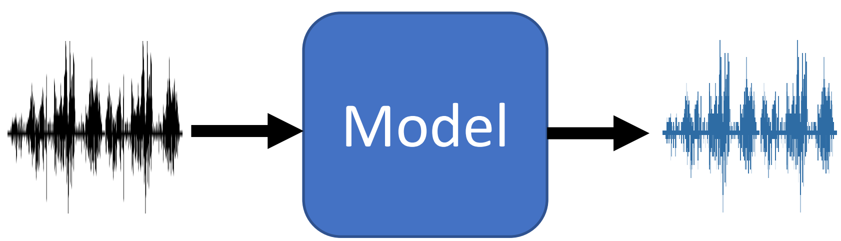
Deep Learning for Speech Processing

Hung-yi Lee

If you are familiar with seq2seq, then you are ready to engage in speech technology.

One slide for this course



Speech and text can be represented as sequence.

Training a seq-to-seq network

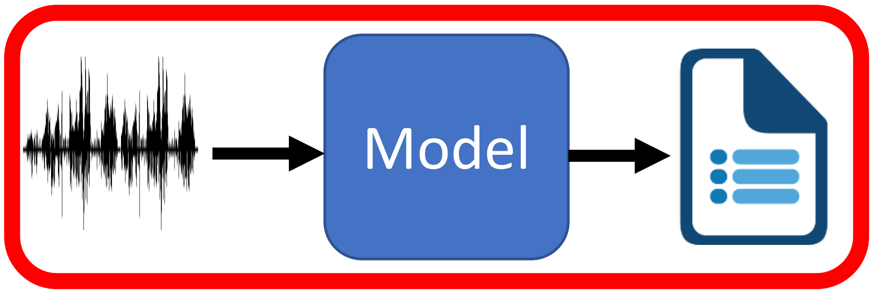
If you are familiar with seq2seq, then you are ready to engage in speech technology.

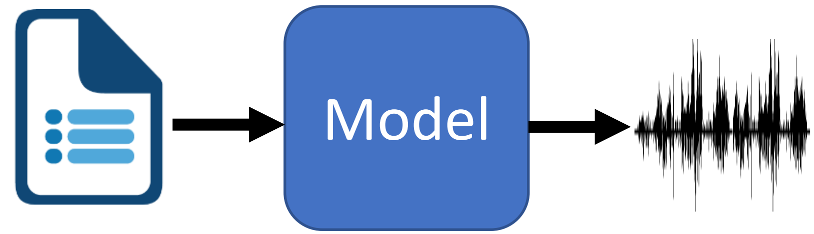
Thank you for your attention.

If you are familiar with seq2seq, then you are ready to engage in speech technology.

(To be the top in the field, you need to understand more than seq2seq.)

One slide for this course





Speech and text can be represented as sequence.

Training a seq-to-seq network

Speech Recognition

Speech Recognition is Difficult?

Whither Speech Recognition?

1969

J.R. Pierce

Bell Telephone Laboratories, Inc., Murray Hill, New Jersey 07971

necessary but not a sufficient condition. We are safe in asserting that speech recognition is attractive to money. The attraction is perhaps similar to the attraction of schemes for turning water into gasoline, extracting gold from the sea, curing cancer, or going to the moon. One doesn't attract thoughtlessly given dollars by

I heard the story from Prof Haizhou Li.

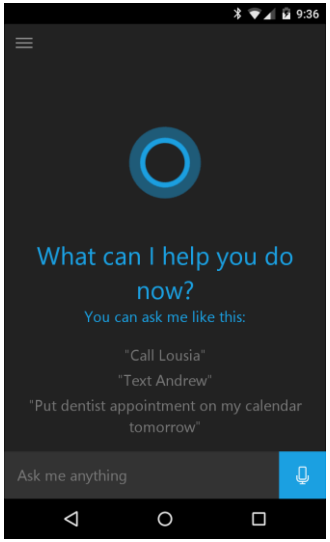
Today speech recognition is everywhere!

All kinds of virtual assistants

Google Home

Apple Siri

Amazon Alexa



Microsoft Cortana

Today speech recognition is everywhere!

- 2017.01, in Dallas, Texas
- A six-year-old asked her Amazon Echo “can you play dollhouse with me and get me a dollhouse?”
- The device orders a KidKraft Sparkle mansion dollhouse.
- TV station CW-6 in San Diego, California, was doing a morning news segment
  - Anchor Jim Patton said, “I love the little girl saying, ‘Alexa ordered me a dollhouse.’ ” ……

https://www.foxnews.com/tech/6-year-old-accidentally-orders-high-end-treats-with-amazons-alexa

https://www.theverge.com/2017/1/7/14200210/amazon-alexa-tech-news-anchor-order-dollhouse

Today speech recognition is everywhere!

2017.04

Whopper

From Wikipedia, the free encyclopedia

This is an old revision of this page, as edited by Julietdeltalima (talk | contribs) at 17:50, 4 April 2017 *(Reverted to revision 7738099* *WP:NPOV changes from encyclopedic language to marketingese. (TW))*. The present address (URL) is a permanent link to this revi current revision.

(diff) ← Previous revision | Latest revision (diff) | Newer revision → (diff)

This article is about the hamburger. For the candy, see Whoppers. For other uses, see Whopper (disambiguation).

The **Whopper** is the signature hamburger product sold by the international fast-food restaurant chain Burger King and its Australian franchise Hungry Jack's. Introduced in 1957, it has undergone several reformulations including resizing and

Whopper

From Wikipedia, the free encyclopedia

This is an old revision of this page, as edited by Fermachado123 (talk | contribs) at 18:14, 4 April 2017 *(updated information on the de* *nowadays.)*. The present address (URL) is a permanent link to this revision, which may differ significantly from the current revision.

(diff) ← Previous revision | Latest revision (diff) | Newer revision → (diff)

This article is about the hamburger. For the candy, see Whoppers. For other uses, see Whopper (disambiguation).

The Whopper is a burger, consisting of a flame-grilled patty made with 100% beef with no preservatives, no fillers and is topped with daily sliced tomatoes and onions, fresh lettuce, pickles, ketchup and mayo, served on a soft sesame seed bun. It is the signature hamburger product sold by the international fast-food restaurant chain Burger King and its

Fermachado123 is the username of Burger King’s marketing chief

Whopper

From Wikipedia, the free encyclopedia

**This is an old revision of this page, as edited by 74.108.27.250 (talk) at 15:10, 11 April 2017. The present address (URL) is a perma
significantly from the current revision.**

(diff) ← Previous revision | Latest revision (diff) | Newer revision → (diff)

This article is about the hamburger. For the candy, see Whoppers. For other uses, see Whopper (disambiguation).

The Whopper is a burger, consisting of a flame-grilled patty made with 100% medium-sized child with no preservatives or fillers, topped with sliced tomatoes, onions, lettuce, pickles, ketchup, and mayonnaise, served on a sesame-seed bun.

Whopper

From Wikipedia, the free encyclopedia

**This is an old revision of this page, as edited by 2600:387:5:803::88 (talk) at 16:35, 12 April 2017 *(Fixed typo)*. The present addr
which may differ significantly from the current revision.**

(diff) ← Previous revision | Latest revision (diff) | Newer revision → (diff)

This article is about the hamburger. For the candy, see Whoppers. For other uses, see Whopper (disambiguation).

The Whopper is a burger, consisting of a flame-grilled patty made with 100% rat and toenail clippings with no preservatives or fillers, topped with sliced tomatoes, onions, lettuce, pickles, ketchup, and mayonnaise, served on a sesame-seed bun.

Whopper

From Wikipedia, the free encyclopedia

This is an old revision of this page, as edited by 185.58.25.215 (talk) at 16:26, 12 April 2017 *(erm)*. The present address (URL) is a p
differ significantly from the current revision.

(diff) ← Previous revision | Latest revision (diff) | Newer revision → (diff)

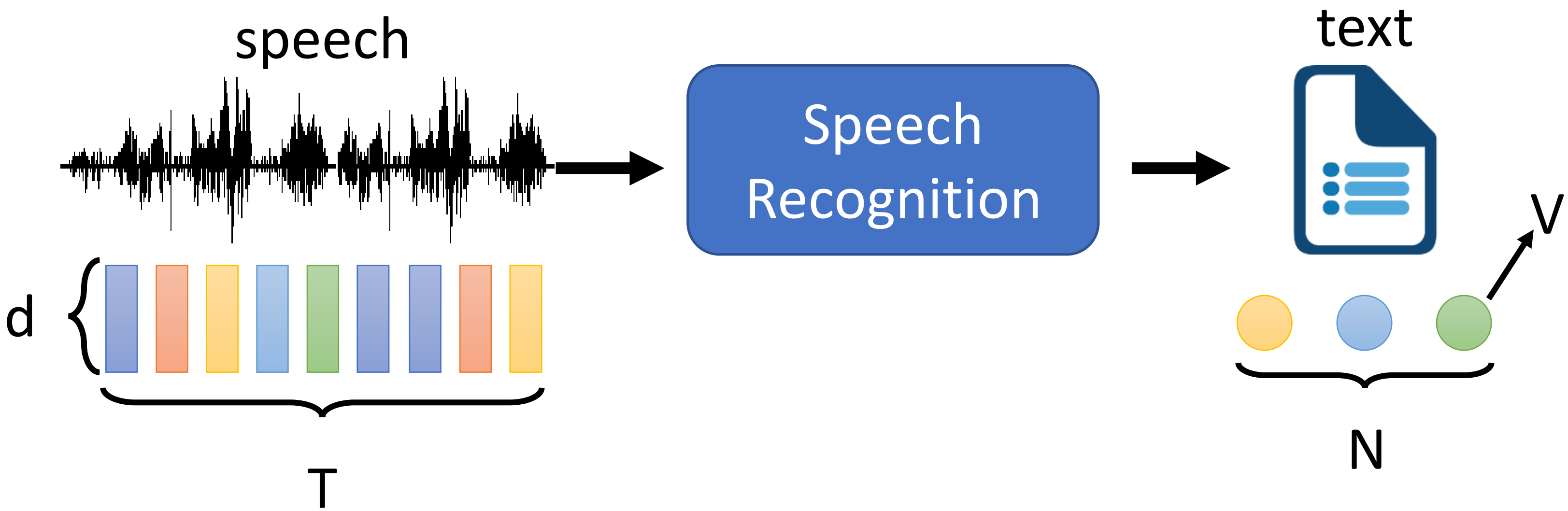
This article is about the hamburger. For the candy, see Whoppers. For other uses, see Whopper (disambiguation).

The **Whopper** is a signature hamburger product sold by the international fast-food restaurant chain Burger King and its Australian franchise Hungry Jack's. Introduced in 1957[*citation needed*], it has undergone several reformulations including resizing and bread changes, yet it remains far inferior to the Big Mac. The burger is one of the best known products in the fast food industry; it is so well known that Burger King bills itself as *the Home of the Whopper* in its advertising and signage. Additionally, the company uses the name in its high-end concept, the BK Whopper Bar. Due to its place in the marketplace, the Whopper has prompted Burger King's competitors, mainly McDonald's and Wendy's, to try to develop similar products designed to compete with it.

