End-to-End Speech Processing: From Pipeline to Integrated Architecture

Shinji Watanabe
Center for Language and Speech Processing
Johns Hopkins University

Joint work with John Hershey, Takaaki Hori, Shigeru Katagiri, Suyoun Kim, Tsubasa Ochiai, Tomoki Hayashi, Hiroshi Seki, Jonathan Le Roux, Murali Karthick Baskar, Ramon Fernandez Astudillo, Xuankai Chang, Aswin Shanmugam Subramanian, etc.

CENTER FOR LANGUAGE AND SPEECH PROCESSING

Frederick Jelinek (1932 –2010)

Statistical speech recognition and machine translation

"Every time I fire a linguist, the performance of the speech recognizer goes up”

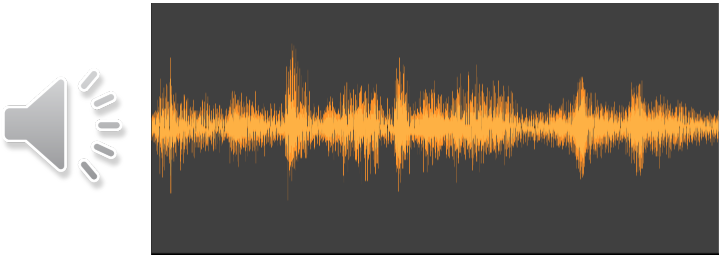
1972 - 1993: IBM

1993 - 2010: JHU and established CLSP

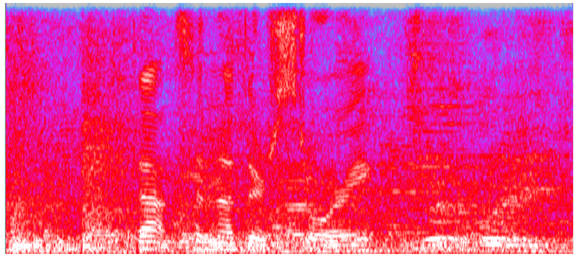
Jelinek methodology (1970s-)

Jelinek methodology (1970s-)

- Automatic Speech Recognition: Mapping ***physical signal sequence*** to ***linguistic symbol sequence***



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



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

"That's another story"

$$W = \{w_n \in \mathcal{V} | n = 1, \dots, N\}$$
$$N = 3$$

Jelinek methodology (1970s-)

$$\arg\max_{W} p(W|X)$$

X: Speech sequence
W: Text sequence

Jelinek methodology (1970s-)

L: Phoneme sequence

$$\arg\max_{W} p(W|X) = \arg\max_{W} p(X|W)p(W)$$
$$\approx \arg\max_{W,L} p(X|L)p(L|W)p(W)$$

- **Speech recognition**
  - $p(X|L)$: Acoustic model (Hidden Markov model)
  - $p(L|W)$: Lexicon
  - $p(W)$: Language model (n-gram)

Jelinek methodology (1970s-)

$$\arg\max_{\mathrm{W}} p(W|X) = \arg\max_{\mathrm{W}} p(X|W)p(W)$$
$$\approx \arg\max_{\mathrm{W,L}} p(X|L)p(L|W)p(W)$$

- **Speech recognition**
  - $p(X|L)$: Acoustic model (Hidden Markov model)
  - $p(L|W)$: Lexicon
  - $p(W)$: Language model (n-gram)

- Factorization
- Conditional independence (Markov) assumptions

Jelinek methodology (1970s-)

$$\arg\max_{W} p(W|X) = \arg\max_{W} p(X|W)p(W)$$

- **Machine translation**
  - $p(X|W)$: Translation model
  - $p(W)$: Language model

Jelinek methodology (1970s-)

$$\arg\max_{\mathrm{W}} p(W|X) = \arg\max_{\mathrm{W}} p(X|W)p(W)$$
$$\approx \arg\max_{\mathrm{W,L}} p(X|L)p(L|W)p(W)$$

- **Speech recognition**
  - $p(X|L)$: Acoustic model (Hidden Markov model)
  - $p(L|W)$: Lexicon
  - $p(W)$: Language model (n-gram)
- Continued 40 years

Jelinek methodology (1970s-)

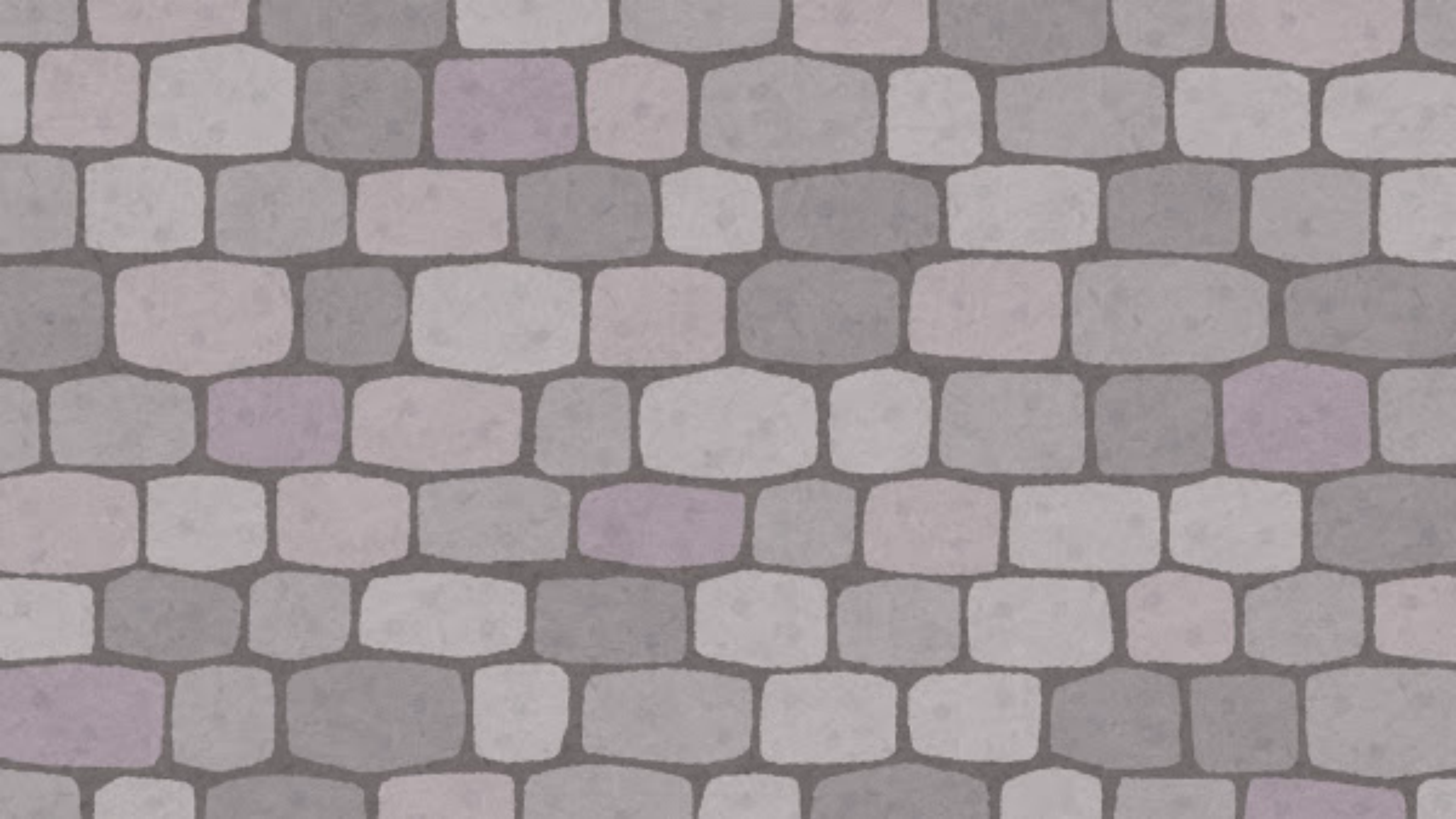
$$\arg\max_{\mathrm{W}} p(W|X) = \arg\max_{\mathrm{W}} p(X|W)p(W)$$
$$\approx \arg\max_{\mathrm{W,L}} p(X|L)p(L|W)p(W)$$

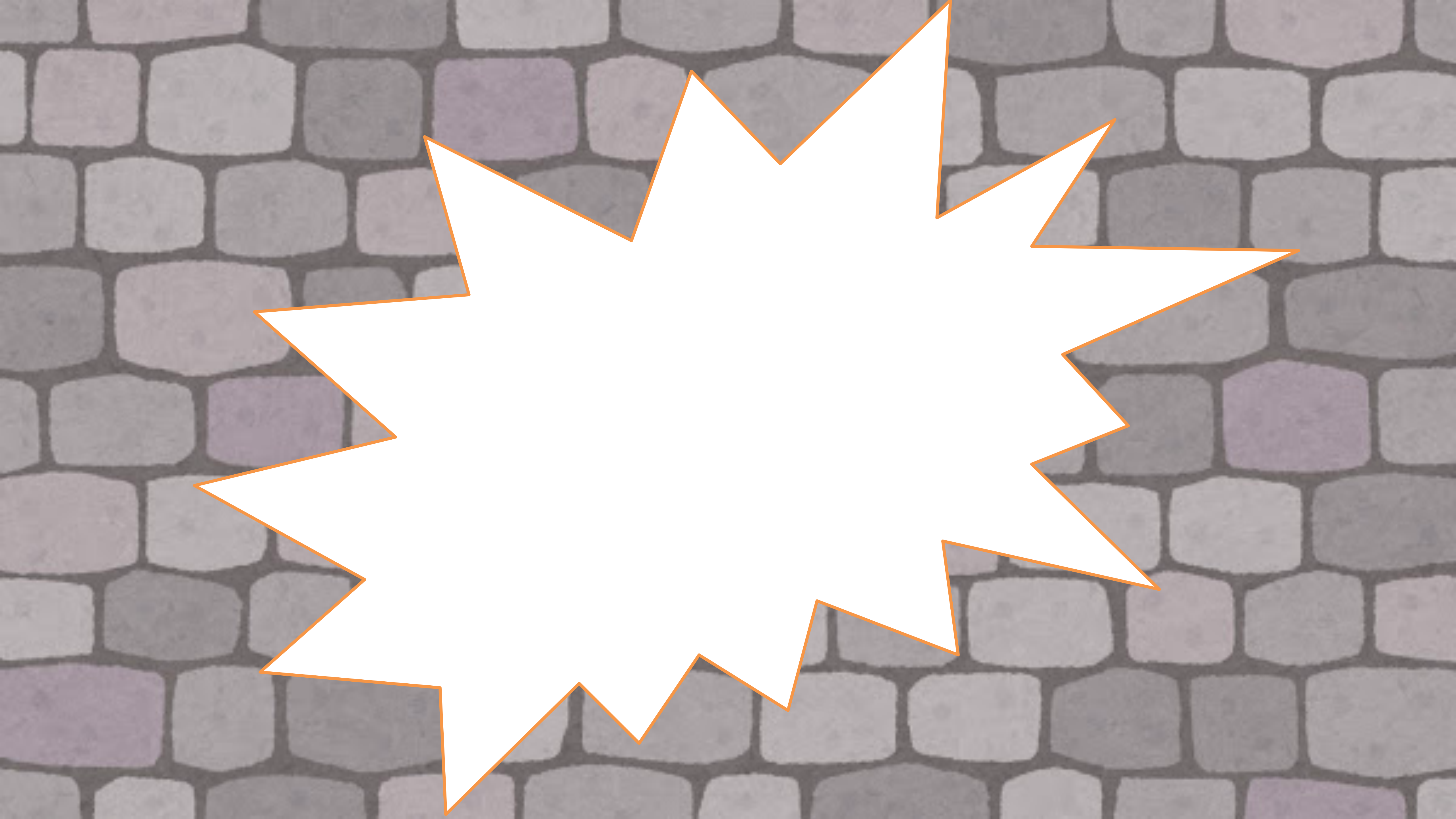
- **Speech recognition**
  - $p(X|L)$: Acoustic model
  - $p(L|W)$: Lexicon
  - $p(W)$: Language model
- Continued 40 years

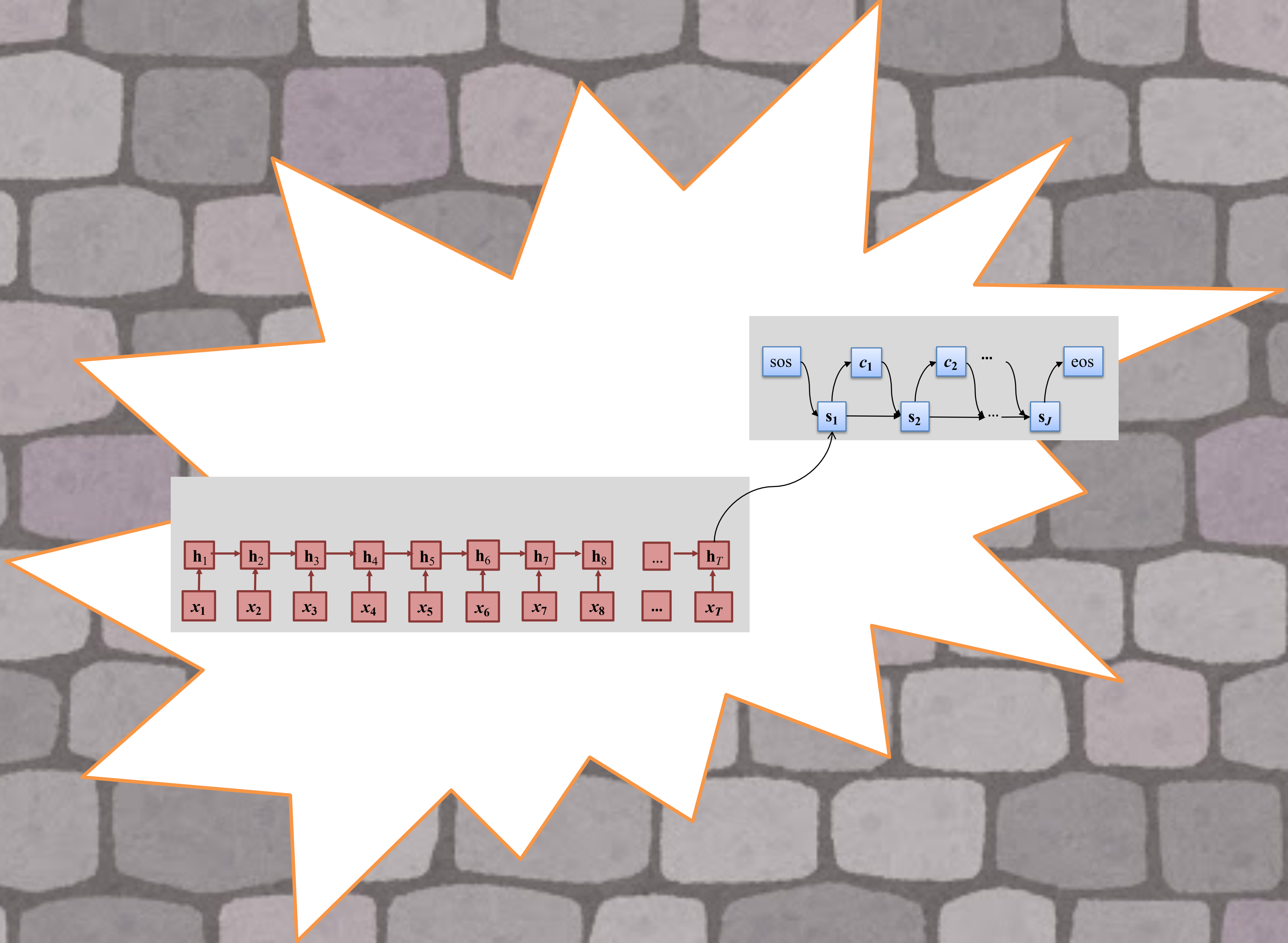
Big barrier:

noisy channel model
HMM
n-gram
etc.

However,

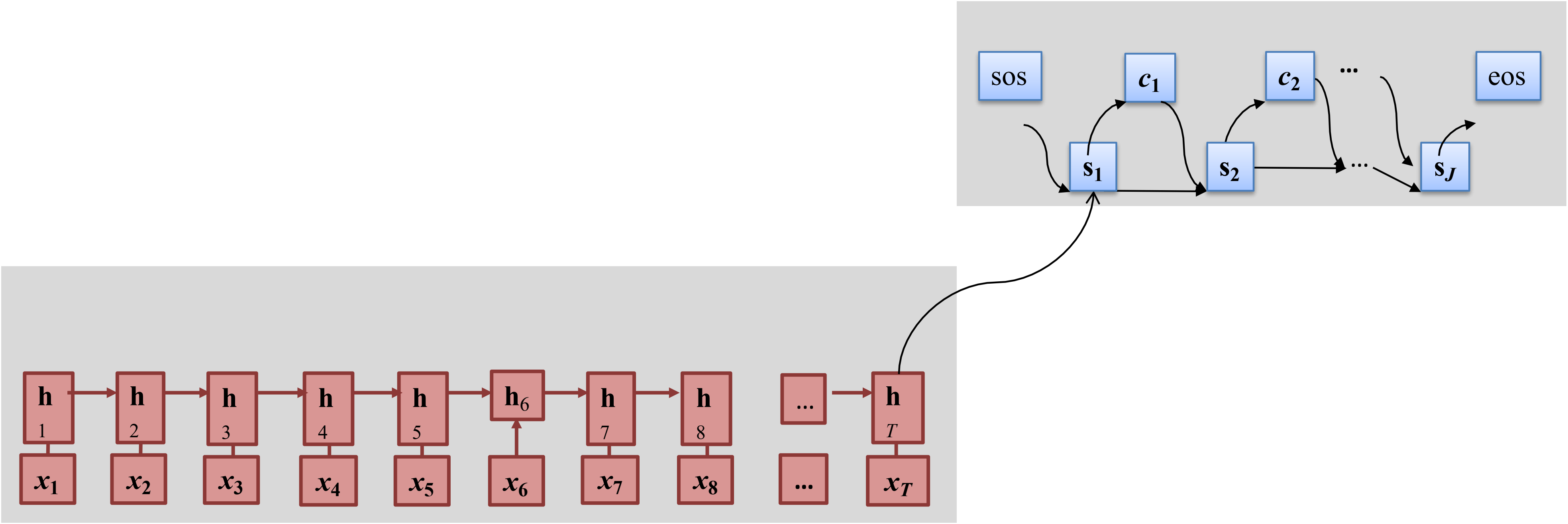






sos
c_1
c_2
...
eos
$\mathbf{s}_1$
$\mathbf{s}_2$
...
$\mathbf{s}_J$
$\mathbf{h}_1$ $\mathbf{h}_2$ $\mathbf{h}_3$ $\mathbf{h}_4$ $\mathbf{h}_5$ $\mathbf{h}_6$ $\mathbf{h}_7$ $\mathbf{h}_8$ ... $\mathbf{h}_T$
x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 ... x_T

"End-to-End" Processing
Using Sequence to Sequence



- Directly model $p(W|X)$ with a **single neural network**
  - **Integrate** acoustic $p(X|L)$, lexicon $p(L|W)$, and language $p(W)$ models
- Great success in neural machine translation

Outline

- End-to-end speech recognition
  - Hybrid CTC/attention-based end-to-end speech recognition
  - Multi-task CTC/attention learning (ICASSP’17)
  - Joint CTC/attention decoding (ACL’17)
- Examples of end-to-end **integrations**
  - Multichannel speech enhancement, dereverberation + speech recognition (ICML’17, MLSP’17, arxiv’19)
  - Multi-lingual speech recognition (ASRU’17, ICASSP’18, JSALT’18)
  - Speech separation and speech recognition (ICASSP’18, ACL’18, ICASSP’19)
  - Speech synthesis and speech recognition (SLT’18, ICASSP’19)
- Open source project
  - ESPnet: End-to-end speech processing toolkit (Interspeech’18)

Challenges

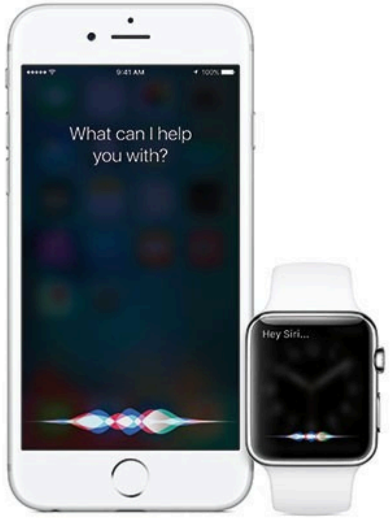
- Can you recognize the following speech?
  - Noisy speech recognition
  - Multilingual code-switching situation
  - Multispeaker situation
  - Multispeaker multilingual code-switching

Challenges

- Can you recognize the following speech?
  - Noisy speech recognition
  - Multilingual code-switching situation
  - Multispeaker situation
  - Multispeaker multilingual code-switching

I will show you how end-to-end models tackle these challenging issues

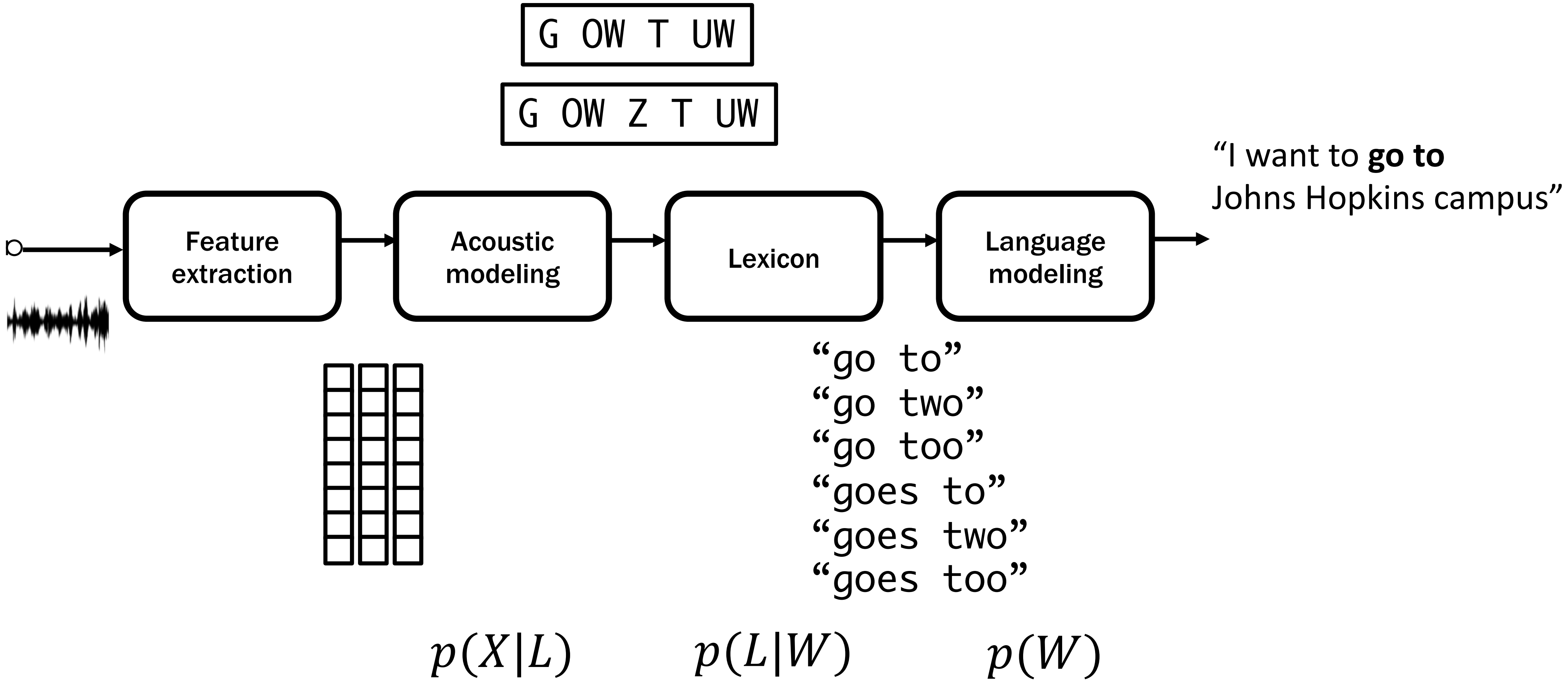
Automatic Speech Recognition (ASR)



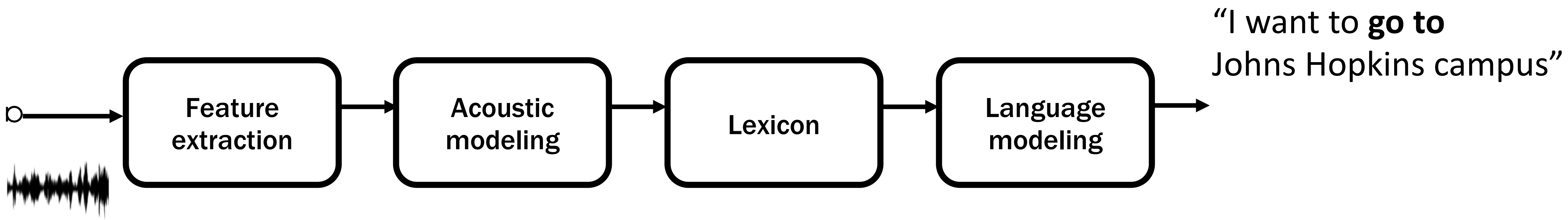
Widely used in many applications!

