Carnegie Mellon University
Language Technologies Institute

Is Everything End-to-End?

Shinji Watanabe
Language Technologies Institute
Carnegie Mellon University

Sphinx Speech Lunch, September 1st, 2022

Sphinx Speech Lunch

- This is an **official start** of the Sphinx Speech Lunch in **2022F**!
- **Biweekly, Thursday 12:30 – 1:30pm**
  - Then, we can eat a lunch and continue a fun discussion **until 2:00pm** (or more!)
  - Pizza will be served around **the end of the talk**
- Please contact Yifan Peng yifanpen@andrew.cmu.edu if you're interested in presenting your work
  - However, we already fixed the speaker line up in the fall semester.. But don't worry! We will also have it in the spring semester (and next year, and forever!)

Sphinx Speech Lunch

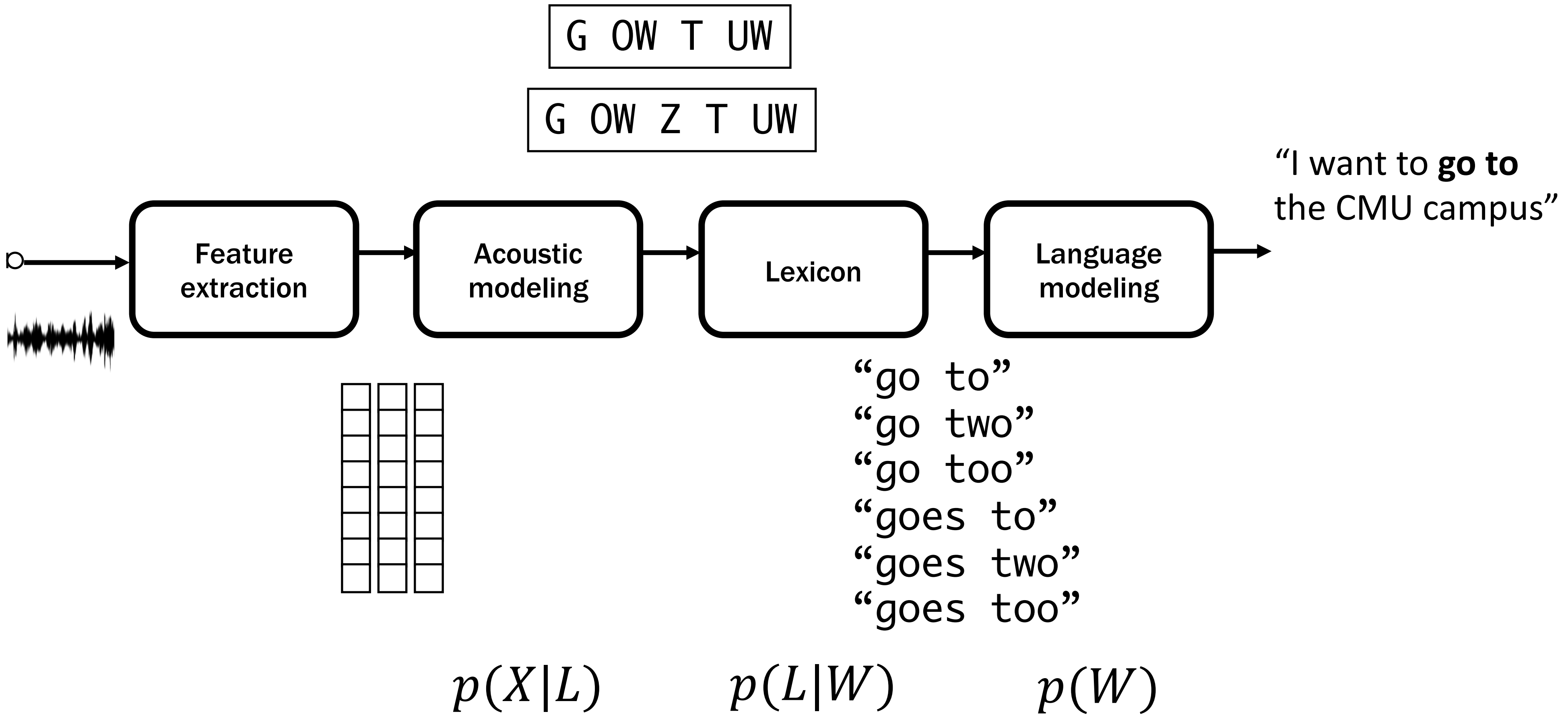
- Sphinx Speech Lunch is an **open space**
  - Mix of **both public and private modes** of the talk
  - Like “openreview” or “GitHub Organization”
  - We can freely discuss the ongoing work even it is under the double-blind review if we make this portion as a private mode

Let’s keep Pittsburgh as a hub for speech research!

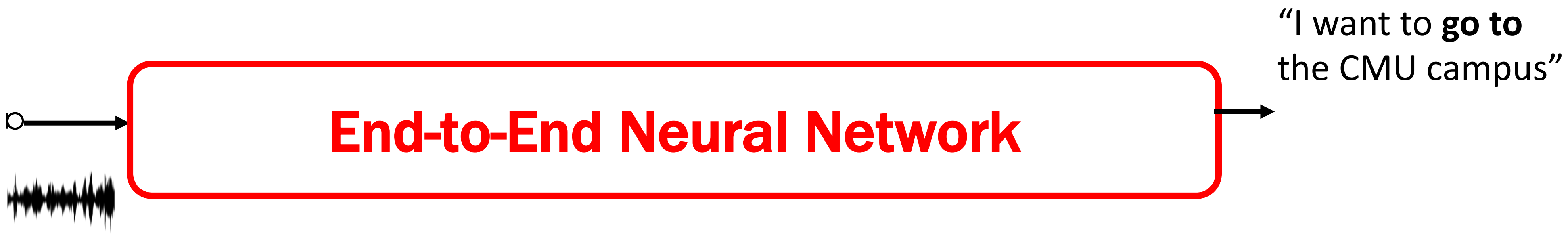
Today's talk

- End-to-End neural network as an **integration tool** for various speech processing.
- In addition to introduce our (or others') previous studies, I would try to make a discussion point of this methodology.
- I want to activate some discussions rather than making some conclusions

Modular system vs. End-to-End system

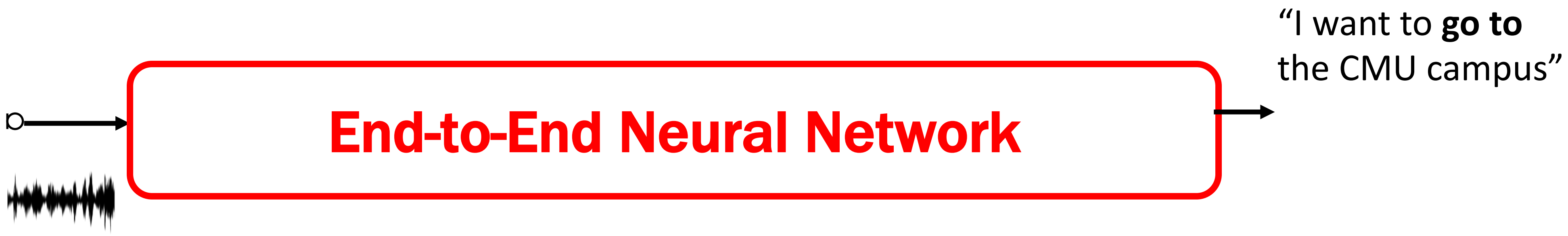


Modular system vs. End-to-End system



- Train a deep network that **directly** maps speech signal to the target letter/word sequence
- Greatly **simplify** the complicated model-building/inference process
- **Integrate** various modules by **optimizing the entire network** with a single objective function

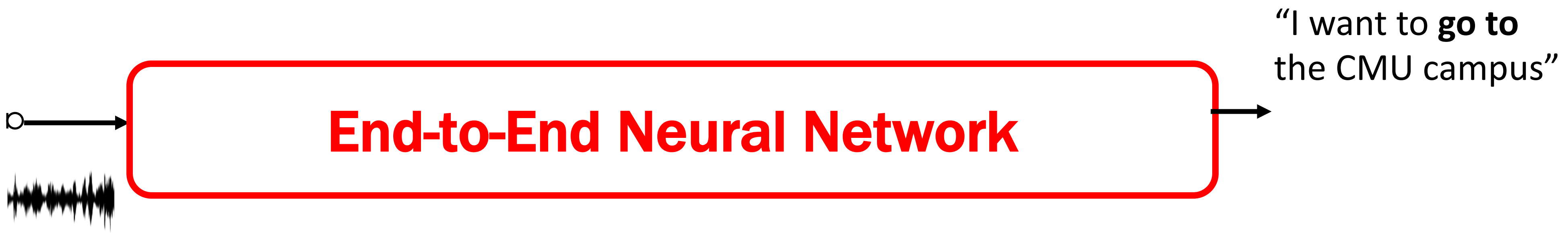
Modular system vs. End-to-End system



- Train a deep network that **directly** maps speech signal to the target letter/word sequence
- Greatly **simplify** the complicated model-building/inference process
- **Integrate** various modules by **optimizing the entire network** with a single objective function

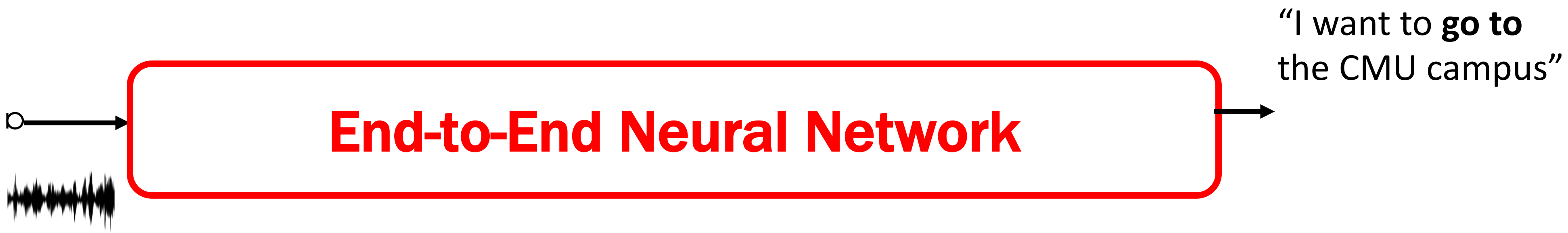
Note that these characteristics always have pros and cons

Modular system vs. End-to-End system



- Train a deep network that **directly** maps speech signal to the target letter/word sequence → **We don't know what's happening. We lose the explainability.**
- Greatly **simplify** the complicated model-building/inference process
- **Integrate** various modules by **optimizing the entire network** with a single objective function → **Difficult to optimize it**

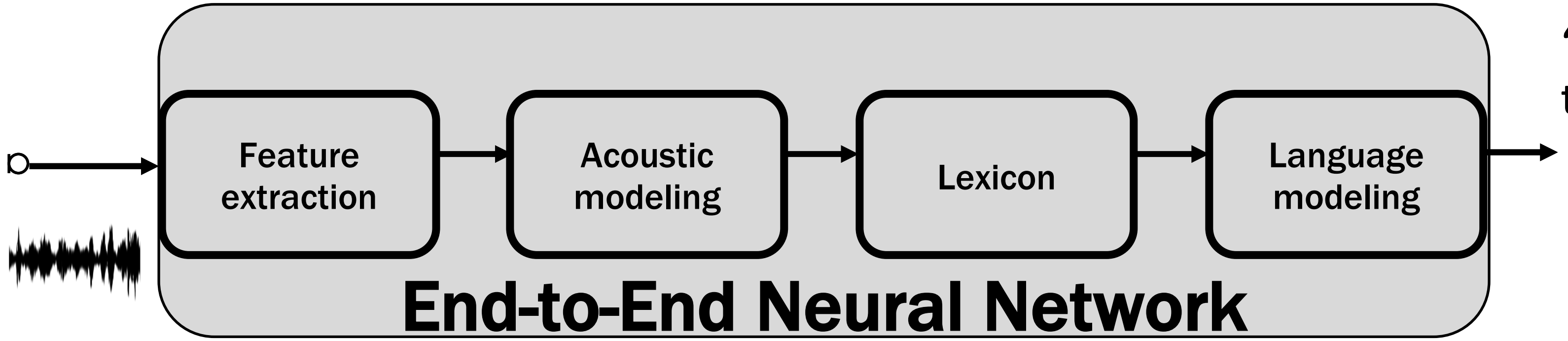
Today's topic



Today's topic
black box

"I want to **go to** the CMU campus"

Today's topic
From black box to transparent box



"I want to **go to** the CMU campus"

- Maintain modularity
- Toward global optimization with back propagation
- **Explainable**

Table of contents

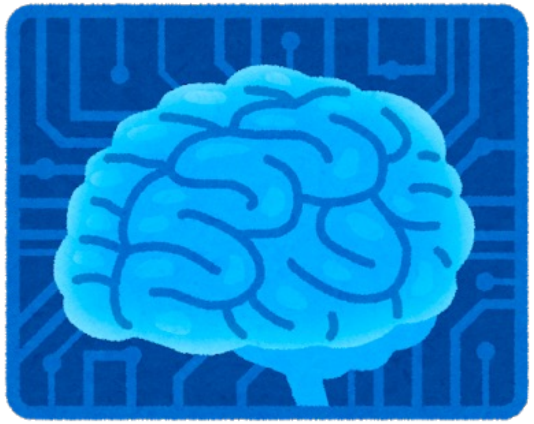
1. **End-to-End Integration** of Speech Recognition and Speech Enhancement

   ```
   X = Speech Recognition
   2) X += Denoising, Dereverberation
   3) X += Speech Separation
   4) X += Speech Localization
   ```

2. **End-to-End Integration** of Speech Recognition and Speech Synthesis
3. Discussion

To solve machine listening

Sound data



- **Machine listening**

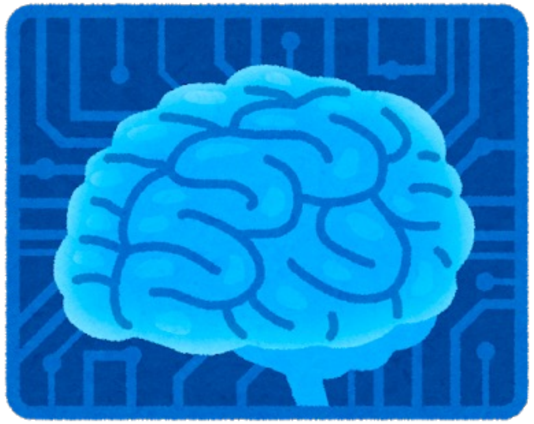
5W1H

Who is speaking when, what, where, how, why

This is one of the ultimate goals of speech and audio processing

To solve machine listening

Sound data



- **Machine listening**

5W1H

Who is speaking when, what, where, how, why

This is one of the ultimate goals of speech and audio processing

Far-field Speech Processing

Close-talking microphone

e.g., voice search

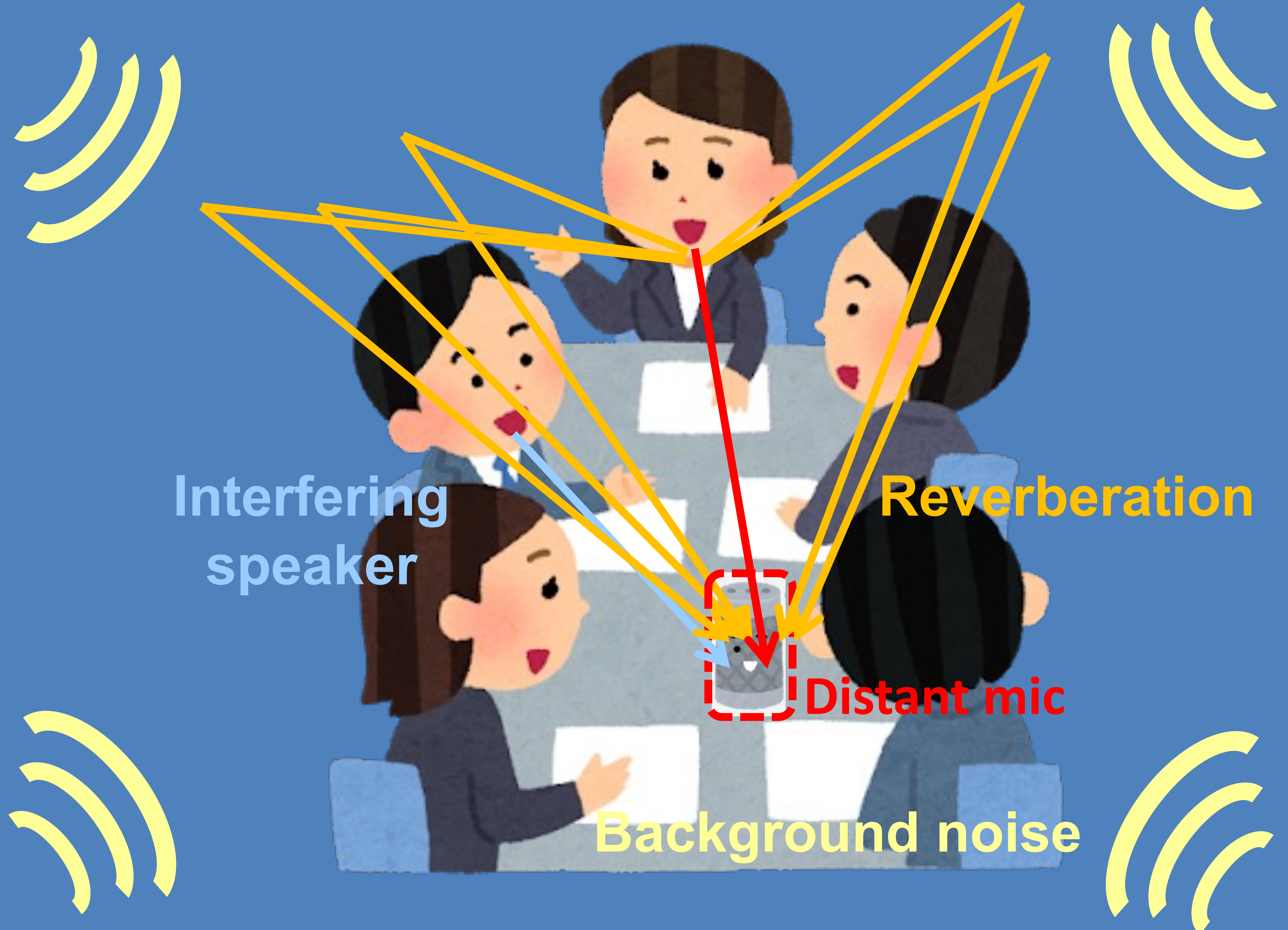
Distant microphone

e.g., Human-human comm.

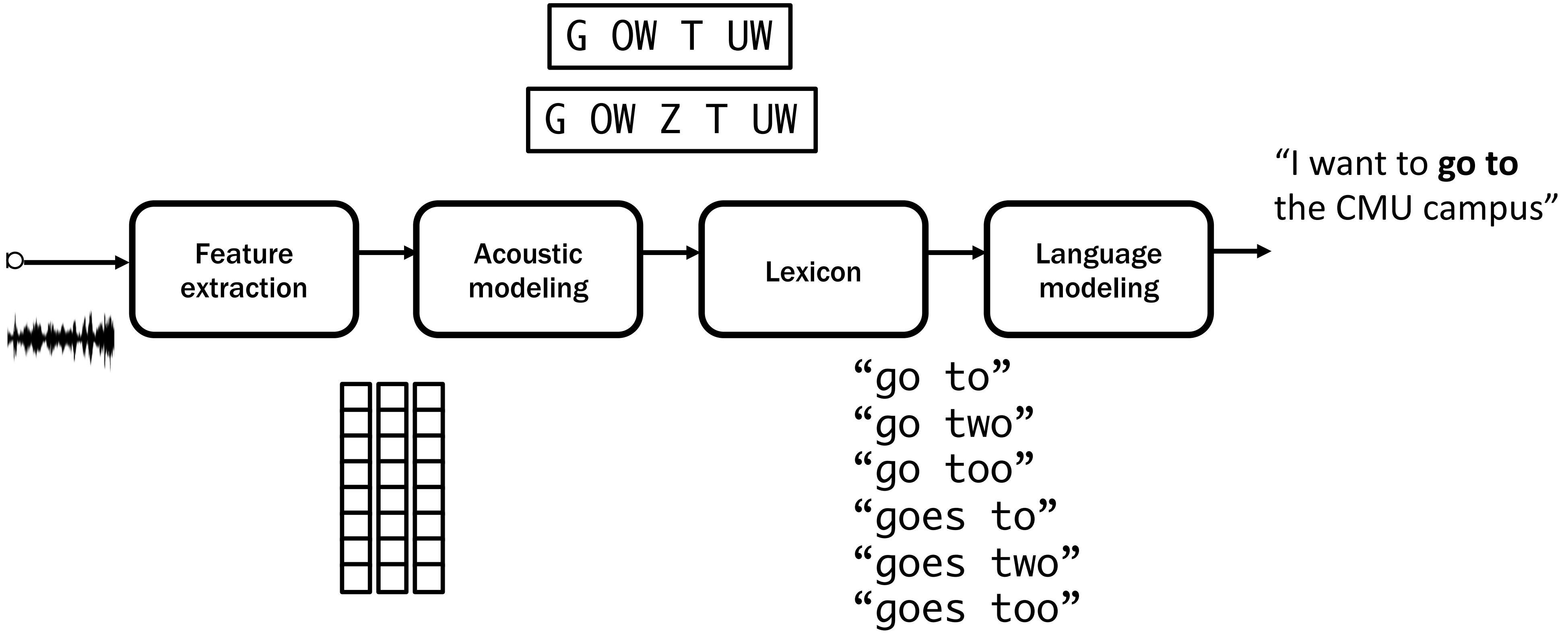
(meeting, conversation analysis)

Human-robot comm.