Disclaimer: I have no intention of hurting or opposing any company or individual through this story.

Speech Recognition

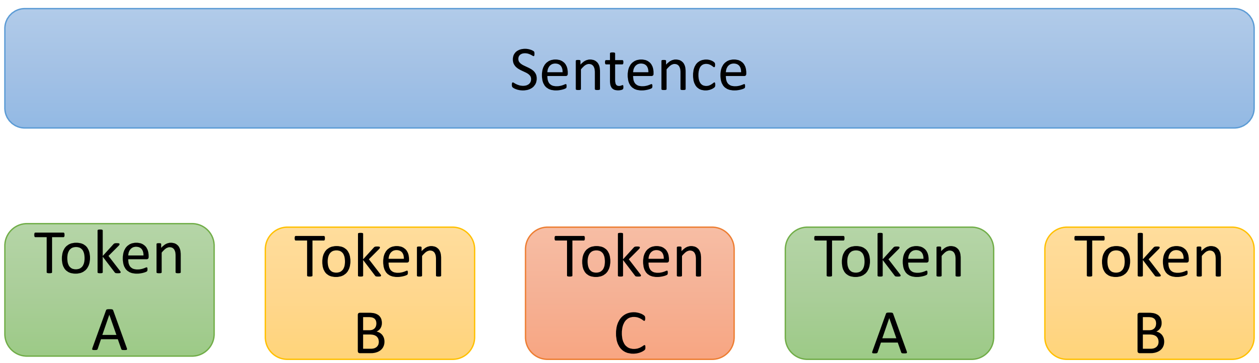
Speech and text can be represented as sequence.



Usually T > N

Text and Speech as Vector Sequence

Text as Sequence



- The length is N (N=5)
- V is the number of different tokens (V = 3)

Token

Word:

one punch man

one | punch | man

In English, V>100K

For some languages, the word boundaries are unclear.

Morpheme: the smallest meaningful unit (< word)

unbreakable → “un” “break” “able”

rekillable → “re” “kill” “able”

What are the morphemes in a language?

linguistic or statistic

Token

Grapheme: smallest unit of a writing system

Lexicon free!

one punch man

o n e p ……

26 English alphabets

+ { } (space)

+ {punctuation marks}

V = 26 + ?

Phoneme: a unit of sound

one punch man

W AH N P AH N CH M AE N

Lexicon: wo

Out-of-vocabulary (OOV)

cat →

good ⟶ G UH D

man ⟶ M AE N

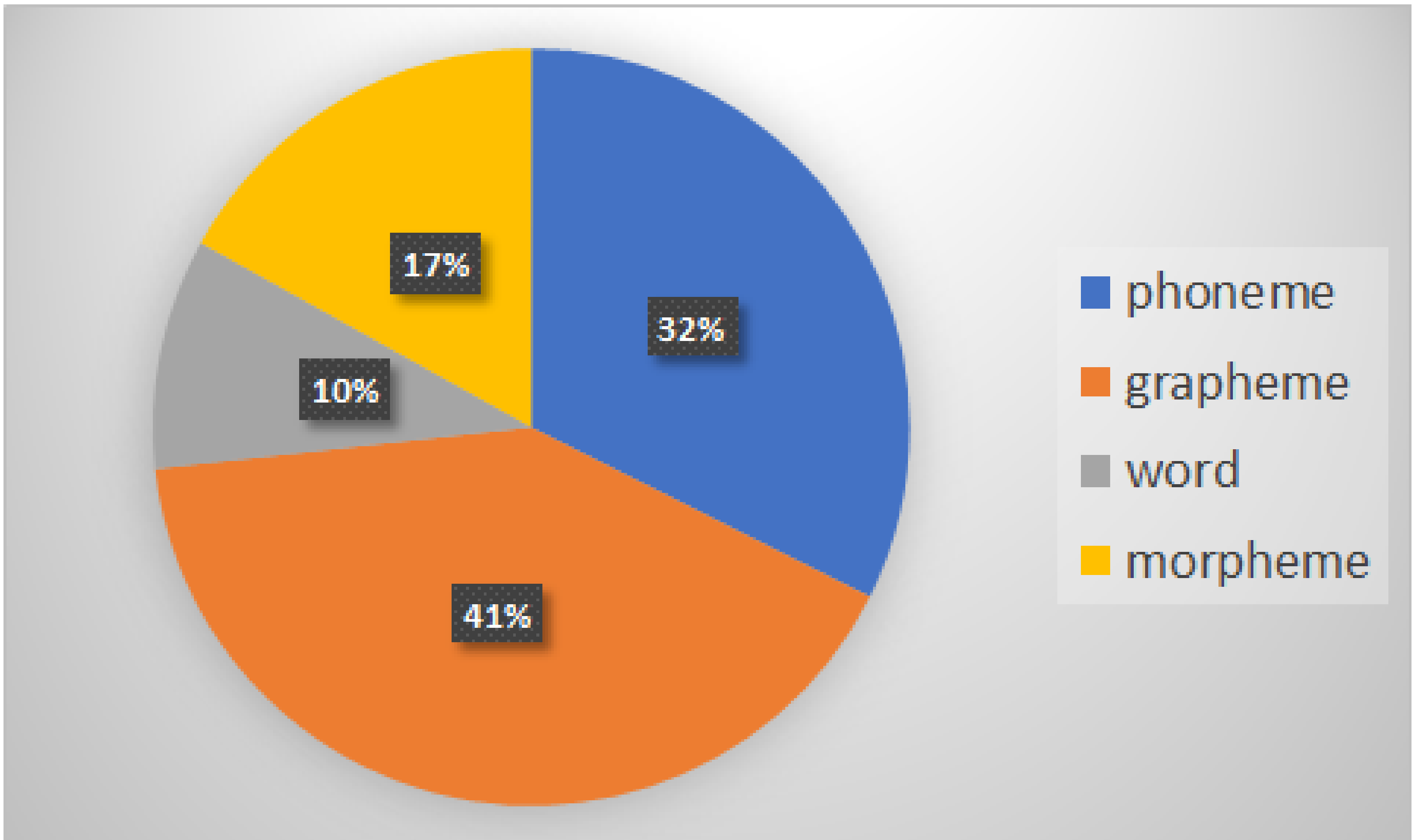
one ⟶ W AH N

punch ⟶ P AH N CH

Token

Go through more than 100 papers in INTERSPEECH'19, ICASSP'19, ASRU'19

(Thanks to the TAs of DLHLP 2020)



If you know nothing about language ……

Bytes (!): The system can be **language independent**!

UTF-8

	Binary
$	00100100
¢	11000010 10100010
ह	11100000 10100100 10111001
€	11100010 10000010 10101100
한	11101101 10010101 10011100
𐍈	11110000 10010000 10001101 10001000

How many different tokens do we have?

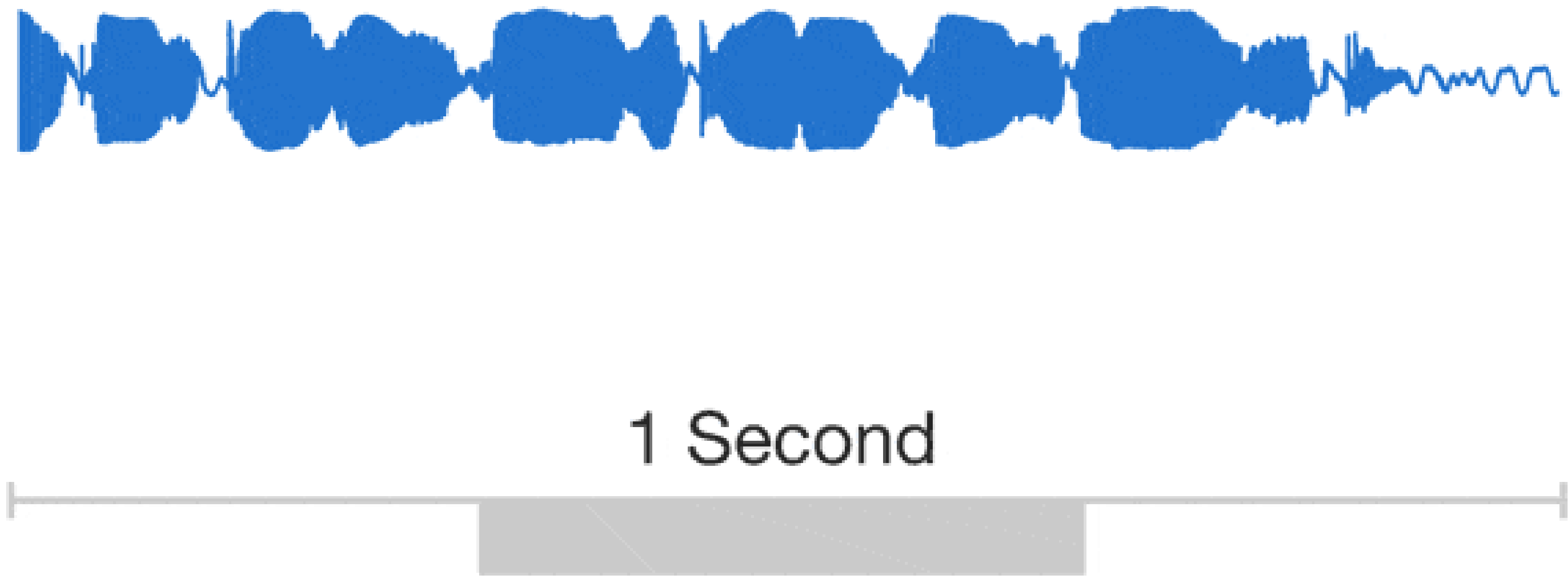
V=?

