



Carnegie Mellon University
Language Technologies Institute

Introduction of ESPnet, end-to-end speech processing toolkit: new features, broadened applications, performance improvements, and future challenges

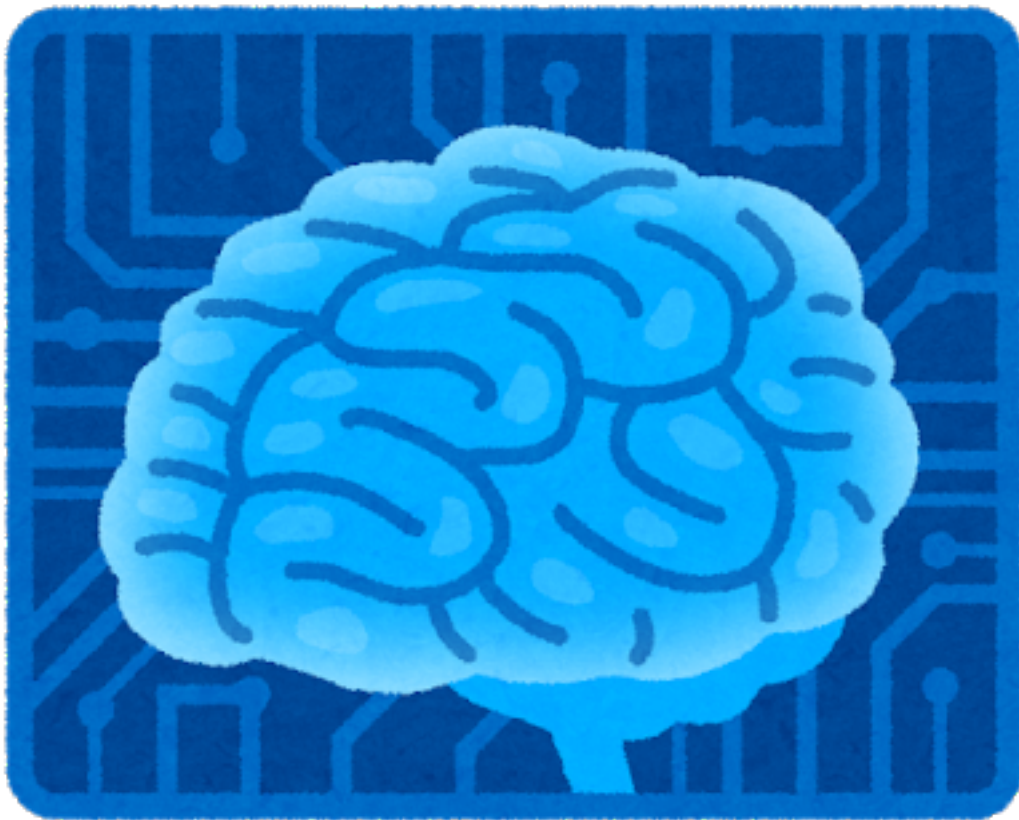


渡部 晋治

Language Technologies Institute
Carnegie Mellon University

第47回AIセミナー】「AIによる音声解析の最先端 ～音声
認識・音声合成・マルチモーダル処理を中心に～」

AI research in Japan



Country distribution of AI 2000 most influential scholars (from <https://www.aminer.cn/ai2000>)

Machine Learning	 81 Scholars	 10 Scholars	 3 Scholars	 2 Scholars	 2 Scholars
Computer Vision	 65 Scholars	 15 Scholars	 5 Scholars	 4 Scholars	 3 Scholars
Natural Language Processing	 80 Scholars	 8 Scholars	 4 Scholars	 2 Scholars	 1 Scholar
Robotics	 59 Scholars	 18 Scholars	 9 Scholars	 3 Scholars	 3 Scholars
Knowledge Engineering	 28 Scholars	 19 Scholars	 15 Scholars	 8 Scholars	 5 Scholars

Country distribution of AI 2000 most influential scholars (from <https://www.aminer.cn/ai2000>)

Machine Learning	 81 Scholars	 10 Scholars	 3 Scholars	 2 Scholars	 2 Scholars
Computer Vision	 65 Scholars	 15 Scholars	 5 Scholars	 4 Scholars	 3 Scholars
Natural Language Processing	 80 Scholars	 8 Scholars	 4 Scholars	 2 Scholars	 1 Scholar
Robotics	 59 Scholars	 18 Scholars	 9 Scholars	 3 Scholars	 3 Scholars
Knowledge Engineering	 28 Scholars	 19 Scholars	 15 Scholars	 8 Scholars	 5 Scholars
Speech Recognition	 57 Scholars	 10 Scholars	 5 Scholars	 5 Scholars	 5 Scholars



音声研究のすゝめ

- **日本は音声・音響分野で国際的に超強い**

- 先駆的研究 (LPC by 板倉, DP matching by 迫江・千葉)
- 国際学会をリード(主要会議であるInterspeechは日本人が作った!)
- ATR, NTT, 名工大などの国際的に有名な研究機関
- IEEE Speech and Language Processing Technical Committee (SLTC)
 - 音声分野における指導的国際的委員会
 - 六十人中**十人以上**が日本人！（2019年12月）
 - 日本で長年研究している（していた）外国人研究者も多数在籍
- 多くの世界的に有名な音声処理ソフトウェア(Julius, HTS)やデータベース

- **本トークでは、もう一つの日本発の音声音響研究における世界的な取り組みである、ESPnetプロジェクトを紹介**

Today's talk

- Introduction of ESPnet, end-to-end speech processing toolkit
- Broadened Applications
 - Automatic speech recognition (ASR)
 - Performance improvement
 - RNN-transducer
 - Non-autoregressive modeling
 - Text to speech (TTS)
 - Voice conversion
 - Speech translation
 - Speech enhancement

ESPnet  **ESPnet**, launched in December 2017

Our initial report at Interspeech 2018

ESPnet: End-to-End Speech Processing Toolkit

*Shinji Watanabe¹, Takaaki Hori², Shigeki Karita³, Tomoki Hayashi⁴, Jiro Nishitoba⁵, Yuya Unno⁶,
Nelson Enrique Yalta Soplin⁷, Jahn Heymann⁸, Matthew Wiesner¹, Nanxin Chen¹, Adithya
Renduchintala¹, Tsubasa Ochiai⁹,*

¹Johns Hopkins University, ²Mitsubishi Electric Research Laboratories, ³NTT Communication
Science Laboratories, ⁴Nagoya University, ⁵Retrieva, Inc., ⁶Preferred Networks, Inc., ⁷Waseda
University, ⁸Paderborn University, ⁹Doshisha University

shinjiw@jhu.edu

Abstract

This paper introduces a new open source platform for end-to-end speech processing named ESPnet. ESPnet mainly focuses on end-to-end automatic speech recognition (ASR), and adopts widely-used dynamic neural network toolkits, Chainer and PyTorch, as a main deep learning engine. ESPnet also follows the Kaldi ASR toolkit style for data processing, feature extraction/format, and recipes to provide a complete setup for speech

network [13, 14, 15, 16]. Attention-based methods use an attention mechanism to perform alignment between acoustic frames and recognized symbols, while CTC uses Markov assumptions to efficiently solve sequential problems by dynamic programming. ESPnet adopts hybrid CTC/attention end-to-end ASR [17], which effectively utilizes the advantages of both architectures in training and decoding. During training, we employ the multiobjective learning framework to improve robustness on irregular alignments and achieve fast convergence. During de-

Our initial report at Interspeech 2018

ESPnet: End-to-End Speech Processing Toolkit

*Shinji Watanabe¹, Takaaki Hori², Shigeki Karita³, Tomoki Hayashi⁴, Jiro Nishitoba⁵, Yuya Unno⁶,
Nelson Enrique Yalta Soplin⁷, Jahn Heymann⁸, Matthew Wiesner¹, Nanxin Chen¹, Adithya
Renduchintala¹, Tsubasa Ochiai⁹,*

¹Johns Hopkins University, ²Mitsubishi Electric Research Laboratories, ³NTT Communication
Science Laboratories, ⁴Nagoya University, ⁵Retrieva, Inc., ⁶Preferred Networks, Inc., ⁷Waseda
University, ⁸Paderborn University, ⁹Doshisha University

shinjiw@jhu.edu

Abstract

This paper introduces a new open source platform for end-to-end speech processing named ESPnet. ESPnet is an end-to-end automatic speech recognition toolkit that integrates widely-used dynamic neural network toolkit PyTorch, as a main deep learning engine. ESPnet also follows the Kaldi ASR toolkit style for data processing, feature extraction/format, and recipes to provide a complete setup for speech

Many contributions from researchers in Japan

lectures in training and decoding. During training, we employ the multiobjective learning framework to improve robustness on irregular alignments and achieve fast convergence. During de-

ESPnet  **ESPnet**, launched in December 2017

- **Open source** (Apache2.0) end-to-end speech processing toolkit
- Major concept
 - Accelerates end-to-end research for speech researchers
- Initially Chainer but later PyTorch based dynamic neural network toolkit as an engine
 - Easily develop novel neural network architecture
- Follows the famous speech recognition (Kaldi) style
 - Data processing, feature extraction/format
 - Recipes to provide a complete setup for speech processing experiments
- **The project is greatly accelerated in these three years**

Our latest report at DSLW 2021

THE 2020 ESPNET UPDATE: NEW FEATURES, BROADENED APPLICATIONS, PERFORMANCE IMPROVEMENTS, AND FUTURE PLANS

*Shinji Watanabe¹, Florian Boyer^{2,3}, Xuankai Chang¹, Pengcheng Guo^{4,1}, Tomoki Hayashi^{5,6},
Yosuke Higuchi⁷, Takaaki Hori⁸, Wen-Chin Huang⁶, Hirofumi Inaguma⁹, Naoyuki Kamo¹⁰,
Shigeki Karita¹¹, Chenda Li¹², Jing Shi¹³, Aswin Shanmugam Subramanian¹, Wangyou Zhang¹²*

¹Johns Hopkins University, ²Airudit, Speech Lab., ³LaBRI, Bordeaux INP, CNRS, UMR 5800,

⁴Northwestern Polytechnical University, ⁵Human Dataware Lab. Co., Ltd., ⁶Nagoya University

⁷Waseda University, ⁸MERL, ⁹Kyoto University, ¹⁰NTT Corporation, ¹¹Google

¹²Shanghai Jiao Tong University, ¹³Institute of Automation, Chinese Academy of Sciences

ABSTRACT

This paper describes the recent development of ESPnet (<https://github.com/espnet/espnet>), an end-to-end speech processing toolkit. This project was initiated in December 2017 to mainly deal with end-to-end speech recognition experiments based on sequence-to-sequence modeling. The project has grown rapidly and now covers a wide range of speech processing applications. Now ESPnet also includes text to speech (TTS), voice conversation (VC), speech translation (ST), and speech enhancement (SE) with support for beamforming, speech separation, denoising, and dere-

1.1. Related framework

There are a number of excellent deep learning frameworks that realize similar functions to what ESPnet covers, e.g., Fairseq [11], RETURNN [12], Lingvo [13], and NeMo [14]. These frameworks provide many AI applications, including various natural language processing (NLP) and speech processing methods, based on sequence-to-sequence modeling. Most frameworks include ASR, Text-to-Speech (TTS), and neural machine translation or speech translation (ST). Compared with them, ESPnet focuses more on a wide range of speech applications, and in addition to the above applications, ESPnet also supports various speech enhancement functions

Our latest report at DSLW 2021

THE 2020 ESPNET UPDATE: NEW FEATURES, BROADENED APPLICATIONS, PERFORMANCE IMPROVEMENTS, AND FUTURE PLANS

*Shinji Watanabe¹, Florian Boyer^{2,3}, Xuankai Chang¹, Pengcheng Guo^{4,1}, Tomoki Hayashi^{5,6},
Yosuke Higuchi⁷, Takaaki Hori⁸, Wen-Chin Huang⁶, Hirofumi Inaguma⁹, Naoyuki Kamo¹⁰,
Shigeki Karita¹¹, Chenda Li¹², Jing Shi¹³, Aswin Shanmugam Subramanian¹, Wangyou Zhang¹²*

¹Johns Hopkins University, ²Airudit, Speech Lab., ³LaBRI, Bordeaux INP, CNRS, UMR 5800,

⁴Northwestern Polytechnical University, ⁵Human Dataware Lab. Co., Ltd., ⁶Nagoya University

⁷Waseda University, ⁸MERL, ⁹Kyoto University, ¹⁰NTT Corporation, ¹¹Google

¹²Shanghai Jiao Tong University, ¹³Institute of Automation, Chinese Academy of Sciences

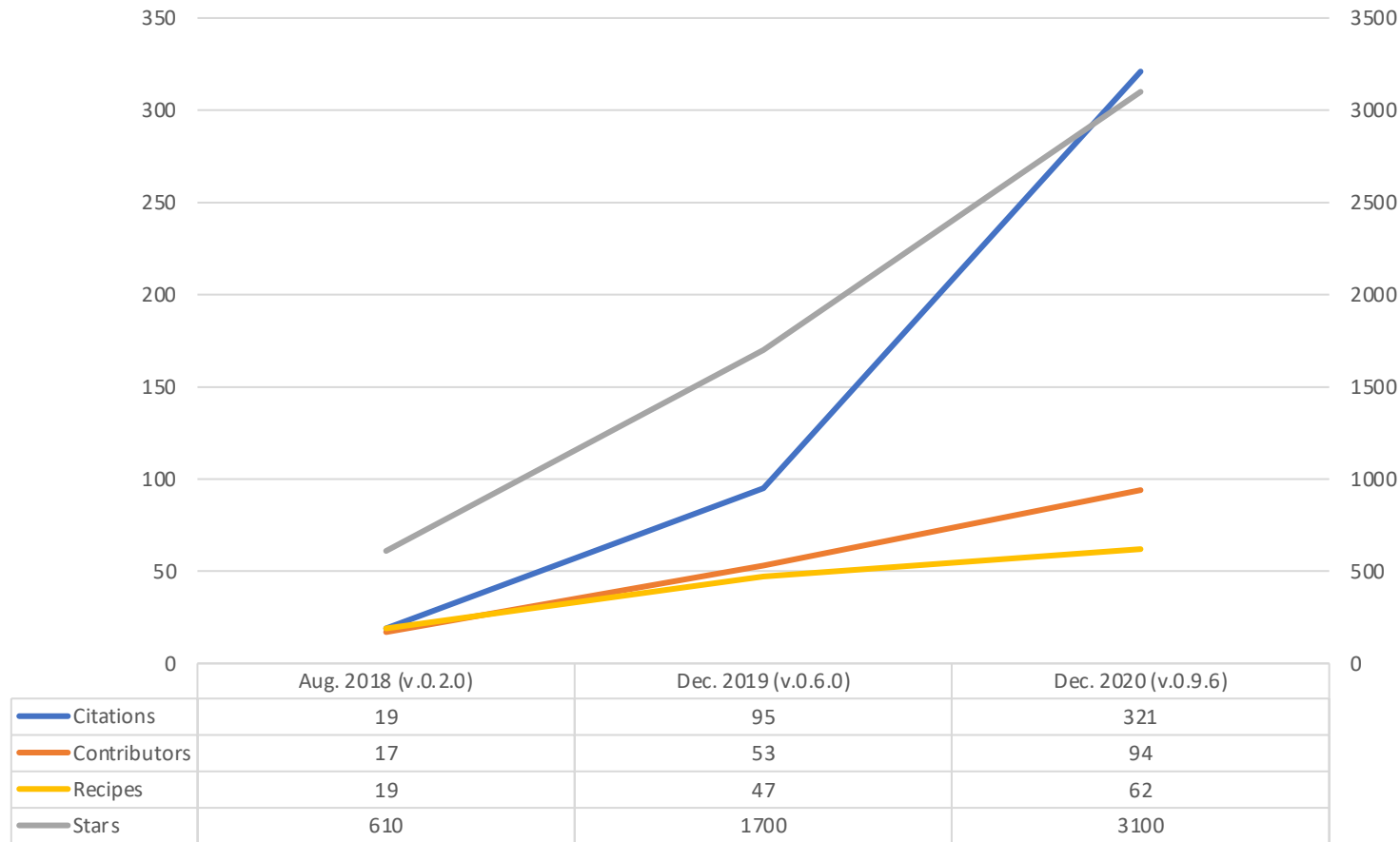
ABSTRACT

This paper describes the recent development of ESPnet (https://github.com/espnet/espnet), an end-to-end speech processing toolkit. This project was initiated by Shinji Watanabe and mainly deal with end-to-end speech recognition based on sequence-to-sequence modeling. The project has grown rapidly and now covers a wide range of speech processing applications. Now ESPnet also includes text to speech (TTS), voice conversation (VC), speech translation (ST), and speech enhancement (SE) with support for beamforming, speech separation, denoising, and dereverberation.