Machine listening

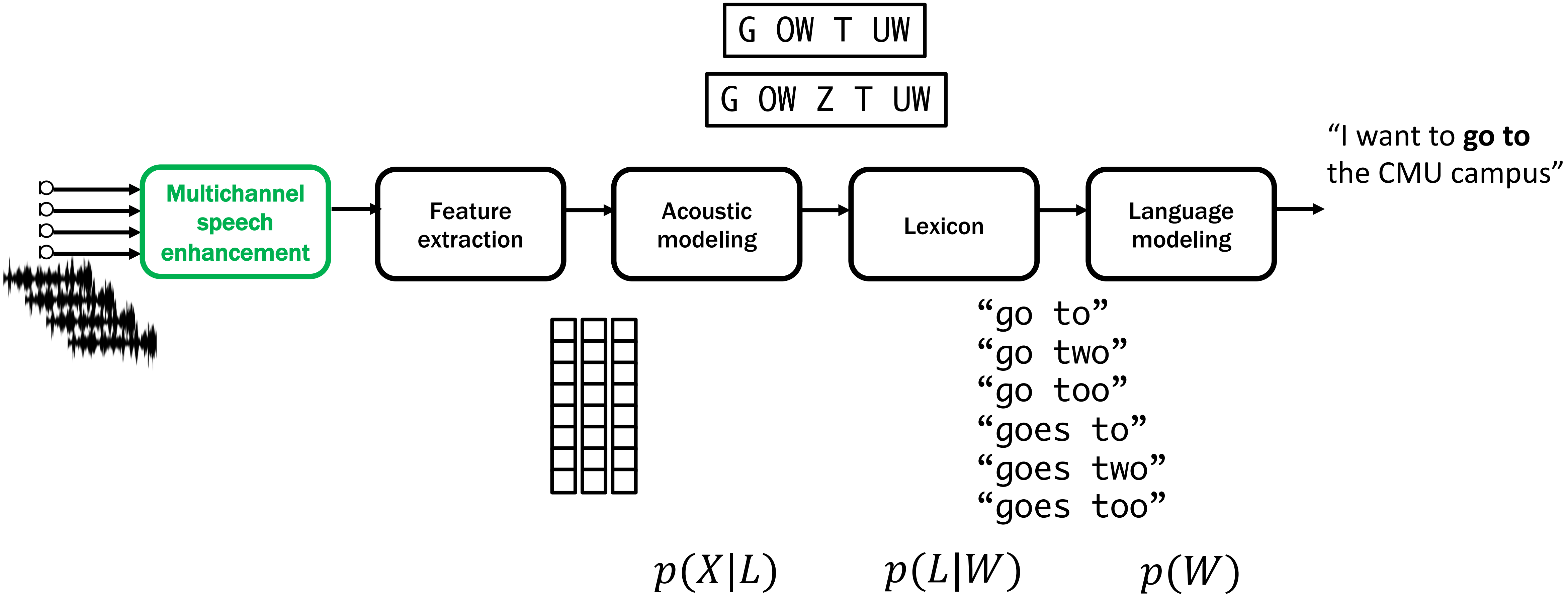

Interfering speaker
Reverberation
Distant mic
Background noise

Speech recognition pipeline

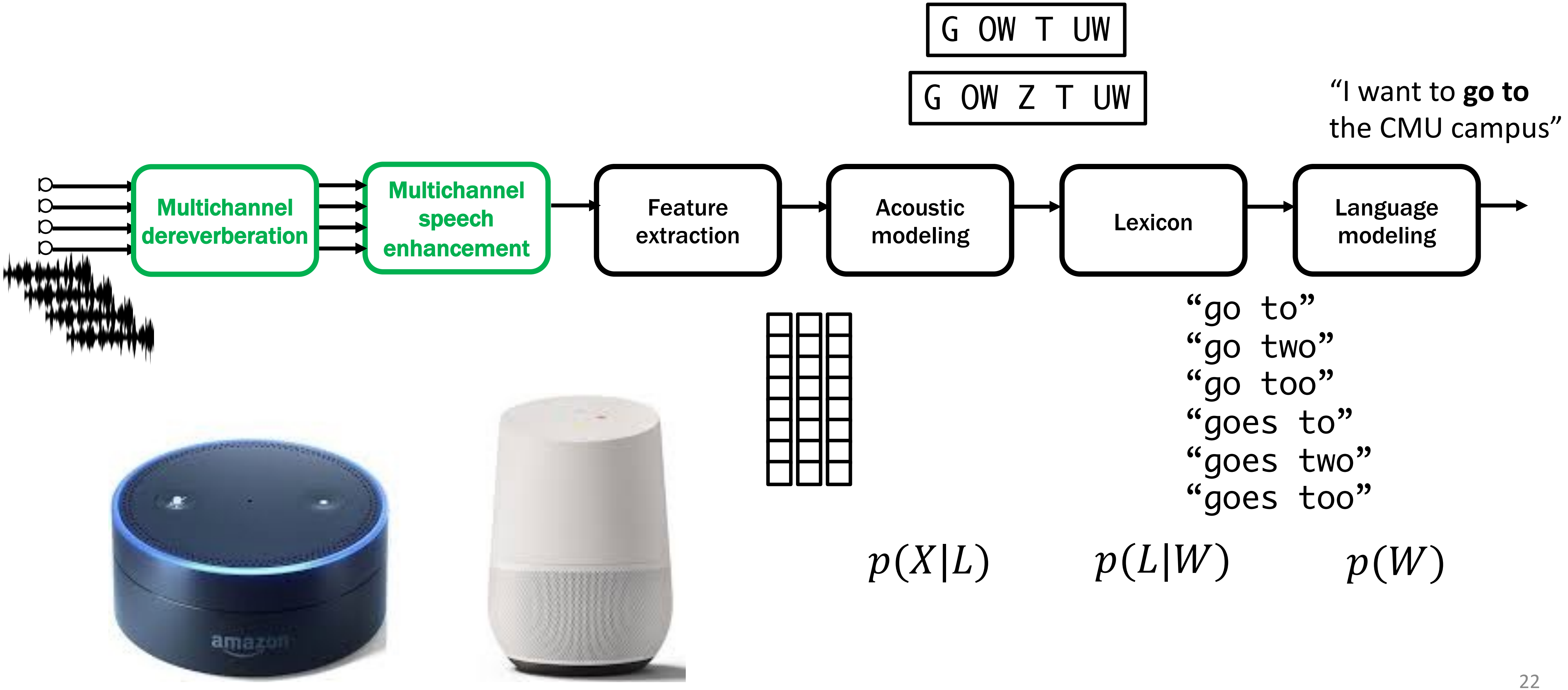


$p(X|L)$ $p(L|W)$ $p(W)$

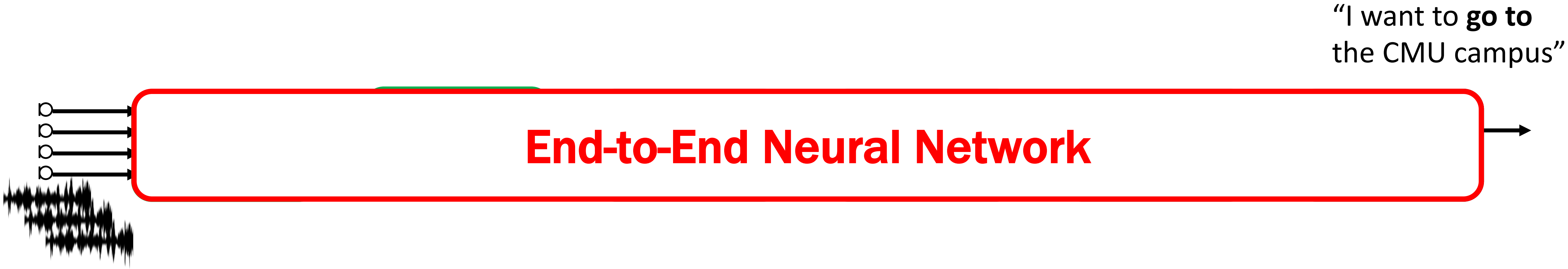
Far-field speech recognition pipeline



Far-field speech recognition pipeline



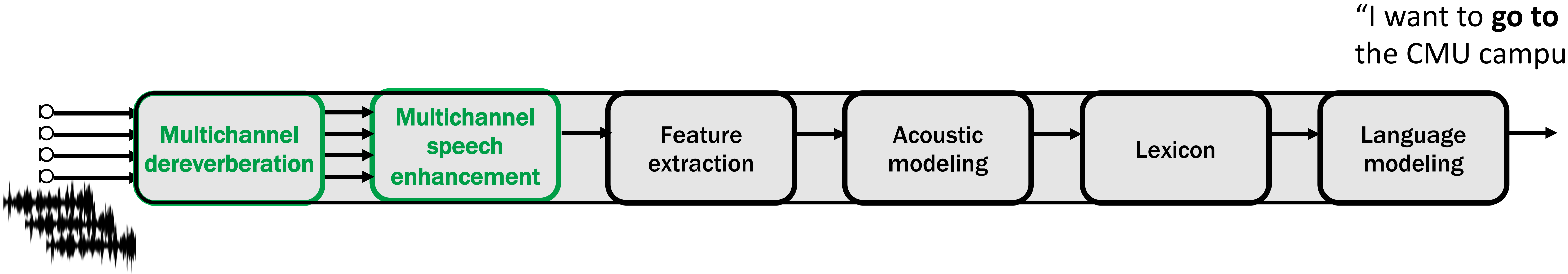
How to design a neural network



How to design a neural network?

- Black box neural network!?

How to design a neural network?



Interpretable neural network

- Keep the original modularity
- Carefully design each module to keep computational graphs
- We can provide interpretations for each sub neural network module

Table of contents

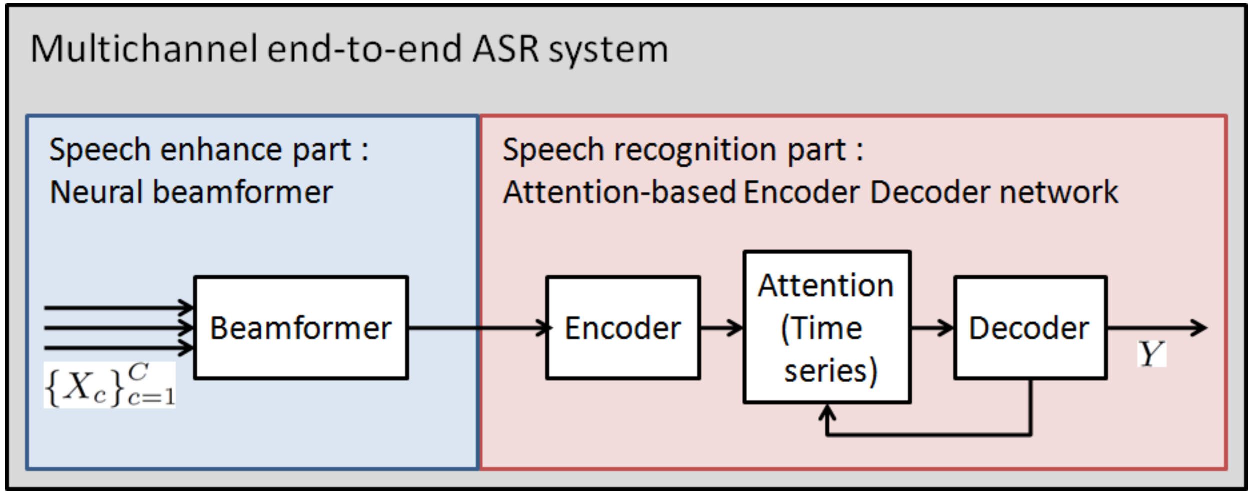
Overview of entire architecture
[Ochiai et al., 2017, ICML]

- Multichannel end-to-end (ME2E) architecture
  - integrates entire process of **speech enhancement (SE)** and **speech recognition (SR)**, by single neural-network-based architecture

↓

SE : Mask-based neural beamformer [Erdogan et al., 2016]

SR : Attention-based encoder-decoder network [Chorowski et al., 2014]



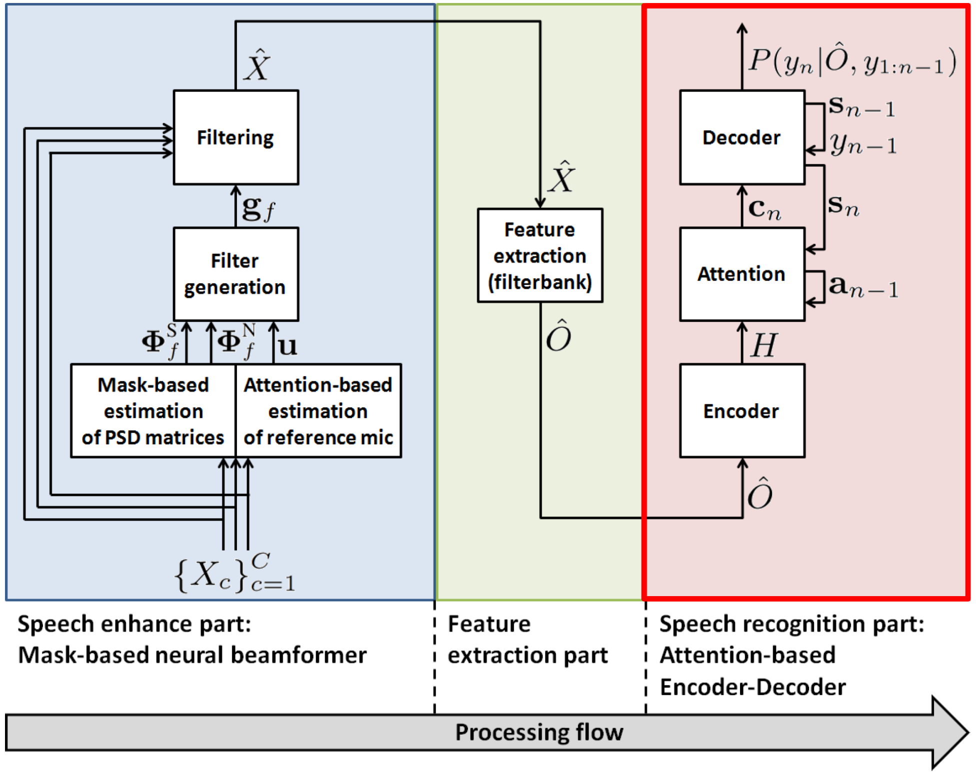
Proposed framework
Overview of entire architecture

- Multichannel end-to-end ASR framework
  - integrates entire process of **speech enhancement (SE)** and **speech recognition (SR)**, by single neural-network-based architecture

↓

SE : Mask-based neural beamformer [Erdogan et al., 2016]

SR : Attention-based encoder-decoder network [Chorowski et al., 2014]



Based on a lot of **signal processing oriented** components!

Beamformer subnetwork

Imitate minimum variance distortionless response (MVDR) beamformer

- Basic equation to obtain enhanced signal $\hat{x}_{t,f}$ at frame t and bin f

$$\hat{x}_{t,f} = \mathbf{g}_f^{\mathrm{H}} \mathbf{x}_{t,f}$$

  - $\mathbf{x}_{t,f} \in \mathbb{C}^M$: observed M multichannel signal
  - $\boldsymbol{g}_f \in \mathbb{C}^M$: beamforming filter coefficients
- Time-invariant beamforming filter with a reference mic $\mathbf{u}$

Mask m is obtained from DNN

$$\mathbf{g}_f = \frac{(\boldsymbol{\Phi}_f^{\mathrm{N}})^{-1} \boldsymbol{\Phi}_f^{\mathrm{S}}}{\mathrm{Tr}((\boldsymbol{\Phi}_f^{\mathrm{N}})^{-1} \boldsymbol{\Phi}_f^{\mathrm{S}})} \mathbf{u},$$

$$\boldsymbol{\Phi}_f^{\mathrm{S}} = \frac{1}{\sum_{t=1}^{T} m_{t,f}^{\mathrm{S}}} \sum_{t=1}^{T} m_{t,f}^{\mathrm{S}} \mathbf{x}_{t,f} \mathbf{x}_{t,f}^{\dagger},$$

$$\boldsymbol{\Phi}_f^{\mathrm{N}} = \frac{1}{\sum_{t=1}^{T} m_{t,f}^{\mathrm{N}}} \sum_{t=1}^{T} m_{t,f}^{\mathrm{N}} \mathbf{x}_{t,f} \mathbf{x}_{t,f}^{\dagger}.$$

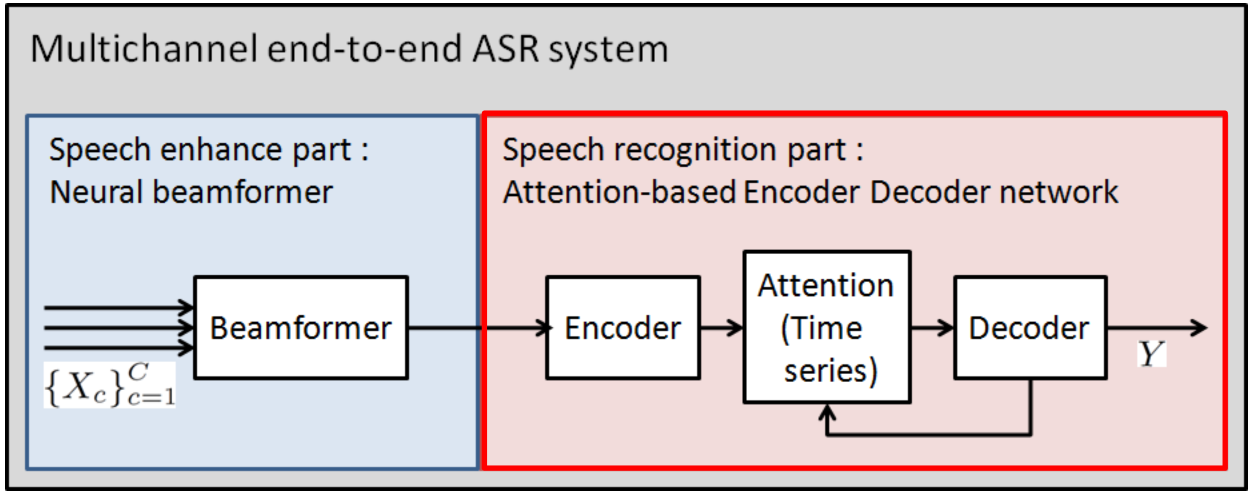
Overview of entire architecture

- Multichannel end-to-end (ME2E) architecture
  - integrates entire process of **speech enhancement (SE)** and **speech recognition (SR)**, by single neural-network-based architecture

↓

SE : Mask-based neural beamformer [Erdogan et al., 2016]

SR : Attention-based encoder-decoder network [Chorowski et al., 2014]



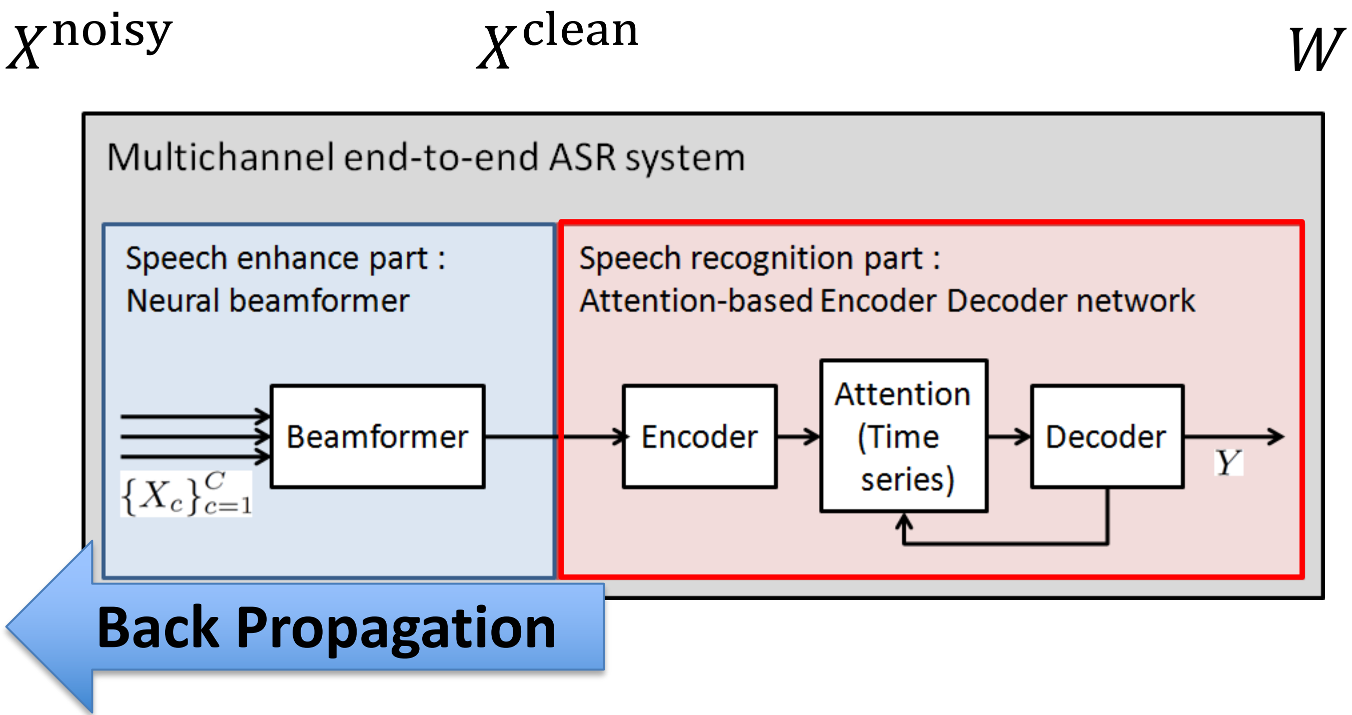
Overview of entire architecture

- **Multichannel end-to-end (ME2E) architecture**
  - integrates entire process of **speech enhancement (SE)** and **speech recognition (SR)**, by single neural-network-based architecture

↓

SE : Mask-based neural beamformer [Erdogan et al., 2016]

SR : Attention-based encoder-decoder network [Chorowski et al., 2014]



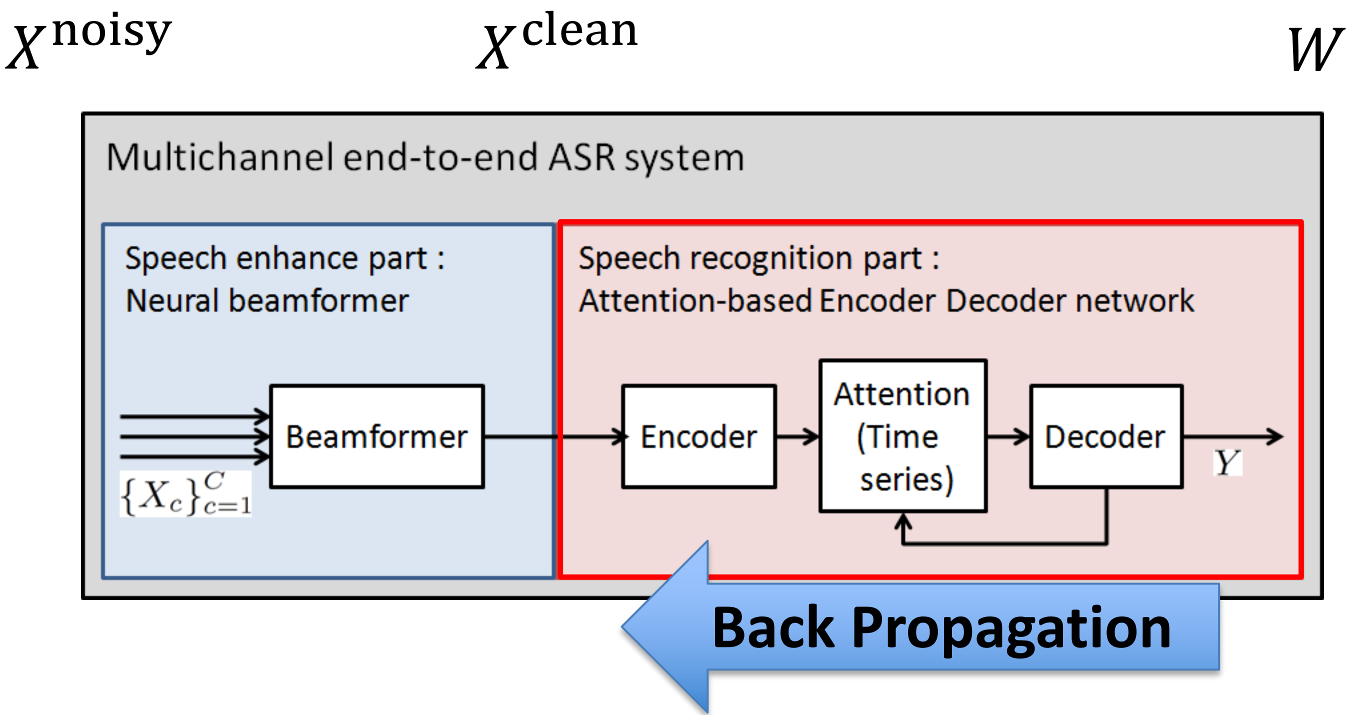
Overview of entire architecture

- Multichannel end-to-end (ME2E) architecture
  - integrates entire process of **speech enhancement (SE)** and **speech recognition (SR)**, by single neural-network-based architecture