[Li, et al., ICASSP'19]

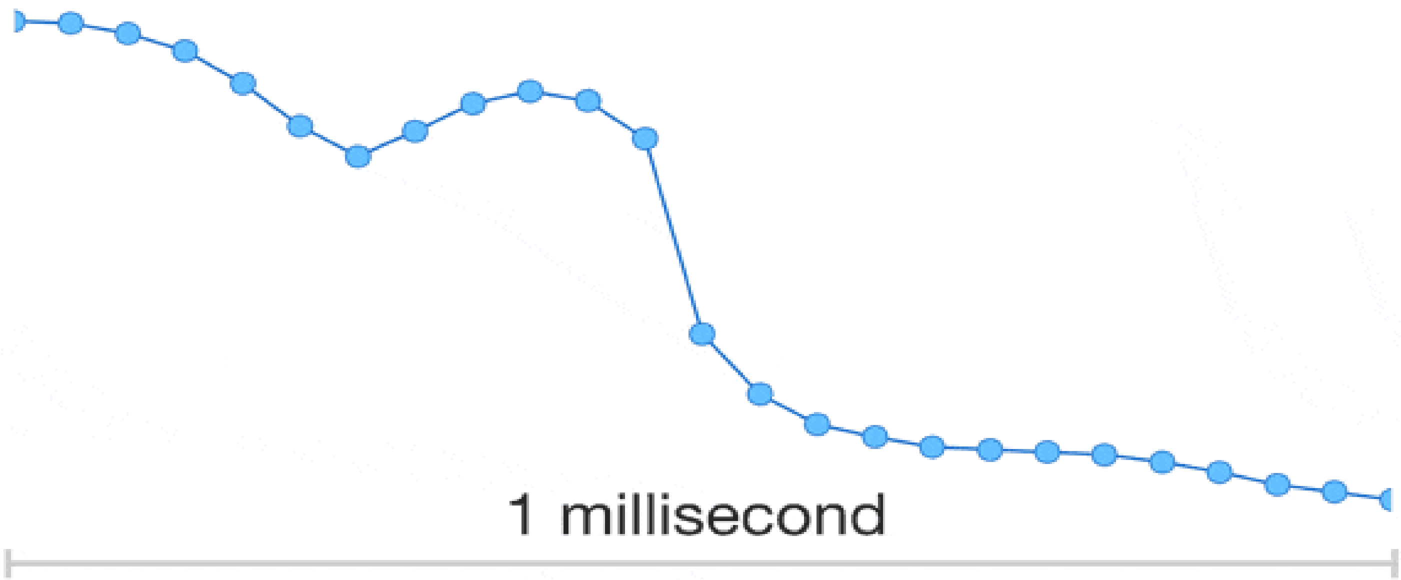
Text and Speech as Vector Sequence

Speech

Source of image: https://deepmind.com/blog/article/wavenet-generative-model-raw-audio



Speech



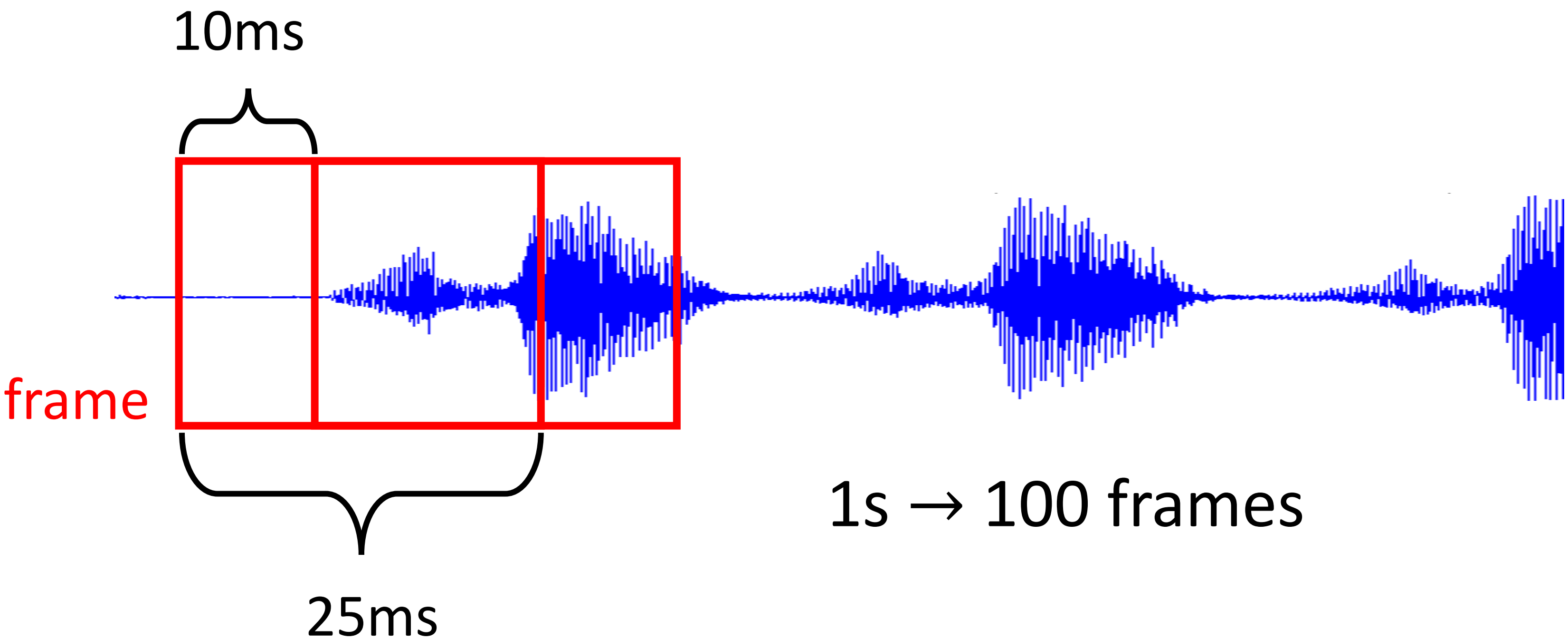
Speech is a sequence of numbers.

Represented as a very long vector sequence, each vector only has one dimension.

The sequence will be very long.

1 second has 16K sample points

Acoustic Feature



Acoustic feature

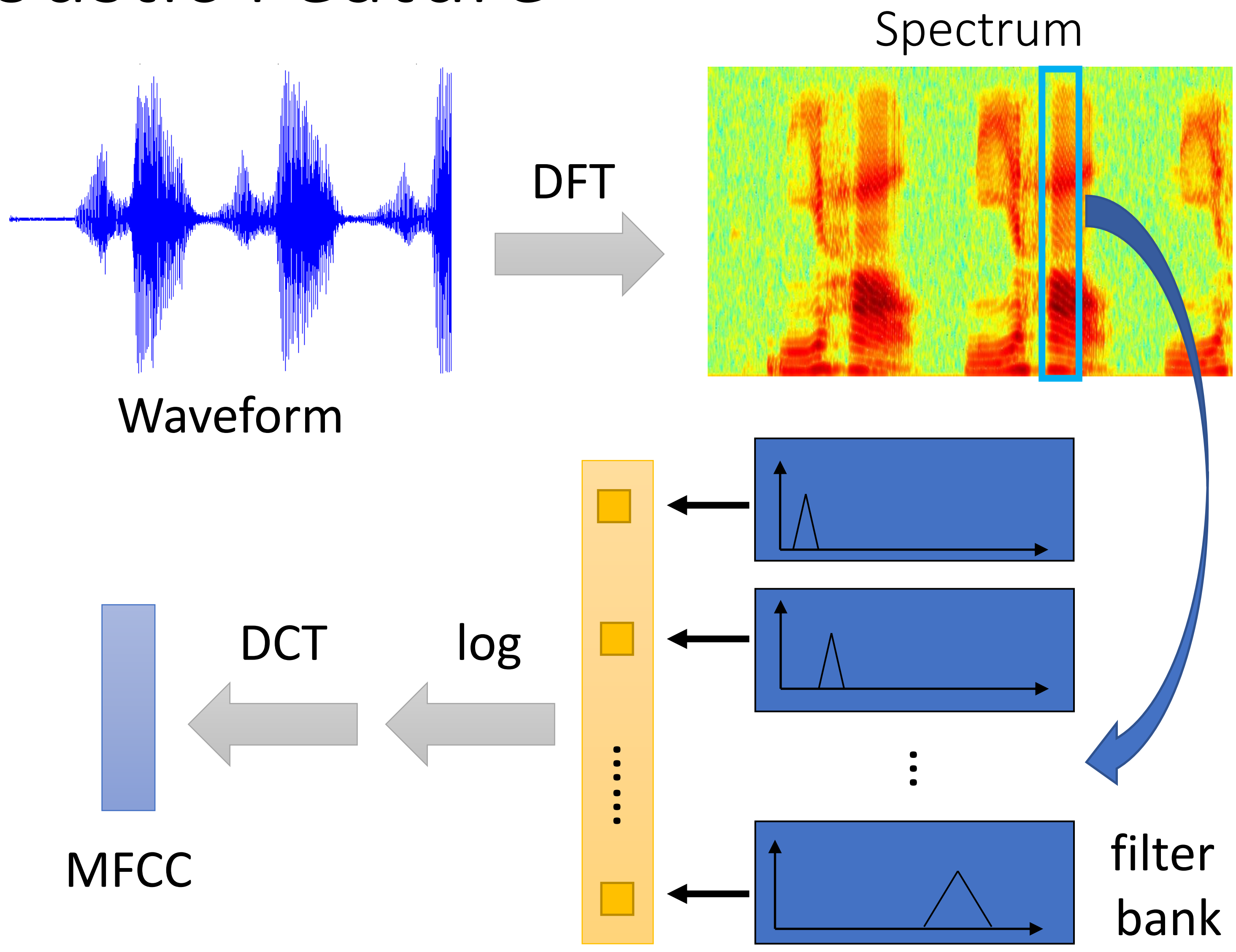
400 sample points (16KHz)
39-dim MFCC
80-dim filter bank output

Hand-crafted Features

https://librosa.org/doc/latest/index.html

https://pytorch.org/audio/stable/index.html

Acoustic Feature



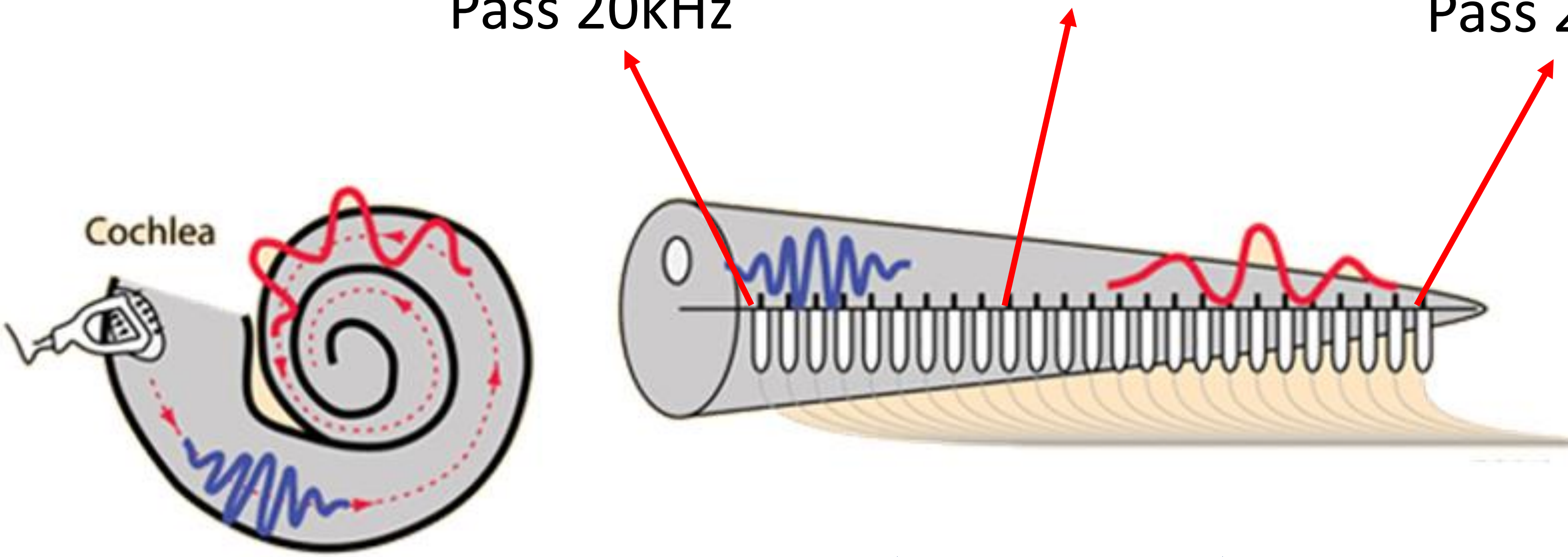
Auditory System

Cochlea

Pass 20kHz

Each neuron only passes a specific frequency.