- Still many contributions from researchers in Japan
- The project becomes bigger and more international
- Today's main talk

to-sequence modeling. Most frameworks include ASR, Text-to-Speech (TTS), and neural machine translation or speech translation (ST). Compared with them, ESPnet focuses more on a wide range of speech applications, and in addition to the above applications, ESPnet also supports various speech enhancement functions

Activity statistics (from 2018 to 2020)



- **Citations**, **contributors**, **recipes** (examples), and **stars** are all growing

i.e.,

- Developers have increasingly supported the development of ESPnet
- has been used in various research groups and contributed a lot to speech research activities
- ESPnet 3.5K stars, Chainer 5.5K stars, Julius 1.3K stars

Major change in the internal framework

From ESPnet1 to ESPnet2

ESPnet2: a new system for DNN training to extend our system from v.0.7.0

Major update to deal with

- **Distributed training, on-the-fly feature extraction** from the raw waveform
- Improved the **scalability**
- Improved software workflow by enhancing the continuous integration, enriching documentation, supporting the docker, pip install, and model zoo functions.
- The migration is ongoing (ASR and TTS are already finished)

Today's talk

- Introduction of ESPnet, end-to-end speech processing toolkit
- **Broadened Applications**
 - Automatic speech recognition (ASR)
 - Performance improvement
 - RNN-transducer
 - Non-autoregressive modeling
 - Text to speech (TTS)
 - Voice conversion
 - Speech translation
 - Speech enhancement

Broadened Applications

- ESPnet (**ASR+X**) covers the following topics complementally



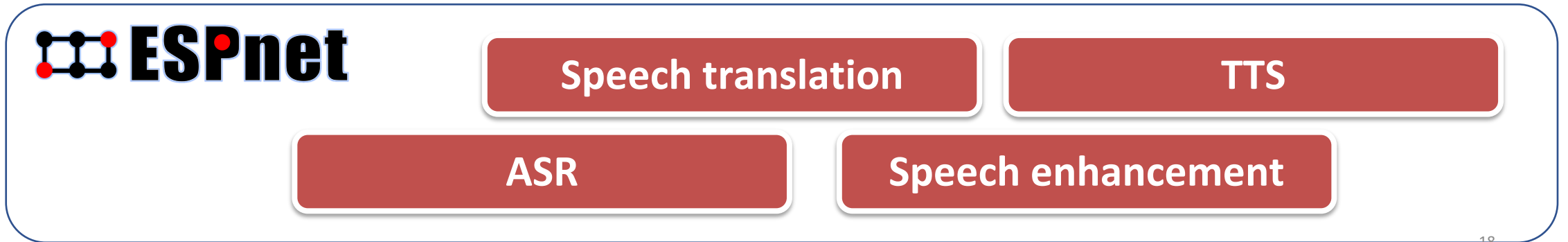
ASR

17

- Why can we support such wide-ranges of applications?

Broadened Applications

- ESPnet (**ASR+X**) covers the following topics complementally



18

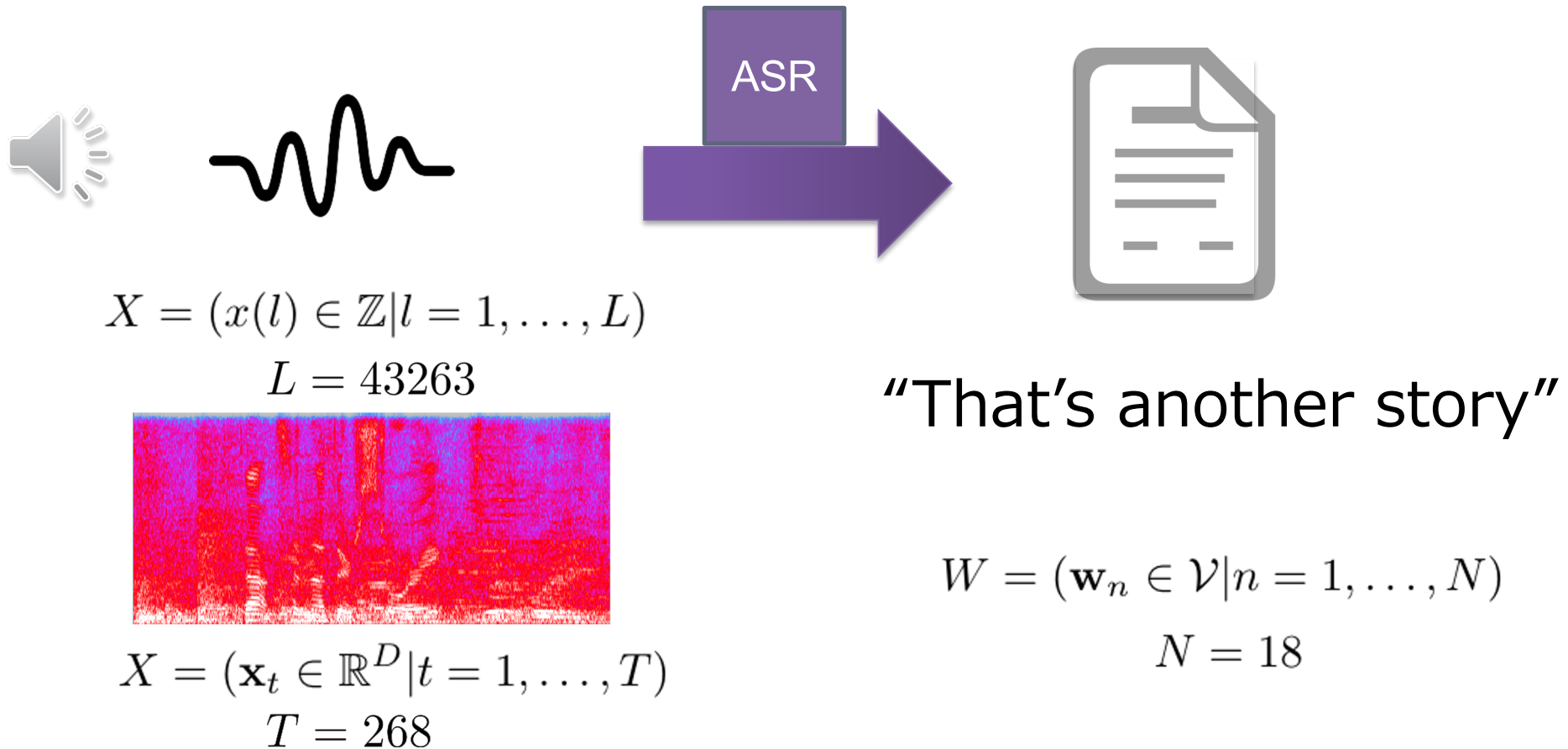
- Why can we support such wide-ranges of applications?

High-level benefit of e2e neural network

- **Unified** views of multiple speech processing applications based on end-to-end neural architecture
- **Integration** of these applications in a single network
- **Implementation** of such applications and their integrations based on an open-source toolkit like ESPnet, nemo, espresso, ctc++, fairseq, speechbrain, opennmt.py, lingvo, etc. etc., in a unified manner

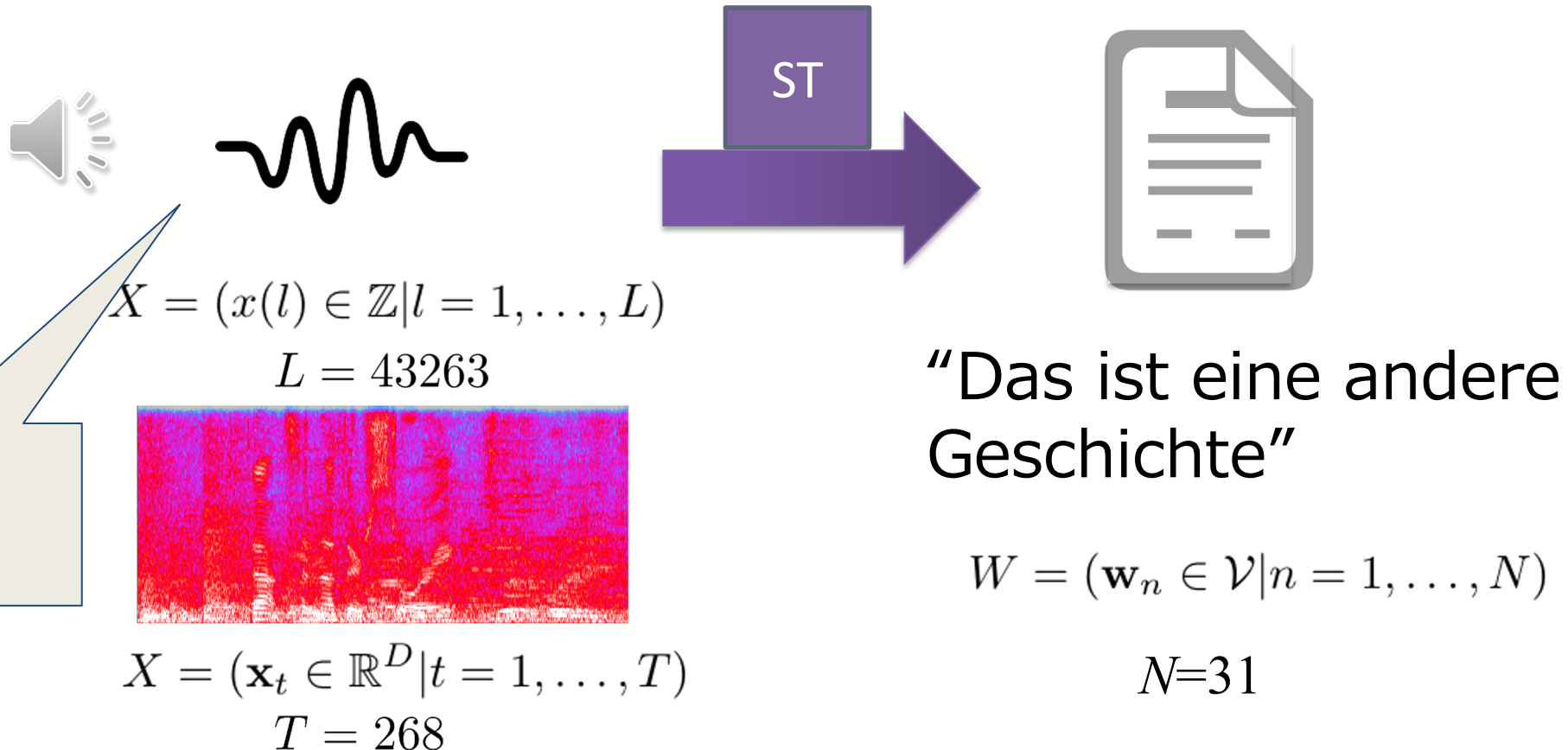
Automatic speech recognition (ASR)

- Mapping **speech** sequence to **character** sequence



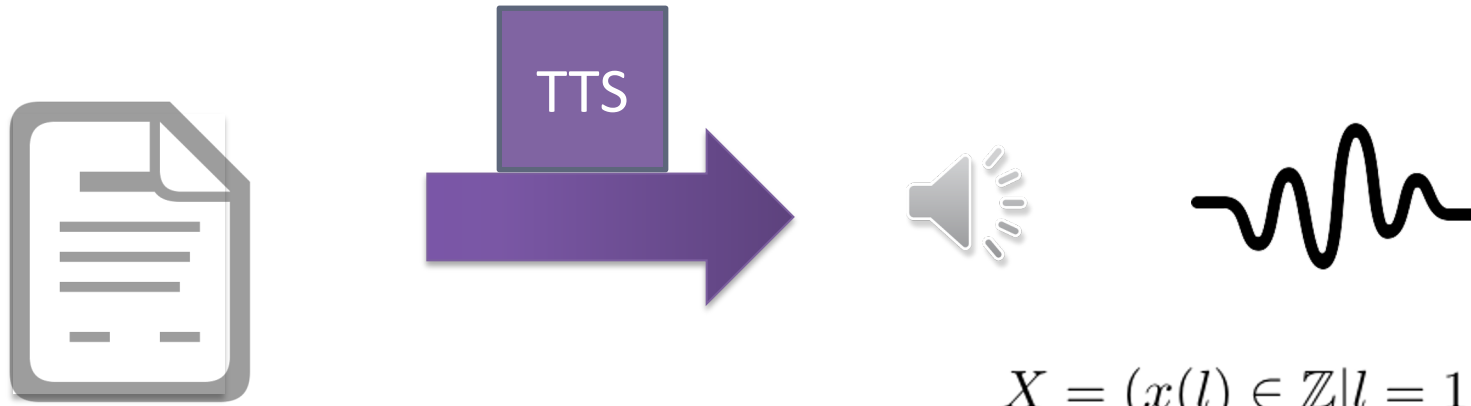
Speech to text translation (ST)

- Mapping **speech** sequence in a **source** language to **character** sequence in a **target** language

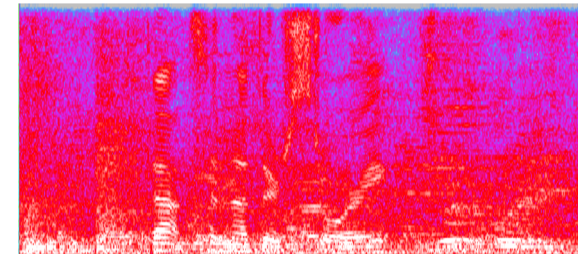


Text to speech (TTS)

