2023)

It is better than using TTS/VC spoofed data

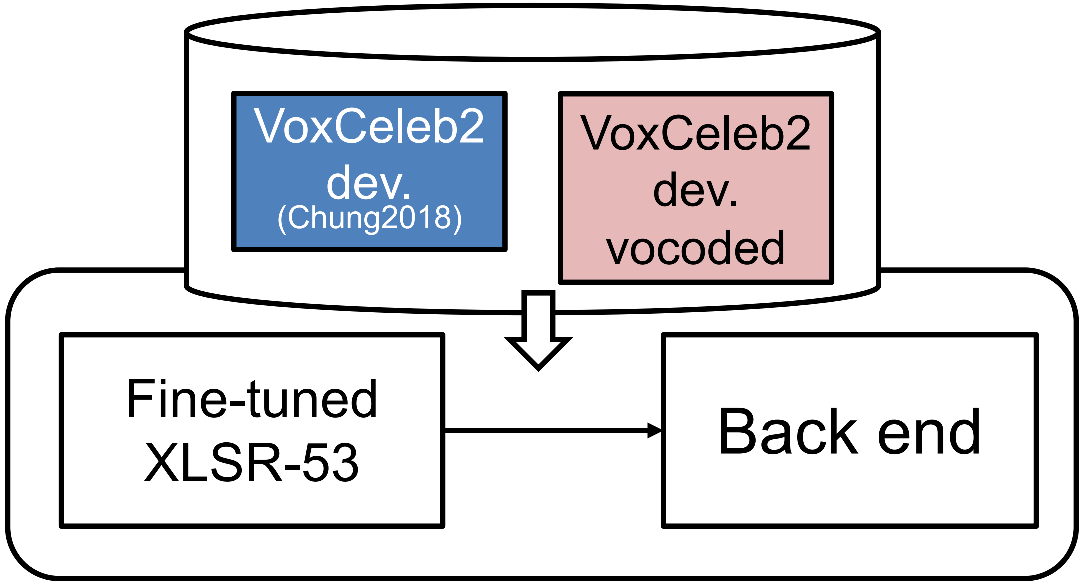
Experiment results

❑ Other results in the paper

- Using two SSLs without distillization?



- Downstream fine-tuning using vocoded voxceleb2?

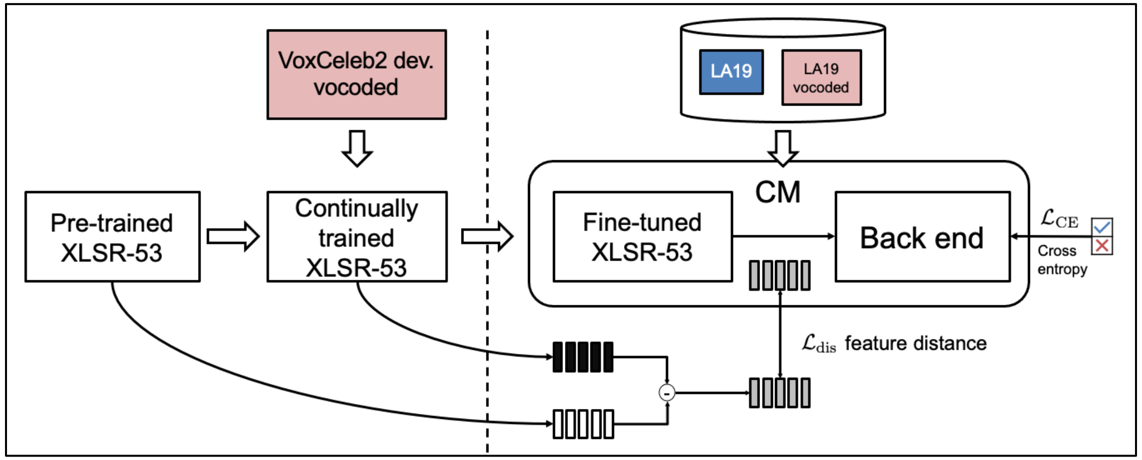


- …

Summary

❑ Method

- Large scale vocoded VoxCeleb2
- Upstream SSL training
- Downstream fine-tuning + distilling



❑ Results

- Slightly outperformed previous work (pooled EER)

- Limitation: only 4 types of vocoders

Thank you

project/10-asvspoof-vocoded-trn-ssl

Appendix

CM	ID	B1	B2	B3	P1	P2	P3	B1-b	P3-b	B1-c	P3-c
	Front end SSL(s)	xlsr	xlsr, w2v	xlsr, w2v	v.vox	xlsr, v.vox	xlsr, v.vox	xlsr	xlsr, v.vox	xlsr	xlsr, v.vox
	SSL distilling	-	×	✓	-	×	✓	-	✓	-	✓
Data for fine-tune CM		voc.LA			voc.LA			LA19trn		voc.VoxCel	
Test sets	LA19eval	3.45	1.97	1.26	2.09	2.01	1.91	0.22	0.13	3.59	3.71
	LA21eval	17.59	13.94	21.09	16.88	14.94	15.92	2.69	3.29	15.22	12.37
	DF21eval	6.53	4.04	14.72	4.34	5.28	5.67	4.27	3.45	5.99	3.31
	LA19etrim	2.69	2.80	3.74	3.33	2.79	3.28	7.37	7.37	2.74	3.63
	LA21hid	13.93	14.05	20.03	16.02	13.95	14.97	15.56	24.23	10.14	9.53
	DF21hid	8.89	9.10	15.27	7.71	8.40	8.84	9.16	13.95	9.03	7.77
	WaveFake	7.33	1.48	5.88	1.94	0.89	1.30	23.75	15.44	13.41	24.17
	InWild	6.78	4.25	13.20	5.84	4.07	6.10	13.52	12.32	6.90	7.00
	Pooled	11.13	12.95	14.06	10.54	9.07	9.98	12.76	12.50	10.92	12.26

Reference

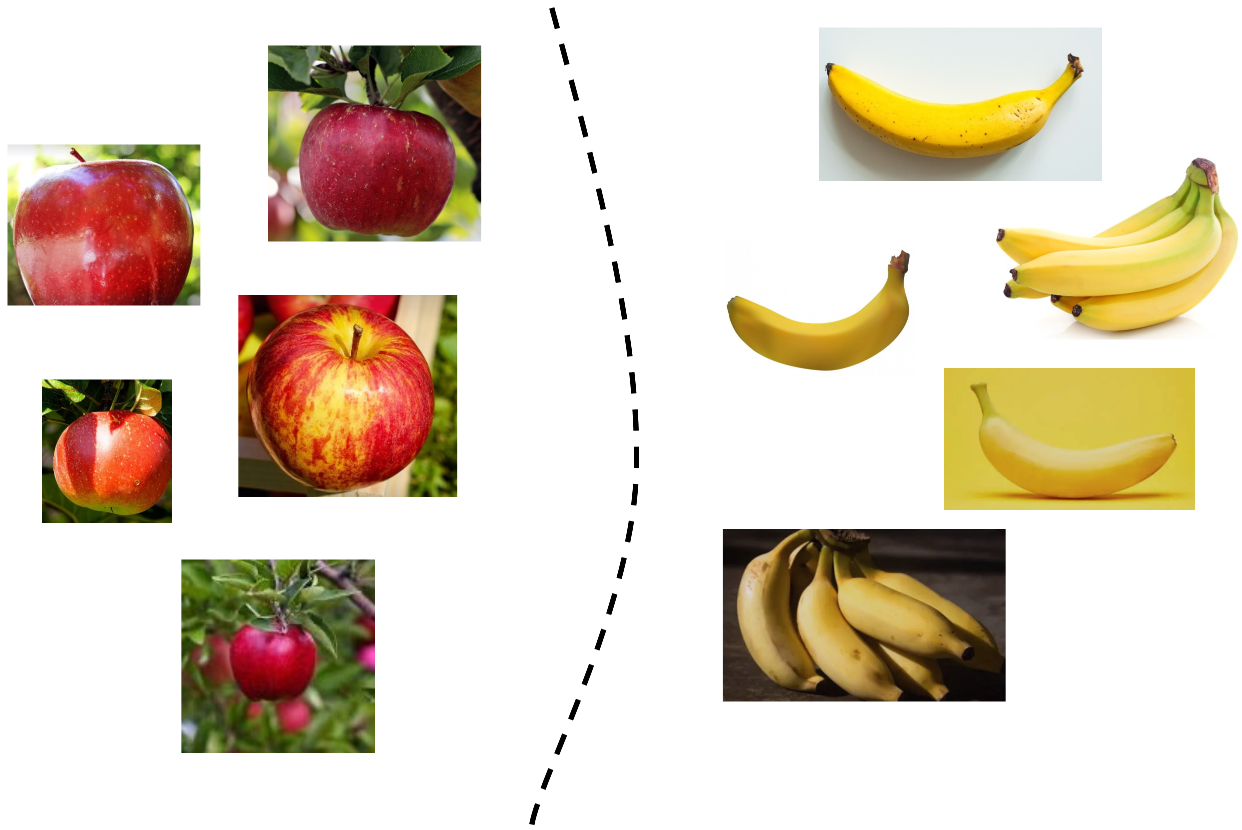
DSP-based vocoders

- Xingming Wang, Xiaoyi Qin, Tinglong Zhu, Chao Wang, Shilei Zhang, and Ming Li. The DKU-CMRI System for the ASVspoof 2021 Challenge: Vocoder Based Replay Channel Response Estimation. In *Proc. ASVspoof challenge workshop*, 16–21. 2021.
- Monisankha Pal, Dipjyoti Paul, and Goutam Saha. Synthetic Speech Detection Using Fundamental Frequency Variation and Spectral Features. *Computer Speech & Language* 48. Elsevier: 31–50. 2018.
- Ibon Saratxaga, Jon Sanchez, Zhizheng Wu, Inma Hernaez, and Eva Navas. Synthetic Speech Detection Using Phase Information. *Speech Communication* 81 (July): 30–41. doi:10.1016/j.specom.2016.04.001. 2016.
- Aleksandr Sizov, Elie Khoury, Tomi Kinnunen, Zhizheng Wu, and Sébastien Marcel. Joint Speaker Verification and Antispoofing in the I-Vector Space. *IEEE Transactions on Information Forensics and Security* 10 (4). IEEE: 821–832. doi:10.1109/TIFS.2015.2407362. 2015.
- Elie Khoury, Tomi Kinnunen, Aleksandr Sizov, Zhizheng Wu, and Sébastien Marcel. Introducing I-Vectors for Joint Anti-Spoofing and Speaker Verification. In *Proc. Interspeech*, 61–65. 2014.
- Jon Sanchez, Ibon Saratxaga, Inma Hernaez, Eva Navas, and Daniel Erro. A Cross-Vocoder Study of Speaker Independent Synthetic Speech Detection Using Phase Information. In *Proc. Interspeech*. 2014.
- Zhizheng Wu, Xiong Xiao, Eng Siong Chng, and Haizhou Li. Synthetic Speech Detection Using Temporal Modulation Feature. In *Proc. ICASSP*, 7234–7238. 2013.

neural vocoders

- Joel Frank, and Lea Schönherr. WaveFake: A Data Set to Facilitate Audio DeepFake Detection. In Proc. NeurIPS Datasets and Benchmarks 2021. 2021.
- Chengzhe Sun, Shan Jia, Shuwei Hou, Ehab AlBadawy, and Siwei Lyu. Exposing AI-Synthesized Human Voices Using Neural Vocoder Artifacts. ArXiv Preprint ArXiv:2302.09198. 2023.

Introduction – one challenge in my opinion



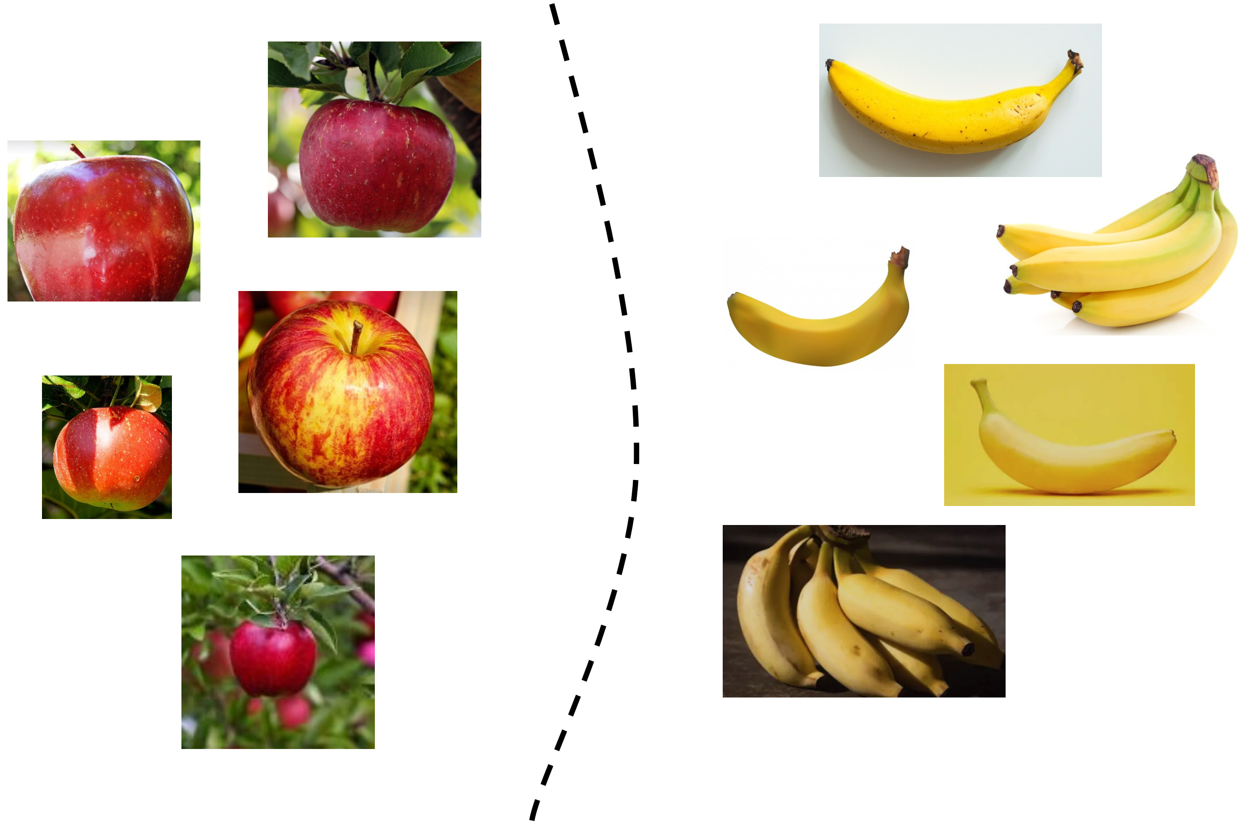
Decision boundary

Color?
Shape?
Background?

Introduction – one challenge in my opinion

Spoofed

Bona fide

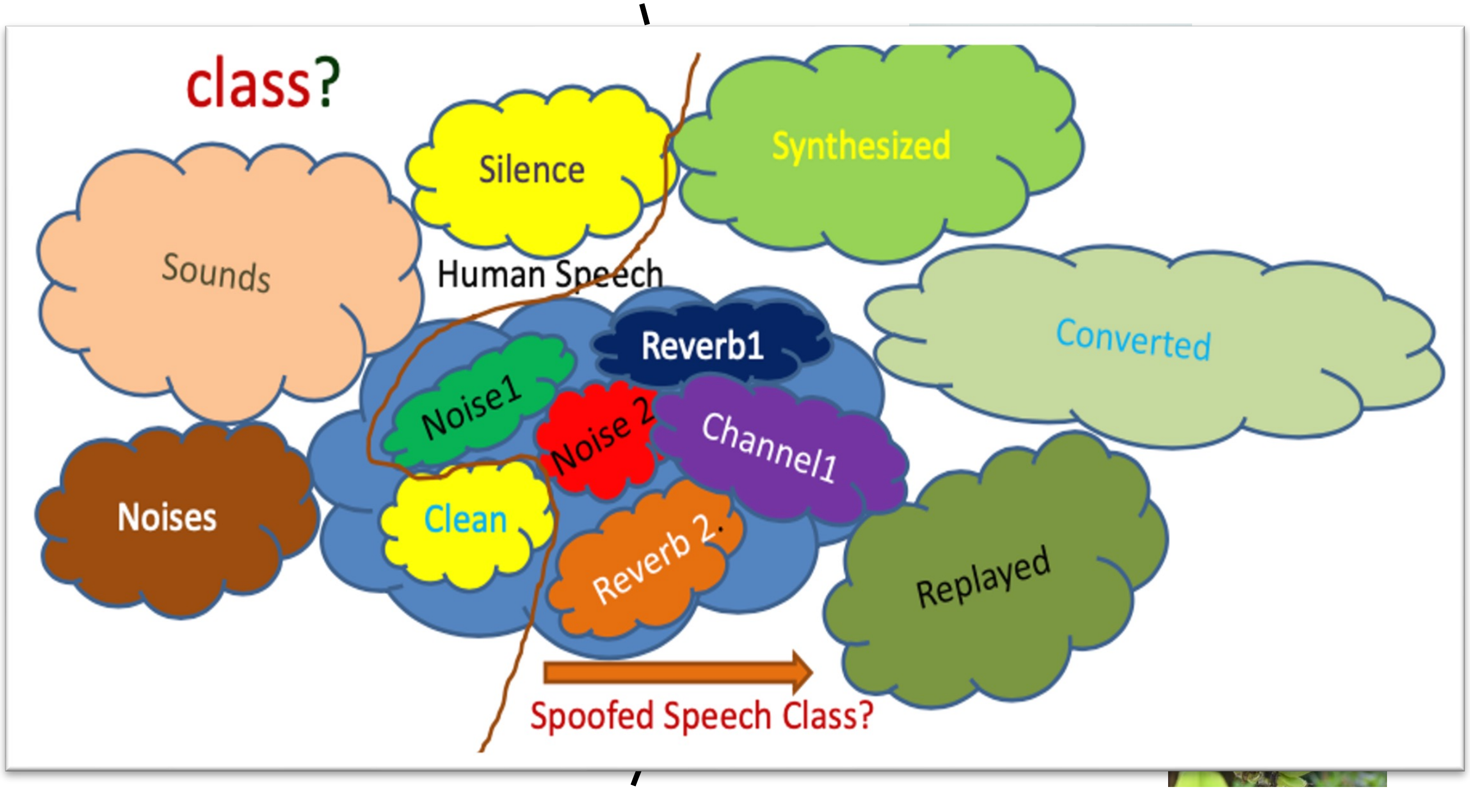


Decision boundary

Non-speech (Muller 2021) ?
Mp3 & additive noise (Shim 2023)?
Single source data domain?

Müller, Nicolas, Franziska Dieckmann, Pavel Czempin, Roman Canals, Konstantin Böttinger, and Jennifer Williams. 2021. “Speech Is Silver, Silence Is Golden: What Do ASVspoof-Trained Models Really Learn?” In *Proc. ASVspoof Challenge Workshop*, 55–60.
Shim, Hye-jin, Rosa Gonzalez Hautamäki, Md Sahidullah, and Tomi Kinnunen. 2023. “How to Construct Perfect and Worse-than-Coin-Flip Spoofing Countermeasures: A Word of Warning on Shortcut Learning.” In Proc. INTERSPEECH 2023, 785–89. https://doi.org/10.21437/Interspeech.2023-1901.

Introduction – one challenge in my opinion



Speech is more complicated

En, Fr, Ch, Jp, …

LJ-speech, Librispeech …

Wav, mp3, m4a …

New speech synthesis methods

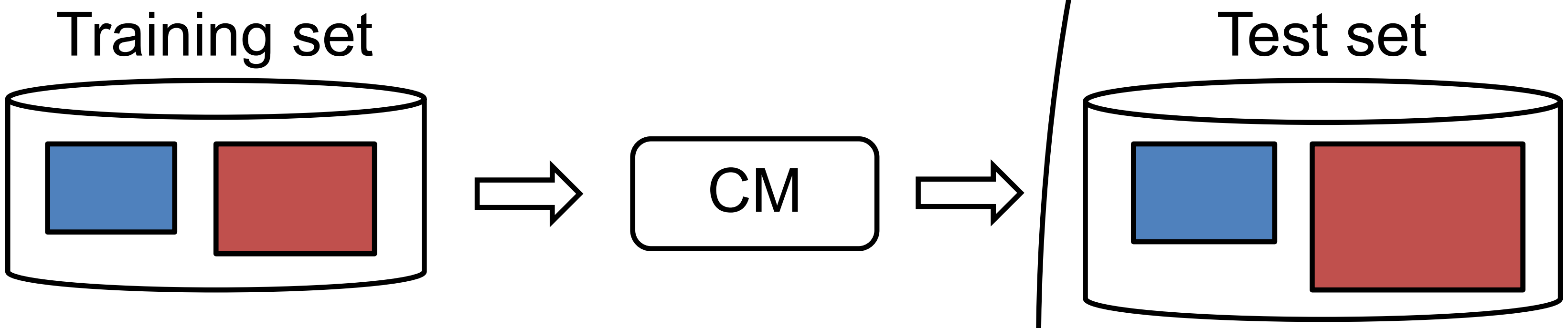
*For spoofed trials, **how is it possible to "well define" something which is unknown?***

From Jean-Francois Bonastre's talk

Mini-tutorial on TTS

Building diverse TTS and VC systems is not that easy

More spoofing data?