Pass 20Hz



to brain

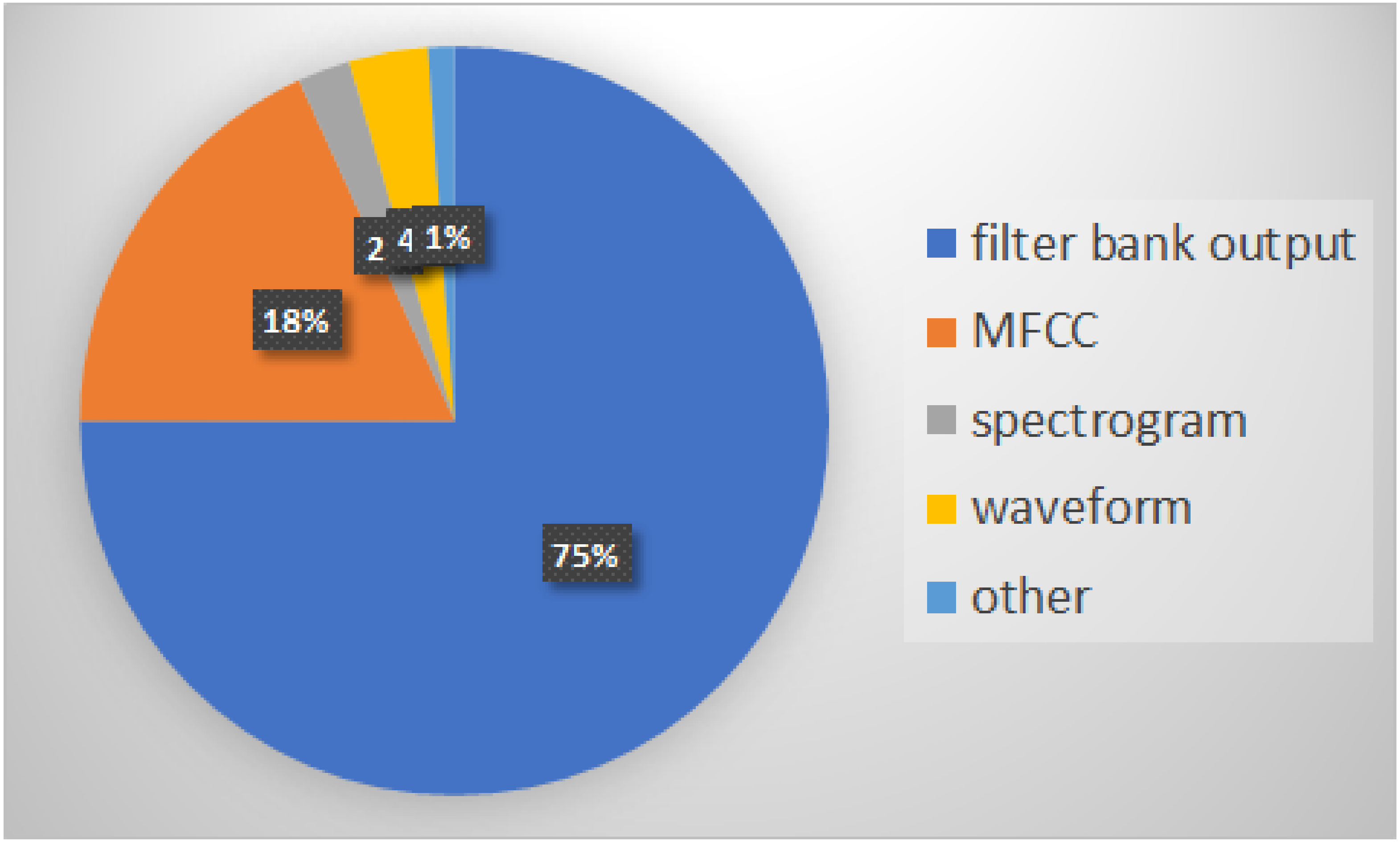
Pass high frequency

Pass low frequency

Acoustic Feature

Go through more than 100 papers in INTERSPEECH’19, ICASSP’19, ASRU’19

(Thanks to the TAs of DLHLP 2020)



How much data do we need?

(English corpora)

MNIST: 28 x 28 x 1 x 60000

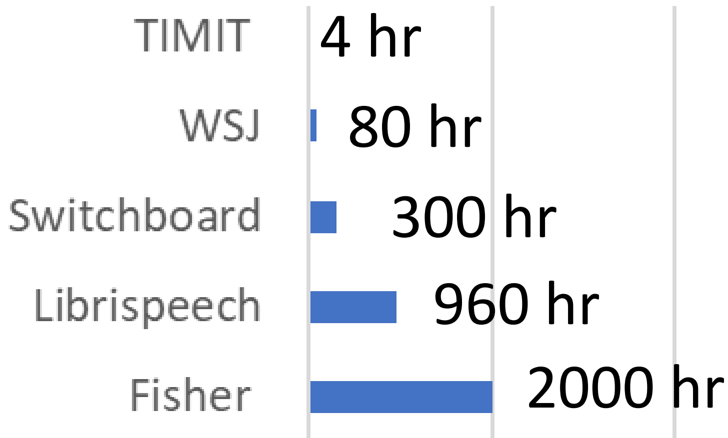
= 47,040,000

= 49 minutes (16kHz)

