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

SPSC Webinar
2023/03/06

Using vocoders to create training data for speech spoofing countermeasures

Xin Wang, Junichi Yamagishi
National Institute of Informatics

wangxin@nii.ac.jp

Slides by Xin Wang
National Institute of Informatics

Contents

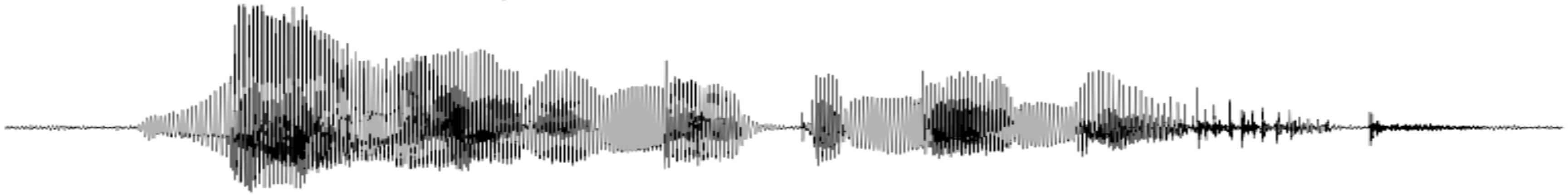
https://arxiv.org/abs/2210.10570

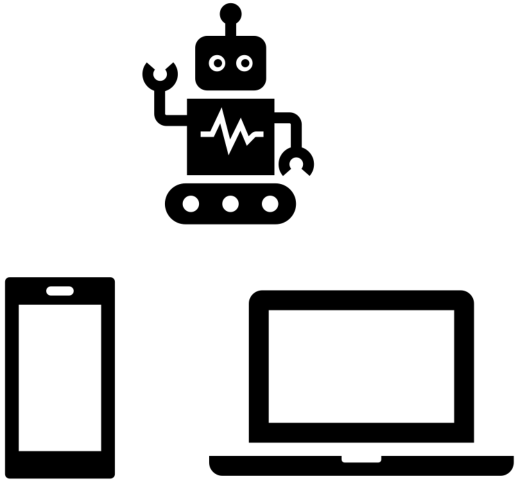
- **Introduction**
  - Text-to-speech synthesis
  - Speech anti-spoofing
- **Method**
  - Copy-synthesized data as spoofed data
  - Contrastive feature loss over bona fide and copy-synthesized data pair
- **Experiment**
- **Summary**

- We use external training data. Please be careful when interpreting the results.
- We take a data-driven approach. Apolgoize that we cannot precisely explain the model behavior.

Introduction

Text-to-speech synthesis

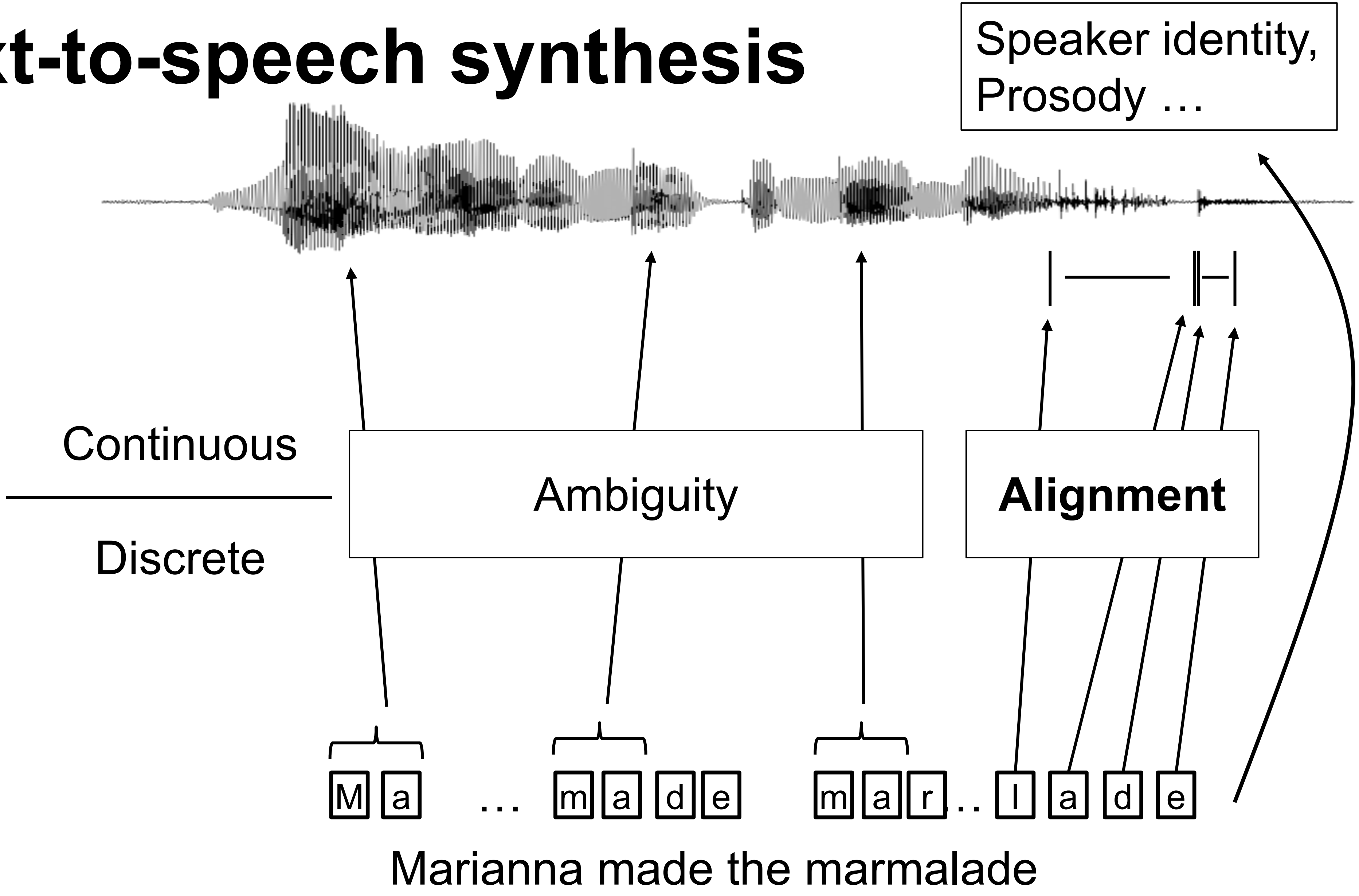




Input text

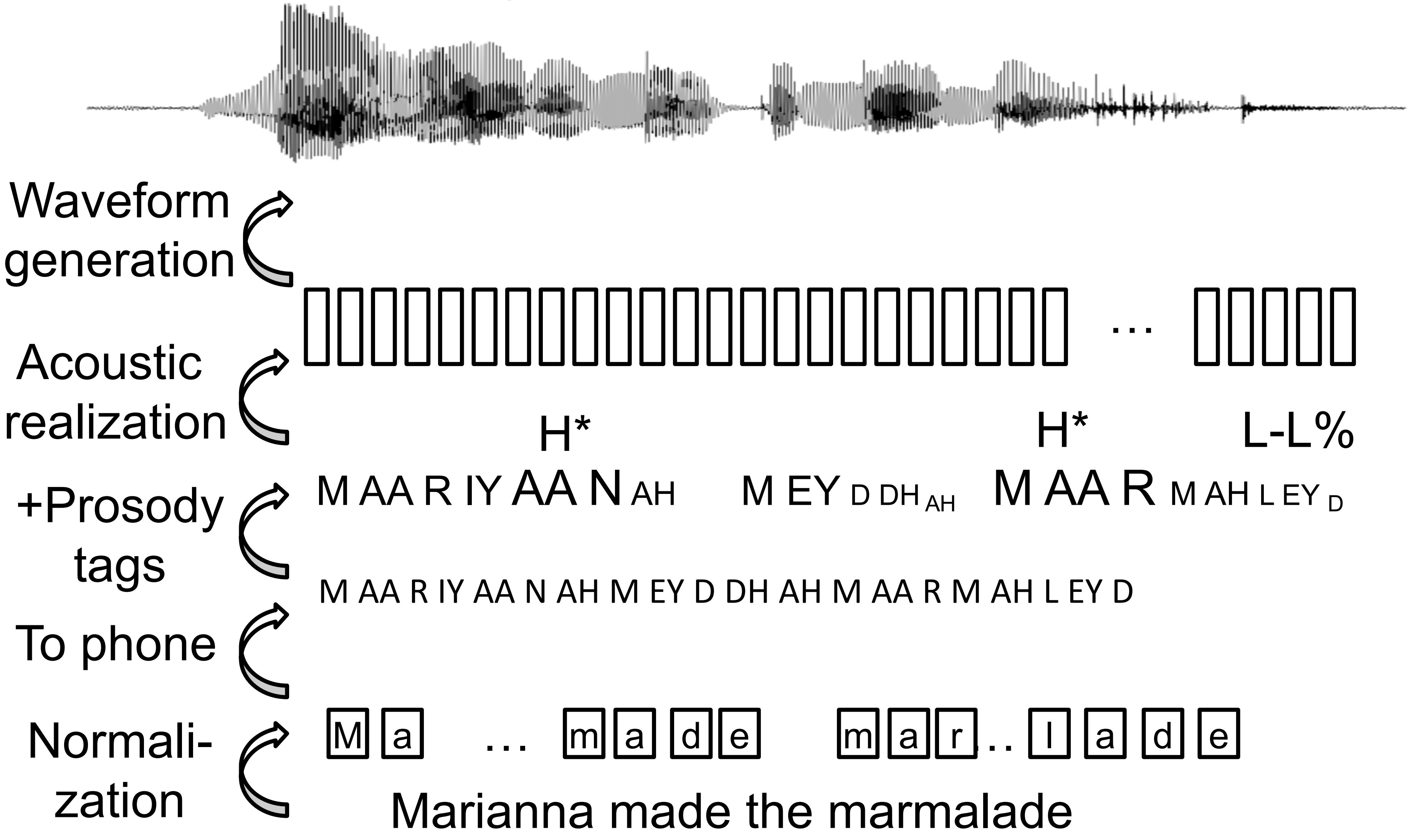
Kewley-Port, D. & M. Nearey, T. Speech synthesizer produced voices for disabled, including Stephen Hawking. J. Acoust. Soc. Am. 148, R1--R2 (2020)

Text-to-speech synthesis



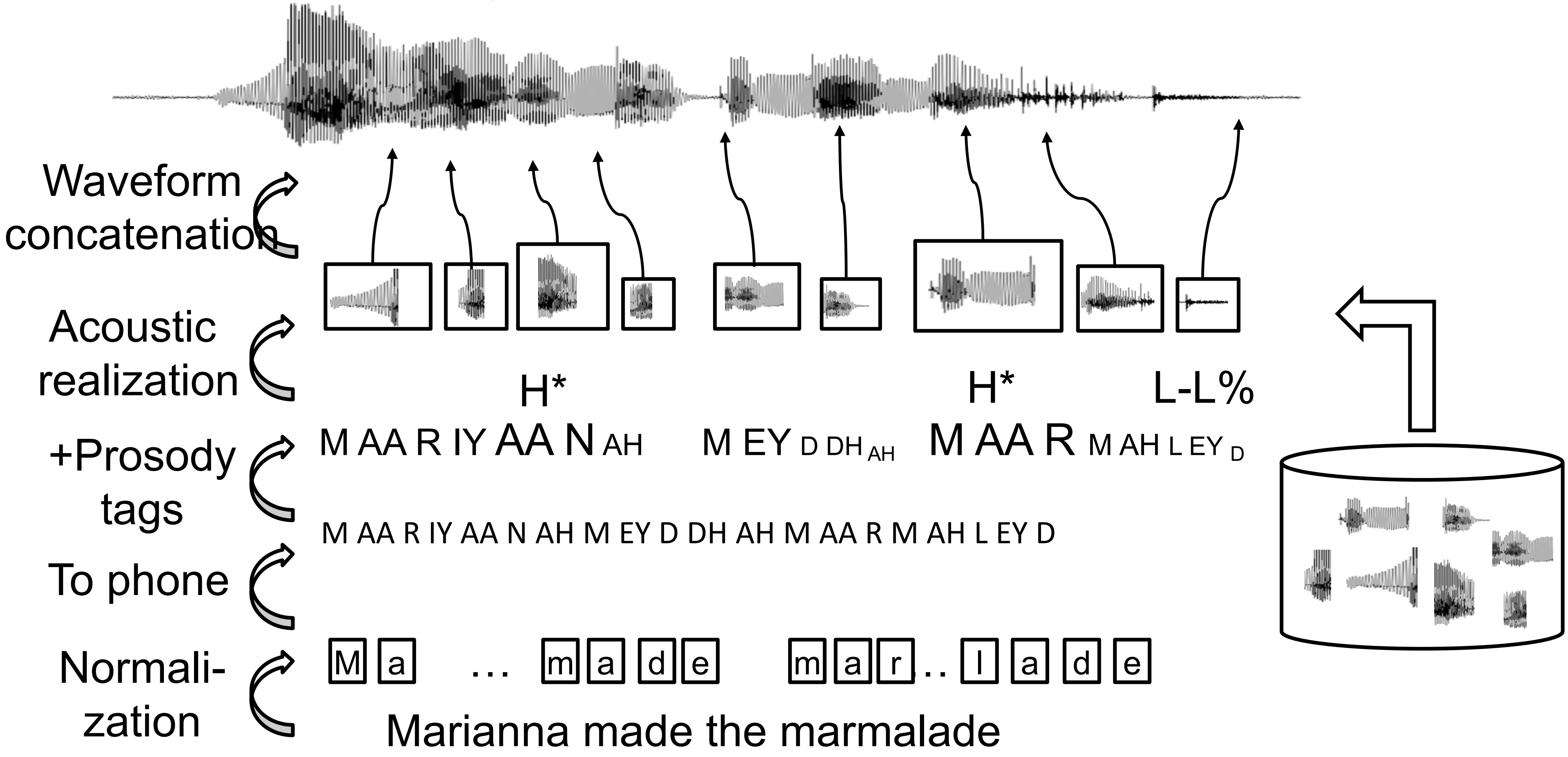
Sentence from: Beckman, M. E. & Ayers, G. Guidelines for ToBI labelling. OSU Res. Found. 3, (1997)