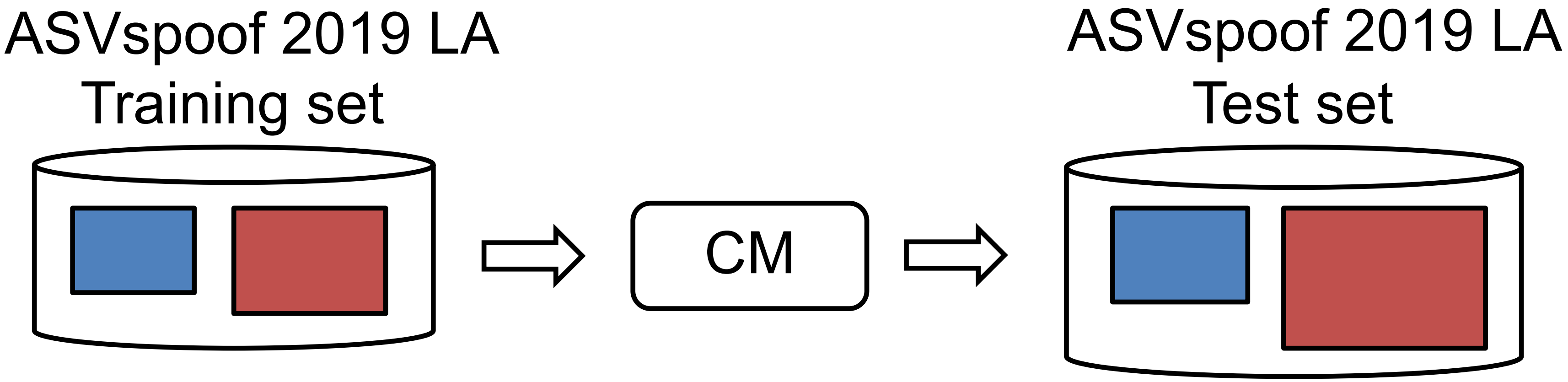
Space of all possible bona fide and spoofed data

En, Fr, Ch, Jp, …

wav, mp3, m4a …

New spoofing algorithms

More spoofing data?



6 TTS/VC

ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech

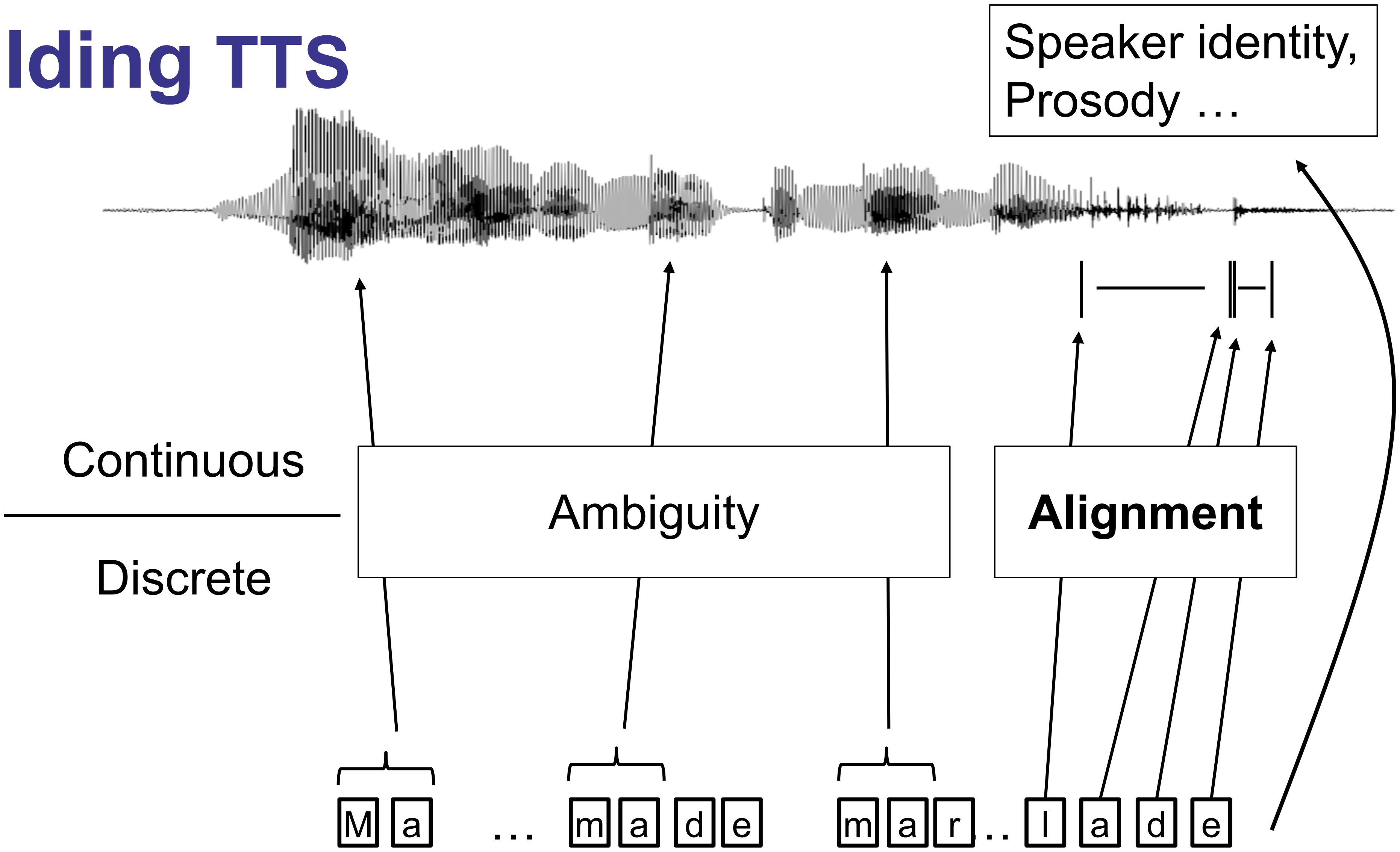
Xin Wang[a,*], Junichi Yamagishi[a,b,**], Massimiliano Todisco[c,**], Héctor Delgado[c,**], Andreas Nautsch[c,**], Nicholas Evans[c,**], Md Sahidullah[d,**], Ville Vestman[e,**], Tomi Kinnunen[e,**], Kong Aik Lee[f,**], Lauri Juvela[g], Paavo Alku[g], Yu-Huai Peng[h], Hsin-Te Hwang[h], Yu Tsao[h], Hsin-Min Wang[h], Sébastien Le Maguer[i], Markus Becker[j], Fergus Henderson[j], Rob Clark[j], Yu Zhang[j], Quan Wang[j], Ye Jia[j], Kai Onuma[k], Koji Mushika[k], Takashi Kaneda[k], Yuan Jiang[l], Li-Juan Liu[l], Yi-Chiao Wu[m], Wen-Chin Huang[m], Tomoki Toda[m], Kou Tanaka[n], Hirokazu Kameoka[n], Ingmar Steiner[o], Driss Matrouf[p], Jean-François Bonastre[p], Avashna Govender[b], Srikanth Ronanki[q], Jing-Xuan Zhang[r], Zhen-Hua Ling[r]

[a] *National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan*
[b] *Centre for Speech Technology Research, University of Edinburgh, UK*
[c] *EURECOM, Campus SophiaTech, 450 Route des Chappes, 06410 Biot, France*
[d] *Université de Lorraine, CNRS, Inria, LORIA, F-54000, Nancy, France*
[e] *University of Eastern Finland, Joensuu campus, Länsikatu 15, FI-80110 Joensuu, Finland*
[f] *NEC Corp., 7-1, Shiba 5-chome Minato-ku, Tokyo 108-8001, Japan*
[g] *Aalto University, Rakentajanaukio 2 C, 00076 Aalto, Finland*
[h] *Academia Sinica, 128, Sec. 2, Academia Road, Nankang, Taipei, Taiwan*
[i] *Sigmedia, ADAPT Centre, School of Engineering, Trinity College Dublin, Ireland*
[j] *Google Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA*
[k] *HOYA, Shinjuku Park Tower 35F, 3-7-1 Nishi-Shinjuku, Shinjuku-ku, Tokyo 163-1035 Japan*
[l] *iFlytek Research, High-tech Development Zone, No. 666 Wangjiang West Road, Hefei, 230088, China*
[m] *Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi 464-8601, Japan*
[n] *NTT Communication Science Laboratories, 3-1, Morinosato Wakamiya Atsugi-shi, Kanagawa, 243-0198 Japan*
[o] *audEERING GmbH, Friedrichshafener Str. 1 82205 Gilching, Germany*
[p] *Avignon University, LIA, 339 Chemin des Meinajariés, 84911 Avignon, France*
[q] *Centre for Speech Technology Research, University of Edinburgh, UK (Currently with Amazon)*
[r] *University of Science and Technology of China, No.96, JinZhai Road Baohe District,Hefei, Anhui, 230026, China*

11 + 2 TTS/VC

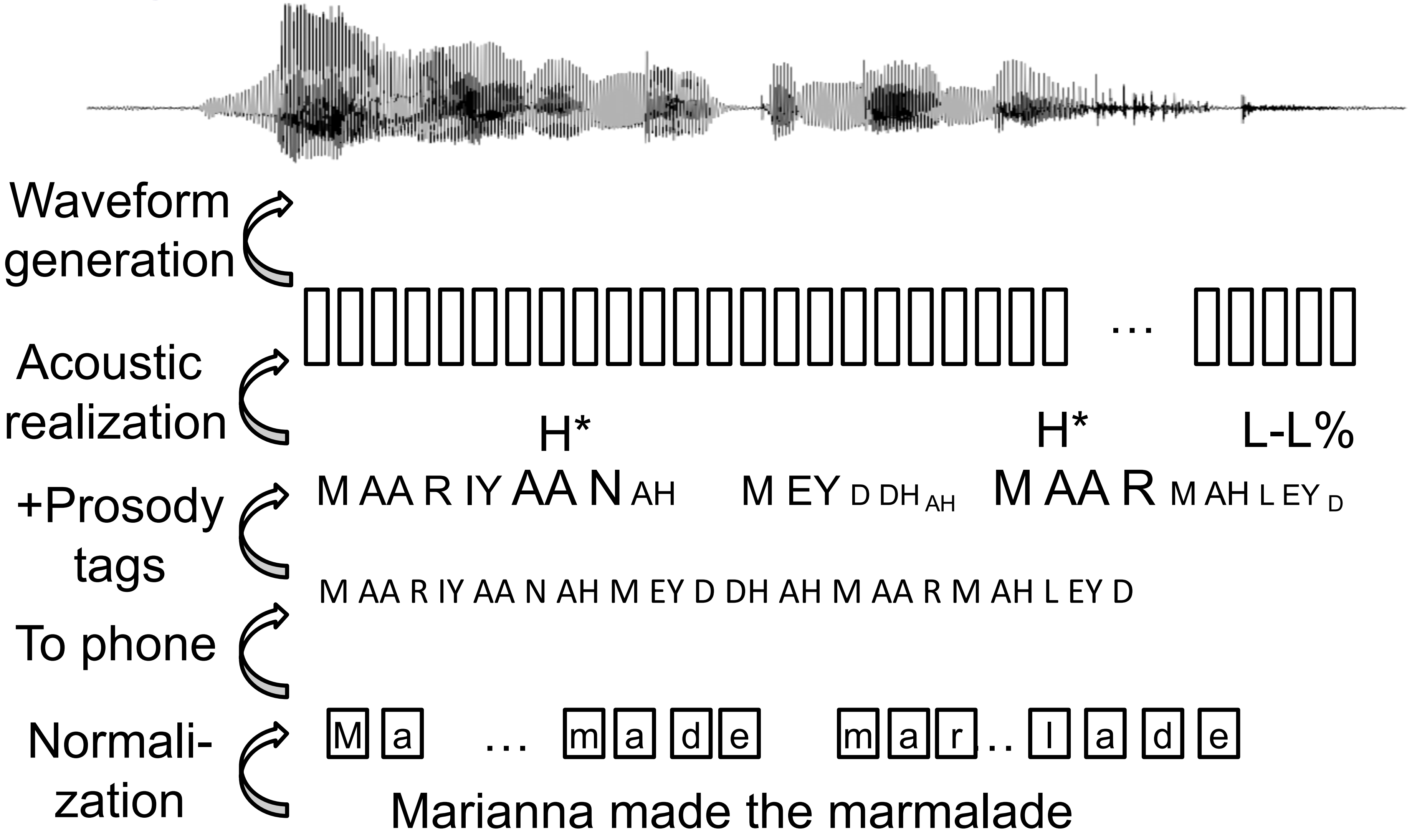
~6 months of work

Building TTS



Marianna made the marmalade

Building TTS



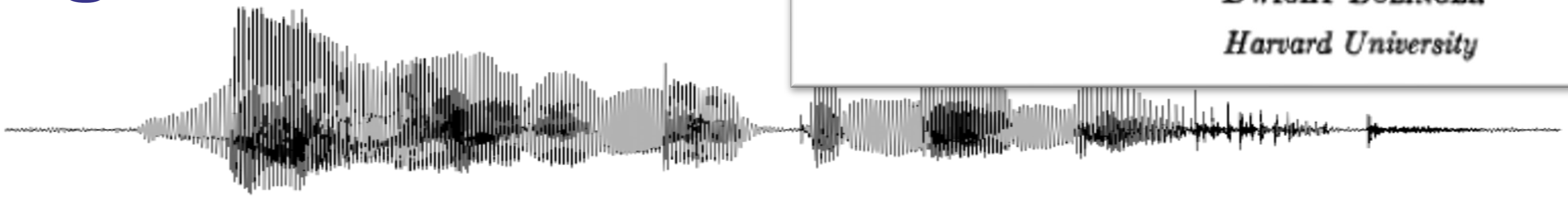
Sentence from: Beckman, M. E. & Ayers, G. Guidelines for ToBI labelling. OSU Res. Found. 3, (1997)
LOGIOS Lexicon tool: http://www.speech.cs.cmu.edu/tools/lextool.html
H*, L-L%: ToBI labels Beckman, M. E. & Ayers, G. Guidelines for ToBI labelling. OSU Res. Found. 3, (1997)

Building TTS

ACCENT IS PREDICTABLE (IF YOU'RE A MIND-READER)

Dwight Bolinger

Harvard University



Speaker A: Who made the marmalade.

Speaker B: Marianna made the marmalade.

Speaker A: Bob made the marmalade.

Speaker B: (No,) Marianna made the marmalade.

Speaker B: Marianna made the marmalade.

H* H* L-L%

+Prosody tags

M AA R IY AA N AH M EY D DH AH M AA R M AH L EY D

To phone

M AA R IY AA N AH M EY D DH AH M AA R M AH L EY D

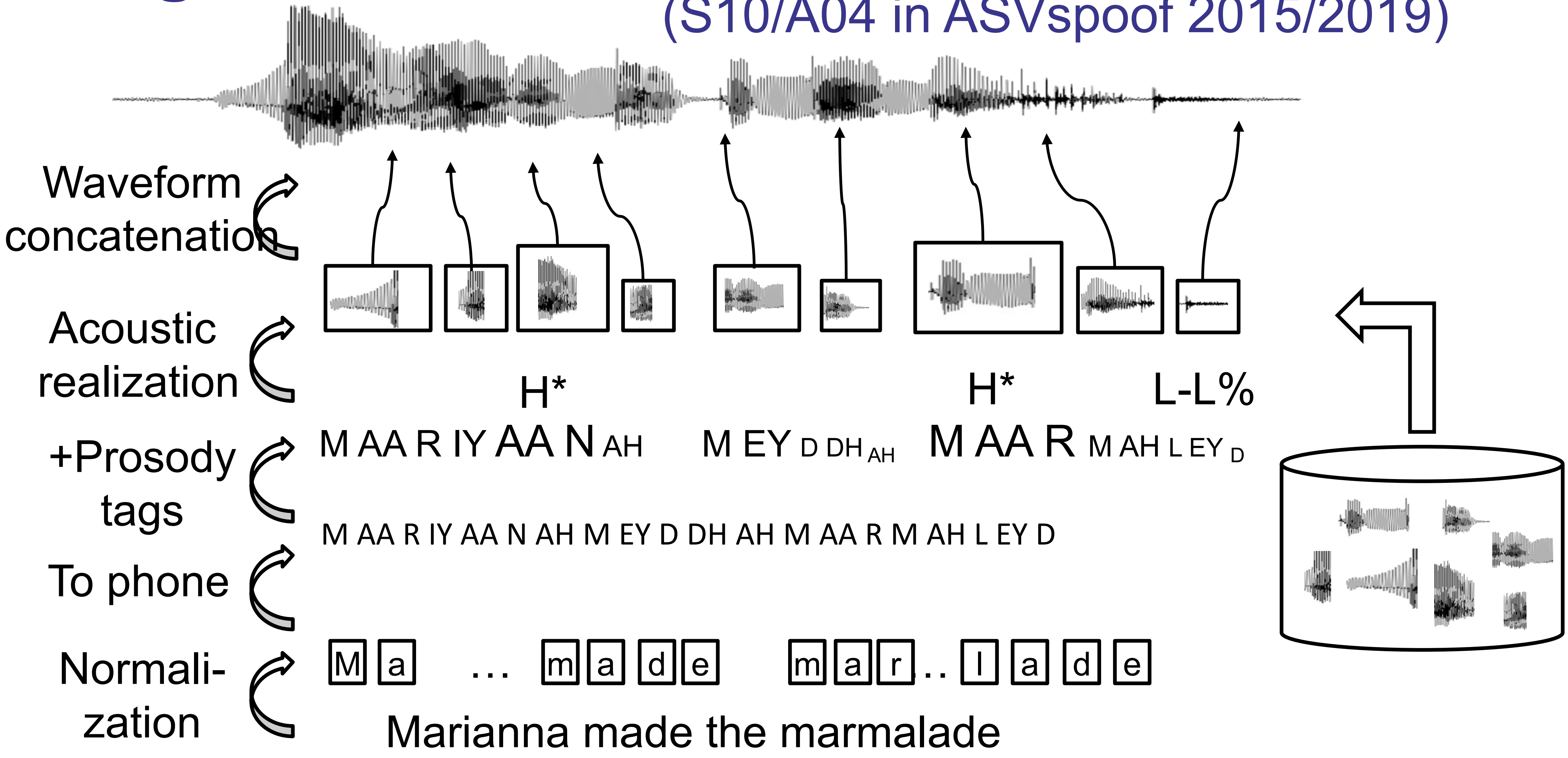
Normali-zation

M a … m a d e m a r .. l a d e

Marianna made the marmalade

Building TTS

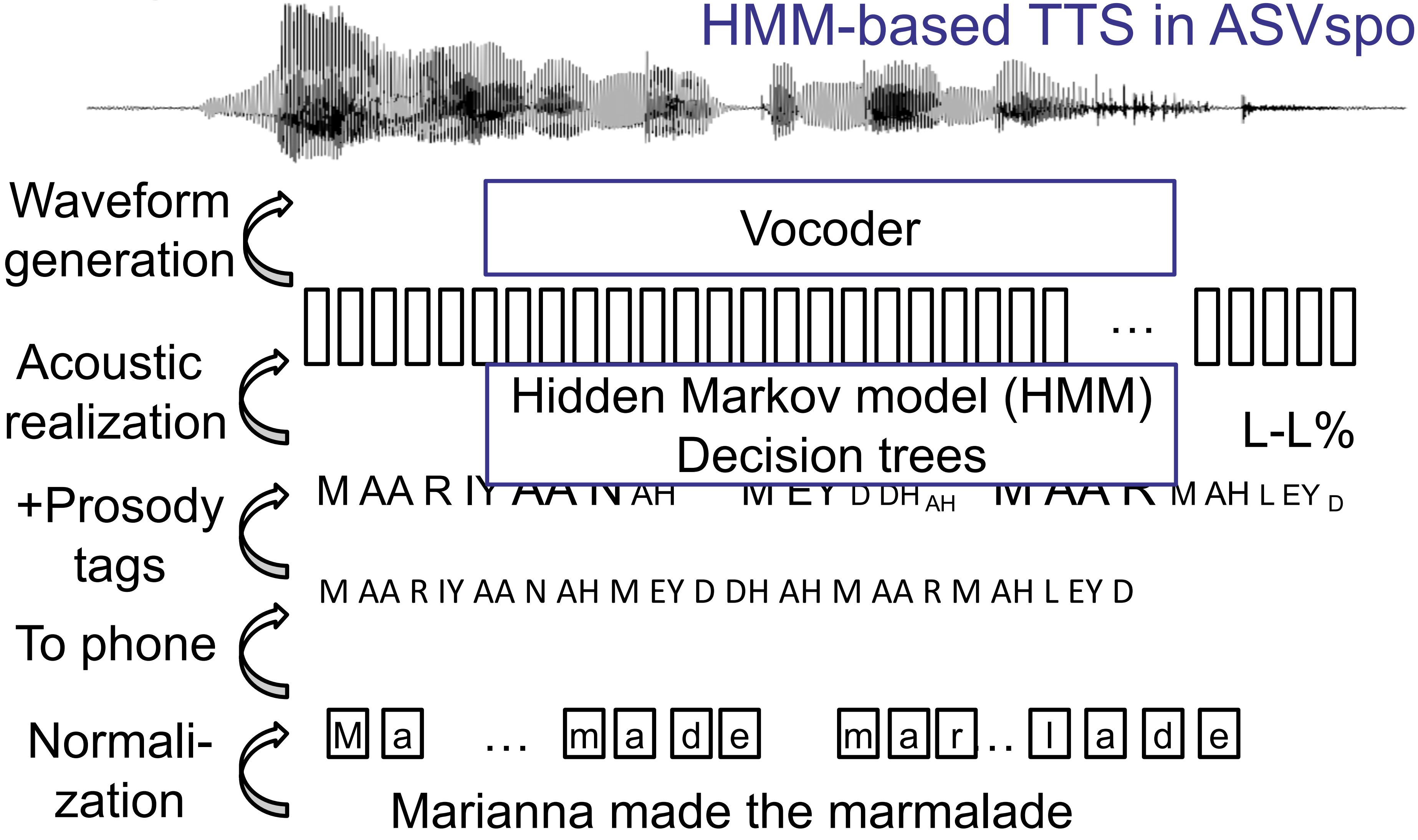
Unit-selection
(S10/A04 in ASVspoof 2015/2019)