Great success based on the Jelinek methodology

Speech recognition pipeline

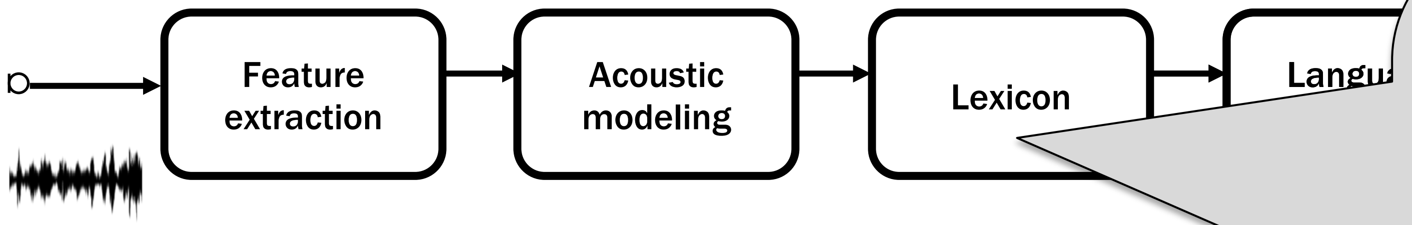


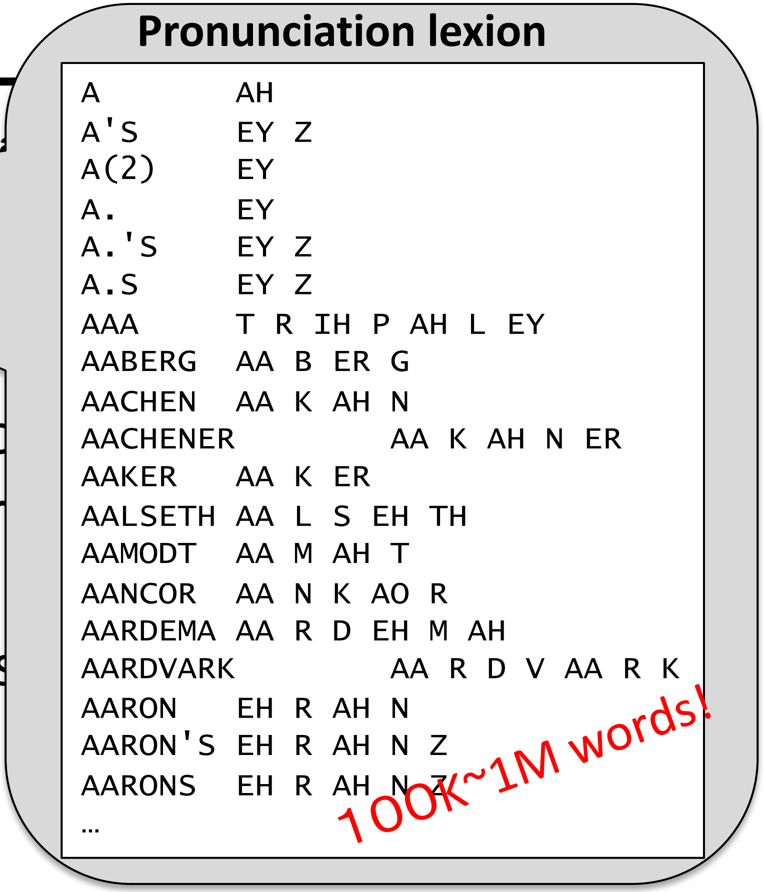
Speech recognition pipeline



- Require **a lot of development** for an acoustic model, a pronunciation lexicon, a language model, and finite-state-transducer decoding
- Require linguistic resources
- Difficult to build ASR systems for non-experts

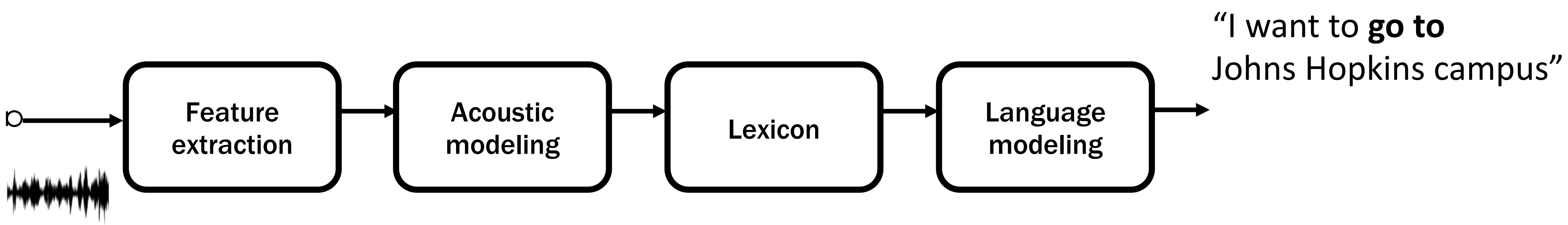
Speech recognition pipeline





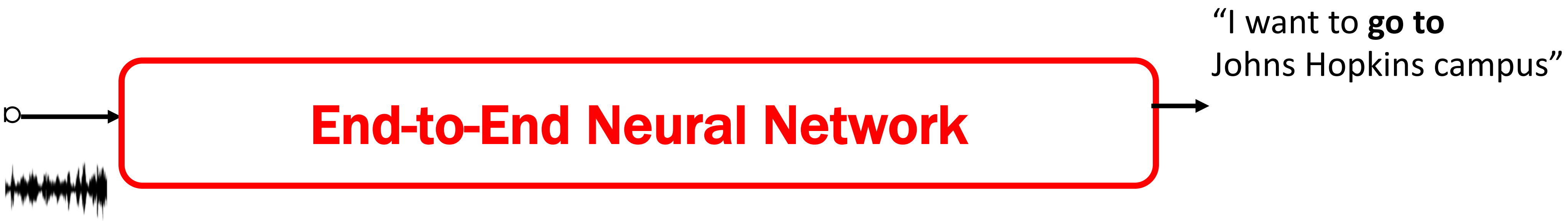
- Require **a lot of development** for an acousti
  lexicon, a language model, and finite-state-tr
- Require linguistic resources
- Difficult to build ASR systems for non-expert

Speech recognition pipeline



- Require **a lot of development** for an acoustic model, a pronunciation lexicon, a language model, and finite-state-transducer decoding
- Require linguistic resources
- Difficult to build ASR systems for **non-experts**

From pipeline to integrated architecture

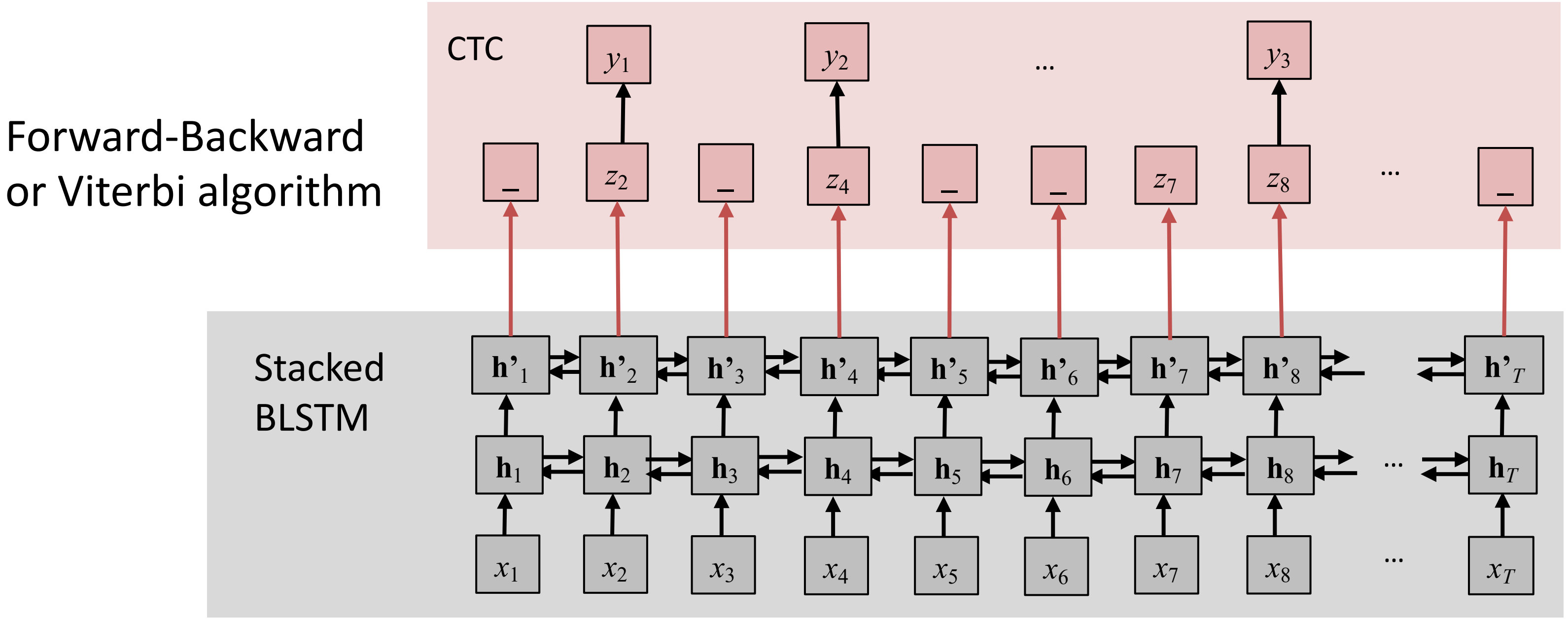


- Train a deep network that directly maps speech signal to the target letter/word sequence
- Greatly simplify the complicated model-building/decoding process
- Easy to build ASR systems for new tasks **without expert knowledge**
- Potential to outperform conventional ASR by **optimizing the entire network** with a single objective function

Connectionist temporal classification (CTC)

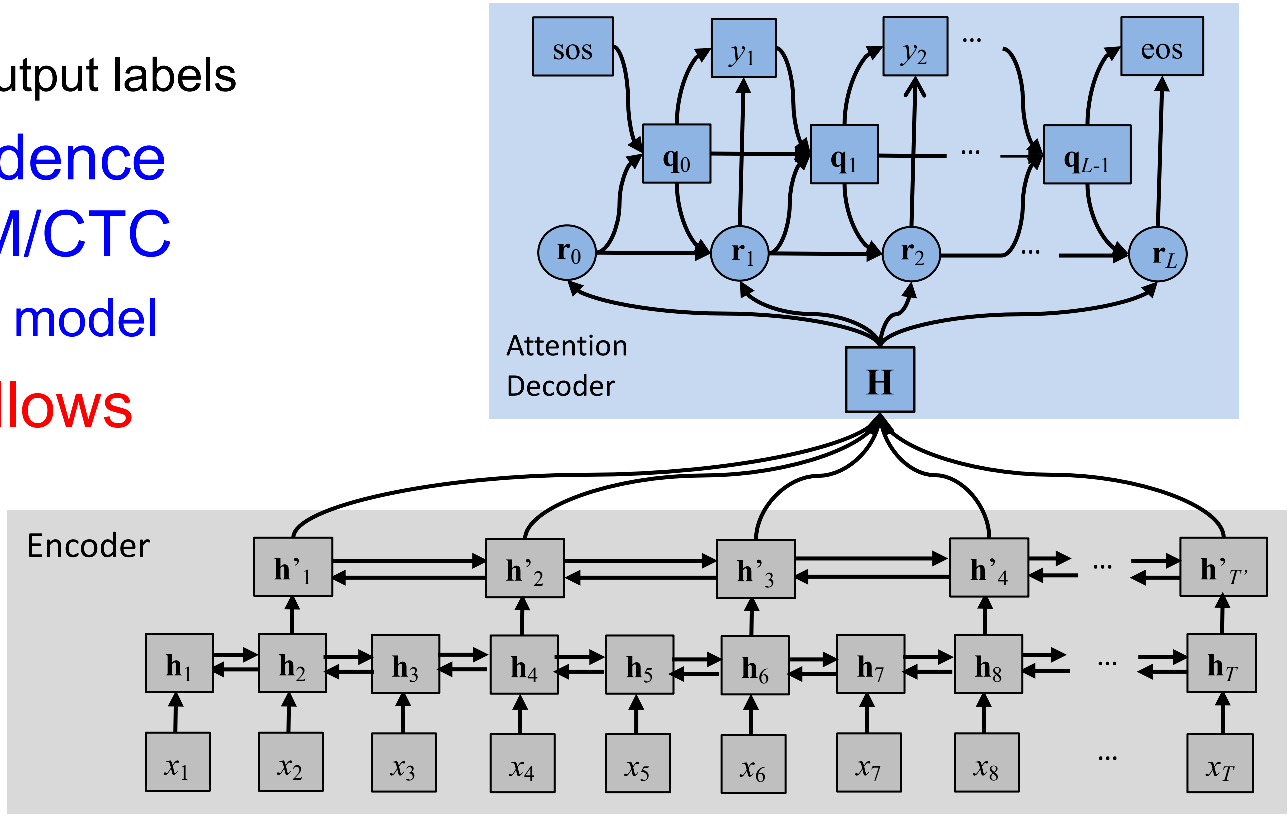
[Graves+ 2006, Graves+ 2014, Miao+ 2015]

- Use bidirectional RNNs to predict frame-based labels including blanks
- Find alignments between X and Y using dynamic programming
- Relying on conditional independence assumptions (similar to HMM)
- Output sequence is not well modeled (no language model)



Attention-based encoder decoder [Chorowski+ 2014, Chan+ 2015]

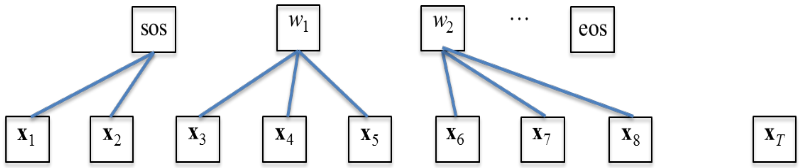
- Combine acoustic and language models in a single architecture
  - Encoder: acoustic model
  - Decoder: language model
  - Attention: align input and output labels
- No conditional independence assumption unlike HMM/CTC
  - More precise seq-to-seq model
- Attention mechanism allows too flexible alignments
  - Hard to train the model from scratch



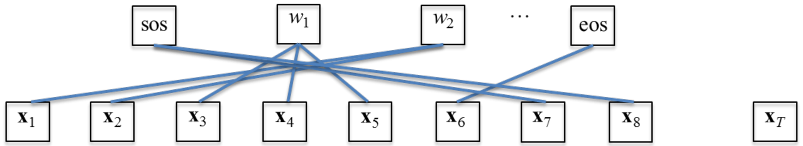
Input/output alignment by temporal attention

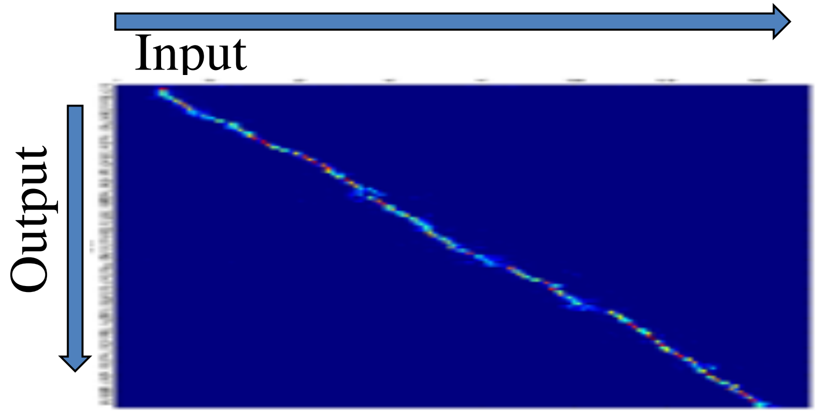
- Unlike CTC, attention model does not preserve order of inputs
- Our desired alignment in ASR task is **monotonic**
- Not regularized alignment makes the model **hard to learn** from scratch

HMM or CTC case

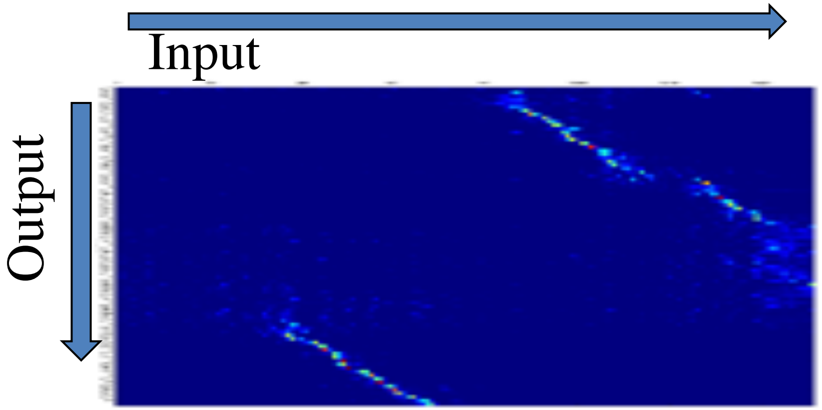


Attention model case





Example of monotonic alignment



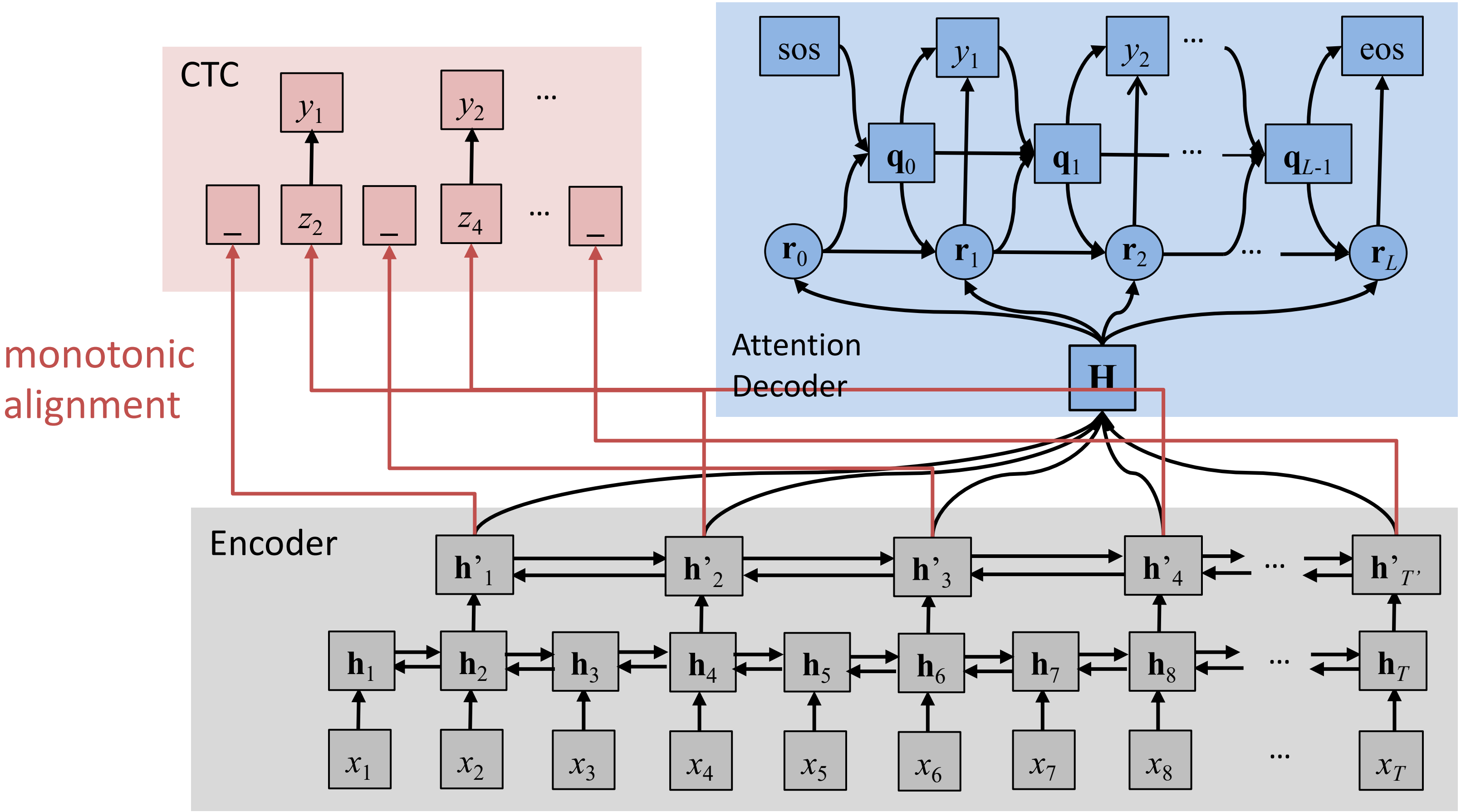
Example of distorted alignment

Outline

- End-to-end speech recognition
  - Hybrid CTC/attention-based end-to-end speech recognition
  - Multi-task CTC/attention learning (ICASSP'17)
  - Joint CTC/attention decoding (ACL'17)
- Examples of end-to-end **integrations**
  - Multichannel speech enhancement, dereverberation + speech recognition (ICML'17, MLSP'17, arxiv'19)
  - Multi-lingual speech recognition (ASRU'17, ICASSP'18, JSALT'18)
  - Speech separation and speech recognition (ICASSP'18, ACL'18, ICASSP'19)
  - Speech synthesis and speech recognition (SLT'18, ICASSP'19)
- Open source project
  - ESPnet: End-to-end speech processing toolkit (Interspeech'18)