Text-to-speech synthesis



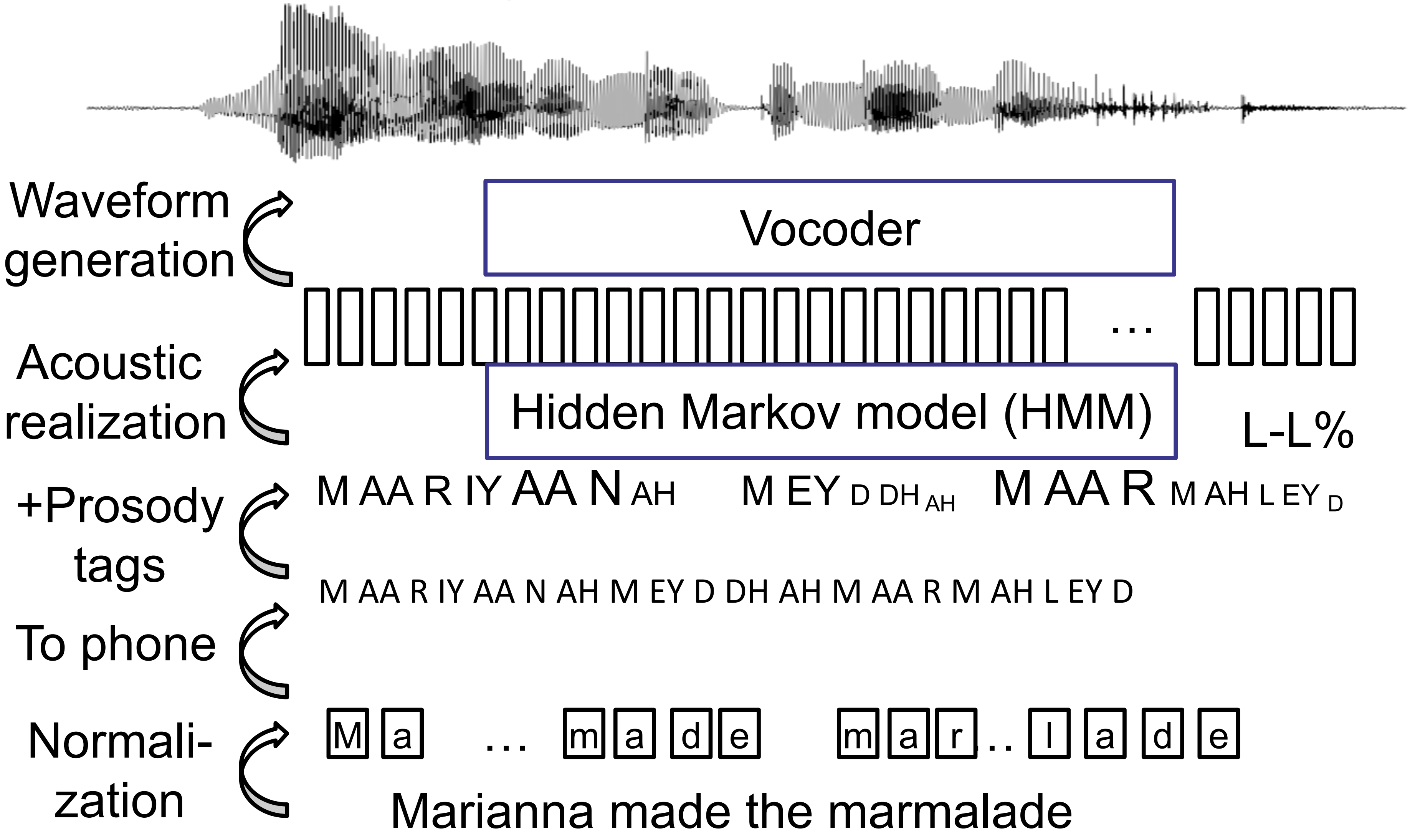
Sentence from: Beckman, M. E. & Ayers, G. Guidelines for ToBI labelling. OSU Res. Found. 3, (1997)
LOGIOS Lexicon tool: http://www.speech.cs.cmu.edu/tools/lextool.html
H*, L-L%: ToBI labels Beckman, M. E. & Ayers, G. Guidelines for ToBI labelling. OSU Res. Found. 3, (1997)

Text-to-speech synthesis – Unit selection



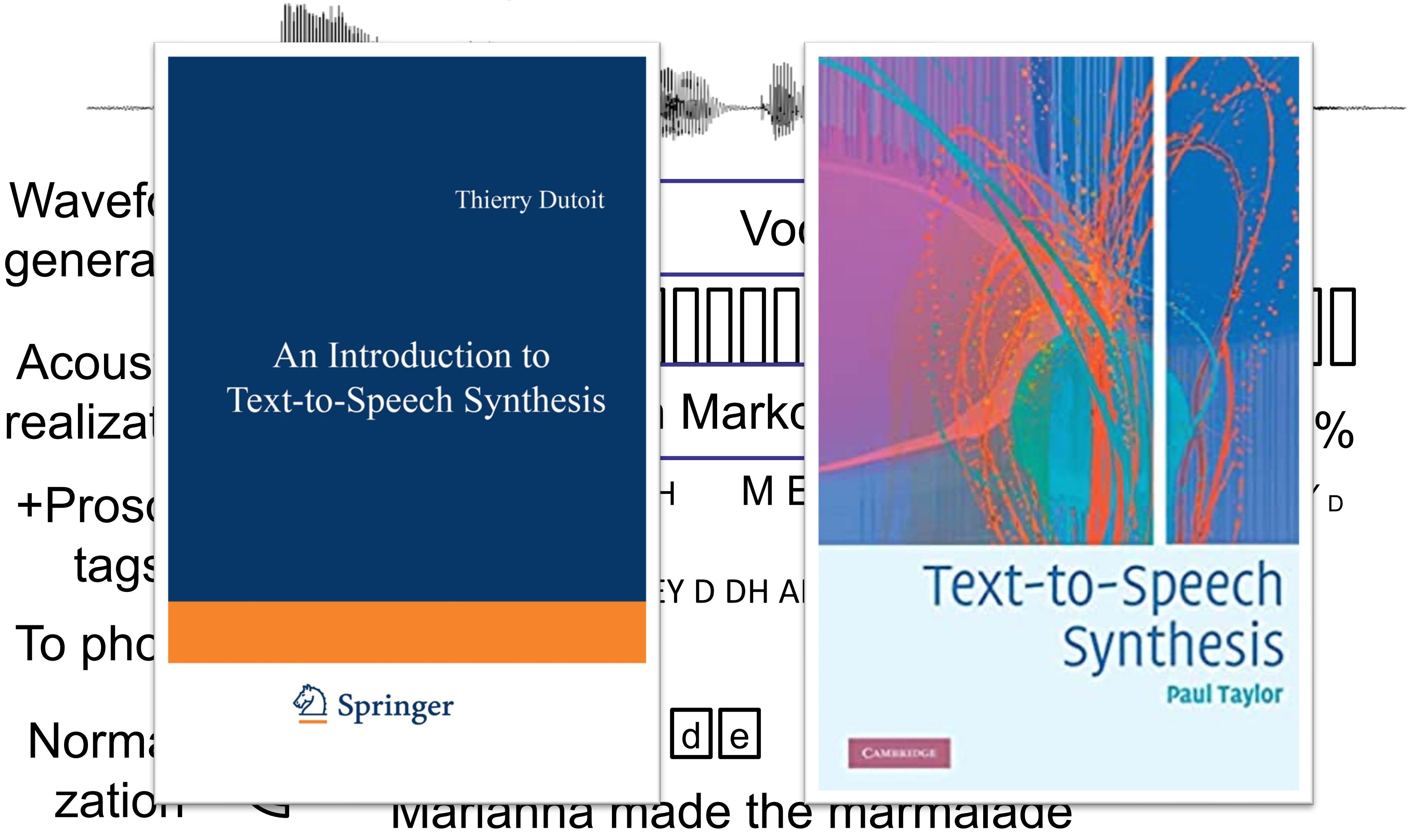
Hunt, A. J. & Black, A. W. Unit selection in a concatenative speech synthesis system using a large speech database. in Proc. ICASSP 373–376 (1996).
Black, A. W. & Taylor, P. A. Automatically clustering similar units for unit selection in speech synthesis. (1997)

Text-to-speech synthesis – HMM



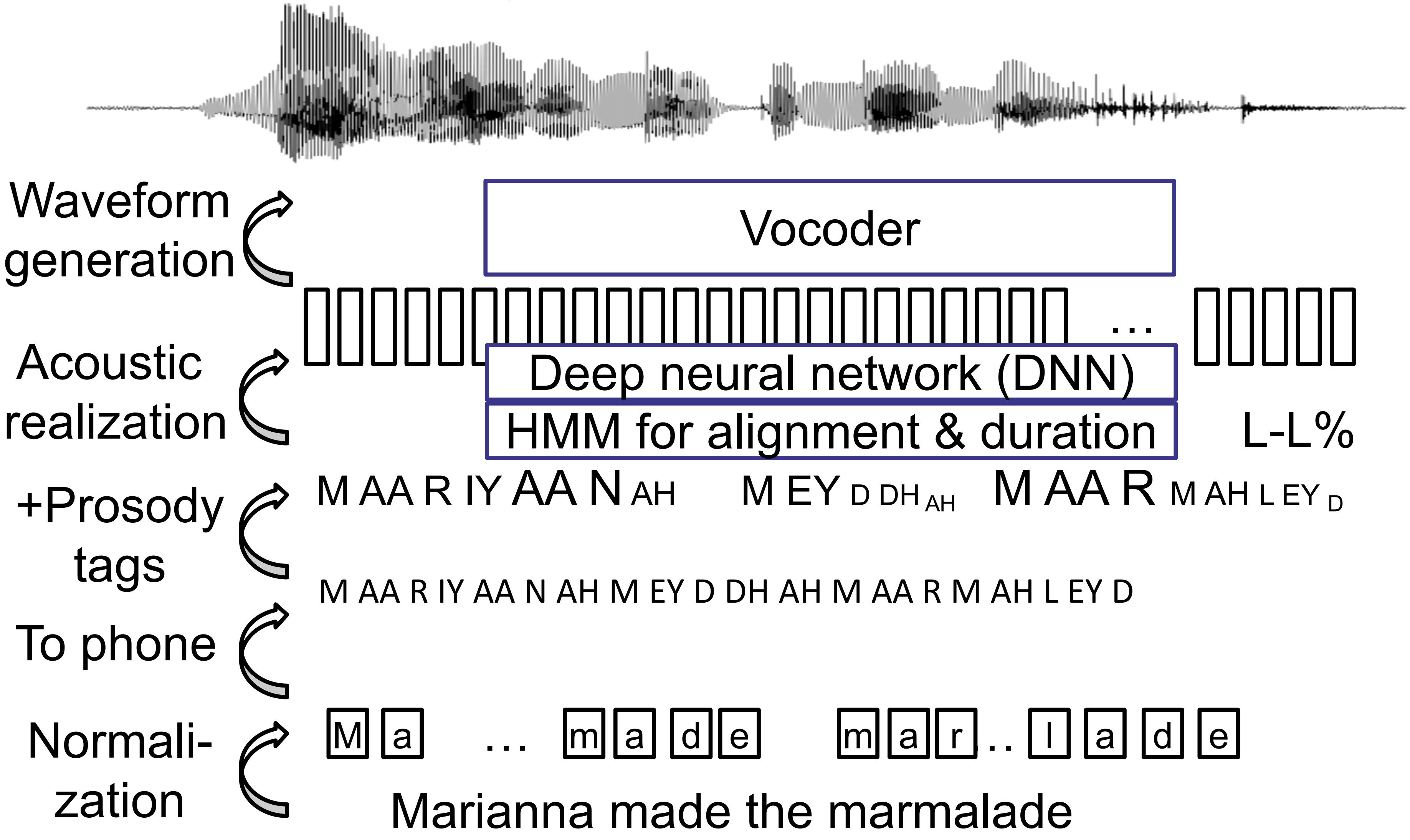
Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T. & Kitamura, T. Speech parameter generation algorithms for HMM-based speech synthesis. in Proc. ICASSP 936–939 (2000).
Keiichi Tokuda, Heiga Zen, and Alan W Black. An HMM-Based Speech Synthesis System Applied to English. In *Proc. SSW*, 227–230. 2002.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

Text-to-speech synthesis



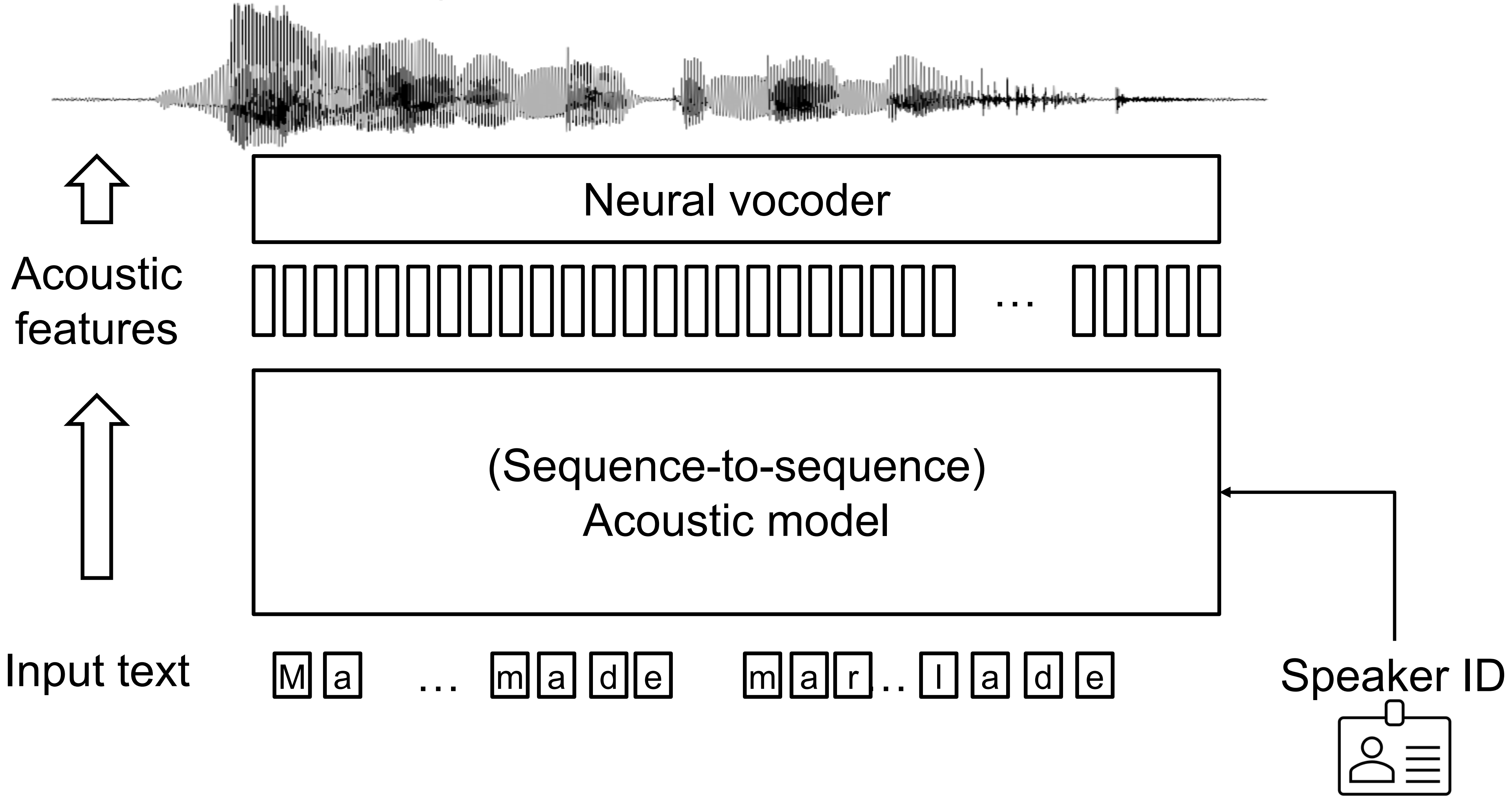
Thierry Dutoit. *An Introduction to Text-to-Speech Synthesis*. Norwell, MA, USA: Kluwer Academic Publishers. 1997.
Paul Taylor. *Text-to-Speech Synthesis*. Cambridge University Press. 2009.