Hunt, A. J. & Black, A. W. Unit selection in a concatenative speech synthesis system using a large speech database. in Proc. ICASSP 373–376 (1996).
Black, A. W. & Taylor, P. A. Automatically clustering similar units for unit selection in speech synthesis. (1997)

Building TTS

HMM-based TTS in ASVspoof 2015

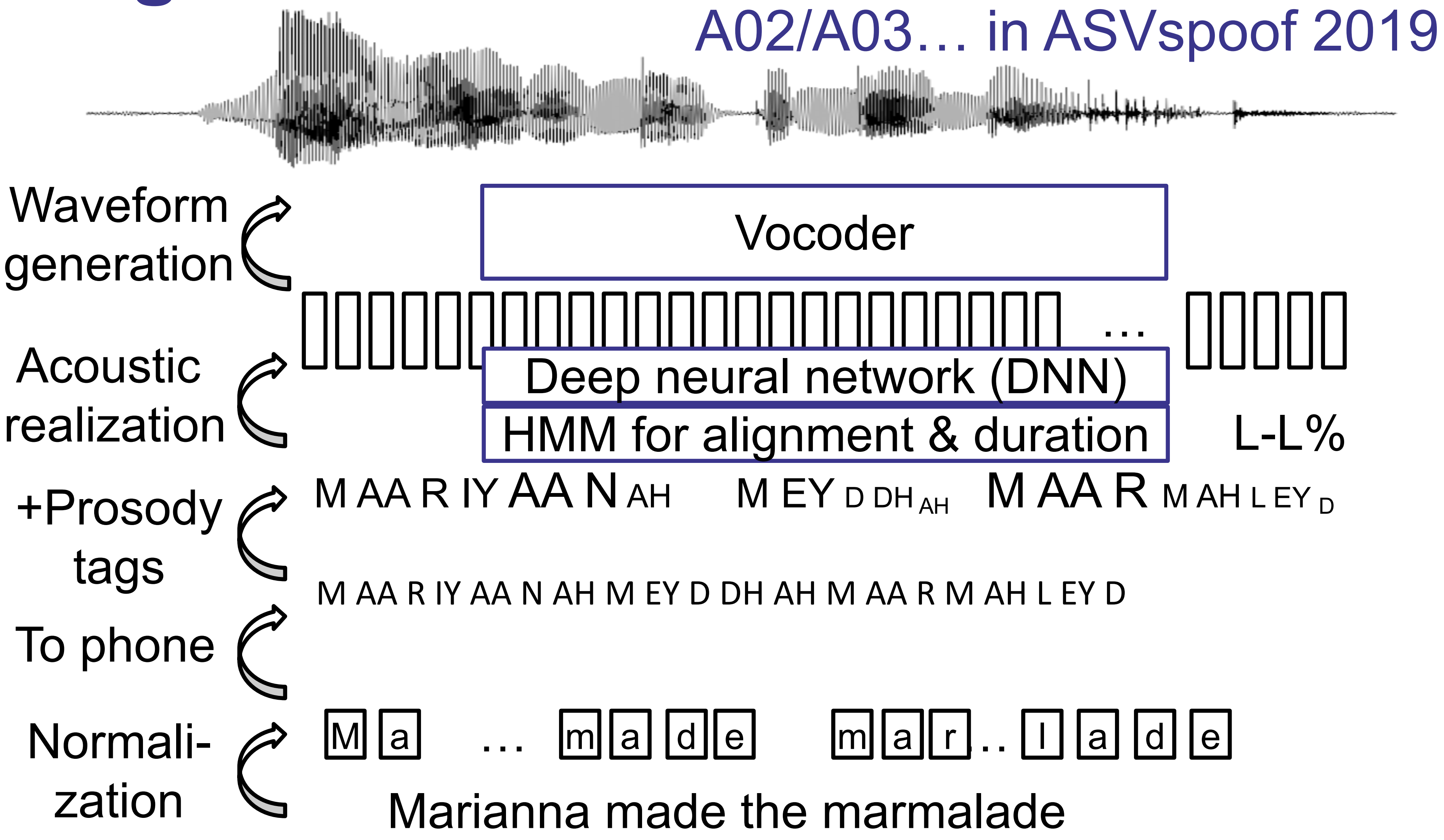


Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T. & Kitamura, T. Speech parameter generation algorithms for HMM-based speech synthesis. in Proc. ICASSP 936–939 (2000).
Keiichi Tokuda, Heiga Zen, and Alan W Black. An HMM-Based Speech Synthesis System Applied to English. In *Proc. SSW*, 227–230. 2002.
HMM-Based Speech Synthesis Toolkit (HTS), home page: http://hts.sp.nitech.ac.jp/?Welcome

Building TTS

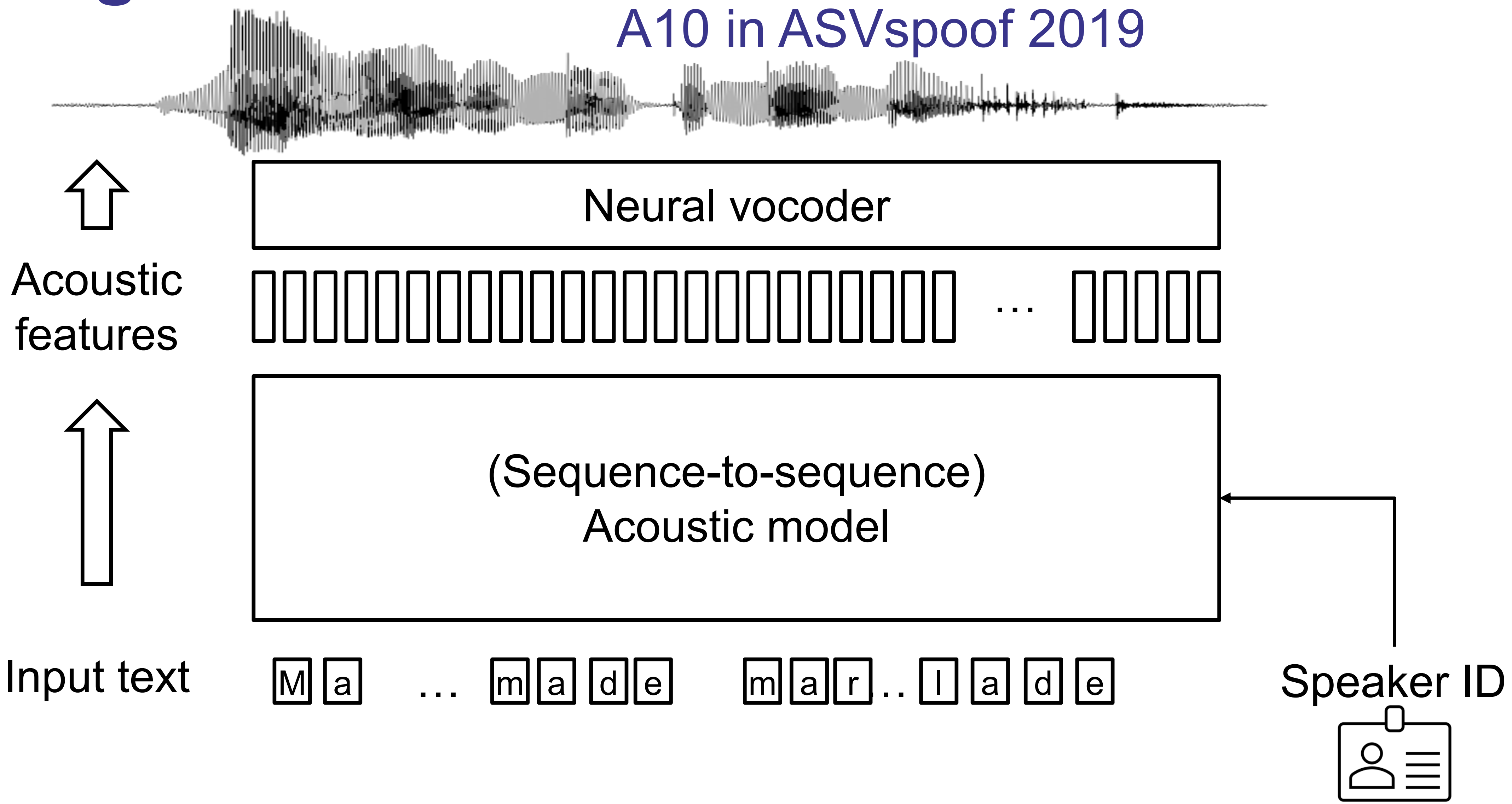
Thierry Dutoit. *An Introduction to Text-to-Speech Synthesis*. Norwell, MA, USA: Kluwer Academic Publishers. 1997.
Paul Taylor. *Text-to-Speech Synthesis*. Cambridge University Press. 2009.

Building TTS



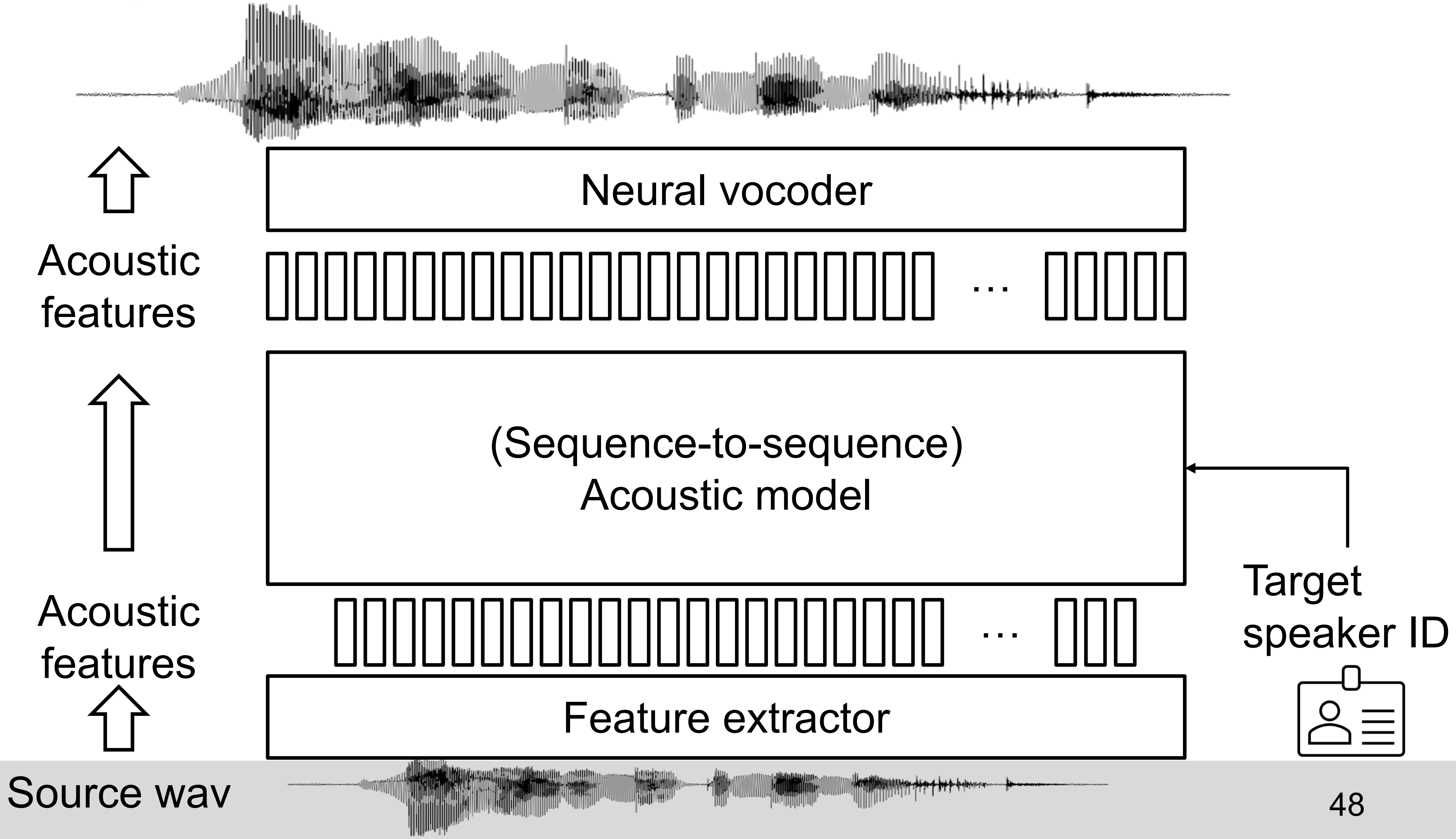
Heiga Zen, Alan Senior, and Martin Schuster. Statistical Parametric Speech Synthesis Using Deep Neural Networks. In *Proc. ICASSP*, 7962–7966. 2013.
Yuchen Fan, Yap Qian, Feilong Xie, and Frank K Soong. TTS Synthesis with Bidirectional LSTM Based Recurrent Neural Networks. In *Proc. Interspeech*, 1964–1968. 2014.
Heiga Zen, and Andrew Senior. Deep Mixture Density Networks for Acoustic Modeling in Statistical Parametric Speech Synthesis. In *Proc. ICASSP*, 3844–3848. 2014.

Building TTS



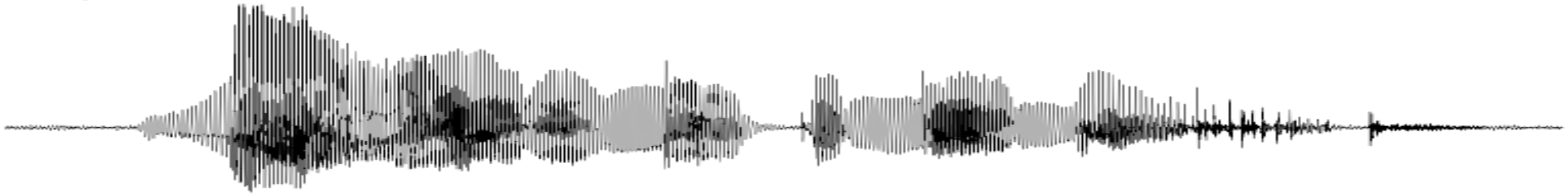
Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. A Survey on Neural Speech Synthesis. *ArXiv Preprint ArXiv:2106.15561*. 2021.
Xin Wang, Neural statistical parametric speech synthesis, ISCA Odyssey 2020, tutorial: https://tonywangx.github.io/slide.html#dec-2020

Building VC



Neural vocoder
Acoustic features
…
(Sequence-to-sequence)
Acoustic model
Target speaker ID
Acoustic features
…
Feature extractor
Source wav

Building TTS/VC is not that easy



Personal opinion

Expert knowledge

Interpretability

Softwares available

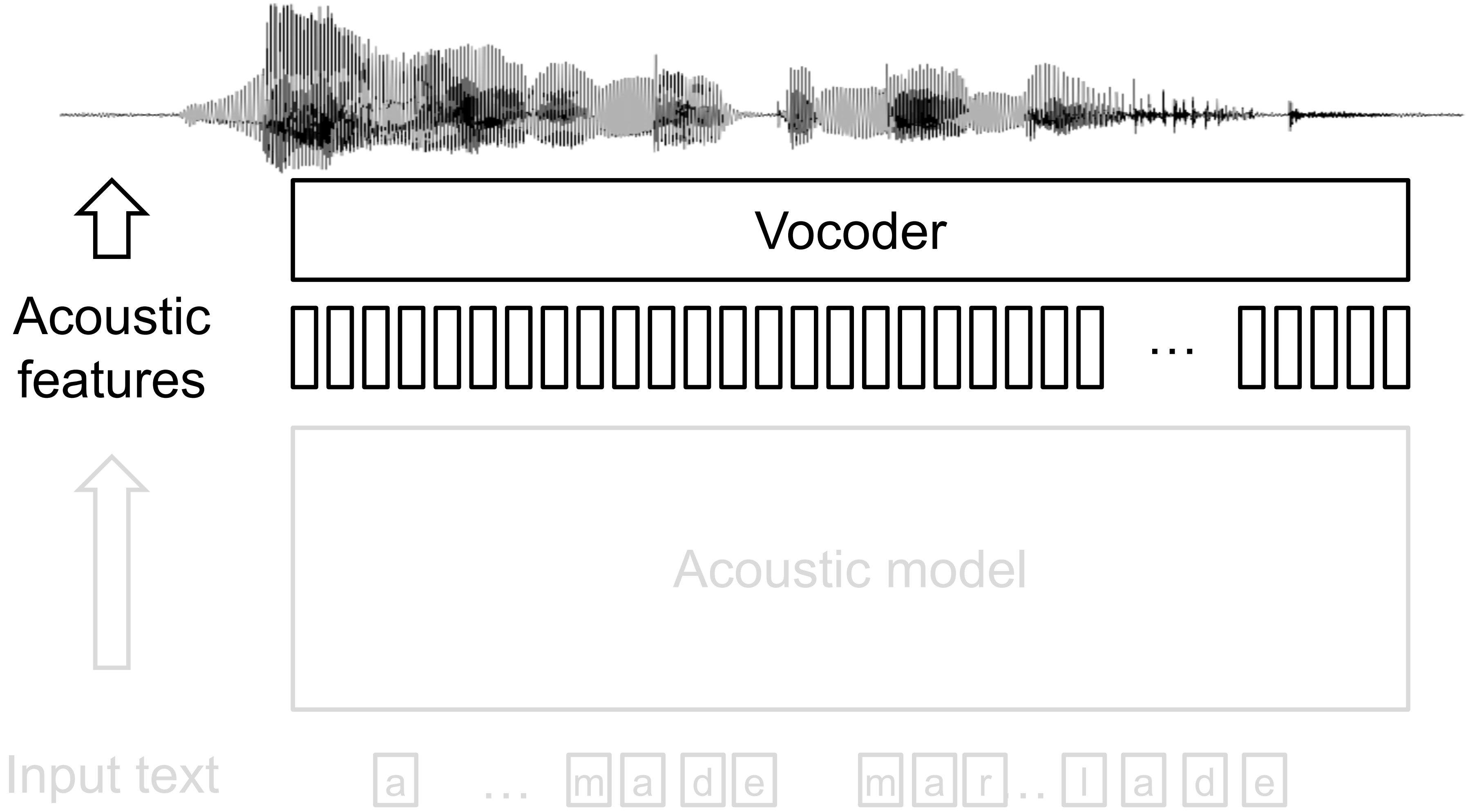
Computation power

Classifical methods → End-to-end methods

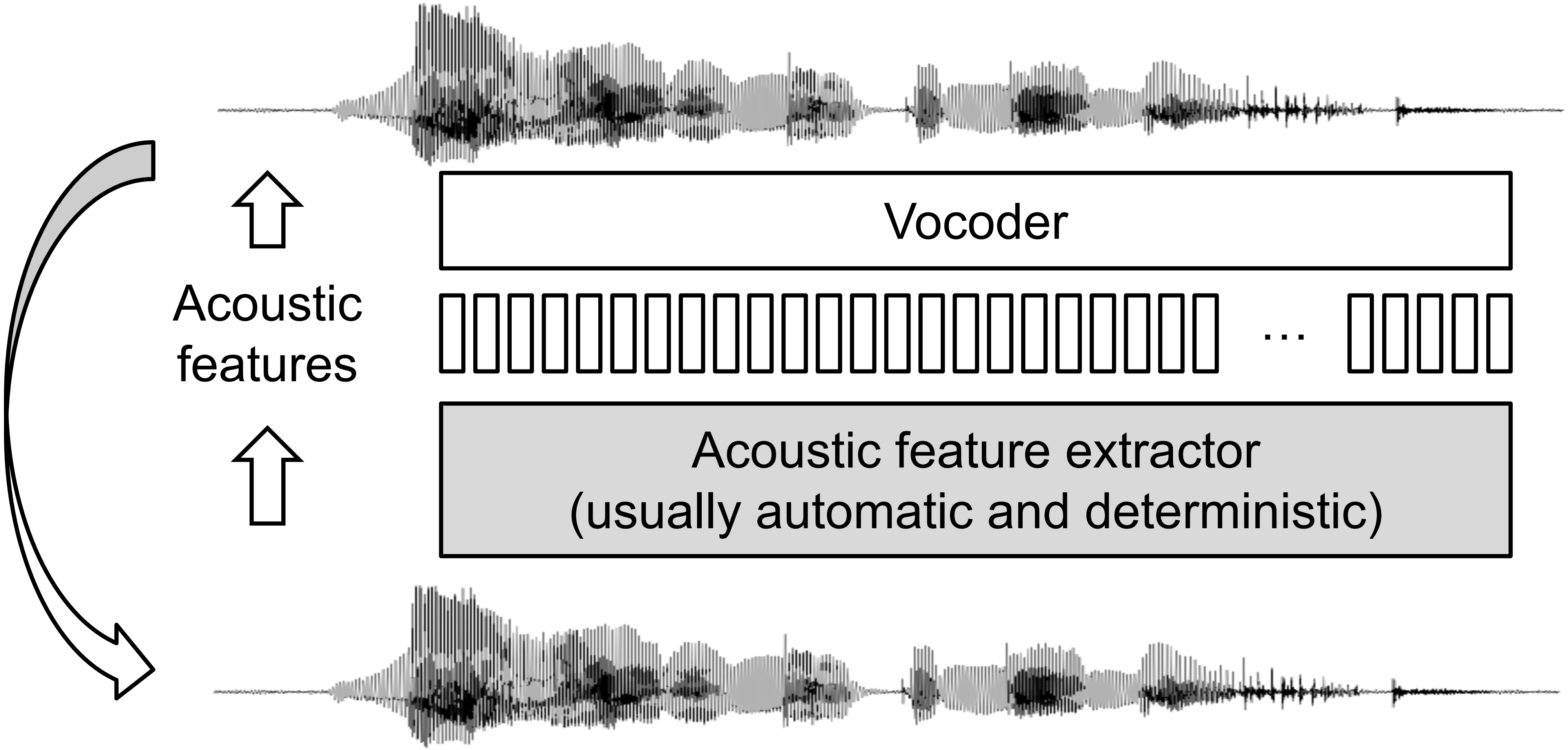
Source wav

Alternative method – vocoding

Vocoding


Vocoder
Acoustic features
…
Acoustic model
Input text
a … m a d e m a r … l a d e

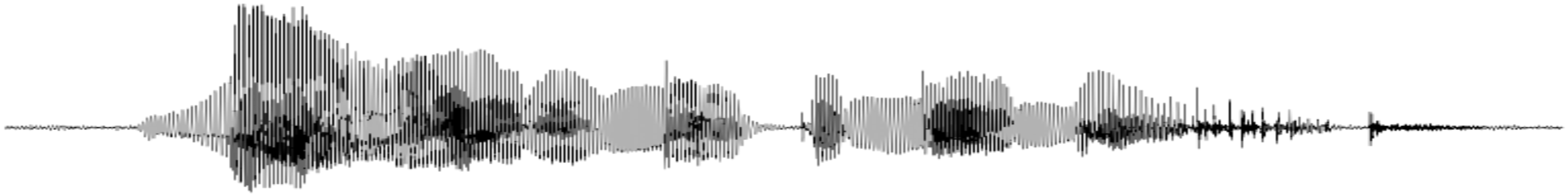
Vocoding



Copy-synthesis, analysis-by-synthesis, copy-resynthesis, …

Paul Taylor. *Text-to-Speech Synthesis*. Cambridge University Press. 2009.
J Holmes. The Influence of Glottal Waveform on the Naturalness of Speech from a Parallel Formant Synthesizer. *IEEE Transactions on Audio and Electroacoustics* 21 (3). IEEE: 298–305. 1973.
Wendy J Holmes. Copy Synthesis of Female Speech Using the JSRU Parallel Formant Synthesiser. In *EUROSPEECH*, 2513–2516. 1989.