- Mapping **character** sequence to **speech** sequence



$$X = (x(l) \in \mathbb{Z} | l = 1, \dots, L)$$
$$L = 43263$$



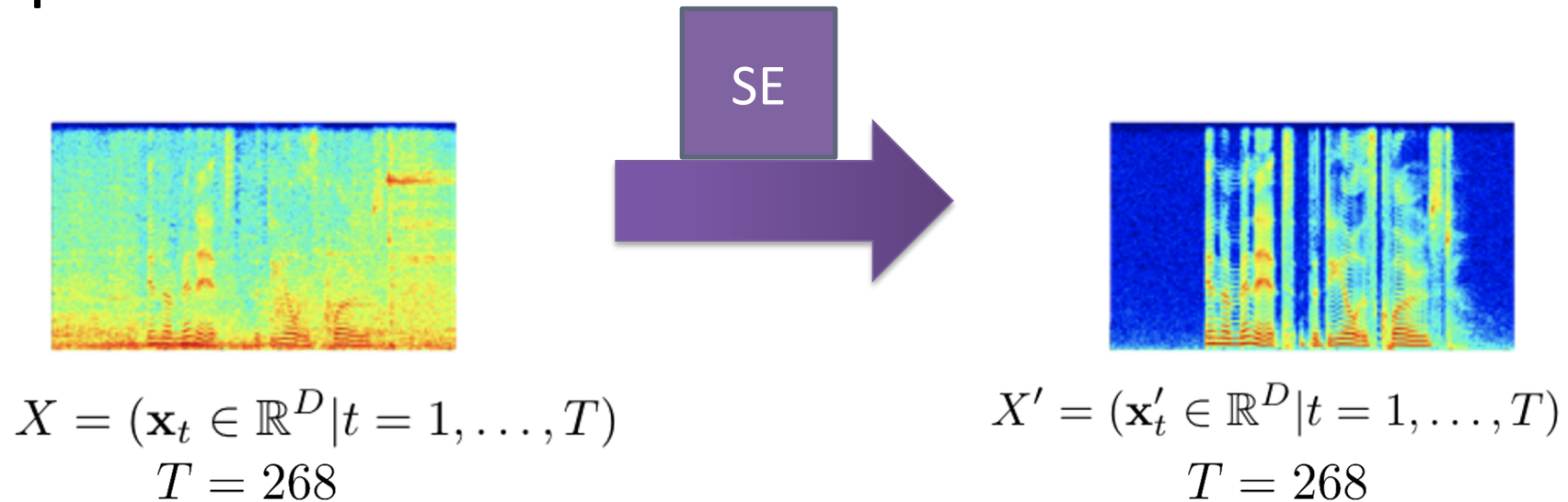
$$X = (\mathbf{x}_t \in \mathbb{R}^D | t = 1, \dots, T)$$
$$T = 268$$

“That’s another story”

$$W = (\mathbf{w}_n \in \mathcal{V} | n = 1, \dots, N)$$
$$N = 18$$

Speech enhancement (SE)

- Mapping **noisy** speech sequence to **clean** speech sequence



All problems are represented as

$$X = (x_1, x_2, \dots, x_T) \xrightarrow{f} Y = (y_1, y_2, \dots, y_N)$$

Unified view with sequence to sequence

- All the above problems: find a mapping function from *sequence to sequence* (**unification**)

$$X = (x_1, x_2, \dots, x_T) \xrightarrow{f} Y = (y_1, y_2, \dots, y_N)$$

- ASR: $X = \text{Speech}$, $Y = \text{Text}$
 - TTS: $X = \text{Text}$, $Y = \text{Speech}$
 - ST: $X = \text{Speech (EN)}$, $Y = \text{Text (JP)}$
 - Speech Enhancement: $X = \text{Noisy speech}$, $Y = \text{Clean speech}$
- Mapping function $f(\cdot)$
 - Sequence to sequence (seq2seq) function
 - ASR as an example

Seq2seq end-to-end ASR

$$X = (x_1, x_2, \dots, x_T) \xrightarrow{f} Y = (y_1, y_2, \dots, y_N)$$

Mapping seq2seq function $f(\cdot)$

1. Connectionist temporal classification (CTC)
2. Attention-based encoder decoder
3. Joint CTC/attention (Joint C/A)
4. RNN transducer (RNN-T)
5. Transformer

Unified view

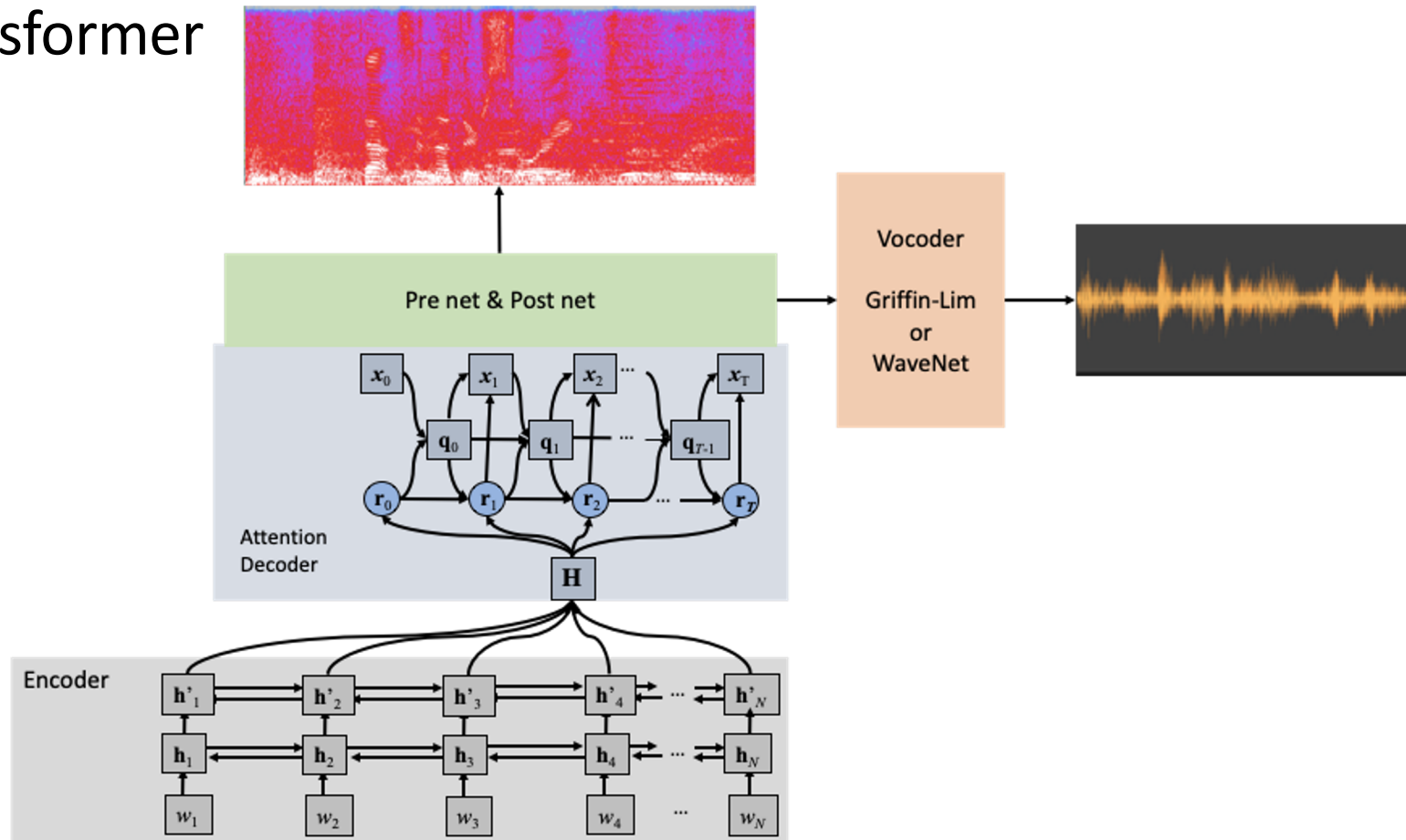
- Target speech processing problems: find a mapping function from *sequence* to *sequence* (**unification**)

$$X = (x_1, x_2, \dots, x_T) \xrightarrow{f} Y = (y_1, y_2, \dots, y_N)$$

- ASR: $X = \text{Speech}$, $Y = \text{Text}$
- TTS: $X = \text{Text}$, $Y = \text{Speech}$
- ...
- Mapping function (f)
 - Attention based encoder decoder
 - Transformer
 - ...

Seq2seq TTS (e.g., Tacotron2) [Shen+ 2018]

- Use seq2seq generate a spectrogram feature sequence
- We can use either attention-based encoder decoder or transformer



Unified view → Unified software design

We design a new speech processing toolkit based on

$$X = (x_1, x_2, \dots, x_T) \xrightarrow{f} Y = (y_1, y_2, \dots, y_N)$$

Unified view → Unified software design

We design a new speech processing toolkit based on

$$X = (x_1, x_2, \dots, x_T)$$

$$Y = (y_1, y_2, \dots, y_N)$$



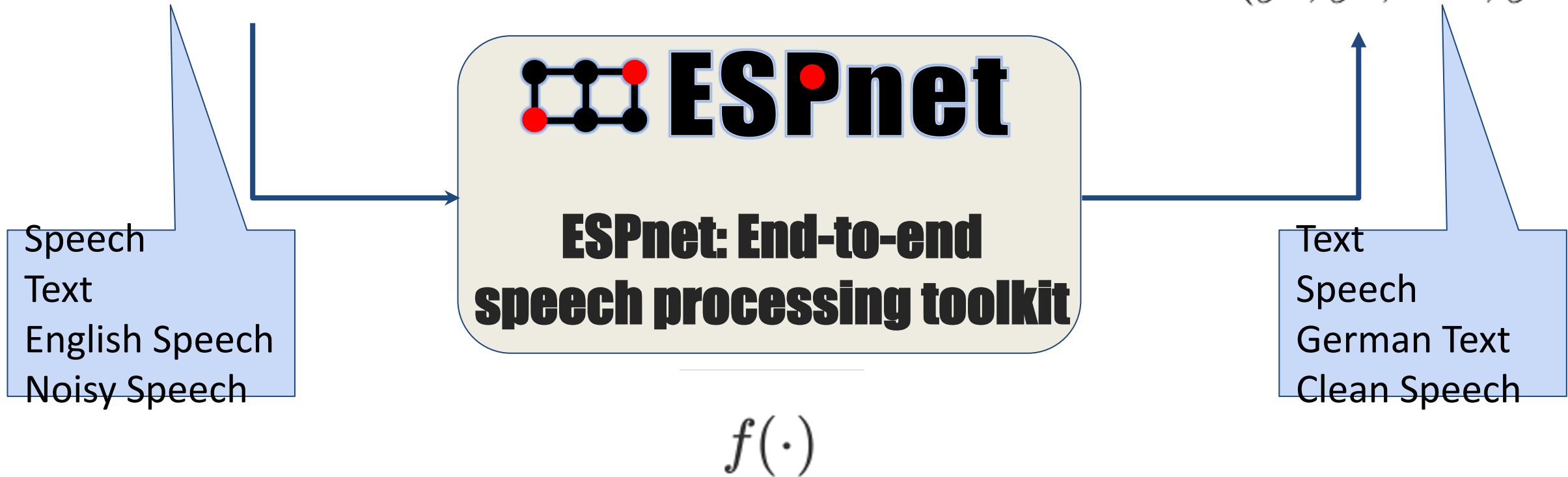
$$f(\cdot)$$

Unified view → Unified software design

We design a new speech processing toolkit based on

$$X = (x_1, x_2, \dots, x_T)$$

$$Y = (y_1, y_2, \dots, y_N)$$

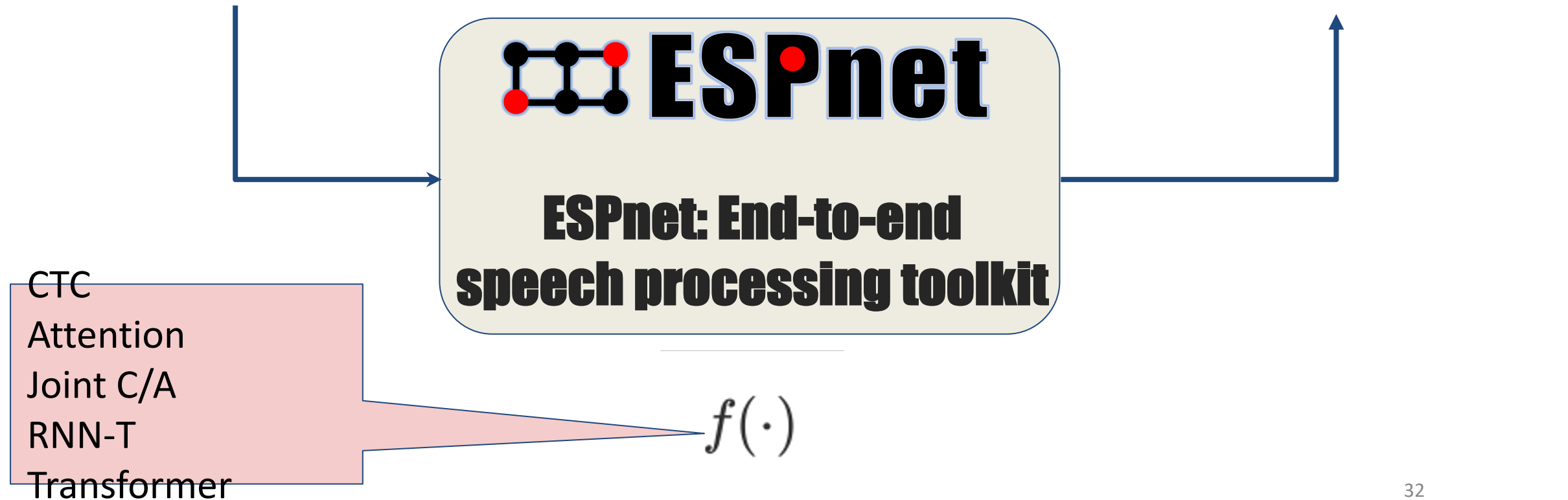


Unified view → Unified software design

We design a new speech processing toolkit based on

$$X = (x_1, x_2, \dots, x_T)$$

$$Y = (y_1, y_2, \dots, y_N)$$



Unified view → Unified software design

We design a new speech processing toolkit based on

$$X = (x_1, x_2, \dots, x_T)$$

$$Y = (y_1, y_2, \dots, y_N)$$



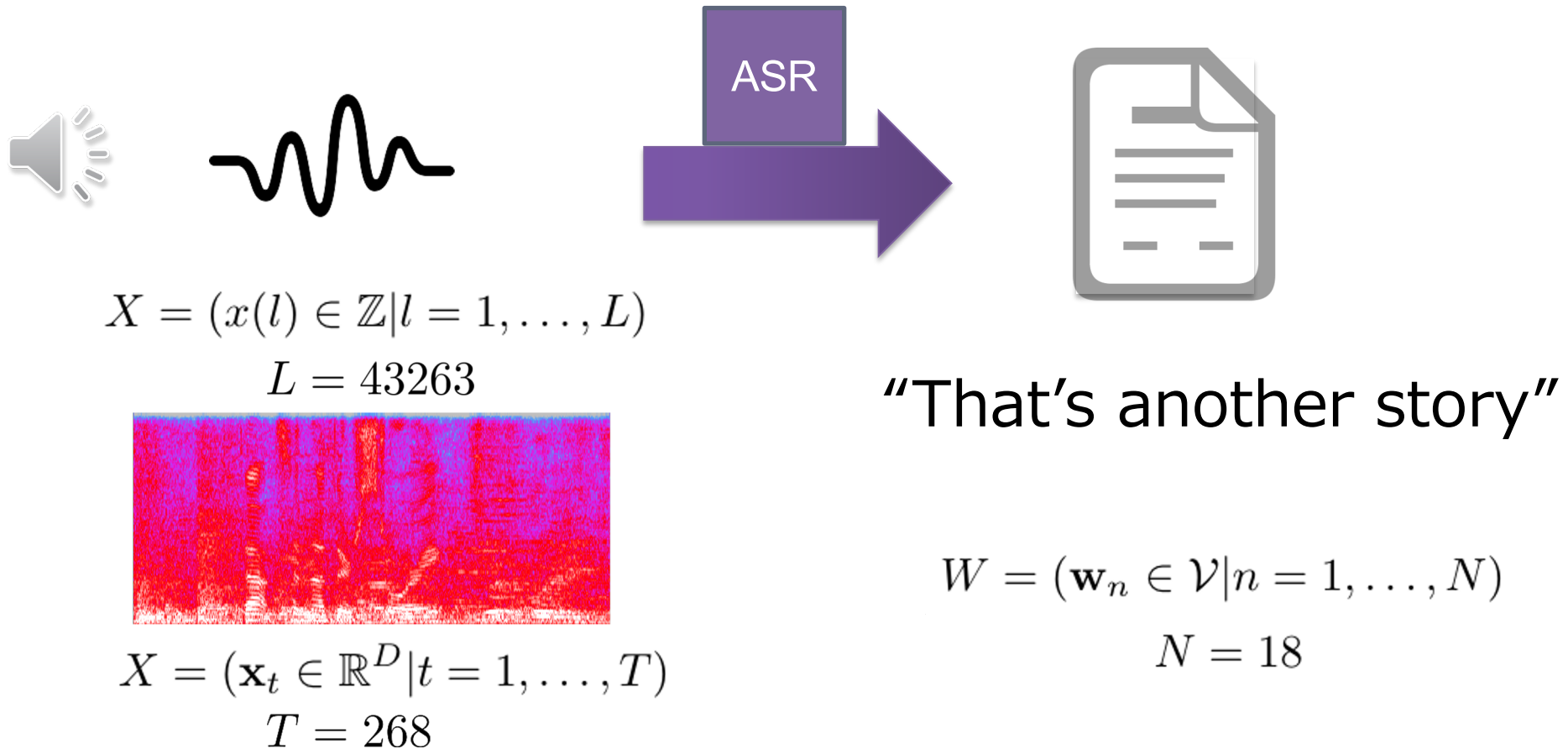
- Many speech processing applications can be **unified** based on seq2seq
- **Nemo, Fairseq, Lingvo, Espresso, SpeechBrain, Asteroid** and other toolkits also fully make use of these functions
- We are closely collaborating/interacting with them