Hybrid CTC/attention network [Kim+'17]

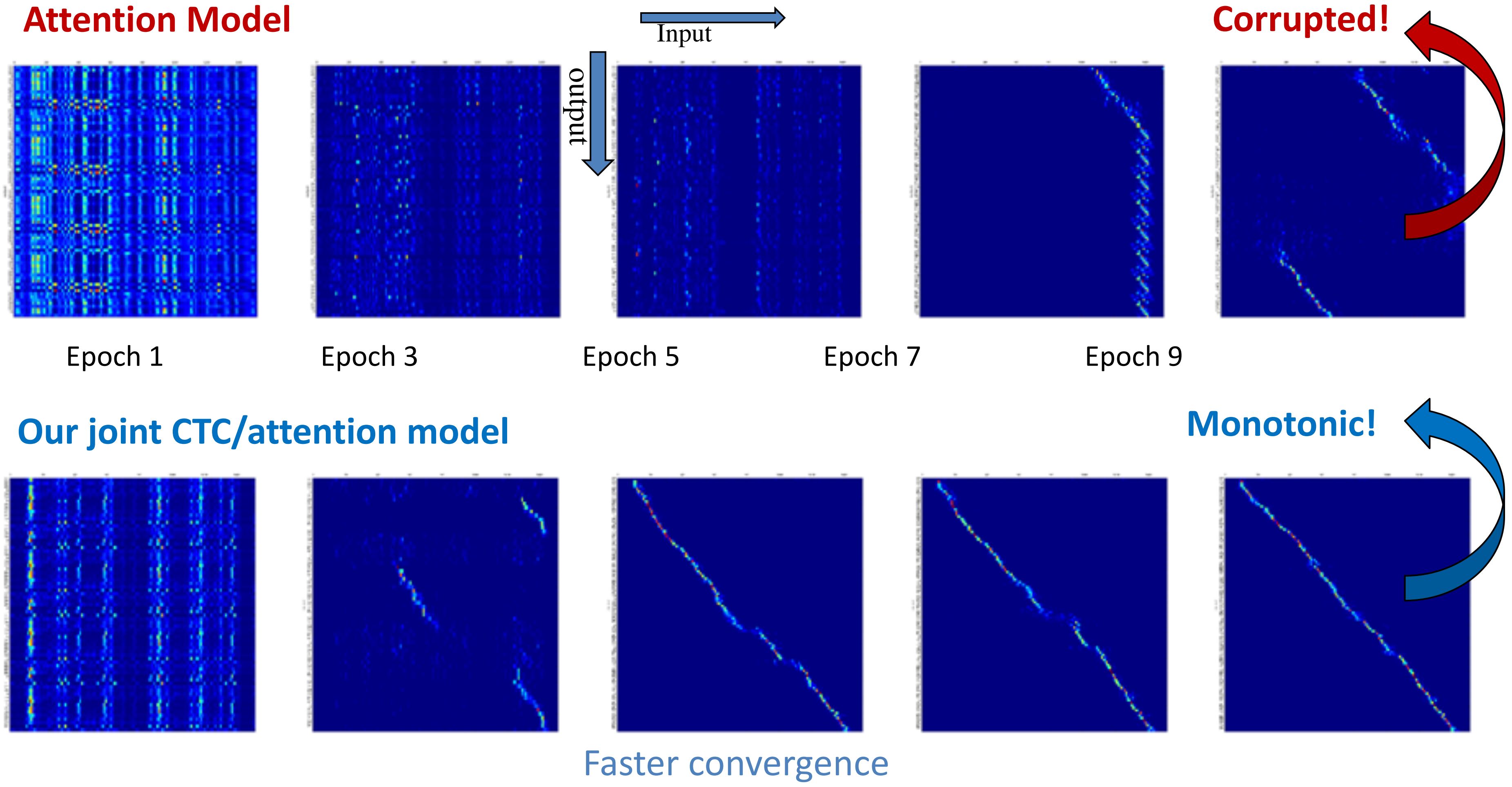
Multitask learning: $\mathcal{L}_{\text{MTL}} = \lambda \mathcal{L}_{\text{CTC}} + (1-\lambda)\mathcal{L}_{\text{Attention}}$ λ: CTC weight



CTC guides attention alignment to be monotonic

More robust input/output alignment of attention

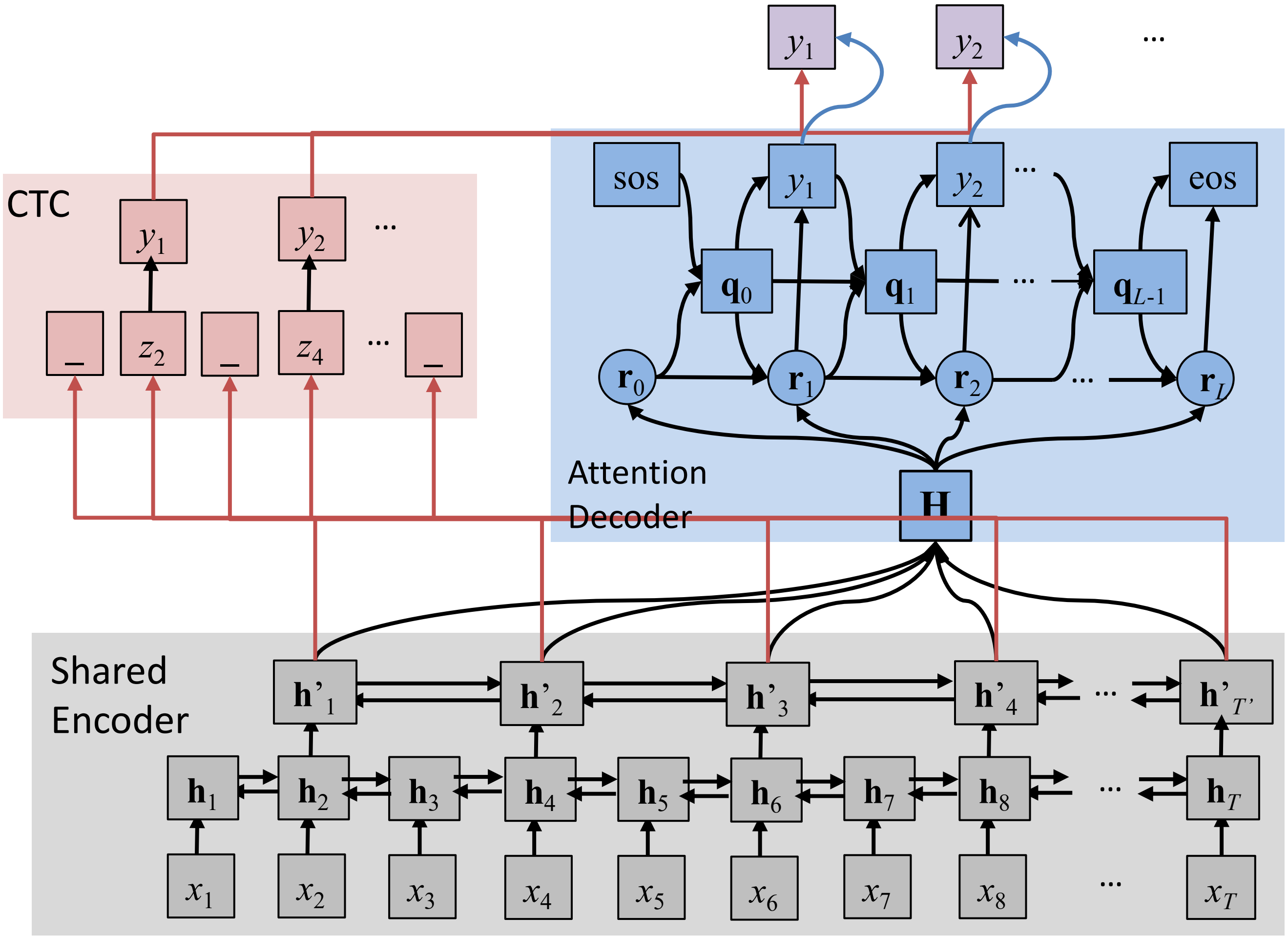
- Alignment of one selected utterance from CHiME4 task



Joint CTC/attention decoding [Hori+'17]

Use CTC for decoding together with the attention decoder

CTC explicitly eliminates non-monotonic alignment



Experimental Results

Character Error Rate (%) in **Mandarin** Chinese Telephone Conversational (HKUST, 167 hours)

Models	Dev.	Eval
Attention model (baseline)	40.3	37.8
CTC-attention learning (MTL)	38.7	36.6
+ Joint decoding	**35.5**	**33.9**

Character Error Rate (%) in Corpus of Spontaneous **Japanese** (CSJ, 581 hours)

Models	Task 1	Task 2	Task 3
Attention model (baseline)	11.4	7.9	9.0
CTC-attention learning (MTL)	10.5	7.6	8.3
+ Joint decoding	**10.0**	**7.1**	**7.6**

Example of recovering insertion errors (HKUST)

id: (20040717_152947_A010409_B010408-A-057045-057837)

Reference

但是如果你想想如果回到了过去你如果带着这个现在的记忆是不是很痛苦啊

Hybrid CTC/attention (w/o joint decoding)

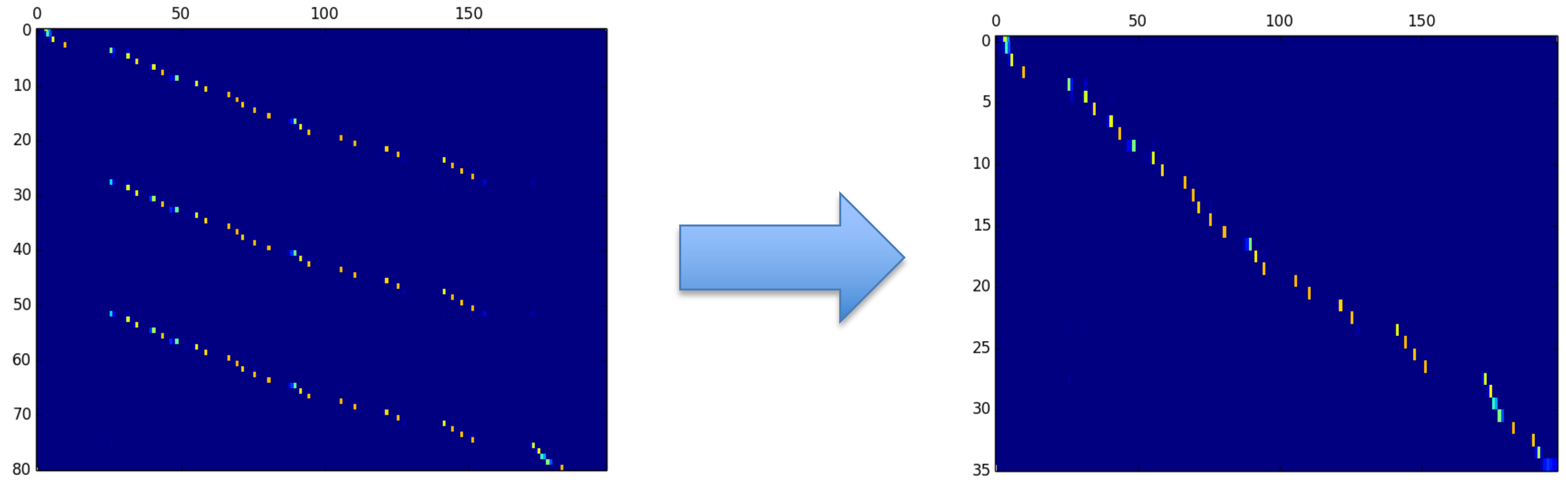
Scores: (#Correctness #Substitution #Deletion #Insertion) 28 2 3 45

但是如果你想想如果回到了过去你如果带着这个现在的节如果你想想如果回到了过去你如果带着这个现在的节如果你想想如果回到了过去你如果带着这个现在的机是不是很・・・

w/ Joint decoding

Scores: (#Correctness #Substitution #Deletion #Insertion) 31 1 1 0

HYP: 但是如果你想想如果回到了过去你如果带着这个现在的・机是不是很痛苦啊



Example of recovering deletion errors (CSJ)

id: (A01F0001_0844951_0854386)

Reference

ま た え 飛 行 時 の エ コ ー ロ ケ ー シ ョ ン 機 能 を よ り 詳 細 に 解 明 す る 為 に 超 小 型 マ イ
ク ロ ホ ン お よ び 生 体 ア ン プ を コ ウ モ リ に 搭 載 す る こ と を 考 え て お り ま す そ う す
る こ と に よ っ て

Hybrid CTC/attention (w/o joint decoding)

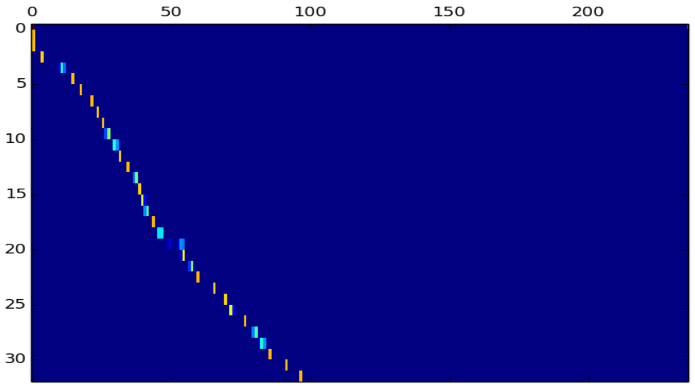
Scores: (#Correctness #Substitution #Deletion #Insertion) 30 0 47 0

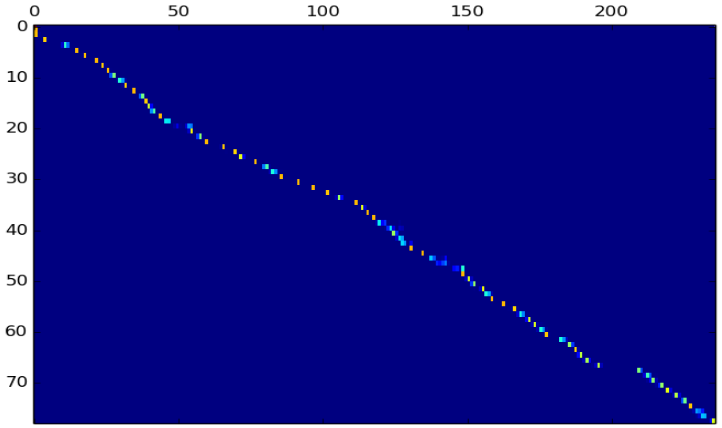
ま た え 飛 行 時 の エ コ ー ロ ケ ー シ ョ ン 機 能 を よ り 詳 細 に 解 明 す る
為 ・
・ ・ ・ ・ ・ ・ ・ ・ ・ ・ に ・ ・ ・

w/ Joint decoding

Scores: (#Correctness #Substitution #Deletion #Insertion) 67 9 1 0

ま た え 飛 行 時 の エ コ ー ロ ケ ー シ ョ ン 機 能 を よ り 詳 細 に 解 明 す る 為 に 長 国 型 マ イ
ク ロ ホ ン お ・ い く 声 単 位 方 を コ ウ モ リ に 登 載 す る こ と を 考 え て お り ま す そ う す
る こ と に よ っ て





Discussions

- Hybrid CTC/attention-based end-to-end speech recognition
  - Multi-task learning during training
  - Joint decoding during recognition

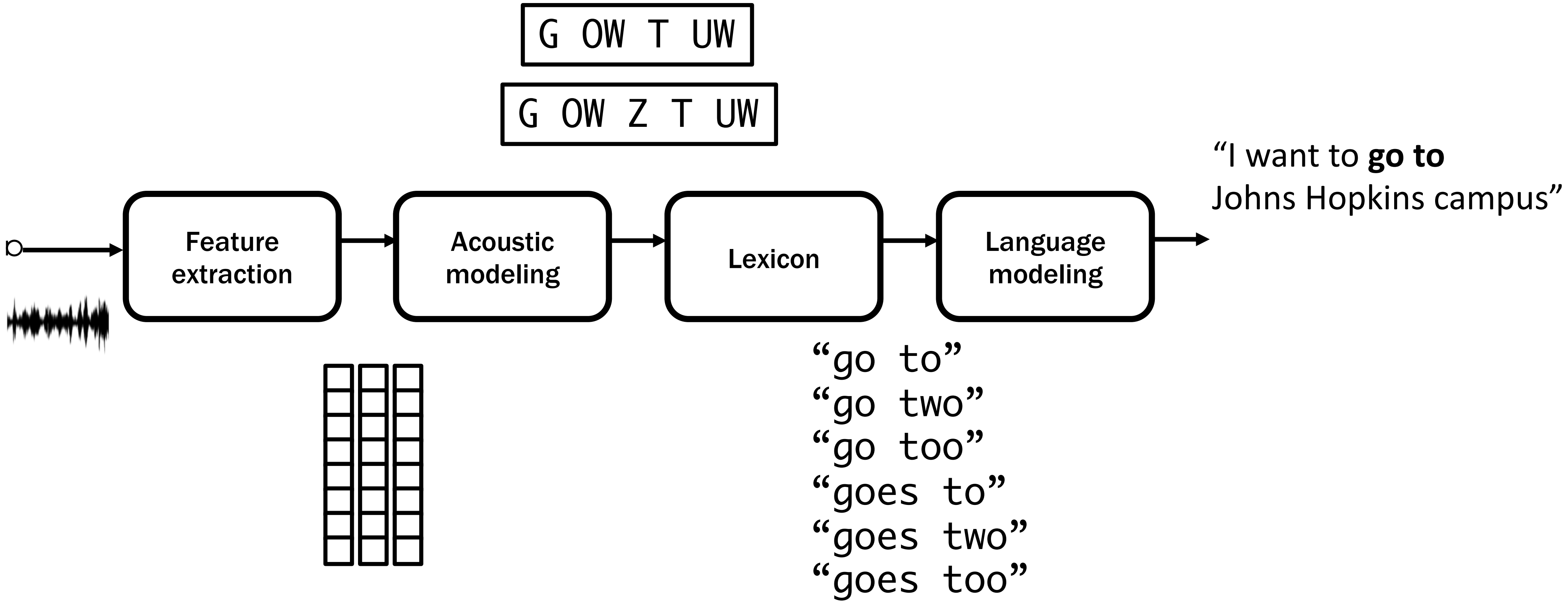
  ➡ **Make use of both benefits, completely solve alignment issues**
- Now we have a good end-to-end ASR tool

  ➡ **Apply several challenging ASR issues**

Outline

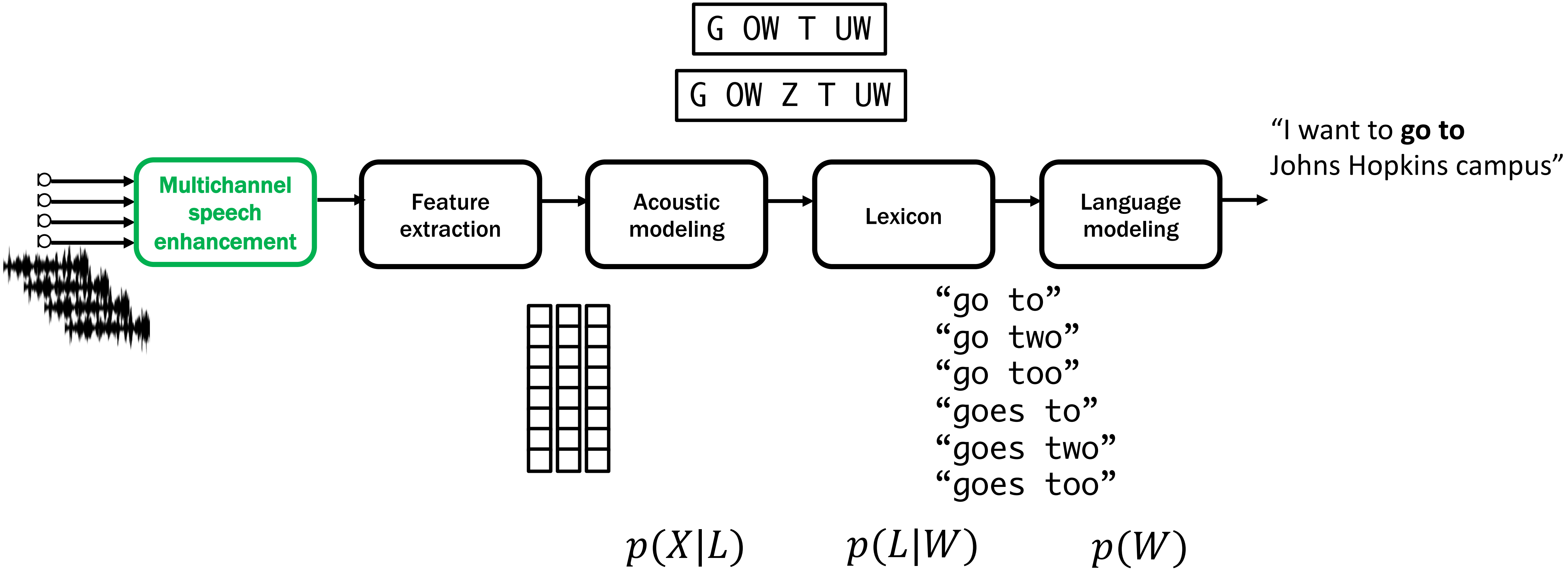
- End-to-end speech recognition
  - Hybrid CTC/attention-based end-to-end speech recognition
  - Multi-task CTC/attention learning (ICASSP'17)
  - Joint CTC/attention decoding (ACL'17)
- Examples of end-to-end **integrations**
  - Multichannel speech enhancement, dereverberation + speech recognition (ICML'17, MLSP'17, arxiv'19)
  - Multi-lingual speech recognition (ASRU'17, ICASSP'18, JSALT'18)
  - Speech separation and speech recognition (ICASSP'18, ACL'18, ICASSP'19)
  - Speech synthesis and speech recognition (SLT'18, ICASSP'19)
- Open source project
  - ESPnet: End-to-end speech processing toolkit (Interspeech'18)