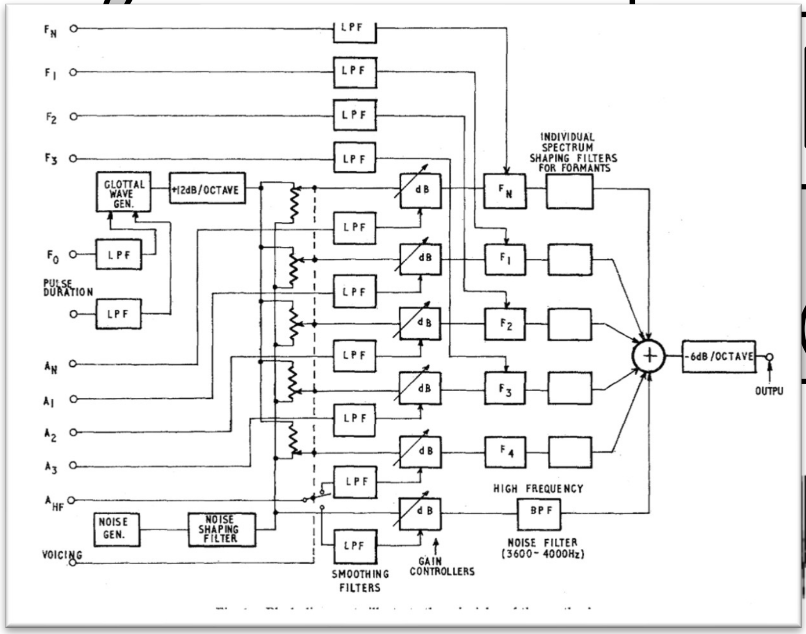
Vocoding



Formant synthesizer (deterministic)

…

Acoustic feature extractor
(usually automatic and deterministic)



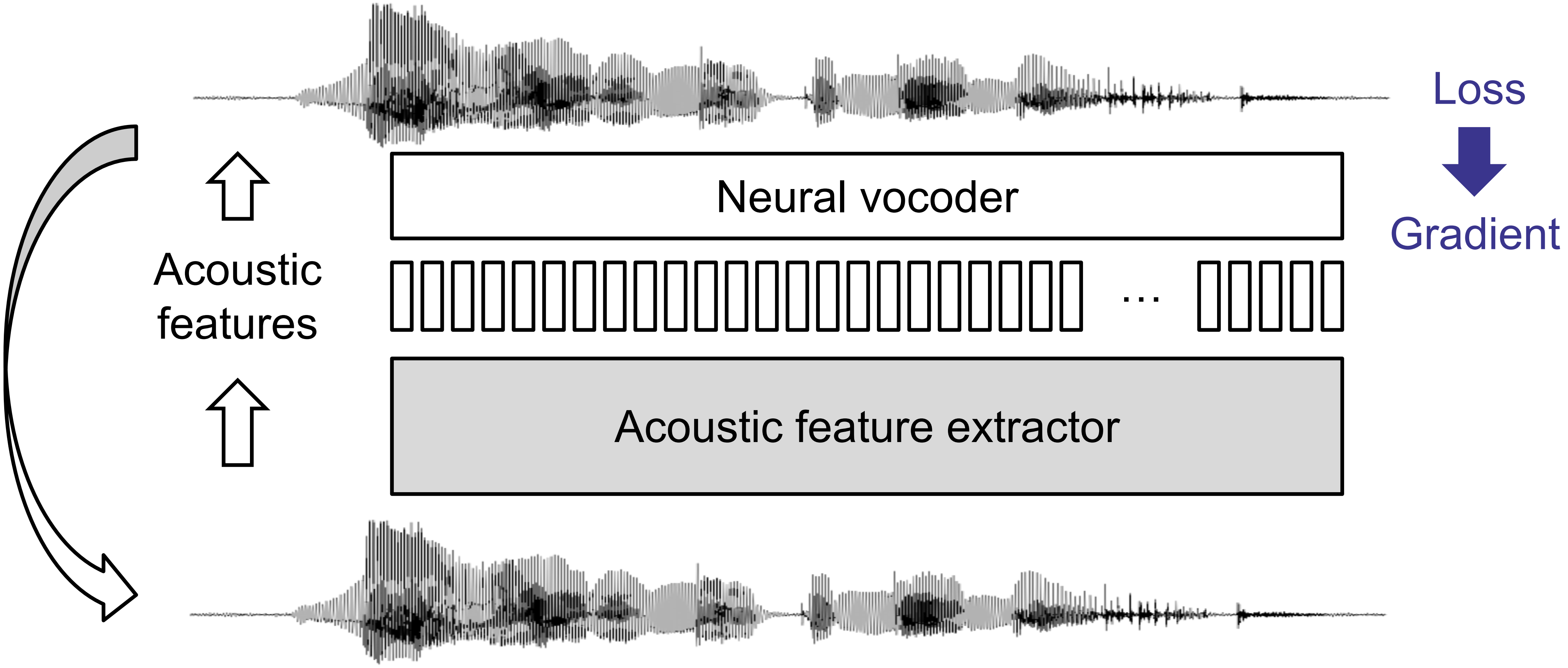
IV. Synthesis of Copies of Natural Utterances

preset bandwidth values. The techniques used are mostly of the "analysis-by-synthesis" type, with a human interpreter of differences between natural and synthetic speech in the feedback loop.

Copy-synthesis, analysis-by-syn

J Holmes. The Influence of Glottal Waveform on the Naturalness of Speech from a Parallel Formant Synthesizer. *IEEE Transactions on Audio and Electroacoustics* 21 (3). IEEE: 298–305. 1973.
Wendy J Holmes. Copy Synthesis of Female Speech Using the JSRU Parallel Formant Synthesiser. In *EUROSPEECH*, 2513–2516. 1989.

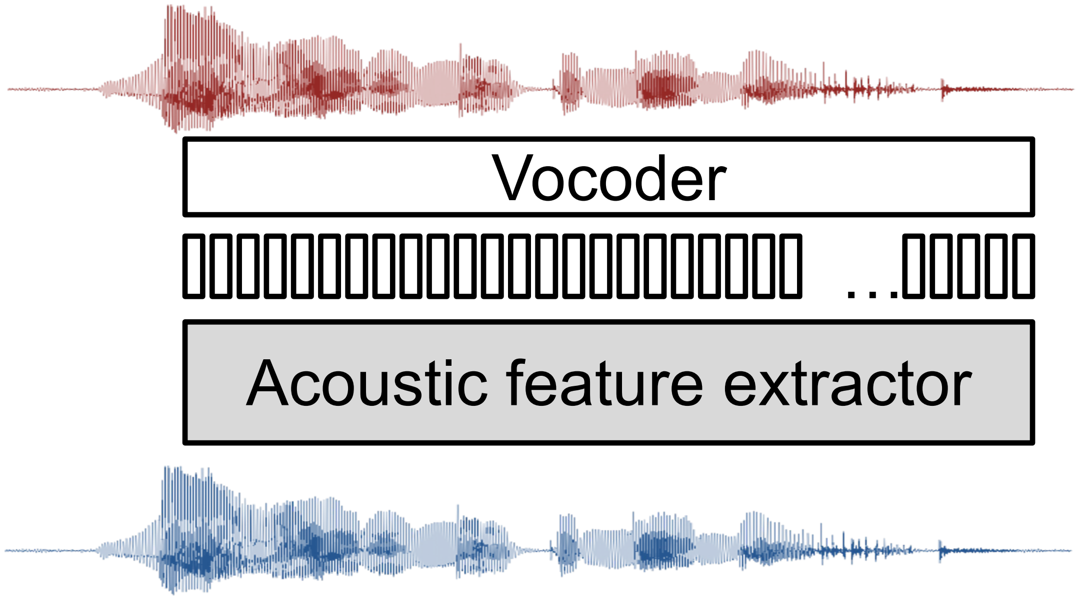
Vocoding using neural vocoder

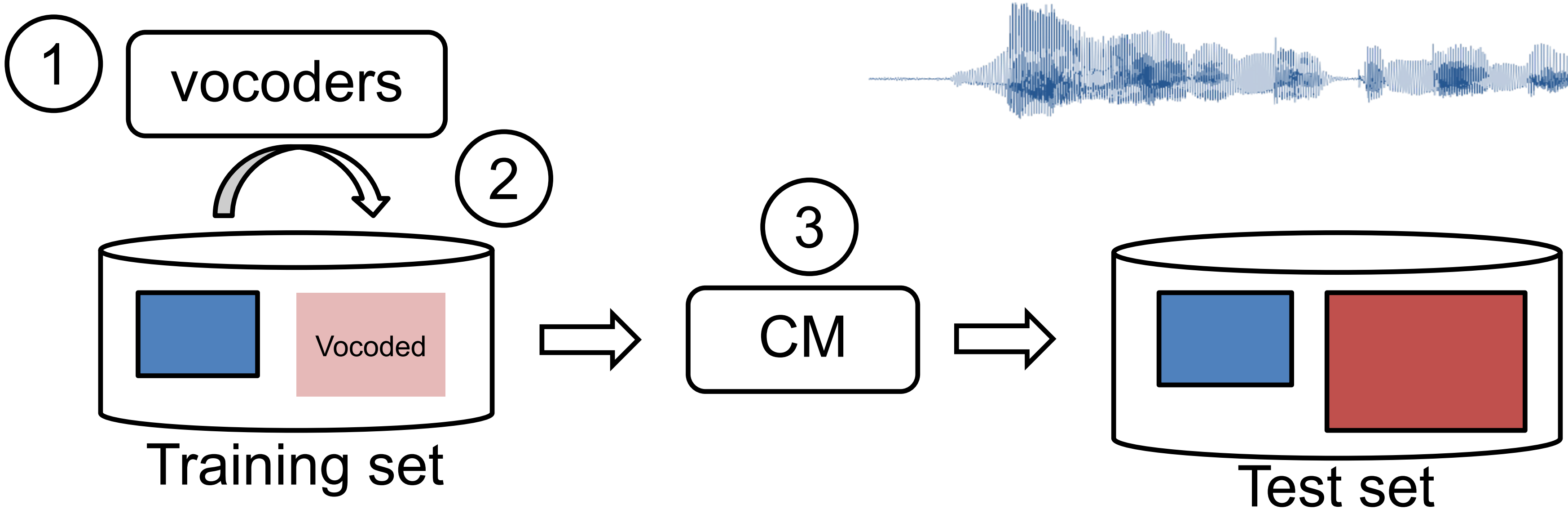


We do copy-synthesis when training the neural vocoders

Creating vocoded spoofed data

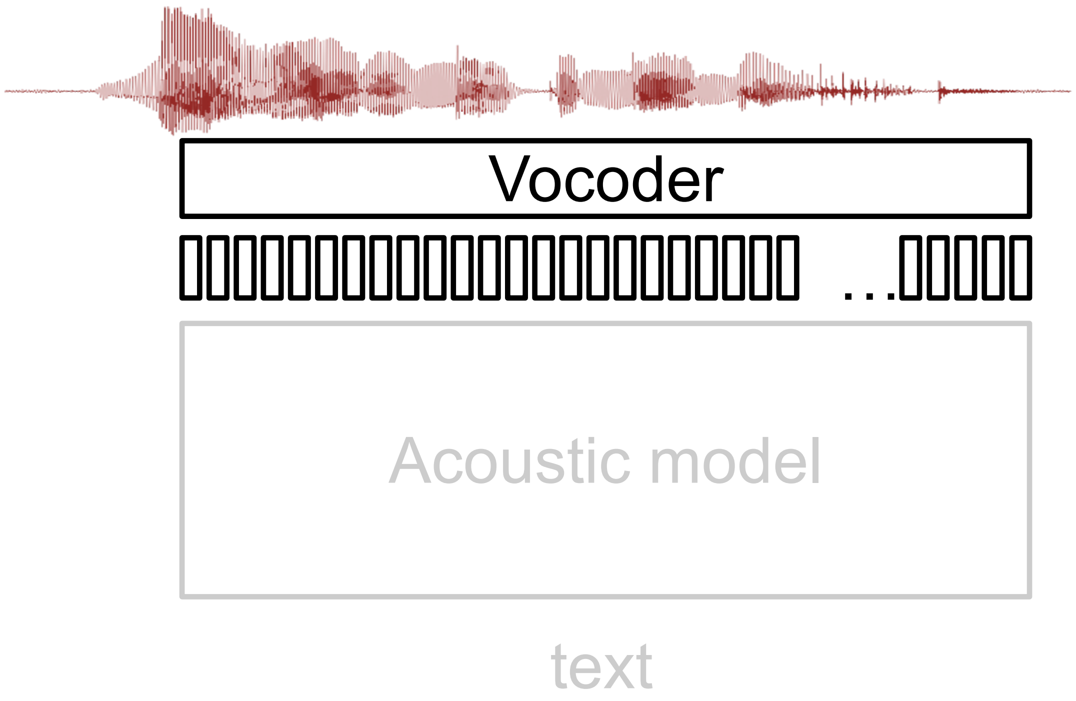
- **Three steps**
  1. Prepare (or training) vocoders
  2. Do vocoding on bona fide data
  3. Train the CM using {bona fide, vocoded spoofed}





Creating vocoded spoofed data

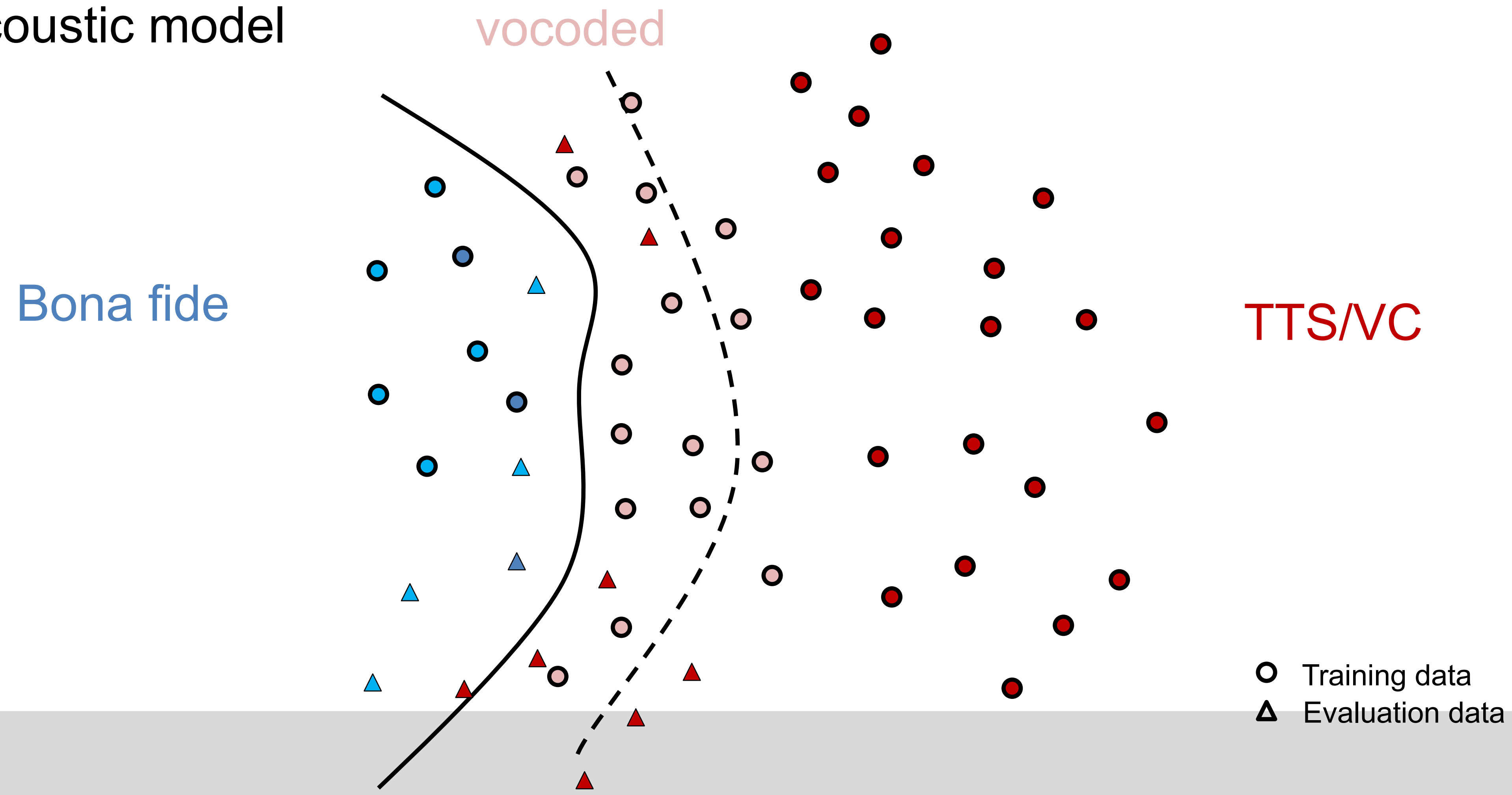
- Assumptions
  - Vocoding is TTS/VC with a perfect acoustic model
    - x artefacts by the acoustic model
    - ✓ artefacts by the vocoder



Creating vocoded spoofed data

❑ Assumptions

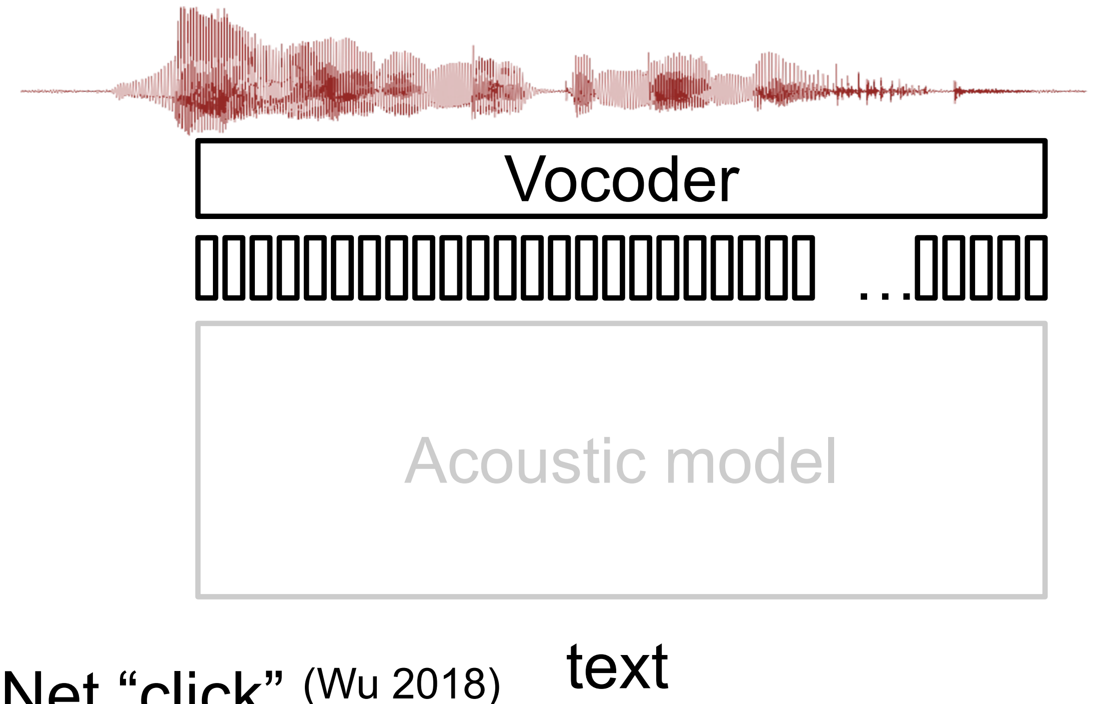
- Vocoding is TTS/VC with a perfect acoustic model