Today's talk

- Introduction of ESPnet, end-to-end speech processing toolkit
- Applications
 - Automatic speech recognition (ASR)
 - Performance improvement
 - RNN-transducer
 - Non-autoregressive modeling
 - Text to speech (TTS)
 - Voice conversion
 - Speech translation
 - Speech enhancement

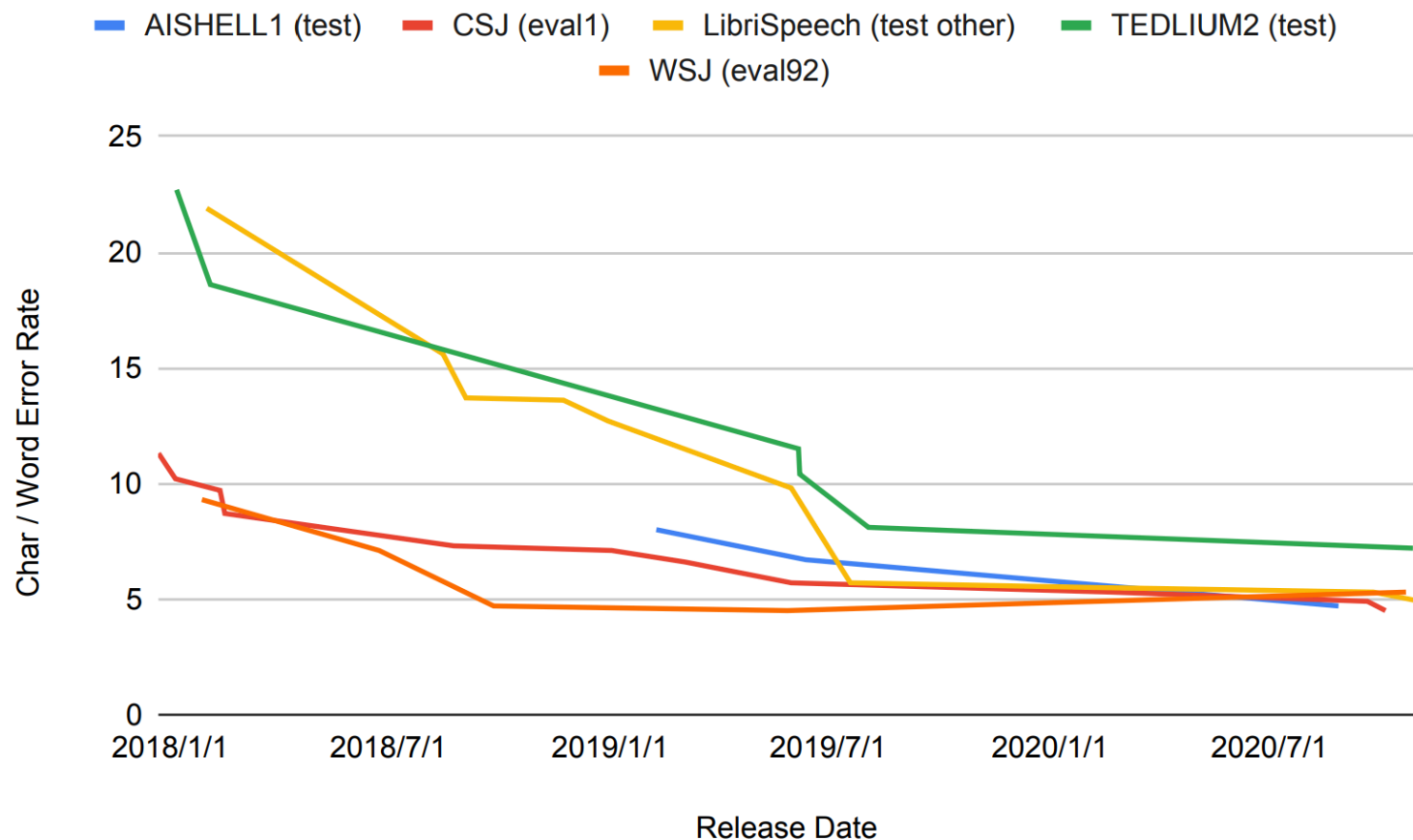
Automatic speech recognition (ASR)

- Mapping **speech** sequence to **character** sequence

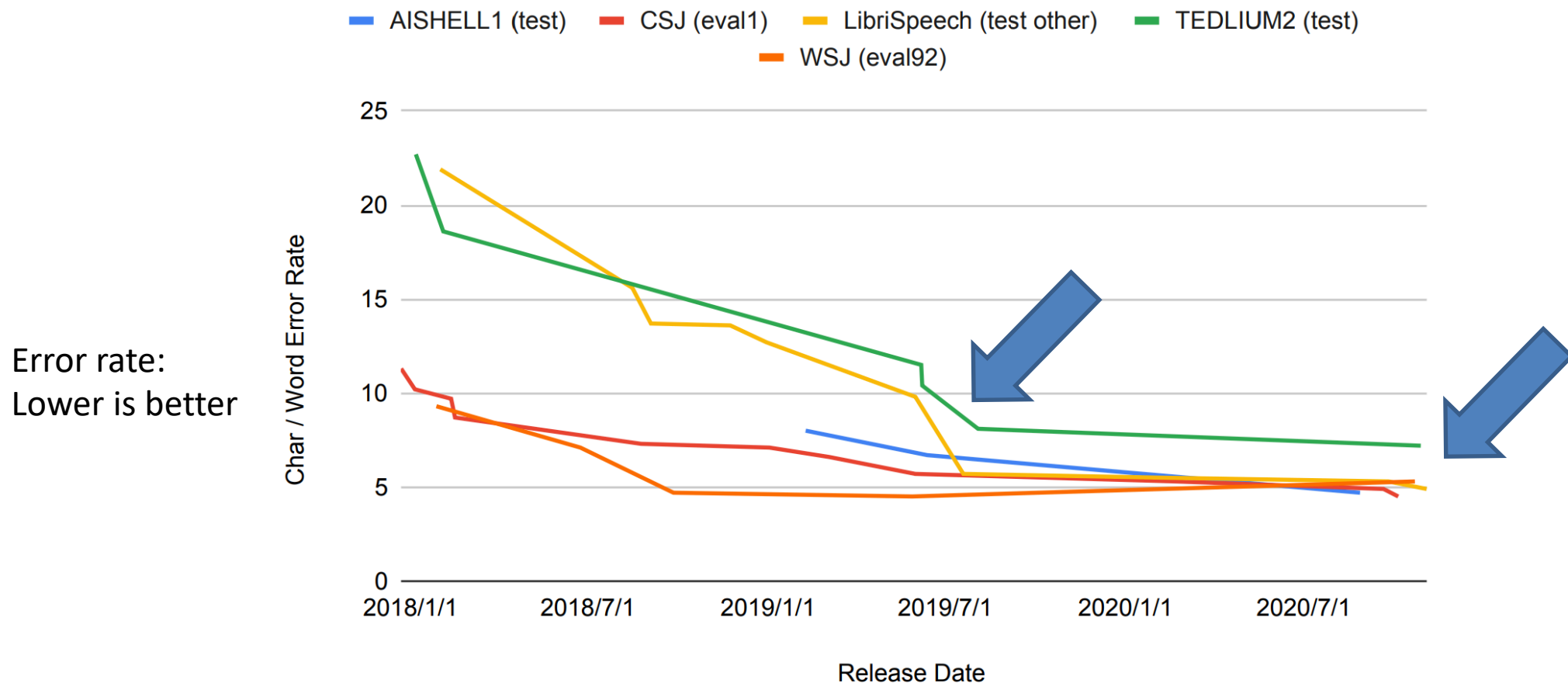


Maintaining state-of-the-art performance in ASR

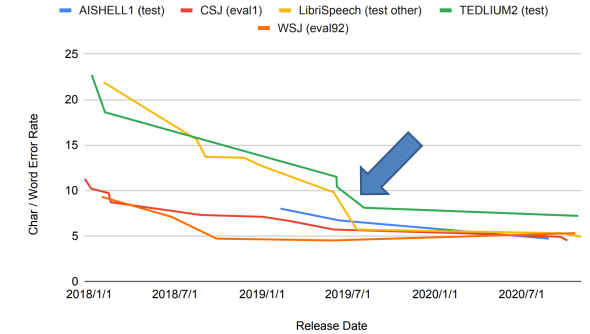
Error rate:
Lower is better



Maintaining state-of-the-art performance in ASR



ESPnet Transformer



A COMPARATIVE STUDY ON TRANSFORMER VS RNN IN SPEECH APPLICATIONS

*Shigeki Karita*¹,

(Alphabetical Order) *Nanxin Chen*³, *Tomoki Hayashi*^{5,6}, *Takaaki Hori*⁷, *Hirofumi Inaguma*⁸, *Ziyan Jiang*³,
*Masao Someki*⁵, *Nelson Enrique Yalta Soplin*², *Ryuichi Yamamoto*⁴, *Xiaofei Wang*³, *Shinji Watanabe*³,
Takenori Yoshimura^{5,6}, *Wangyou Zhang*⁹

¹NTT Communication Science Laboratories, ²Waseda University, ³Johns Hopkins University,

⁴LINE Corporation, ⁵Nagoya University, ⁶Human Dataware Lab. Co., Ltd.,

⁷Mitsubishi Electric Research Laboratories, ⁸Kyoto University, ⁹Shanghai Jiao Tong University

ABSTRACT

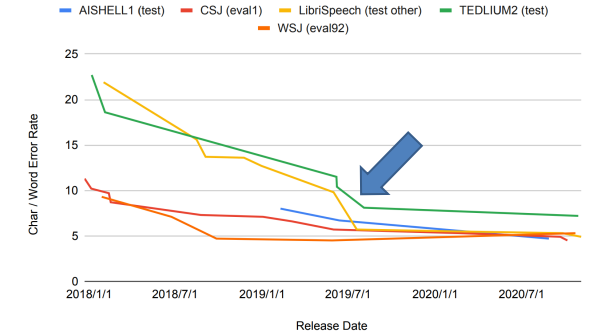
Sequence-to-sequence models have been widely used in end-to-end speech processing, for example, automatic speech recognition (ASR), speech translation (ST), and text-to-speech (TTS). This paper focuses on an emergent sequence-to-sequence model called Transformer, which achieves state-of-the-art performance in neural machine translation and other natural language processing applications. We undertook intensive studies in which we experimentally compared and analyzed Transformer and conventional recurrent

In our speech application experiments, we investigate several aspects of Transformer and RNN-based systems. For example, we measure the word/character/regression error from the ground truth, training curve, and scalability for multiple GPUs.

The contributions of this work are:

- We conduct a large-scale comparative study on Transformer and RNN with significant performance gains especially for the ASR related tasks.
- We explain our training tips for Transformer in speech applications: ASR, TTS and ST.

ESPnet Transformer



A COMPARATIVE STUDY ON TRANSFORMER VS RNN IN SPEECH APPLICATIONS

*Shigeki Karita*¹,

(Alphabetical Order) *Nanxin Chen*³, *Tomoki Hayashi*^{5,6}, *Takaaki Hori*⁷, *Hirofumi Inaguma*⁸, *Ziyan Jiang*³,
*Masao Someki*⁵, *Nelson Enrique Yalta Soplin*², *Ryuichi Yamamoto*⁴, *Xiaofei Wang*³, *Shinji Watanabe*³,
Takenori Yoshimura^{5,6}, *Wangyou Zhang*⁹

¹NTT Communication Science Laboratories, ²Waseda University, ³Johns Hopkins University,

⁴LINE Corporation, ⁵Nagoya University, ⁶Human Dataware Lab. Co., Ltd.,

⁷Mitsubishi Electric Research Laboratories, ⁸Kyoto University, ⁹Shanghai Jiao Tong University

ABSTRACT

Sequence-to-sequence models have been widely used in end speech processing, for example, automatic speech recognition (ASR), speech translation (ST), and text-to-speech synthesis. This paper focuses on an emergent sequence-to-sequence model, the Transformer, which achieves state-of-the-art performance in machine translation and other natural language processing applications. We undertook intensive studies in which we experimentally compared and analyzed Transformer and conventional recurrent

- One of the first success in the speech areas
- The performance was boosted

and RNN with significant performance gains especially for the ASR related tasks.