Speech recognition pipeline

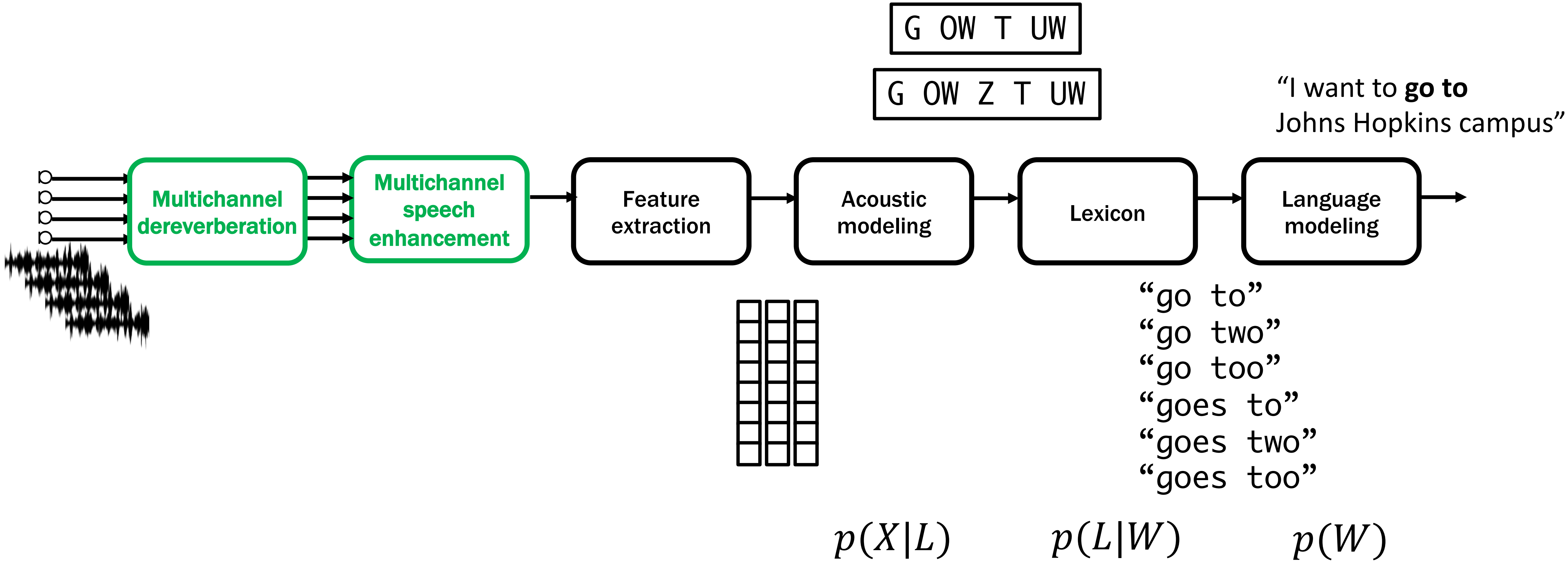


$p(X|L)$ $p(L|W)$ $p(W)$

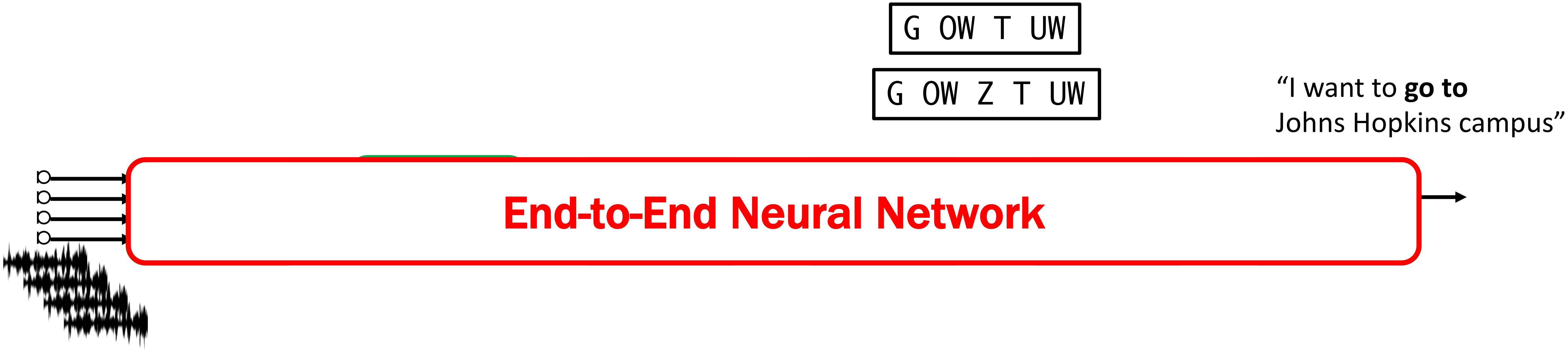
Multichannel speech recognition pipeline



Multichannel speech recognition pipeline

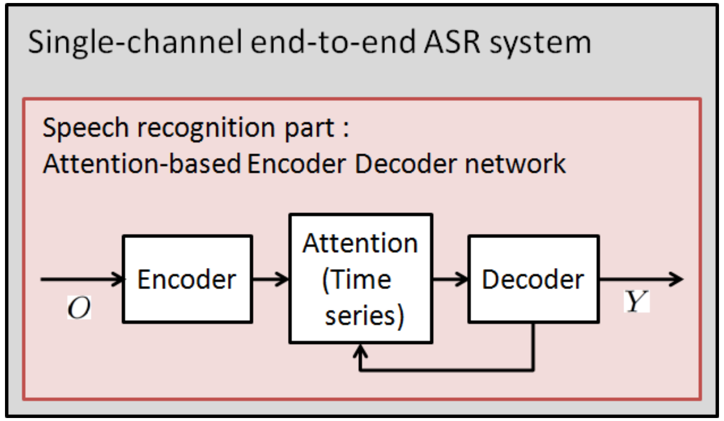


Multichannel speech recognition pipeline

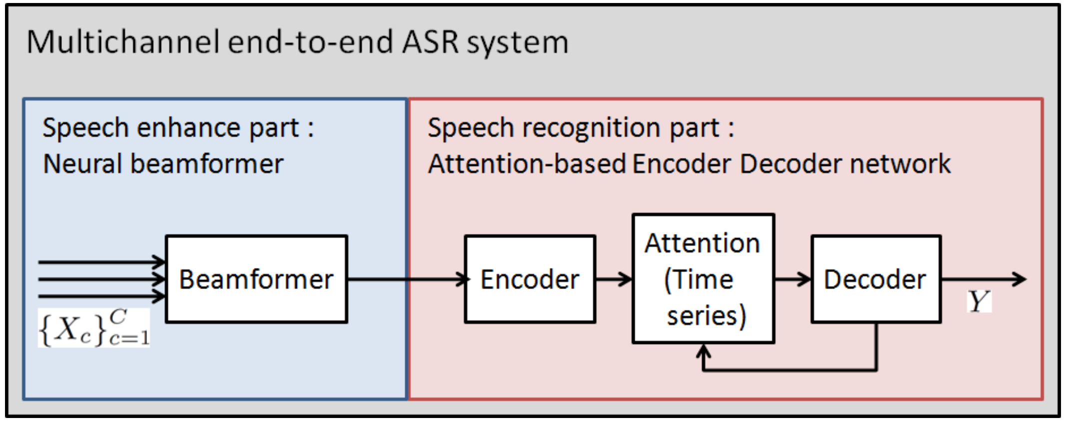


Multichannel end-to-end ASR architecture
[Ochiai et al., 2017, ICML]

Single-channel (Conventional)



Multichannel (Proposed)



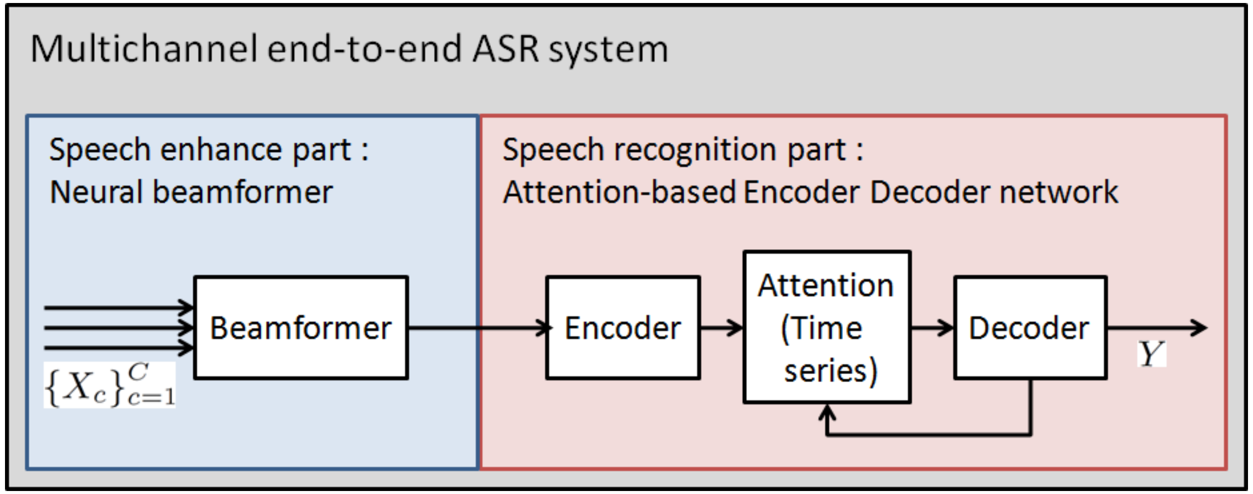
Overview of entire architecture

- 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]



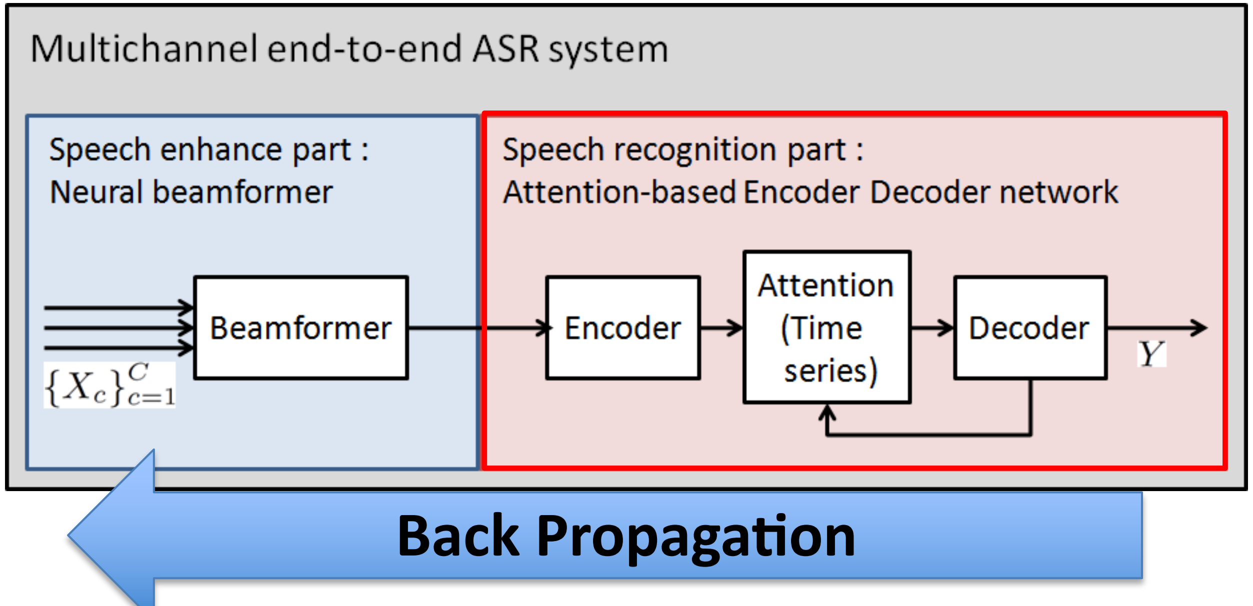
Overview of entire architecture

- 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]



- **No pre-training**
- **No signal-level supervision**

(only require trans. + noisy speech)

Experimental Results
[espnet #596]

- Noisy speech recognition task (CHiME-4)
  - Sigle-channel E2E + beamforming (pipeline)
  - Multichannel E2E (integration of speech enhancement and recognition)

model	Word error rate (dev real)	Word error rate (test real)
Single-channel E2E +Beamforming (**pipeline**)	10.1	19.8
Multichannel E2E (**integration**)	**8.5**	**16.4**

Obtained noise robustness through end-to-end training

Further extension
Dereverberation + beamforming + ASR [espnet #596]

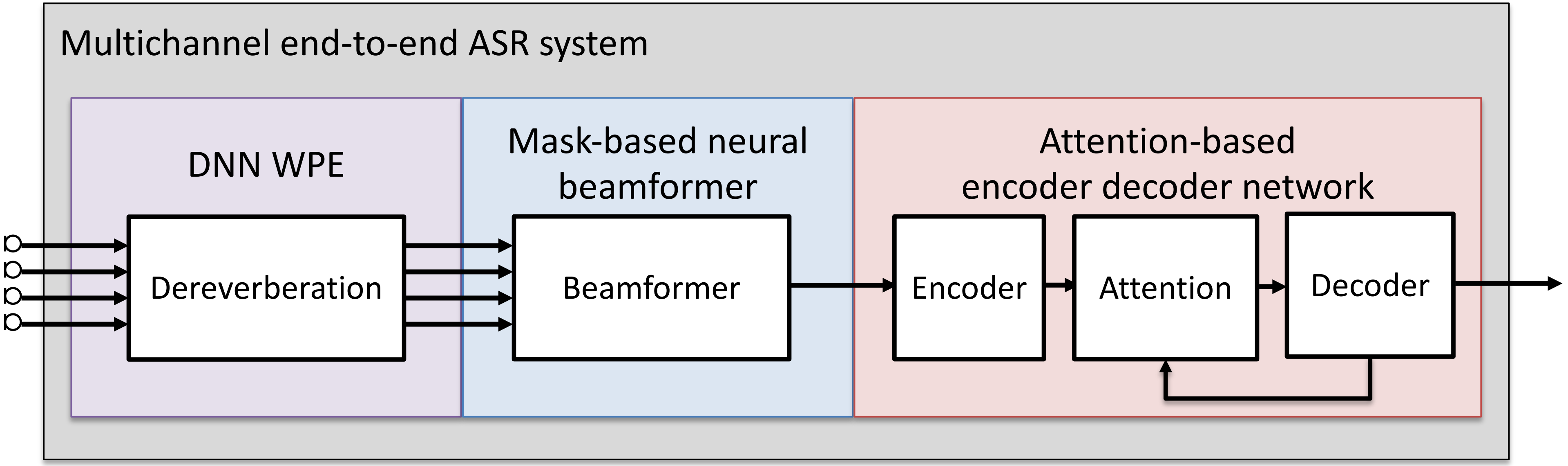
- Multichannel end-to-end ASR framework
  - integrates entire process of **speech dereverberation (SD)**, **beamforming (SB)**and **speech recognition (SR)**, by single neural-network-based architecture

↓

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 [espnet #596]

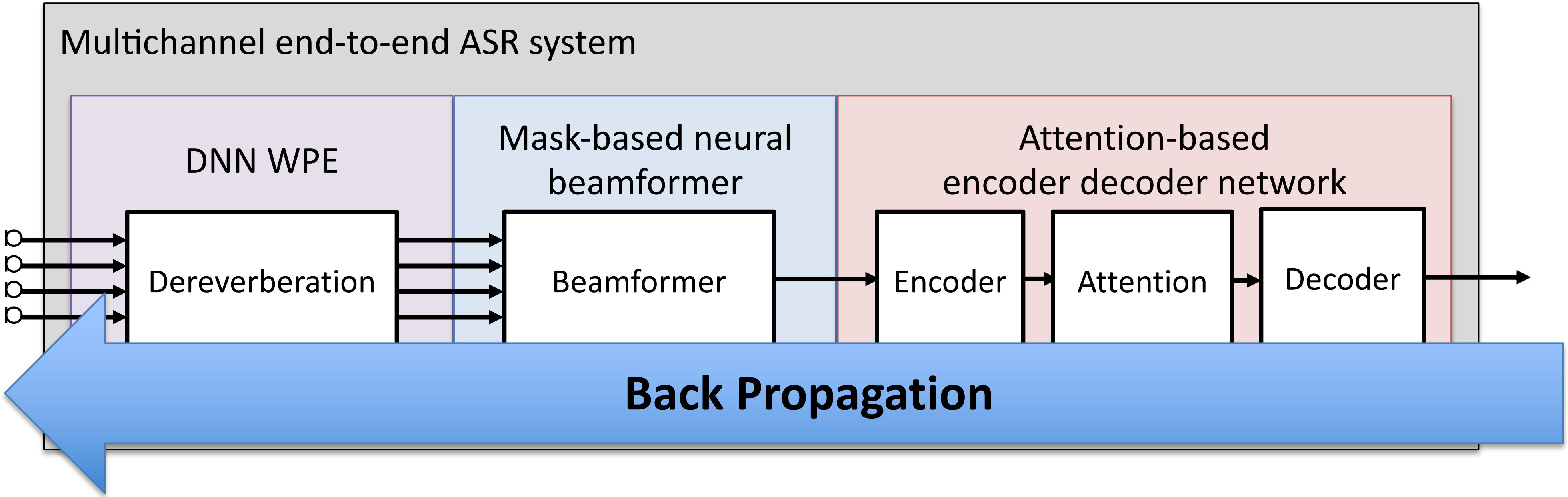
- Multichannel end-to-end ASR framework
  - integrates entire process of **speech dereverberation (SD)**, **beamforming (SB)** and **speech recognition (SR)**, by single neural-network-based architecture

↓

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]



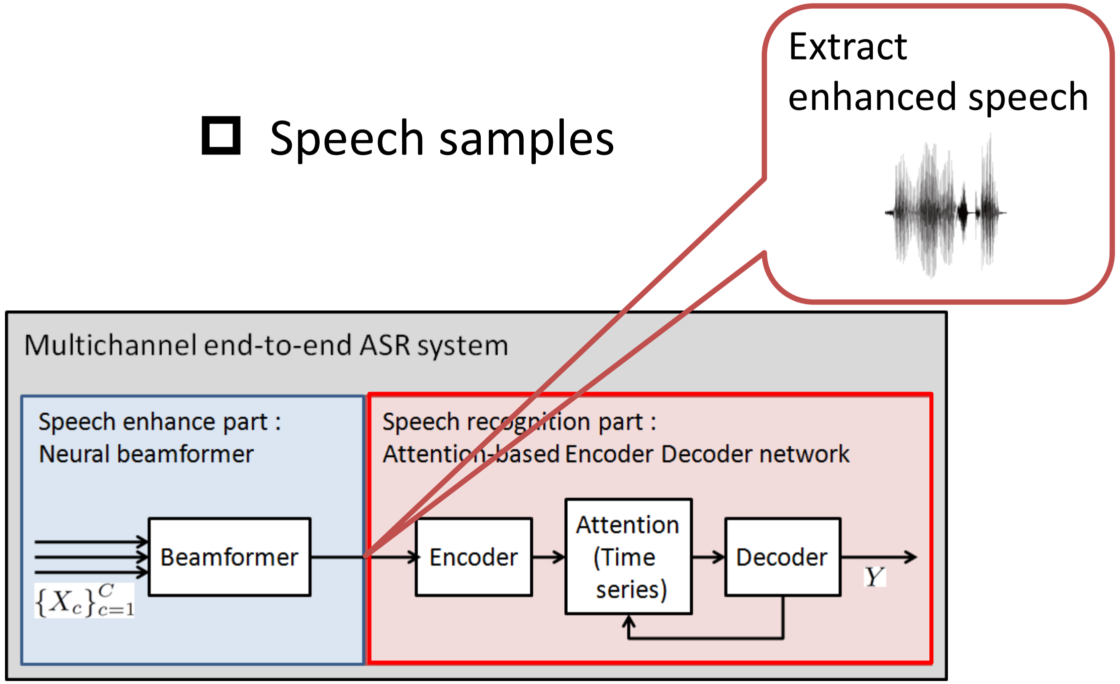
Experimental Results
[Subramanian et al (2019)]

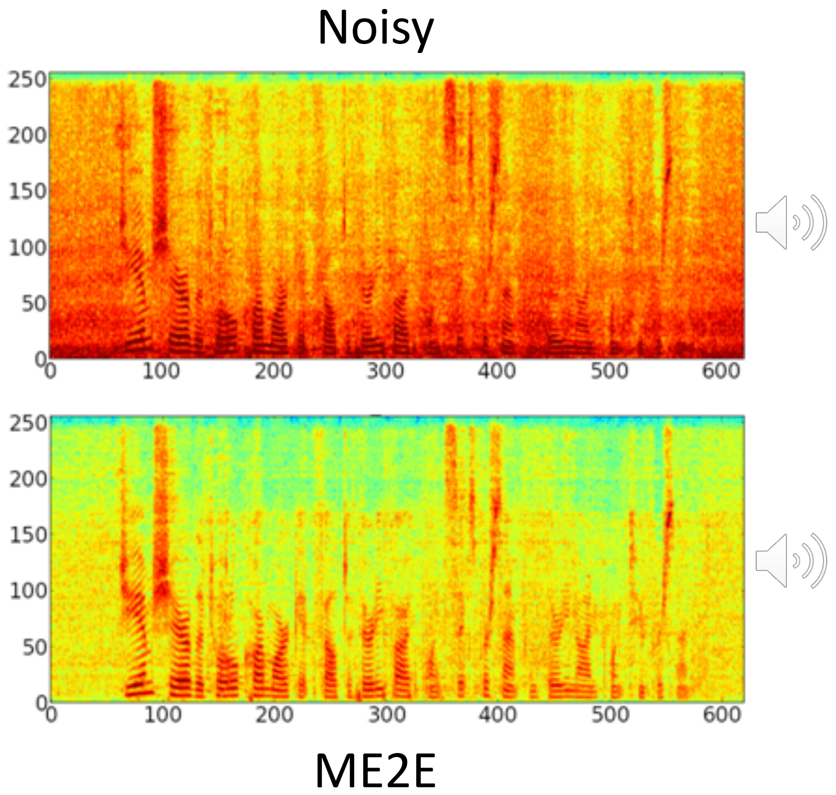
- Noisy reverberant speech recognition task (REVERB and DIRHA-WSJ)
  - Sigle-channel E2E + dereverberation + beamforming (pipeline)
  - Multichannel E2E (integration of speech enhancement and recognition)

model	REVERB Room1 Near	REVERB Room1 Far	DIRHA WSJ Real
Single-channel E2E + Dereverberation + Beamforming (**pipeline**)	11.0	**10.8**	31.3
Multichannel E2E (**integration**)	**8.7**	12.4	**29.1**

It works as speech enhancement!