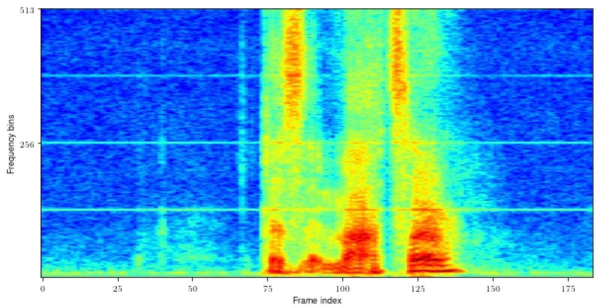
Creating vocoded spoofed data

❑ Assumptions

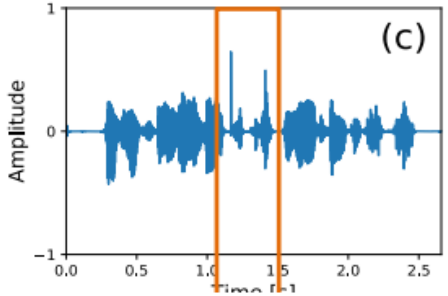
- Vocoding is TTS/VC with a perfect acoustic model
- Actual TTS/VC spoofed data contain artefacts by the vocoder



WaveGlow "bar" (Prenger 2019)



WaveNet "click" (Wu 2018)



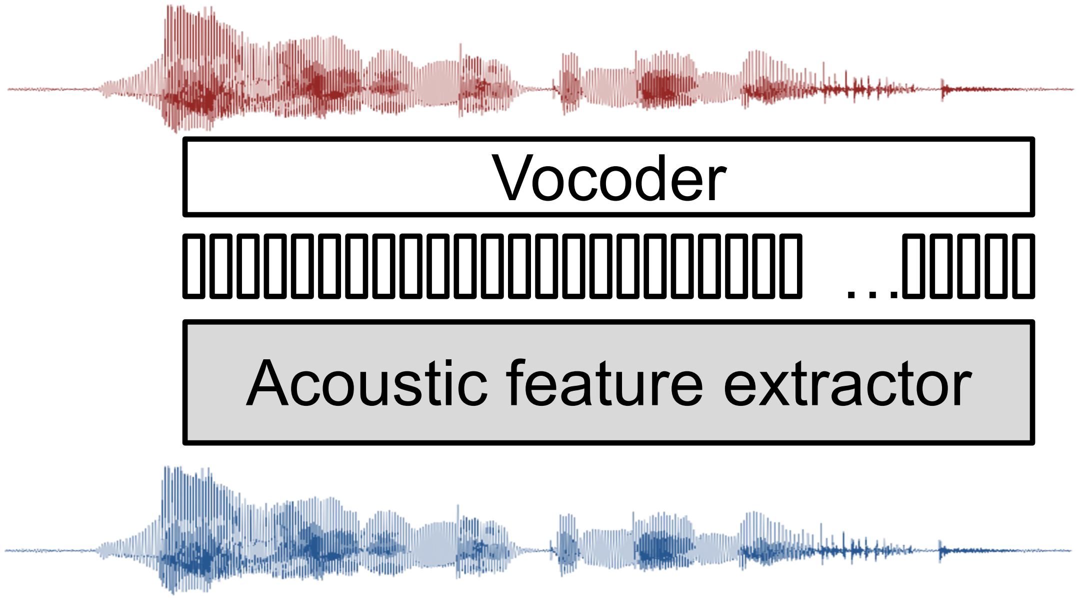
…

Ryan Prenger, Rafael Valle, and Bryan Catanzaro. WaveGlow: A Flow-Based Generative Network for Speech Synthesis. In *Proc. ICASSP*, 3617–3621. 2019.
Yi-Chiao Wu, Kazuhiro Kobayashi, Tomoki Hayashi, Patrick Lumban Tobing, and Tomoki Toda. Collapsed Speech Segment Detection and Suppression for WaveNet Vocoder. In *Interspeech 2018*, 1988–1992. ISCA: ISCA. doi:10.21437/Interspeech.2018-1210. 2018.

Creating vocoded spoofed data

❑ Potential benefits

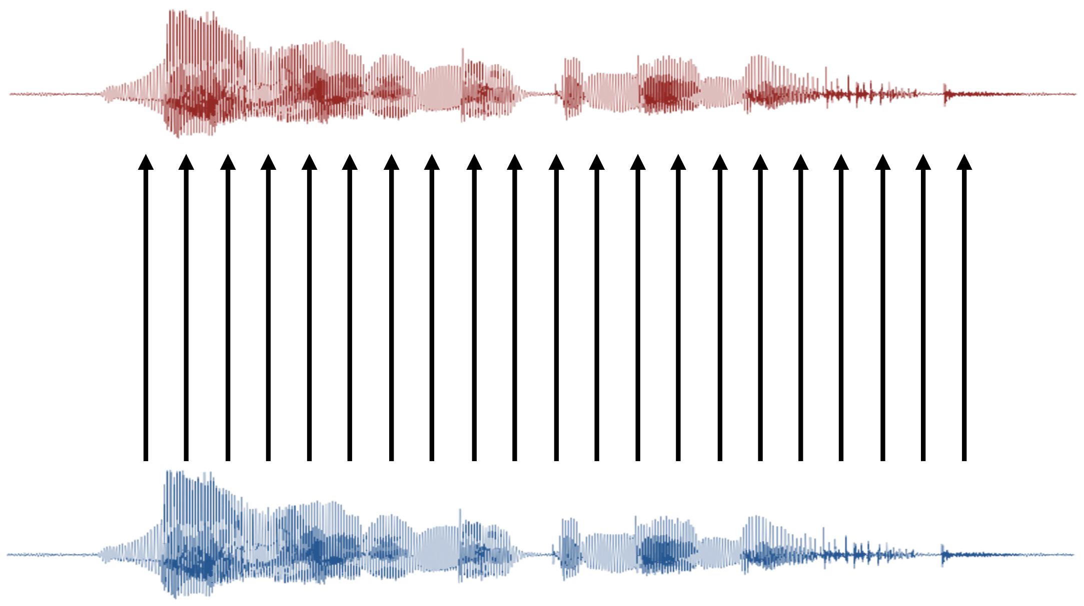
- Preparing vocoders is easier
  - ~~linguistic knowledge~~
  - ~~transcription / annotation~~
  - ~~speaker embedding~~
  - …



Creating vocoded spoofed data

❑ Potential benefits

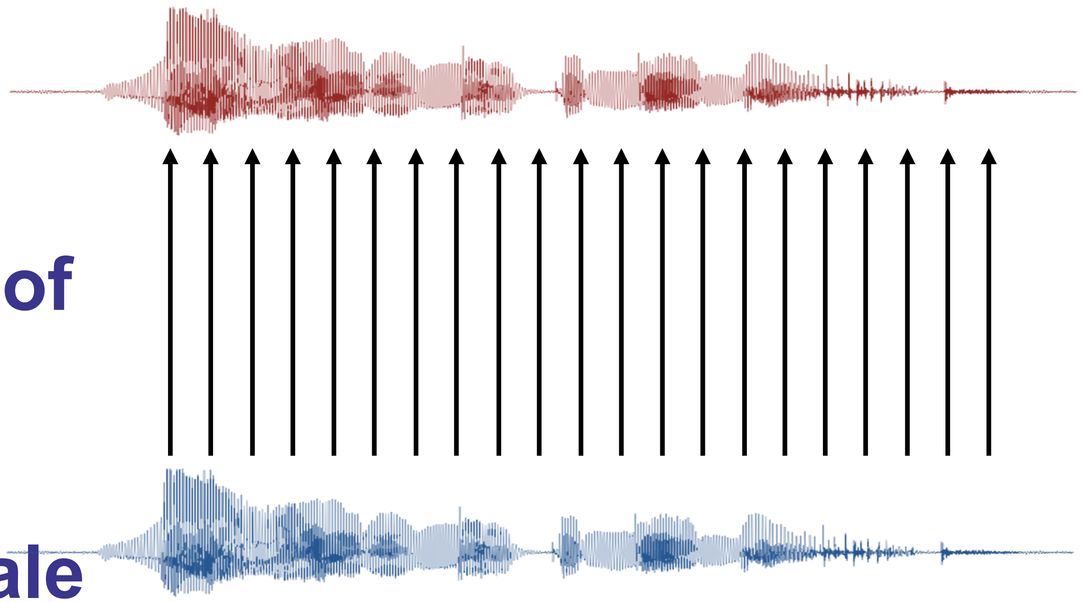
- Preparing vocoders is easier
  - ~~linguistic knowledge~~
  - ~~transcription / annotation~~
  - ~~speaker embedding~~
  - …



- Bona fide and vocoded waveforms are aligned in time

Questions

- Which vocoders? How to train?
  - pre-trained?
  - fine tuning?



- How to exploit the aligned pair of {bona fide, vocoded spoofed}?
- Improvement of using large scale vocoded data?

Vocoded spoofed data is a not new idea

- Using DSP-based vocoders

- Xingming Wang, Xiaoyi Qin, Tinglong Zhu, Chao Wang, Shilei Zhang, and Ming Li. The DKU-CMRI System for the ASVspoof 2021 Challenge: Vocoder Based Replay Channel Response Estimation. In *Proc. ASVspoof challenge workshop*, 16–21. 2021.
- Monisankha Pal, Dipjyoti Paul, and Goutam Saha. Synthetic Speech Detection Using Fundamental Frequency Variation and Spectral Features. *Computer Speech & Language* 48. Elsevier: 31–50. 2018.
- Ibon Saratxaga, Jon Sanchez, Zhizheng Wu, Inma Hernaez, and Eva Navas. Synthetic Speech Detection Using Phase Information. *Speech Communication* 81 (July): 30–41. doi:10.1016/j.specom.2016.04.001. 2016.
- Aleksandr Sizov, Elie Khoury, Tomi Kinnunen, Zhizheng Wu, and Sébastien Marcel. Joint Speaker Verification and Antispoofing in the I-Vector Space. *IEEE Transactions on Information Forensics and Security* 10 (4). IEEE: 821–832. doi:10.1109/TIFS.2015.2407362. 2015.
- Elie Khoury, Tomi Kinnunen, Aleksandr Sizov, Zhizheng Wu, and Sébastien Marcel. Introducing I-Vectors for Joint Anti-Spoofing and Speaker Verification. In *Proc. Interspeech*, 61–65. 2014.
- Jon Sanchez, Ibon Saratxaga, Inma Hernaez, Eva Navas, and Daniel Erro. A Cross-Vocoder Study of Speaker Independent Synthetic Speech Detection Using Phase Information. In *Proc. Interspeech*. 2014.
- Zhizheng Wu, Xiong Xiao, Eng Siong Chng, and Haizhou Li. Synthetic Speech Detection Using Temporal Modulation Feature. In *Proc. ICASSP*, 7234–7238. 2013.

- Using neural vocoders

- Joel Frank, and Lea Schönherr. WaveFake: A Data Set to Facilitate Audio DeepFake Detection. In Proc. NeurIPS Datasets and Benchmarks 2021. 2021.
- Chengzhe Sun, Shan Jia, Shuwei Hou, Ehab AlBadawy, and Siwei Lyu. Exposing AI-Synthesized Human Voices Using Neural Vocoder Artifacts. ArXiv Preprint ArXiv:2302.09198. 2023.

Question 1

- ❑ Which vocoders? How to train?

Which vocoders?

❑ Options

Not necessary

- ~~Digital-signal-processing (DSP)~~
- ~~Autoregressive DNN~~
- Non-autoregressive DNN+DSP

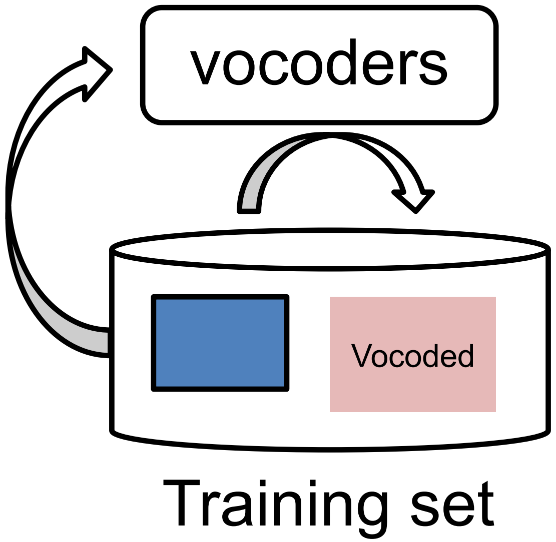
❑ Constraints

- Fast generation speed, much faster than real-time speed

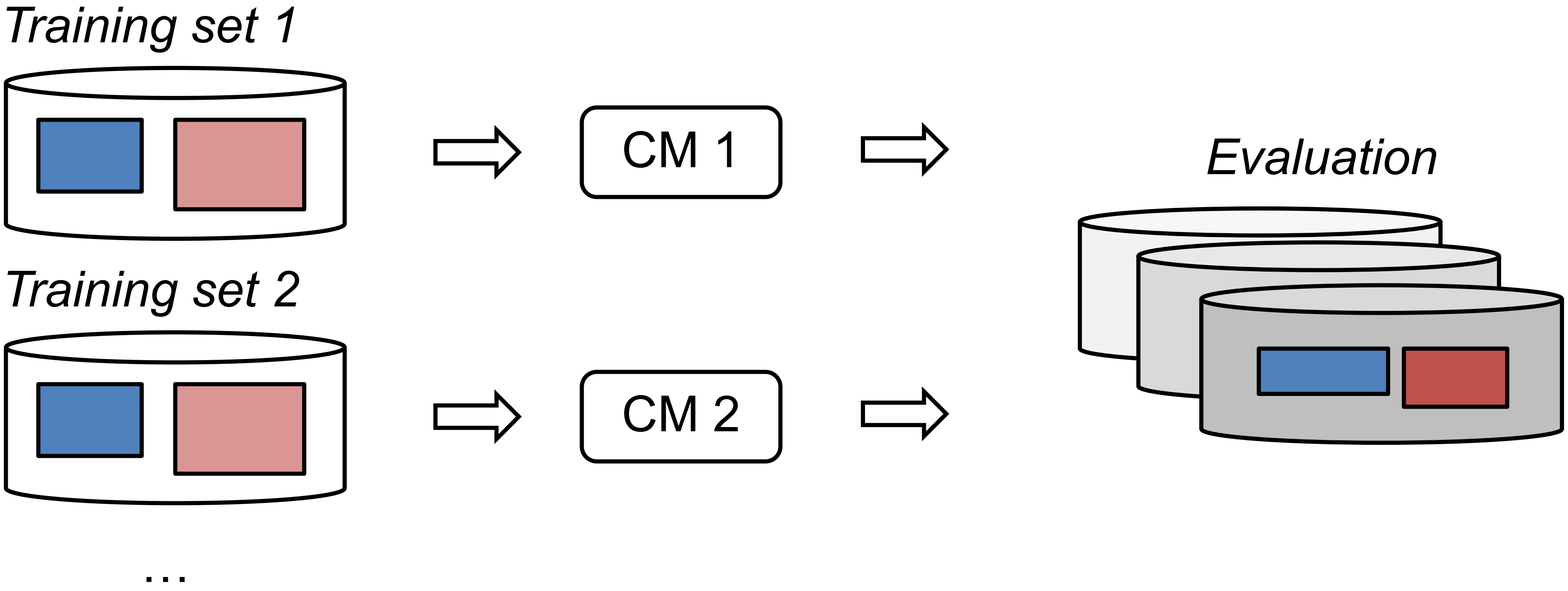
How to train?

❑ Options

- Pre-trained, off-the-shelf
- Pre-trained, fine-tuning
- Trained from scratch

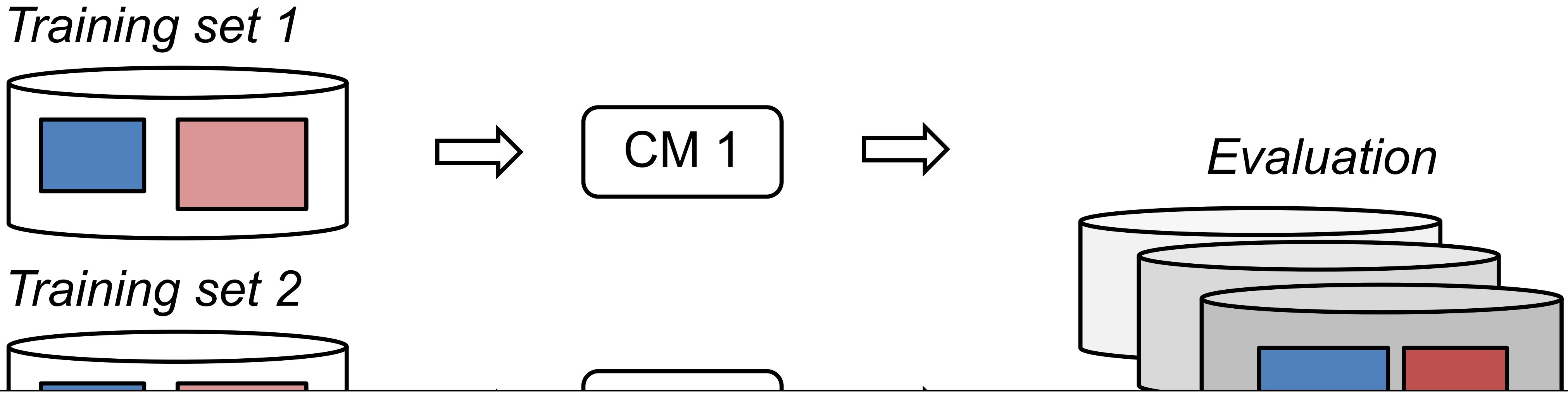


Experiment



Variable of interest

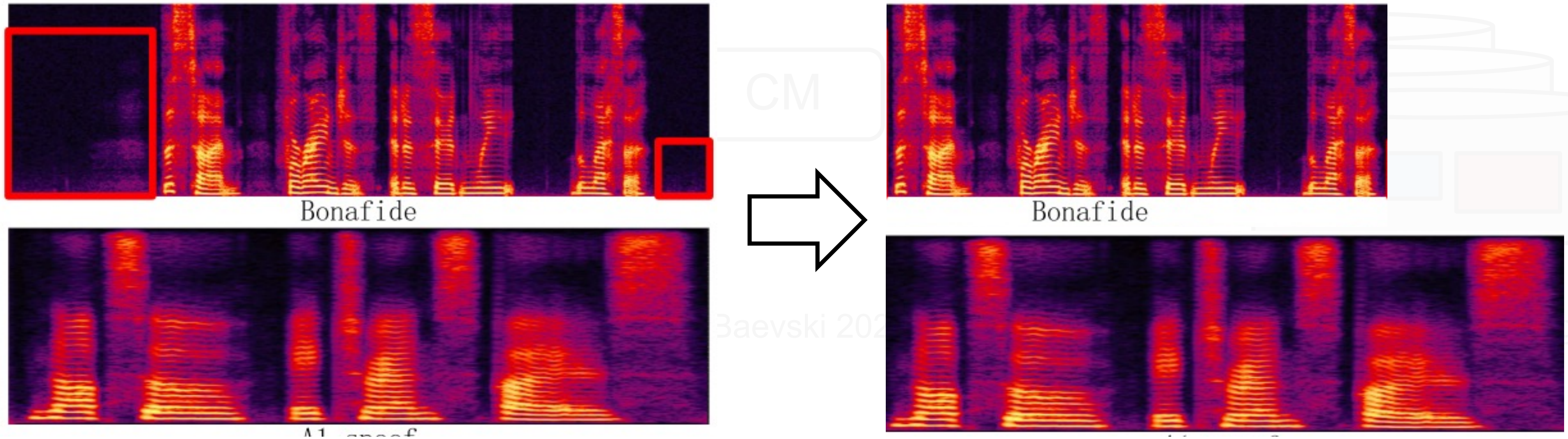
Experiment



- CM: Wav2vec2.0 front end (Baevski 2020) + pooling + linear output (Wang 2022)
- Evaluation on
  - ASVspoof 2019 LA test set, 2021 LA & DF eval sets
  - ASVspoof 2019 LA test set w/o non-speech, 2021 LA & DF hidden track
  - WaveFake (Frank 2021), In-the-Wild (Müller 2022)

Xin Wang, and Junichi Yamagishi. Investigating Self-Supervised Front Ends for Speech Spoofing Countermeasures. In *Proc. Odyssey*, 100–106. doi:10.21437/Odyssey.2022-14. 2022.
Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Proc. NIPS*, 33:12449–12460. 2020.
Joel Frank, and Lea Schönherr. WaveFake: A Data Set to Facilitate Audio DeepFake Detection. In *Proc. NeurIPS Datasets and Benchmarks 2021*. 2021.
Nicolas M Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger. Does Audio Deepfake Detection Generalize? In *Proc. Interspeech*, 2783–2787. 2022.