Text-to-speech synthesis – HMM&DNN



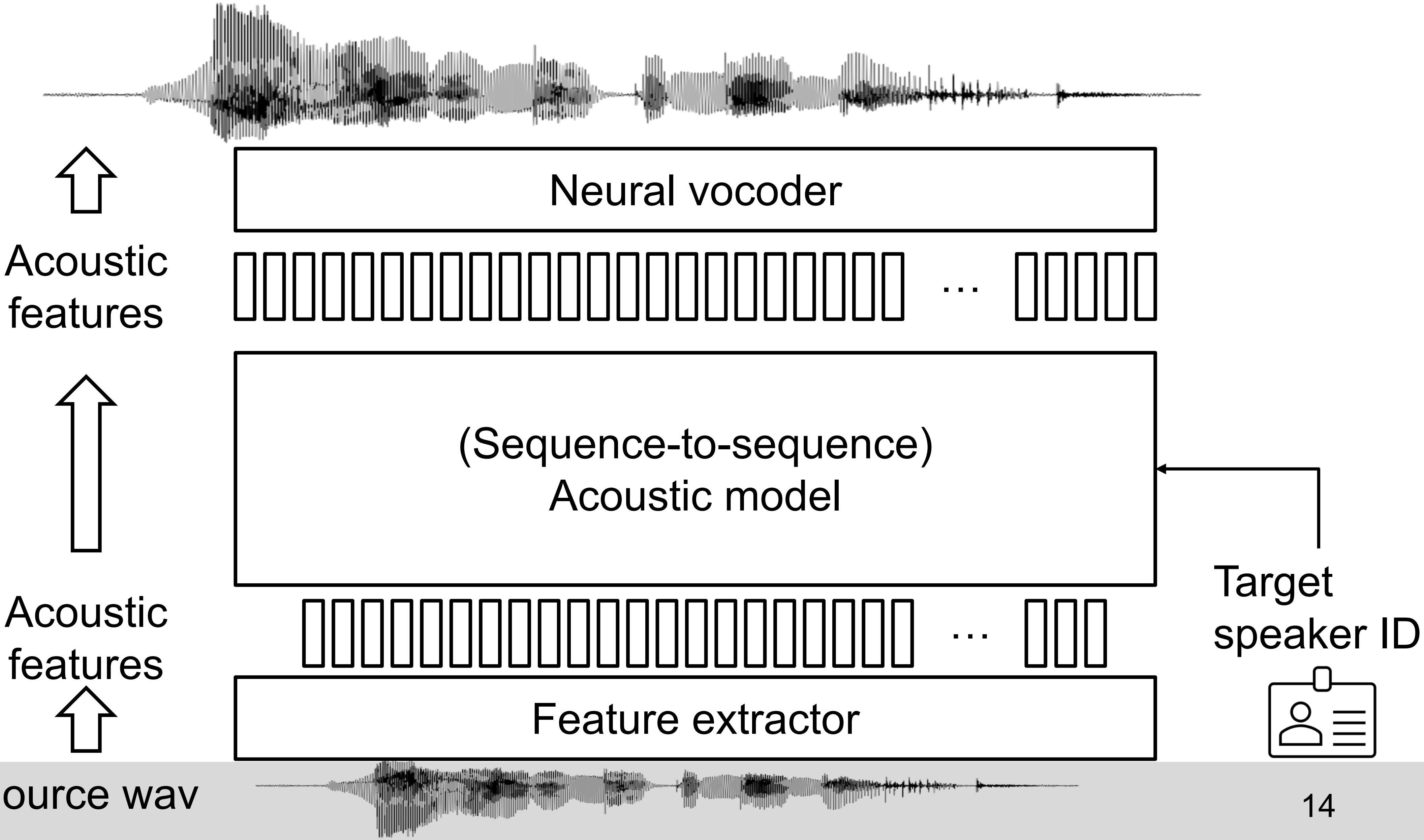
Heiga Zen, Alan Senior, and Martin Schuster. Statistical Parametric Speech Synthesis Using Deep Neural Networks. In *Proc. ICASSP*, 7962–7966. 2013.
Yuchen Fan, Yap Qian, Feilong Xie, and Frank K Soong. TTS Synthesis with Bidirectional LSTM Based Recurrent Neural Networks. In *Proc. Interspeech*, 1964–1968.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

Text-to-speech synthesis – Recent methods

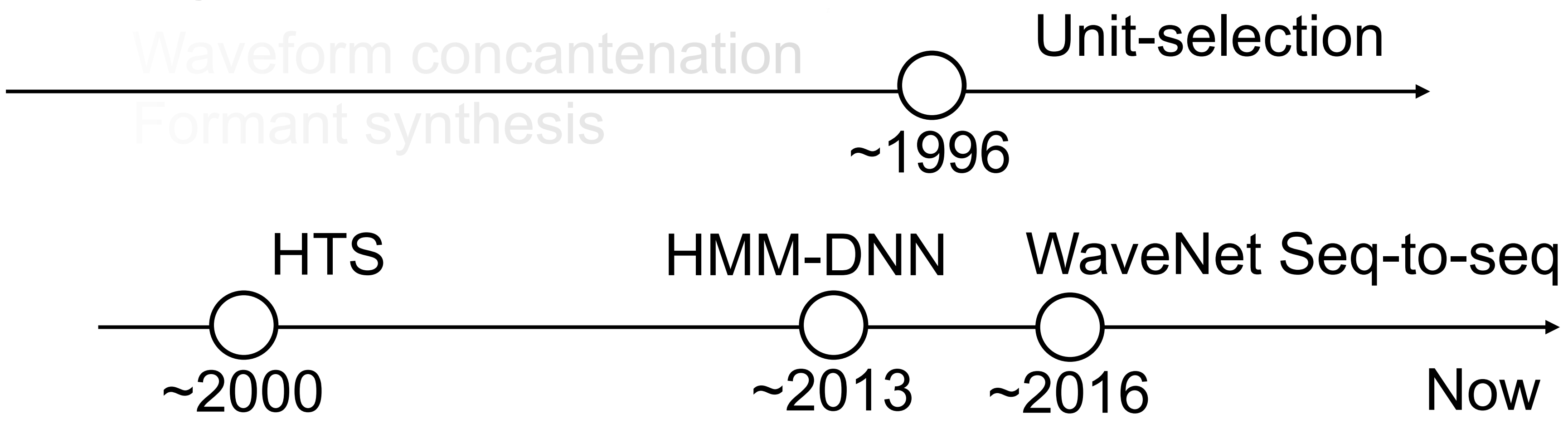


Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. A Survey on Neural Speech Synthesis. *ArXiv Preprint ArXiv:2106.15561*. 2021.
Xin Wang, Neural statistical parametric speech synthesis, ISCA Odyssey 2020, tutorial: https://tonywangx.github.io/slide.html#dec-2020

Voice conversion – Recent methods



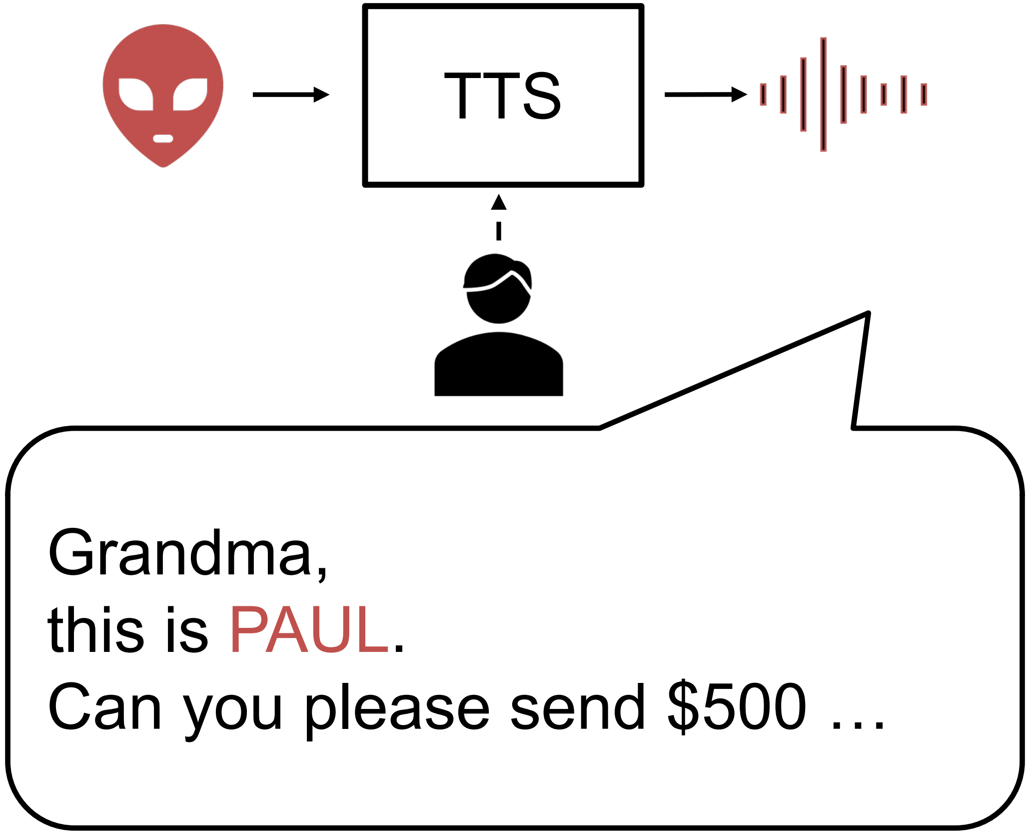
Rapid progress of TTS



Unit-selection	HMM-DNN (close to HMM)	HMM-DNN (close to DNN)	Seq2seq	Natural speech

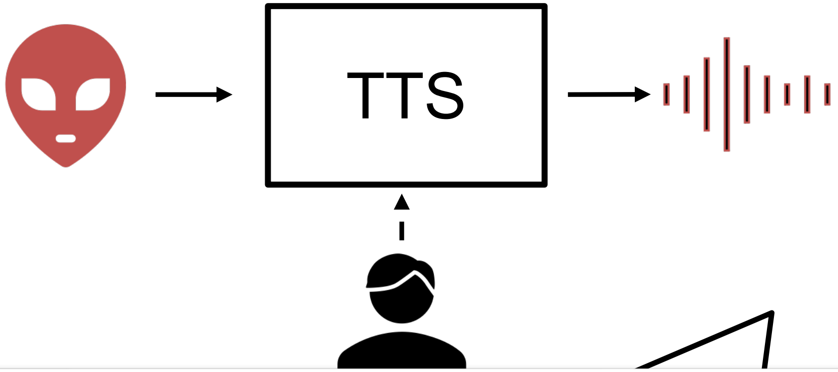
More samples: Xin Wang, Neural statistical parametric speech synthesis, ISCA Odyssey 2020, tutorial:
https://tonywangx.github.io/slide.html#dec-2020

TTS/VC may be misused



Example from Deepfakes : High-tech Illusions to Trick the Human Brain University of Zurich (UZH), National Institute of Informatics (NII), Science & Technology Office Tokyo, Embassy of Switzerland in Japan https://www.jst.go.jp/sis/scienceagora/2021/session/11-a17.html

TTS/VC may be misused

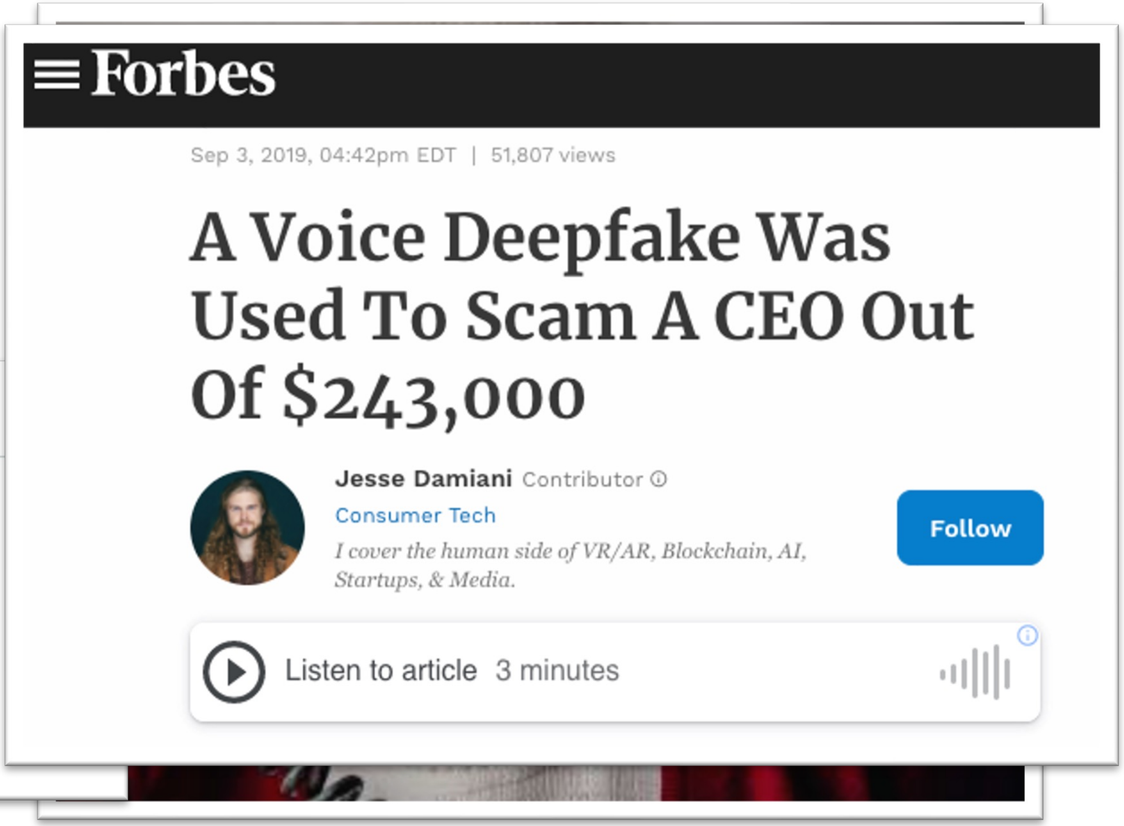



THE WALL STREET JOURNAL.

PRO CYBER NEWS

Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case

Scams using artificial intelligence are a new challenge for companies


Forbes

Sep 3, 2019, 04:42pm EDT | 51,807 views

A Voice Deepfake Was Used To Scam A CEO Out Of $243,000

Jesse Damiani Contributor

Consumer Tech

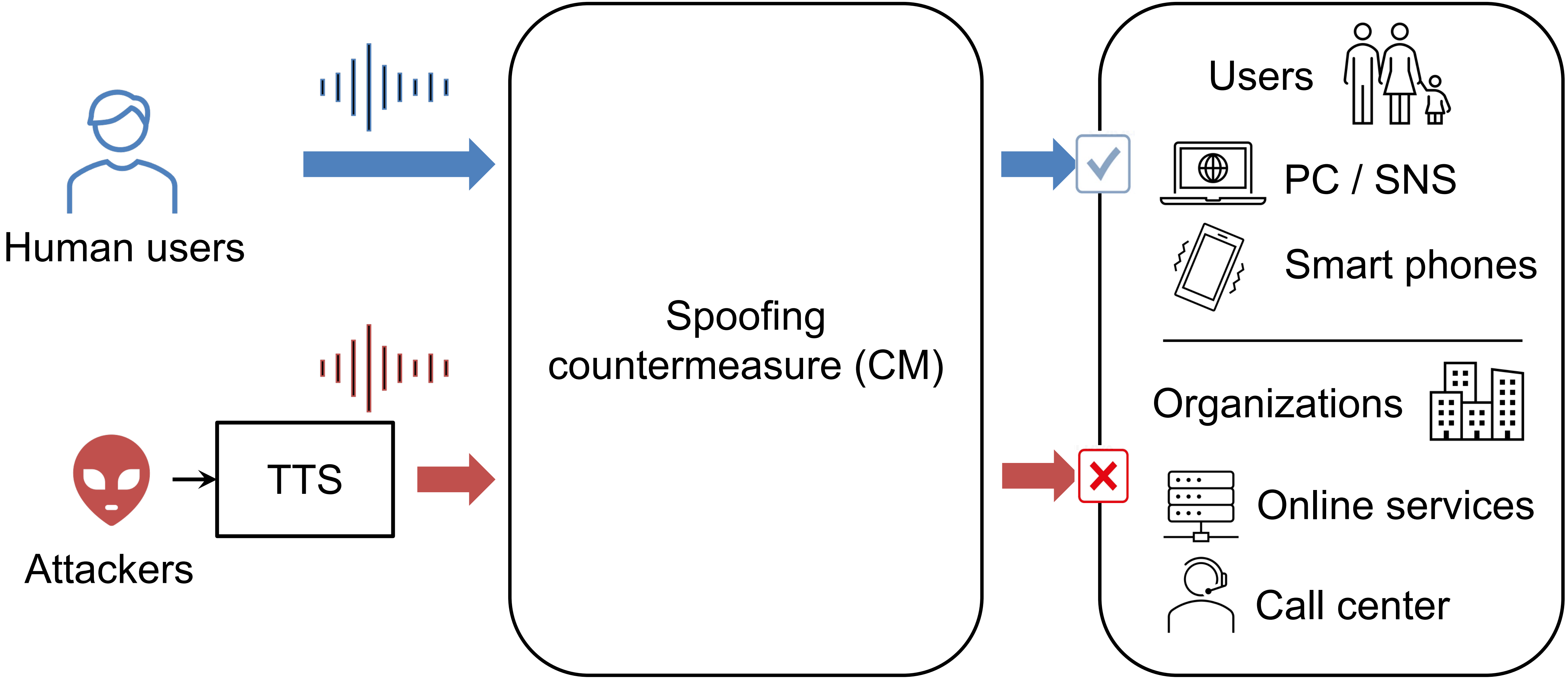
I cover the human side of VR/AR, Blockchain, AI, Startups, & Media.