- Speech samples



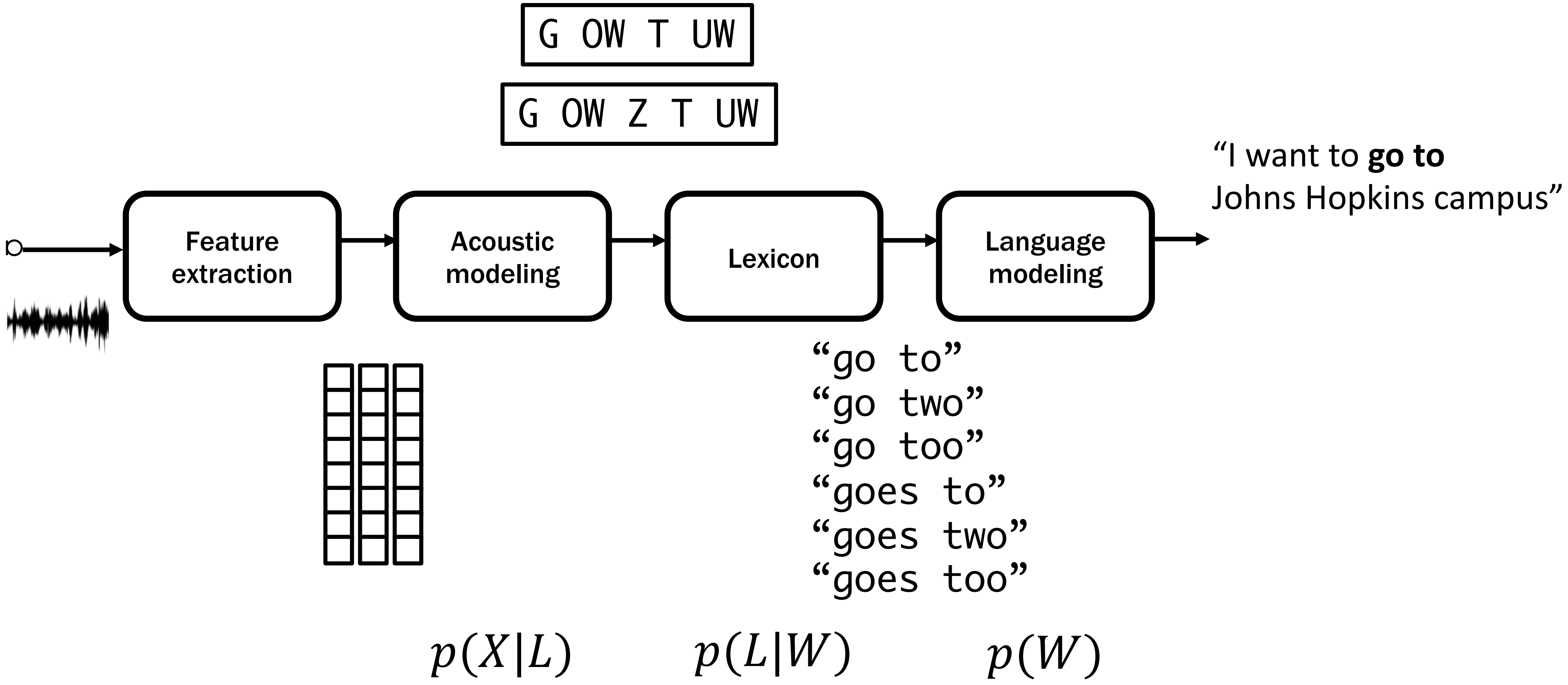


- **Entire network** are **consistently optimized with ASR-level objective including speech enhancement part**
- **Pairs of parallel clean and noisy data are not required for training → SE can be optimized only with noisy signals and their transcripts**

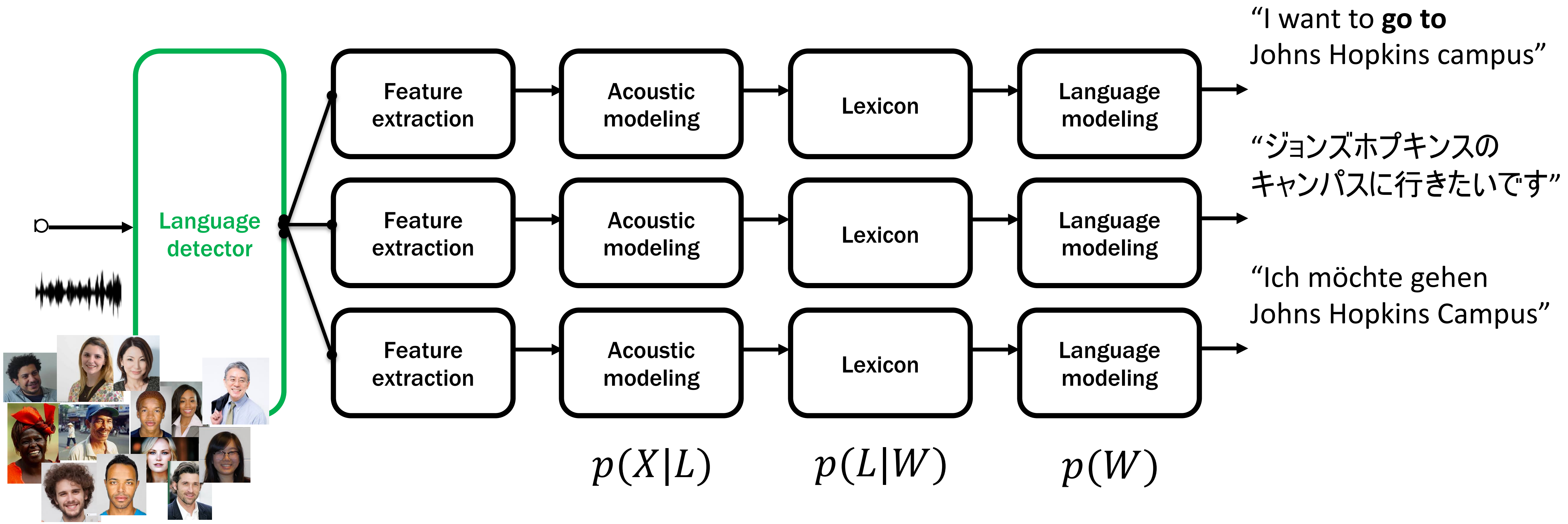
Outline

- End-to-end speech recognition
  - Hybrid CTC/attention-based end-to-end speech recognition
  - Multi-task CTC/attention learning (ICASSP'17)
  - Joint CTC/attention decoding (ACL'17)
- Examples of end-to-end **integrations**
  - Multichannel speech enhancement, dereverberation + speech recognition (ICML'17, MLSP'17, arxiv'19)
  - Multi-lingual speech recognition (ASRU'17, ICASSP'18, JSALT'18)
  - Speech separation and speech recognition (ICASSP'18, ACL'18, ICASSP'19)
  - Speech synthesis and speech recognition (SLT'18, ICASSP'19)
- Open source project
  - ESPnet: End-to-end speech processing toolkit (Interspeech'18)

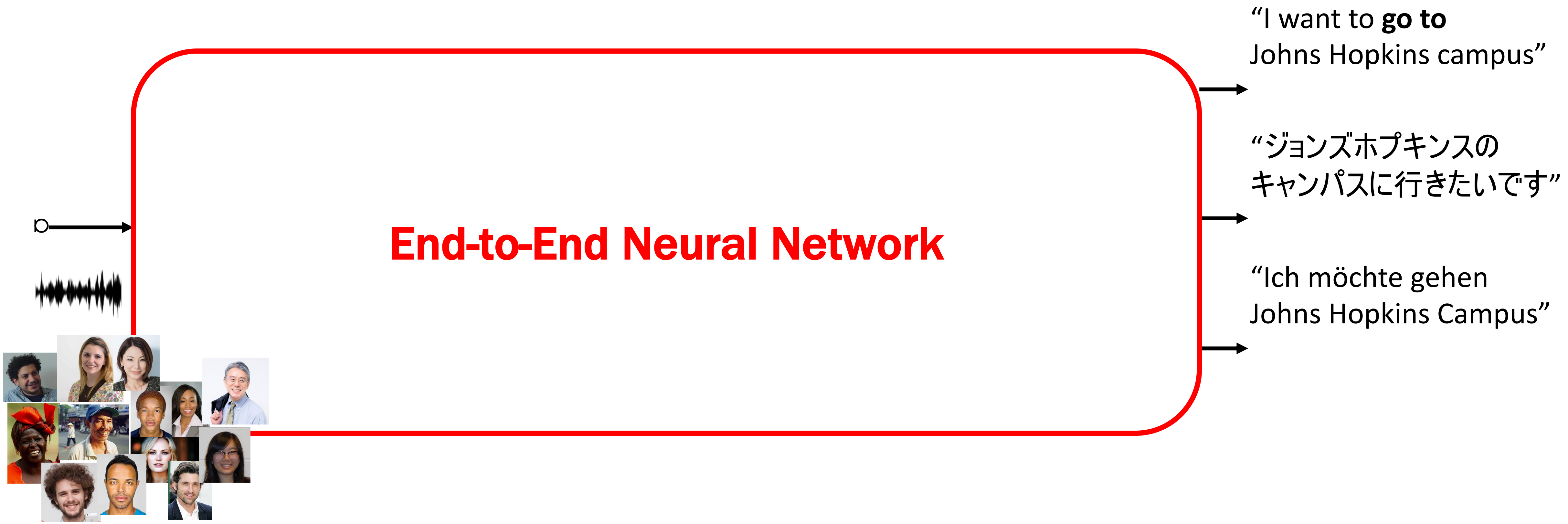
Speech recognition pipeline



Multilingual speech recognition pipeline



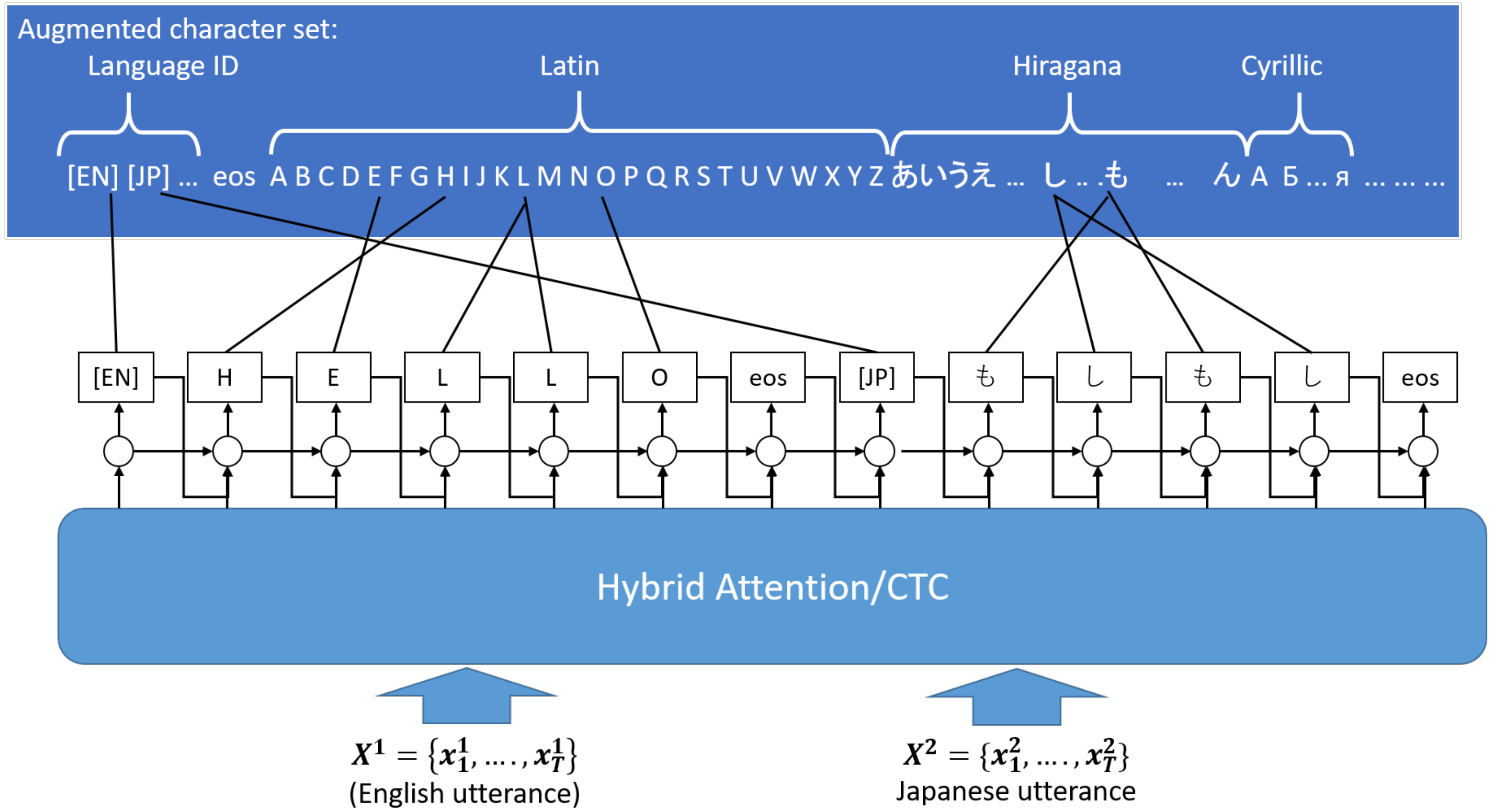
Multilingual speech recognition pipeline



Multi-lingual end-to-end speech recognition

[Watanabe+'17, Seki+'18]

- Learn a single model with multi-language data (10 languages)
- **Integrates** language identification and 10-language speech recognition systems
- **No pronunciation lexicons**



Include all language characters and language ID for final softmax to accept all target languages


DICTIONARY
漢字


DICTIO
NARY

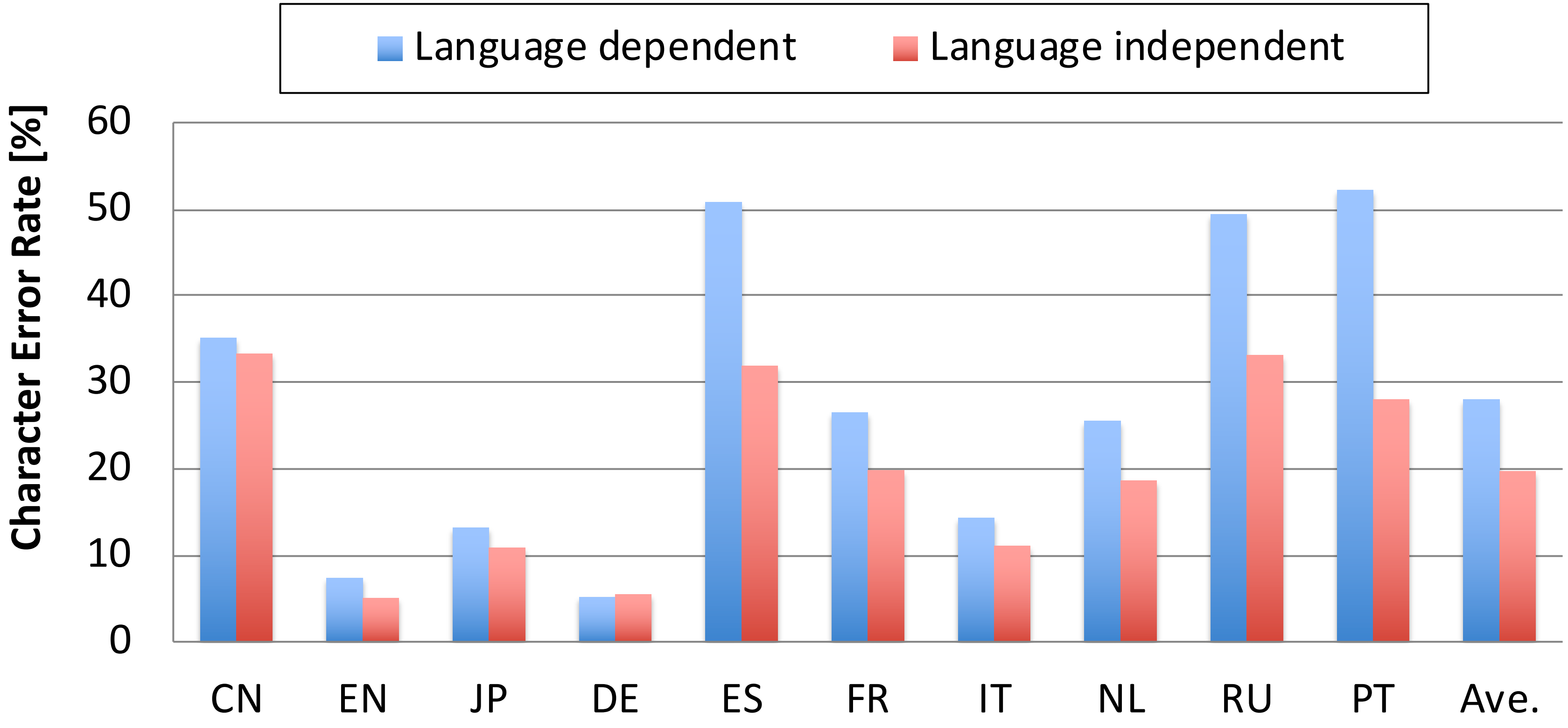
ASR performance for 10 languages

- Comparison with language dependent systems
- One language per utterance (w/o code switching)

	# systems	7 languages CER(%)	10 languages CER(%)
Language-***dependent*** E2E	7 or 10 Given Language ID	22.7	27.4
Language-***independent*** E2E (small model)	1	20.3	--
Language-***independent*** E2E (large model)	1	**16.6**	**21.4**

ASR performance for 10 languages

- Comparison with language dependent systems
- Language-independent single end-to-end ASR works well!



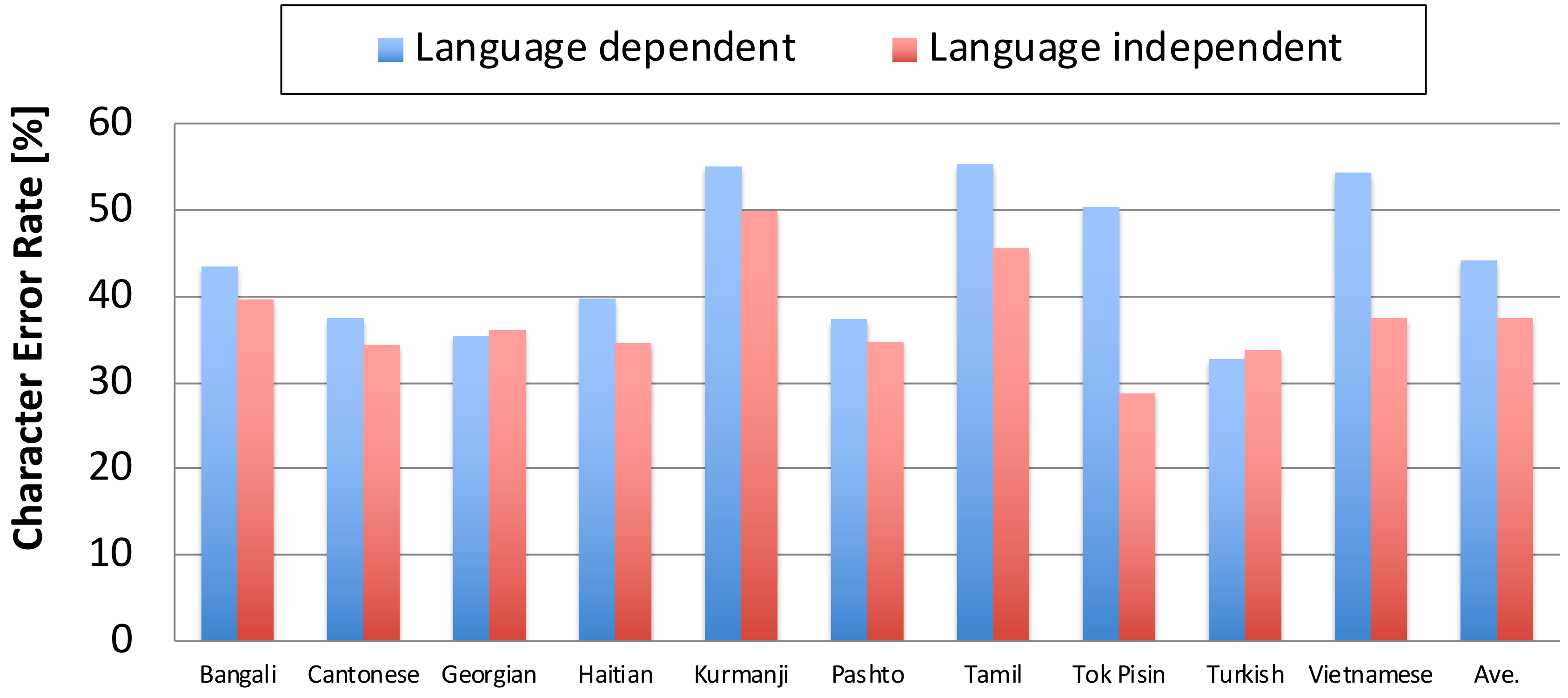
你好
Hello
こんにちは
Hallo
Hola
Bonjour
Ciao
Hallo
Привет
Olá

Language recognition performance

		CH	EN	JP	DE	ES	FR	IT	NL	RU	PT
CH	train_dev	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	dev	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
EN	test_eval92	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	test_dev93	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
JP	eval1_jpn	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	eval2_jpn	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	eval3_jpn	0.0	0.0	99.9	0.0	0.0	0.0	0.1	0.0	0.0	0.0
DE	et_de	0.0	0.0	0.0	99.7	0.0	0.0	0.0	0.3	0.0	0.0
	dt_de	0.0	0.0	0.0	99.7	0.0	0.0	0.0	0.3	0.0	0.0
ES	dt_es	0.0	0.0	0.0	0.0	67.9	0.0	31.9	0.0	0.0	0.2
	et_es	0.0	0.0	0.0	0.1	91.1	0.0	8.4	0.1	0.0	0.2
FR	dt_fr	0.0	0.0	0.0	0.1	0.0	99.4	0.0	0.2	0.0	0.3
	et_fr	0.0	0.0	0.0	0.1	0.0	99.5	0.0	0.1	0.0	0.3
IT	dt_it	0.0	0.0	0.0	0.0	0.3	0.4	99.1	0.0	0.0	0.3
	et_it	0.0	0.0	0.0	0.0	0.4	0.4	98.3	0.2	0.1	0.7
NL	dt_nl	0.0	0.0	0.0	1.3	0.0	0.1	0.1	97.2	0.0	1.3
	et_nl	0.0	0.0	0.0	1.0	0.0	0.2	0.2	97.6	0.0	0.9
RU	dt_ru	0.2	0.0	0.0	0.0	0.2	0.6	0.5	0.0	97.9	0.8
	et_ru	0.0	0.0	0.0	0.2	0.2	0.3	4.3	0.0	94.7	0.3
PT	dt_pt	0.0	0.0	0.0	0.3	0.3	2.6	1.7	3.4	0.6	91.2
	et_pt	0.0	0.3	0.0	0.3	0.0	0.0	3.9	3.6	0.3	91.5

ASR performance for **law-resource** 10 languages

- Comparison with language dependent systems



হ্যালো
你好
გამარჯობა
hello
???
سلام
வணக்கம்
???
Merhaba
xin chào

Actually it was one of the easiest studies in my work

Q. How many people were involved in the development?

A. 1 person

Q. How long did it take to build a system?

A. Totally ~1 or 2 day efforts with bash and python scripting (no change of main e2e ASR source code), **then I waited 10 days to finish training**