↓

SE : Mask-based neural beamformer [Erdogan et al., 2016]

SR : Attention-based encoder-decoder network [Chorowski et al., 2014]



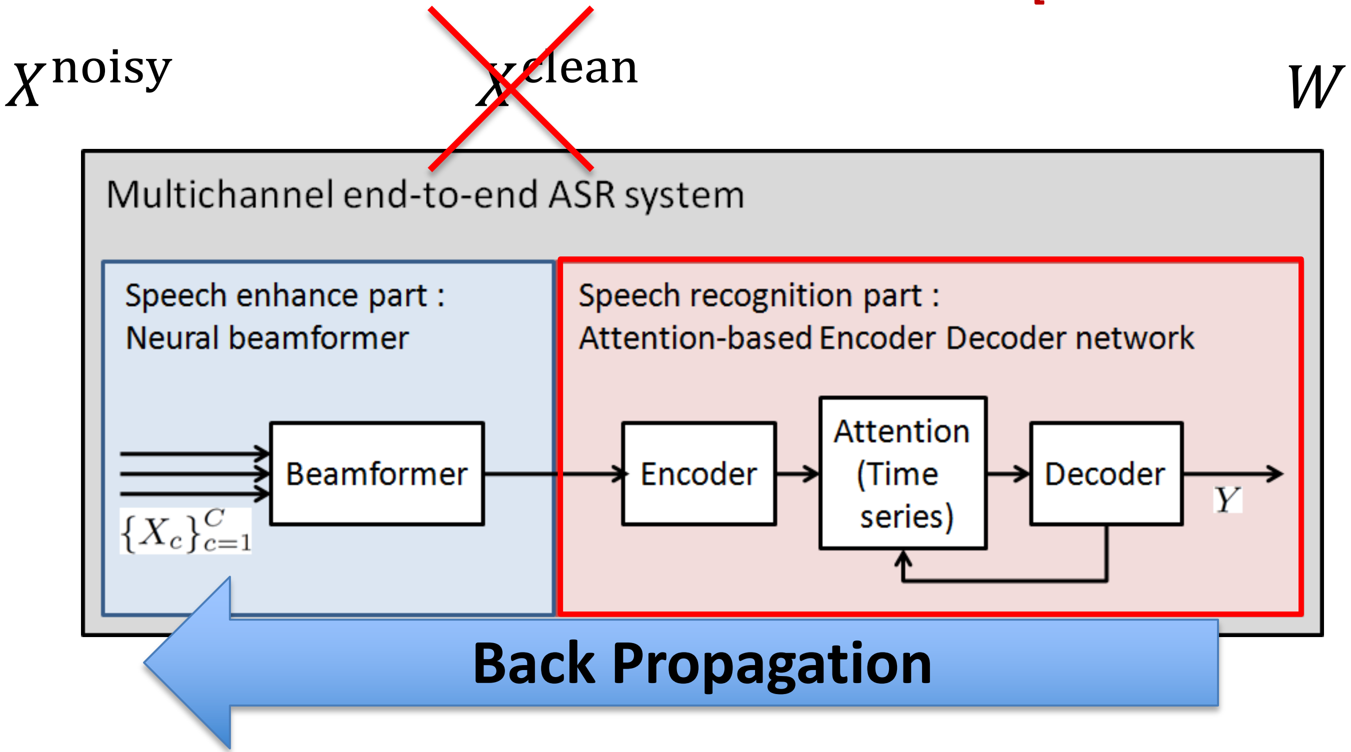
Overview of entire architecture

- Multichannel end-to-end (ME2E) architecture
  - integrates entire process of **speech enhancement (SE)** and **speech recognition (SR)**, by single neural-network-based architecture

↓

SE : Mask-based neural beamformer [Erdogan et al., 2016]

SR : Attention-based encoder-decoder network [Chorowski et al., 2014]



- We don't need pair data of noisy and clean data
- We can train both beamforming and ASR with the ASR criterion

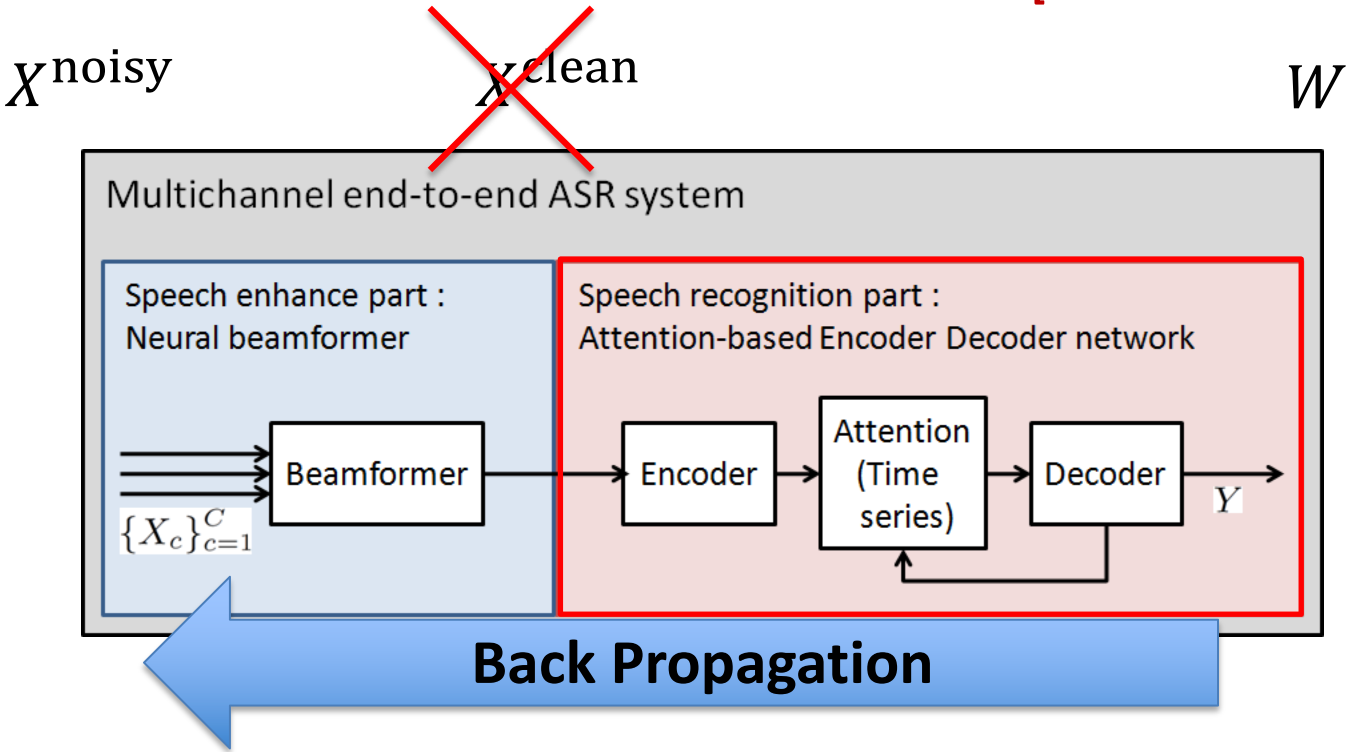
Overview of entire architecture

- Multichannel end-to-end (ME2E) architecture
  - integrates entire process of **speech enhancement (SE)** and **speech recognition (SR)**, by single neural-network-based architecture

↓

SE : Mask-based neural beamformer [Erdogan et al., 2016]

SR : Attention-based encoder-decoder network [Chorowski et al., 2014]



- We don't need pair data of noisy and clean data
- We can train both beamforming and ASR with the ASR criterion
- **NOTE**: It's not new! This methodology was already established in LIMABEAM [Seltzer et al., 2004]

Further extension
Dereverberation + beamforming + ASR

https://github.com/nttcslab-sp/dnn_wpe, [Subramanian'19]

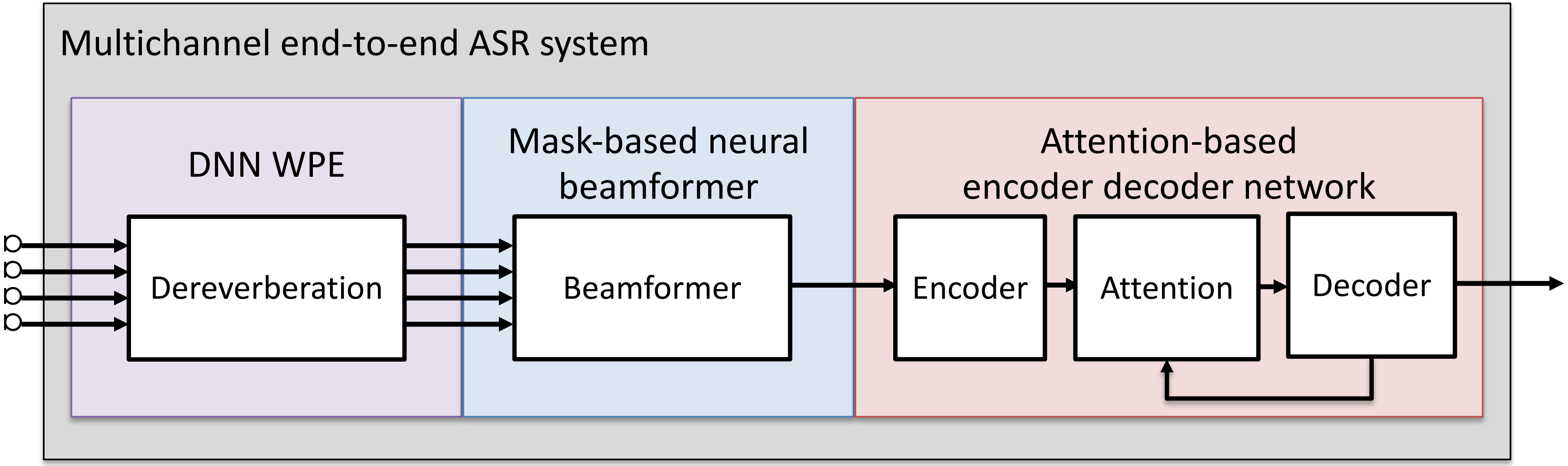
- Multichannel end-to-end ASR framework
  - integrates entire process of **speech dereverberation (SD), beamforming (SB)** and **speech recognition (SR)**, by single neural-network-based architecture

↓

SD : DNN-based weighted prediction error (DNN-WPE) [Kinoshita et al., 2016]

SB : Mask-based neural beamformer [Erdogan et al., 2016]

SR : Attention-based encoder-decoder network [Chorowski et al., 2014]



Further extension
Dereverberation + beamforming + ASR

https://github.com/nttcslab-sp/dnn_wpe, [Subramanian'19]

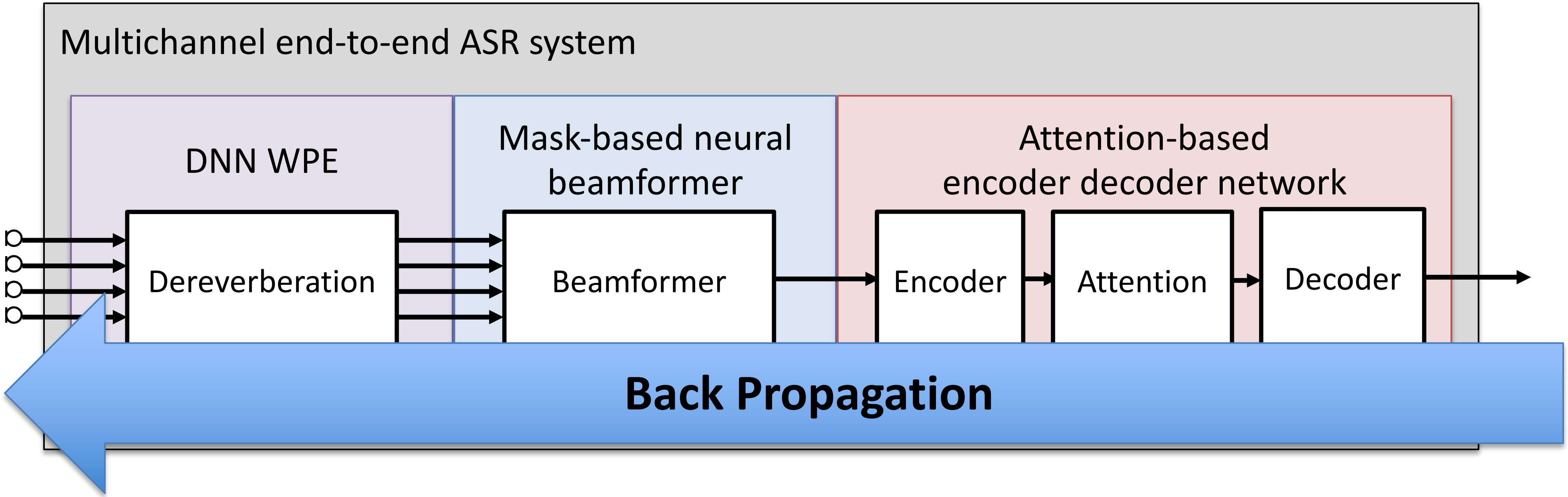
- Multichannel end-to-end ASR framework
  - integrates entire process of **speech dereverberation (SD)**, **beamforming (SB)** and **speech recognition (SR)**, by single neural-network-based architecture

↓

SD : DNN-based weighted prediction error (DNN-WPE) [Kinoshita et al., 2016]

SB : Mask-based neural beamformer [Erdogan et al., 2016]

SR : Attention-based encoder-decoder network [Chorowski et al., 2014]



Dereverberation subnetwork

Imitate multichannel linear prediction filtering [Nakatani+(2010)]

- Basic equation (Δ delayed linear prediction)

$$\mathbf{d}(t,b) = \mathbf{y}(t,b) - \mathbf{G}^{\mathrm{H}}(b)\tilde{\mathbf{y}}(t-\Delta,b),$$

  - $\mathbf{y}(t,b) \in \mathbb{C}^M$ is the observed multichannel signal (b: frequency bin, M: # channels)
  - Filter: $\mathbf{G}^{\mathrm{H}}(b)$, history of the observation signal: $\tilde{\mathbf{y}}(t-\Delta,b)$

- Update equations (well-known maximum likelihood solutions)

$$\lambda(t,b) = \frac{1}{M}\sum_m |\bar{d}(t,b,m)|^2,$$

Obtained from DNN

[Kinoshita et al., 2016]

$$\mathbf{R}(b) = \sum_t \frac{\tilde{\mathbf{y}}(t-\Delta,b)\tilde{\mathbf{y}}^{\mathrm{H}}(t-\Delta,b)}{\lambda(t,b)},$$

$$\mathbf{P}(b) = \sum_t \frac{\tilde{\mathbf{y}}(t-\Delta,b)\mathbf{y}^{\mathrm{H}}(t,b)}{\lambda(t,b)} \in \mathbb{C}^{ML\times M},$$

$$\mathbf{G}(b) = \mathbf{R}(b)^{-1}\mathbf{P}(b) \in \mathbb{C}^{ML\times M},$$

Experimental Results

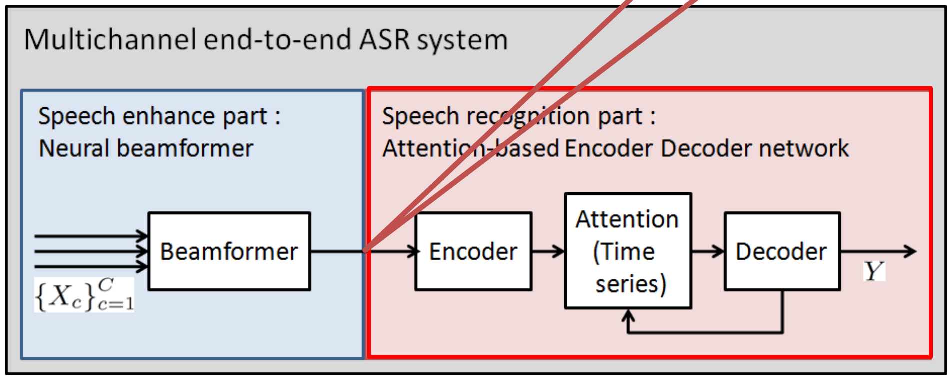
https://github.com/nttcslab-sp/dnn_wpe, [Subramanian'19]

- Noisy reverberant speech recognition task (REVERB and DIRHA-WSJ)
  - Sigle-channel E2E + dereverberation + beamforming (**pipeline**)
  - Multichannel E2E (**integration** of speech enhancement and recognition)

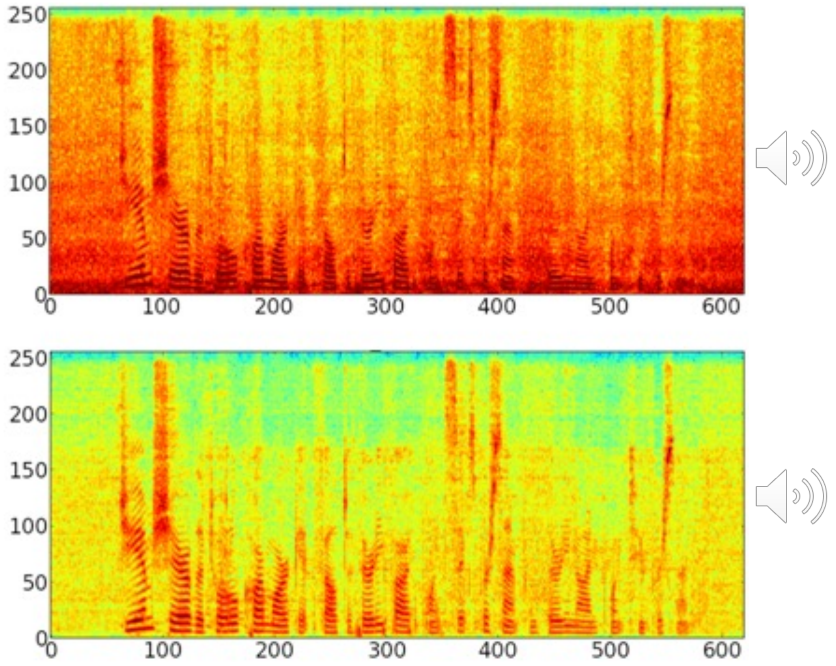
model	REVERB Room1 Near	REVERB Room1 Far	DIRHA WSJ Real
E2E baseline (no enhancement)	23.9	26.8	55.3
Single-channel E2E + Dereverberation + Beamforming (**pipeline**)	11.0	**10.8**	31.3
Multichannel E2E (**end-to-end**)	**8.7**	12.4	**29.1**

It works as speech enhancement!

- Speech samples



Noisy

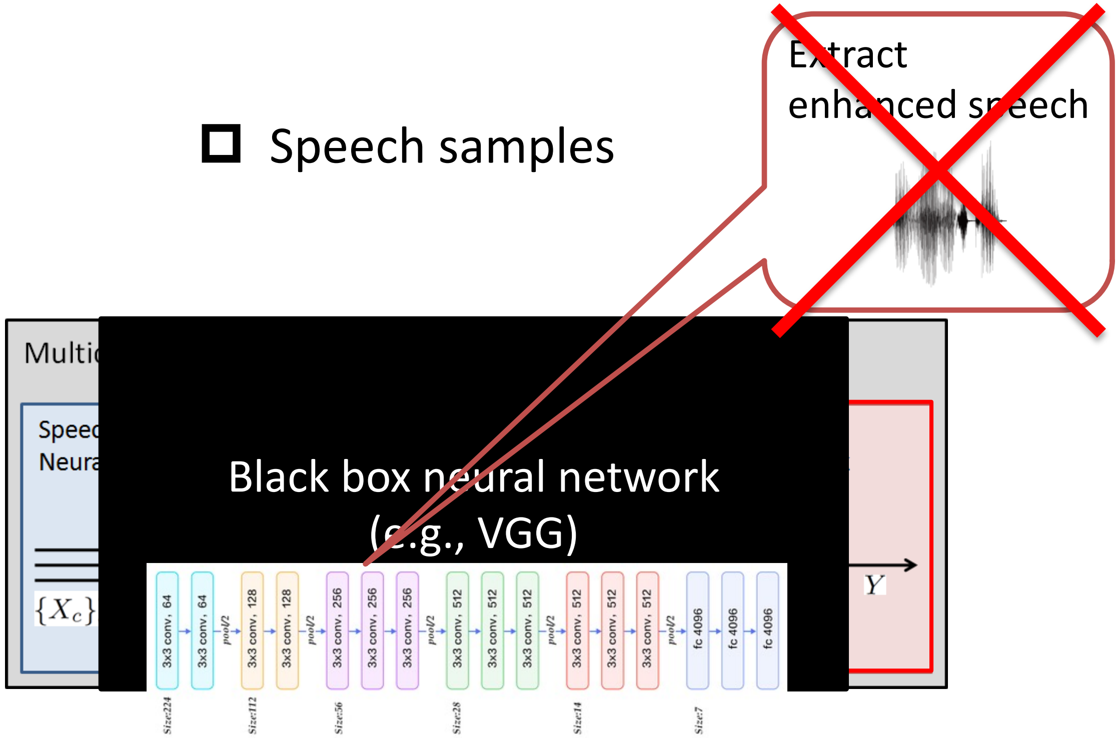


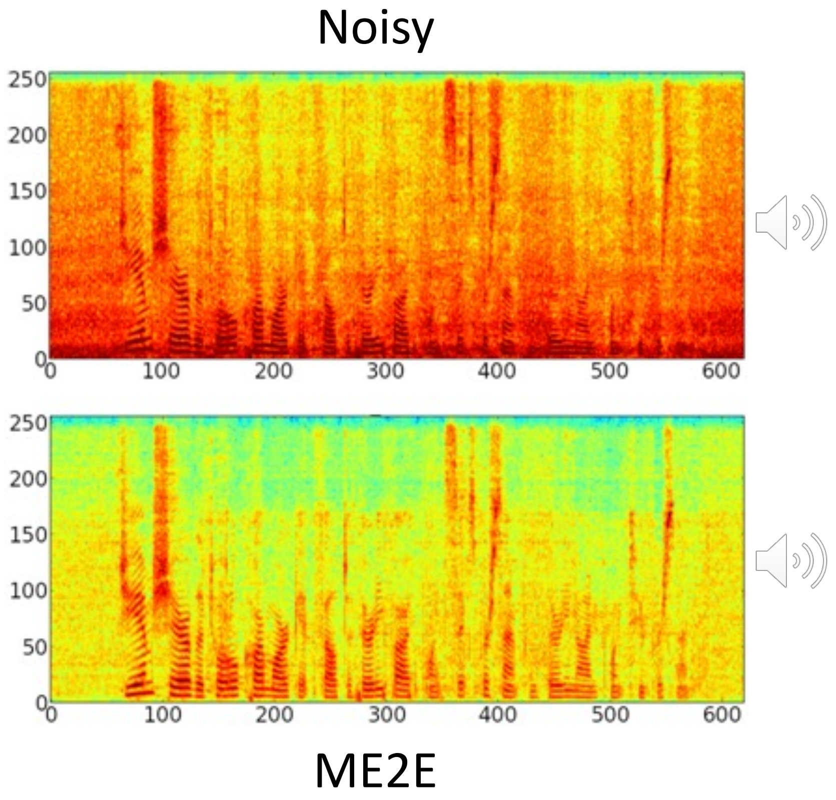
ME2E

- **Entire network** are **consistently optimized with ASR-level objective including speech enhancement part**
- **Pairs of parallel clean and noisy data are not required for training → SE can be optimized only with noisy signals and their transcripts**

It works as speech enhancement!