Experiment



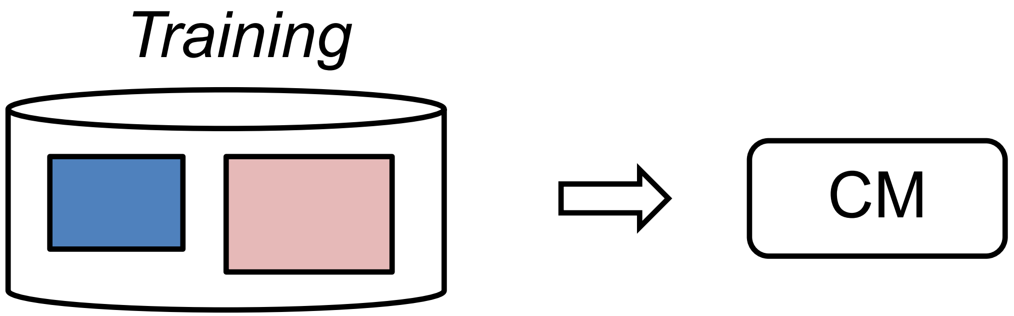
– ASVspoof 2019 LA test set w/o non-speech, 2021 LA & DF hidden track

I personally recommend using both versions of ASVspoof test sets

Xin Wang, and Junichi Yamagishi. Investigating Self-Supervised Front Ends for Speech Spoofing Countermeasures. In *Proc. Odyssey*, 100–106. doi:10.21437/Odyssey.2022-14. 2022.
Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Proc. NIPS*, 33:12449–12460. 2020.
Joel Frank, and Lea Schönherr. WaveFake: A Data Set to Facilitate Audio DeepFake Detection. In *Proc. NeurIPS Datasets and Benchmarks 2021*. 2021.
Nicolas M Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger. Does Audio Deepfake Detection Generalize? In *Proc. Interspeech*, 2783–2787. 2022.

Experiment

❑ CM training sets in comparison

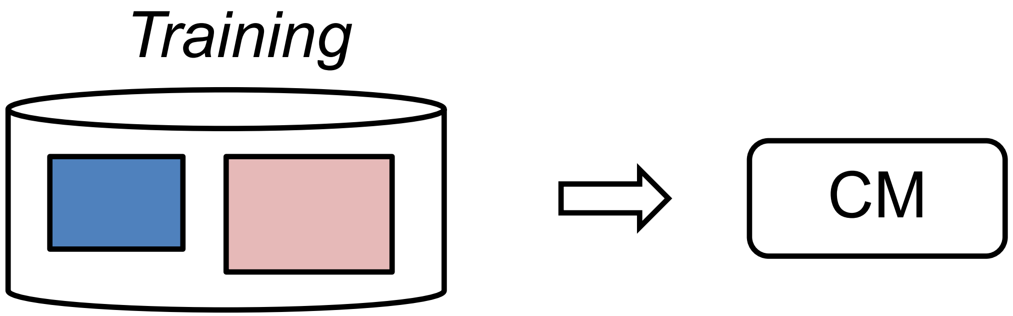


ID	#. Spr.	#. Bona.	#. Spoof.	Vocoder type	Implementation	Vocoder train/fine-tune data
LA19trn	20	2,580	22,800	-	-	-
`WFtrn`	1	3,930	15,720	HiFiGAN, MB-MelGAN, PWG, WaveGlow	ESPNet toolkit	LJSpeech / -
`Voc.v1`	20 same as `LA19trn`	2,580	10,320	HiFiGAN, MB-MelGAN, PWG, StyleMelGAN	ESPNet toolkit	LibriTTS / -
`Voc.v2`				HiFiGAN, NSF-HiFiGAN, Hn-NSF, WaveGlow	in-house	LibriTTS / -
`Voc.v3`				HiFiGAN, NSF-HiFiGAN, Hn-NSF, WaveGlow	in-house	`LA19trn` bona. / -
`Voc.v4`				HiFiGAN, NSF-HiFiGAN, Hn-NSF, WaveGlow	in-house	LibriTTS / `LA19trn` bona.

- `LA19trn`: ASVspoof 2019 LA training set (bona fide + TTS/VC)
- `WFtrn`: WaveFake English subset (bona fide + vocoded)
- `Voc.v*`: ASVspoof 2019 LA training set bona fide data + vocoded

Experiment

❑ CM training sets in comparison



ID	#. Spr.	#. Bona.	#. Spoof.	Vocoder type	Implementation	Vocoder train/fine-tune data
LA19trn	20	2,580	22,800	-	-	-
WFtrn	1	3,930	15,720	HiFiGAN, MB-MelGAN, PWG, WaveGlow	ESPNet toolkit	LJSpeech / -
Voc.v1	20 same as LA19trn	2,580	10,320	HiFiGAN, MB-MelGAN, PWG, StyleMelGAN	ESPNet toolkit	LibriTTS / -
Voc.v2				HiFiGAN, NSF-HiFiGAN, Hn-NSF, WaveGlow	in-house	LibriTTS / -
Voc.v3				HiFiGAN, NSF-HiFiGAN, Hn-NSF, WaveGlow	in-house	LA19trn bona. / -
Voc.v4				HiFiGAN, NSF-HiFiGAN, Hn-NSF, WaveGlow	in-house	LibriTTS / LA19trn bona.

How to train the vocoder

- LA19trn: ASVspoof 2019 LA training set (bona fide + TTS/VC)
- WFtrn: WaveFake English subset (bona fide + vocoded)
- Voc.v*: ASVspoof 2019 LA training set bona fide data + vocoded

Experiment results

❑ EER (%, mean of three runs)

☺ Low EER ☹ High EER

	Test sets	LA19 trn	WF trn	Training set Voc. v1	Voc. v2	Voc. v3	Voc. v4
ASVspoof 2019 LA →	LA19eval	2.98	44.48	5.78	5.32	8.74	4.36
ASVspoof 2021 LA →	LA21eval	7.53	41.57	26.30	17.98	19.29	24.39
ASVspoof 2021 DF →	DF21eval	6.67	24.26	11.95	11.54	9.71	13.31
	LA19etrim	15.56	31.62	23.29	16.16	14.99	9.52
	LA21hid	28.80	27.60	28.30	19.49	17.62	21.43
	DF21hid	23.62	26.18	22.01	13.92	13.50	16.99
	WaveFake	15.76	-	39.27	34.05	17.10	10.89
	InWild	26.65	19.98	41.06	36.46	22.26	19.45
Single EER threshold →	Pooled	14.24	-	36.57	39.95	19.39	16.35

ID	#. Spr.	#. Bona.	#. Spoof.	Vocoder type	Implementation	Vocoder train/fine-tune data
LA19trn	20	2,580	22,800	-	-	-
WFtrn	1	3,930	15,720	HiFiGAN, MB-MelGAN, PWG, WaveGlow	ESPNet toolkit	LJSpeech / -
Voc.v1	20 same as LA19trn	2,580	10,320	HiFiGAN, MB-MelGAN, PWG, StyleMelGAN	ESPNet toolkit	LibriTTS / -
Voc.v2				HiFiGAN, NSF-HiFiGAN, Hn-NSF, WaveGlow	in-house	LibriTTS / -
Voc.v3				HiFiGAN, NSF-HiFiGAN, Hn-NSF, WaveGlow	in-house	LA19trn bona. / -
Voc.v4				HiFiGAN, NSF-HiFiGAN, Hn-NSF, WaveGlow	in-house	LibriTTS / LA19trn bona.

		Training set					
		LA19 trn	WF trn	Voc. v1	Voc. v2	Voc. v3	Voc. v4
Test sets	LA19eval	2.98	44.48	5.78	5.32	8.74	4.36
	LA21eval	7.53	41.57	26.30	17.98	19.29	24.39
	DF21eval	6.67	24.26	11.95	11.54	9.71	13.31
	LA19etrim	15.56	31.62	23.29	16.16	14.99	9.52
	LA21hid	28.80	27.60	28.30	19.49	17.62	21.43
	DF21hid	23.62	26.18	22.01	13.92	13.50	16.99
	WaveFake	15.76	-	39.27	34.05	17.10	10.89
	InWild	26.65	19.98	41.06	36.46	22.26	19.45
	Pooled	14.24	-	36.57	39.95	19.39	16.35

Xin Wang, and Junichi Yamagishi. Investigating Self-Supervised Front Ends for Speech Spoofing Countermeasures. In *Proc. Odyssey*, 100–106. doi:10.21437/Odyssey.2022-14. 2022.
Nicolas Müller, Franziska Dieckmann, Pavel Czempin, Roman Canals, Konstantin Böttinger, and Jennifer Williams. Speech Is Silver, Silence Is Golden: What Do ASVspoof-Trained Models Really Learn? In *Proc. ASVspoof Challenge Workshop*, 55–60. doi:10.21437/ASVSPOOF.2021-9. 2021.
Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, and Kong Aik Lee. ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild. *ArXiv Preprint ArXiv:2210.02437*. 2022.

ID	#. Spr.	#. Bona.	#. Spoof.	Vocoder type	Implementation	Vocoder train/fine-tune data
LA19trn	20	2,580	22,800	-	-	-
WFtrn	1	3,930	15,720	HiFiGAN, MB-MelGAN, PWG, WaveGlow	ESPNet toolkit	LJSpeech / -
Voc.v1	20 same as LA19trn	2,580	10,320	HiFiGAN, MB-MelGAN, PWG, StyleMelGAN	ESPNet toolkit	LibriTTS / -
Voc.v2				HiFiGAN, NSF-HiFiGAN, Hn-NSF, WaveGlow	in-house	LibriTTS / -
Voc.v3				HiFiGAN, NSF-HiFiGAN, Hn-NSF, WaveGlow	in-house	LA19trn bona. / -
Voc.v4				HiFiGAN, NSF-HiFiGAN, Hn-NSF, WaveGlow	in-house	LibriTTS / LA19trn bona.

Test sets	Training set: LA19 trn	WF trn	Voc. v1	Voc. v2	Voc. v3	Voc. v4
LA19eval	2.98	44.48	5.78	5.32	8.74	4.36
LA21eval	7.53	41.57	26.30	17.98	19.29	24.39
DF21eval	6.67	24.26	11.95	11.54	9.71	13.31
LA19etrim	15.56	31.62	23.29	[illegible]	[illegible]	[illegible]
LA21hid	28.80	27.60	28.30	[illegible]	[illegible]	[illegible]
DF21hid	23.62	26.18	22.01	13.92	13.50	16.99
WaveFake	15.76	-	39.27	34.05	17.10	10.89
InWild	26.65	19.98	41.06	36.46	22.26	19.45
Pooled	14.24	-	36.57	39.95	19.39	16.35

Vocoders pre-trained by ESPNet (Hayashi 2020)

: (

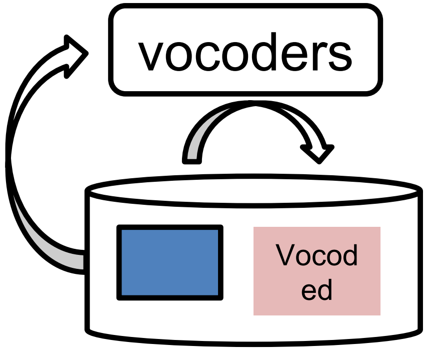
ID	#. Spr.	#. Bona.	#. Spoof.	Vocoder type	Implementation	Vocoder train/fine-tune data
LA19trn	20	2,580	22,800	-	-	-
WFtrn	1	3,930	15,720	HiFiGAN, MB-MelGAN, PWG, WaveGlow	ESPNet toolkit	LJSpeech / -
Voc.v1	20 same as LA19trn	2,580	10,320	HiFiGAN, MB-MelGAN, PWG, StyleMelGAN	ESPNet toolkit	LibriTTS / -
Voc.v2				HiFiGAN, NSF-HiFiGAN, Hn-NSF, WaveGlow	in-house	LibriTTS / -
Voc.v3				HiFiGAN, NSF-HiFiGAN, Hn-NSF, WaveGlow	in-house	LA19trn bona. / -
Voc.v4				HiFiGAN, NSF-HiFiGAN, Hn-NSF, WaveGlow	in-house	LibriTTS / LA19trn bona.

pretrained

trained from scratch

finetuned

		...g set					
		LA19 trn	WF trn	Voc. v1	Voc. v2	Voc. v3	Voc. v4
Test sets	LA19eval	2.98	44.48	5.78	5.32	8.74	4.36
	LA21eval	7.53	41.57	26.30	17.98	19.29	24.39
	DF21eval	6.67	24.26	11.95	11.54	9.71	13.31
	LA19etrim	15.56	31.62	23.29	16.16	14.99	9.52
	LA21hid	28.80	27.60	28.30	19.49	17.62	21.43
	DF21hid	23.62	26.18	22.01	13.92	13.50	16.99
	WaveFake	15.76	-	39.27	34.05	17.10	10.89
	InWild	26.65	19.98	41.06	36.46	22.26	19.45
	Pooled	14.24	-	36.57	39.95	19.39	16.35



Training set

ID	#. Spr.	#. Bona.	#. Spoof.	Vocoder type	Implementation	Vocoder train/fine-tune data
LA19trn	20	2,580	22,800	-	-	-
`WFtrn`	1	3,930	15,720	HiFiGAN, MB-MelGAN, PWG, WaveGlow	ESPNet toolkit	LJSpeech / -
`Voc.v1`	20 same as `LA19trn`	2,580	10,320	HiFiGAN, MB-MelGAN, PWG, StyleMelGAN	ESPNet toolkit	LibriTTS / -
`Voc.v2`				HiFiGAN, NSF-HiFiGAN, Hn-NSF, WaveGlow	in-house	LibriTTS / -
`Voc.v3`				HiFiGAN, NSF-HiFiGAN, Hn-NSF, WaveGlow	in-house	`LA19trn` bona. / -
`Voc.v4`				HiFiGAN, NSF-HiFiGAN, Hn-NSF, WaveGlow	in-house	LibriTTS / `LA19trn` bona.

Test sets	Training set: `LA19 trn`	`WF trn`	`Voc. v1`	`Voc. v2`	`Voc. v3`	`Voc. v4`
`LA19eval`	2.98	44.48	5.78	5.32	8.74	4.36
`LA21eval`	7.53	[illegible]	[illegible]	[illegible]	19.29	24.39
`DF21eval`	6.67	[illegible]	[illegible]	[illegible]	9.71	13.31
`LA19etrim`	15.56	[illegible]	[illegible]	[illegible]	14.99	9.52
`LA21hid`	28.80	27.60	28.30	19.49	17.62	21.43
`DF21hid`	23.62	[illegible]	[illegible]	[illegible]	13.50	16.99
`WaveFake`	15.76	-	[illegible]	[illegible]	17.10	10.89
`InWild`	26.65	19.98	41.06	36.46	22.26	19.45
`Pooled`	14.24	-	36.57	39.95	19.39	16.35

We cannot exploit non-speech length

Reasonablly good

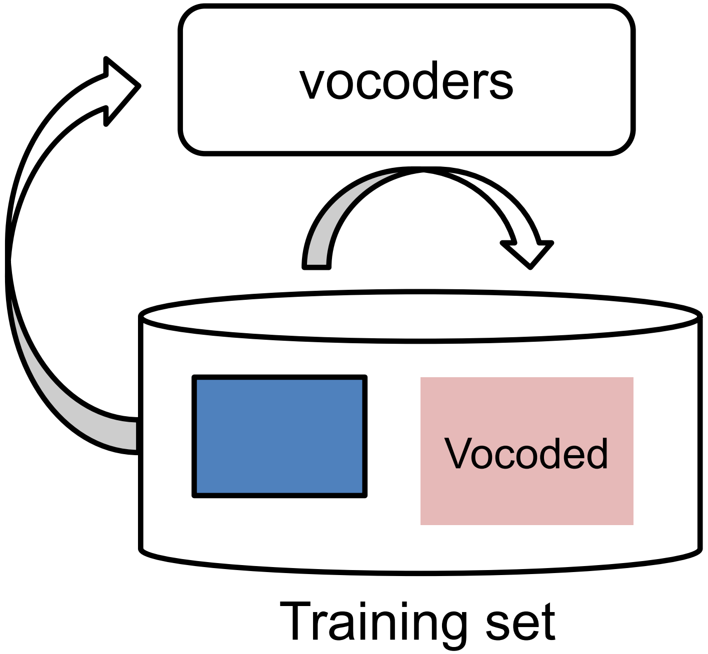
Vocoding data are not useless

❑ Which vocoders?

- practical choice – non-autogressive neural vocoders
- more analysis later

❑ How to train?

- expose vocoder to the data to be vocoded



Vocoding data are not useless

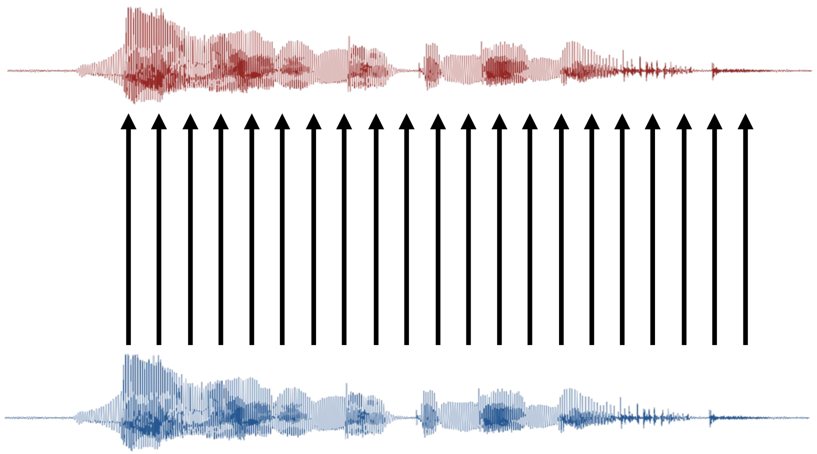
❑ CAUTION !

- Vocoder overlap with test sets?
  - partially in ASVspoof DF 2021
  - unknown in In-the-wild

- Other CMs performed poorly
  https://arxiv.org/abs/2210.10570

		LA19 trn	WF trn	Voc. v1	Voc. v2	Voc. v3	Voc. v4
LFCC−LCNN	LA19trn	0.10	41.69	14.25	40.72	22.83	28.08
	LA15eval	32.42	23.82	49.89	28.34	23.90	37.60
	LA19eval	3.32	49.80	23.52	40.79	37.71	29.68
	LA21eval	23.38	57.18	36.43	62.69	49.17	49.84
	DF21eval	29.45	47.40	40.27	52.76	47.22	51.76
	LA19etrim	21.18	40.10	33.23	41.97	40.94	40.56
	LA21hid	39.27	49.53	44.59	49.07	44.18	49.52
	DF21hid	35.86	45.12	43.05	48.52	45.18	49.81
	WaveFake	45.85	-	22.17	21.15	14.38	26.16
	InWild	72.19	91.28	84.33	45.93	61.38	28.60
	Pooled	37.68	-	43.90	51.16	46.26	51.35

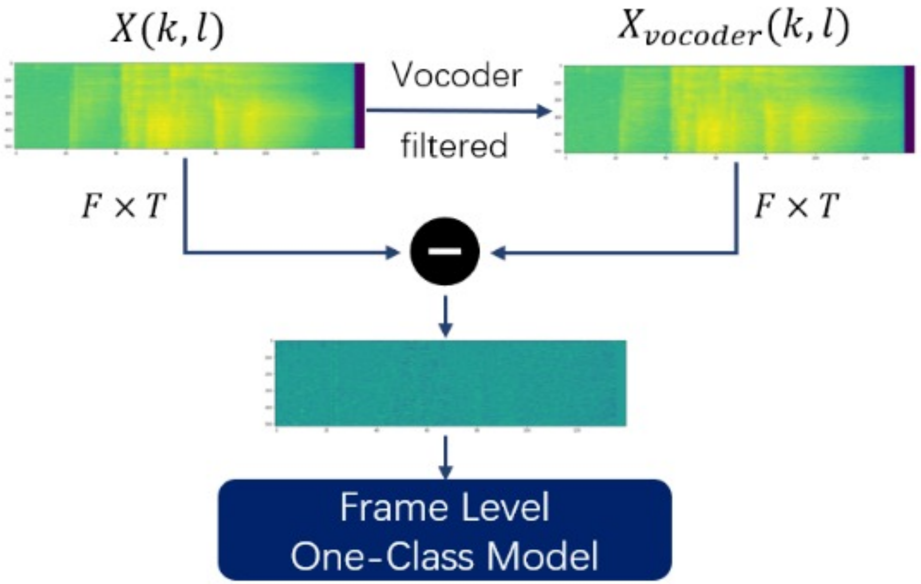
- How to exploit the aligned pair of {bona fide, vocoded spoofed}?