Q. What kind of linguistic knowledge did you require?

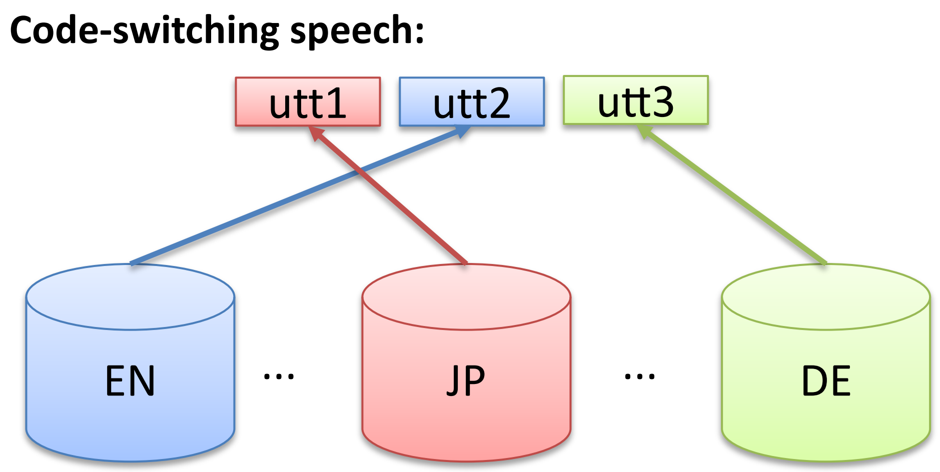
A. Unicode (because python2 Unicode treatment is tricky. If I used python3, I would not even have to consider it)

ASRU'17 best paper **candidate**

Data generation for multi-lingual **code-switching** speech
[Seki+ (2018)]

- **Don't change any architecture**, but change the training data preparation
- Concatenation of utterances from 10 language corpora

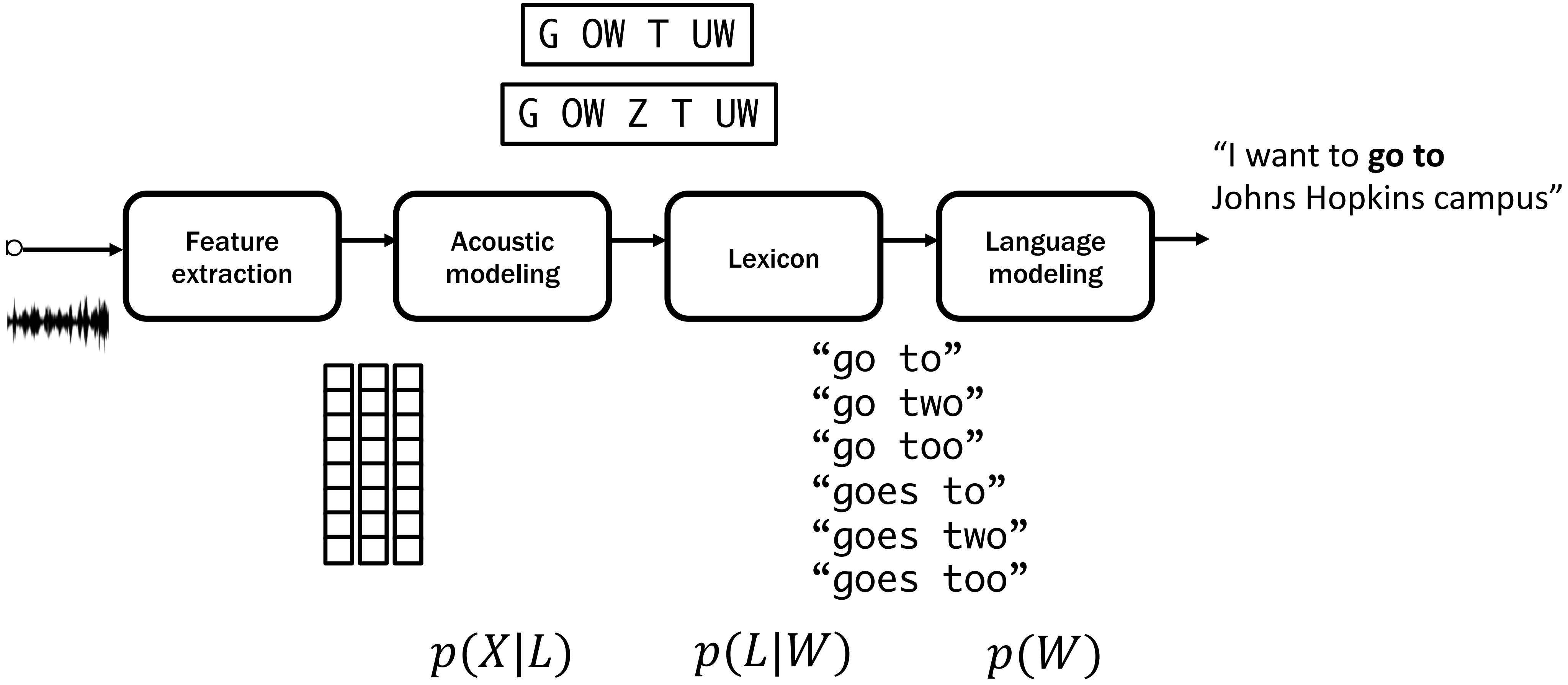
1) Select number to concat (1, 2, or 3)
2) Sample language and utterance:
3) Repeat generation to reach the duration of the original corpora



Outline

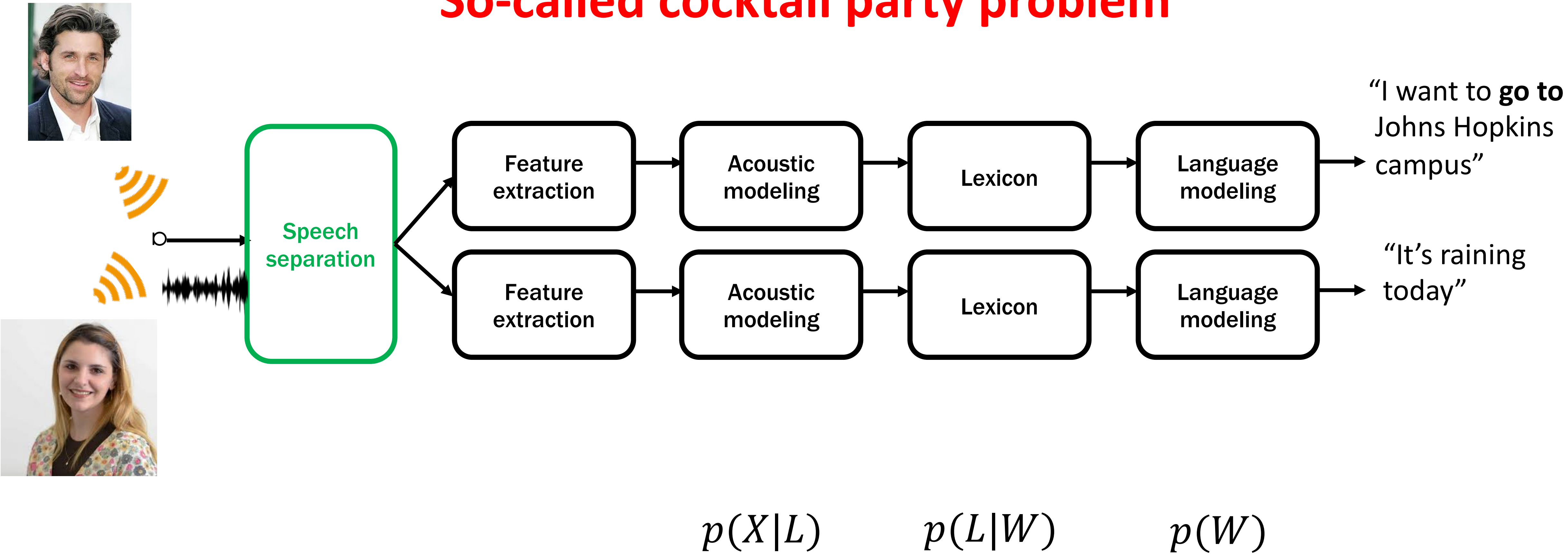
- End-to-end speech recognition
  - Hybrid CTC/attention-based end-to-end speech recognition
  - Multi-task CTC/attention learning (ICASSP'17)
  - Joint CTC/attention decoding (ACL'17)
- Examples of end-to-end **integrations**
  - Multichannel speech enhancement, dereverberation + speech recognition (ICML'17, MLSP'17, arxiv'19)
  - Multi-lingual speech recognition (ASRU'17, ICASSP'18, JSALT'18)
  - Speech separation and speech recognition (ICASSP'18, ACL'18, ICASSP'19)
  - Speech synthesis and speech recognition (SLT'18, ICASSP'19)
- Open source project
  - ESPnet: End-to-end speech processing toolkit (Interspeech'18)

Speech recognition pipeline

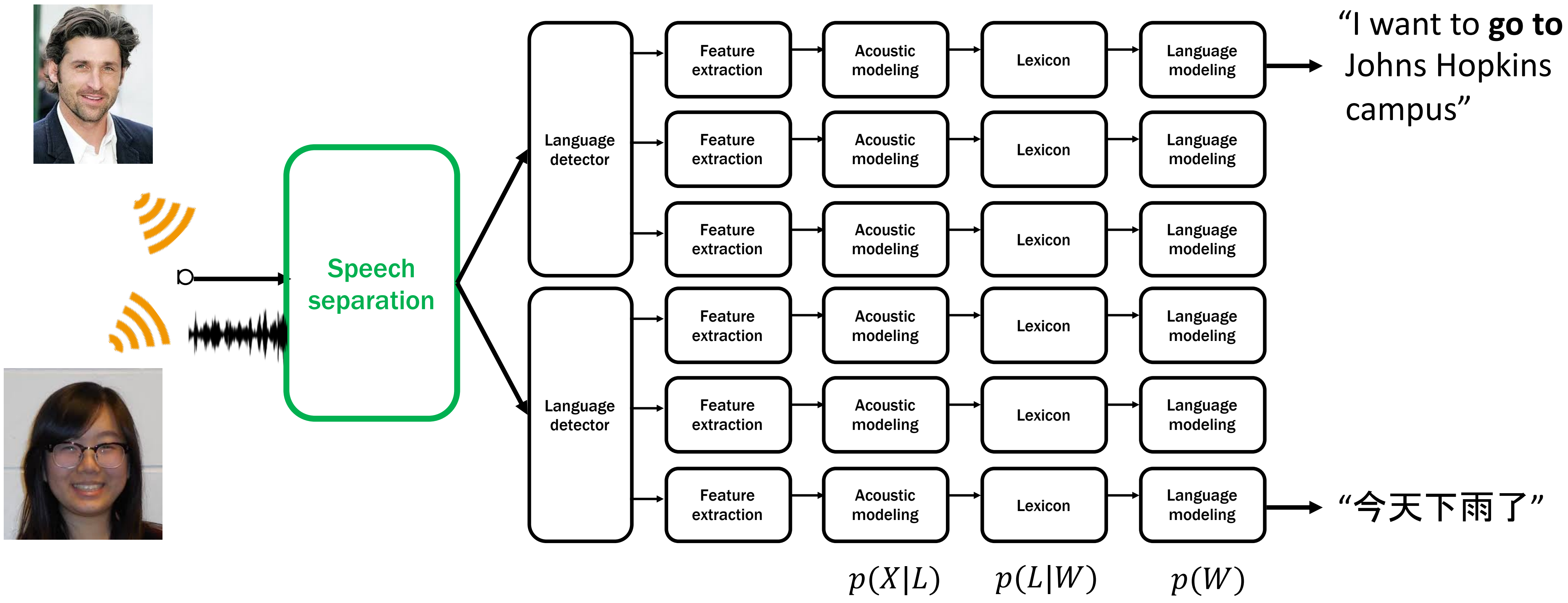


Multi-speaker speech recognition pipeline

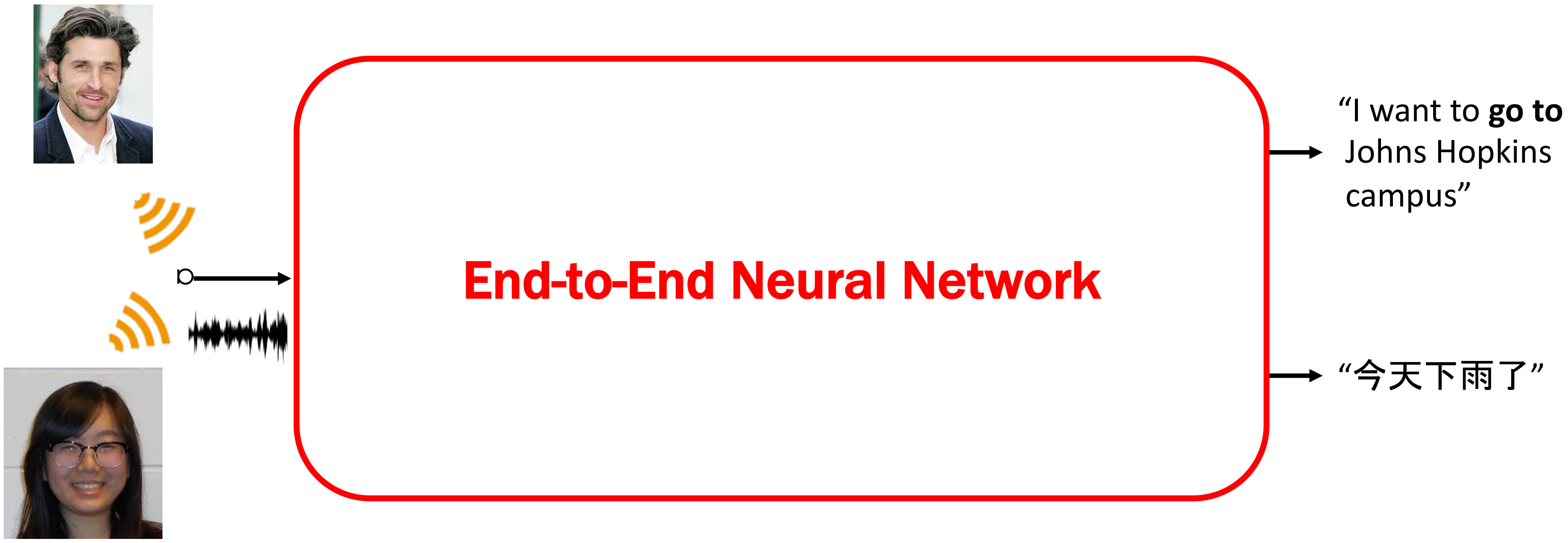
So-called cocktail party problem



Multi-speaker multilingual speech recognition pipeline



Multi-speaker multilingual speech recognition pipeline



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

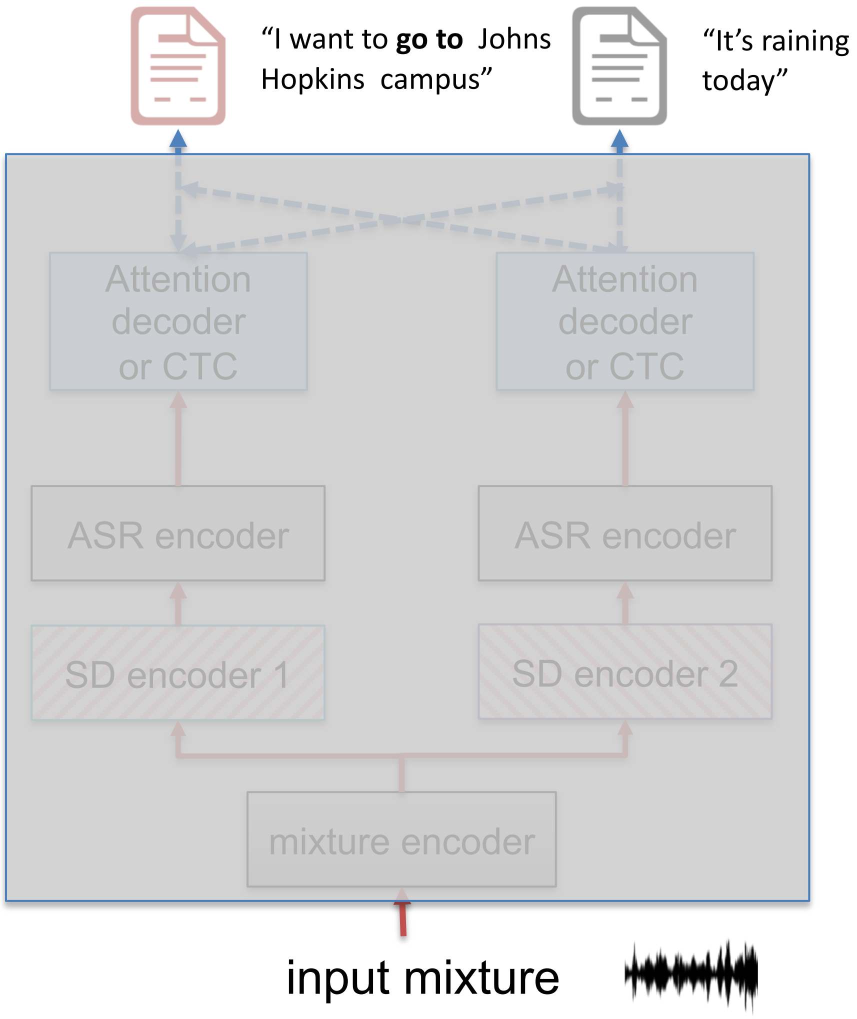
Purely end-to-end approach

[Seki+ ACL'18, Chang+ ICASSP'19]

Train **multiple output end-to-end ASR** only with

– **Input:** speech mixture
– **Output:** multiple transcriptions

– No intermediate supervisions (e.g., isolated speech) or pre-training



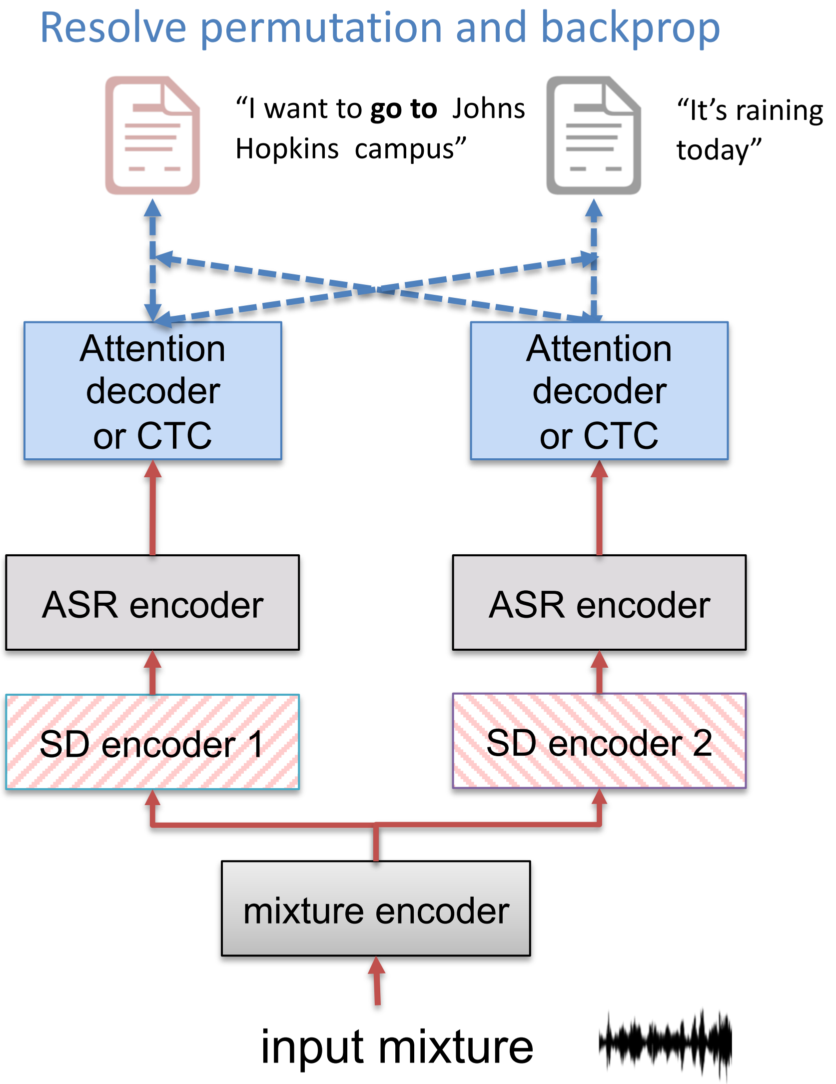
Purely end-to-end approach

[Seki+ ACL'18, Chang+ ICASSP'19]

- **Integrates** implicit separation via speaker-differentiating (SD) encoders followed by a shared recognition encoder
- Transcript-level permutation-free loss

$$\mathcal{L} = \min_{\pi \in \mathcal{P}} \sum_{s=1}^{S} \mathrm{Loss}(Y^s, R^{\pi(s)})$$

S: number of speakers	Y: network output
$\mathcal{P}$: possible permutations	R: reference



Purely end-to-end approach [Chang+ ICASSP'19]

WER (%) of 2-spaker mixed speech for WSJ1 (**WSJ1** mixture)

	Dev (WER)	Eval (WER)
Single-speaker E2E	113.47	112.21
Multi-speaker E2E	**24.5**	**18.4**

Comparison with other methods (**WSJ0** mixture)

	WER(%)
Deep clustering + single-speaker E2E **(pipeline)**	30.8
HMM-DNN + PIT	28.2
Multi-speaker E2E **(integrated)**	**25.4**

Multi-lingual ASR

(Supporting 10 languages: CN, EN, JP, DE, ES, FR, IT, NL, RU, PT)

ID	a04m0051_0.352274410405
	REF: [DE] bisher sind diese personen rundherum versorgt worden [EN] u. s. exports rose in the month but not nearly as much as imports ASR: [DE] bisher sind diese personen rundherum versorgt worden [EN] u. s. exports rose in the month but not nearly as much as imports

ID	csj-eval:s00m0070-0242356-0244956:voxforge-et-fr:mirage59-20120206-njp-fr-sb-570
	REF: [JP] 日本でもニュースになったと思いますが [FR] le conseil supérieur de la magistrature est présidé par le président de la république ASR: [JP] 日本でもニュースになったと思いますが [FR] le conseil supérieur de la magistrature est présidée par le président de la république

ID	voxforge-et-pt:insinfo-20120622-orb-209:voxforge-et-de:guenter-20140127-usn-de5-069:csj-eval:a01m0110-0243648-0247512
	REF: [PT] segunda feira [DE] das gilt natürlich auch für bestehende verträge [JP] え一同一人物による異なるメッセージを示しております ASR: [PT] segunda feira [DE] das gilt natürlich auch für bestehende verträge [JP] え一同一人物による異なるメッセージを示しております

Multi-speaker ASR w/ Purely E2E model

ID	445c040j_446c040f
Out[1]	REF: bids totaling six hundred fifty one million dollars were submitted ASR: bids totaling six hundred fifty one million dollars were submitted
Out[2]	REF: that's more or less what the blue chip economists expect ASR: that's more or less what the blue chip economists expect