- Speech samples

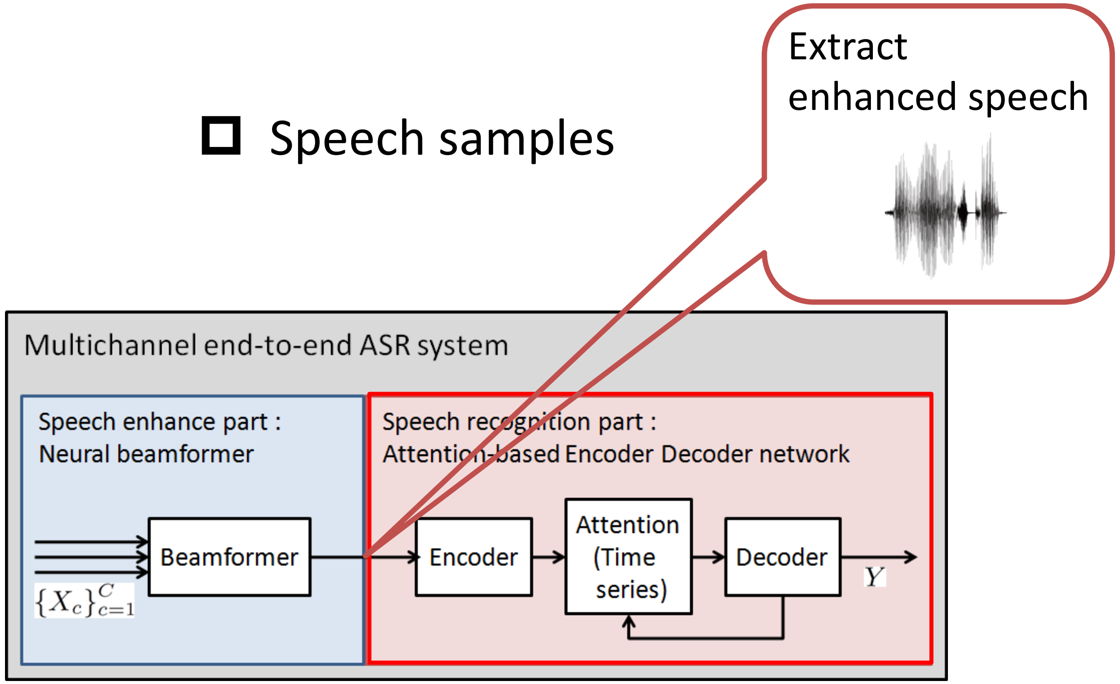


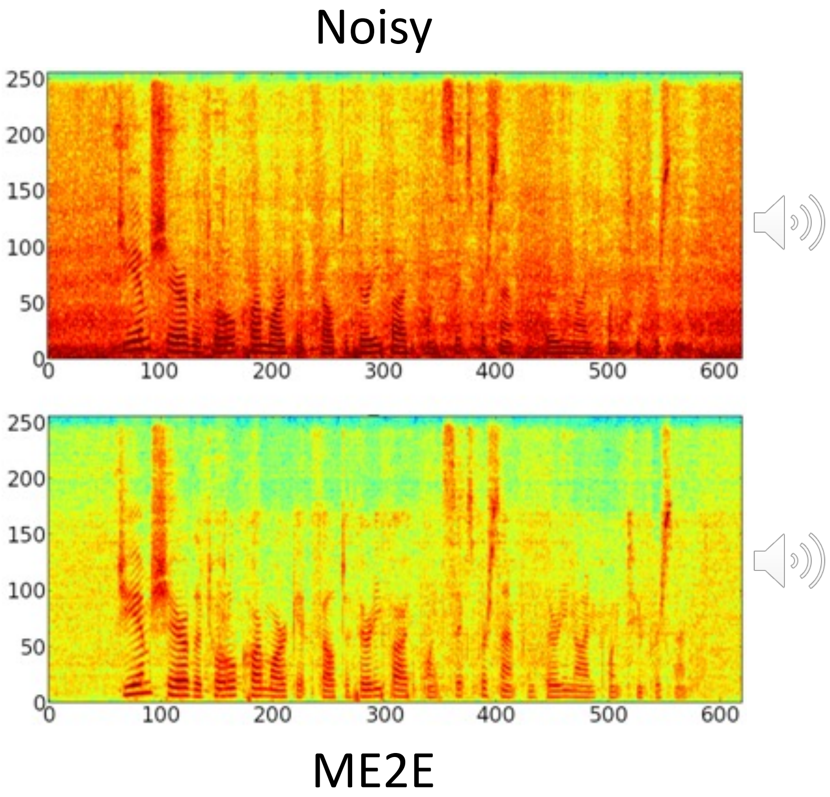


- **Entire network** are **consistently optimized with ASR-level objective including speech enhancement part**
- **Pairs of parallel clean and noisy data are not required for training → SE can be optimized only with noisy signals and their transcripts**

It works as speech enhancement!

- Speech samples

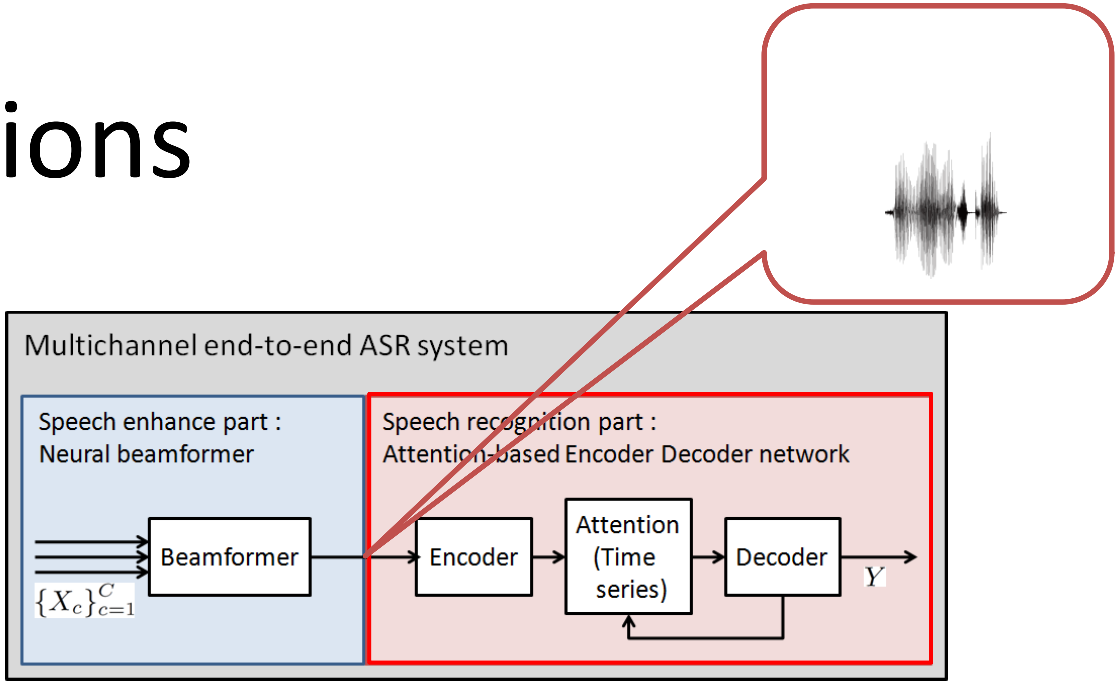




Explainable neural network thanks to ***signal proceeding motivated*** architecture

Discussions

- Is it really better?
- The rich sound information was “**projected**” to the enhanced (clean) speech space
  - The sound event and room acoustic information were **disappeared**.
  - We need to provide supplemental information or original information to avoid this projection problem
  - Taking over the drawback of the modular system
- Why stick to the flat start?



It’s like a measurement problem in the quantum theory(?)

Make it visible
Interpretable

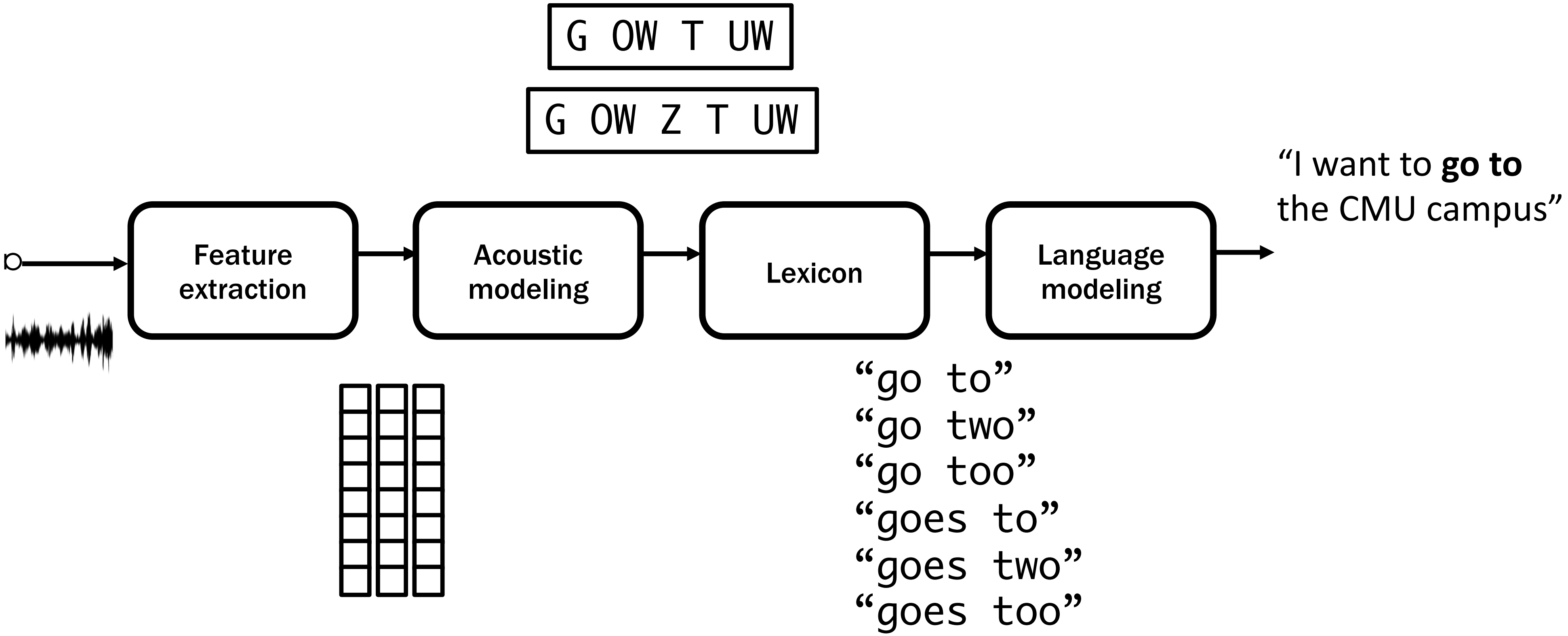
Table of contents

1. **End-to-End Integration** of Speech Recognition and Speech Enhancement

```
X = Speech Recognition
2) X += Denoising, Dereverberation
3) X += Speech Separation
4) X += Speech Localization
```

2. **End-to-End Integration** of Speech Recognition and Speech Synthesis
3. Discussion

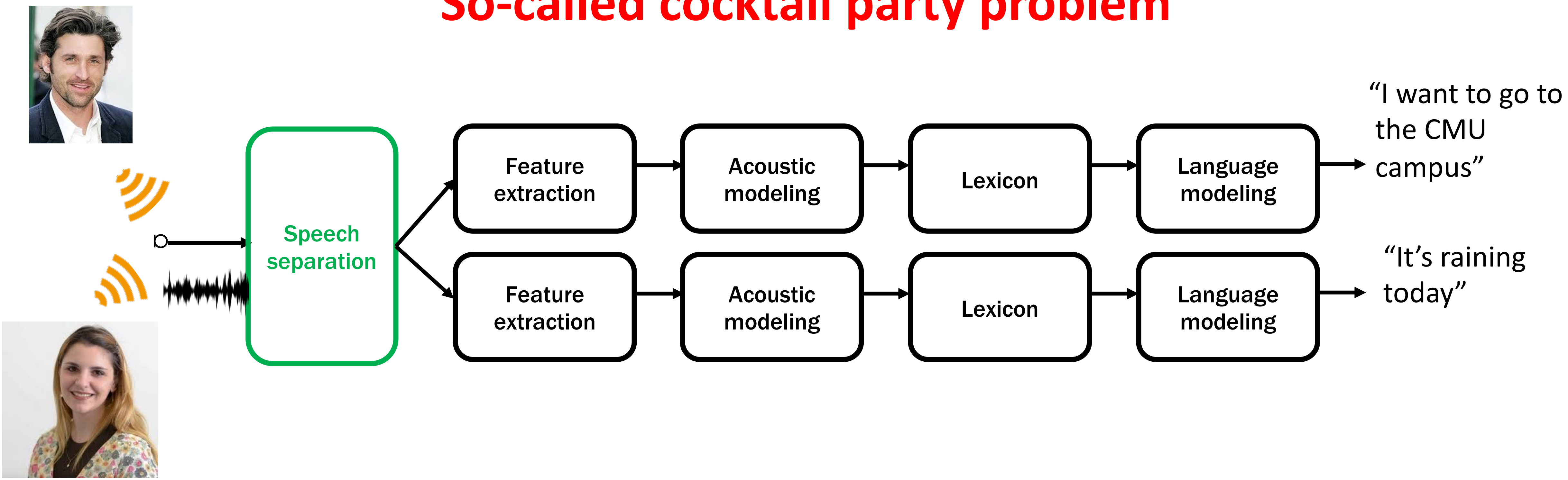
Speech recognition pipeline



$p(X|L)$ $p(L|W)$ $p(W)$

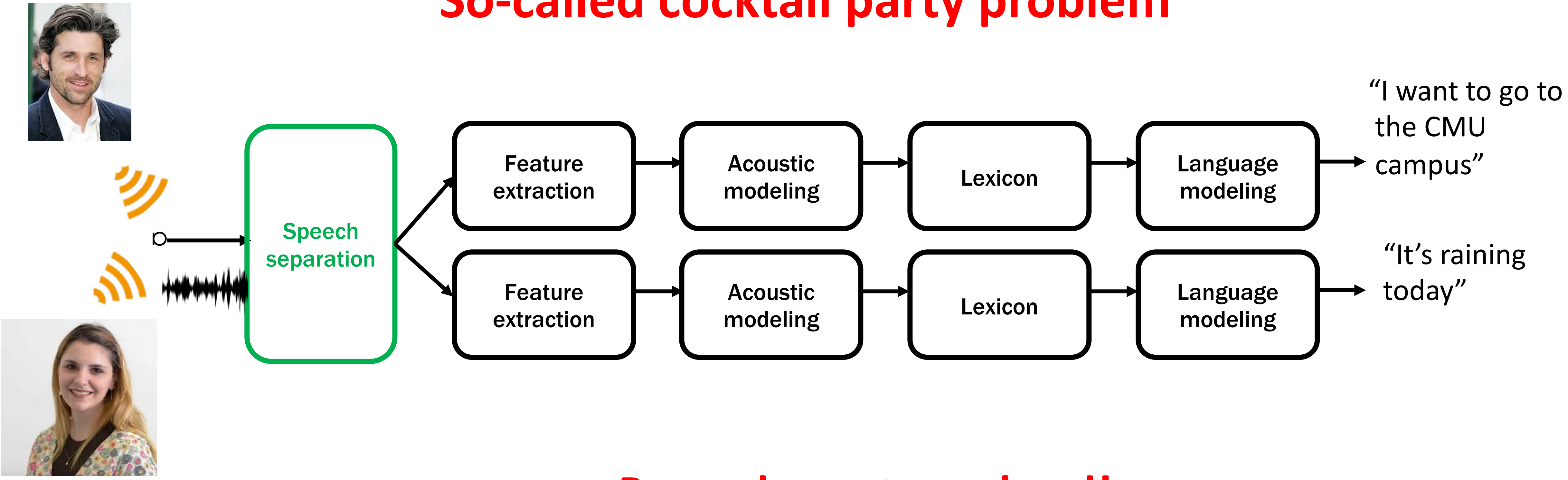
Multi-speaker speech recognition pipeline

So-called cocktail party problem



Multi-speaker speech recognition pipeline

So-called cocktail party problem

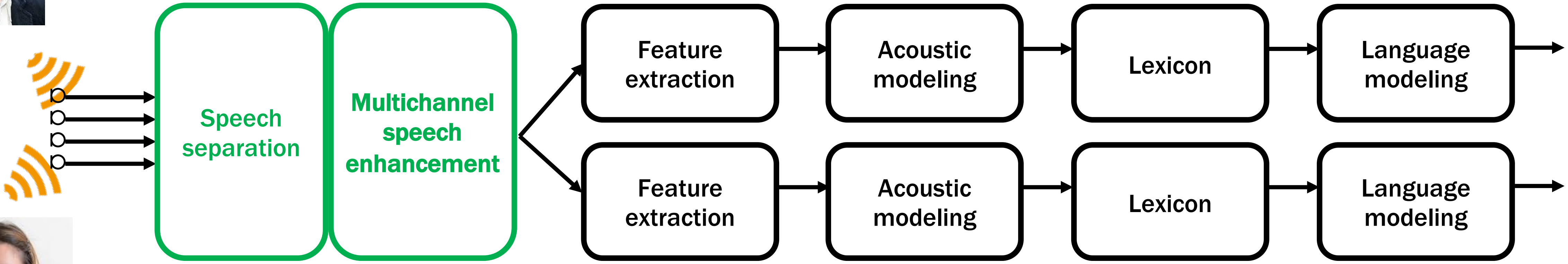


Beyond smart speakers!!

Multi-channel Multi-speaker speech recognition pipeline

So-called cocktail party problem

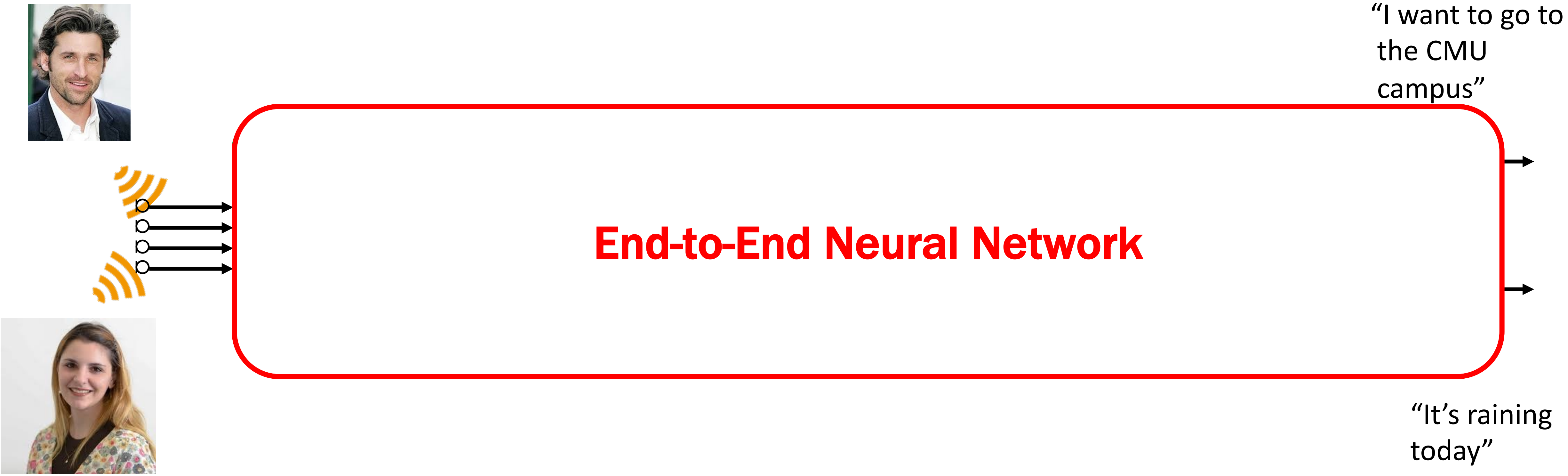
“I want to go to the CMU campus”



“It’s raining today”

Beyond smart speakers!!