CIFAR-10: 32 x 32 x 3 x 50000

= 153,600,000

= 2 hours 40 minutes

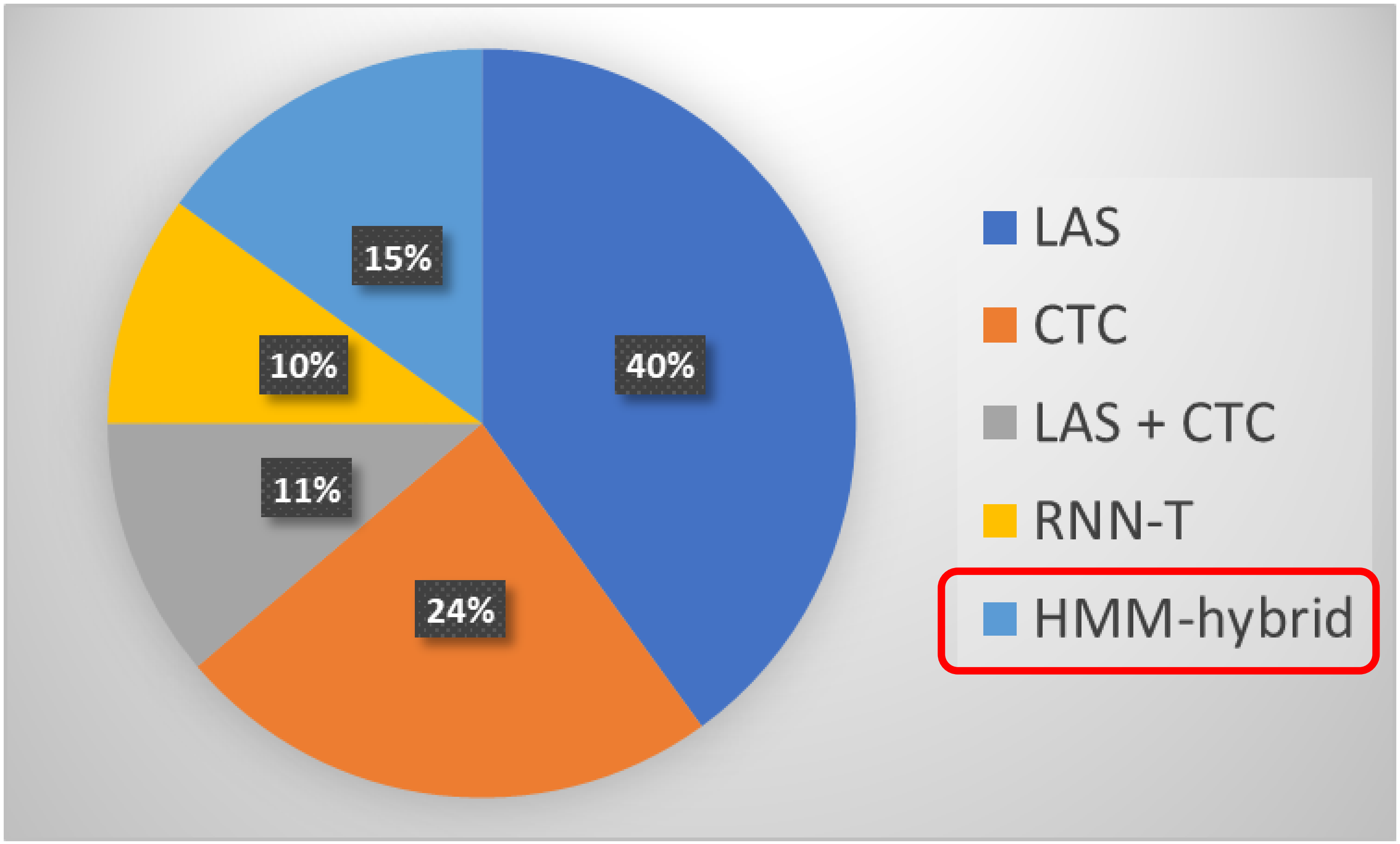


The commercial systems use more than that ……

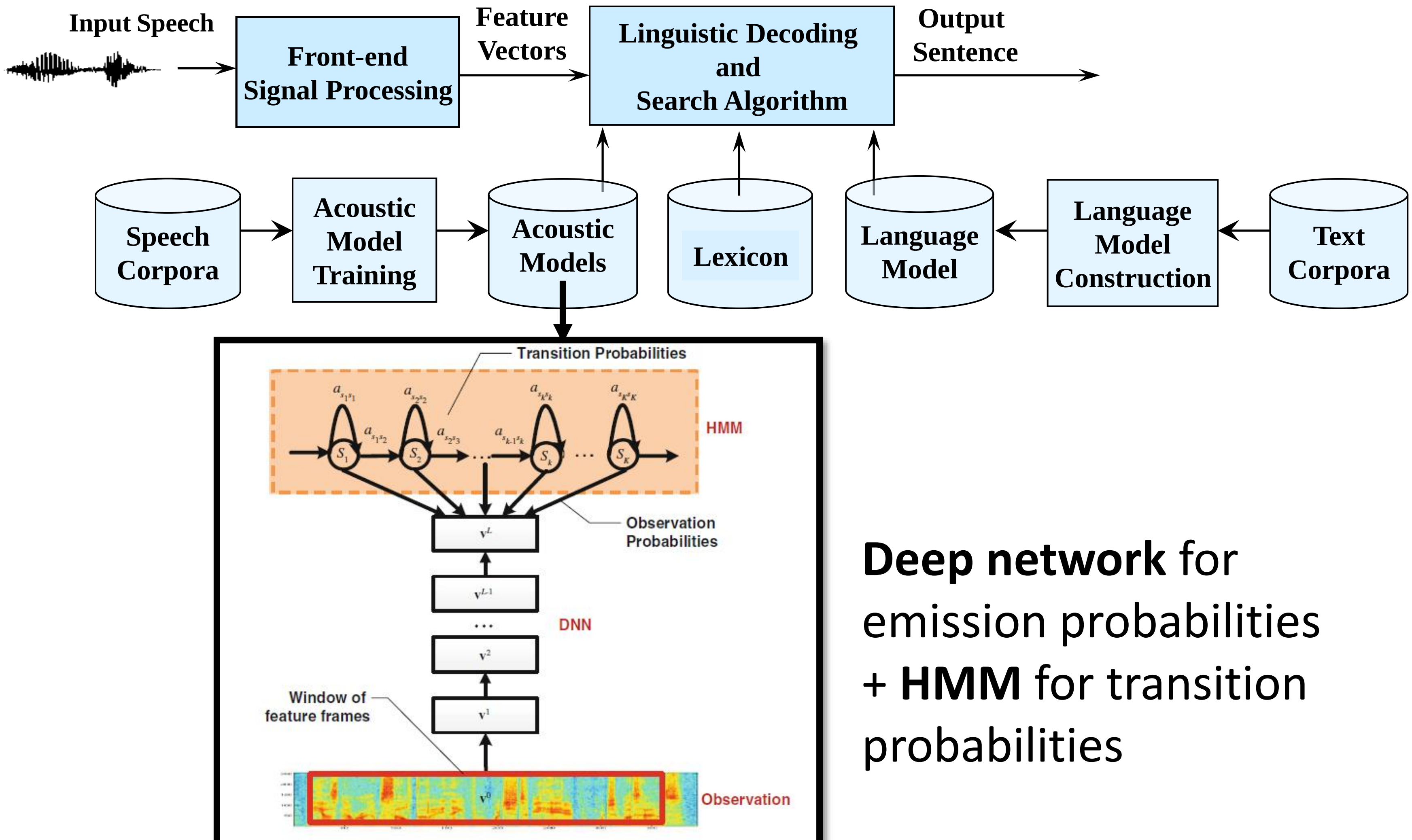
Models

Go through more than 100 papers in INTERSPEECH'19, ICASSP'19, ASRU'19

(Thanks to the TAs of DLHLP 2020)



Hybrid model

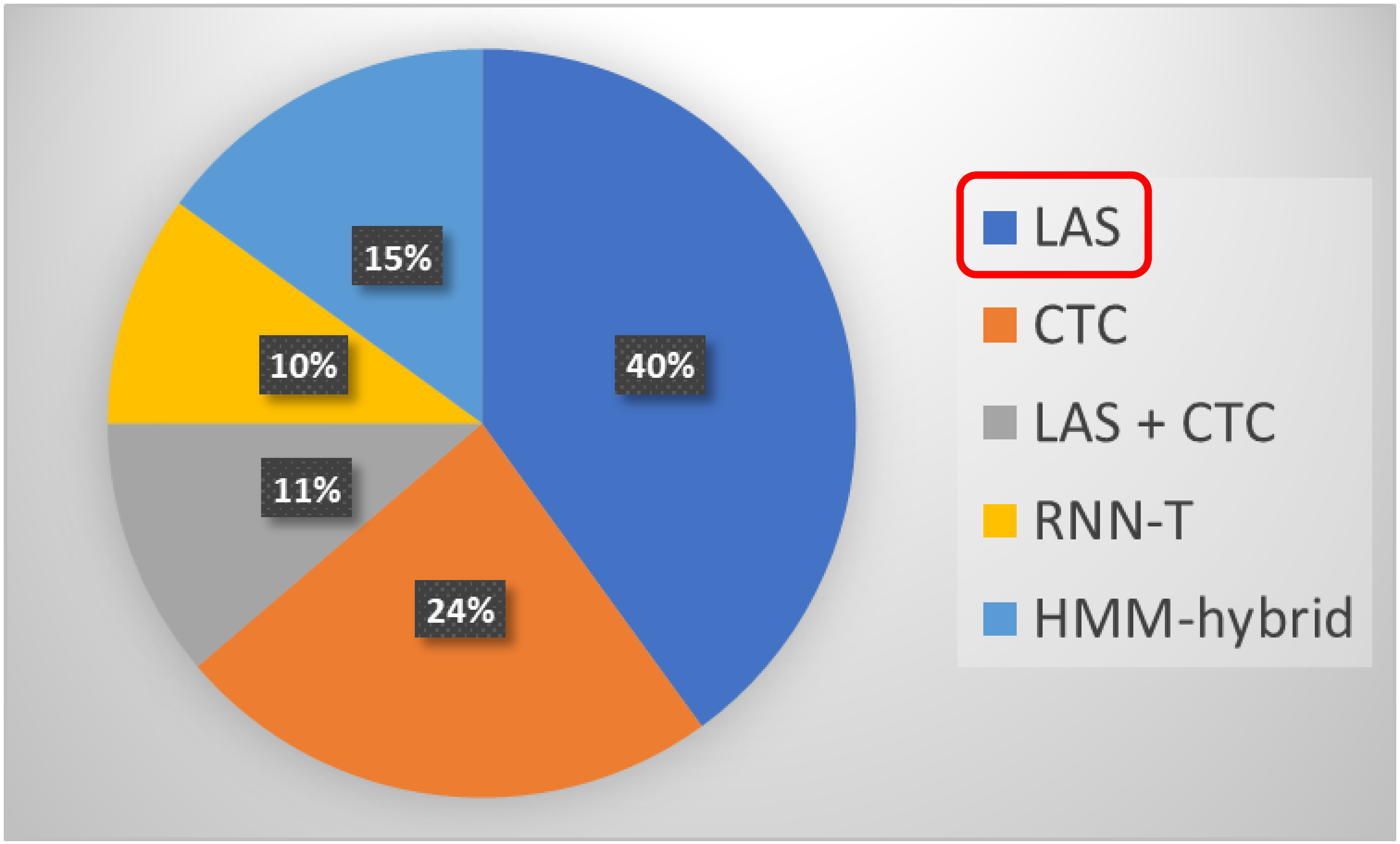


Deep network for emission probabilities + **HMM** for transition probabilities

Models

Go through more than 100 papers in INTERSPEECH'19, ICASSP'19, ASRU'19

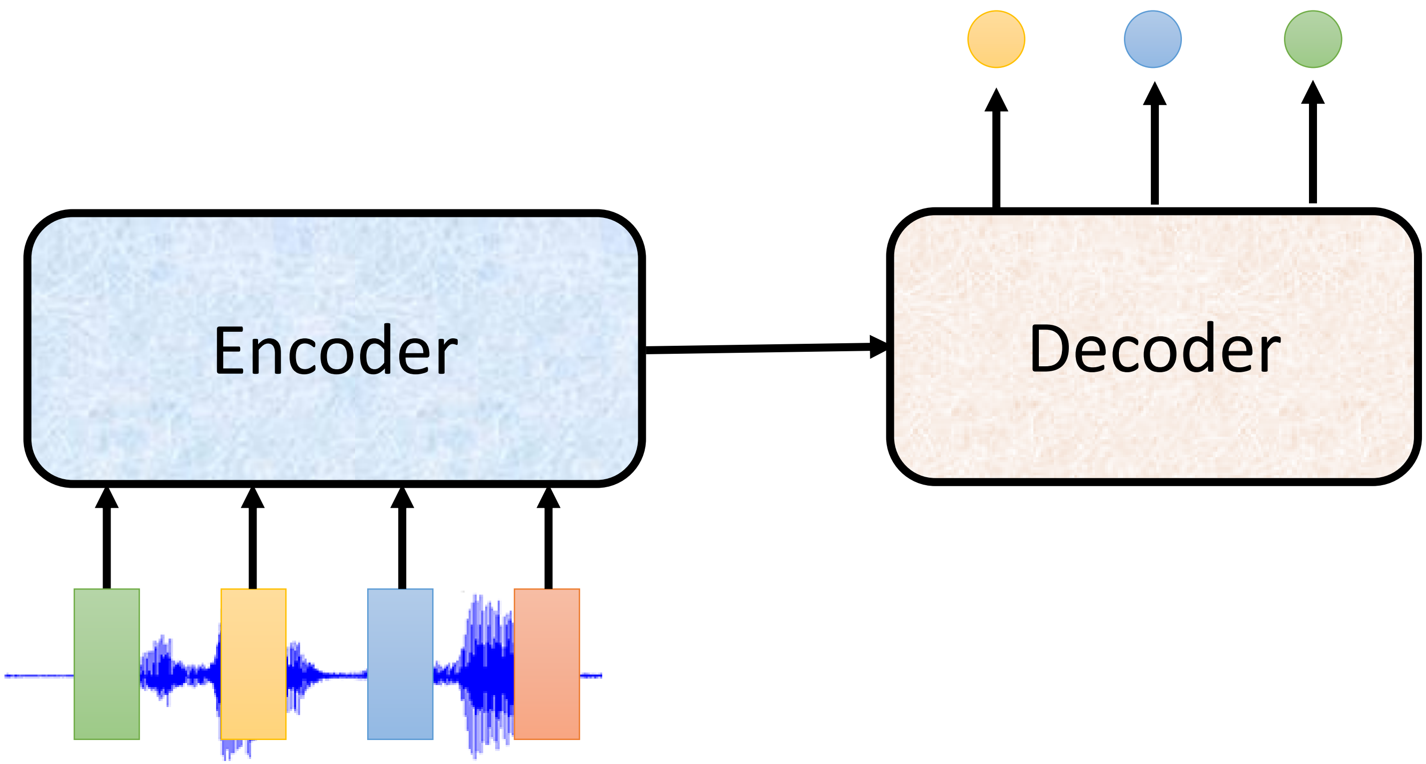
(Thanks to the TAs of DLHLP 2020)



Listen, Attend, and Spell (LAS)

- It is the typical seq2seq with attention.

[Chorowski. et al., NIPS'15]



LAS – Does it work?

Model	Dev	Test
Baseline Model	15.9%	18.7%
Baseline + Conv. Features	16.1%	18.0%
Baseline + Conv. Features + Smooth Focus	15.8%	**17.6%**
RNN Transducer [16]	N/A	17.7%
HMM over Time and Frequency Convolutional Net [25]	13.9%	16.7%

TIMIT

[Chorowski. Et al., NIPS'15]

Step	Splicing	Space	CHM	SWB	Avg
1	±5	feature	62.7	47.6	55.2
2	±5	feature	61.3	40.8	51.1
3	±5	feature	59.9	**38.8**	**49.4**
4	±5	feature	60.2	41.7	51.0
1	±7	feature	65.5	47.6	56.6
2	±7	feature	59.9	41.7	50.9
3	±7	feature	59.8	40.3	50.1
4	±7	feature	60.0	43.0	51.6
2	±5	hidden	60.7	42.3	51.5
3	±5	hidden	**58.9**	41.7	50.3

10.4% on SWB …

[Soltau, et al., ICASSP'14]

300 hours

[Lu, et al., INTERSPEECH'15]

LAS – Yes, it works!