- We explain our training tips for Transformer in speech applications: ASR, TTS and ST

Transformer boosted the performance

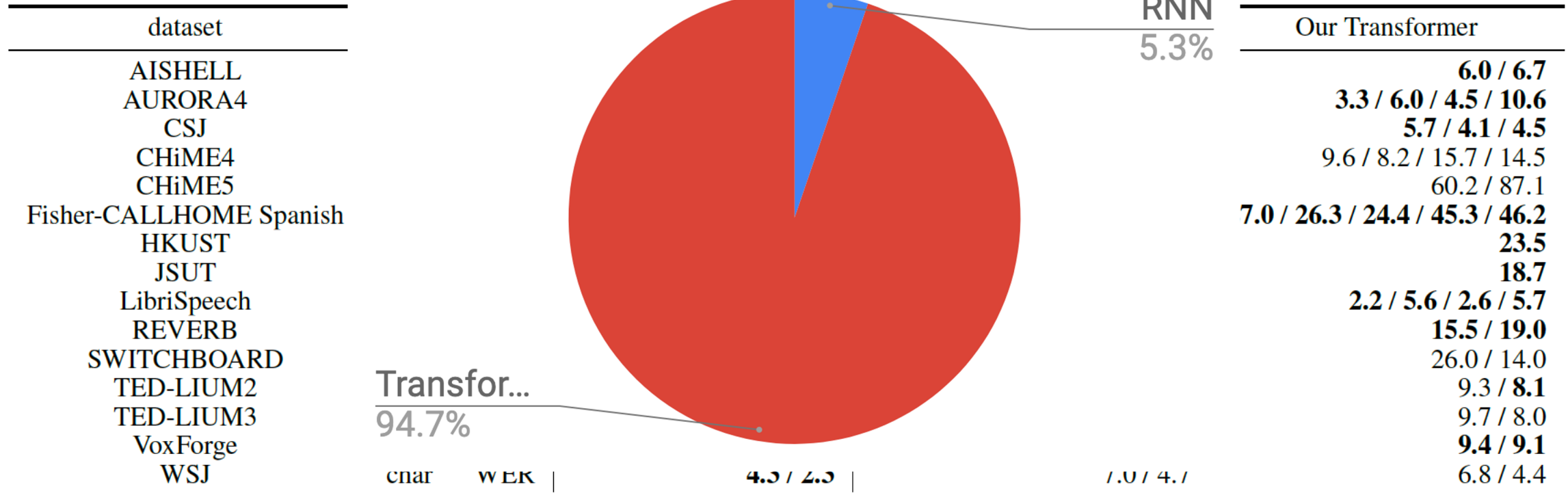
- Improve the performance from RNN with **13** ASR tasks among 15 tasks

dataset	token	error	Kaldi	Our RNN	Our Transformer
AISHELL	char	CER	N/A / 7.4	6.8 / 8.0	6.0 / 6.7
AURORA4	char	WER	(*) 3.6 / 7.7 / 10.0 / 22.3	3.5 / 6.4 / 5.1 / 12.3	3.3 / 6.0 / 4.5 / 10.6
CSJ	char	CER	(*) 7.5 / 6.3 / 6.9	6.6 / 4.8 / 5.0	5.7 / 4.1 / 4.5
CHiME4	char	WER	6.8 / 5.6 / 12.1 / 11.4	9.5 / 8.9 / 18.3 / 16.6	9.6 / 8.2 / 15.7 / 14.5
CHiME5	char	WER	47.9 / 81.3	59.3 / 88.1	60.2 / 87.1
Fisher-CALLHOME Spanish	char	WER	N/A	27.9 / 27.8 / 25.4 / 47.2 / 47.9	27.0 / 26.3 / 24.4 / 45.3 / 46.2
HKUST	char	CER	23.7	27.4	23.5
JSUT	char	CER	N/A	20.6	18.7
LibriSpeech	BPE	WER	3.9 / 10.4 / 4.3 / 10.8	3.1 / 9.9 / 3.3 / 10.8	2.2 / 5.6 / 2.6 / 5.7
REVERB	char	WER	18.2 / 19.9	24.1 / 27.2	15.5 / 19.0
SWITCHBOARD	BPE	WER	18.1 / 8.8	28.5 / 15.6	26.0 / 14.0
TED-LIUM2	BPE	WER	9.0 / 9.0	11.2 / 11.0	9.3 / 8.1
TED-LIUM3	BPE	WER	6.2 / 6.8	14.3 / 15.0	9.7 / 8.0
VoxForge	char	CER	N/A	12.9 / 12.6	9.4 / 9.1
WSJ	char	WER	4.3 / 2.3	7.0 / 4.7	6.8 / 4.4

Transformer boosted the performance

- Improve the performance from RNN with **13** ASR tasks among 15 tasks

RNN vs. Transformer



Experiments (~ 1000 hours)

Librispeech

Toolkit	dev_clean	dev_other	test_clean	test_other
Facebook wav2letter++	3.1	10.1	3.4	11.2
RWTH RASR	2.9	8.8	3.1	9.8
Nvidia Jasper	2.6	7.6	2.8	7.8
Google SpecAug.	N/A	N/A	2.5	5.8

- Very sensational results by Google

Experiments (~ 1000 hours)

Librispeech

Toolkit	dev_clean	dev_other	test_clean	test_other
Facebook wav2letter++	3.1	10.1	3.4	11.2
RWTH RASR	2.9	8.8	3.1	9.8
Nvidia Jasper	2.6	7.6	2.8	7.8
Google SpecAug.	N/A	N/A	2.5	5.8
ESPnet	2.2	5.6	2.6	5.7

- **Reached Google's best performance by community-driven efforts**



GAFAM



GAFAM



 **ESPnet**



Good example of “Collapetition”
= Collaboration + Competition

Experiments (~ **1000** hours) in August 2019

Librispeech

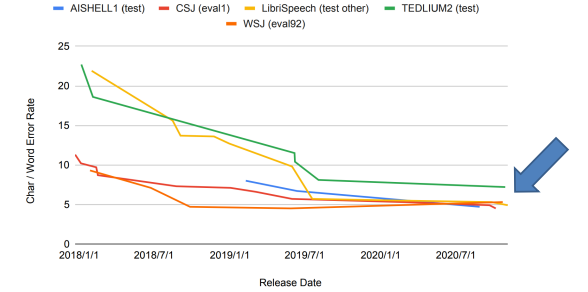
Toolkit	dev_clean	dev_other	test_clean	test_other
Facebook wav2letter++	3.1	10.1	3.4	11.2
RWTH RASR	2.9	8.8	3.1	9.8
Nvidia Jasper	2.6	7.6	2.8	7.8
Google SpecAug.	N/A	N/A	2.5	5.8
ESPnet	2.2	5.6	2.6	5.7

Experiments (~ **1000** hours) in March 2020

Librispeech

Toolkit	dev_clean	dev_other	test_clean	test_other
Facebook wav2letter++	3.1	10.1	3.4	11.2
RWTH RASR	2.9	8.8	3.1	9.8
Nvidia Jasper	2.6	7.6	2.8	7.8
Google SpecAug.	N/A	N/A	2.5	5.8
ESPnet	2.2	5.6	2.6	5.7
Google Conformer	N/A	N/A	1.9	3.9

ESPnet Conformer



RECENT DEVELOPMENTS ON ESPNET TOOLKIT BOOSTED BY CONFORMER

*Pengcheng Guo^{1,4}, Florian Boyer^{2,3}, Xuankai Chang⁴, Tomoki Hayashi⁵, Yosuke Higuchi⁶
Hirofumi Inaguma⁷, Naoyuki Kamo⁸, Chenda Li⁹, Daniel Garcia-Romero⁴, Jiatong Shi⁴
Jing Shi^{4,10}, Shinji Watanabe⁴, Kun Wei¹, Wangyou Zhang⁹, Yuekai Zhang⁴*

¹Northwestern Polytechnical University, ²LaBRI, University of Bordeaux, ³ Airudit

⁴Johns Hopkins University, ⁵Human Dataware Lab. Co., Ltd.

⁶Waseda University, ⁷Kyoto University, ⁸NTT Corporation ⁹Shanghai Jiao Tong University

¹⁰Institute of Automation, Chinese Academy of Sciences

ABSTRACT

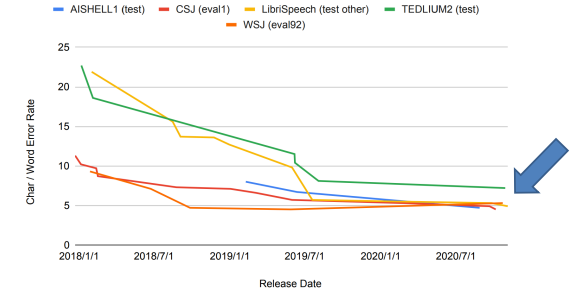
In this study, we present recent developments on ESPnet: End-to-End Speech Processing toolkit, which mainly involves a recently proposed architecture called Conformer, Convolution-augmented Transformer. This paper shows the results for a wide range of end-to-end speech processing applications, such as automatic speech recognition (ASR), speech translations (ST), speech separation (SS) and text-to-speech (TTS). Our experiments reveal various training tips and significant performance benefits obtained with the Conformer on different tasks. These results are competitive or even outperform the current state-of-art Transformer models. We are preparing to release all-in-one recipes using open source and publicly available corpora for all the above tasks with pre-trained

of publicly available corpora and try our best to share the practical guides (e.g., learning rate, hyper-parameters, network structure) on the use of Conformer. We also prepare to release the reproducible recipes and state-of-the-art setups to the community to succeed our exciting outcomes.

The contributions of this study include:

- We extend the Conformer architecture to various end-to-end speech processing applications and conduct comparative experiments with Transformer.
- We share our practical guides for the training of Conformer, like learning rate, kernel size of Conformer block, and model architectures, etc.

ESPnet Conformer



RECENT DEVELOPMENTS ON ESPNET TOOLKIT BOOSTED BY CONFORMER

*Pengcheng Guo^{1,4}, Florian Boyer^{2,3}, Xuankai Chang⁴, Tomoki Hayashi⁵, Yosuke Higuchi⁶
Hirofumi Inaguma⁷, Naoyuki Kamo⁸, Chenda Li⁹, Daniel Garcia-Romero⁴, Jiatong Shi⁴
Jing Shi^{4,10}, Shinji Watanabe⁴, Kun Wei¹, Wangyou Zhang⁹, Yuekai Zhang⁴*

¹Northwestern Polytechnical University, ²LaBRI, University of Bordeaux, ³ Airudit

⁴Johns Hopkins University, ⁵Human Dataware Lab. Co., Ltd.

⁶Waseda University, ⁷Kyoto University, ⁸NTT Corporation ⁹Shanghai Jiao Tong University

¹⁰Institute of Automation, Chinese Academy of Sciences

ABSTRACT

In this study, we present recent developments of the ESPnet End Speech Processing toolkit, which mainly introduces a proposed architecture called Conformer, Conformer-Transformer. This paper shows the results for various end-to-end speech processing applications, such as automatic speech recognition (ASR), speech translations (ST), speech-to-speech (S2S) and text-to-speech (TTS). Our experiments reveal interesting tips and significant performance benefits obtained with the Conformer on different tasks. These results are competitive or even outperform the current state-of-art Transformer models. We are preparing to release all-in-one recipes using open source and publicly available corpora for all the above tasks with pre-trained

of publicly available corpora and try our best to share the practical

- We try to follow Google's conformer work
- Also apply conformer to ST, TTS, as well as ASR

experiments with Transformer.

- We share our practical guides for the training of Conformer, like learning rate, kernel size of Conformer block, and model architectures, etc.

Experiments (~ **1000** hours) in October 2020

Librispeech

Toolkit	dev_clean	dev_other	test_clean	test_other
Facebook wav2letter++	3.1	10.1	3.4	11.2
RWTH RASR	2.9	8.8	3.1	9.8
Nvidia Jasper	2.6	7.6	2.8	7.8
Google SpecAug.	N/A	N/A	2.5	5.8
ESPnet	2.2	5.6	2.6	5.7
Google Conformer	N/A	N/A	1.9	3.9
ESPnet Conformer	1.9	4.6	2.1	4.7

We continue to work on catching up SOTA

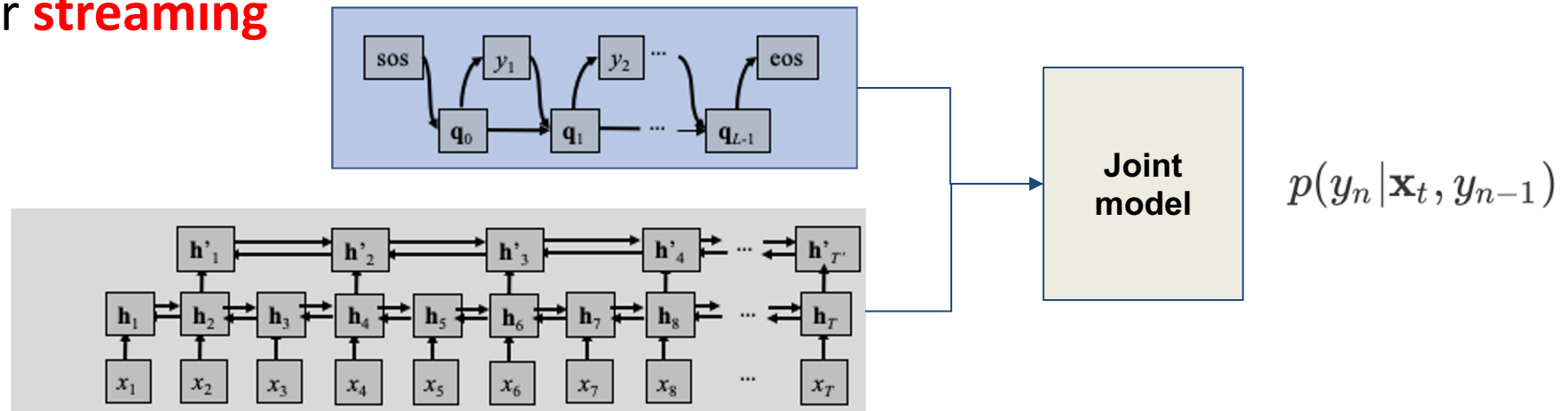
Today's talk

- Introduction of ESPnet, end-to-end speech processing toolkit
- Broadened Applications
 - Automatic speech recognition (ASR)
 - Performance improvement
 - RNN-transducer
 - Non-autoregressive modeling
 - Text to speech (TTS)
 - Voice conversion
 - Speech translation
 - Speech enhancement

RNN/Transformer Transducer [Boyer+(2021)]

- RNN or transformer transducer

- Good for **streaming**



- ESPnet has **various architecture supports** (LSTM, CNN, Transformer, conformer)
- Also supports **various beam search algorithms**
- We are now summarizing our efforts as a report, stay tuned!