Multi-channel Multi-speaker speech recognition pipeline



Integrates separation and recognition with a single end-to-end network

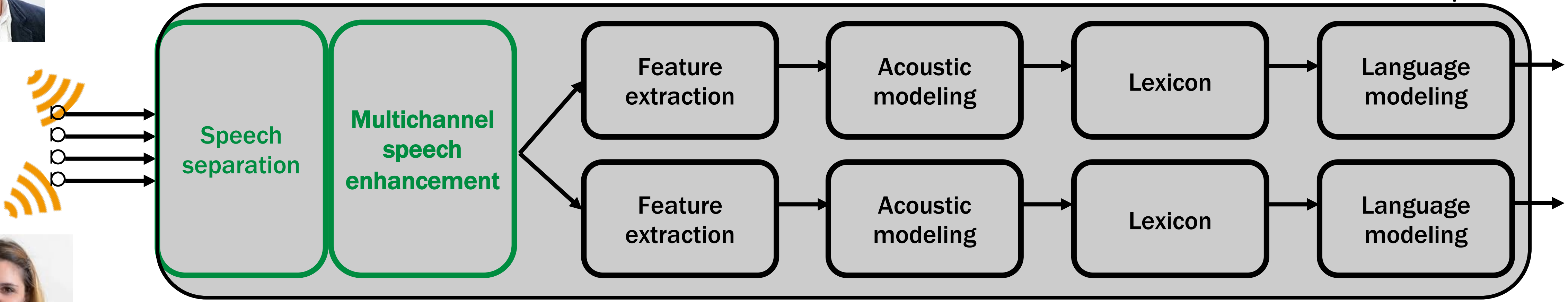
Multi-channel Multi-speaker speech recognition pipeline

Integrates separation and recognition with a single end-to-end network

Multi-channel Multi-speaker speech recognition pipeline

Interpretable neural network

“I want to go to the CMU campus”



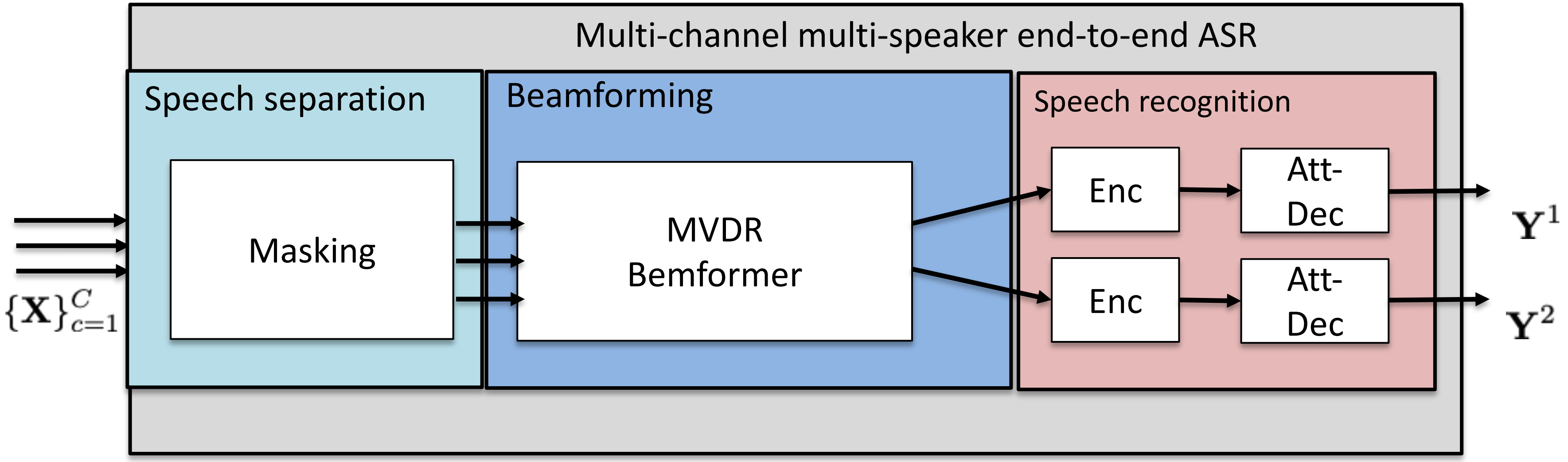
“It’s raining today”

Integrates separation and recognition with a single end-to-end network

Overview of entire architecture

[Xuankai Chang., 2019, ASRU]

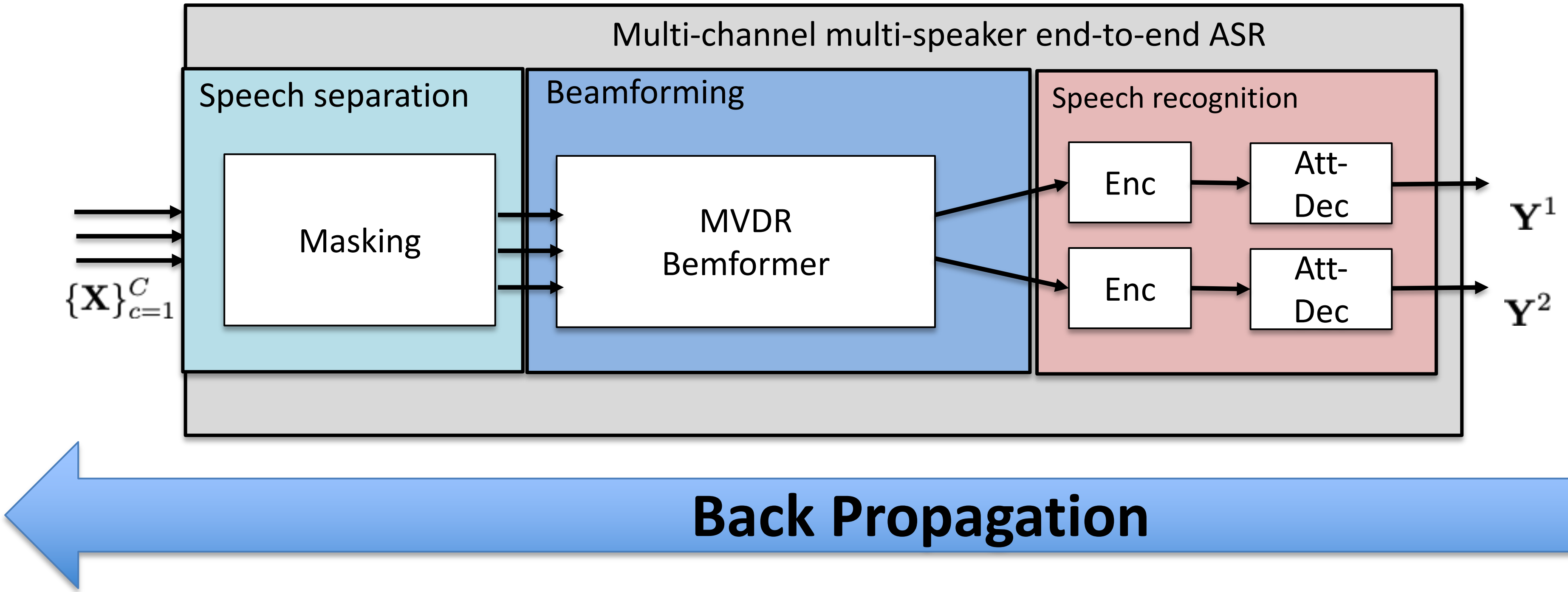
- Multi-channel (MI) multi-speaker (MO) end-to-end architecture (**MIMO-Speech**)
  - Extend our previous model to **multispeaker** end-to-end network based on **permutation invariant training in the ASR reference level**
  - Integrate the ***beamforming-based speech enhancement and separation networks*** inside the neural network



Overview of entire architecture

[Xuankai Chang., 2019, ASRU]

- ❑ Multi-channel (MI) multi-speaker (MO) end-to-end architecture (**MIMO-Speech**)
  - Extend our previous model to **multispeaker** end-to-end network based on **permutation invariant training in the ASR reference level**
  - Integrate the ***beamforming-based speech enhancement and separation networks*** inside the neural network

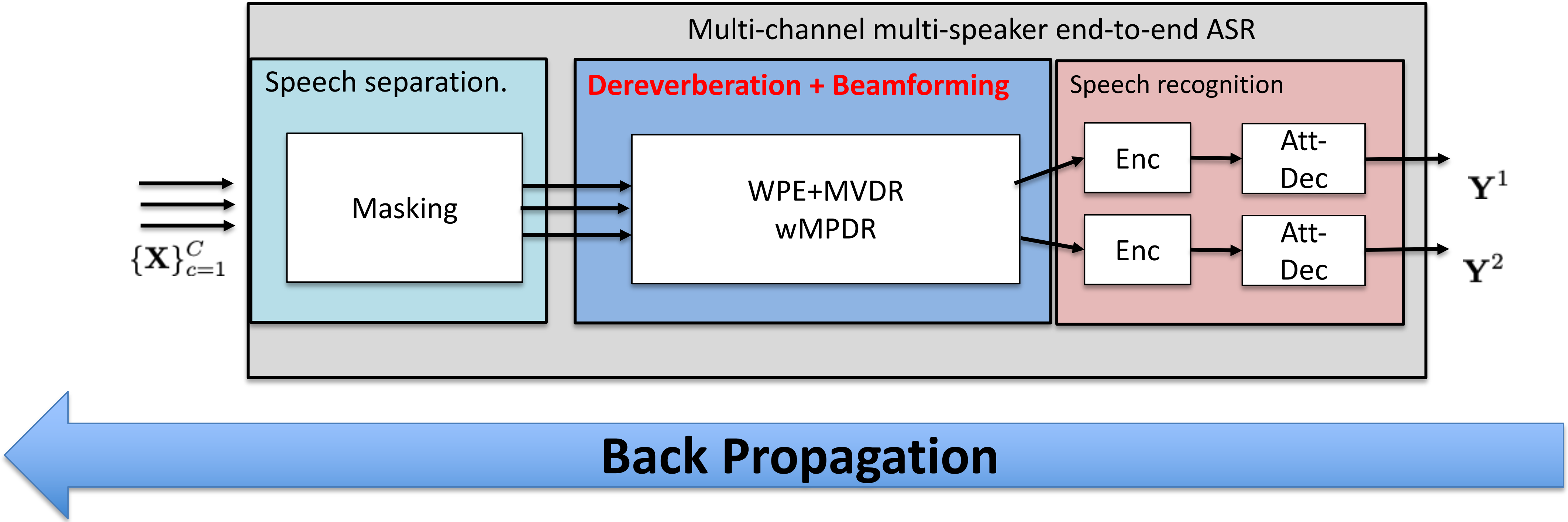


Extensions with with Improved Numerical Stability and Advanced Frontend [Wangyou Zhang, 2021, ICASSP]

❑ Extension of MIMO-speech

- Improved numerical stability (Diagonal loading, mask flooring, precision)
- **Joint dereverberaton and beamforming** (WPE+MVDR or wMPDR)

MIMO speech is now robustly working **under noisy reverberant conations.**



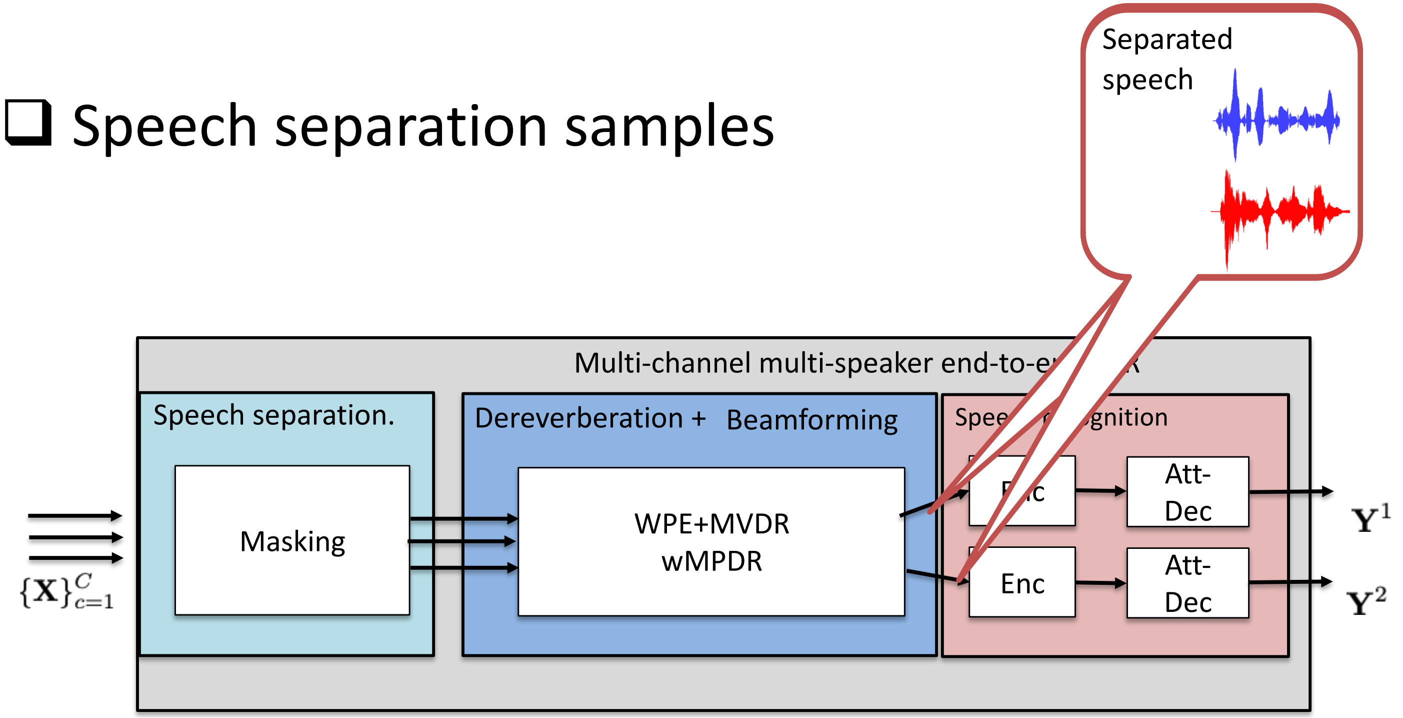
Experimental Results

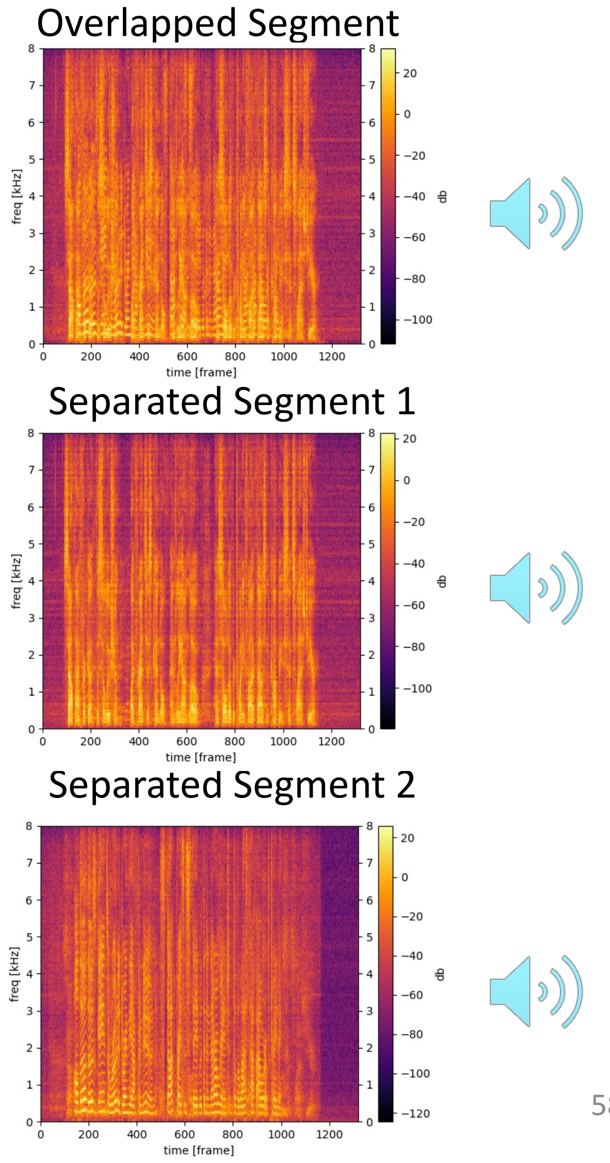
- Multi-speaker speech recognition task (Spatialized wsj-2mix corpus)

Model	Word error rate (WER) (eval)
single-channel multi-speaker (noisy speech)	29.43
single-channel multi-speaker (with beamforming, pipeline)	21.75
MIMO-Speech with joint dereverberation/beamforming (end-to-end)	**15.01**

Neural beamformer learns separation ability!

- ❑ Speech separation samples



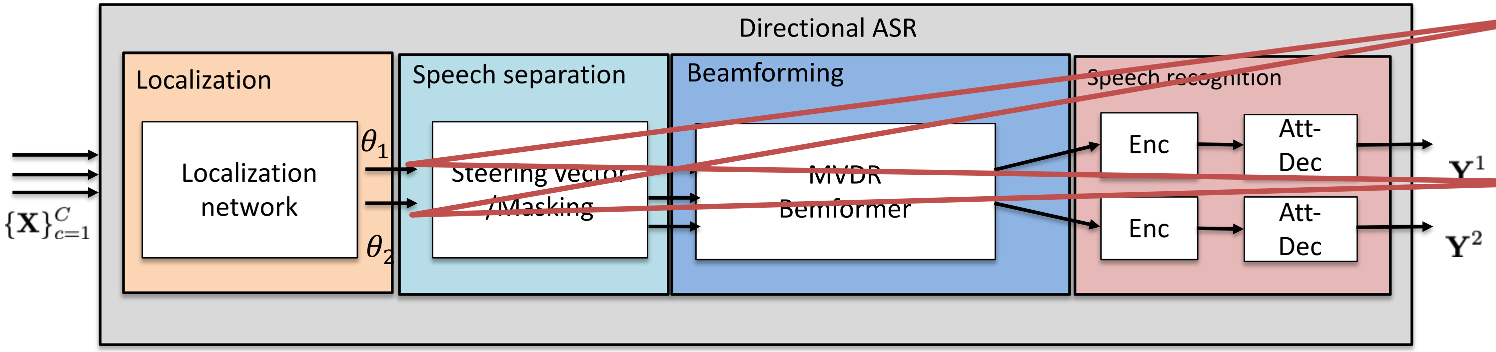


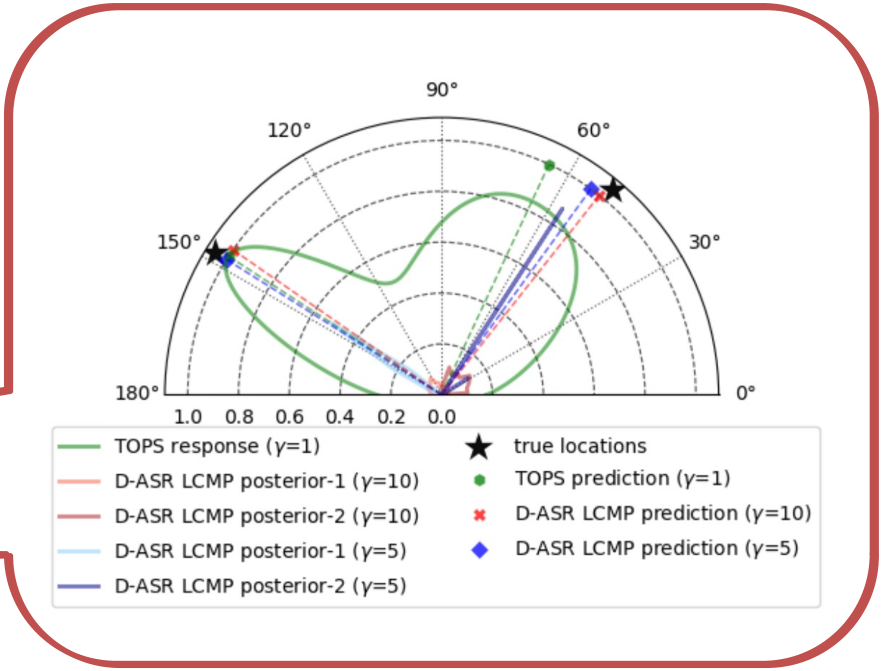
- ❑ The mask-based neural beamformer and speech recognition are jointly optimized via **ASR objective.**
- ❑ No explicit speech separation criterion is required
- ❑ **Explainable**

Table of contents

Directional ASR learns localization ability!

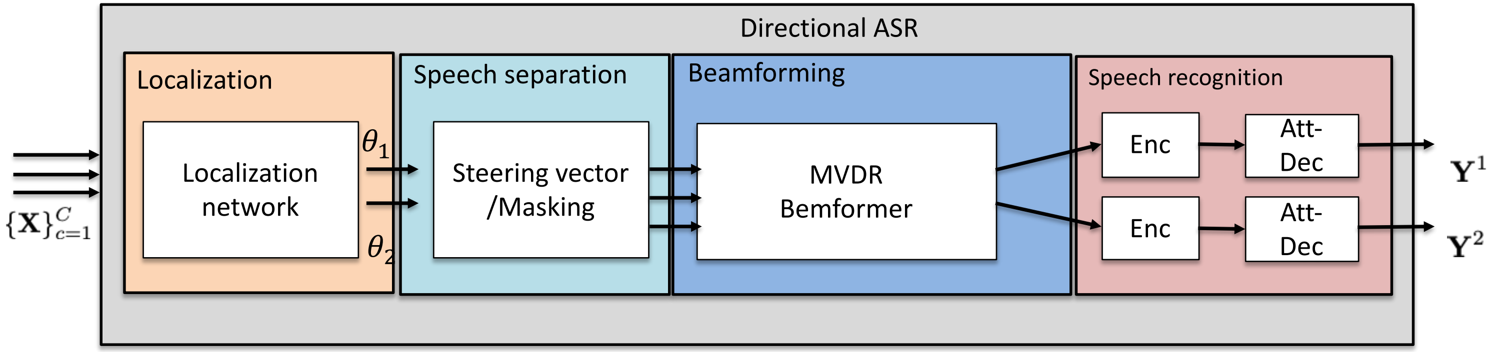
- ❑ Localization samples





- ❑ The localization network, mask-based neural beamformer and speech recognition are jointly optimized via **ASR objective**
- ❑ **Explainable**
- ❑ We can realize "who is speaking when, what, and **where**"

Further integrations?



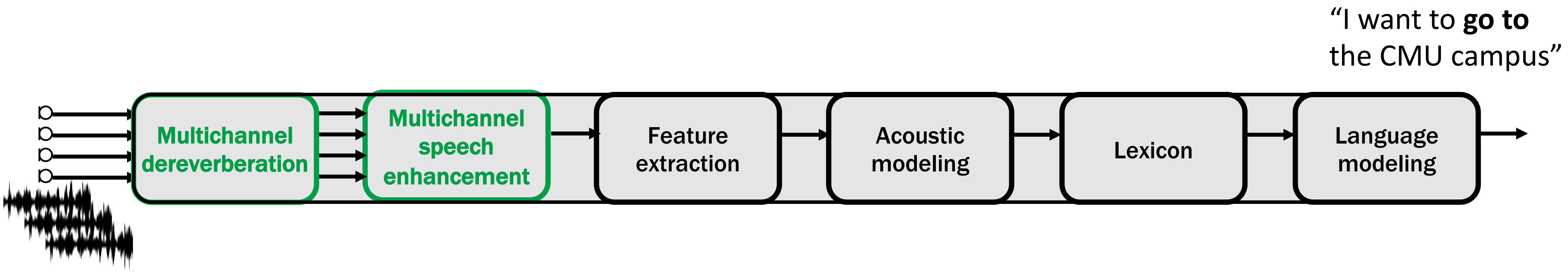
- Number of speakers?
- Audio event classification/detection?
- Emotion/Sentiment recognition?
- Room information?
- Spoken language understanding?
- Any idea?