Model	Clean WER	Noisy WER
CLDNN-HMM [22]	8.0	8.9
LAS	14.1	16.5
LAS + LM Rescoring	10.3	12.0

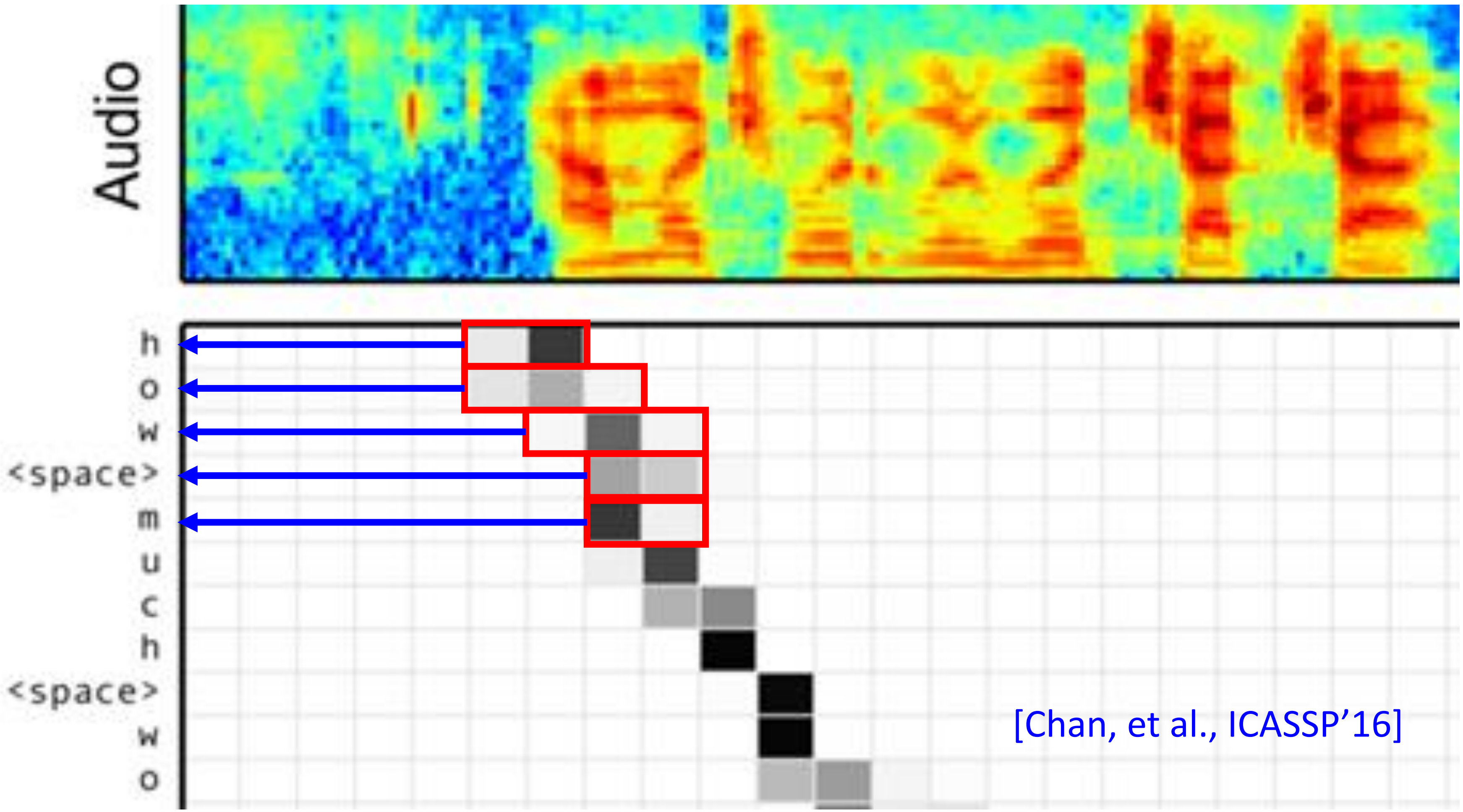
2000 hours

[Chan, et al., ICASSP'16]

Exp-ID	Model	VS/D	1st pass Model Size
E8	Proposed	**5.6/4.1**	**0.4 GB**
E9	Conventional LFR system	6.7/5.0	0.1 GB (AM) + 2.2 GB (PM) + 4.9 GB (LM) = 7.2GB

12500 hours

[Chiu, et al., ICASSP, 2018]



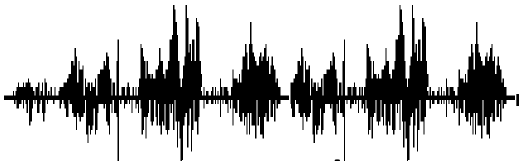
Beam	Text	Log Probability	WER
Truth	call aaa roadside assistance	-	-
1	call aaa roadside assistance	-0.5740	0.00
2	call triple a roadside assistance	-1.5399	50.00
3	call trip way roadside assistance	-3.5012	50.00
4	call xxx roadside assistance	-4.4375	25.00

[Chan, et al., ICASSP'16]

More than Speech Recognition …

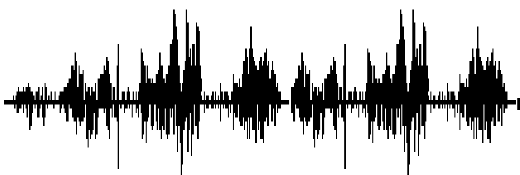
- Only 56% languages have written form (Ethnologue, 21st edition)
- The existing writing systems may not be widely used.

More than Speech Recognition …



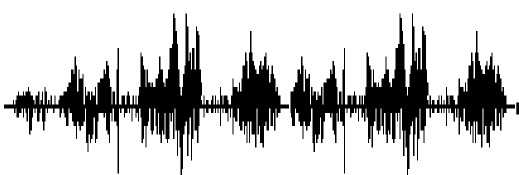
one punch man → Speech Recognition → 一 拳 超 人

+ Translation



one ticket to Taipei on March 2nd → Speech Recognition → <buy ticket>

+ Intent classification



one ticket to Taipei on March 2nd → Speech Recognition → Location: Taipei, Time: March 2nd

+ Slot filling

Comparison

Hybrid Model

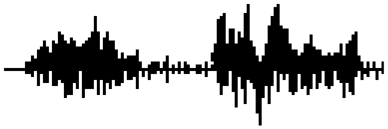
- Less data ☺
- Easy to add new token ☺ (modify lexicon)
- Easy for teamwork ☺
- Larger model ☹
- Relative difficult to implement ☹
- Commercial system

End-to-end

- More data required ☹
- How to add new token? ☹
- There is only one model …… ☹
- Smaller model ☺
- Easy to implement ☺
- Usually for research

Limitation of LAS

- LAS outputs the first token after listening the whole input.
- Users expect on-line speech recognition.



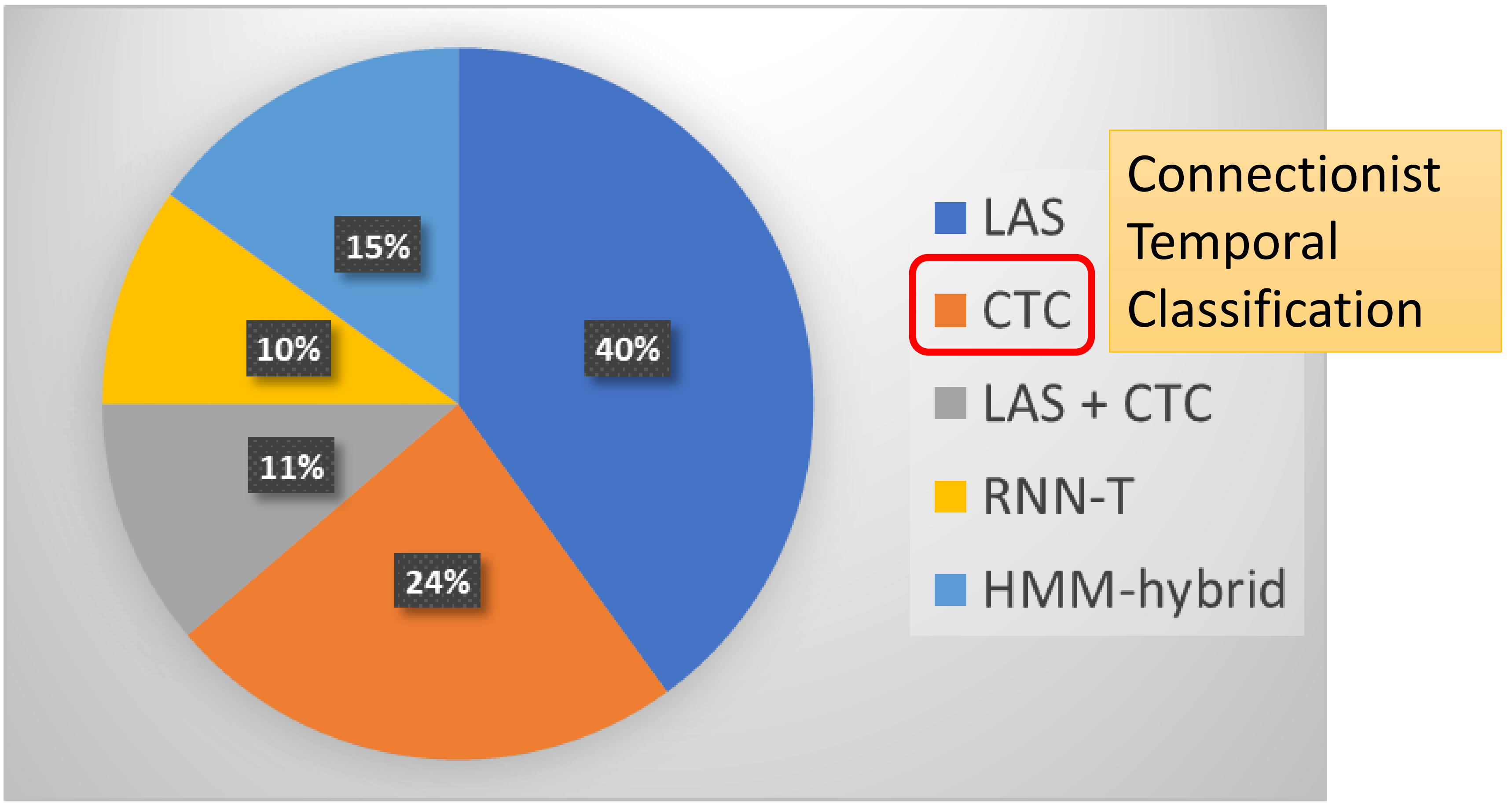
今 天 的 天 氣 非 常 好

LAS is not the final solution of speech recognition!