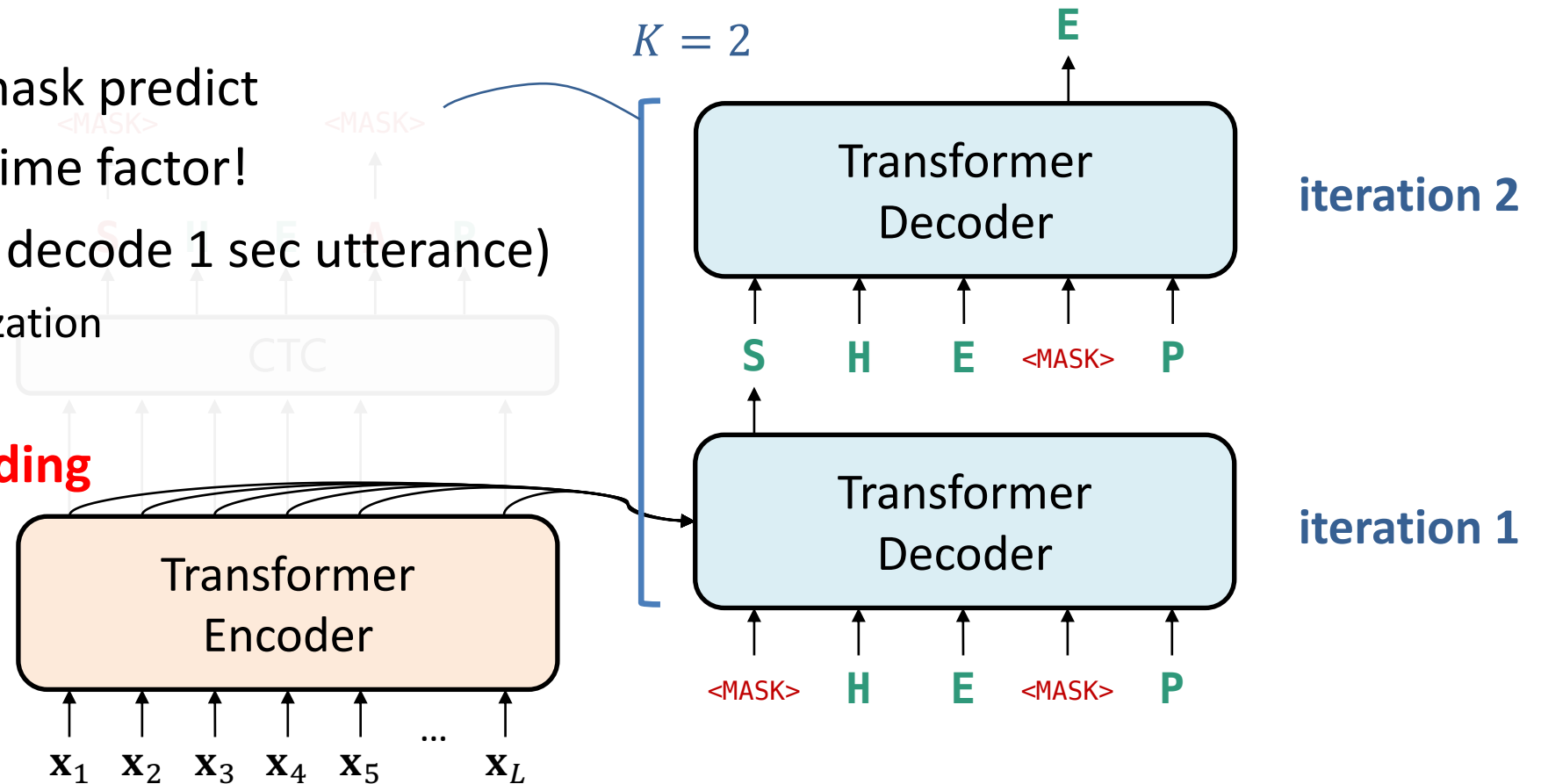
Non-Autoregressive modeling [Higuchi+(2020)]

- The most complicated part in ASR: **Left-to-right beam search (several hundreds of lines)**
- BERT-like iterative mask predict
- Achieved 0.07 real time factor!

(Takes only 70ms to decode 1 sec utterance)

- No software optimization
- Just CPU

- **Only 20 lines for coding**

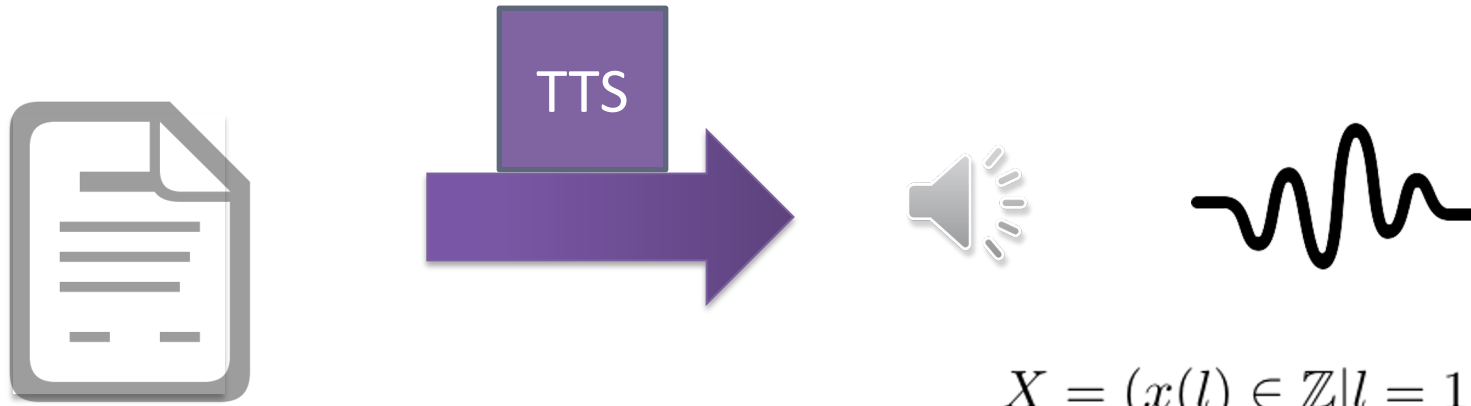


Today's talk

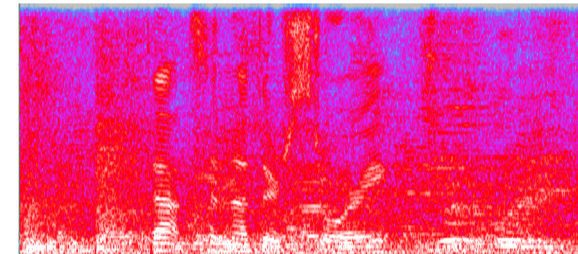
- Introduction of ESPnet, end-to-end speech processing toolkit
- Broadened Applications
 - Automatic speech recognition (ASR)
 - Performance improvement
 - RNN-transducer
 - Non-autoregressive modeling
 - Text to speech (TTS)
 - Voice conversion
 - Speech translation
 - Speech enhancement

Text to speech (TTS)

- Mapping **character** sequence to **speech** sequence



$$X = (x(l) \in \mathbb{Z} | l = 1, \dots, L)$$
$$L = 43263$$



$$X = (\mathbf{x}_t \in \mathbb{R}^D | t = 1, \dots, T)$$
$$T = 268$$

“That’s another story”

$$W = (\mathbf{w}_n \in \mathcal{V} | n = 1, \dots, N)$$
$$N = 18$$

ESPnet TTS

ESPNET-TTS: UNIFIED, REPRODUCIBLE, AND INTEGRATABLE OPEN SOURCE END-TO-END TEXT-TO-SPEECH TOOLKIT

*Tomoki Hayashi^{1,2}, Ryuichi Yamamoto³, Katsuki Inoue⁴, Takenori Yoshimura^{1,2},
Shinji Watanabe⁵, Tomoki Toda¹, Kazuya Takeda¹, Yu Zhang⁶, and Xu Tan⁷*

¹Nagoya University, ²Human Dataware Lab. Co., Ltd., ³LINE Corp.,

⁴Okayama University, ⁵Johns Hopkins University, ⁶Google AI, ⁷Microsoft Research

ABSTRACT

This paper introduces a new end-to-end text-to-speech (E2E-TTS) toolkit named ESPnet-TTS, which is an extension of the open-source speech processing toolkit ESPnet. The toolkit supports state-of-the-art E2E-TTS models, including Tacotron 2, Transformer TTS, and FastSpeech, and also provides recipes inspired by the Kaldi automatic speech recognition (ASR) toolkit. The recipes are based on the design unified with the ESPnet ASR recipe, providing high reproducibility. The toolkit also provides pre-trained models and samples of all of the recipes so that users can use it as a baseline. Furthermore, the unified design enables the integration of ASR functions with TTS, e.g., ASR-based objective evaluation and semi-supervised learning with both ASR and TTS models. This paper describes the

research purpose to make E2E-TTS systems more user-friendly and to accelerate research in this field. The toolkit not only supports state-of-the-art E2E-TTS models such as Tacotron 2 [6], Transformer TTS [8], and FastSpeech [9] but also provides Kaldi automatic speech recognition (ASR) toolkit [17] style recipes. The recipe is based on the design unified with the ASR recipe and includes all of the procedures required to reproduce the results. The toolkit provides a number of recipes for more than ten languages, which include single-speaker TTS as well as multi-speaker one and speaker adaptation. Pre-trained models and generated samples of all of the recipes are also provided so that users can easily use it as a baseline or perform TTS demonstrations. Furthermore, thanks to the unified design among TTS and ASR, we can easily integrate ASR functions with TTS, for example, ASR-based objective evaluation

ESPnet TTS

- Mainly focuses on the development of **text to mel-spectrogram** (text2mel) models.
- It supports Tacotron2 and Transformer-TTS (AR), and FastSpeech and FastSpeech2 (non-AR)
- **Multi-speaker extensions** with X-vector and global style token.
- Of course, we can easily switch to transformer or conformer
- Users can quickly develop the **state-of-the-art baseline systems** for the research purpose
- A lot of **examples and demonstration systems**, which works in **real-time** for various languages, including English, Mandarin, and Japanese

https://colab.research.google.com/github/espnet/notebook/blob/master/espnet2_tts_realtime_demo.ipynb

- Special thanks to Dr. Heiga Zen at Google for his valuable comments for this work!!!

▼ Synthesis

```
▶ # decide the input sentence by yourself
print(f"Input your favorite sentence in {lang}.")
x = input()

# synthesis
with torch.no_grad():
    start = time.time()
    wav, c, *_ = text2speech(x)
    wav = vocoder.inference(c)
    rtf = (time.time() - start) / (len(wav) / fs)
    print(f"RTF = {rtf:5f}")

# let us listen to generated samples
from IPython.display import display, Audio
display(Audio(wav.view(-1).cpu().numpy(), rate=fs))
```

download (1).wav

Input your favorite sentence in Japanese.
第47回AIセミナー
RTF = 0.024083

▶ 0:02 / 0:02 ———— 🔊 ⋮



Voice conversion challenge 2020 baseline system

The Sequence-to-Sequence Baseline for the Voice Conversion Challenge 2020: Cascading ASR and TTS

Wen-Chin Huang¹, Tomoki Hayashi¹, Shinji Watanabe², Tomoki Toda¹

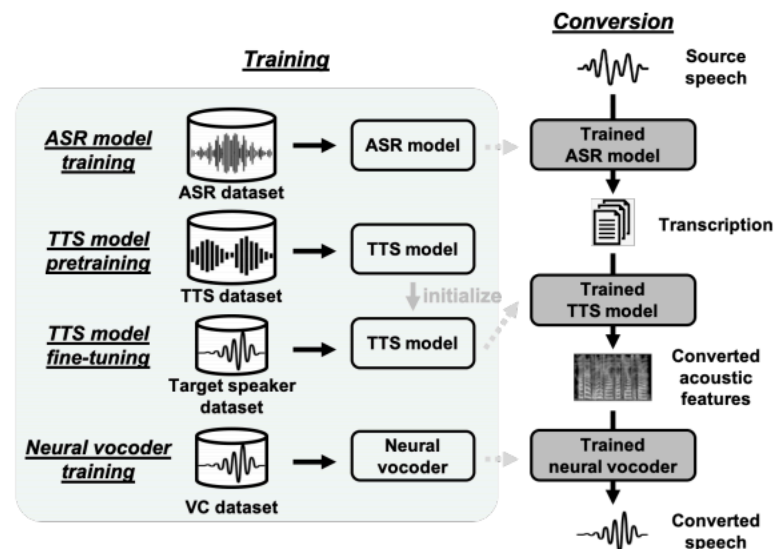
¹Nagoya University, Japan

²Johns Hopkins University, USA

wen.chinhuang@g.sp.m.is.nagoya-u.ac.jp

Abstract

This paper presents the sequence-to-sequence (seq2seq) baseline system for the voice conversion challenge (VCC) 2020. We consider a naive approach for voice conversion (VC), which is to first transcribe the input speech with an automatic speech recognition (ASR) model, followed using the transcriptions to generate the voice of the target with a text-to-speech (TTS) model. We revisit this method under a sequence-to-sequence (seq2seq) framework by utilizing ESPnet, an open-source end-to-end speech processing toolkit, and the many well-configured pretrained models provided by the community. Official evaluation results show that our system comes out top among the participating systems in terms of conversion similarity, demonstrating the promising ability of seq2seq models to convert speaker identity. The implementation is



Voice conversion challenge 2020 baseline system

The Sequence-to-Sequence Baseline for the Voice Conversion Challenge 2020: Cascading ASR and TTS

Wen-Chin Huang¹, Tomoki Hayashi¹, Shinji Watanabe², Tomoki Toda¹

¹Nagoya University, Japan

²Johns Hopkins University, USA

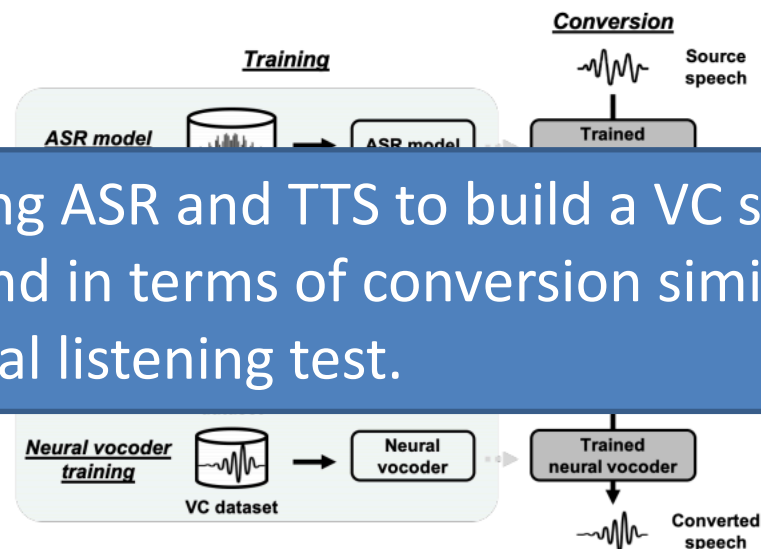
wen.chinhuang@g.sp.m.is.nagoya-u.ac.jp

Abstract

This paper presents the sequence-to-sequence (seq2seq) baseline system for the voice conversion challenge (VCC) 2020.

We consider a naive approach for voice conversion which is to first transcribe the input speech with an automatic speech recognition (ASR) model, followed by text-to-speech (TTS) synthesis using a text-to-speech (TTS) model. We revisit this method under the sequence (seq2seq) framework by utilizing the source end-to-end speech processing toolkit and well-configured pretrained models provided by the community. Official evaluation results show that our system comes out top among the participating systems in terms of conversion similarity, demonstrating the promising ability of seq2seq models to convert speaker identity. The implementation is

- Combining ASR and TTS to build a VC system
- Placed 2nd in terms of conversion similarity in the official listening test.