Discussions
Modular system vs. End-to-End system

"I want to **go to** the CMU campus"

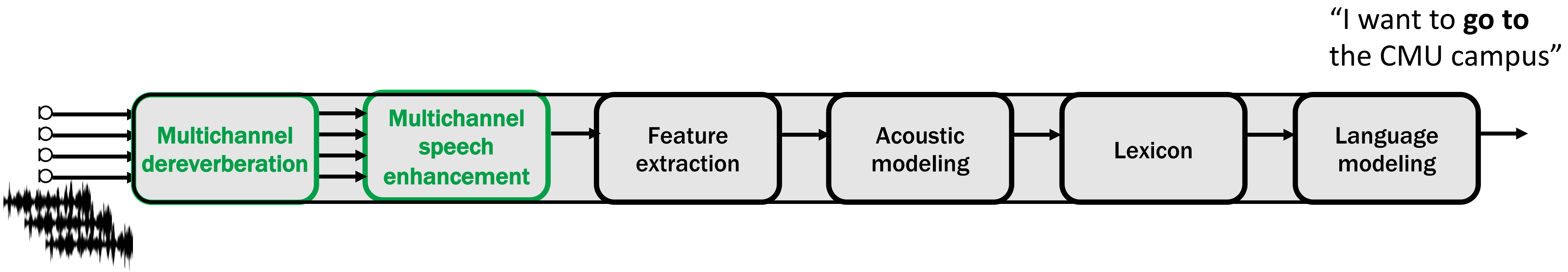
- Train a deep network that **directly** maps speech signal to the target letter/word sequence → **We don't know what's happening. We lose the explainability.**
- Greatly **simplify** the complicated model-building/inference process
- **Integrate** various modules by **optimizing the entire network** with a single objective function → **Difficult to optimize it**

Discussions
Explainable neural network



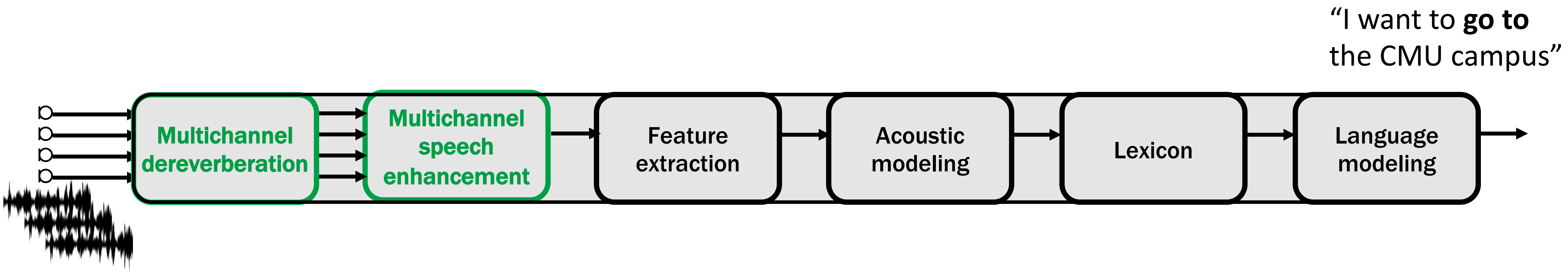
- Train a deep network that **directly** maps speech signal to the target letter/word sequence → **We don’t know what’s happening. We lose the explainability.** → **We can keep the explainability**
- Greatly **simplify** the complicated model-building/inference process
- **Integrate** various modules by **optimizing the entire network** with a single objective function → **Difficult to optimize it** → **Easy to optimize with model constraint, pre-training, ease of debugging with the explainability**

Discussions
Explainable neural network



- Train a deep network that **directly** maps speech signal to the target letter/word sequence → **We don't know what's happening. We lose the explainability.** → **We can keep the explainability**
- Greatly **simplify** the complicated model-building/inference process → **Very complicated (again) ☹**
- **Integrate** various modules by **optimizing the entire network** with a single objective function → **Difficult to optimize it** → **Easy to optimize with model constraint, pre-training, ease of debugging with the explainability**

Discussions
Explainable neural network



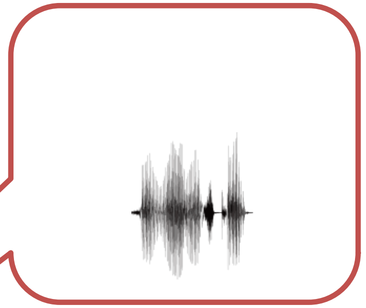
- Train a deep network that **directly** maps speech signal to the target letter/word sequence → **We don't know what's happening. We lose the explainability.** → **We can keep the explainability**
- Greatly **simplify** the complicated model-building/inference process → **Very complicated (again) ☹**
- **Integrate** various modules by **optimizing the entire network** with a single objective function → **Difficult to optimize it** → **Easy to optimize with model constraint, pre-training, ease of debugging with the explainability**

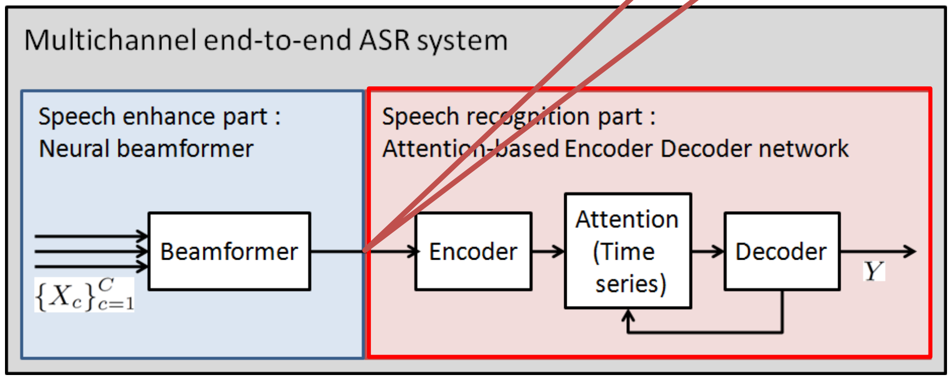
ESPnet

Open source is one solution for dealing with reproducibility, but…

Discussions

- Is it really better?
- The rich sound information was "**projected**" to the enhanced (clean) speech space
  - The sound event and room acoustic information were **disappeared**.
  - We need to provide supplemental information or original information to avoid this projection problem
  - Taking over the drawback of the modular system
- Let's go to the next topic!





It's like a measurement problem in the quantum theory(?)

Make it visible
Interpretable

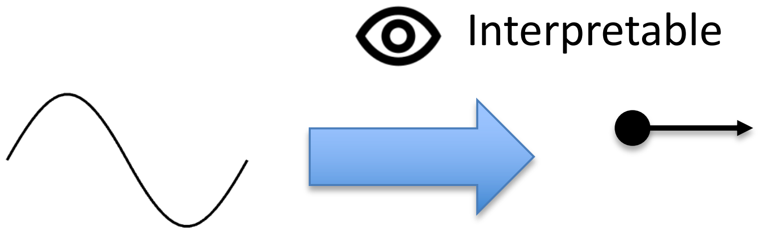
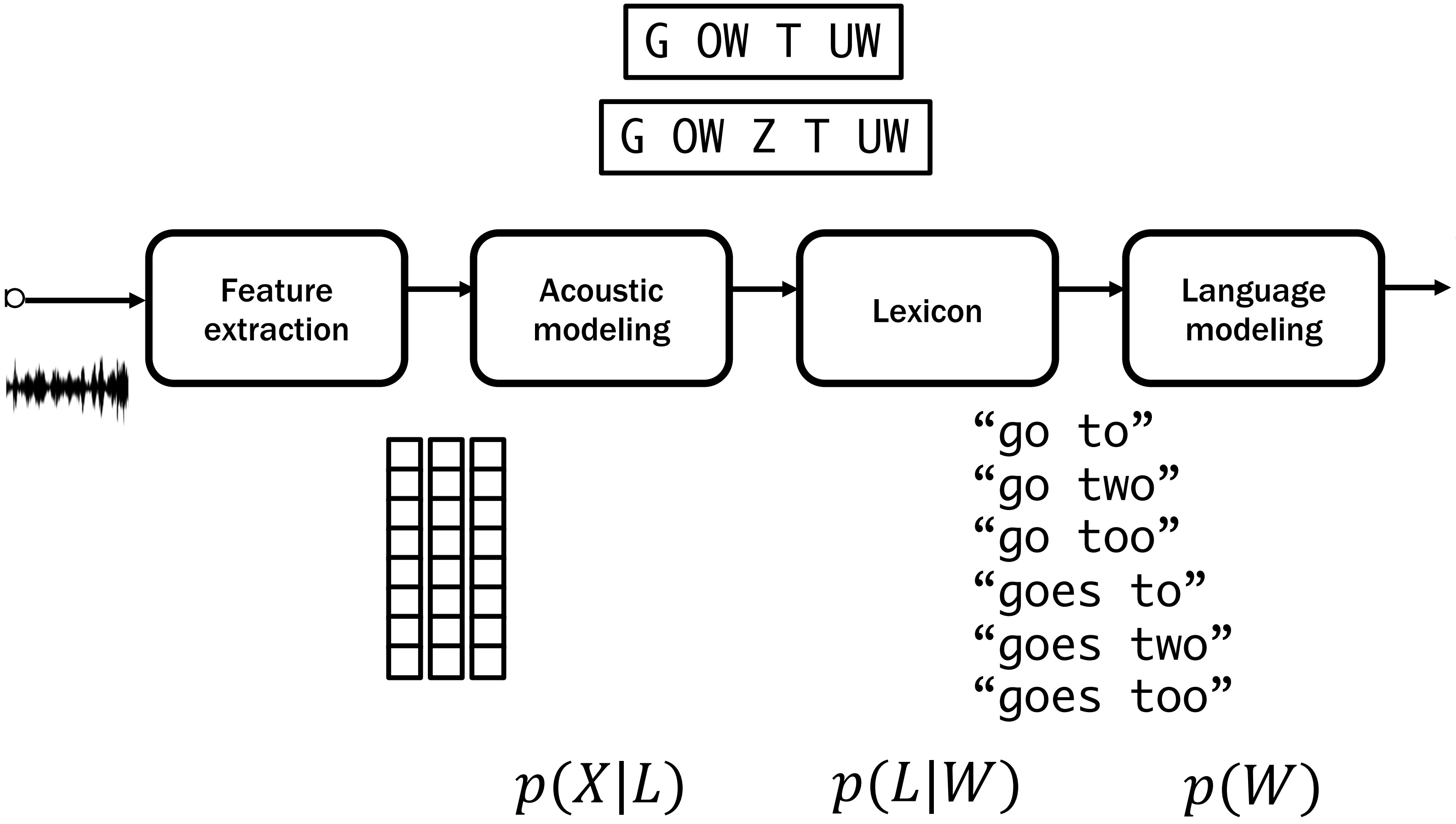


Table of contents

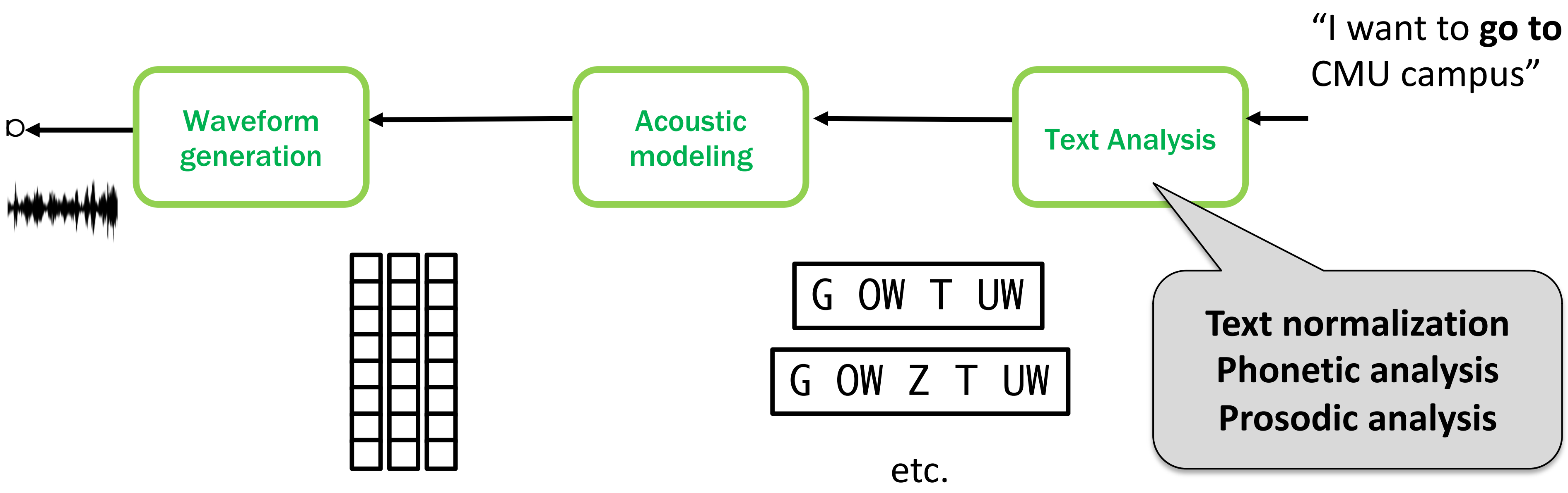
1. **End-to-End Integration** of Speech Recognition and Speech Enhancement
   X = Speech Recognition
   2) X += Denoising, Dereverberation
   3) X += Speech Separation
   4) X += Speech Localization
2. **End-to-End Integration** of Speech Recognition and Speech Synthesis
3. Discussion

Speech recognition pipeline

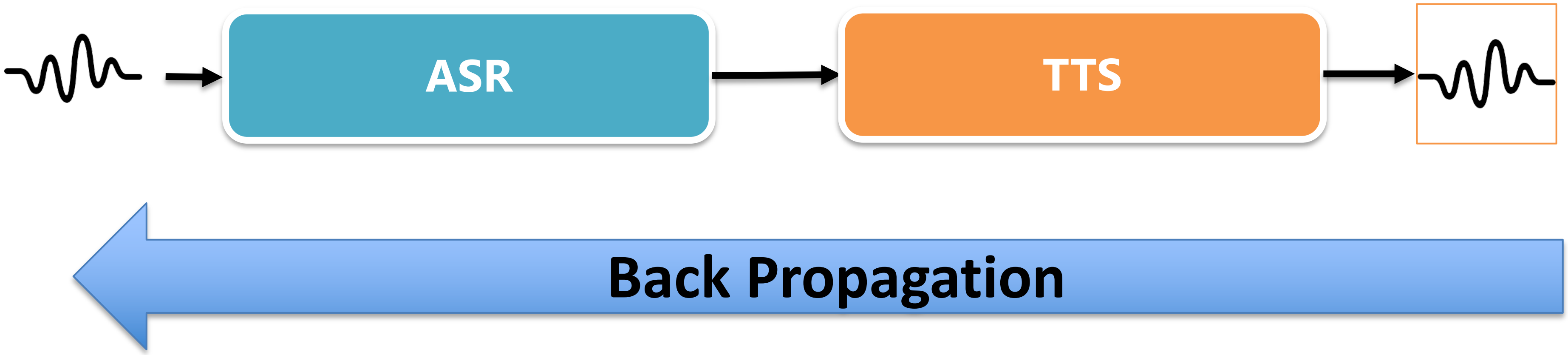


“I want to **go to** CMU campus”

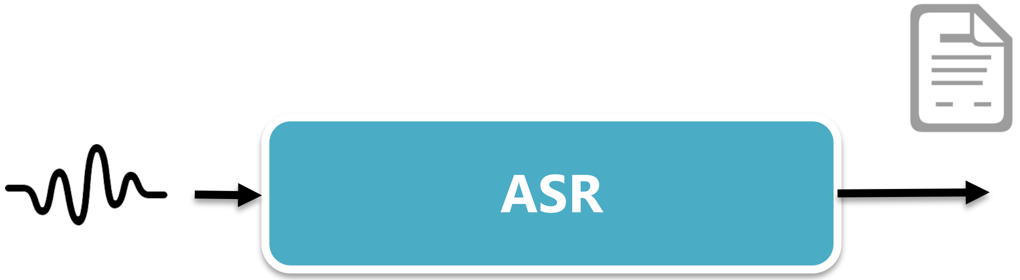
Speech synthesis pipeline
(or Text To Speech, TTS)



ASR + TTS feedback loop
→Unpaired data training

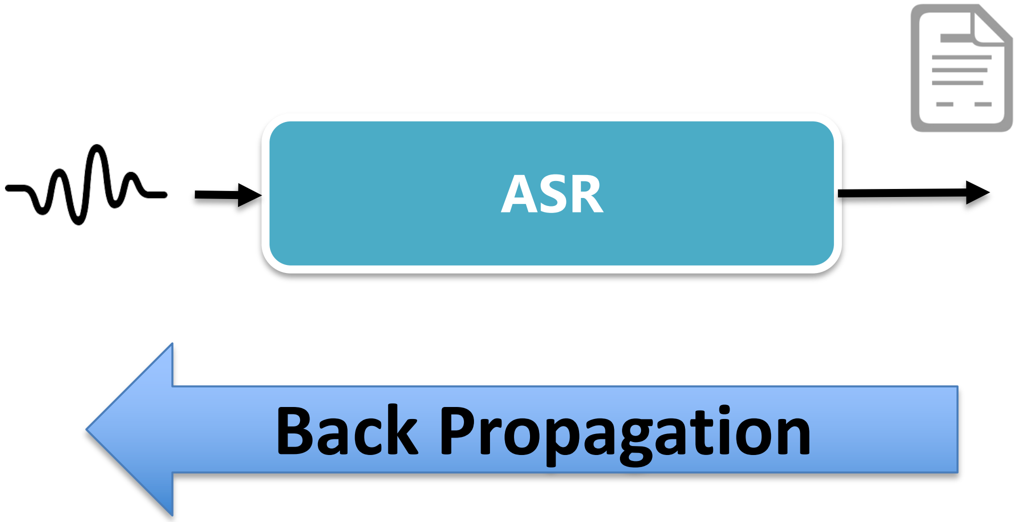


ASR + TTS feedback loop
→Unpaired data training



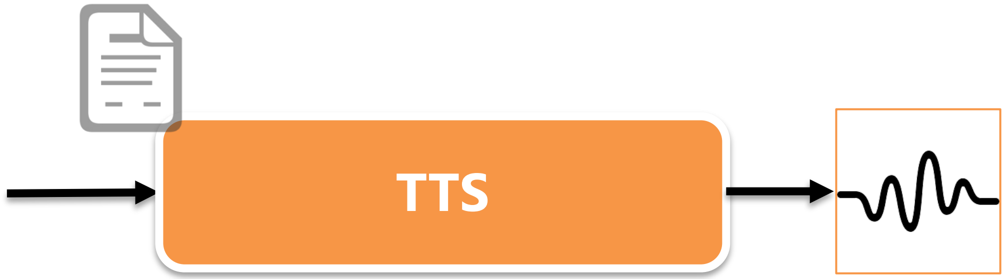
Train ASR with the *pair of audio and text*

ASR + TTS feedback loop
→Unpaired data training



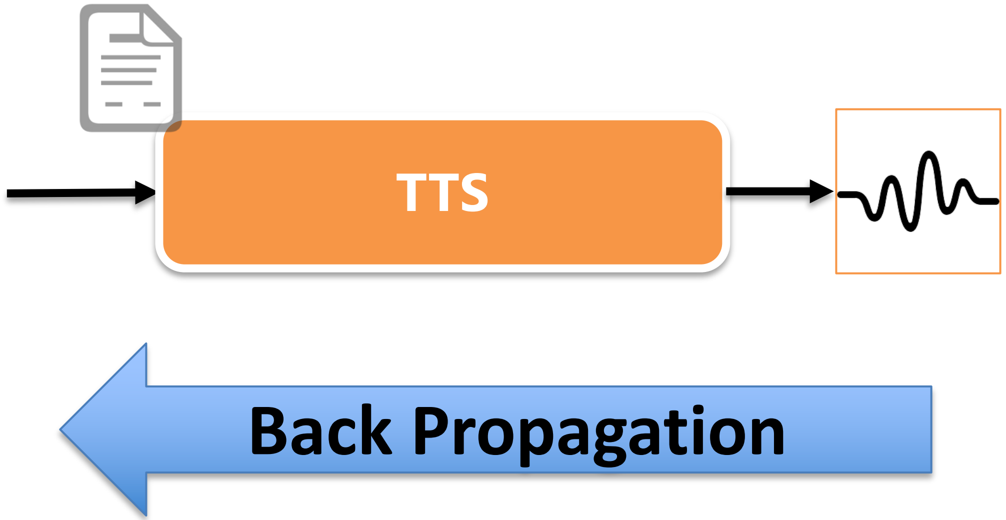
Train ASR with the *pair of audio and text*

ASR + TTS feedback loop →Unpaired data training



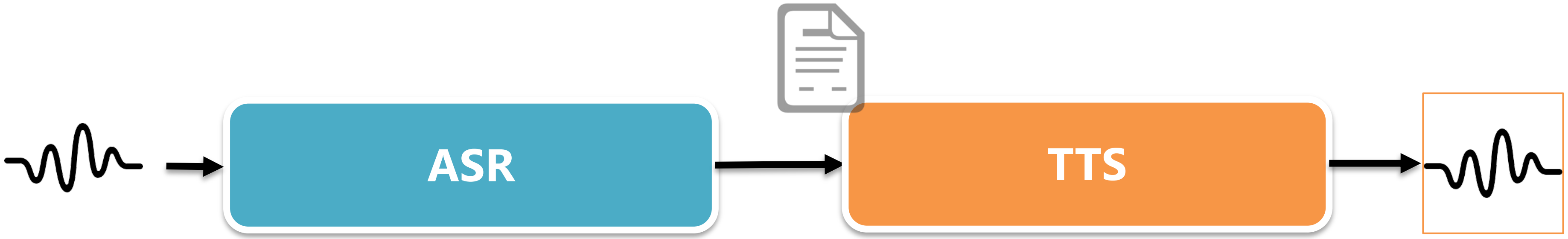
Train TTS with the ***pair of audio and text***

ASR + TTS feedback loop
→Unpaired data training



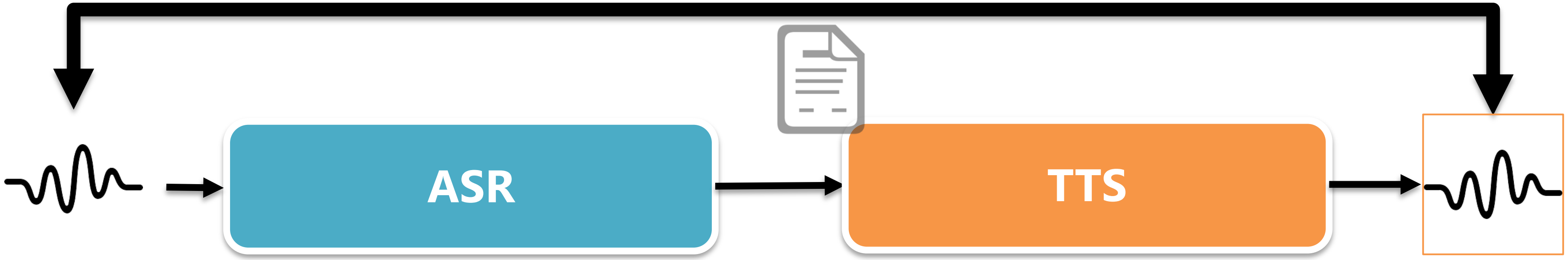
Train TTS with the *pair of audio and text*

ASR + TTS feedback loop
→Unpaired data training



Only audio data to train both ASR and TTS

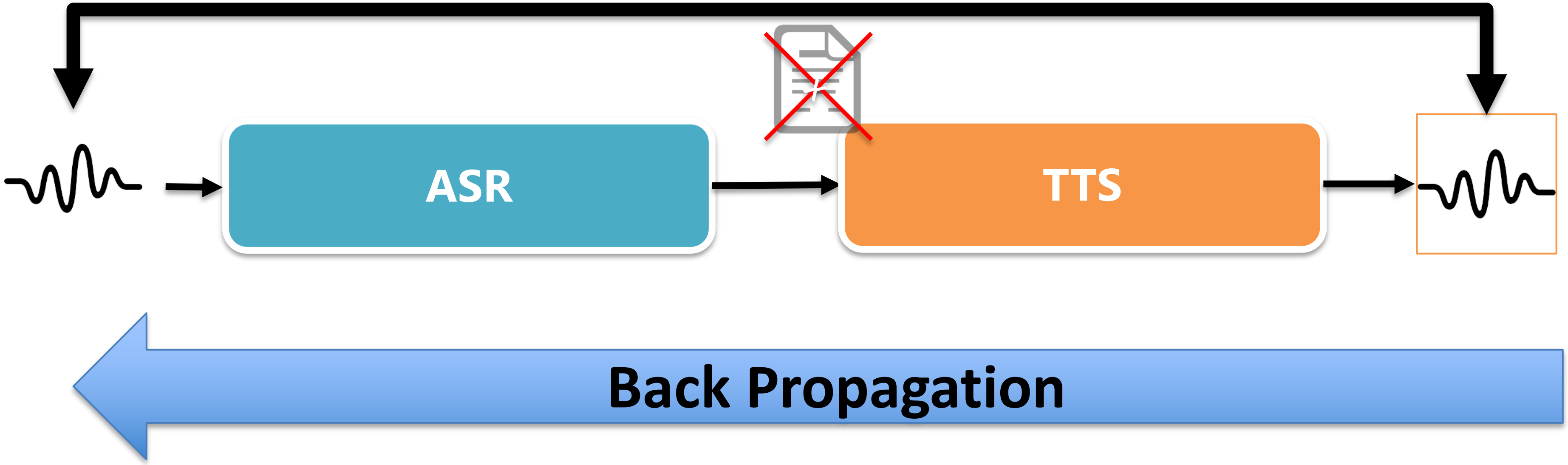
ASR + TTS feedback loop
→Unpaired data training