Contrastive learning

❑ Use contrastive features (Wang 2021)

- Vocoder is needed during testing

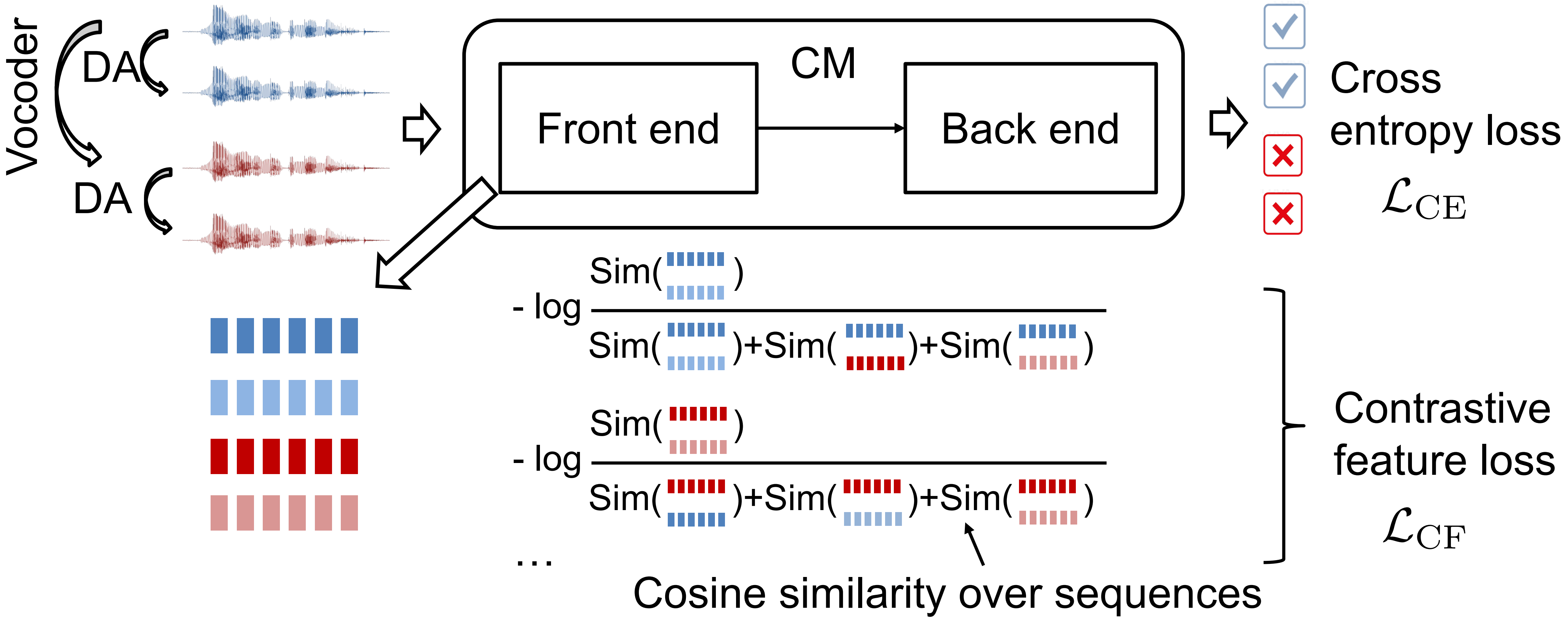


❑ Use contrastive learning

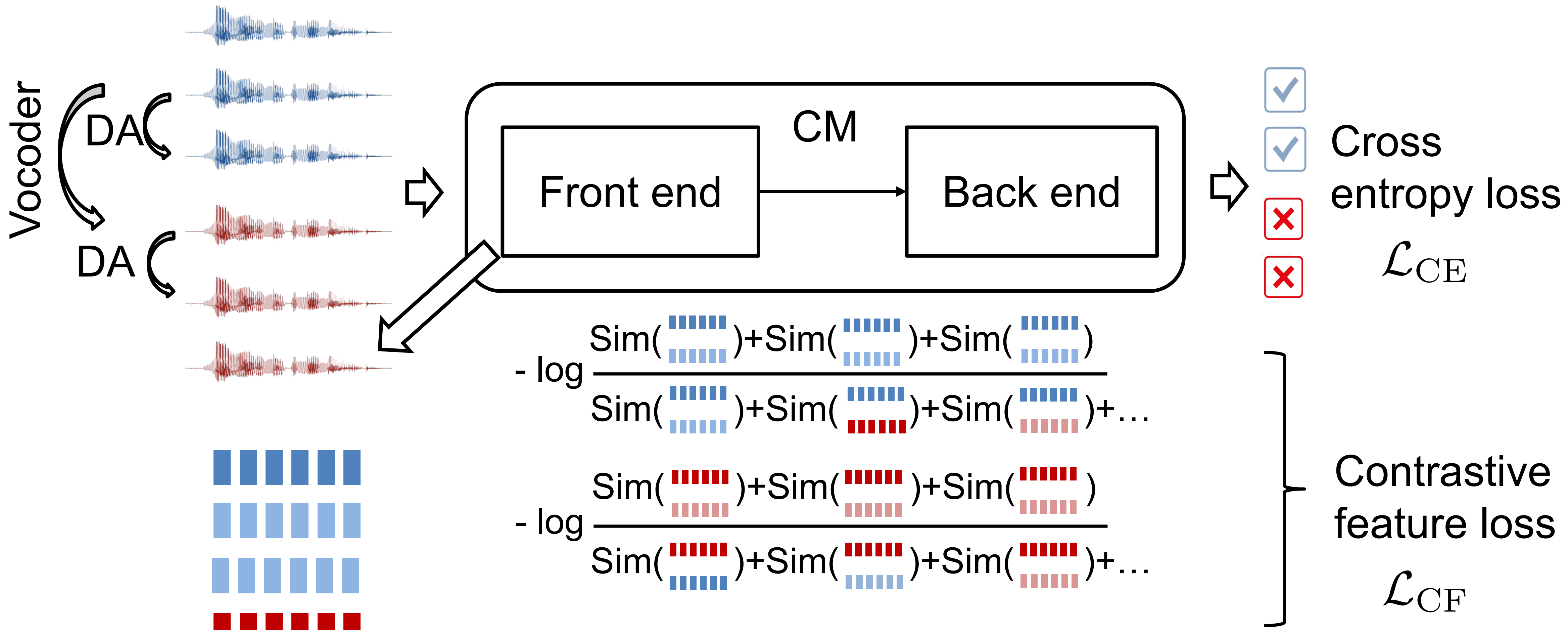
- Supervised contrastive loss (Khosla 2020)
- Vocoder is NOT needed during testing
- What is needed:
  - an additional CM training loss
  - data augmentation (DA) to create multi-view, e.g., RawBoost (Tak 2022)

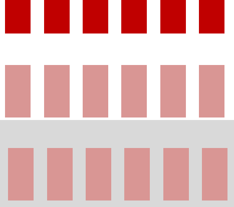
Xingming Wang, Xiaoyi Qin, Tinglong Zhu, Chao Wang, Shilei Zhang, and Ming Li. The DKU-CMRI System for the ASVspoof 2021 Challenge: Vocoder Based Replay Channel Response Estimation. In *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 16–21. doi:10.21437/ASVSPOOF.2021-3. 2021.
Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised Contrastive Learning. In *Proc. NIPS*, 18661–18673. 2020.
Hemlata Tak, Madhu R Kamble, Jose Patino, Massimiliano Todisco, and Nicholas W D Evans. RawBoost: A Raw Data Boosting and Augmentation Method Applied to Automatic Speaker Verification Anti-Spoofing. In *Proc. ICASSP*, 6382–6386. 2022.

Contrastive learning



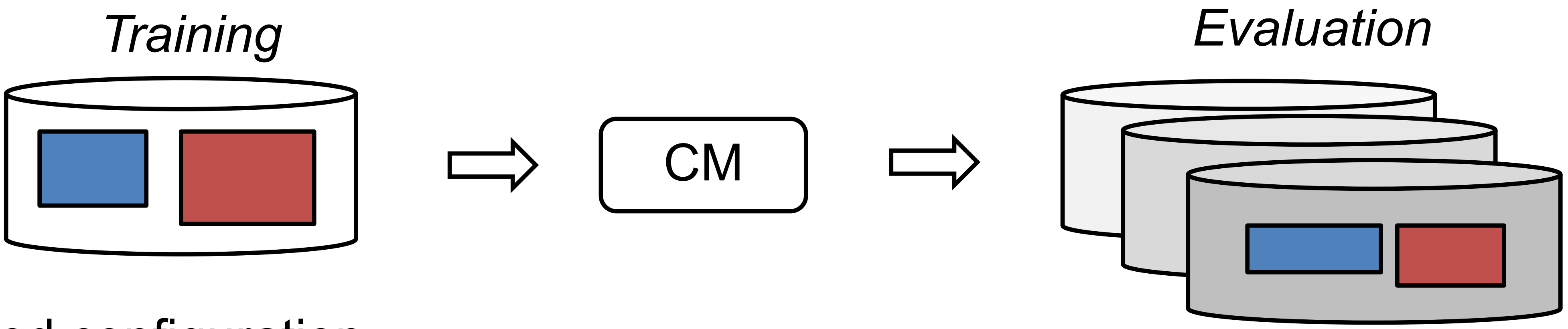
Contrastive learning





More views are helpful (appendix in https://arxiv.org/abs/2210.10570)

Experiment



- Fixed configuration
  - training data **Voc.v4** (ASVspoof 2019 LA trn bona fide + vocoded)
  - Wav2vec-based CM
  - multiple test sets

- Variable of interest: how CM is trained

Hemlata Tak, Madhu R Kamble, Jose Patino, Massimiliano Todisco, and Nicholas W D Evans. RawBoost: A Raw Data Boosting and Augmentation Method Applied to Automatic Speaker Verification Anti-Spoofing. In *Proc. ICASSP*, 6382–6386. 2022.

Experiment results

from Experiment I | control groups | best

+Contrastive loss

	Training criterion	$\mathcal{L}_{CE}$				$\mathcal{L}_{CE} + \mathcal{L}_{CF}$		
	Data augmentation	×		RawBoost		RawBoost		
	Training set	LA19 trn	Voc. v4	LA19 trn	Voc. v4	LA19 trn	Voc. v4	Voc. v4
	Bona-spoof paired	×	×	×	×	×	×	✓
	ID	①	②	③	④	⑤	⑥	⑦
Test sets	LA19eval	2.98	4.36	0.22	3.46	0.21	2.63	2.21
	LA21eval	7.53	24.39	3.63	16.55	3.30	16.67	17.90
	DF21eval	6.67	13.31	3.65	9.60	4.12	6.92	5.04
	LA19etrim	15.56	9.52	9.16	6.09	9.00	4.48	3.79
	LA21hid	28.80	21.43	21.18	19.37	26.98	15.05	14.57
	DF21hid	23.62	16.99	13.64	14.29	16.85	8.17	7.78
	WaveFake	15.76	10.89	26.37	6.87	24.62	4.03	2.50
	InWild	26.65	19.45	16.17	12.08	17.07	9.37	7.55
	Pooled	14.24	16.35	13.12	13.13	13.68	13.15	11.27

Experiment results

from Experiment I | control groups | best

Training criterion	$\mathcal{L}_{\mathrm{CE}}$				$\mathcal{L}_{\mathrm{CE}} + \mathcal{L}_{\mathrm{CF}}$		
Data augmentation	×		RawBoost		RawBoost		
Training set	LA19 trn	Voc. v4	LA19 trn	Voc. v4	LA19 trn	Voc. v4	Voc. v4
Bona-spoof paired	×	×	×	×	×	×	✓
ID	①	②	③	④	⑤	⑥	⑦
Test sets: LA19eval	2.98	4.36	0.22	3.46	0.21	2.63	2.21
LA21eval	7.53	24.39	3.63	16.55	3.30	16.67	17.90
DF21eval	6.67	13.31	3.65	9.60	4.12	6.92	5.04
LA19etrim	15.56	9.52	9.16	6.09	9.00	4.48	3.79
LA21hid	28.80	21.43	21.18	19.37	26.98	15.05	14.57
DF21hid	23.62	16.99	13.64	14.29	16.85	8.17	7.78
WaveFake	15.76	10.89	26.37	6.87	24.62	4.03	2.50
InWild	26.65	19.45	16.17	12.08	17.07	9.37	7.55
Pooled	14.24	16.35	13.12	13.13	13.68	13.15	11.27

best

② vs ④
RawBoost is useful

④ vs ⑦
Contrastive feature loss is useful

Experiment results

from Experiment I | control groups | best

Training criterion	$\mathcal{L}_{\text{CE}}$				$\mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{CF}}$		
Data augmentation	×		RawBoost		RawBoost		
Training set	LA19 trn	Voc. v4	LA19 trn	Voc. v4	LA19 trn	Voc. v4	Voc. v4
Bona-spoof paired	×	×	×	×	×	×	✓
ID	①	②	③	④	⑤	⑥	⑦
Test sets: LA19eval	2.98	4.36	0.22	3.46	0.21	2.63	2.21
LA21eval	7.53	24.39	3.63	16.55	3.30	16.67	17.90
DF21eval	6.67	13.31	3.65	9.60	4.12	6.92	5.04
LA19etrim	15.56	9.52	9.16	6.09	9.00	4.48	3.79
LA21hid	28.80	21.43	21.18	19.37	26.98	15.05	14.57
DF21hid	23.62	16.99	13.64	14.29	16.85	8.17	7.78
WaveFake	15.76	10.89	26.37	6.87	24.62	4.03	2.50
InWild	26.65	19.45	16.17	12.08	17.07	9.37	7.55
Pooled	14.24	16.35	13.12	13.13	13.68	13.15	11.27

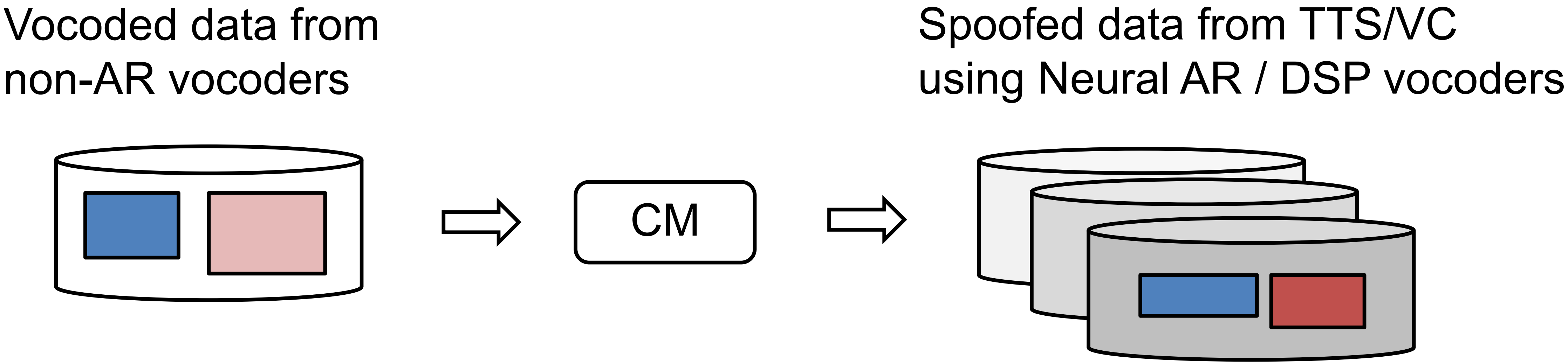
vocoded data
+ contrastive learning is helpful

Analysis

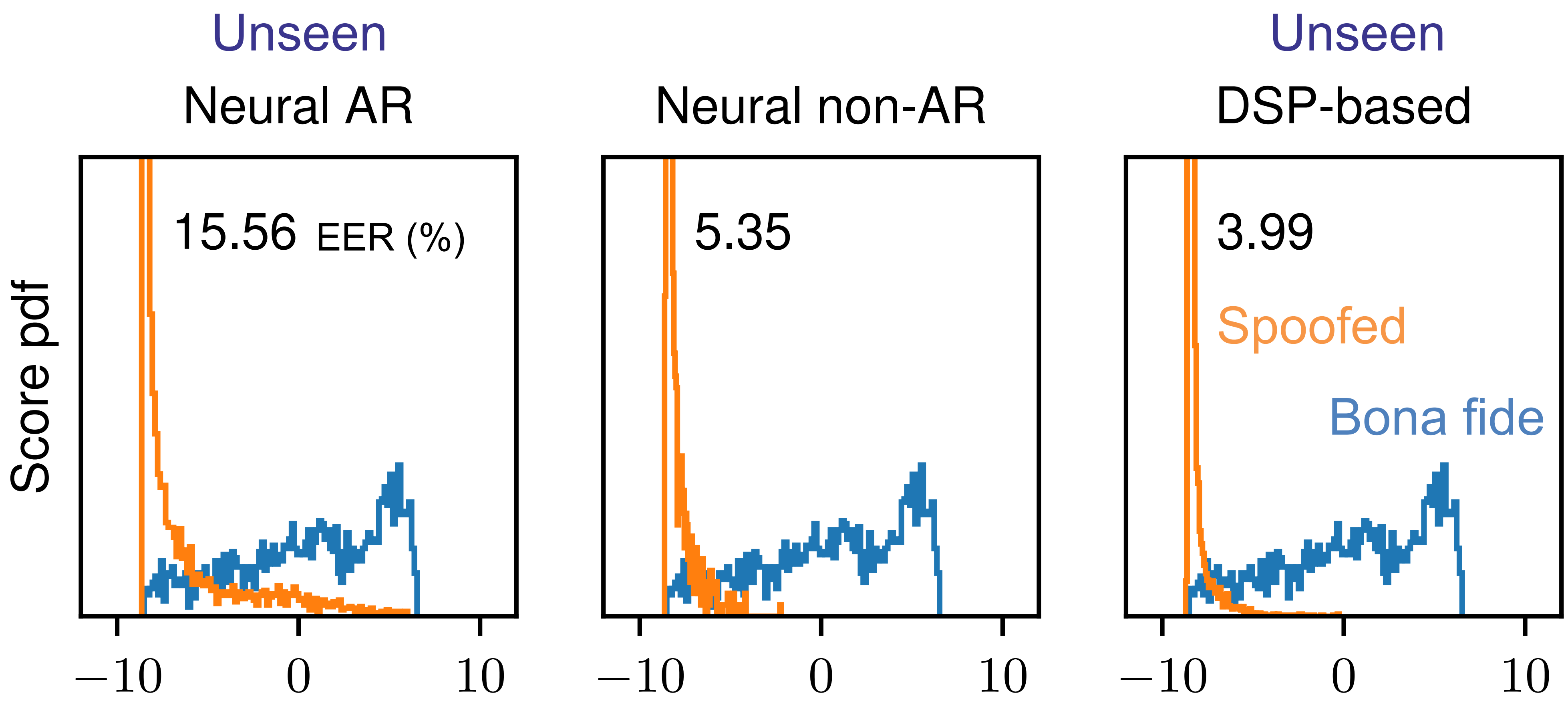
Can we detect TTS/VC w/ unseen vocoders?

❑ Options

- ~~DSP vocoders~~
- ~~Neural autoregressive (AR) vocoders~~
- Non-autoregressive DNN+DSP

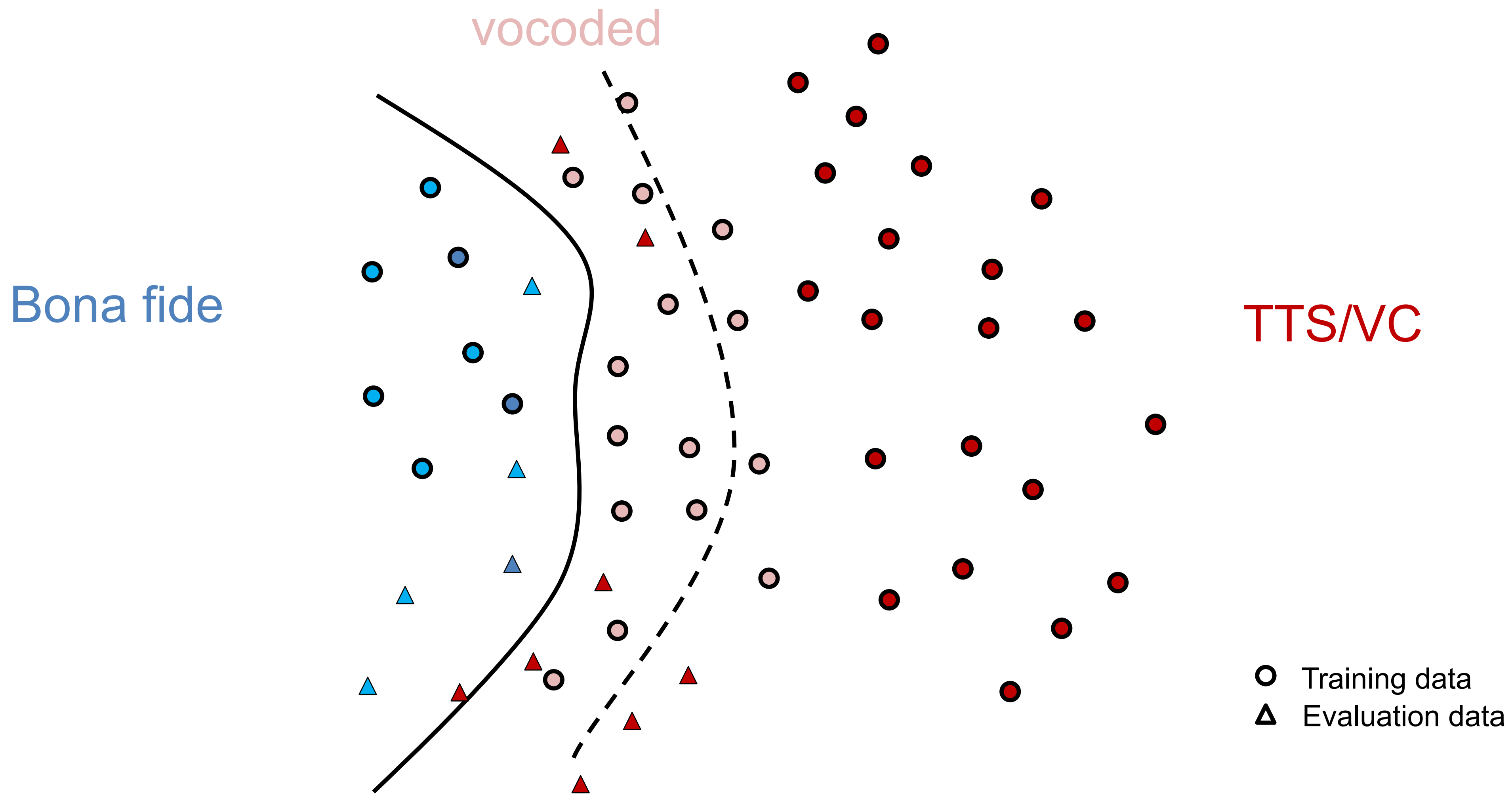


Can we detect TTS/VC w/ unseen vocoders?