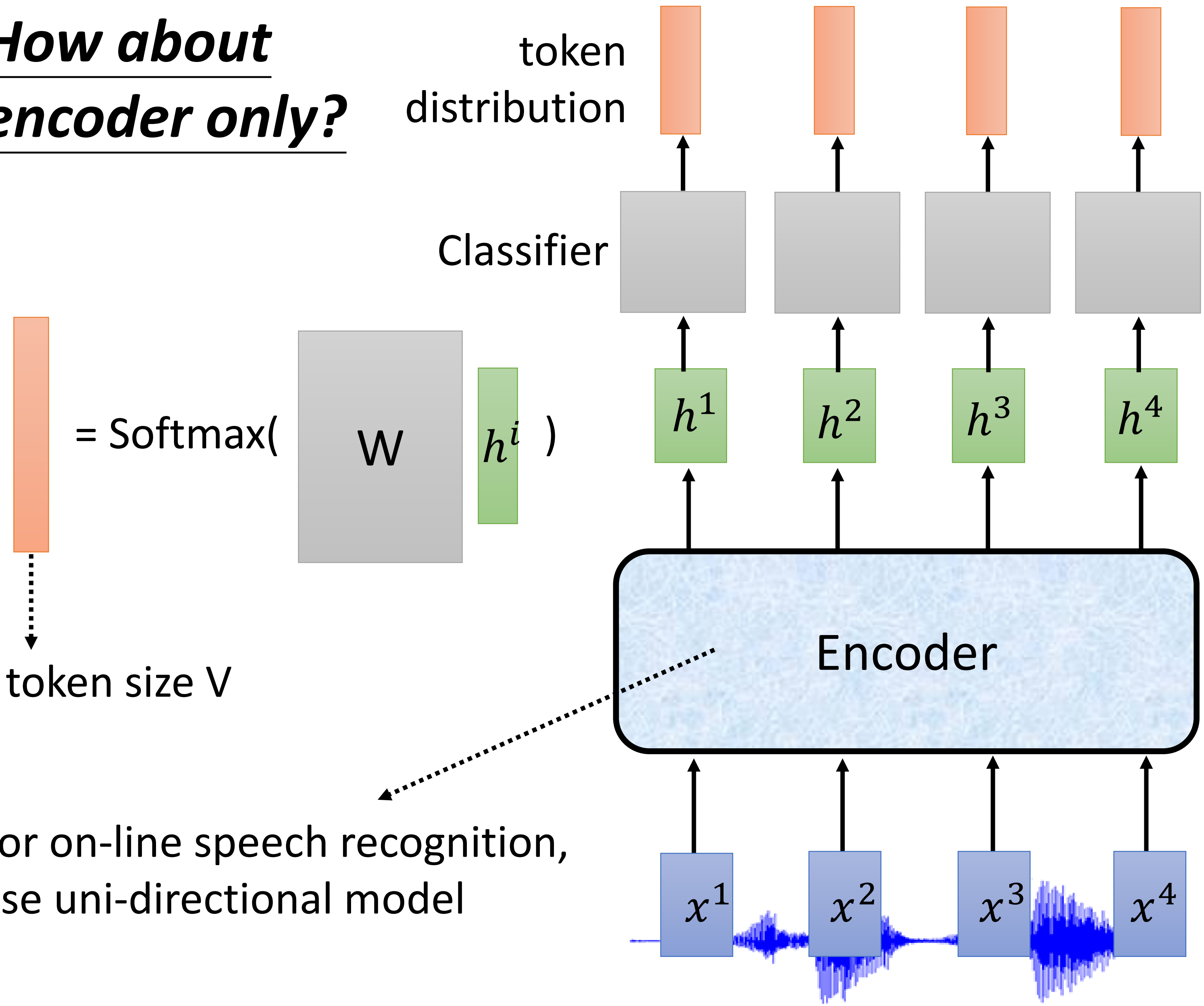
Models

Go through more than 100 papers in INTERSPEECH’19, ICASSP’19, ASRU’19

(Thanks to the TAs of DLHLP 2020)



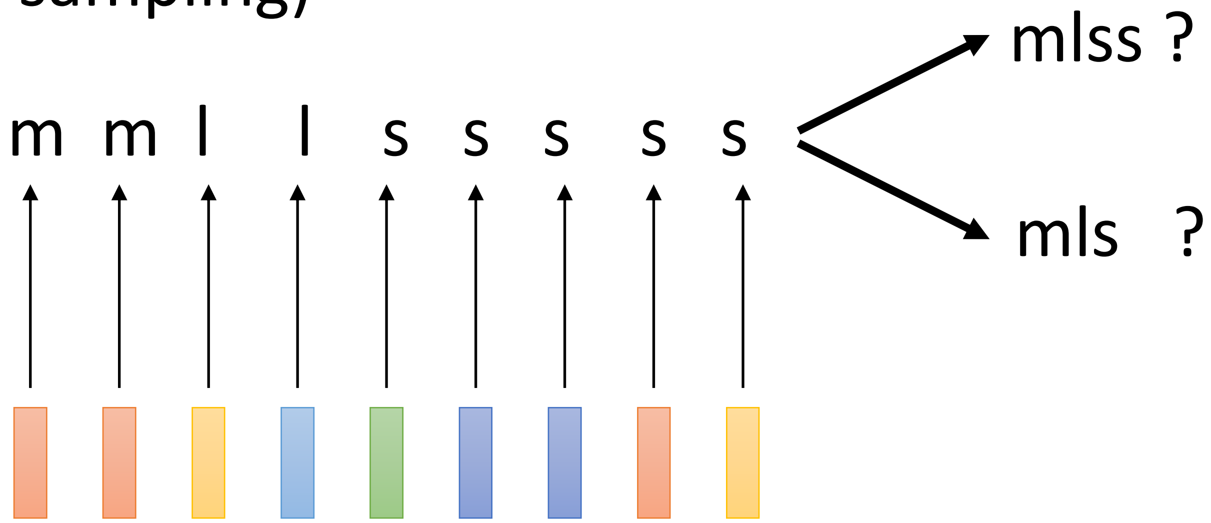
How about encoder only?



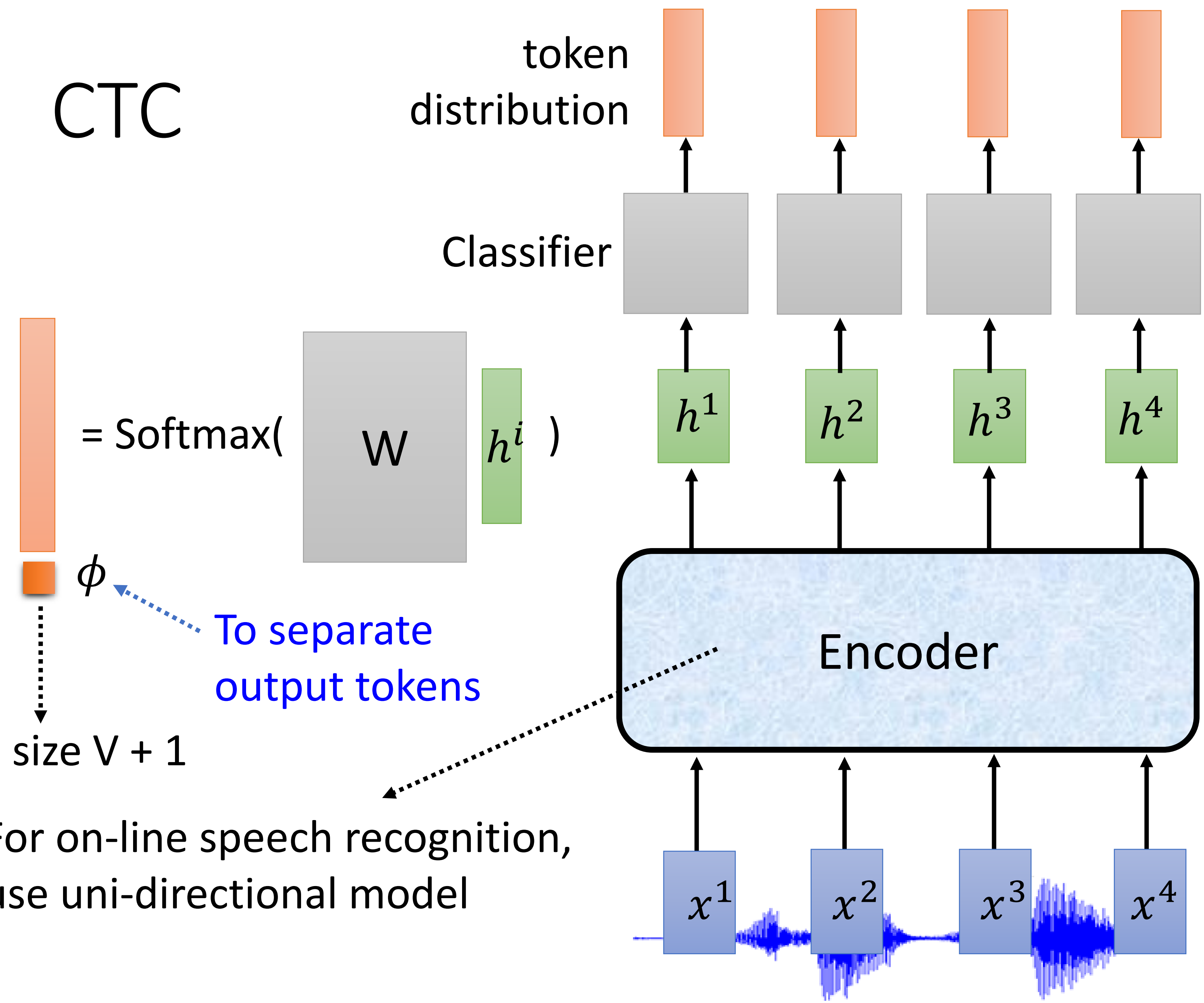
For on-line speech recognition, use uni-directional model

How about Encoder Only?