Should be similar

Only audio data to train both ASR and TTS

ASR + TTS feedback loop
→Unpaired data training

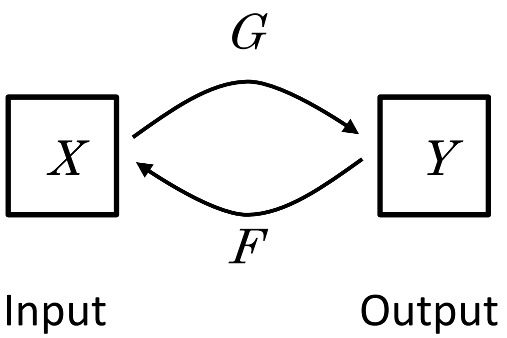


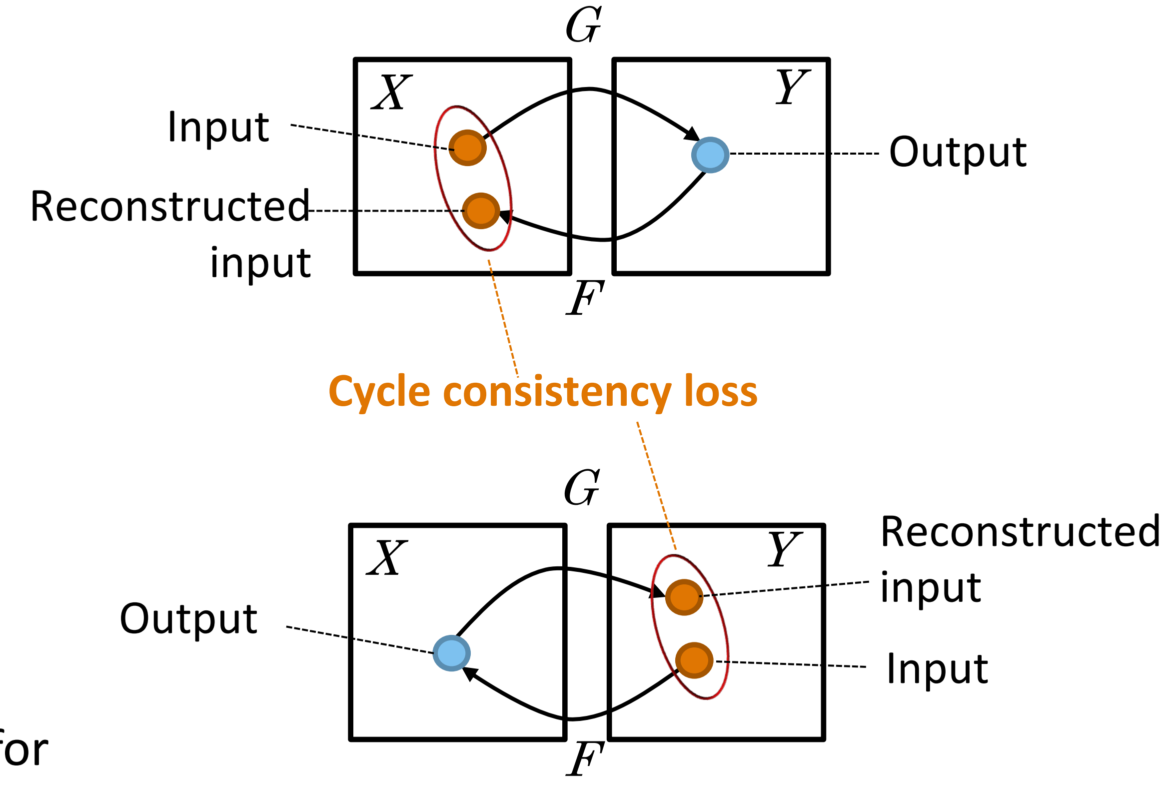
Only audio data to train both ASR and TTS

We do not need a pair data!!!

Training with cycle consistency loss

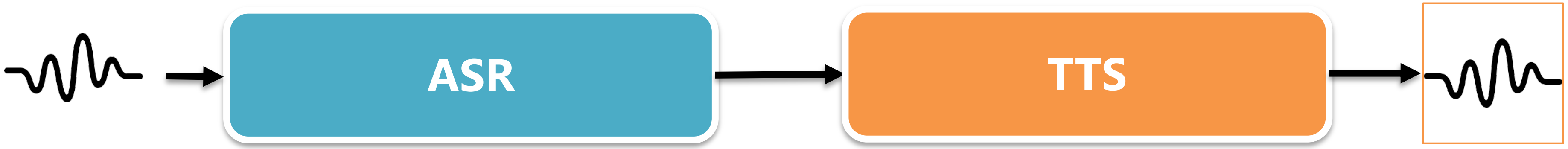
- Input and reconstruction should be similar
- No need for paired data

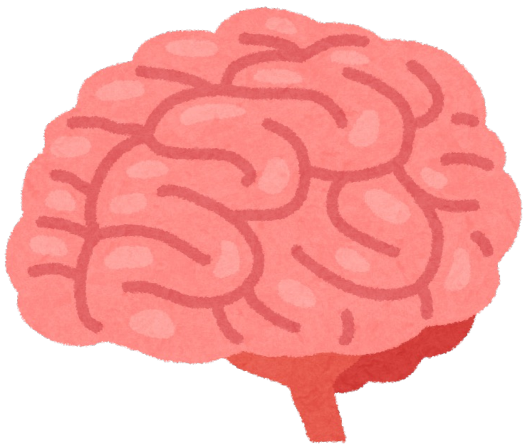




The idea has been proposed for machine translation [Xia+’16] and image-to-image transformation [Zhu+’18].

Joint modeling of ASR and TTS is quite natural for human learning



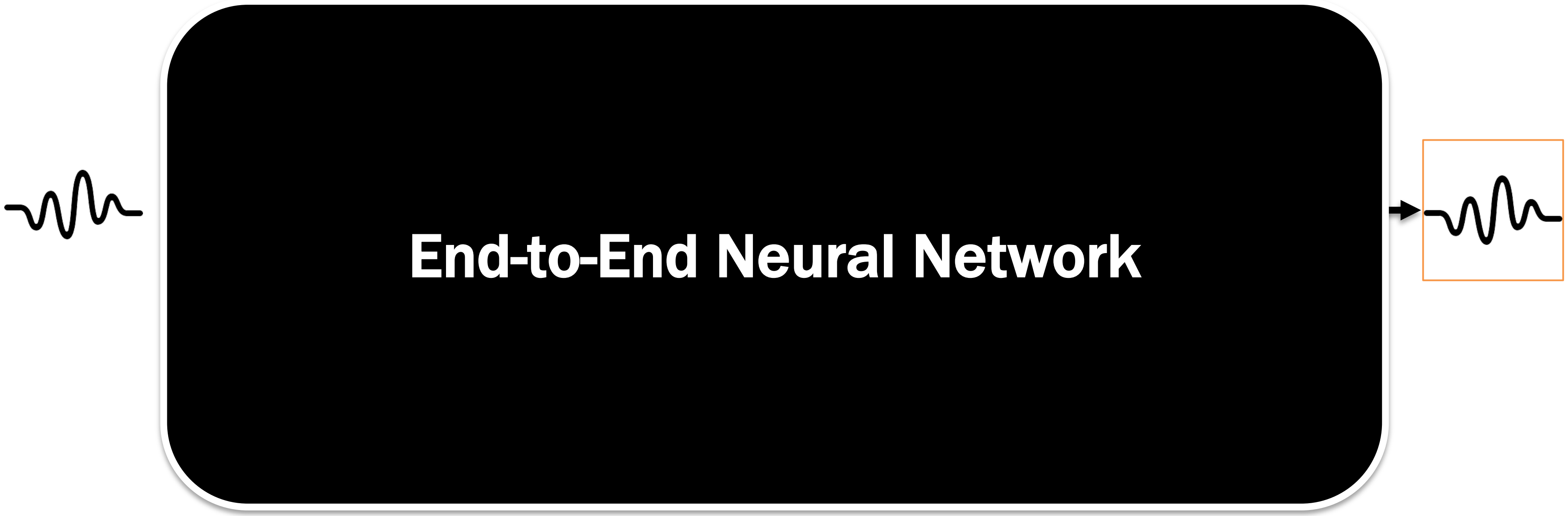


Joint modeling of speech recognition and synthesis is a very important concept in **neuroscience**

- **Phonological loop**
- **Speech chain**
- **Motor theory**

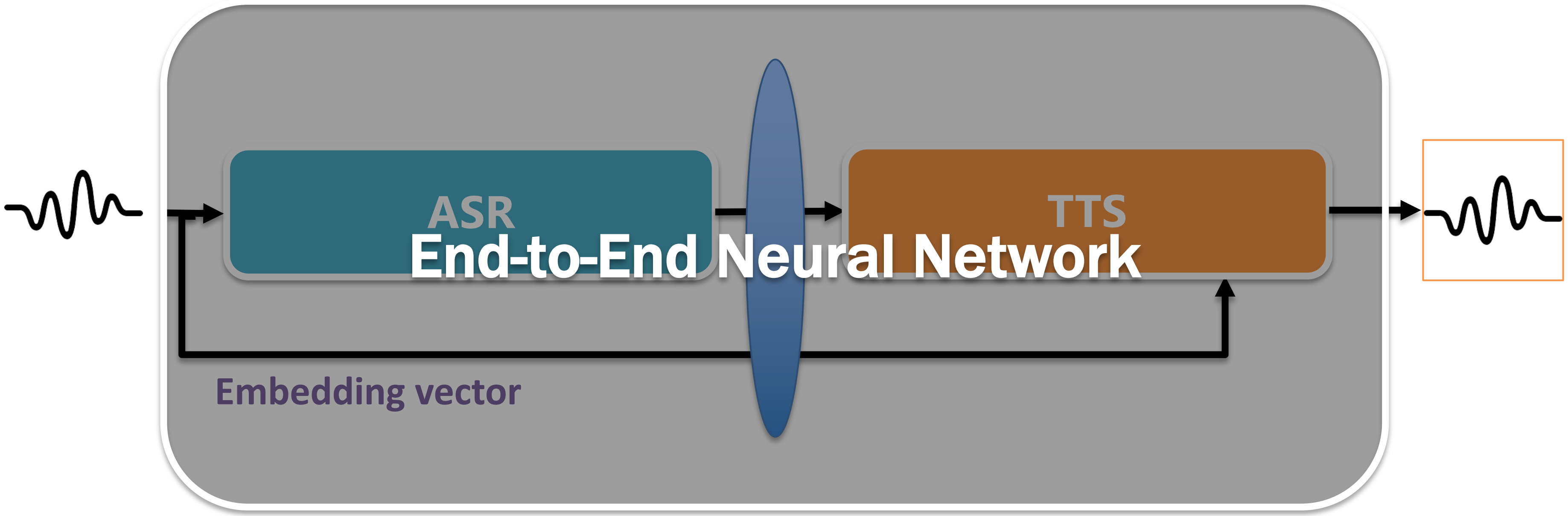
[Tjandra (2017)]
[Hori (2019)]
[Baskar (2021)]