Results on ASVspoof 2021 DF test set

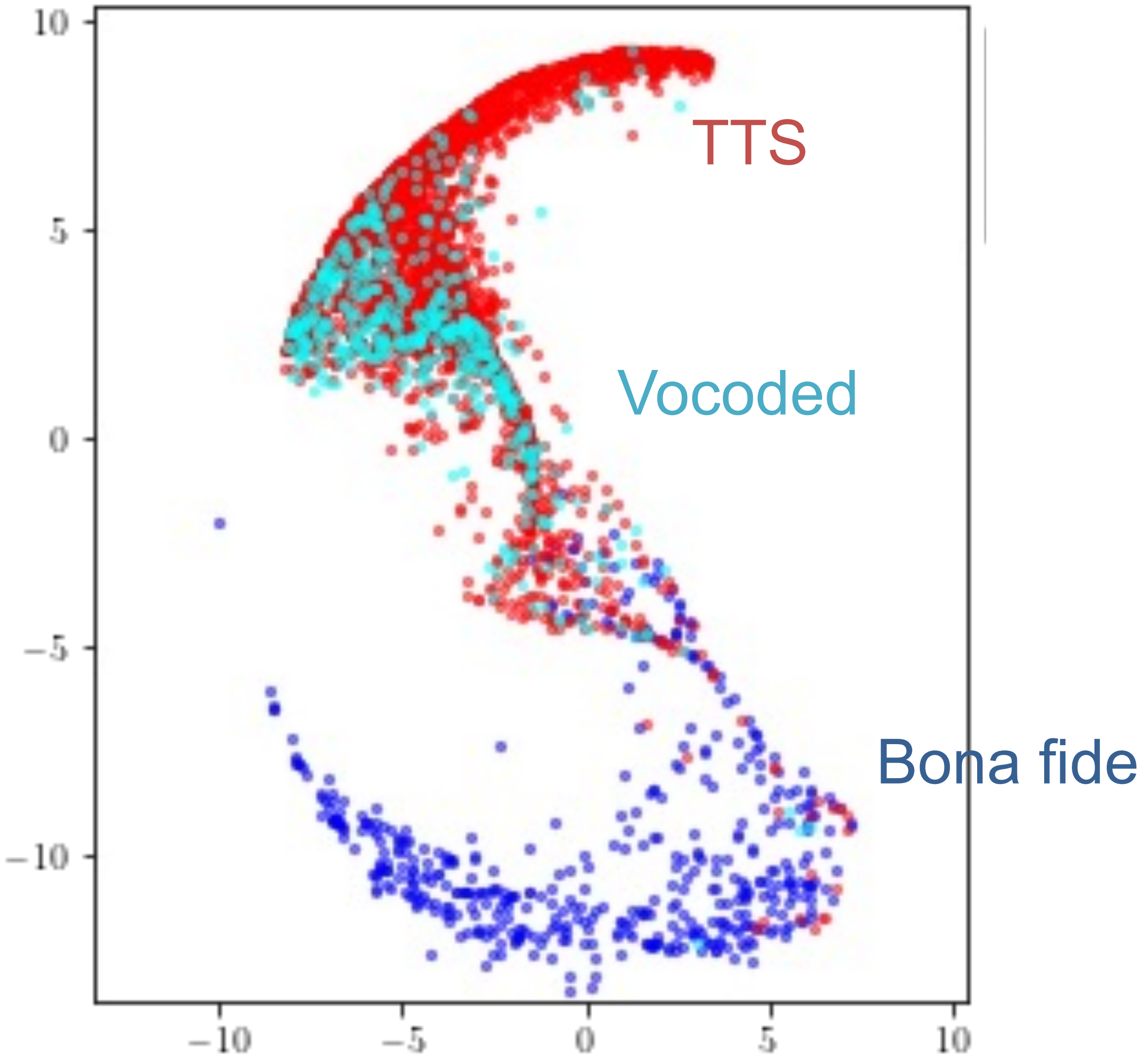
Is vocoded data closer to bona fide data?



Is vocoded data closer to bona fide data?

- Yes

  - bonafide: LJspeech data
  - vocoded: full-band MelGAN
  - TTS: full-band MelGAN + Fastspeech2 / Tacotron2



NII

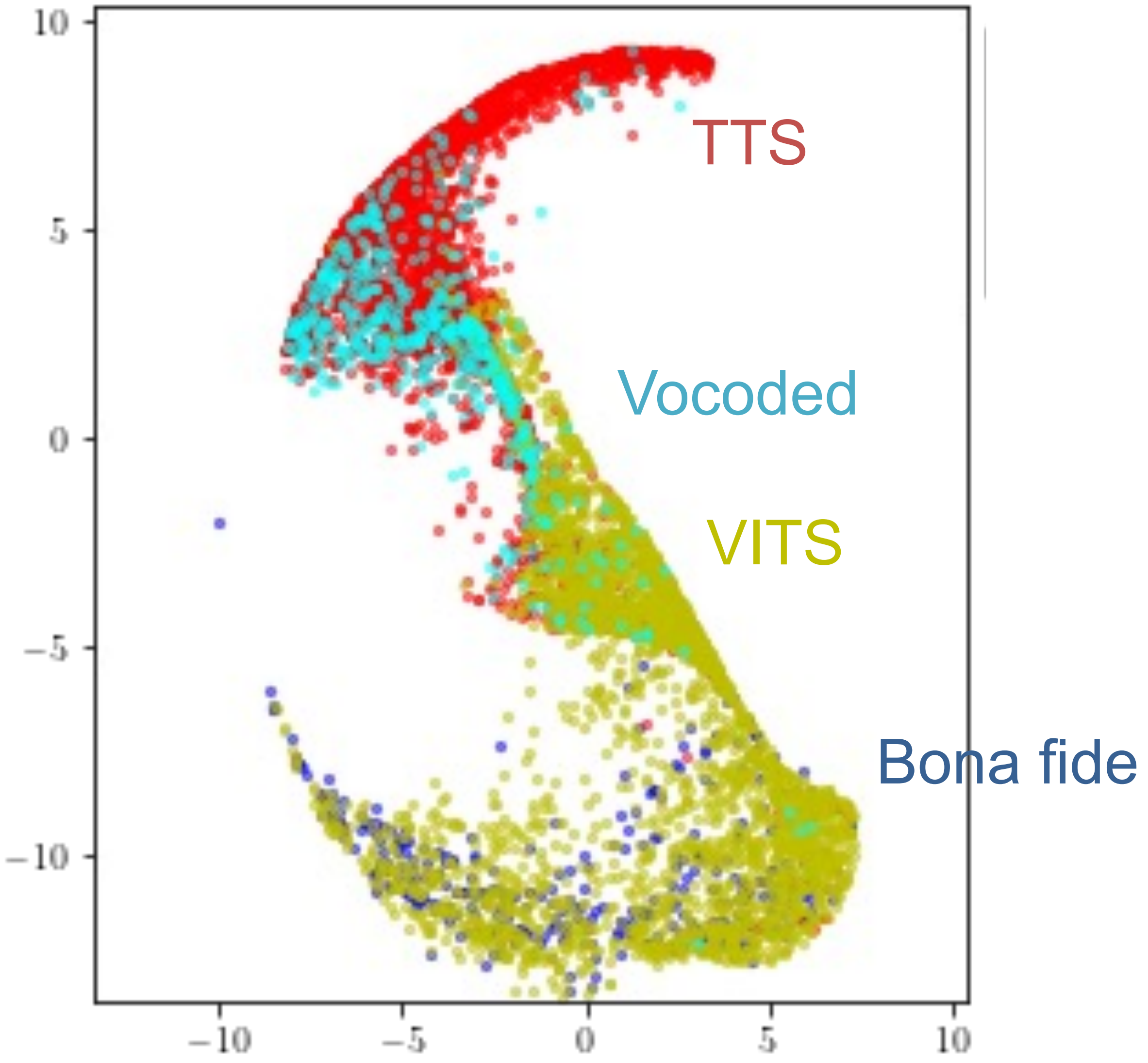
Is vocoded data closer to bona fide data?

❑ **Yes**

- bonafide: LJspeech data
- vocoded: full-band MelGAN
- TTS: full-band MelGAN + Fastspeech2 / Tacotron2

❑ **but not always**

- end-to-end TTS: VITS



Limitations

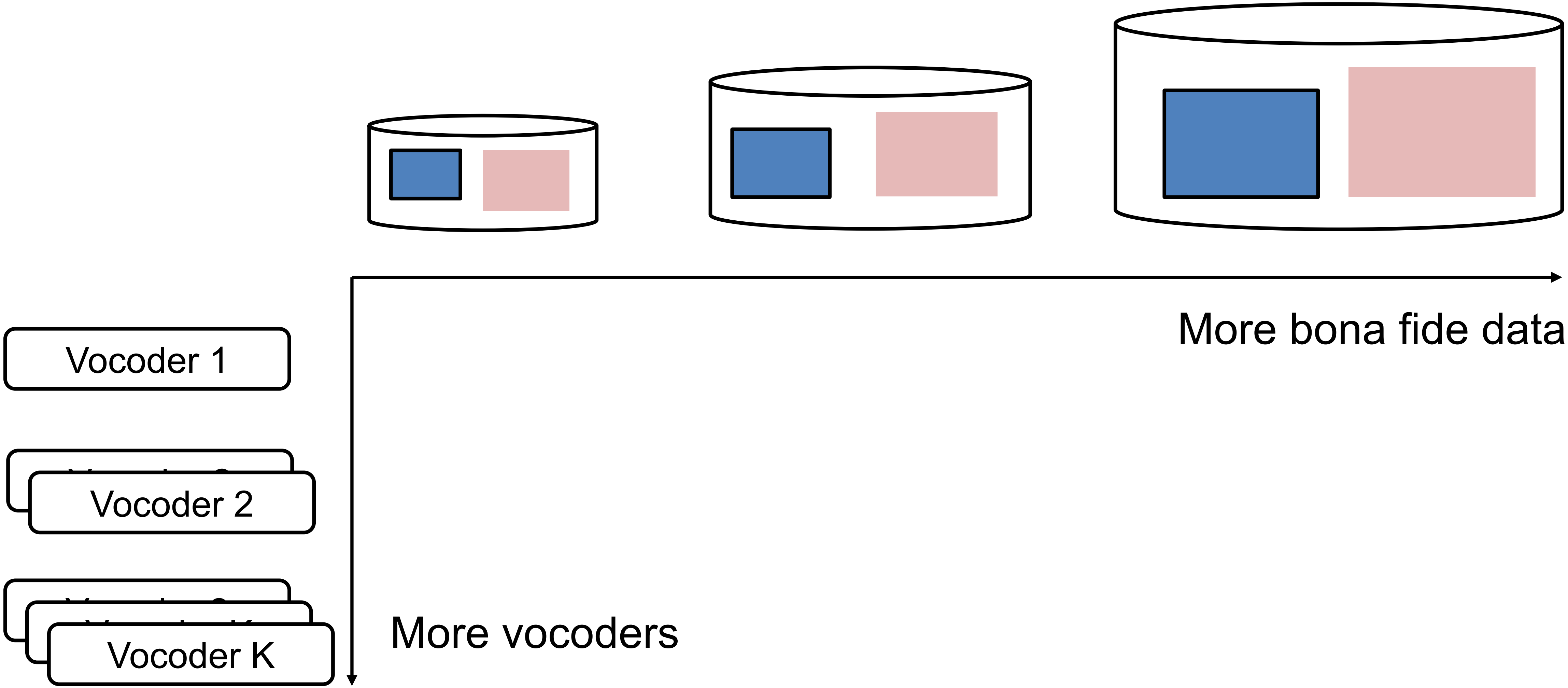
- **Generalization?**
  - ✓ DSP-based vocoders
  - ? Neural AR vocoders

- **Generalization to end-to-end TTS?**
  - X VITS

Question 3

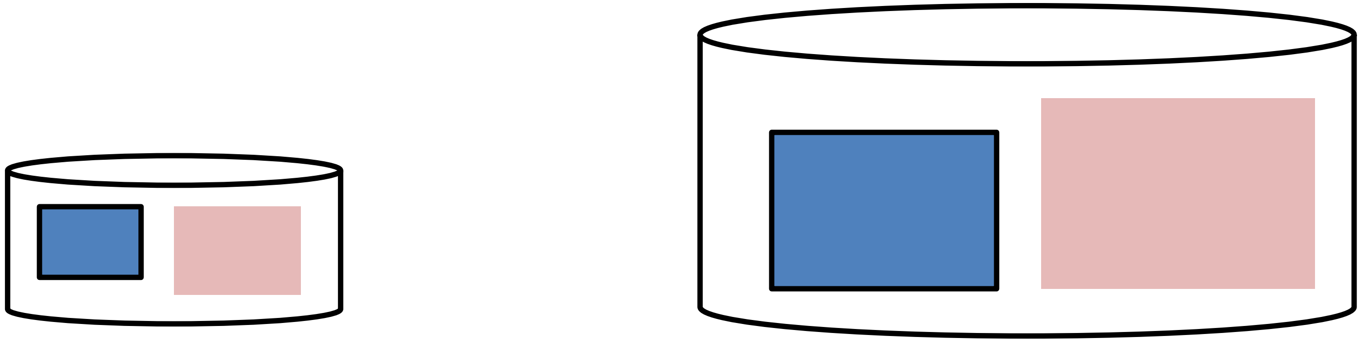
- Benefit of large-scale vocoded data?

Get much more data



Get much more data

❑ Vocoding VoxCeleb2

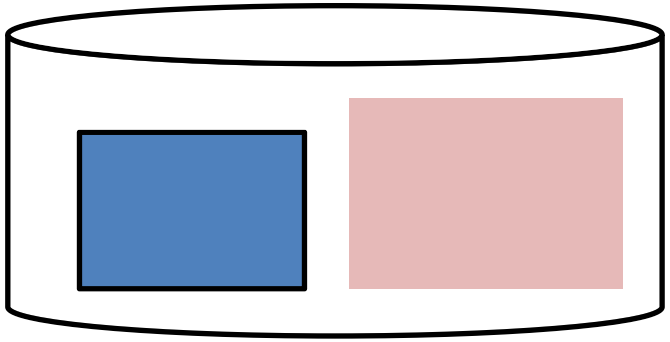


	ASVspoof 2019 LA	VoxCeleb2 dev
Bona fide:	2.42 hours	2360 hours
Vocoded:	2.42 x 4	2360 x 4

Get much more data using VoxCeleb2

❑ Usage 1: train CM as usual

- Practical issue – data size is too large
  - random sampling data to train CM



VoxCeleb2 dev

Get much more data using VoxCeleb2

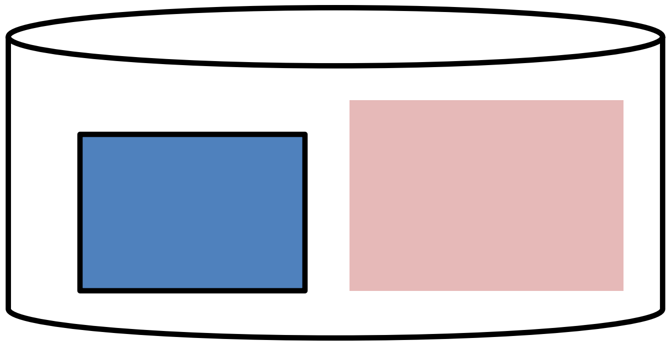
the best in Experiment II

	LA19 trn.	Voc.v4	Vocoded VoxCeleb2 dev					
Training data size	-	$\times 1$	$\times 0.3$	$\times 0.6$	$\times 1.2$	$\times 2.4$	$\times 6.0$	$\times 12.0$
LA19eval	0.21	2.21	8.20	7.04	5.40	6.53	5.40	5.63
LA21eval	3.30	17.90	20.19	16.73	14.33	18.10	17.44	17.84
DF21eval	4.12	5.04	7.49	5.41	5.39	6.00	5.65	5.64
LA19etrim	9.00	3.79	6.17	5.53	5.16	5.25	5.22	5.14
LA21hid	26.98	14.57	13.98	12.34	11.47	11.64	11.37	11.43
DF21hid	16.85	7.78	11.02	9.71	9.90	10.05	9.99	10.04
WaveFake	24.62	2.50	14.94	10.39	8.38	5.52	4.88	4.94
InWild	17.07	7.55	16.12	15.63	14.19	13.43	13.32	13.77
Pooled	13.68	11.27	13.52	11.79	9.98	9.01	8.36	8.37

Get much more data using VoxCeleb2

❑ Usage 1: train CM as usual

- Practical issue – data size is too large
  - random sampling data to train CM



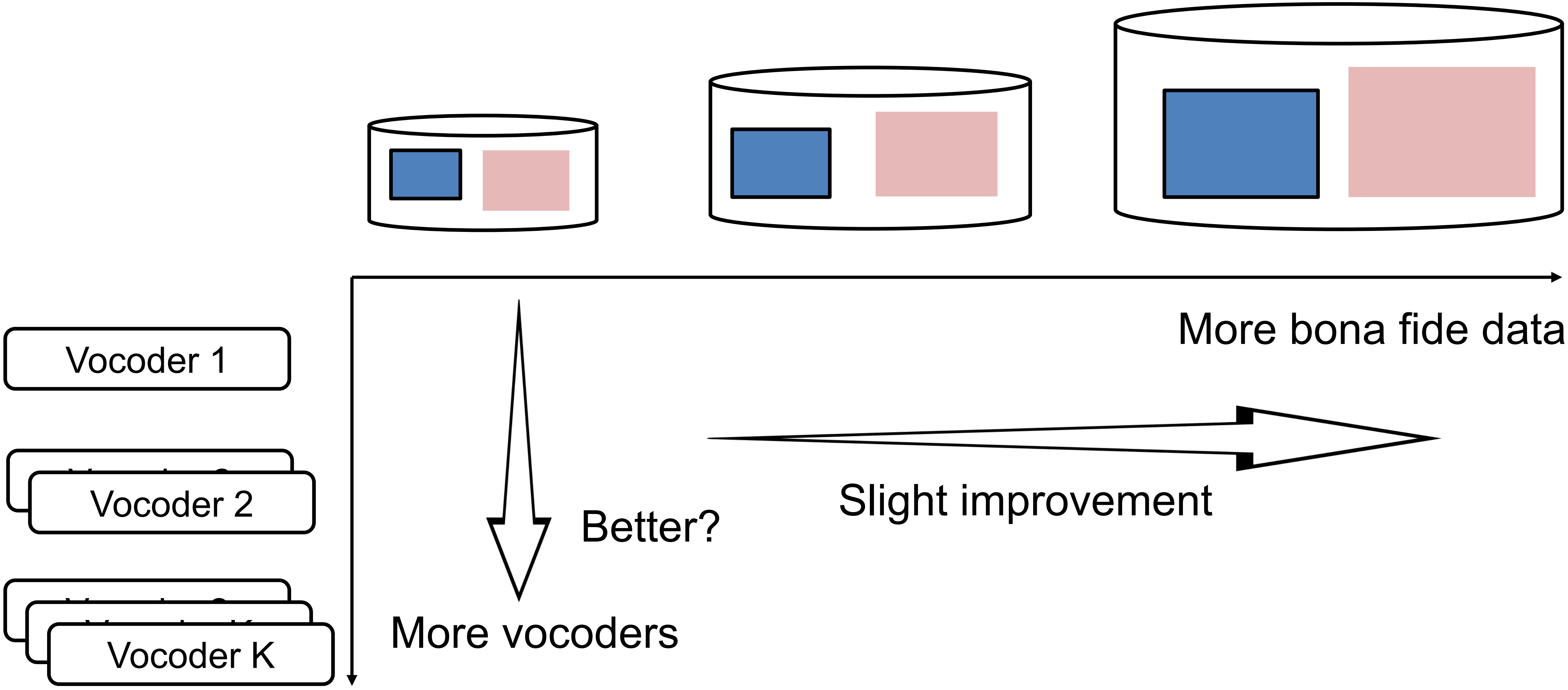
VoxCeleb2 dev

❑ Usage 2: train CM feature extractor

- wav2vec2.0 ..
- …

Limited improvement
https://arxiv.org/abs/2309.06014

Get much more data



Summary

Vocoding

Contrastive learning