Follow

Listen to article 3 minutes

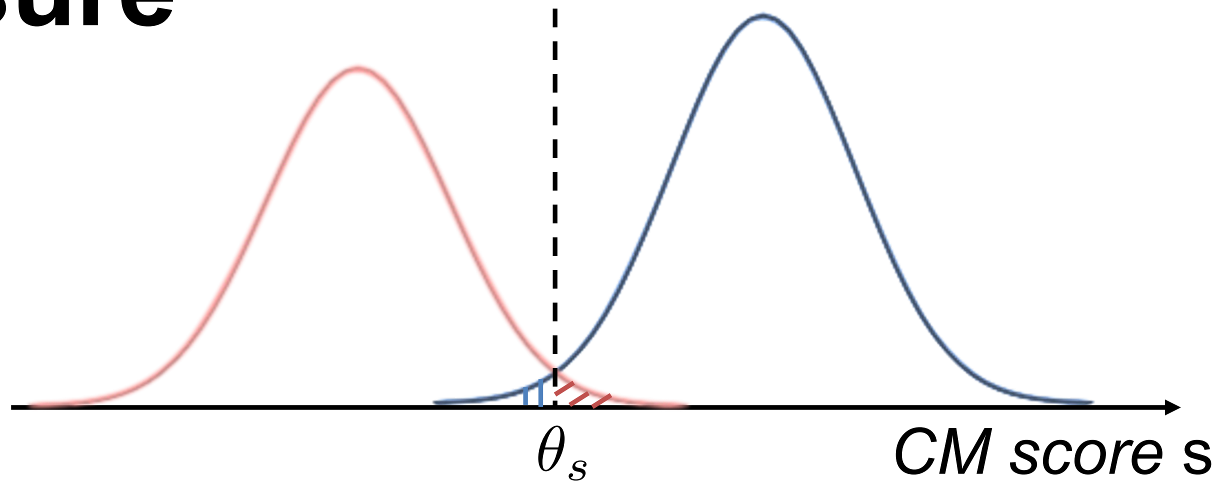
https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402
https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/?sh=1081d7802241

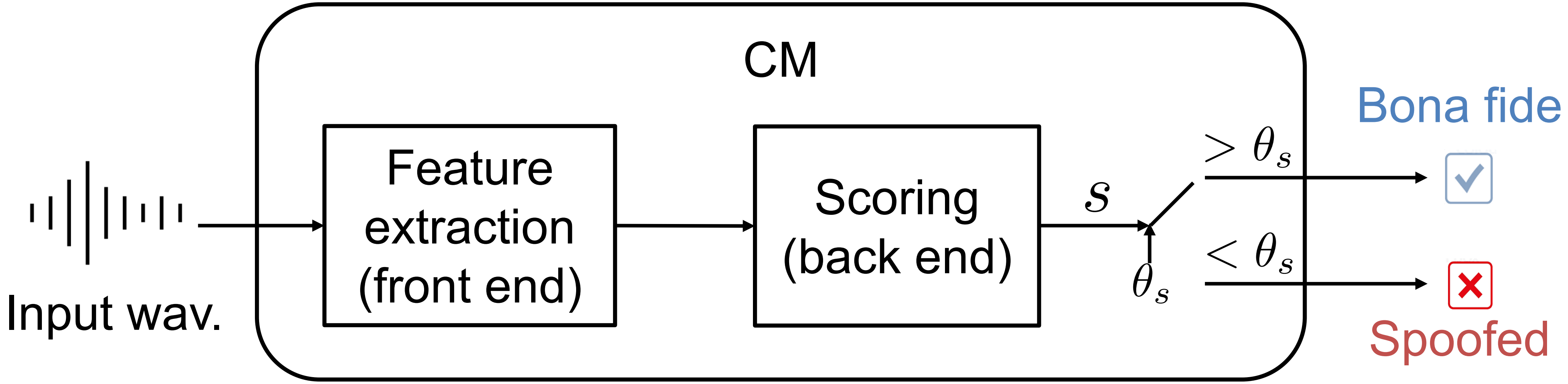
Spoofing countermeasure



Spoofing countermeasure

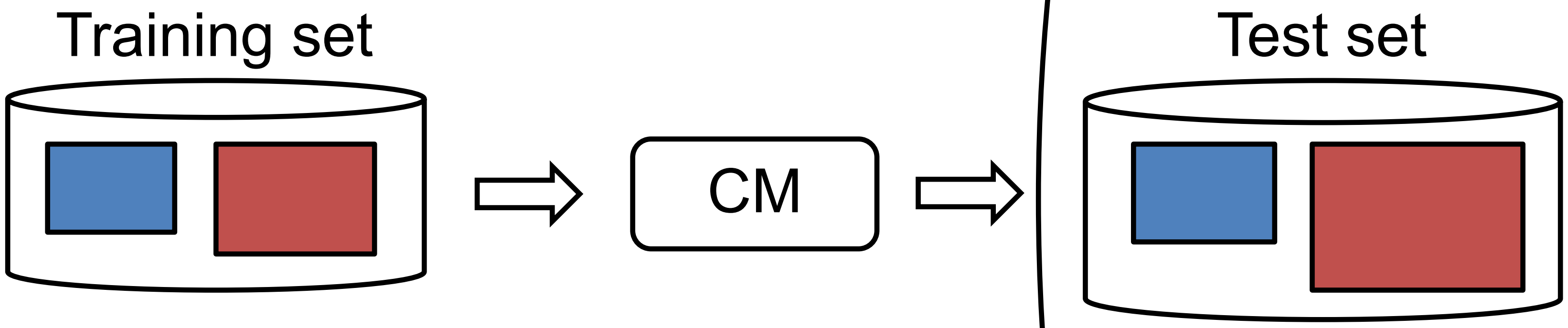
- **Model construction**
  - a binary classification model
- **Model evaluation**
  - accuracy
  - **equal error rate (EER)**
  - **tandem detection cost function (t-DCF)** **(Kinnunen 2020)**





Speech anti-spoofing

Space of all possible bona fide and spoofed data



- training and test sets from a database
  - ASVspoof www.asvspoof.org
  - BTAS2016
  - FMFCC-A
  - ...

En, Fr, Ch, Jp, …

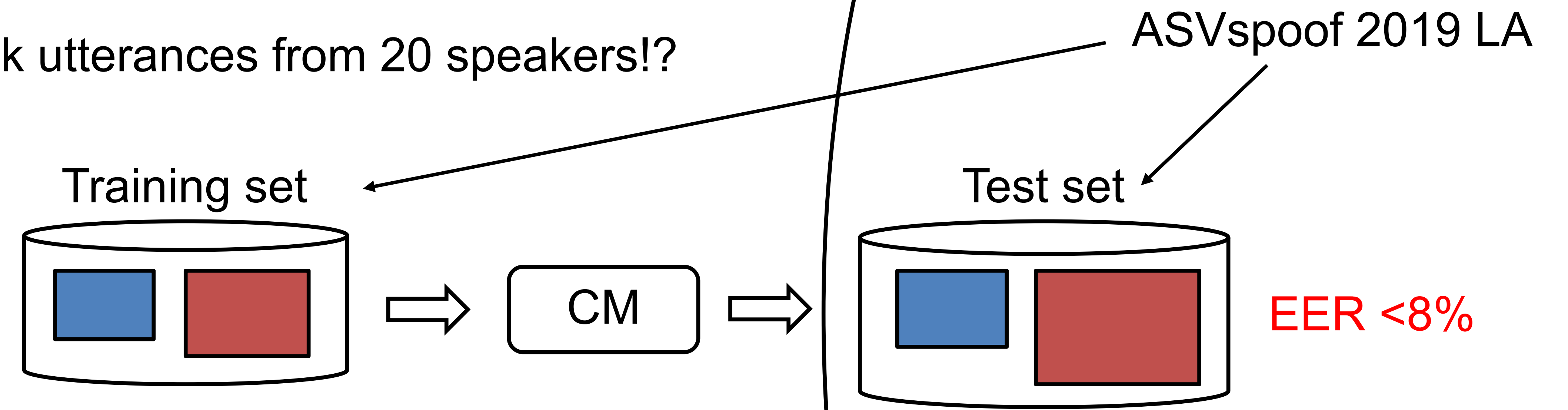
Wav, mp3, m4a …

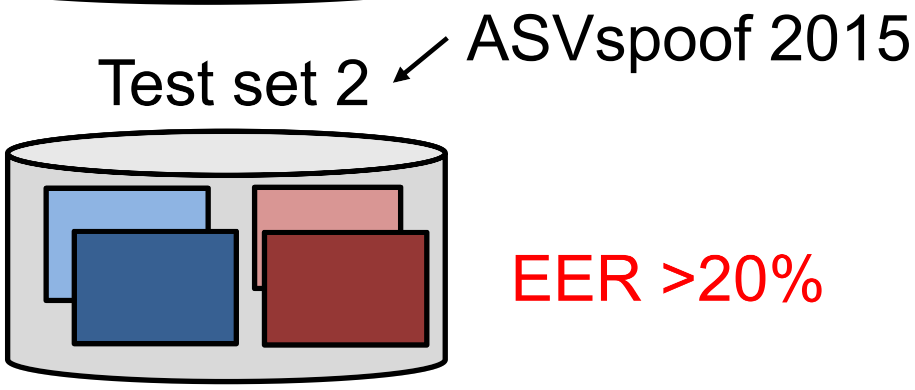
New TTS/VC methods

Pavel Korshunov, Sébastien Marcel, Hannah Muckenhirn, André R Gonçalves, A G Souza Mello, R P Velloso Violato, Flávio O Simoes, M Uliani Neto, Marcus de Assis Angeloni, José Augusto Stuchi, and others. Overview of BTAS 2016 Speaker Anti-Spoofing Competition. In *Proc. BTAS*, 1–6. doi:10.1109/BTAS.2016.7791200. 2016.
Zhenyu Zhang, Yewei Gu, Xiaowei Yi, and Xianfeng Zhao. FMFCC-A: A Challenging Mandarin Dataset for Synthetic Speech Detection. In *Digital Forensics and Watermarking. IWDW 2021*, 117–131. doi:10.1007/978-3-030-95398-0_9. 2022.

Speech anti-spoofing

Just 25k utterances from 20 speakers!?

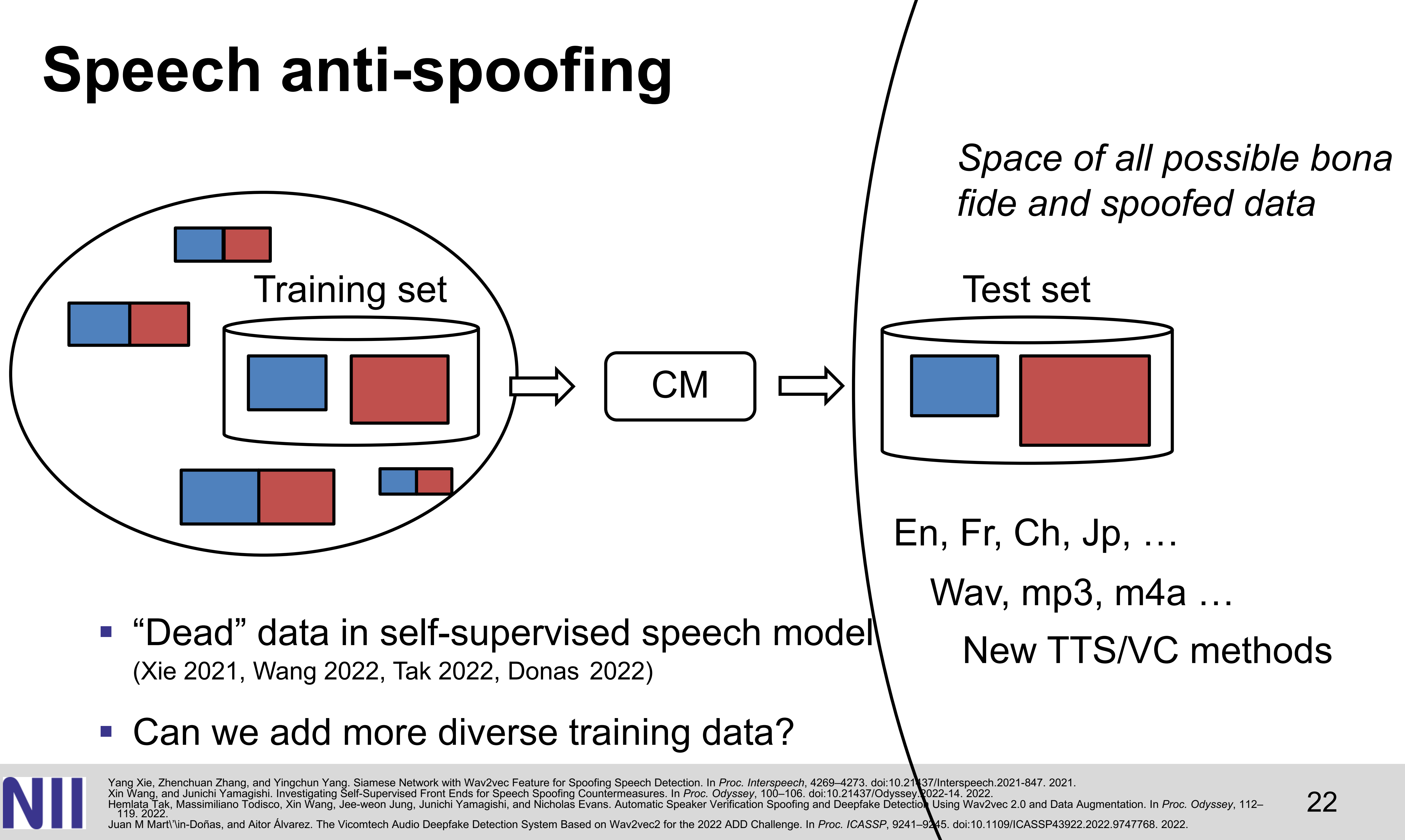




- poor generalization (Das 2020)
- dataset-specific bias?
  - biased dist. of non-speech (Müller 2021, Liu 2022)
  - artefacts in high-frequency band (Wang 2022)
  - …

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.

Speech anti-spoofing



- "Dead" data in self-supervised speech model
  (Xie 2021, Wang 2022, Tak 2022, Donas 2022)
- Can we add more diverse training data?

Yang Xie, Zhenchuan Zhang, and Yingchun Yang. Siamese Network with Wav2vec Feature for Spoofing Speech Detection. In *Proc. Interspeech*, 4269–4273. doi:10.21437/Interspeech.2021-847. 2021.
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.
Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee-weon Jung, Junichi Yamagishi, and Nicholas Evans. Automatic Speaker Verification Spoofing and Deepfake Detection Using Wav2vec 2.0 and Data Augmentation. In *Proc. Odyssey*, 112–119. 2022.
Juan M Mart\'\in-Doñas, and Aitor Álvarez. The Vicomtech Audio Deepfake Detection System Based on Wav2vec2 for the 2022 ADD Challenge. In *Proc. ICASSP*, 9241–9245. doi:10.1109/ICASSP43922.2022.9747768. 2022.