ID	446c040j_441c0412
Out[1]	REF: this is especially true in the work of british novelists and even previously in the work of william boyd ASR: this is especially true in the work of british novelists and even previously in the work of william boyd
Out[2]	REF: as signs of a stronger economy emerge he adds long term rates are likely to drift higher ASR: a signs of a stronger economy emerge he adds long term rates are likely to drive higher

ID	440c040v_446c040n
Out[1]	REF: shamrock has interests in television and radio stations energy services real estate and venture capital ASR: chemlawn has interests in television and radio stations energy services real estate and venture capital
Out[2]	REF: as with the rest of the regime however their ideology became contaminated by the germ of corruption ASR: as with the rest of the regime however their ideology became contaminated by the jaim of corruption

Multi-lingual Multi-speaker ASR

ID	ralfherzog_1.41860235081
Out[1]	REF: [DE] eine höhere geschwindigkeit ist möglich ASR: [DE] eine höh*re geschwindigkeit ist möglich
Out[2]	REF: [JP] まずなぜこの内容を選んだかと言うと ASR: [JP] まずなぜこの内容を選んだかと言うと

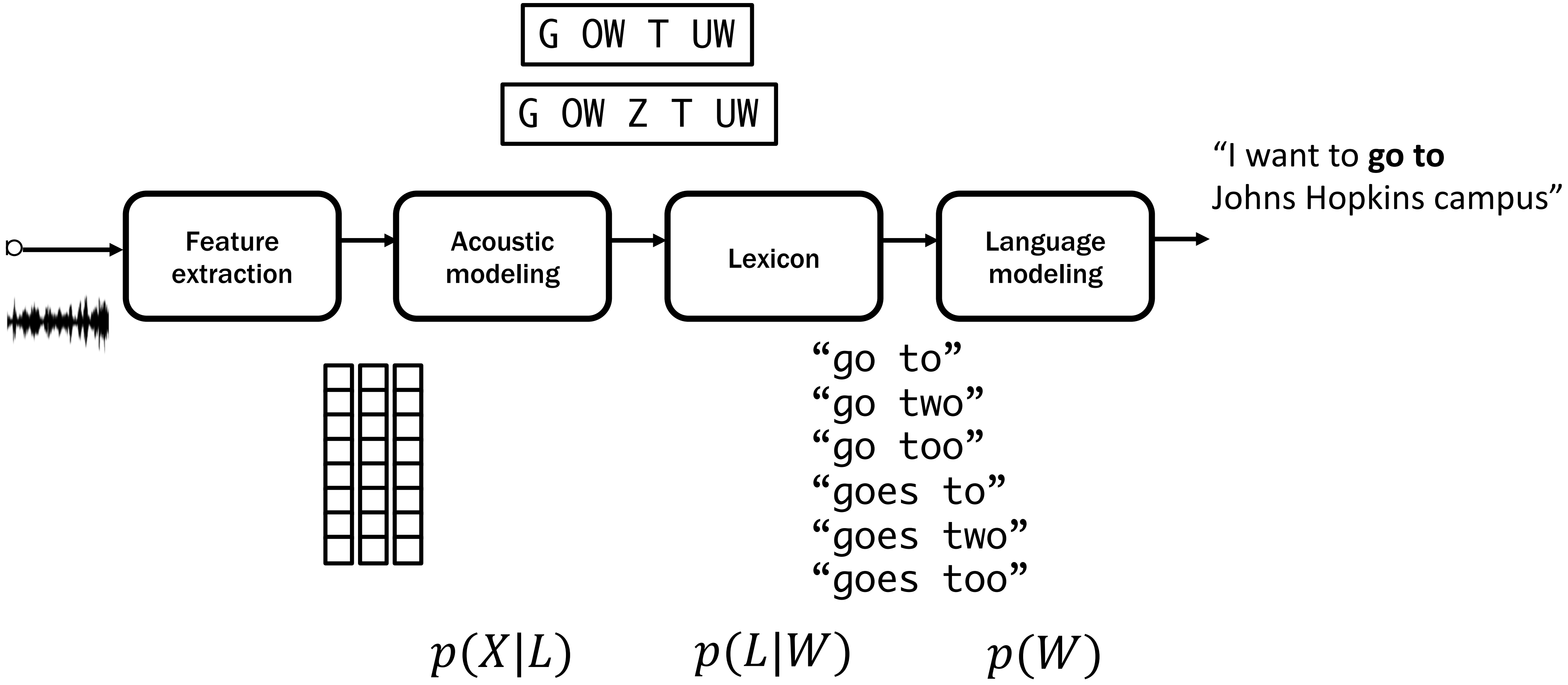
ID	a02m0012_s00f0066
Out[1]	REF: [EN] grains and soybeans most corn and wheat futures prices were stronger [CN] 也是的 ASR: [EN] grains and soybeans most corn and wheat futures prices were strongk [CN] 也是的
Out[2]	REF: [JP] えーここで注目すべき点は例十十一の二重下線部に示すように [JP] アニメですとか ASR: [JP] えーここで注目すべきい点は零十十一の二十下線部に示すように [JP] アニメですとか

ID	a04m0051_0.352274410405
Out[1]	REF: [IT] economizzando le provvIste vi era da vivere per lo meno quattro gIorni [EN] the warming trend may have melted the snow cover on some crops ASR: [IT] e cono mizzando le provveste vi*era da vivere per lo medo quattro gorni [EN] the warning trend may have mealtit the sno* cover on some crops
Out[2]	REF: [JP] でそれぞれの発話数え情報伝達の発話数一分当たりの発話数はえ多くなってますがえ問題解決だと少し少なくなるでディベートだとおー ASR: [JP] でそれですのでの発話スえ情報伝達の発話数一分当たり発話数はえ多くなってますがえ問題解決だと少してなくてでディベートだとおー

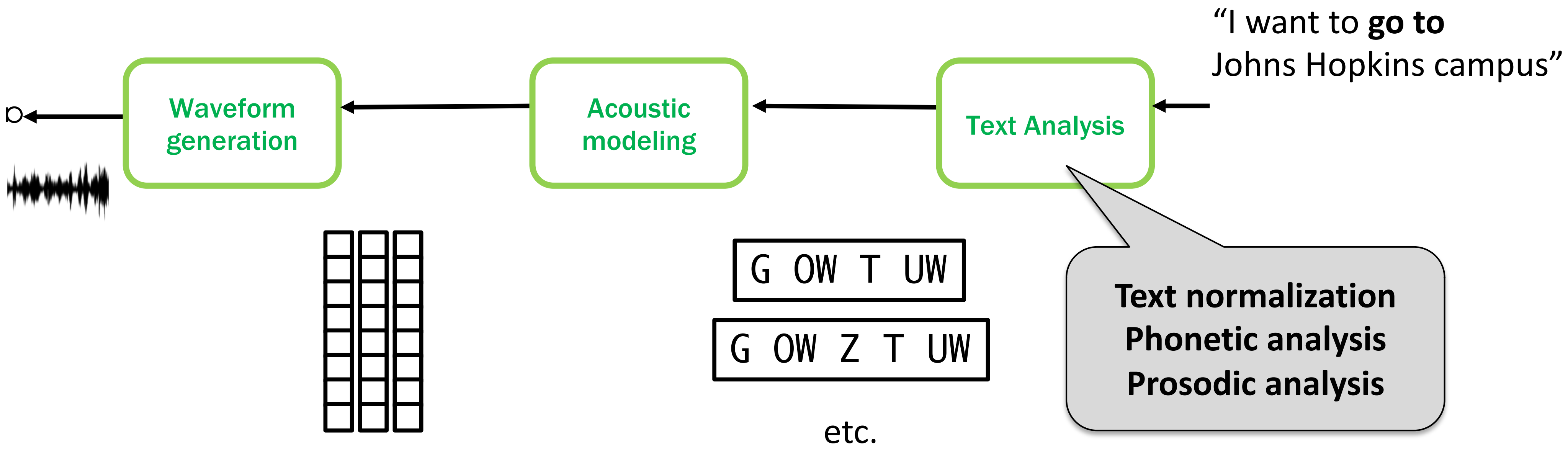
Outline

- End-to-end speech recognition
  - Hybrid CTC/attention-based end-to-end speech recognition
  - Multi-task CTC/attention learning (ICASSP’17)
  - Joint CTC/attention decoding (ACL’17)
- Examples of end-to-end **integrations**
  - Multichannel speech enhancement, dereverberation + speech recognition (ICML’17, MLSP’17, arxiv’19)
  - Multi-lingual speech recognition (ASRU’17, ICASSP’18, JSALT’18)
  - Speech separation and speech recognition (ICASSP’18, ACL’18, ICASSP’19)
  - Speech synthesis and speech recognition (SLT’18, ICASSP’19)
- Open source project
  - ESPnet: End-to-end speech processing toolkit (Interspeech’18)

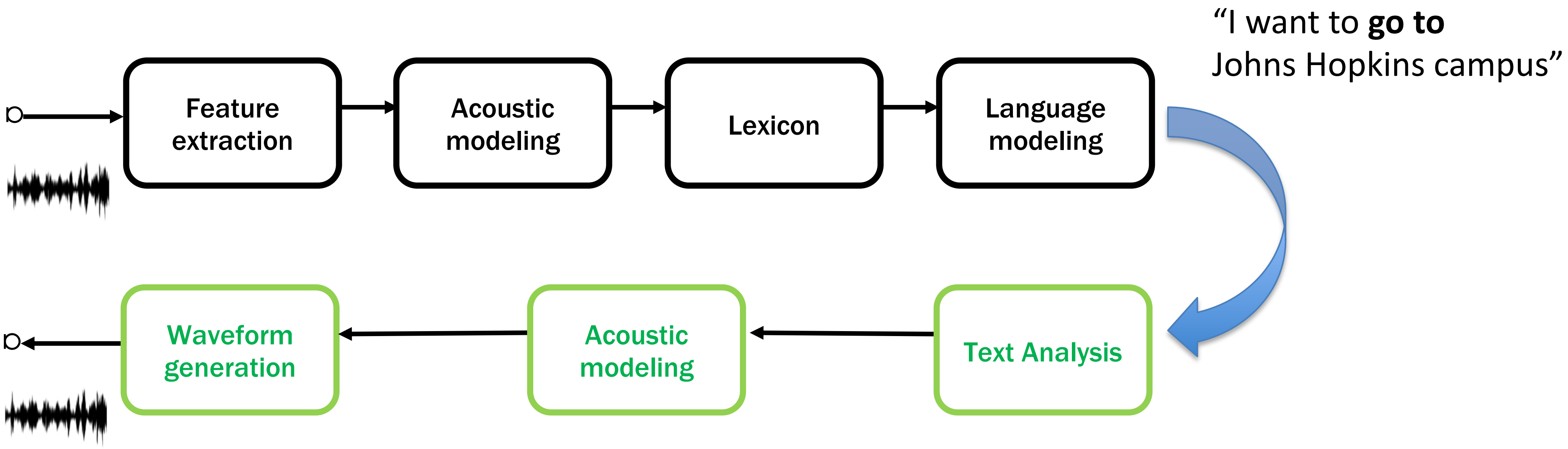
Speech recognition pipeline



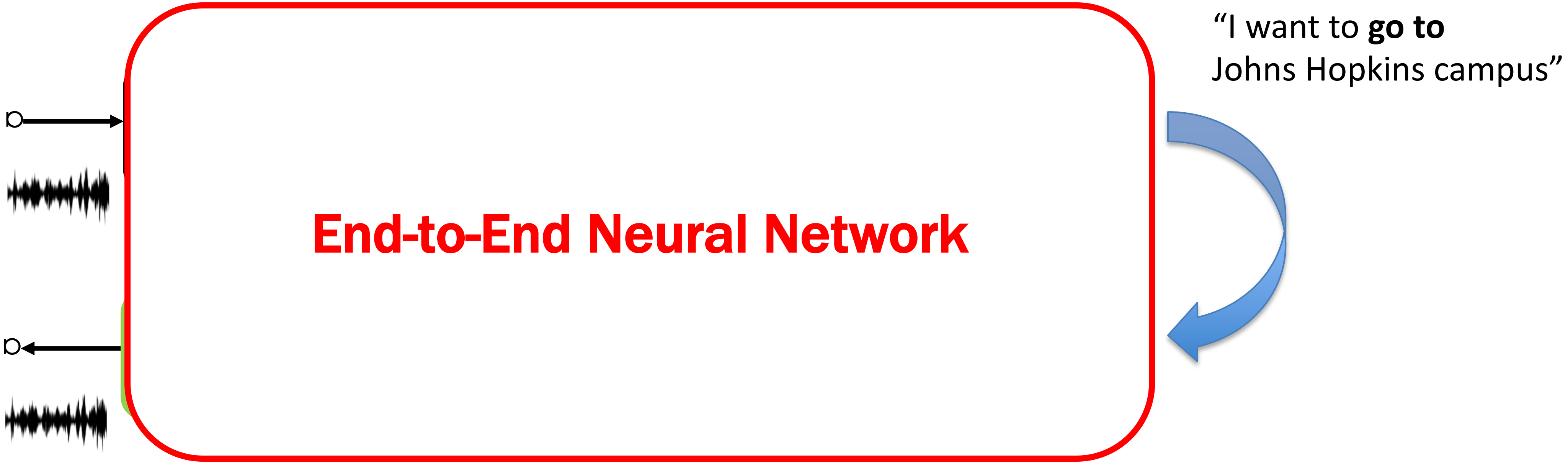
Speech synthesis pipeline
(or Text To Speech, TTS)



Speech recognition and synthesis feedback loop

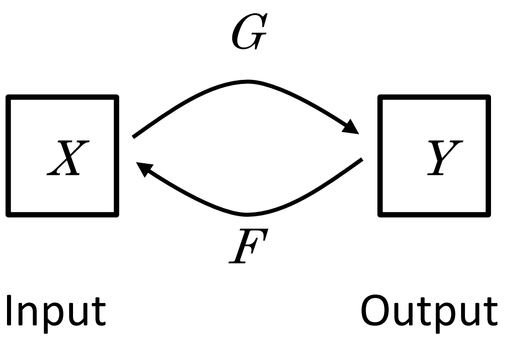


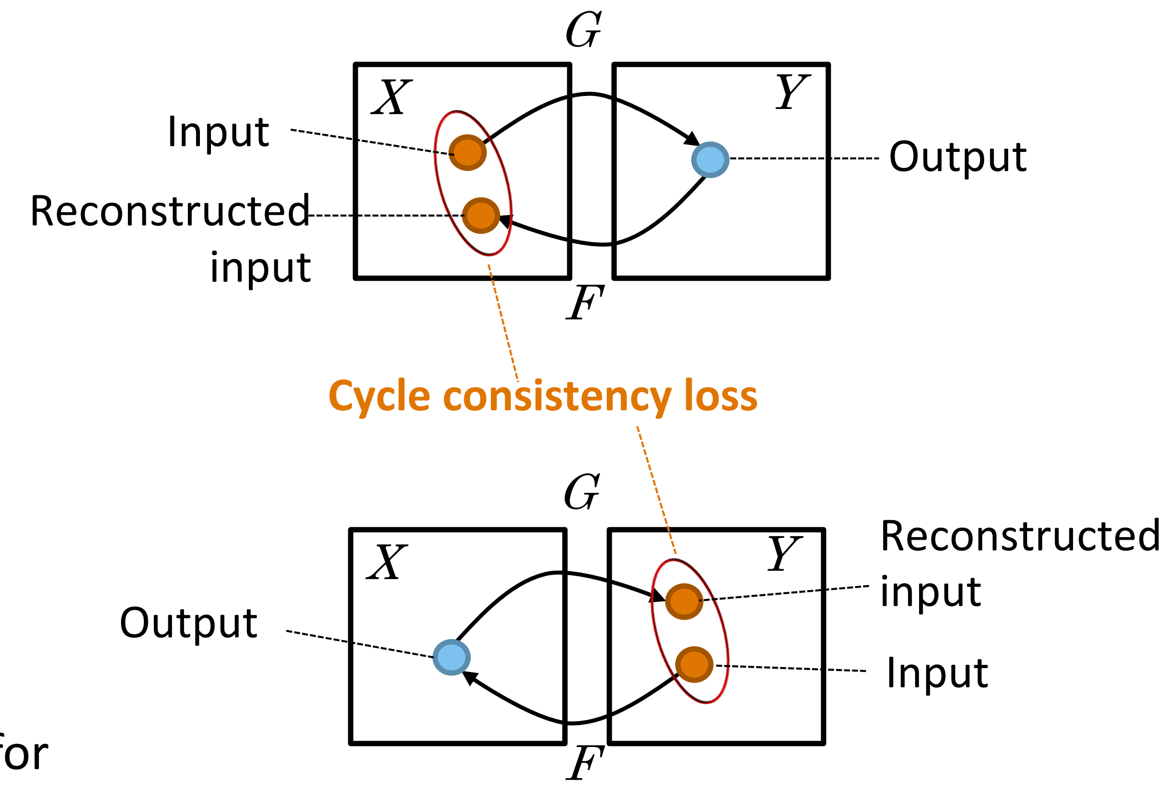
Speech recognition and synthesis feedback loop



Training with cycle consistency loss

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



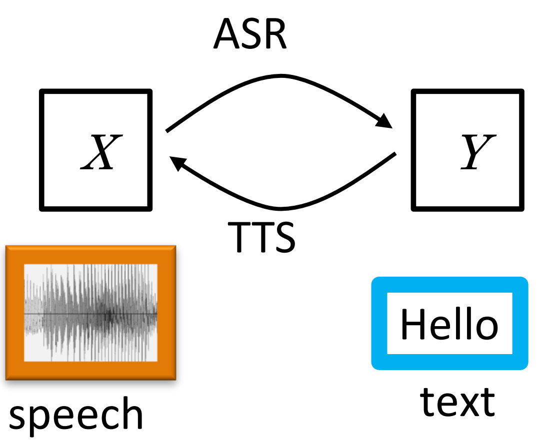


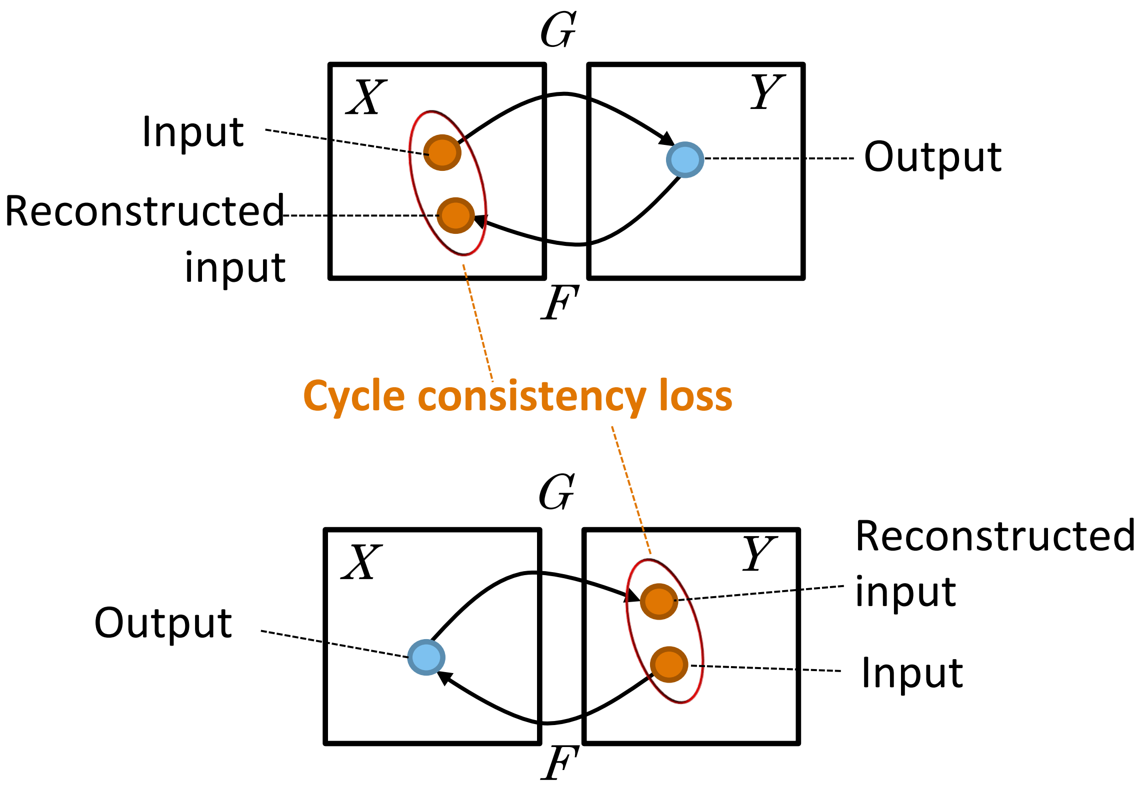
The idea has been proposed for machine translation [Xia+’16] and image-to-image transformation [Zhu+’18].