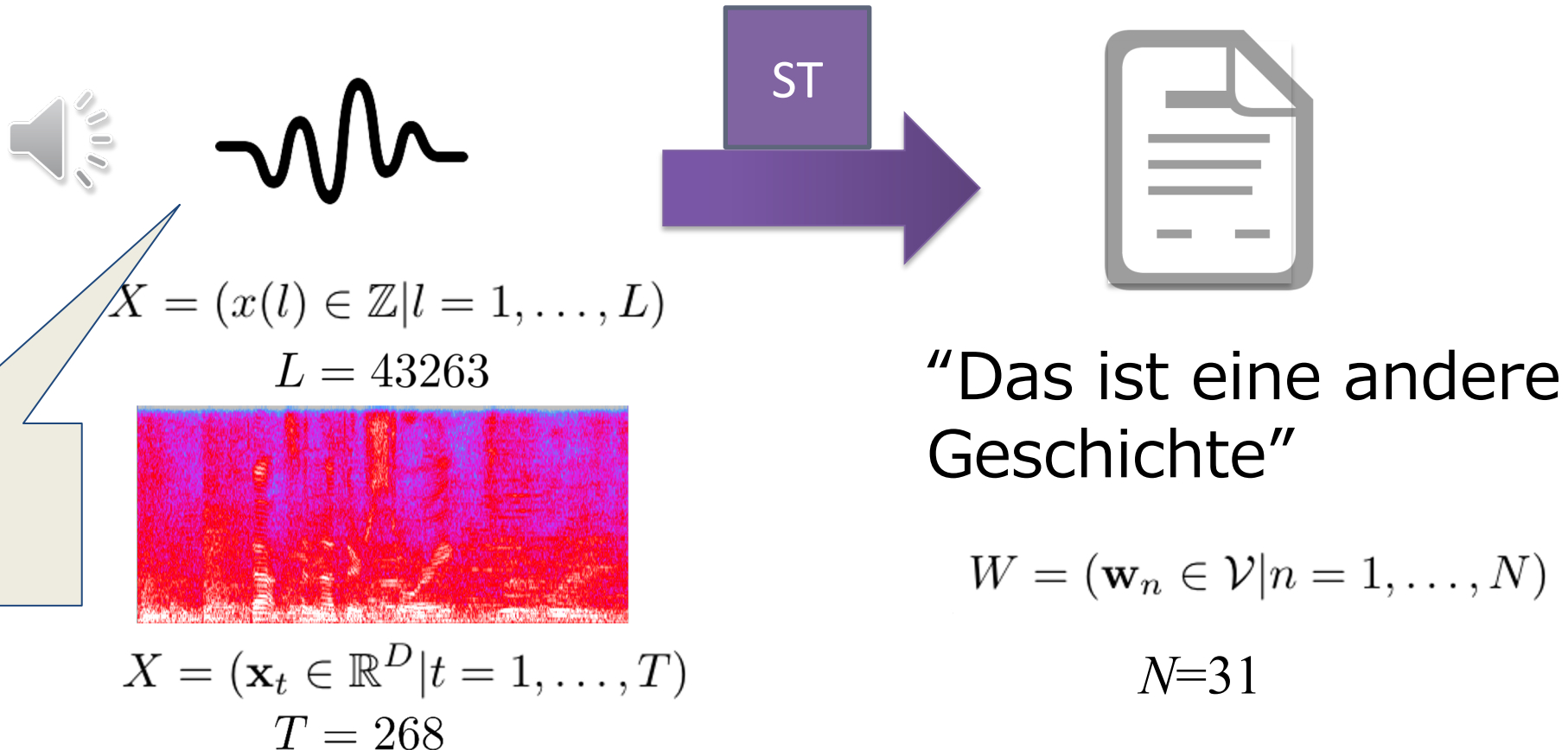


Today's talk

- Introduction of ESPnet, end-to-end speech processing toolkit
- Broadened Applications
 - Automatic speech recognition (ASR)
 - Performance improvement
 - RNN-transducer
 - Non-autoregressive modeling
 - Text to speech (TTS)
 - Voice conversion
 - **Speech translation**
 - Speech enhancement

Speech to text translation (ST)

- Mapping **speech** sequence in a **source** language to **character** sequence in a **target** language



ESPnet-ST

ESPnet-ST: All-in-One Speech Translation Toolkit

Hirofumi Inaguma¹ Shun Kiyono² Kevin Duh³ Shigeki Karita⁴

Nelson Yalta⁵ Tomoki Hayashi^{6,7} Shinji Watanabe³

¹ Kyoto University ² RIKEN AIP ³ Johns Hopkins University

⁴ NTT Communication Science Laboratories ⁵ Waseda University

⁶ Nagoya University ⁷ Human Dataware Lab. Co., Ltd.

inaguma@sap.ist.i.kyoto-u.ac.jp

Abstract

We present *ESPnet-ST*, which is designed for the quick development of speech-to-speech translation systems in a single framework. *ESPnet-ST* is a new project inside end-to-end speech processing toolkit ESPnet which

can reduce latency at inference time, which is useful for time-critical use cases like simultaneous interpretation. (2) A single model enables back-propagation training in an end-to-end fashion, which mitigates the risk of error propagation by cascaded modules. (3) In certain use cases

ESPnet-ST

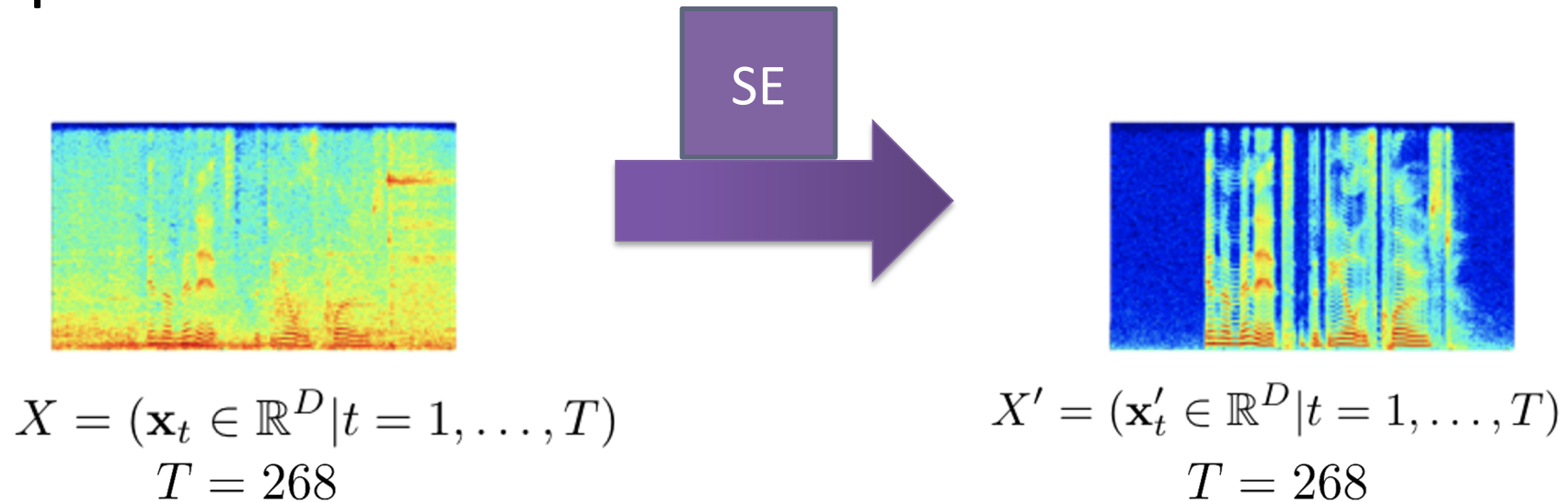
- Support the speech translation (ST) task with both the traditional **pipeline** approach (ASR + NMT) and **end-to-end (E2E)** approach
 - ESPnet also supports NMT and comparable performance to the other toolkit
- We demonstrated the **state-of-the-art translation performance** in standard ST benchmarks
 - MUST-C, IWSLT, Fisher Callhome Spanish
- Again, new progresses in ASR can be easily transferred to ST performance improvement, e.g., conformer, non-AR modeling

Today's talk

- Introduction of ESPnet, end-to-end speech processing toolkit
- Broadened Applications
 - Automatic speech recognition (ASR)
 - Performance improvement
 - RNN-transducer
 - Non-autoregressive modeling
 - Text to speech (TTS)
 - Voice conversion
 - Speech translation
 - **Speech enhancement**

Speech enhancement (SE)

- Mapping **noisy** speech sequence to **clean** speech sequence



ESPnet-SE

ESPNET-SE: END-TO-END SPEECH ENHANCEMENT AND SEPARATION TOOLKIT DESIGNED FOR ASR INTEGRATION

Chenda Li^{1}, Jing Shi^{2,3*}, Wangyou Zhang^{1*}, Aswin Shanmugam Subramanian³, Xuankai Chang³,
Naoyuki Kamo, Moto Hira⁴, Tomoki Hayashi^{5,6}, Christoph Boeddeker⁷, Zhuo Chen⁸, Shinji Watanabe³*

¹Shanghai Jiao Tong University, ²Institute of Automation, Chinese Academy of Sciences,

³Johns Hopkins University, ⁴Facebook AI, ⁵Nagoya University,

⁶Human Dataware Lab. Co., Ltd., ⁷Paderborn University, ⁸Microsoft Research

ABSTRACT

We present ESPnet-SE, which is designed for the quick development of speech enhancement and speech separation systems in a single framework, along with the optional downstream speech recognition module. ESPnet-SE is a new project which integrates rich automatic speech recognition related models, resources and systems to support and validate the proposed front-end implementation (i.e. speech enhancement and separation). It is capable of processing both single-channel and multi-channel data, with various functionalities including dereverberation, denoising and source separation. We provide all-in-one recipes including data pre-processing, feature extraction, training and evaluation pipelines for a wide range of benchmark tasks. This work is a joint effort of several institutions.

In this paper, we introduce a new E2E-SE toolkit named ESPnet-SE¹, which is an extension of the open-source speech processing toolkit ESPnet [8]. ESPnet-SE fully considers the various forms of speech input in the front-end scenes and meanwhile flexibly and organically integrated with the downstream automatic speech recognition (ASR) task, making it a user-friendly toolkit to easily build totally end-to-end robust ASR systems, even without need for clean speech signals. The toolkit provides adaptability to different speech data, including (1) single and multiple speakers, (2) single and multiple channels, (3) anechoic and reverberant conditions. Moreover, thanks to the ripe and efficient ASR modules in ESPnet, rich speech recognition related models, resources and systems can be optionally concatenated after the E2E-SE system, enabling evaluation and joint

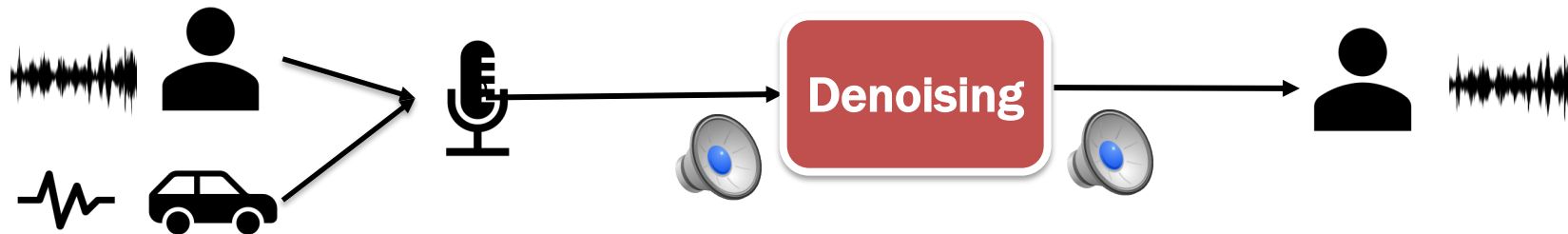
ESPnet-SE

- One of the biggest changes in ESPnet
- Include **ALL** speech enhancement functions

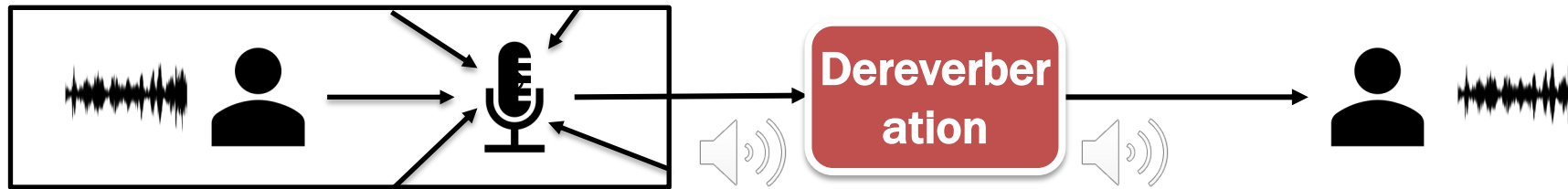
Speech enhancement

Several types of problems

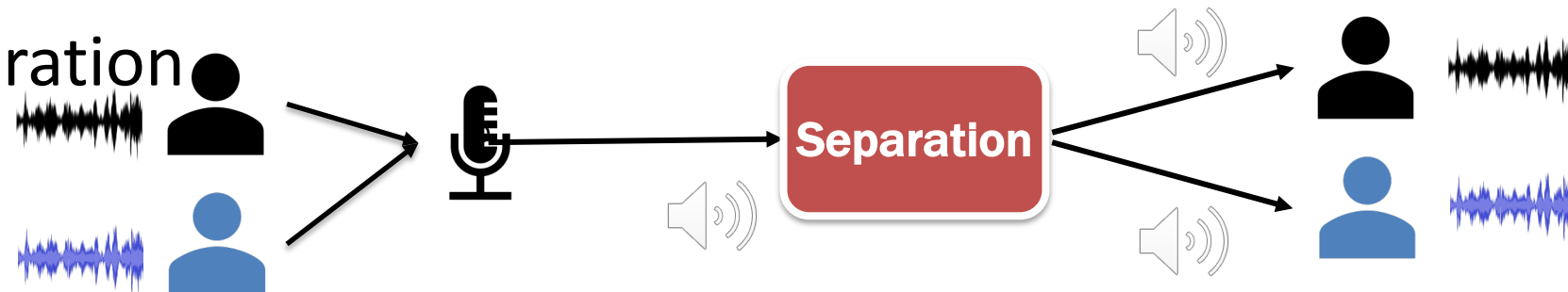
- Denoising (people mainly call it speech enhancement)



- Dereverberation



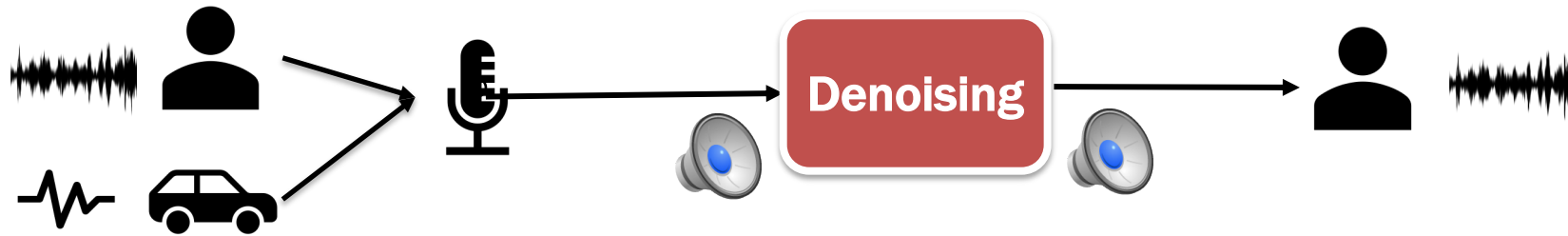
- Separation



Speech enhancement

Several types of problems

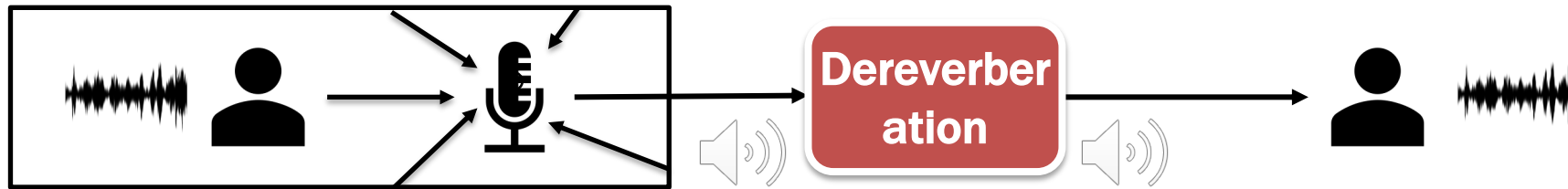
- Denoising (people mainly call it speech enhancement)