Speech anti-spoofing

Space of all possible bona fide and spoofed data

Training set

Test set

- Building various TTS and VC systems are time-consuming.
- We need some efficient ways to create spoofed training data

Wav, mp3, m4a …

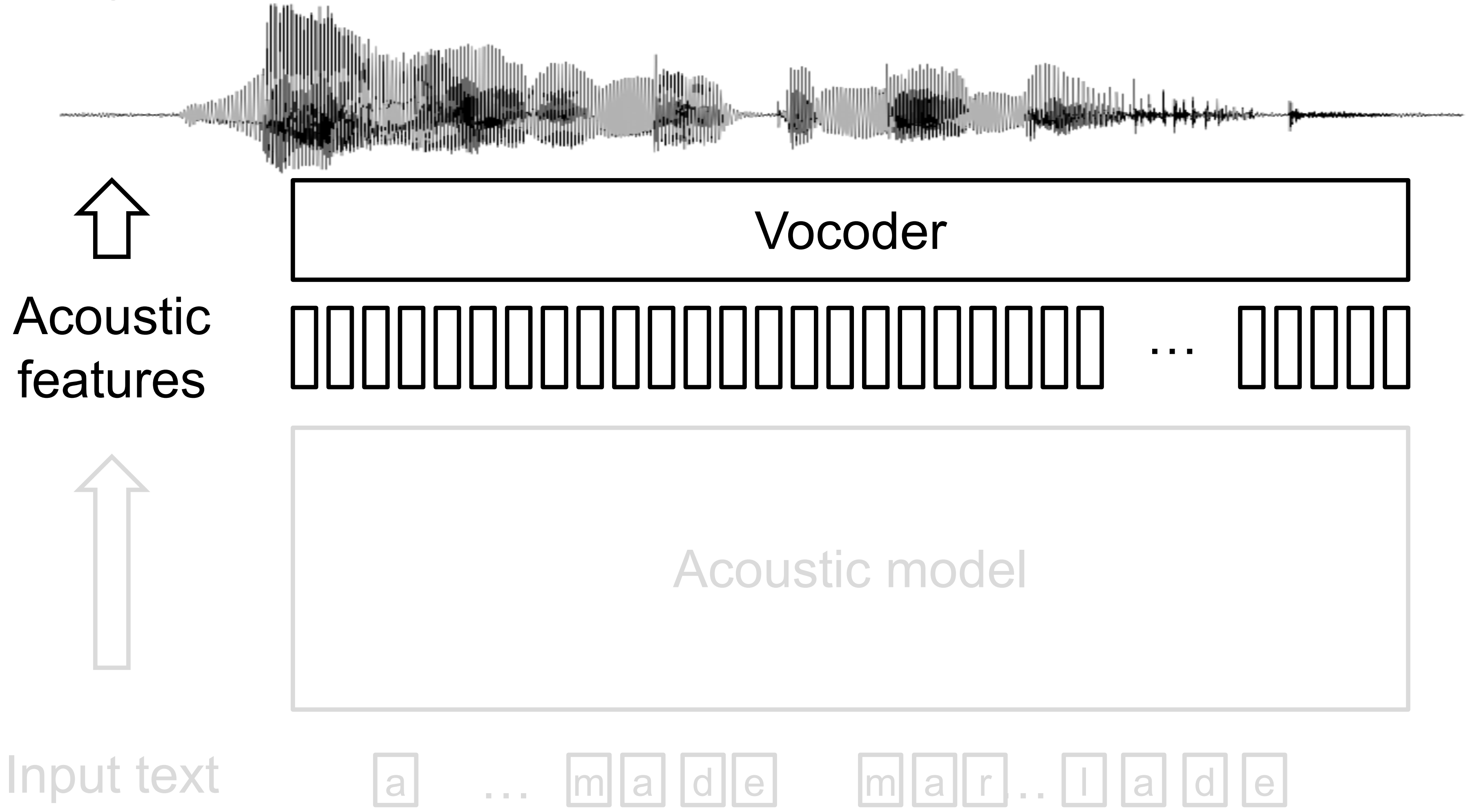
New TTS/VC methods

- "Dead" data in self-supervised speech model
  (Xie 2021, Wang 2022, Tak 2022, Donas 2022)
- Can we add more diverse training data?

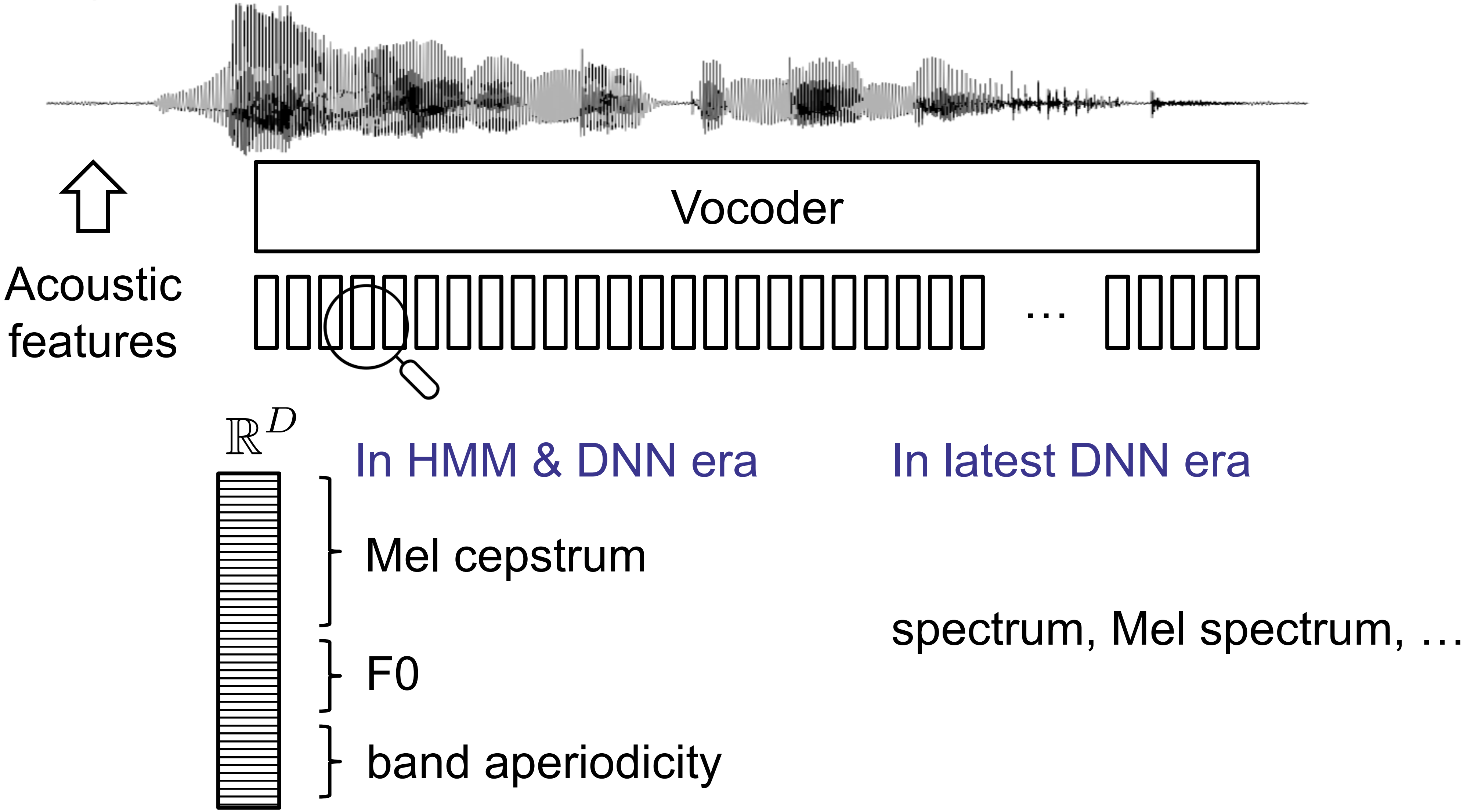
Yang Xie, Zhenchuan Zhang, and Yingchun Yang. Siamese Network with Wav2vec Feature for Spoofing Speech Detection. In *Proc. Interspeech*, 4269–4273. doi:10.21437/Interspeech.2021-847. 2021.
Xin Wang, and Junichi Yamagishi. Investigating Self-Supervised Front Ends for Speech Spoofing Countermeasures. In *Proc. Odyssey*, 100–106. doi:10.21437/Odyssey.2022-14. 2022.
Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee-weon Jung, Junichi Yamagishi, and Nicholas Evans. Automatic Speaker Verification Spoofing and Deepfake Detection Using Wav2vec 2.0 and Data Augmentation. In *Proc. Odyssey*, 112–119. 2022.
Juan M Mart\'\in-Doñas, and Aitor Álvarez. The Vicomtech Audio Deepfake Detection System Based on Wav2vec2 for the 2022 ADD Challenge. In *Proc. ICASSP*, 9241–9245. doi:10.1109/ICASSP43922.2022.9747768. 2022.

Method

Copy synthesis

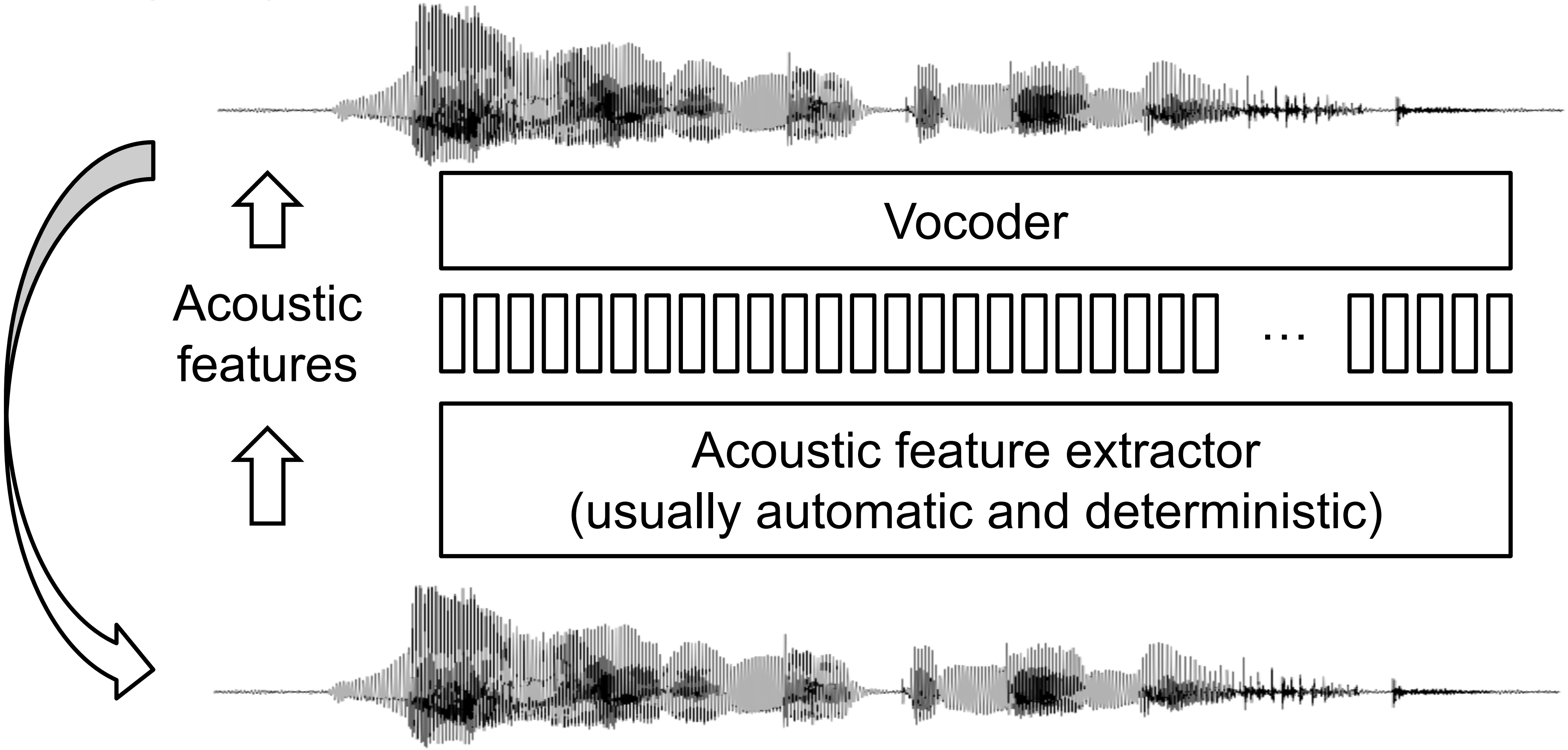


Copy synthesis



Takayoshi Yoshimura, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadasgu Kitamura. Simultaneous Modeling of Spectrum, Pitch and Duration in {HMM}-Based Speech Synthesis. In *Proc. Eurospeech*, 2347–2350. 1999.
Yuxuan Wang, R J Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A Saurous. Tacotron: Towards End-to-End Speech Synthesis. In *Proc. Interspeech*, 4006–4010. 2017.

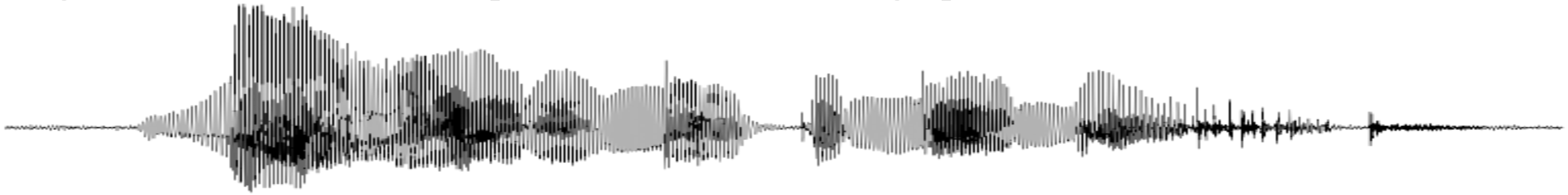
Copy synthesis

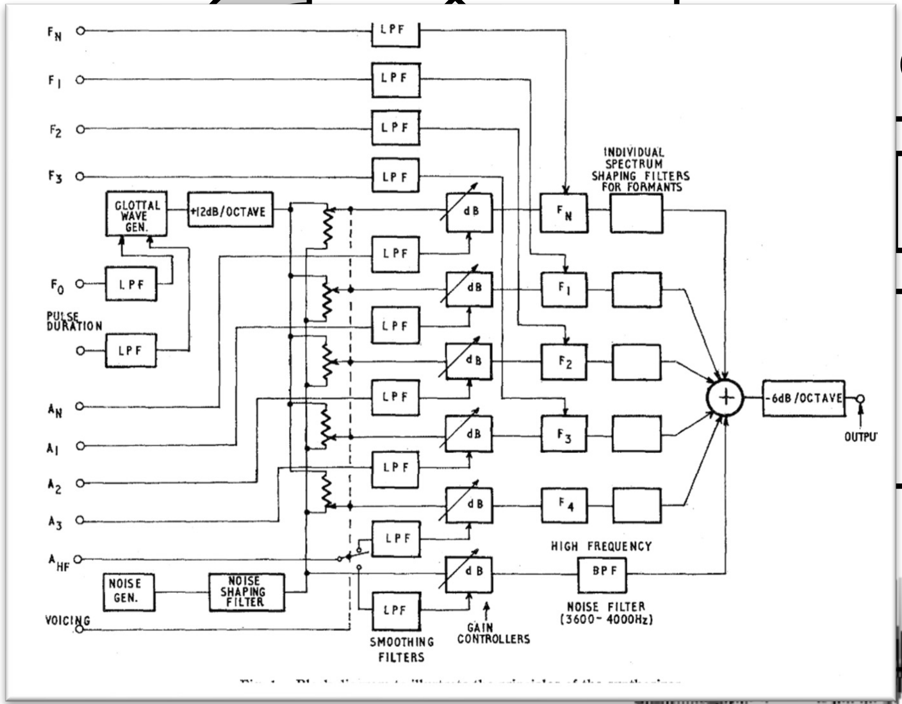


Copy-synthesis, analysis-by-synthesis, copy-resynthesis, vocoding …

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.