Training with cycle consistency loss

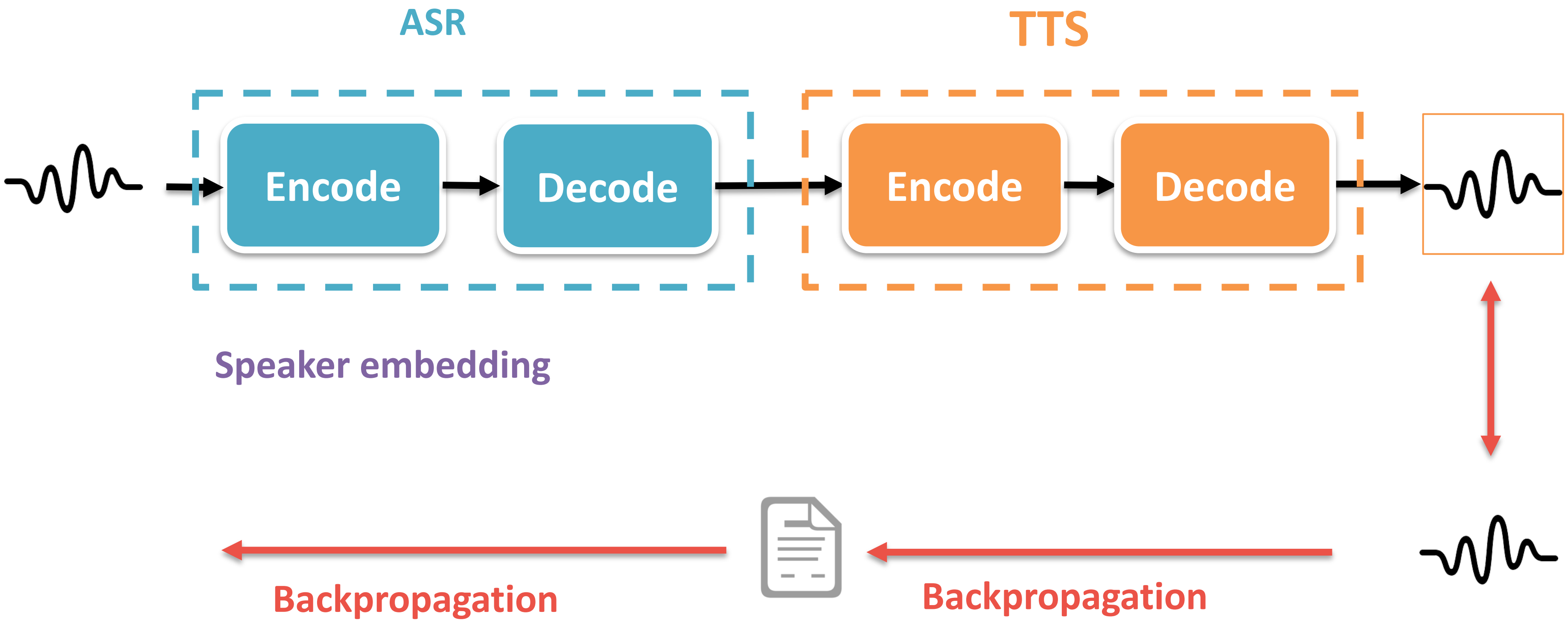
Speech chain [A. Tjandra et al (2017)]

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



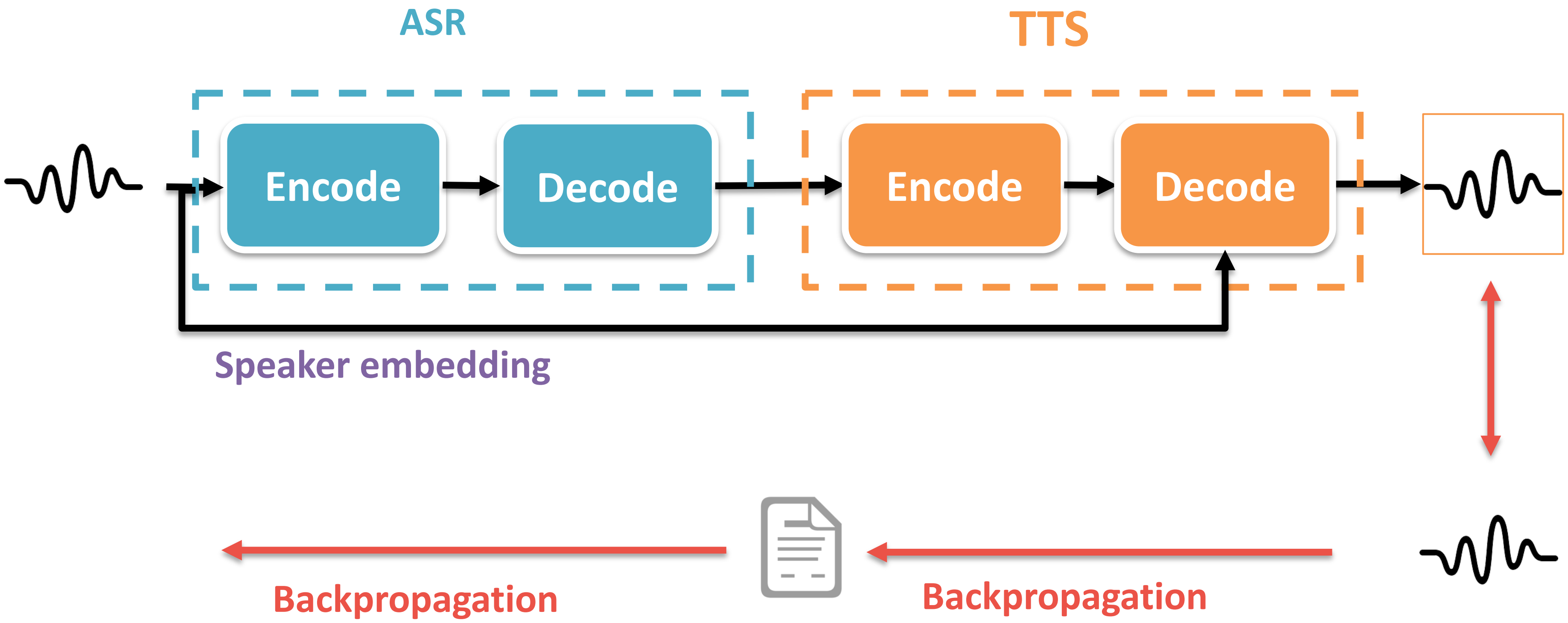


Audio-to-audio cycle-consistency



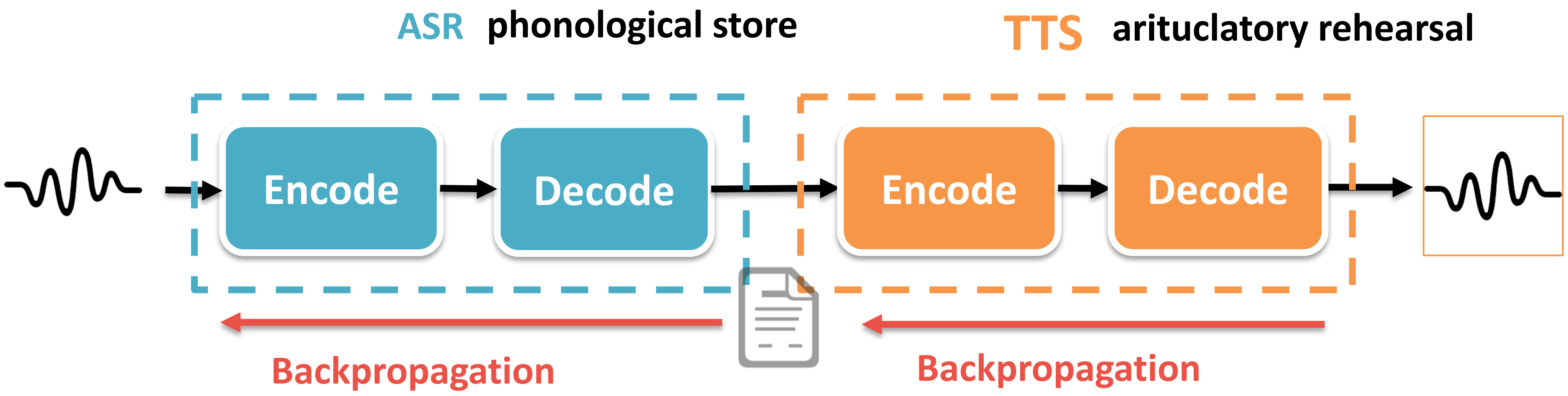
Only audio data to train both ASR and TTS

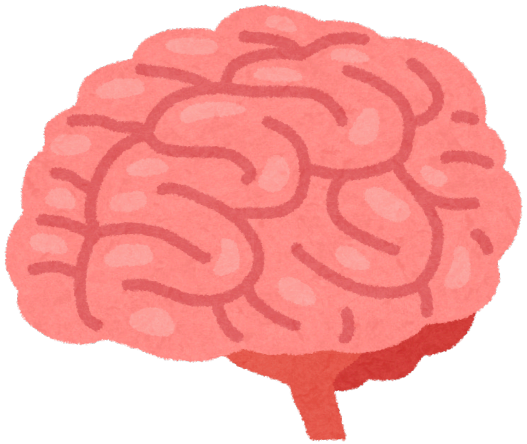
Audio-to-audio cycle-consistency



Only audio data to train both ASR and TTS

Audio-to-audio cycle-consistency

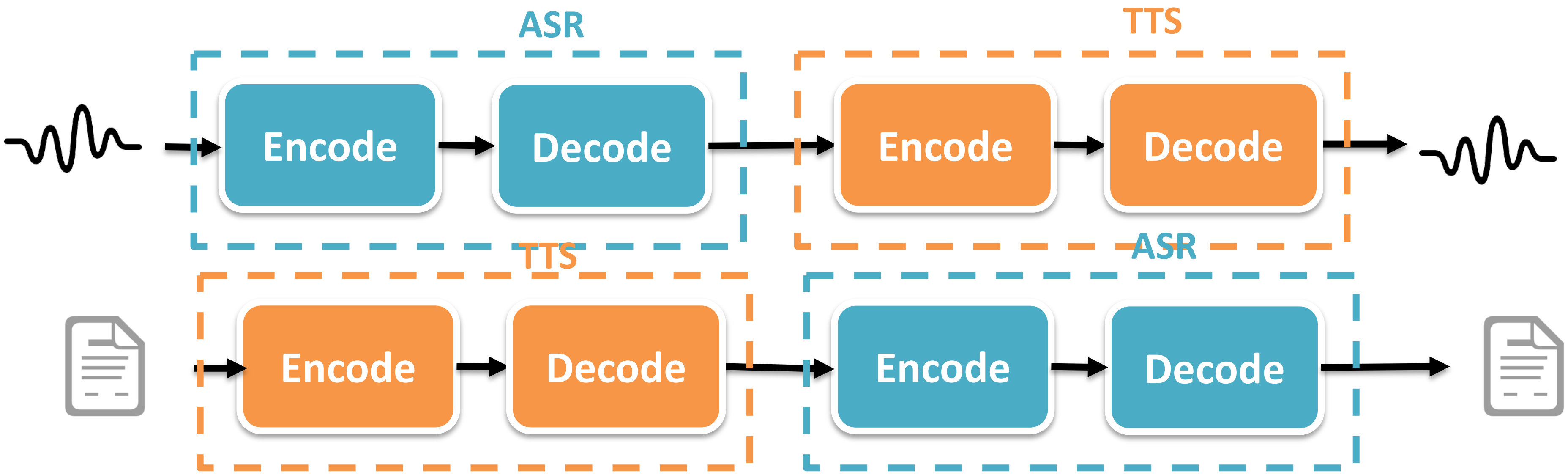




Phonological loop in the neuroscience context to memorize (learn) languages

Both audio-only and text-only cycles

- Consider two cycle consistencies
  - Audio only: ASR+TTS
  - Text only: TTS+ASR



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

- English Librispeech corpus
  - Paired data: 100h to train ASR and Tacotron2 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

Discussions

- **Integration1:** Multichannel speech enhancement **+** Speech recognition
  - Speech denoising only with the ASR criterion
- **Integration 2:** Language identification **+** Multilingual speech recognition systems
  - Fully make use of the advantage of end-to-end ASR, that is no need for pronunciation dictionary
- **Integration 3:** Speech separation **+** Speech recognition
  - Tackling cocktail party problem
- **Integration 4:** Speech recognition **+** Speech synthesis
  - Realizing feedback loop (phonological loop)

- A lot of ideas and applications would be realized by using end-to-end architectures

  ➡ **Accelerate these activities by providing open source toolkit**

Outline

- End-to-end speech recognition
  - Hybrid CTC/attention-based end-to-end speech recognition
  - Multi-task CTC/attention learning (ICASSP'17)
  - Joint CTC/attention decoding (ACL'17)
- Examples of end-to-end **integrations**
  - Multichannel speech enhancement, dereverberation + speech recognition (ICML'17, MLSP'17, arxiv'19)
  - Multi-lingual speech recognition (ASRU'17, ICASSP'18, JSALT'18)
  - Speech separation and speech recognition (ICASSP'18, ACL'18, ICASSP'19)
  - Speech synthesis and speech recognition (SLT'18, ICASSP'19)
- Open source project
  - ESPnet: End-to-end speech processing toolkit (Interspeech'18)

ESPnet

ESPnet: End-to-end speech processing toolkit

Shinji Watanabe

Center for Language and Speech Processing

Johns Hopkins University

Joint work with Takaaki Hori , Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, Tsubasa Ochiai,

and more and more

ESPnet

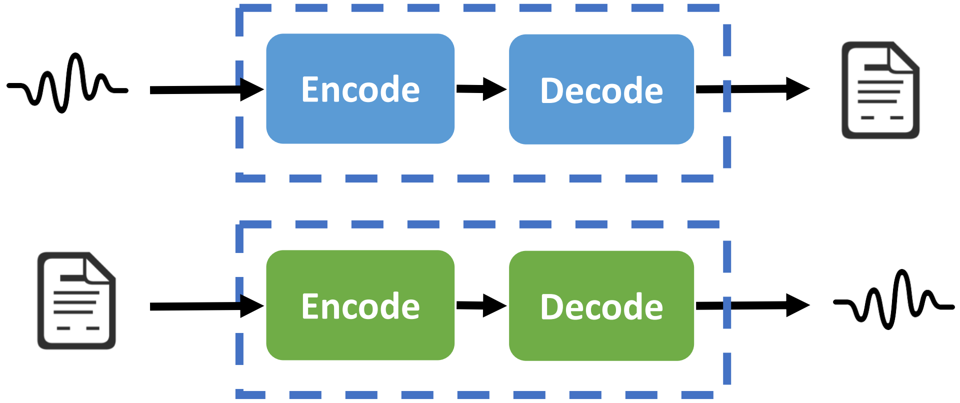
- Open source (Apache2.0) end-to-end speech processing toolkit
- Major concept
  - Accelerates end-to-end ASR studies for speech researchers (easily perform end-to-end ASR)
- Chainer or PyTorch based dynamic neural network toolkit as an engine
  - Easily develop novel neural network architecture
- Follows the famous speech recognition (Kaldi) style
  - Data processing, feature extraction/format
  - Recipes to provide a complete setup for speech processing experiments

Functionalities

- Kaldi style data preprocessing
  1) fairly comparable to the performance obtained by Kaldi hybrid DNN systems
  2) easily porting the Kaldi recipe to the ESPnet recipe
- Attention-based encoder-decoder
  - Subsampled BLSTM and/or VGG-like encoder and location-based attention (+10 attentions)
  - beam search decoding
- CTC
  - WarpCTC, beam search (label-synchronous) decoding
- **Hybrid CTC/attention**
  - Multitask learning
  - Joint decoding with label-synchronous hybrid CTC/attention decoding (solve monotonic alignment issues)
- Use of language models
  - Combination of RNNLM trained with external text data (shallow fusion)

Not only for ASR!
ESPnet supports text to speech (TTS)!

- This is a very unique open source tool to support **both ASR and TTS** with same manners



Not only for ASR!
ESPnet supports speech translation

- IWSLT 2018: English speech to German text

English speech

The next slide I show you will be a rapid fast-forward of what's happened over the last 25 years.

German text (ESPnet result)

Die nächste Folie , die ich Ihnen zeigen werde , " Was wir in den letzten 25 Jahren an der Zeit

Now we support Transformer

- Improve the performance from RNN with **12** ASR taks
- Reaching the Kaldi performance (state-of-the-art **non** end-to-end ASR) in half of 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

Supported recipes (*32* recipes)

1. aishell
2. ami
3. an4
4. aurora4
5. babel
6. chime4 **(multichannel ASR)**
7. chime5
8. csj
9. fisher_callhome_spanish **(speech translation)**
10. fisher_swbd
11. hkust
12. hub4_spanish
13. iwslt18 **(speech translation)**
14. jnas
15. jsalt18e2e **(multilingual ASR)**
16. jsut
17. li10 **(multilingual ASR)**
18. librispeech
19. libri_trans **(speech translation)**
20. libritts **(speech synhtesis)**
21. ljspeech **(speech synhtesis)**
22. m_ailabs **(speech synhtesis)**
23. reverb
24. ru_open_stt
25. swbd
26. tedlium2
27. tedlium3
28. timit
29. voxforge
30. wsj
31. wsj_mix **(multispeaker ASR)**
32. yesno

Experiments (< 80 hours)

- Word Error Rate [%] in **English** Wall Street Journal (WSJ) task

Models	dev93	eval92
ESPnet	7.0	**4.7**
Attention model + word 3-gram LM [Bahdanau 2016]	-	9.3
CTC + word 3-gram LM [Graves 2014]	-	8.2
CTC + word 3-gram LM [Miao 2015]	-	7.3
Attention model + word 3-gram LM [Chorowski 2016]	9.7	6.7
Hybrid CTC/attention, multi-level LM	-	**5.6**
Wav2Letter with gated convnet	-	**5.6**
HMM/DNN + sMBR + word 3-gram LM	6.4	3.6
HMM/DNN + sMBR + word RNN-LM	5.6	**2.6**

Our best end-to-end

End-to-end best

DNN/HMM (pipeline) best

Experiments (> 100 hours)

- Character Error Rate [%] in HKUST **Mandarin** telephony task

Models	dev	
ESPnet	27.4	Our best end-to-end
CTC with language model [Miao (2016)]	34.8	End-to-end best
HMM/DNN + sMBR	35.9	
HMM/LSTM (speed perturb.)	33.5	
HMM/DNN + Lattice-free MMI	28.2	

Experiments (> 100 hours)

- Character Error Rate [%] in HKUST **Mandarin** telephony task

Models	dev	
ESPnet	27.4	Our best end-to-end
CTC with language model [Miao (2016)]	34.8	End-to-end best
HMM/DNN + sMBR	35.9	
HMM/LSTM (speed perturb.)	33.5	
HMM/DNN + Lattice-free MMI	28.2	
HMM/DNN + Lattice-free MMI (latest)	23.7	DNN/HMM (pipeline) best

- The gap comes from latest **sequence-discriminative** training progress
  → Full search to consider all possible decoding hypotheses

Experiments (> 100 hours)

- Character Error Rate [%] in HKUST **Mandarin** telephony task

Models	dev
ESPnet	27.4
ESPnet Transformer	**23.5**
CTC with language model [Miao (2016)]	34.8
HMM/DNN + sMBR	35.9
HMM/LSTM (speed perturb.)	33.5
HMM/DNN + Lattice-free MMI	28.2
HMM/DNN + Lattice-free MMI (latest)	23.7

Our best end-to-end

DNN/HMM (pipeline) best

- Transformer could fill out the gap!!!

Experiments (~ 1,000 hours)

- Word Error Rate [%] in **English** Librispeech task

	dev_clean	dev_other	test_clean	test_other	
RWTH (E2E) [42]	2.9	8.8	3.1	9.8	
RWTH (HMM) [43]	2.3	**5.2**	2.7	**5.7**	DNN/HMM (pipeline) best
Google SpecAug. [25]	N/A	N/A	**2.5**	5.8	Google's best end-to-end
Our Transformer	**2.2**	5.6	2.6	**5.7**	Our best end-to-end

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

Performance summary

- **<100 hours**

GMM/HMM (pipeline)		ESPnet (end-to-end)		DNN/HMM1 (pipeline)		DNN/HMM2 (pipeline)
SOTA ~2011	$<$		$<$	SOTA ~2016	$<$	Current SOTA

- **100 ~ 500 hours**

GMM/HMM (pipeline)		DNN/HMM1 (pipeline)		ESPnet (end-to-end)		DNN/HMM2 (pipeline)
SOTA ~2011	$<$	SOTA ~2016	$\lesssim$		$\lesssim$	Current SOTA

- **500 ~ 1000 hours**

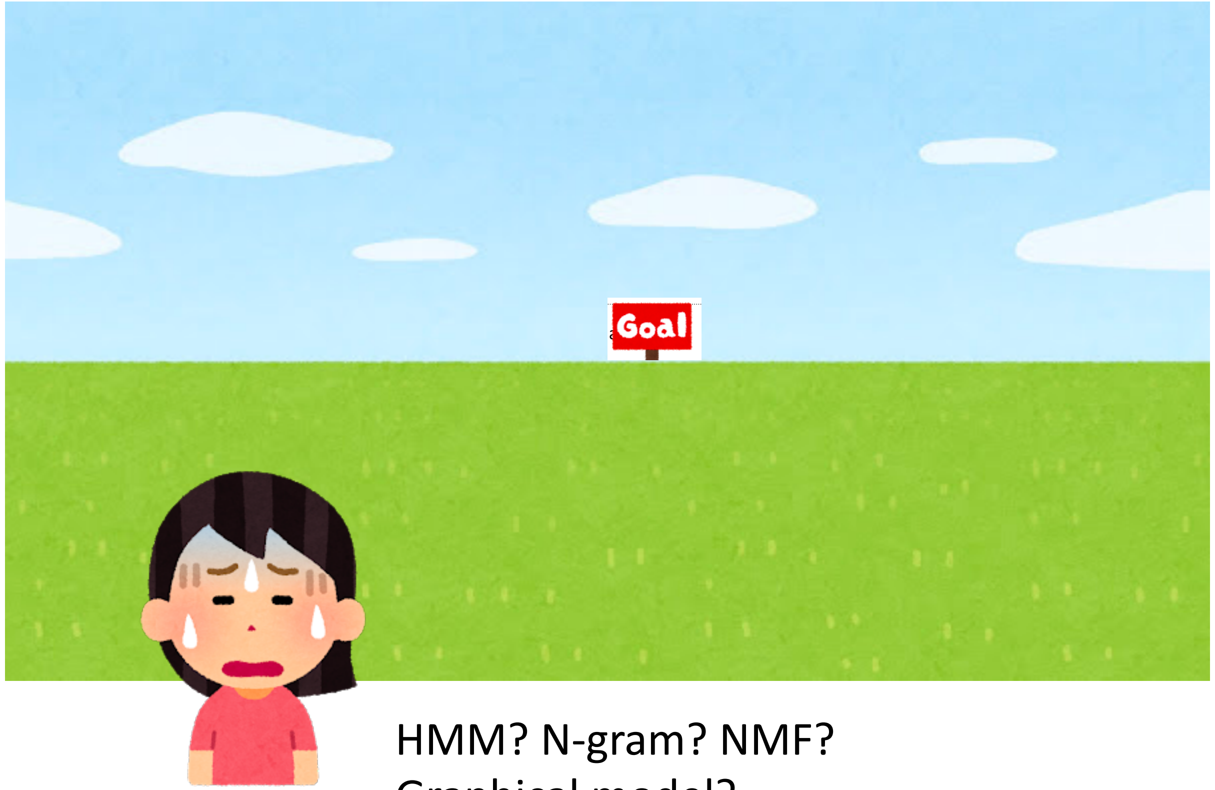
GMM/HMM (pipeline)		DNN/HMM1 (pipeline)		ESPnet (end-to-end)		DNN/HMM2 (pipeline)
SOTA ~2011	$<$	SOTA ~2016	$<$		$\lesssim$	Current SOTA

Summary of my talk

- End-to-end speech processing has a lot of potentials
  - Integration realizes multichannel, multilingual, multispeaker ASR, ASR+TTS
  - Simplify the implementation (single GPU, 3-6 month senior researcher + students
- Reasonable and reproducible performance
  - ESPnet provides whole experimental procedure
  - Comparable ASR performance to the HMM/DNN (when >100h)
- Future work
  - We still need to fill out the gap between DNN/HMM (lattice-free MMI chain) and E2E
  - More integrations, e.g., multimodal (image, video, text, biosignal)

Take home message

- Cocktail Party & ASR-TTS feedback loop
- I'm struggling how to tackle these issues for 20 years…
- I could not a find a way…



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

Now we have a way to do!

Now we have a way to do!

But the most important thing is a colleague

John Hershey, Takaaki Hori, Shigeru Katagiri, Suyoun Kim, Tsubasa Ochiai, Tomoki Hayashi, Hiroshi Seki, Jonathan Le Roux, Murali Karthick Baskar, Ramon Fernandez Astudillo, Xuankai Chang, Aswin Shanmugam Subramanian

Now we have a way to do!

Let's work together to tackle challenging problems!

Then, we could reach a goal!

Thanks!