Audio Disentanglement



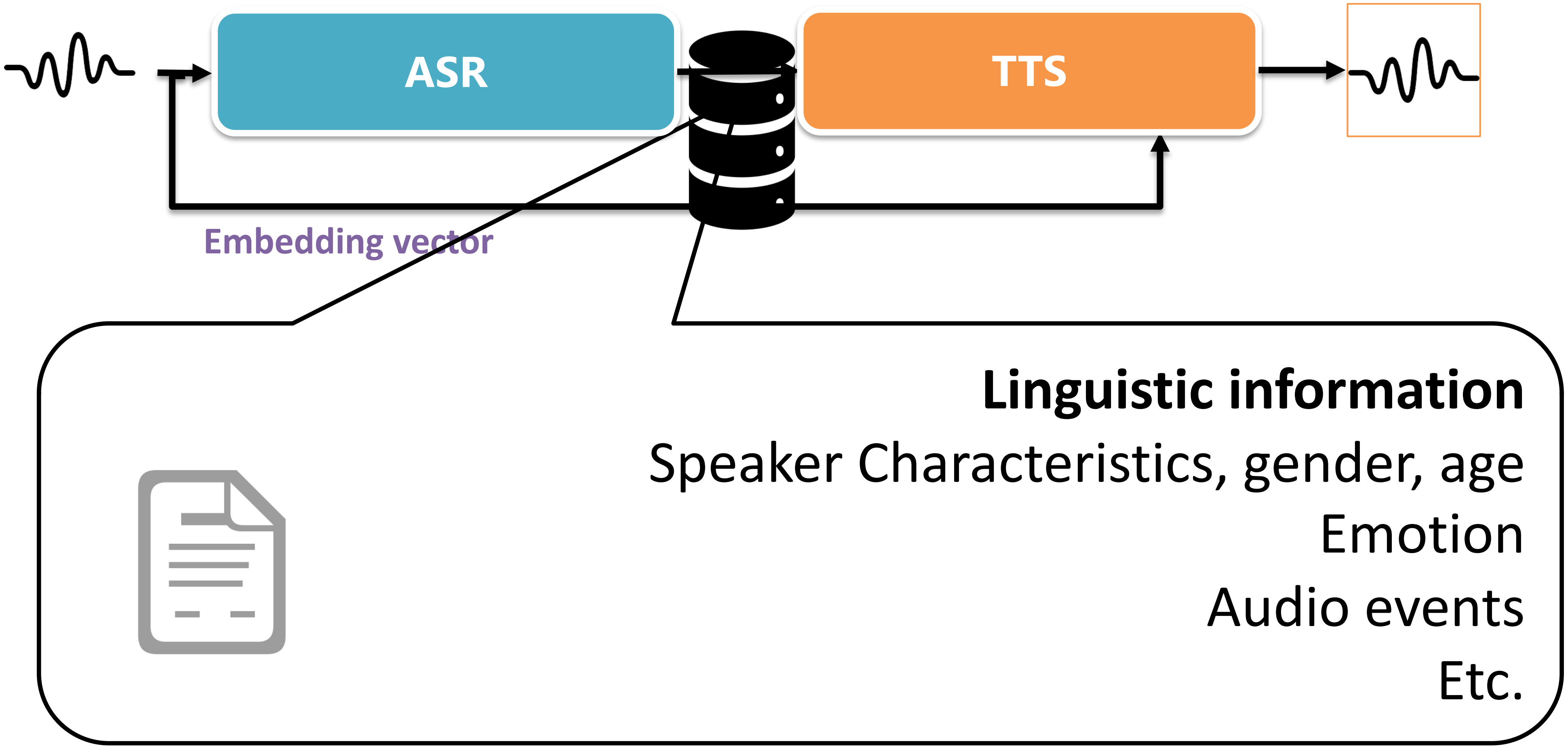
Autoencoder

Audio Disentanglement

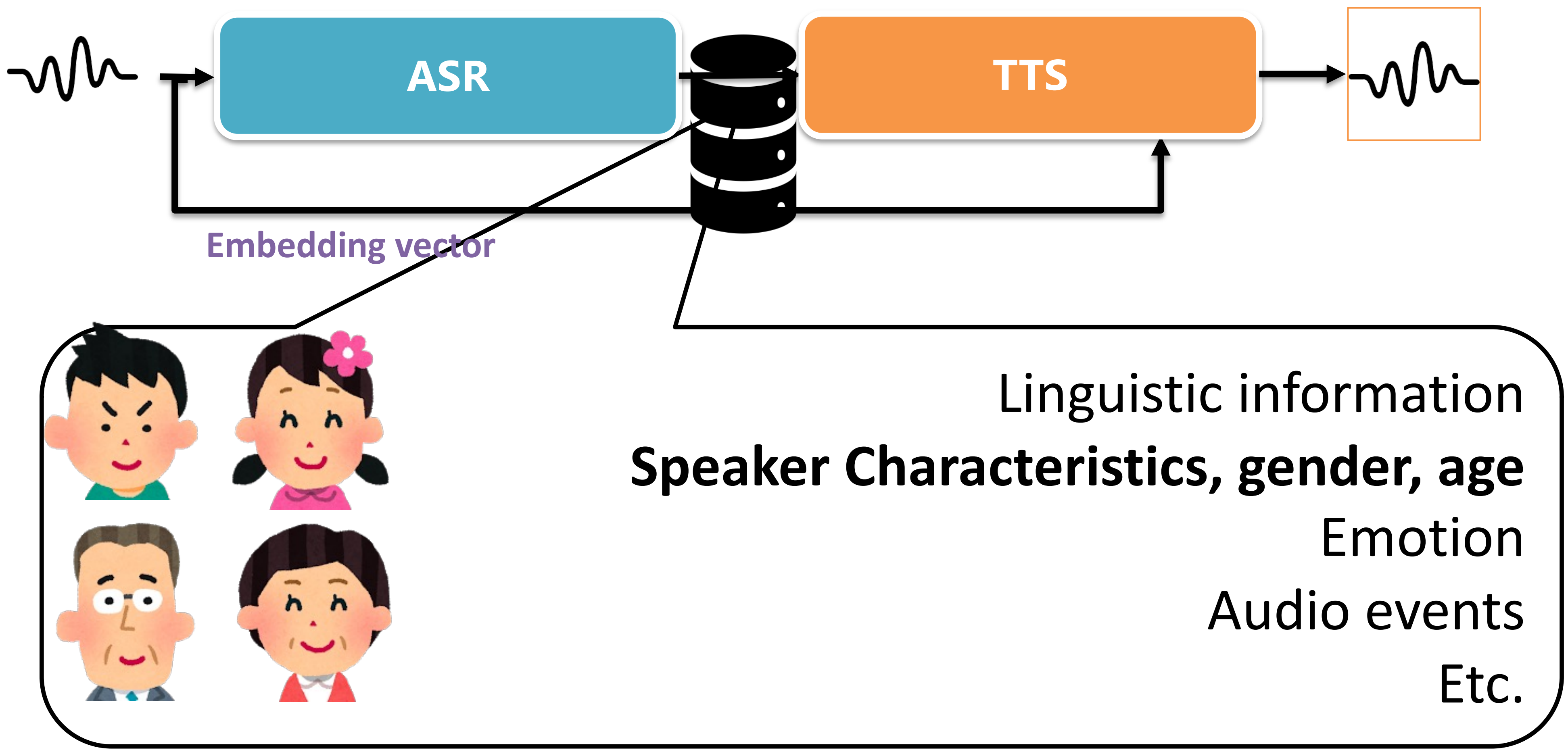


Interpretable neural network

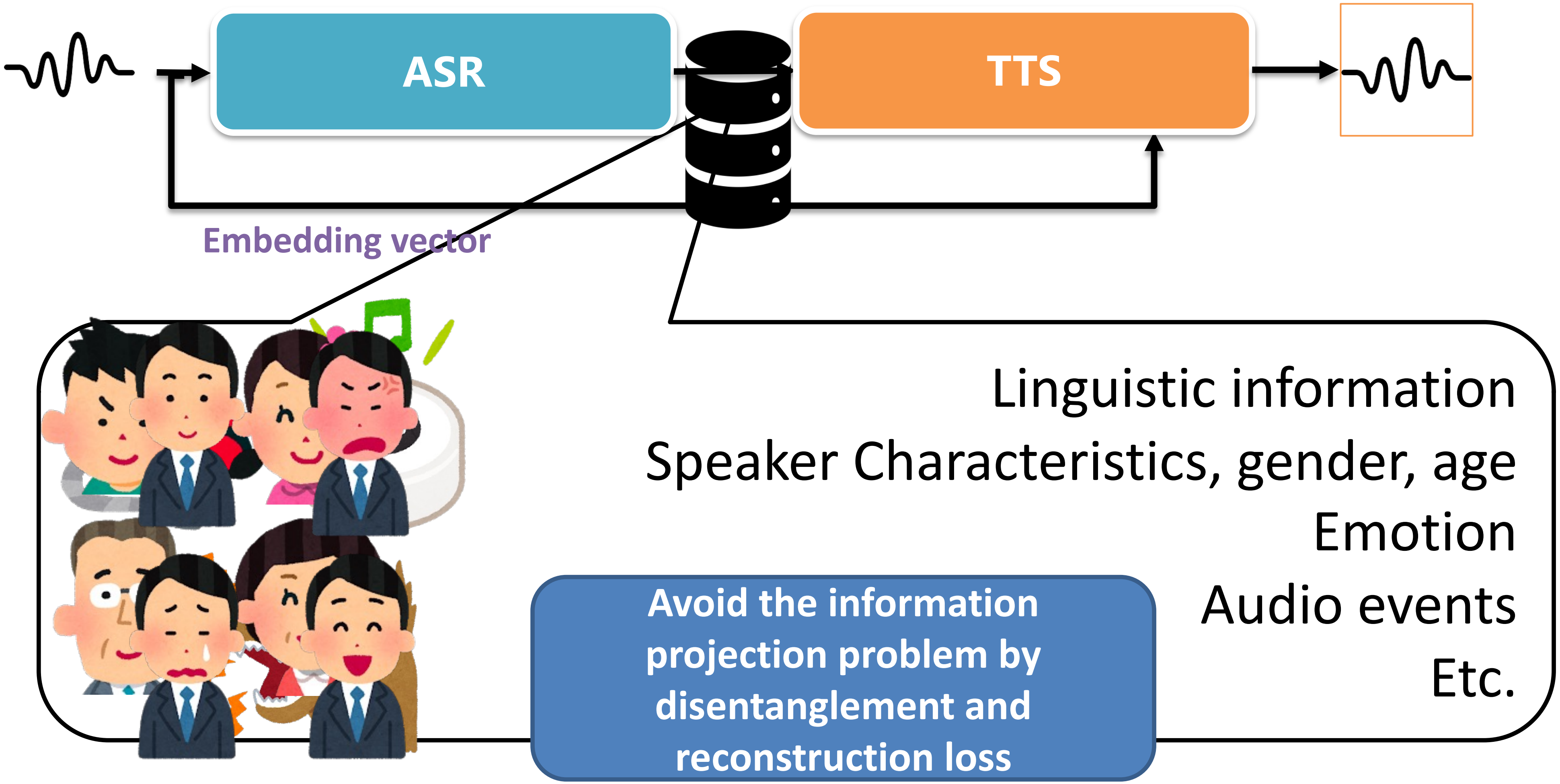
Audio Disentanglement



Audio Disentanglement

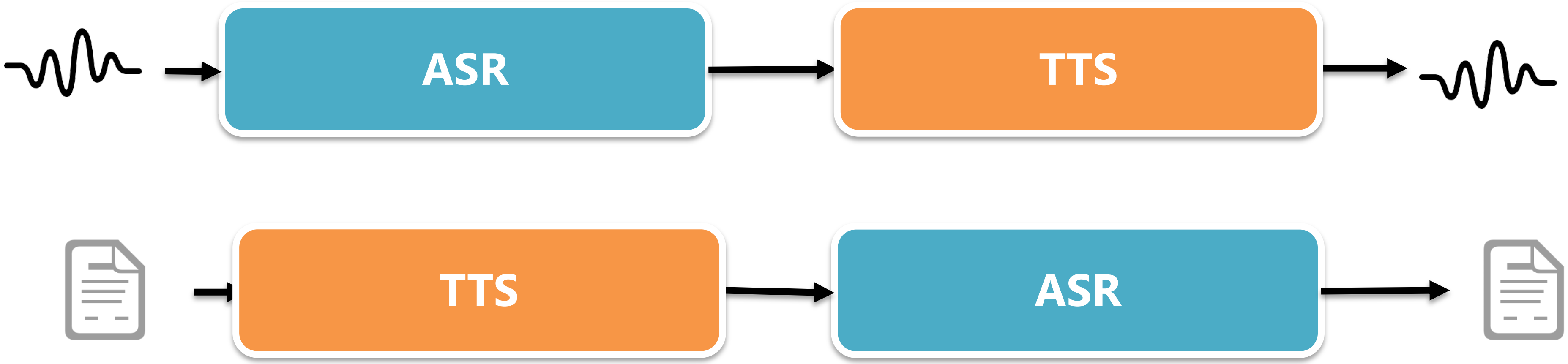


Audio Disentanglement

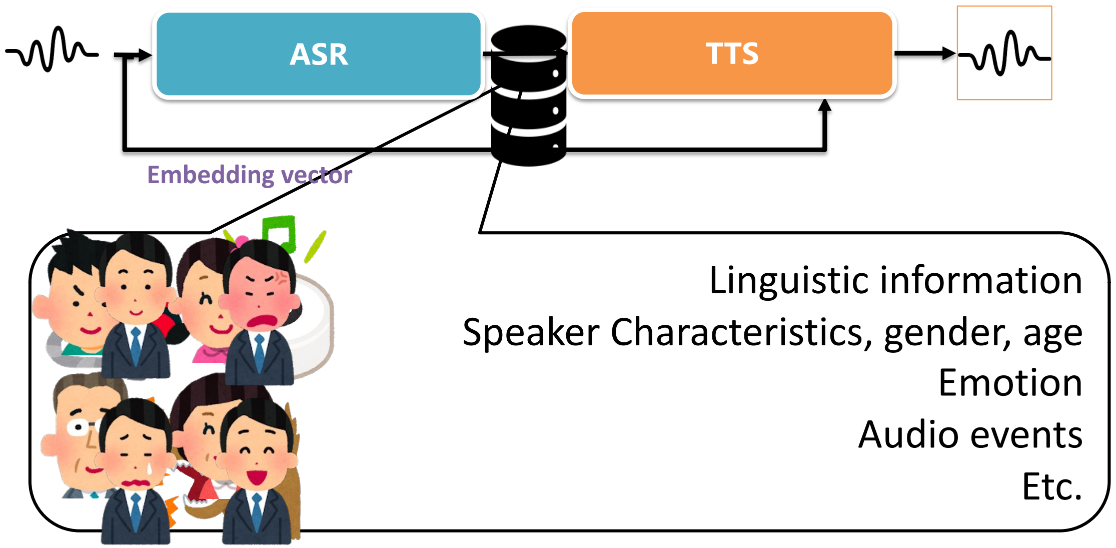


Both audio-only and text-only cycles

- Consider two cycle consistencies (**duality**)
  - Audio only: ASR+TTS
  - Text only: TTS+ASR

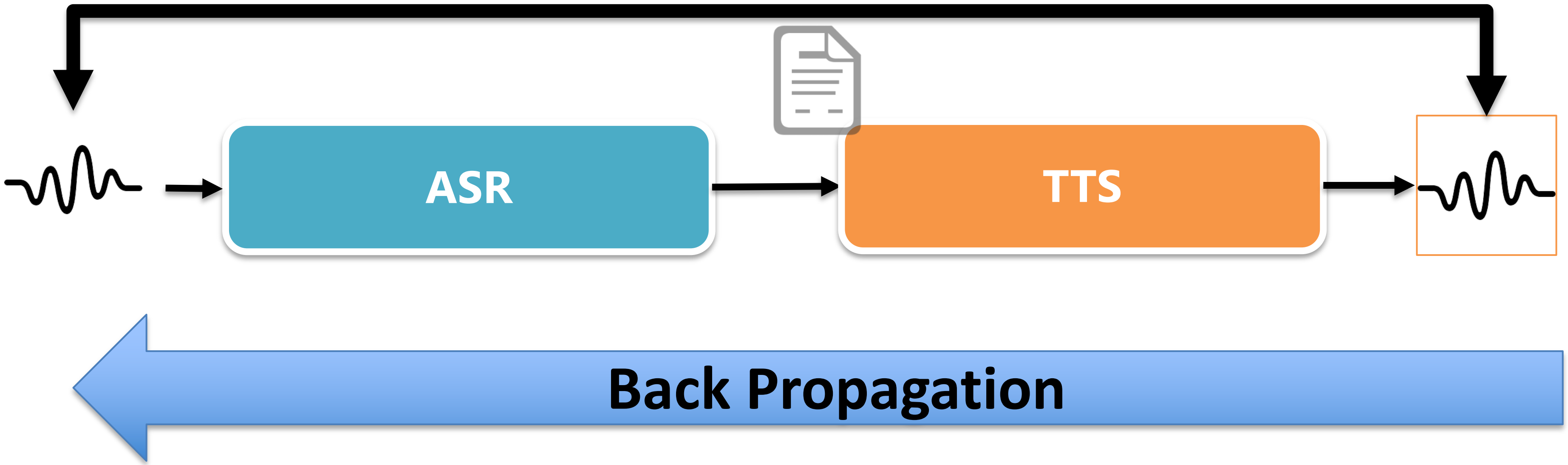


Audio Disentanglement

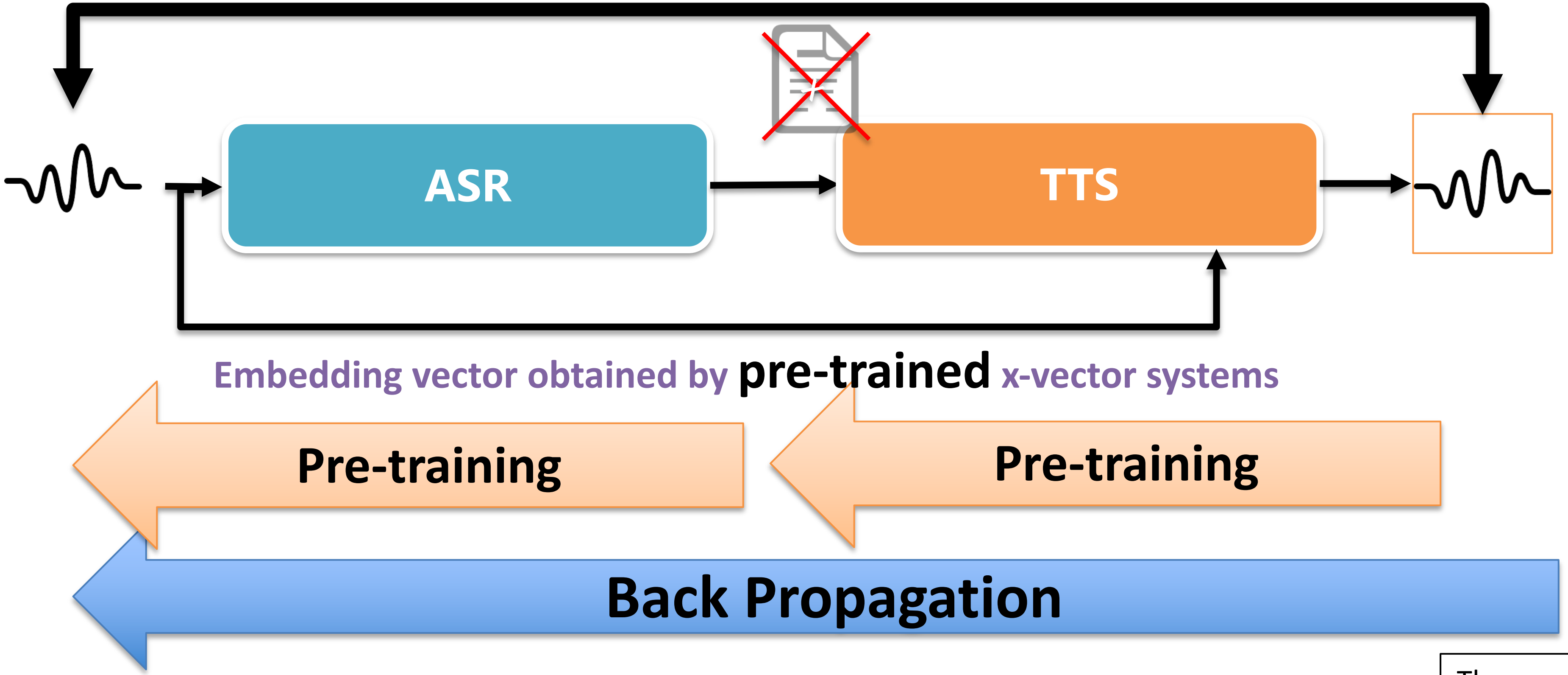


- One of the dream technologies
- No pair data
- All speech processing models are **integrated** and jointly trained
  - ASR, TTS, SID, Emotion, Audio event, etc.

Current realization



Current realization



The argmax problem is handled by REINFORCE or gumbel softmax

- We deal with three models, ASR, TTS, and SID (embedding)
- Pre-train all three models and back propagation with speech only data for ASR and TTS models (SID part is fixed)

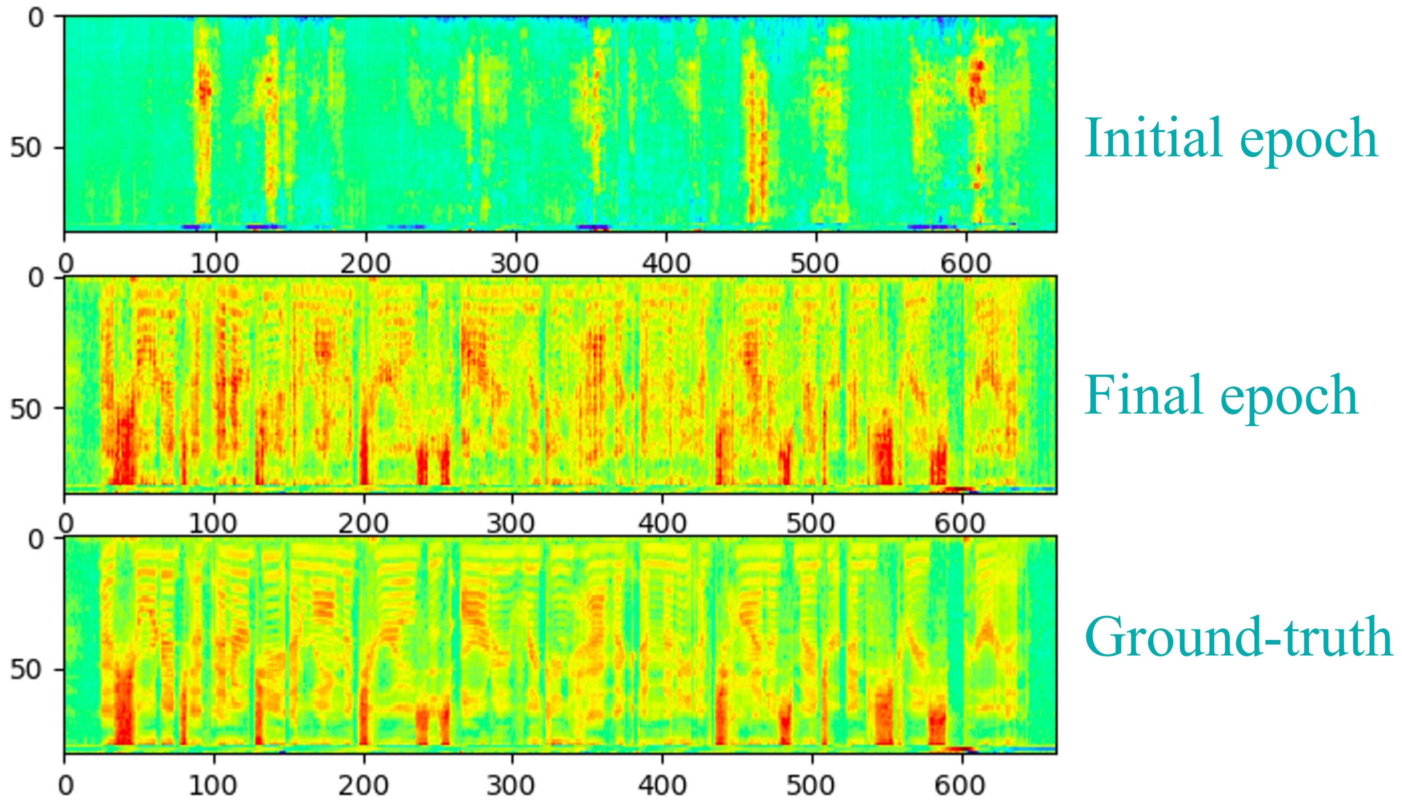
Current realization
Experimental results [Hori+(2019), Baskar+(2019)]

- English Librispeech corpus (Audio book)
  - Paired data: 100h to train ASR and TTS [Shen+ (2018)] models first
  - Unpaired data: 360h (only audio and/or text only): cycle consistency training

Model	**Eval-clean CER / WER [%]**
Baseline	8.8 / 20.7
+ text-only cycle E2E	8.0 / 17.0
+ both audio-only/text-only cycle E2E	**7.6 / 16.6**

Cycle-consistency E2E improved the ASR performance

Improving TTS quality as well!



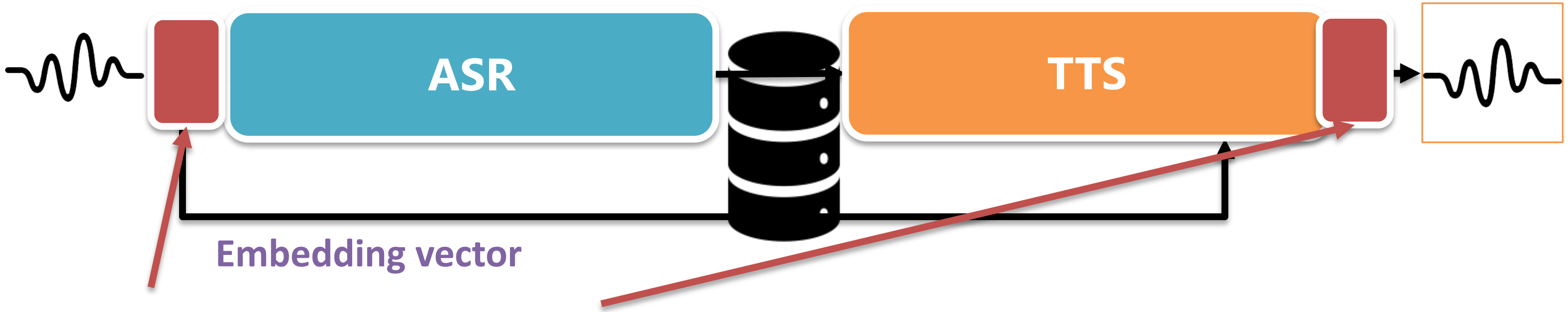
REFERENCE TEXT:

"has never been surpassed"

- Initial epoch
- Final epoch

Future directions

- Incorporate more self-supervised learning ideas
  - It's the same problem setup
  - This direction has a **duality** (Speech → Speech, Text → Text)
- More integrations

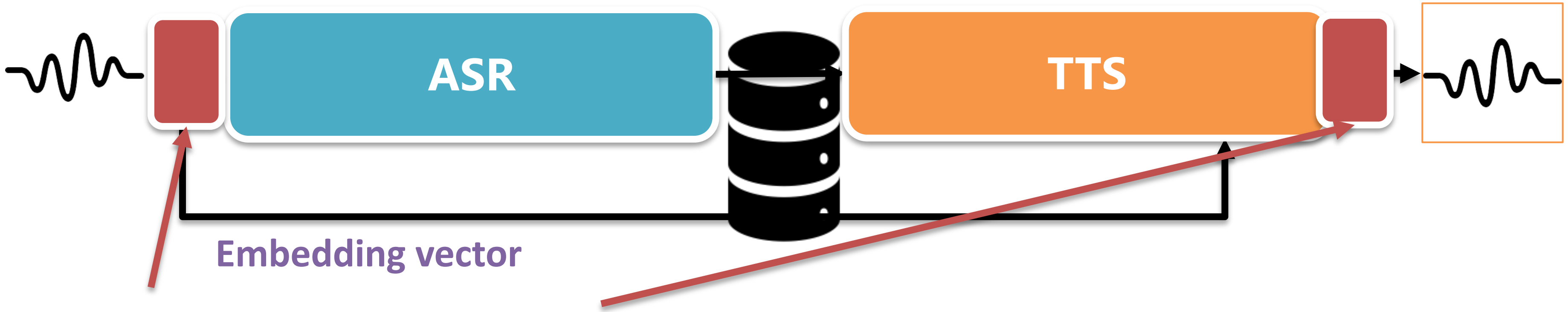


  - Enhancement + Audio generation (Connect part 1 and part 2)

- It is too difficult to make it train from scratch unlike speech enhancement + speech recognition

Future directions

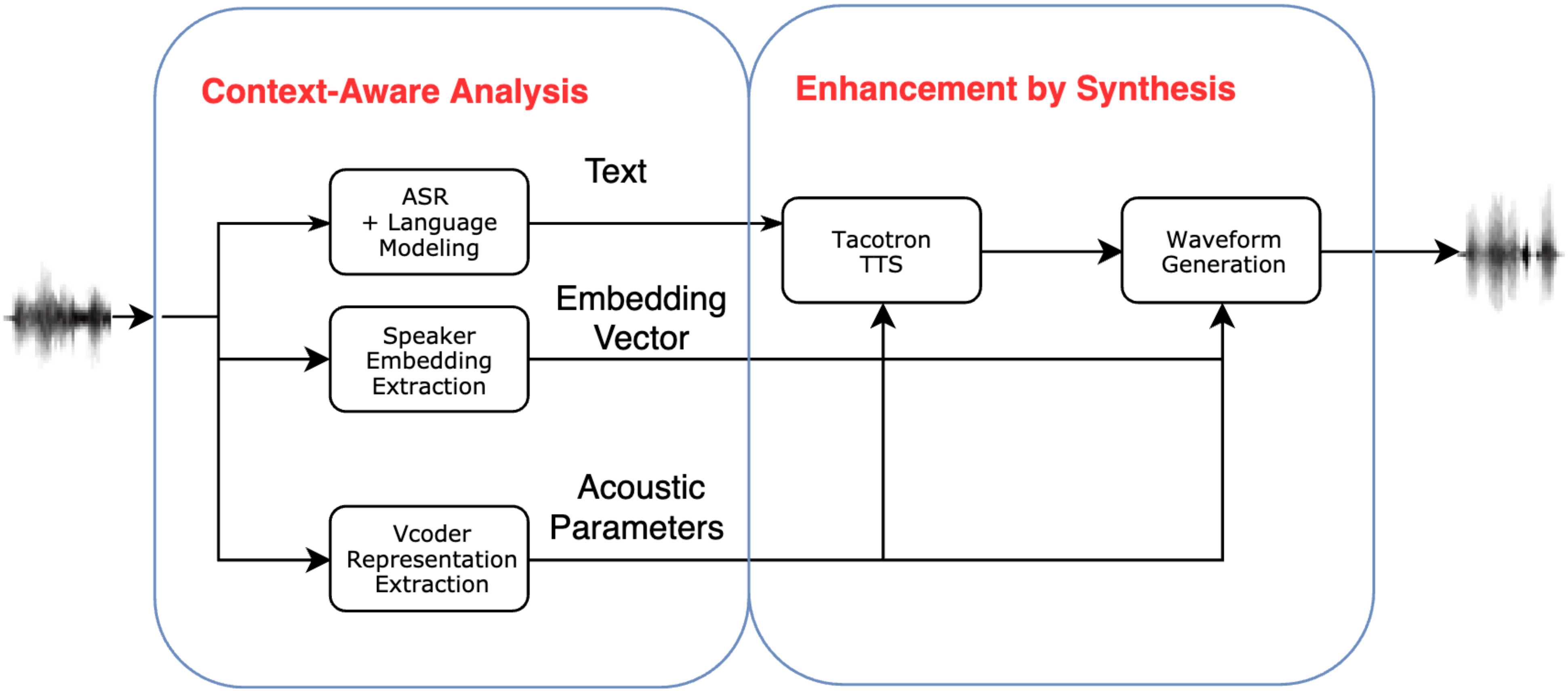
- Incorporate more self-supervised learning ideas
  - It's the same problem setup
  - This direction has a **duality** (Speech → Speech, Text → Text)
- More integrations



  - Enhancement + Audio generation (Connect part 1 and part 2)

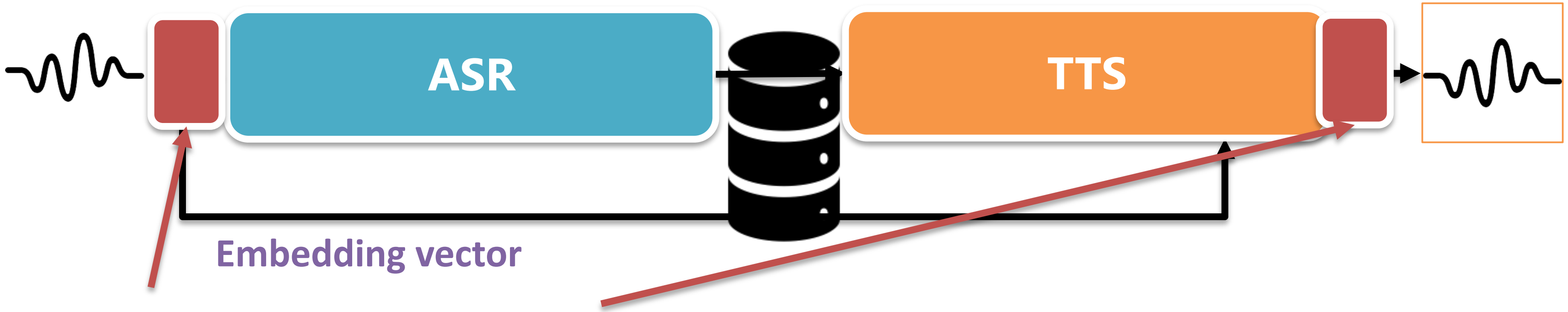
- It is too difficult to make it train from scratch unlike speech enhancement + speech recognition
  - Finding a student and sponsor

Complete disentanglements of speech signals (One of my rejected NSF proposals)



Future directions

- Incorporate more self-supervised learning ideas
  - It's the same problem setup
  - This direction has a **duality** (Speech → Speech, Text → Text)
- More integrations



  - Enhancement + Audio generation (Connect part 1 and part 2)
- It is too difficult to make it train from scratch unlike speech enhancement + speech recognition
  - Finding a student and sponsor

Thanks a lot!