Copy synthesis (in history)





ormant synthesizer (deterministic)

…

Acoustic feature extractor
(manual w/ trial & error)

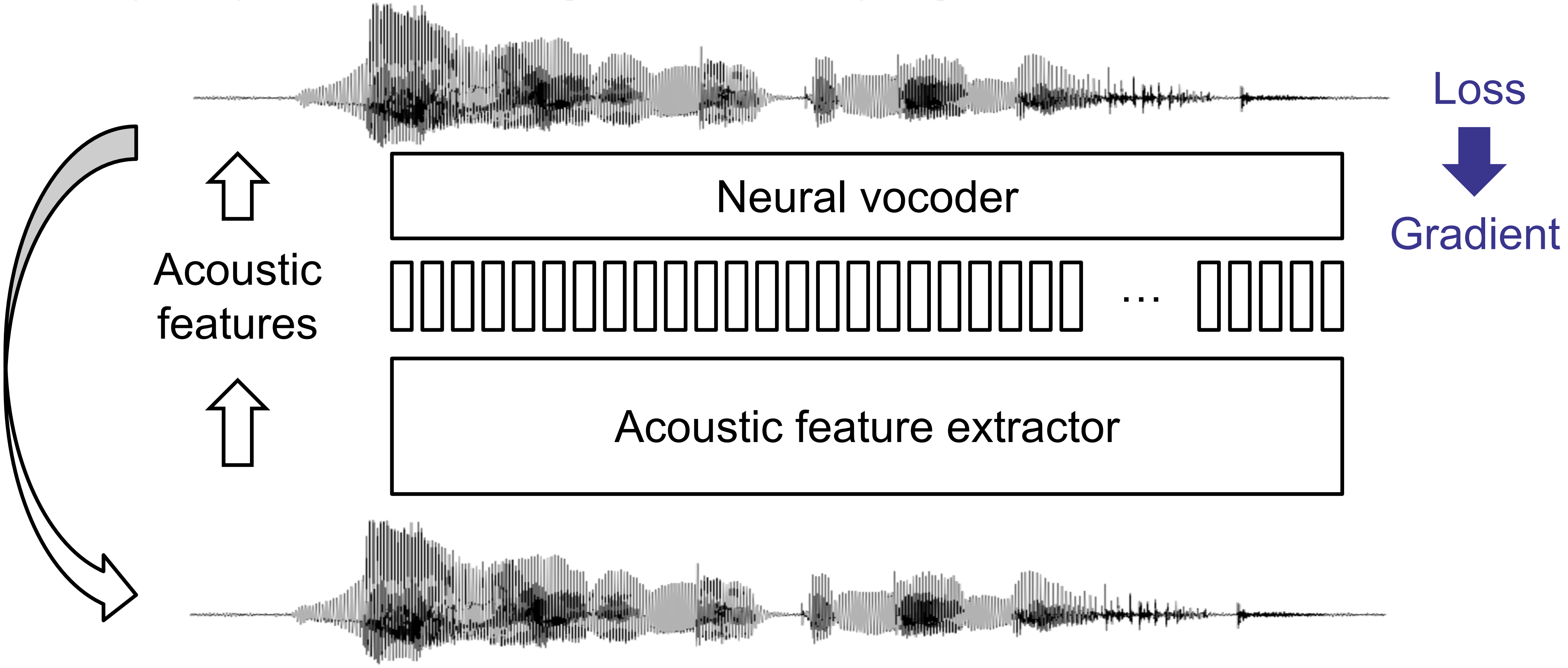
IV. Synthesis of Copies of Natural Utterances

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.

Copy synthesis (nowadays)

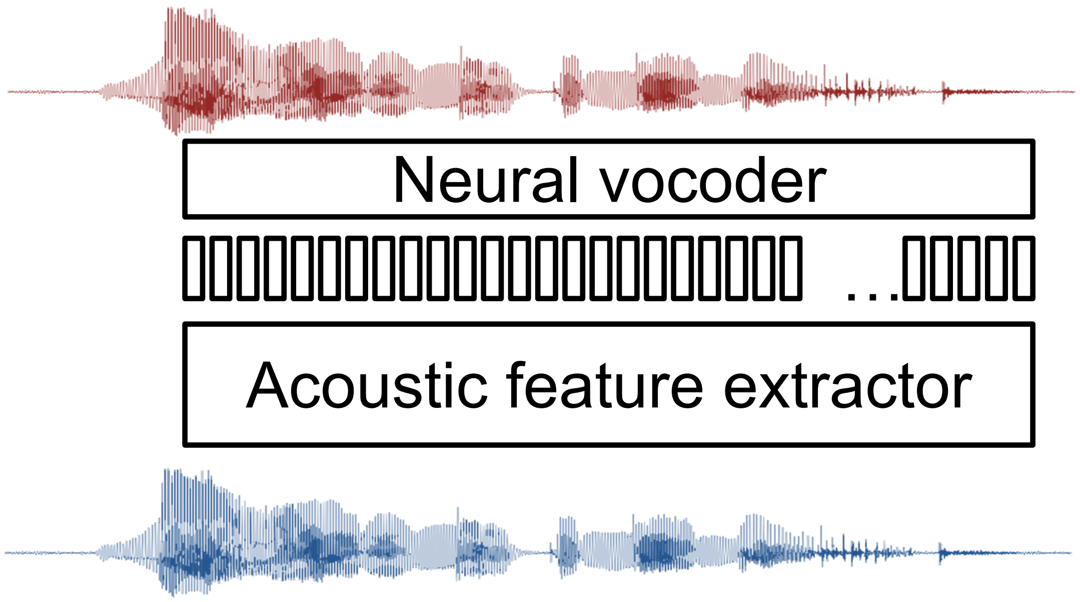


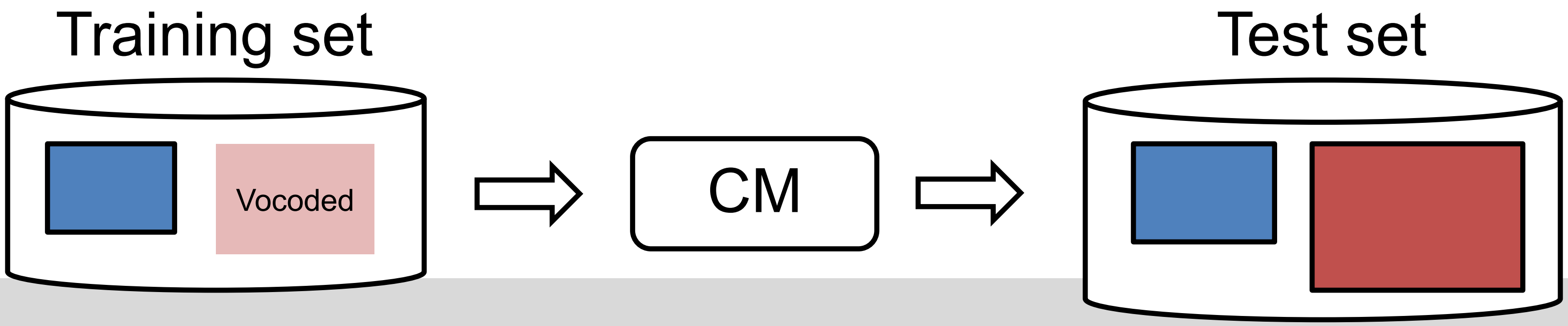
We do copy-synthesis when training the neural vocoders

Creating copy-synthesized spoofed data

Steps

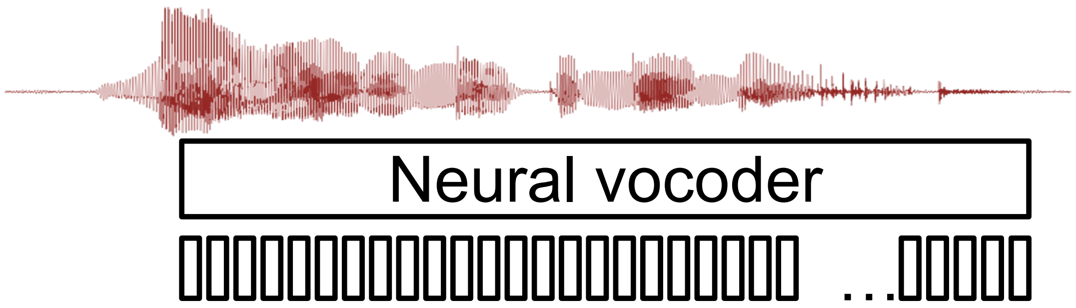
- prepare (or training) vocoders
  - not necessarily neural vocoders
- Do copy synthesis on bona fide data
- use output as copy-synthesized (or vocoded) spoofed data
- train a CM using {bona fide, copy-synthesized spoofed}



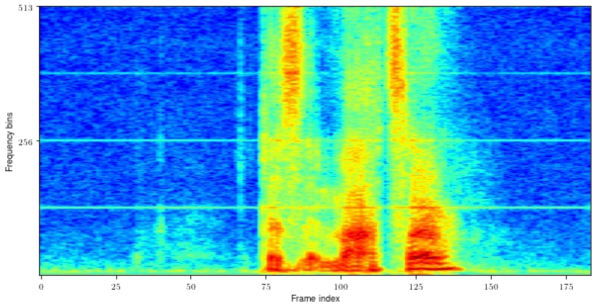


Creating copy-synthesized spoofed data

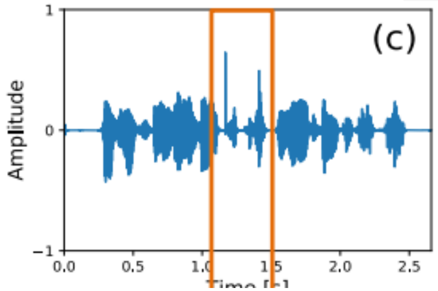
- **Hypothesis**
  - *Copy-synthesis is TTS/VC with a perfect acoustic model*
    - x artefacts by the acoustic model
    - ✓artefacts by the vocoder
  - TTS/VC spoofed data may contain artefacts by the vocoder



WaveGlow “bar” (Prenger 2019)



WaveNet “click” (Wu 2018)



More artefacts not perceptible to ears

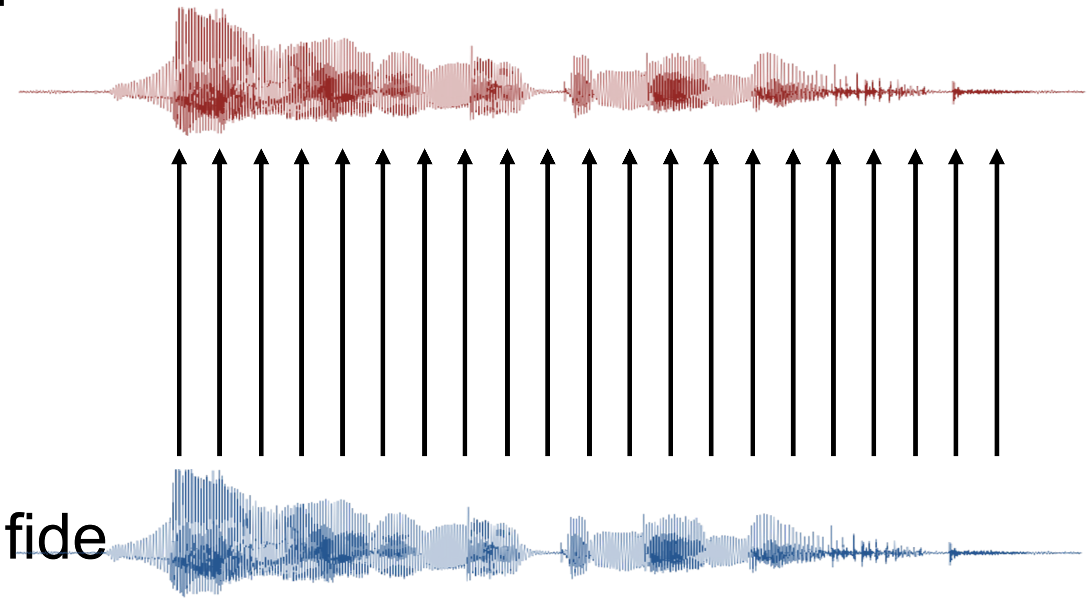
Ryan Prenger, Rafael Valle, and Bryan Catanzaro. WaveGlow: A Flow-Based Generative Network for Speech Synthesis. In *Proc. ICASSP*, 3617–3621. 2019.
Yi-Chiao Wu, Kazuhiro Kobayashi, Tomoki Hayashi, Patrick Lumban Tobing, and Tomoki Toda. Collapsed Speech Segment Detection and Suppression for WaveNet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

Creating copy-synthesized spoofed data

Questions

- How to prepare or train the vocoder?
  - pre-trained vocoder(s)?
  - Fine tuning?

Experiment I



- Can we better use the aligned bona fide and spoofed data pairs?

Experiment II

Creating copy-synthesized spoofed data

Related studies 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.

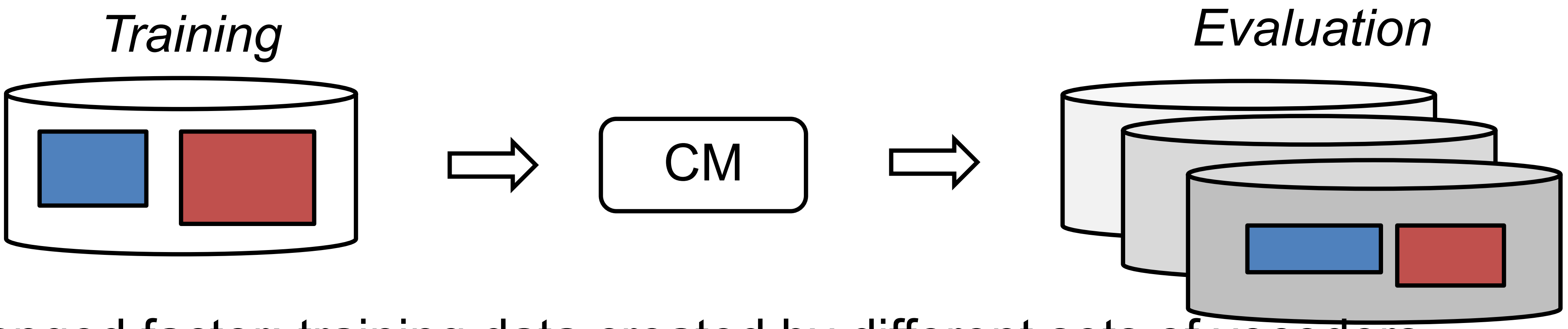
Related studies using neural vocoders

- 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.

Experiment I
How to prepare or train the vocoder?

Experiment I

❑ Design



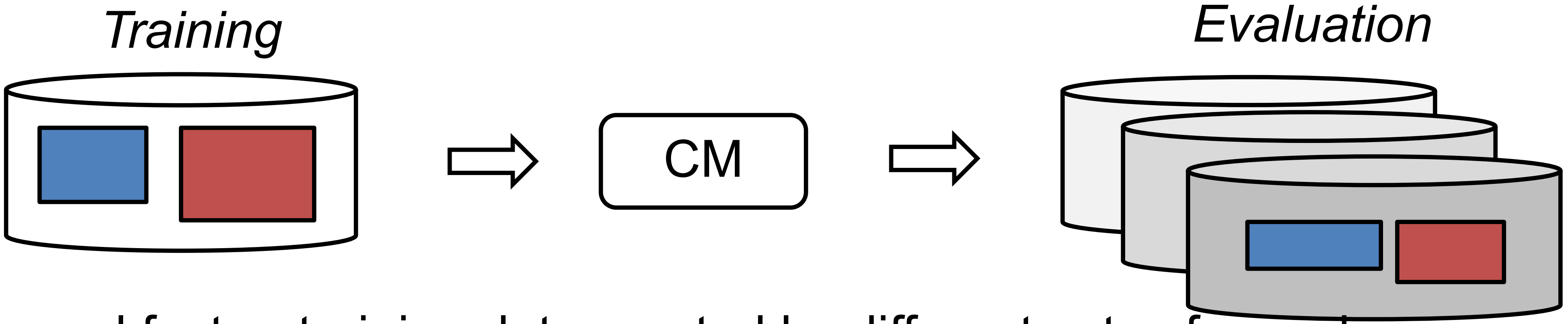
- changed factor: training data created by different sets of vocoders
- unchanged: CM (Wang 2022) using a Wav2vec2.0-based front end (Baevski 2020)
- unchanged: multiple test sets

- three independent training & evaluation rounds
- averaged EER

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.