Speech enhancement

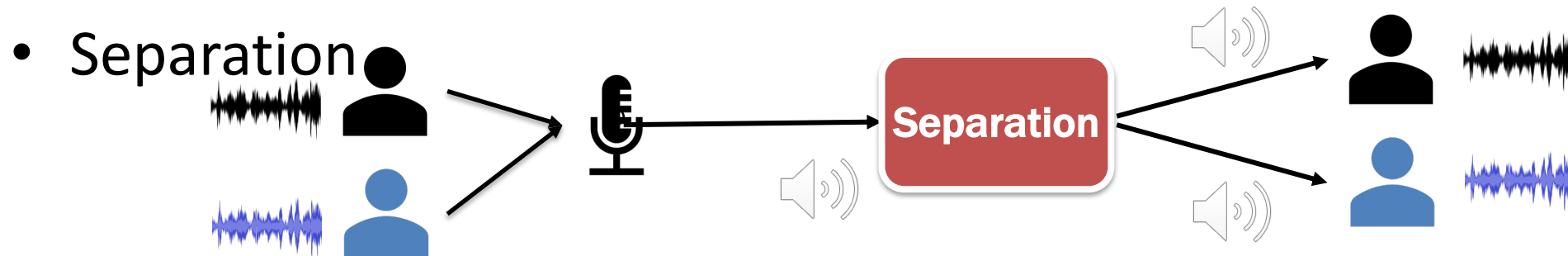
Several types of problems

- Dereverberation



Speech enhancement

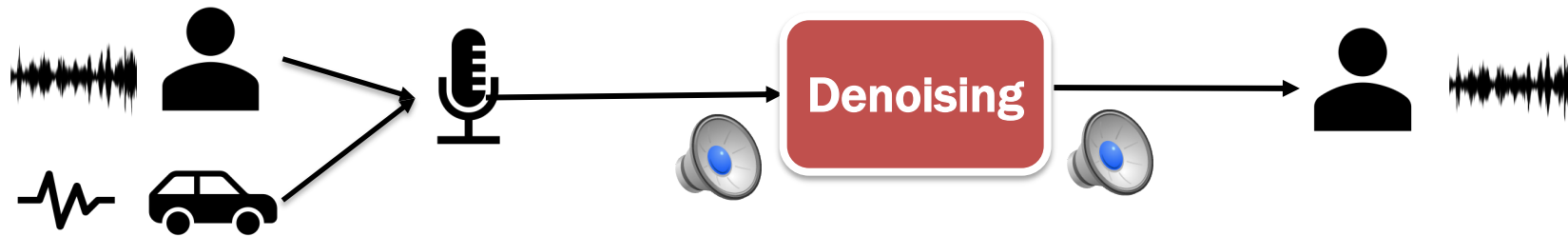
Several types of problems



Microphone array processing

Single to multiple microphones

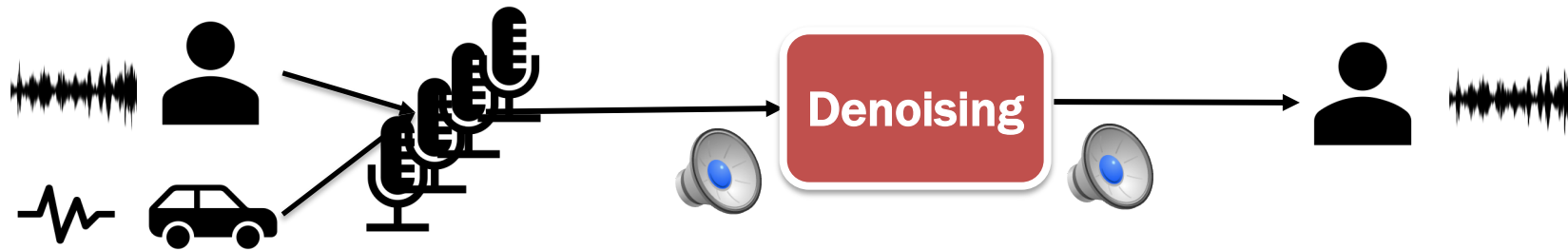
- Denoising (people mainly call it speech enhancement)



Microphone array processing

Single to multiple microphones

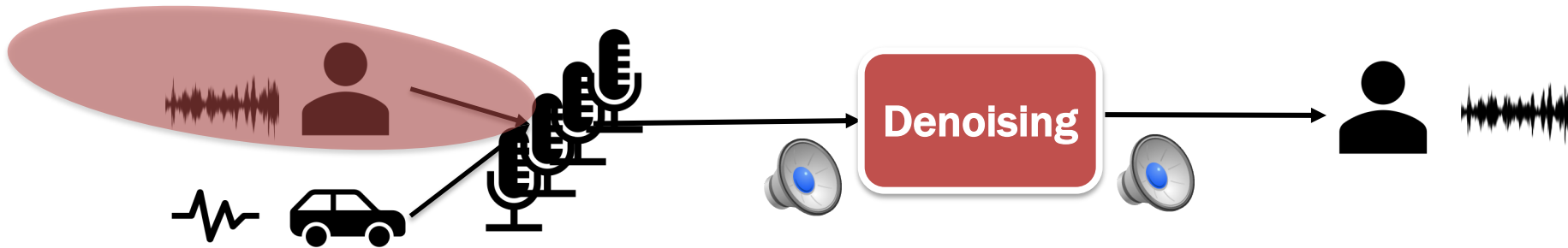
- Denoising (people mainly call it speech enhancement)



Microphone array processing

Single to multiple microphones

- Denoising (people mainly call it speech enhancement)

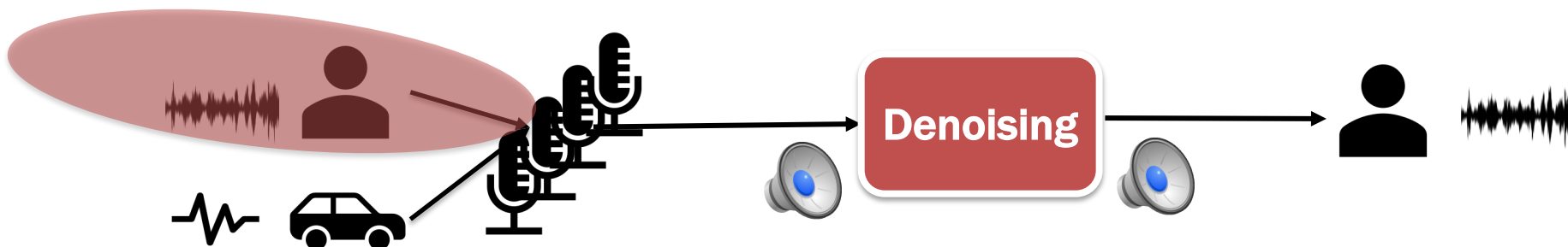


Make a spatial **beam** (beamforming)
to only pick up desired signals

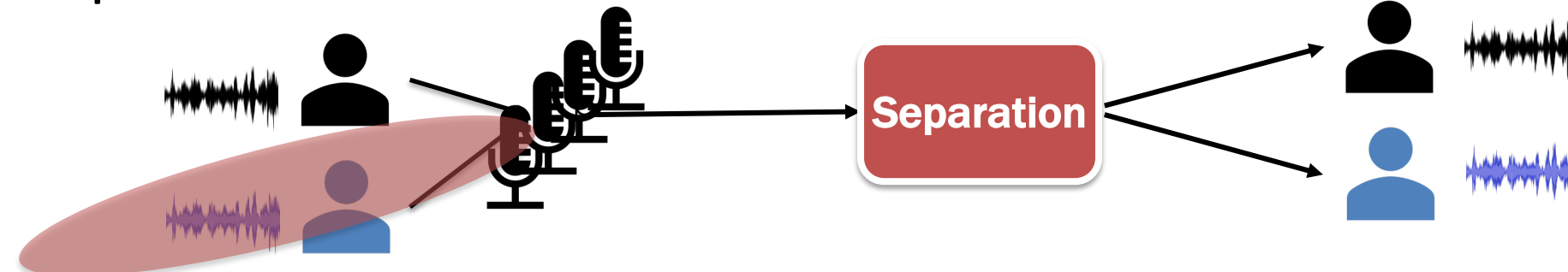
Microphone array processing

Single to multiple microphones

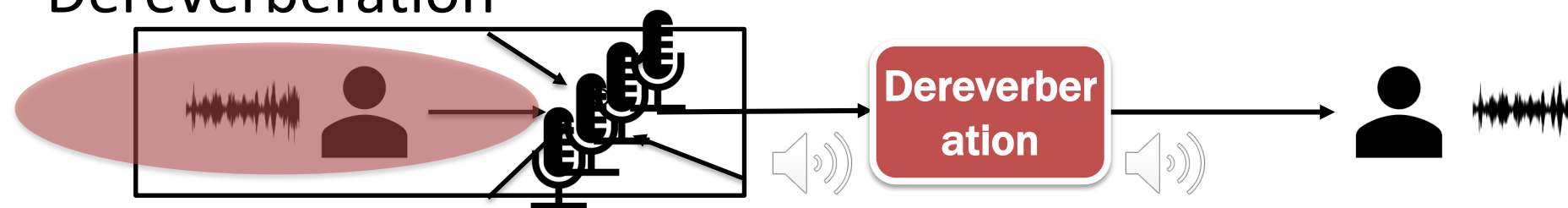
- Denoising (people mainly call it speech enhancement)



- Separation

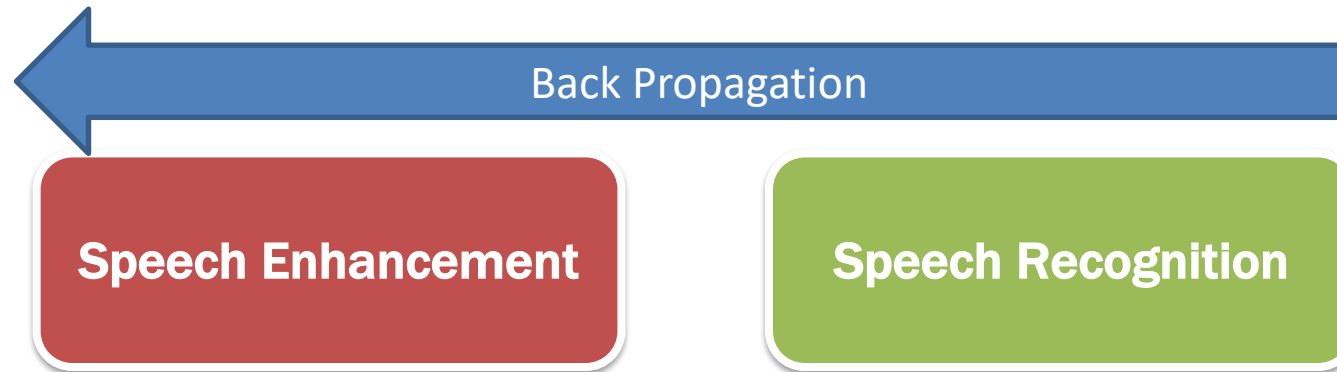


- Dereverberation



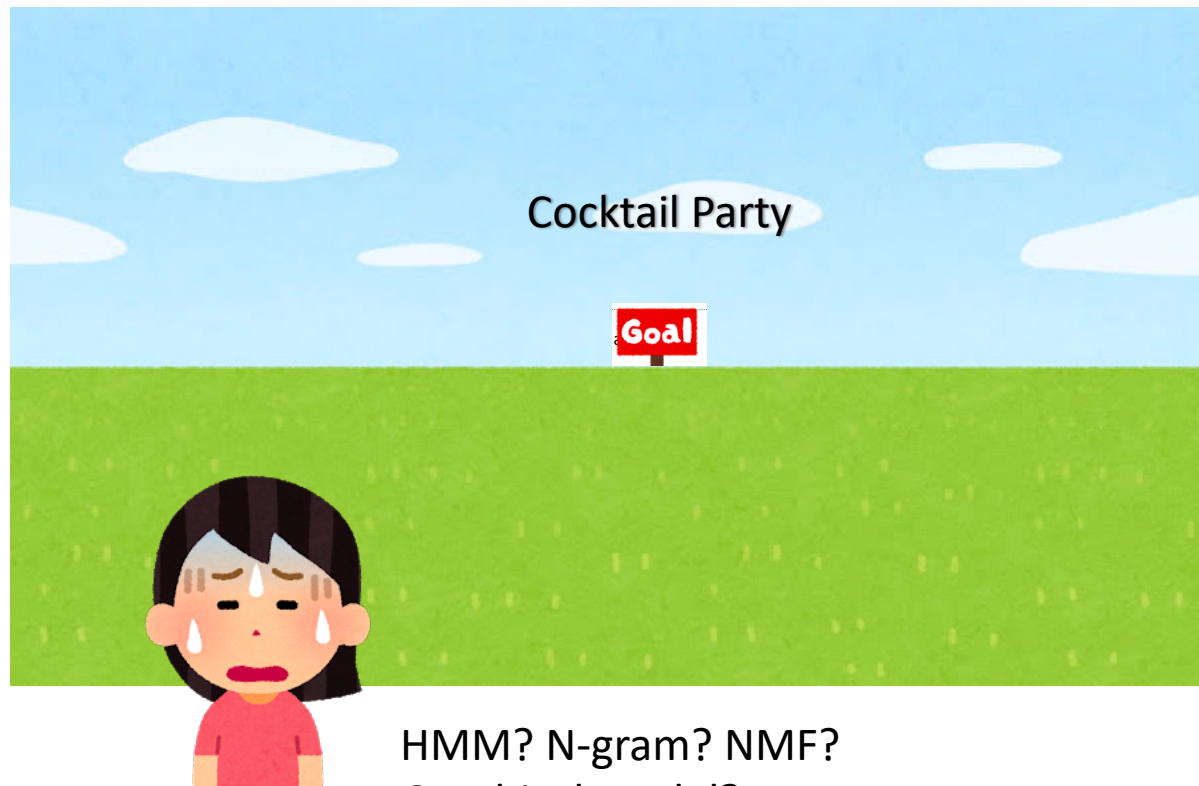
Differentiable speech enhancement frontend

- ESPnet SE can be used as an **independent** enhancement module
 - Denoising, Dereverberation, Separation
 - Single channel or multichannel
- ESPnet SE can be used as a **differentiable** enhancement module



- We can realize a cocktail party effect by a machine

- Cocktail Party is one of my first research topics when I started speech research
- I'm struggling how to tackle these issues for 20 years...
- I could not find a way...



HMM? N-gram? NMF?
Graphical model?
Bayesian? Discriminative?

Cocktail Party

ASR-TTS

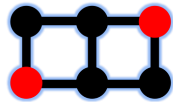
Goal

Now we have a way
to do!

Neural net
GPU
Open source
Great colleagues



Summary and future work

- End-to-end speech processing has a lot of potentials  **ESPnet**
- ESPnet provides **state-of-the-art** and **reproducible** research
- Future work
 - More applications, e.g., speaker diarization (initial version is already included), audio event detection, speaker verification
 - Real-time/streaming applications (including RNN transducer)
 - Continues to follow the state-of-the art performance

**We are working on this project for the
community contribution!!!**



- 求む！共同開発者！！！！
- どんな形の貢献でも結構です
- 日本から世界に発信しましょう！
- まずは使ってみてください！！！！

<https://github.com/espnet/espnet>

- 興味のある方は遠慮なく
shinjiw@ieee.org までご連絡を！！

Thanks a lot!!!