Experiment I

❑ Design



- changed factor: training data created by different sets of vocoders
- unchanged: CM (Wang 2022) using a Wav2vec2.0-based front end (Baevski 2020)
- unchanged: multiple test sets
    - ASVspoof 2019 LA test set, 2021 LA&DF eval track, 2015 test set
    - ASVspoof 2019 LA test set w/o non-speech, 2021 LA & DF hidden track
    - WaveFake(Frank 2021), In-the-Wild (Müller 2022)

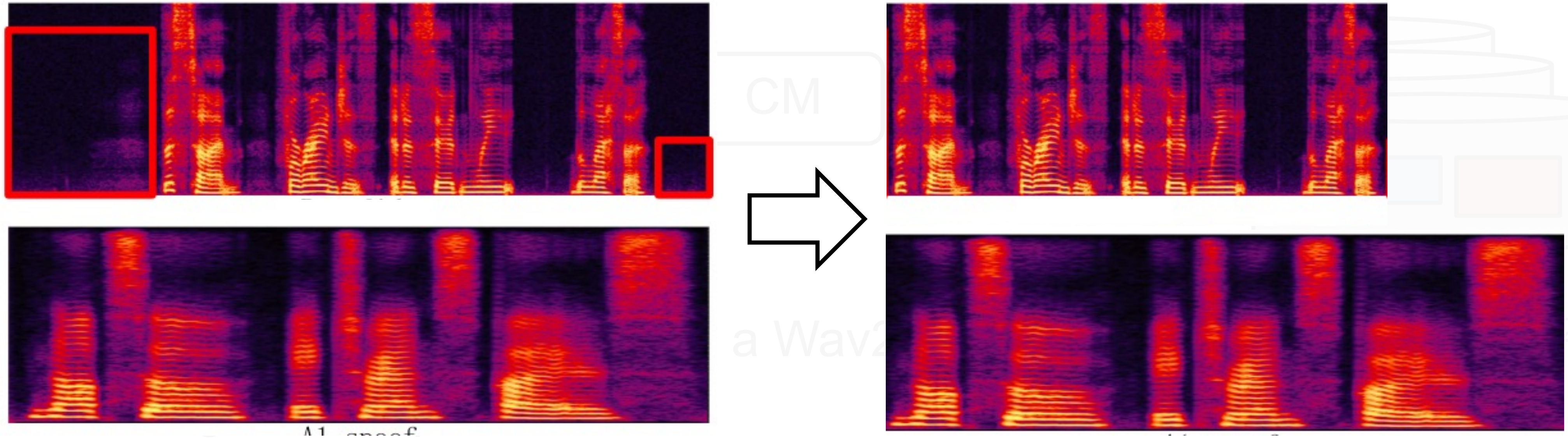
Joel Frank, and Lea Schönherr. WaveFake: A Data Set to Facilitate Audio DeepFake Detection. In *Proc. NeurIPS Datasets and Benchmarks 2021*. 2021.
Nicolas M Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger. Does Audio Deepfake Detection Generalize? In *Proc. Interspeech*,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

Experiment I

❑ Design

Original test trials

Non-speech trimmed test trials



(figure from Zhang 2021)

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

We recommend testing on both versions of ASVspoof test sets

Yuxiang Zhang, Wenchao Wang, and Pengyuan Zhang. The Effect of Silence and Dual-Band Fusion in Anti-Spoofing System. In *Proc. Interspeech*, 4279–4283. doi:10.21437/Interspeech.2021-1281. 2021.

Experiment I

❑ Training data

ID	#. Spr.	#. Bona.	#. Spoof.	Vocoder type	Implementation	Vocoder train/fine-tune data	Vocoder SR
LA19trn	20	2,580	22,800	-	-	-	16 kHz
`WFtrn`	1	3,930	15,720	HiFiGAN, MB-MelGAN, PWG, WaveGlow	ESPNet toolkit	LJSpeech / -	24 kHz
`Voc.v1`	20 same as `LA19trn`	2,580	10,320	HiFiGAN, MB-MelGAN, PWG, StyleMelGAN	ESPNet toolkit	LibriTTS / -	24 kHz
`Voc.v2`				HiFiGAN, NSF-HiFiGAN, Hn-NSF, WaveGlow	in-house	LibriTTS / -	16 kHz
`Voc.v3`				HiFiGAN, NSF-HiFiGAN, Hn-NSF, WaveGlow	in-house	`LA19trn` bona. / -	16 kHz
`Voc.v4`				HiFiGAN, NSF-HiFiGAN, Hn-NSF, WaveGlow	in-house	LibriTTS / `LA19trn` bona.	16 kHz

- `LA19trn`: ASVspoof 2019 LA training set
- `WFtrn`: WaveFake English subset, down-sampled to 16kHz
- `Voc.v*`: ASVspoof 2019 LA training set bona fide data + **its vocoded data**

LibriTTS subset: 45 hours
ASVspoof 2019 LA training set bona fide: 5 hours

Experiment I

Low EER High EER

- Results in EER (%, mean of three runs)

		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

⇦ Single threshold

Experiment I

- Results in EER (%, mean of three runs)

Test sets	LA19 trn	WF trn	Voc. v1	Voc. v2	Voc. v3	Voc. v4
	Training set					
LA19eval	2.98	44.48	5.78	5.32	8.74	4.36
LA21eval	7.53	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
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

Similar results to our previous work (Wang 2022)

Same trend as previous studies (Müller 2021, Liu 2022)

Not 50% :)

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	Vocoder SR
LA19trn	20	2,580	22,800	-	-	-	16 kHz
WFtrn	1	3,930	15,720	HiFiGAN, MB-MelGAN, PWG, WaveGlow	ESPNet toolkit	LJSpeech / -	24 kHz
Voc.v1	20 same as LA19trn	2,580	10,320	HiFiGAN, MB-MelGAN, PWG, StyleMelGAN	ESPNet toolkit	LibriTTS / -	24 kHz
Voc.v2				HiFiGAN, NSF-HiFiGAN, Hn-NSF, WaveGlow	in-house	LibriTTS / -	16 kHz
Voc.v3				HiFiGAN, NSF-HiFiGAN, Hn-NSF, WaveGlow	in-house	LA19trn bona. / -	16 kHz
Voc.v4				HiFiGAN, NSF-HiFiGAN, Hn-NSF, WaveGlow	in-house	LibriTTS / LA19trn bona.	16 kHz

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)

: (

Tomoki Hayashi, Ryuichi Yamamoto, Katsuki Inoue, Takenori Yoshimura, Shinji Watanabe, Tomoki Toda, Kazuya Takeda, Yu Zhang, and Xu Tan. Espnet-TTS: Unified, Reproducible, and Integratable Open Source End-to-End Text-to-Speech Toolkit. In *Proc. ICASSP*, 7654–7658. 2020.

ID	#. Spr.	#. Bona.	#. Spoof.	Vocoder type	Implementation	Vocoder train/fine-tune data	Vocoder SR
LA19trn	20	2,580	22,800	-	-	-	16 kHz
WFtrn	1	3,930	15,720	HiFiGAN, MB-MelGAN, PWG, WaveGlow	ESPNet toolkit	LJSpeech / -	24 kHz
Voc.v1	20 same as LA19trn	2,580	10,320	HiFiGAN, MB-MelGAN, PWG, StyleMelGAN	ESPNet toolkit	LibriTTS / -	24 kHz
Voc.v2				HiFiGAN, NSF-HiFiGAN, Hn-NSF, WaveGlow	in-house	LibriTTS / -	16 kHz
Voc.v3				HiFiGAN, NSF-HiFiGAN, Hn-NSF, WaveGlow	in-house	LA19trn bona. / -	16 kHz
Voc.v4				HiFiGAN, NSF-HiFiGAN, Hn-NSF, WaveGlow	in-house	LibriTTS / LA19trn bona.	16 kHz

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	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

Was vocoder trained on the bona fide data?

No

Yes

ID	#. Spr.	#. Bona.	#. Spoof.	Vocoder type	Implementation	Vocoder train/fine-tune data	Vocoder SR
LA19trn	20	2,580	22,800	-	-	-	16 kHz
WFtrn	1	3,930	15,720	HiFiGAN, MB-MelGAN, PWG, WaveGlow	ESPNet toolkit	LJSpeech / -	24 kHz
Voc.v1	20 same as LA19trn	2,580	10,320	HiFiGAN, MB-MelGAN, PWG, StyleMelGAN	ESPNet toolkit	LibriTTS / -	24 kHz
Voc.v2				HiFiGAN, NSF-HiFiGAN, Hn-NSF, WaveGlow	in-house	LibriTTS / -	16 kHz
Voc.v3				HiFiGAN, NSF-HiFiGAN, Hn-NSF, WaveGlow	in-house	LA19trn bona. / -	16 kHz
Voc.v4				HiFiGAN, NSF-HiFiGAN, Hn-NSF, WaveGlow	in-house	LibriTTS / LA19trn bona.	16 kHz

		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	[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

Experiment I

How to prepare or train the vocoder?

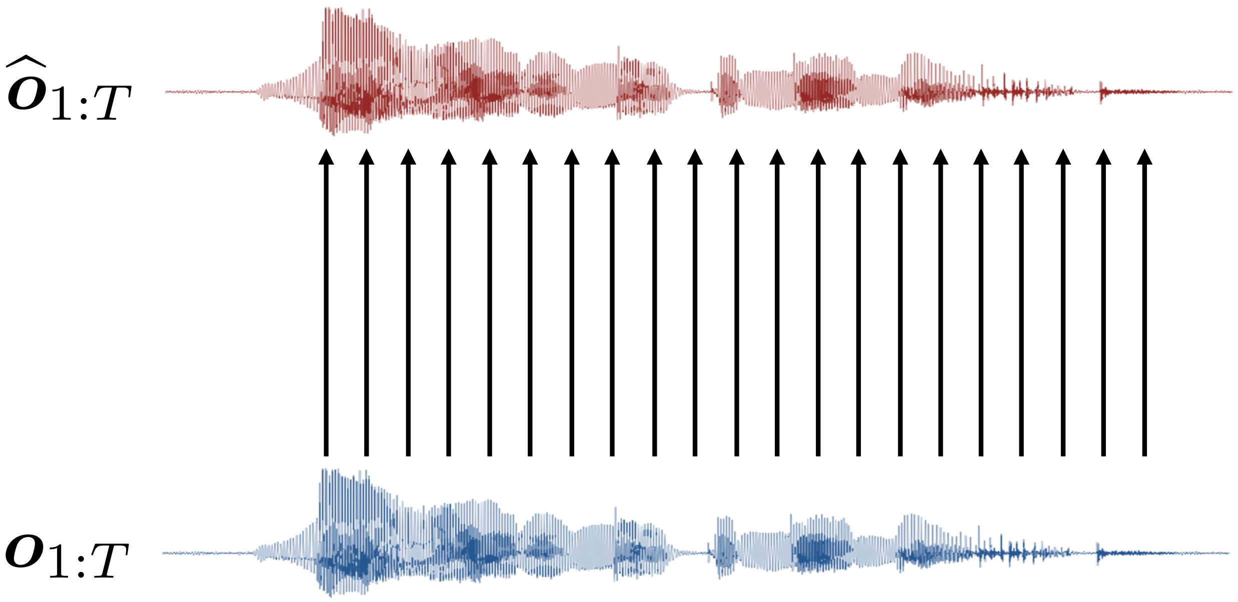
- Pre-trained vocoders may not work --- WFtrn, Voc.v1
- It is better to fine tune vocoders on the bona fide to be copy-synthesized

ID	#. Spr.	#. Bona.	#. Spoof.	Vocoder type	Implementation	Vocoder train/fine-tune data	Vocoder SR
LA19trn	20	2,580	22,800	-	-	-	16 kHz
WFtrn	1	3,930	15,720	HiFiGAN, MB-MelGAN, PWG, WaveGlow	ESPNet toolkit	LJSpeech / -	24 kHz
Voc.v1	20 same as LA19trn	2,580	10,320	HiFiGAN, MB-MelGAN, PWG, StyleMelGAN	ESPNet toolkit	LibriTTS / -	24 kHz
Voc.v2				HiFiGAN, NSF-HiFiGAN, Hn-NSF, WaveGlow	in-house	LibriTTS / -	16 kHz
Voc.v3				HiFiGAN, NSF-HiFiGAN, Hn-NSF, WaveGlow	in-house	LA19trn bona. / -	16 kHz
Voc.v4				HiFiGAN, NSF-HiFiGAN, Hn-NSF, WaveGlow	in-house	LibriTTS / LA19trn bona.	16 kHz

CAUTION: other CMs (e.g., LCNN and AASIST) did not work well on vocoded data

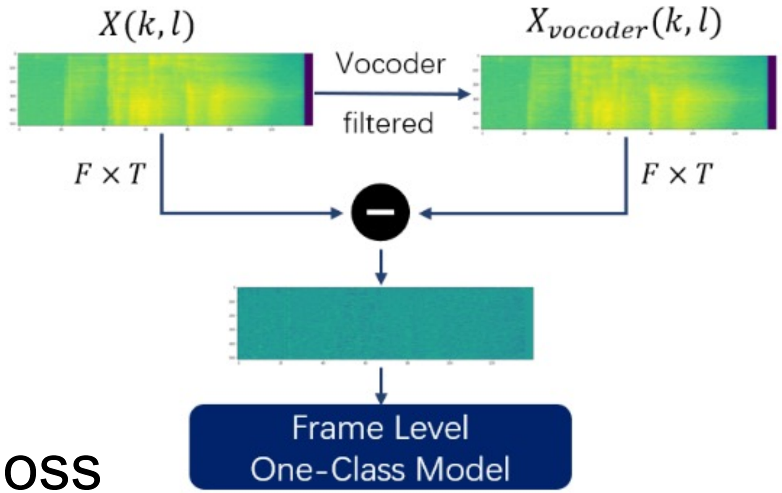
- See results in Appendix, https://arxiv.org/abs/2210.10570

Experiment II
How to make good use of the aligned bona fide and vocoded data pair?



Experiment II

- **Method: make use of the aligned bona fide and vocoded pair**
  - Existing method: use their differences in frequency domain (Wang 2021)
    - Vocoders are needed during inference
    - Too slow to score the test sets

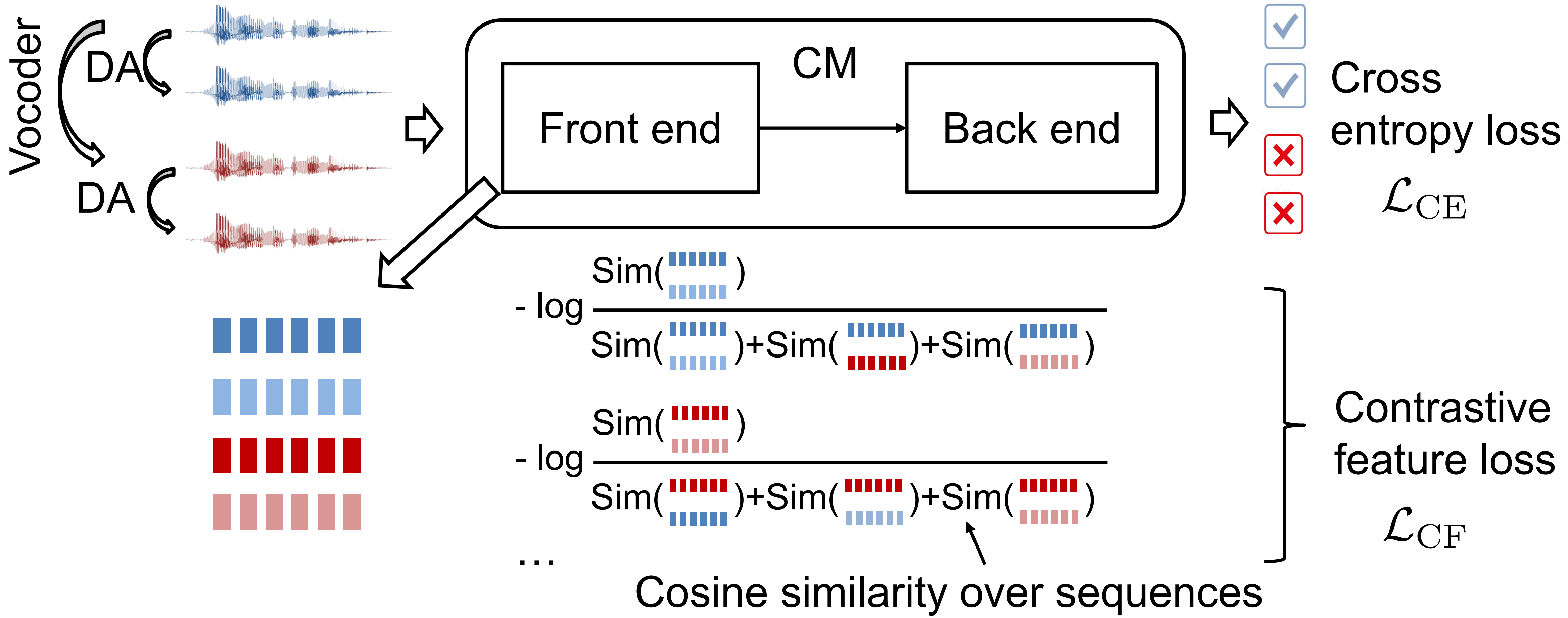


  - We proposed an auxiliary contrastive feature loss
    - It is based on supervised contrastive loss (Khosla 2020)
    - It needs data augmentation (DA)
    - No need to run vocoders during inference

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.

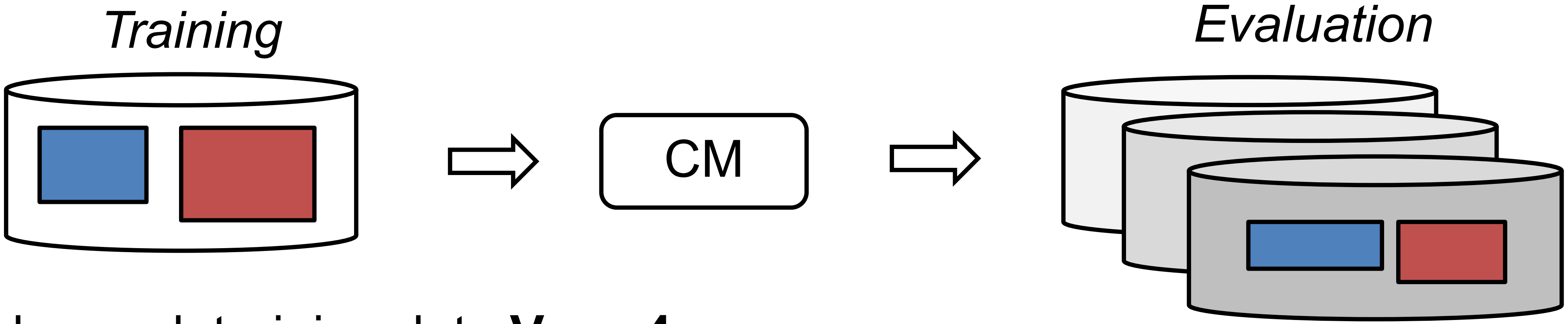
Experiment II

- Method: an auxiliary contrastive feature loss



Experiment II

❑ Design



- unchanged: training data **Voc.v4**
- unchanged: same CM architecture as Experiment I
- unchanged: multiple test sets

- changed: training criterion and method
  - data augmentation based on RawBoost (Tak 2022)

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 II

❑ 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

2 vs 4
RawBoost is useful

4 vs 7
Contrastive feature loss is useful

Experiment II

❑ Results

from Experiment I | Control groups | Best

		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

② vs ④
RawBoost is useful

④ vs ⑦
Contrastive feature loss is useful

⑥ vs ⑦
Use aligned {bona, vocoded} pairs!

Summary

Summary

- **Spoofed training data can be created using neural vocoders**
- **Recommendations (from this study)**
  - Fine-tune vocoders on the bona fide data to be copy-synthesized
  - Exploit the the aligned {bona fide, vocoded} pairs
    - for example, by using contrastive feature loss
  - The best trained CM showed promising generalization performance

Summary

- **Additional findings (see https://arxiv.org/abs/2210.10570)**
  - Do we generalize to old TTS and VC?
    - Yes, EER on the ASVspoof 2015 test set < 1%
  - Is the CM sufficiently generalizable?
    - No, TTS/VC with autoregressive (AR) vocoders are challenging

Unseen Seen Unseen

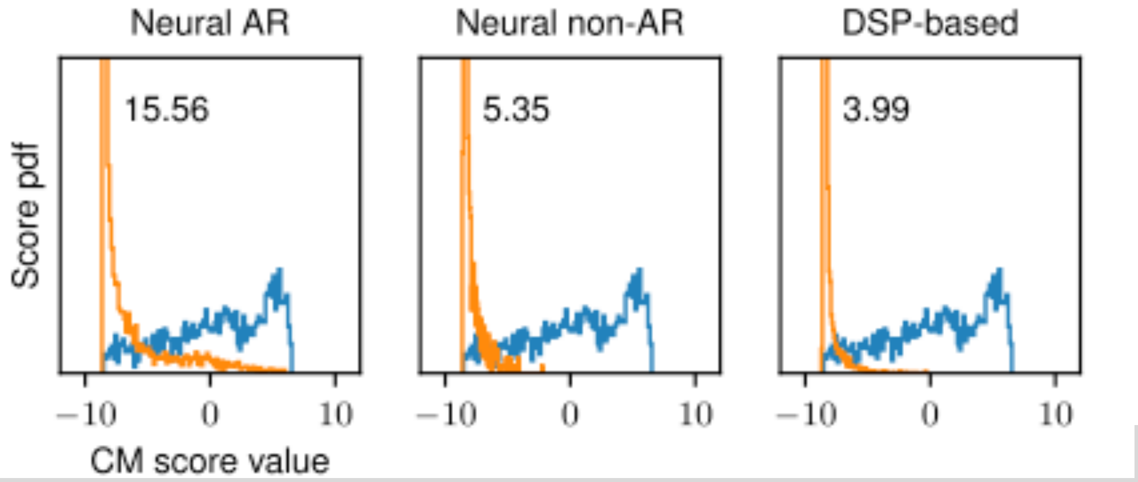


Fig. 1: Score distributions of ⑦ on bona fide and spoofed trials in DF21hid. Number in each sub-figure is EER (%).

In progress

❑ Scale to large data

- CM: same as Experiment II
- ASVspoof 2019 LA trn bona fide + vocoded data: 2 * 5 hours
- VoxCeleb2 dev set + vocoded data: 2300 * 5 hours

	ASVspoof 2019 trn	VoxCeleb2 dev					
Training steps	N	0.3N	0.6N	1.2N	2.4N	6.0N	12.0N
LA19eval	2.21	8.20	7.04	5.40	6.53	5.40	5.63
LA21eval	17.90	20.19	16.73	14.33	18.10	17.44	17.84
DF21eval	5.04	7.49	5.41	5.39	6.00	5.65	5.64
LA19etrim	3.79	6.17	5.53	5.16	5.25	5.22	5.14
LA21hid	14.57	13.98	12.34	11.47	11.64	11.37	11.43
DF21hid	7.78	11.02	9.71	9.90	10.05	9.99	10.04
WaveFake	2.50	14.94	10.39	8.38	5.52	4.88	4.94
InWild	7.55	16.12	15.63	14.19	13.43	13.32	13.77
Pooled	11.27	13.52	11.79	9.98	9.01	8.36	8.37

Resources

- **Code, vocoded data, vocoders, and trained CMs** Git

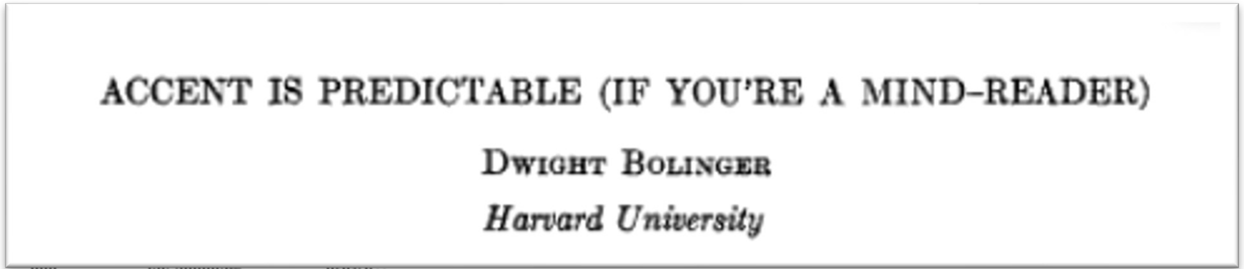
- **Tutorials on neural vocoders (AR, flow, GAN, DSP …)** Git
  - Jupyter notebooks & pre-trained models

- **ASVspoof 2021 hidden track**
  - They are already in the data packages you've downloaded
  - You just need to find them using the official meta labels
  - See https://github.com/asvspoof-challenge/2021

Thank you!

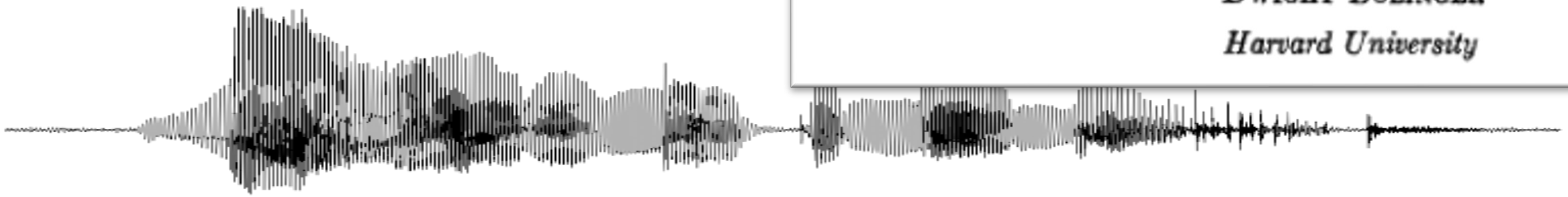
Appendix

Why TTS is difficult?


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 B: Marianna made the marmalade.

Speaker A: Bob made the marmalade.

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

+Prosody tags

H* H* L-L%

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

To phone

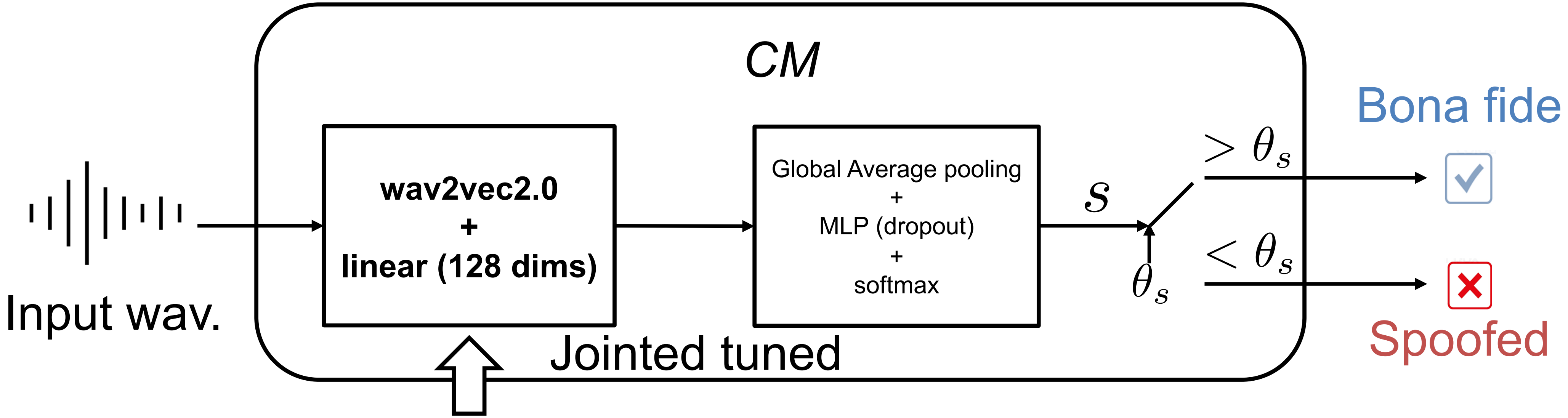
M AA R IY AA N AH M EY D DH AH M AA R M AH L EY D

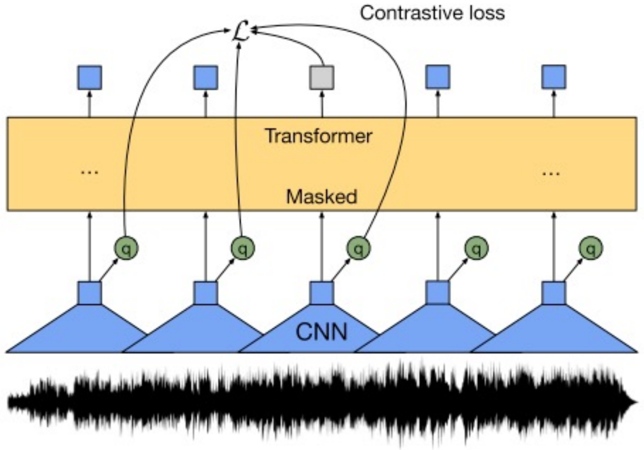
Normali-zation

M a … m a d e m a r .. l a d e

Marianna made the marmalade

CM architecture





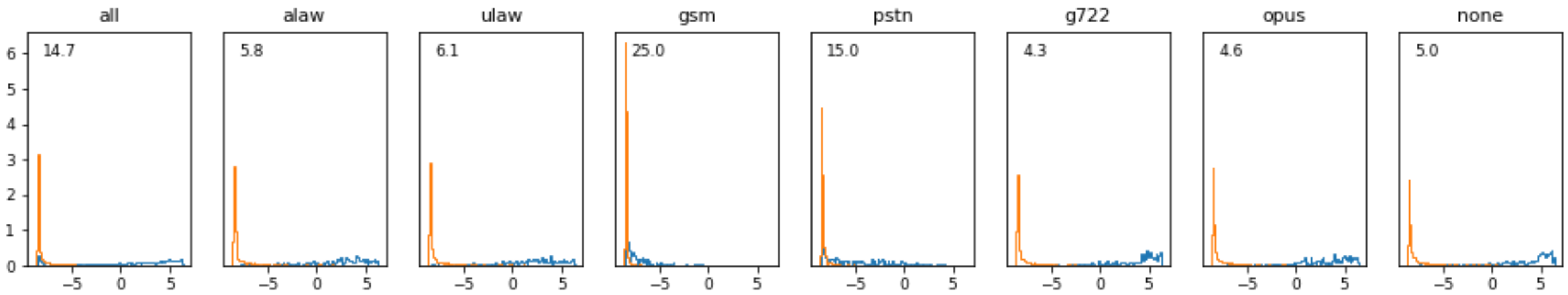
Pre-trained wav2vec 2.0 model
(XLSR-53, pre-trained on multi-lingual data)

Model	Architecture	Hours	Languages	Datasets	Model
XLSR-53	Large	56k	53	MLS, CommonVoice, BABEL	download

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.
https://github.com/pytorch/fairseq/tree/main/examples/wav2vec

Analysis

System 7 on LA 2021



System 7 on DF 2021