- Input T acoustic features, output T tokens (ignoring down sampling)



CTC



For on-line speech recognition, use uni-directional model

CTC

- Input T acoustic features, output T tokens (ignoring down sampling)

m m l l s s s s s → mlss

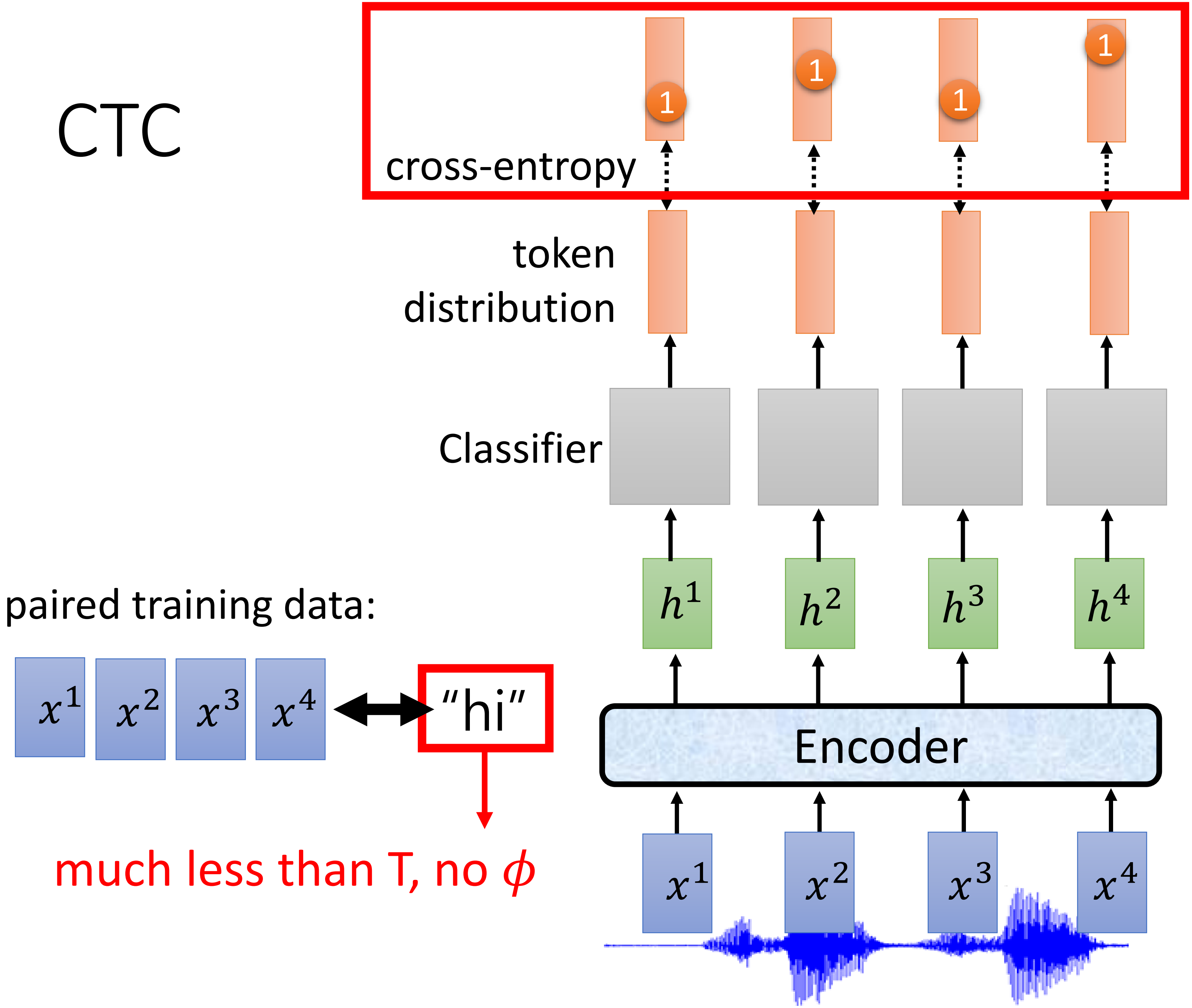
m m l l s s s s s → mls

- Output tokens including ϕ, merging duplicate tokens, removing ϕ

m m ϕ l l ϕ ϕ s s ⟶ m l s

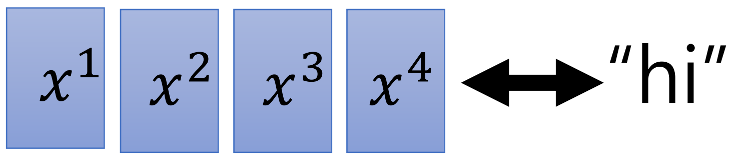
m m ϕ l ϕ s ϕ ϕ s ⟶ m l s s

CTC


cross-entropy
1
1
1
1
token
distribution
Classifier
h^1
h^2
h^3
h^4
Encoder
x^1
x^2
x^3
x^4
paired training data:
x^1
x^2
x^3
x^4
"hi"
much less than T, no ϕ

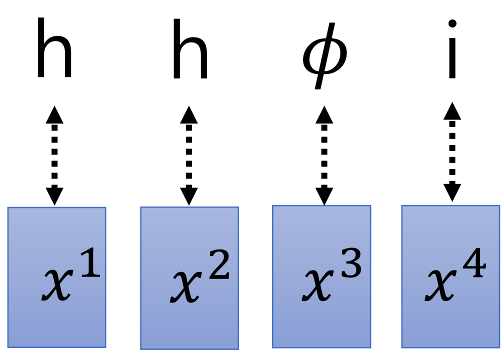
CTC – Training

paired training data:

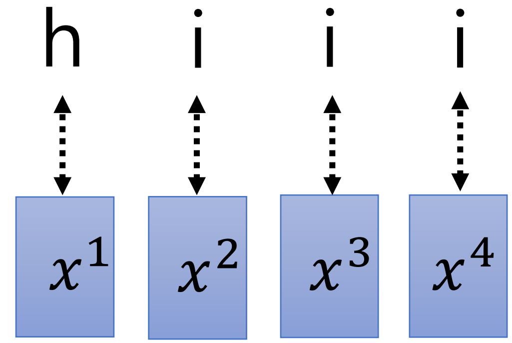


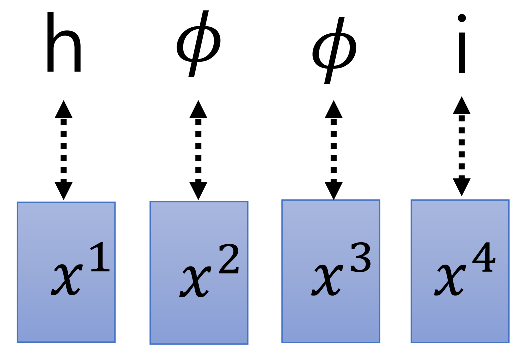
All of them are used in training! (How?!)

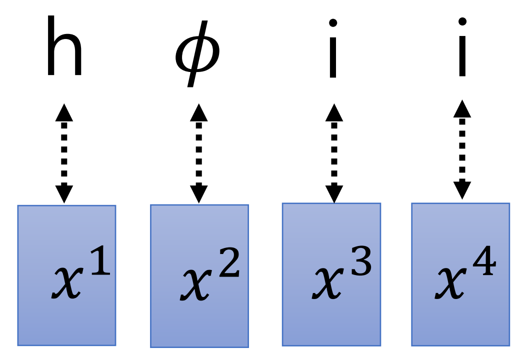
[Graves, et al., ICML'14]



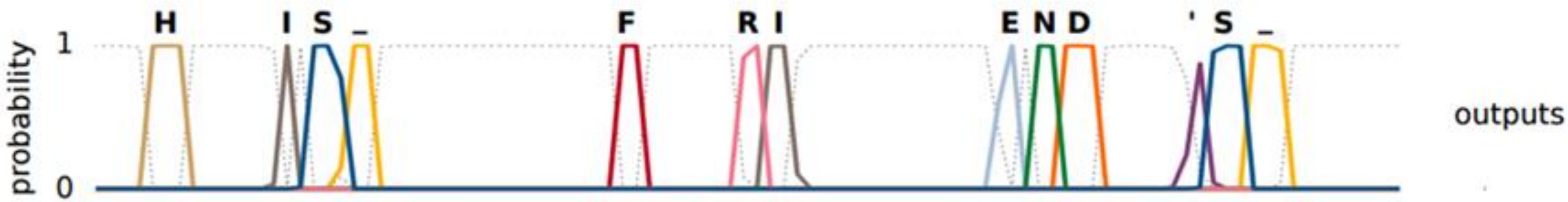
alignment



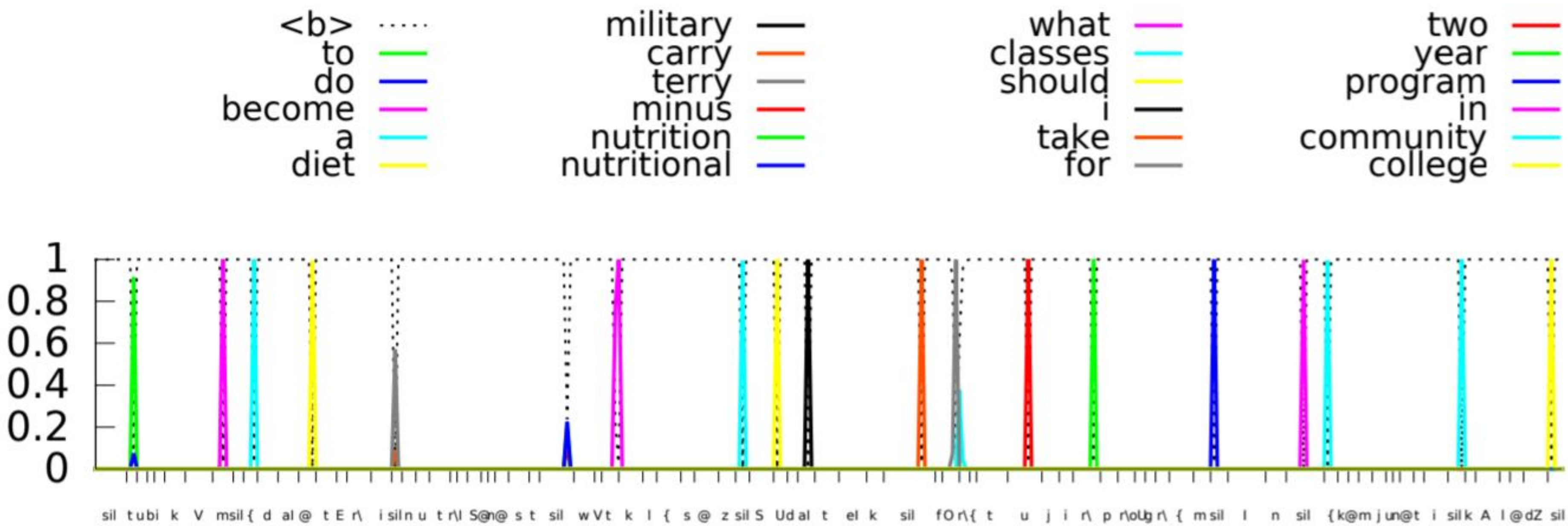




Does CTC work?



[Graves, et al., ICML'14]



[Sak, et al., INTERSPEECH'15]

Does CTC work?

Model	CER	WER
Encoder-Decoder	6.4	18.6
Encoder-Decoder + bigram LM	5.3	11.7
Encoder-Decoder + trigram LM	4.8	10.8
Encoder-Decoder + extended trigram LM	3.9	9.3
Graves and Jaitly (2014)		
CTC	9.2	30.1
CTC, expected transcription loss	8.4	27.3
Hannun et al. (2014)		
CTC	10.0	35.8
CTC + bigram LM	5.7	14.1
Miao et al. (2015),		
CTC for phonemes + lexicon	-	26.9
CTC for phonemes + trigram LM	-	7.3
CTC + trigram LM	-	9.0

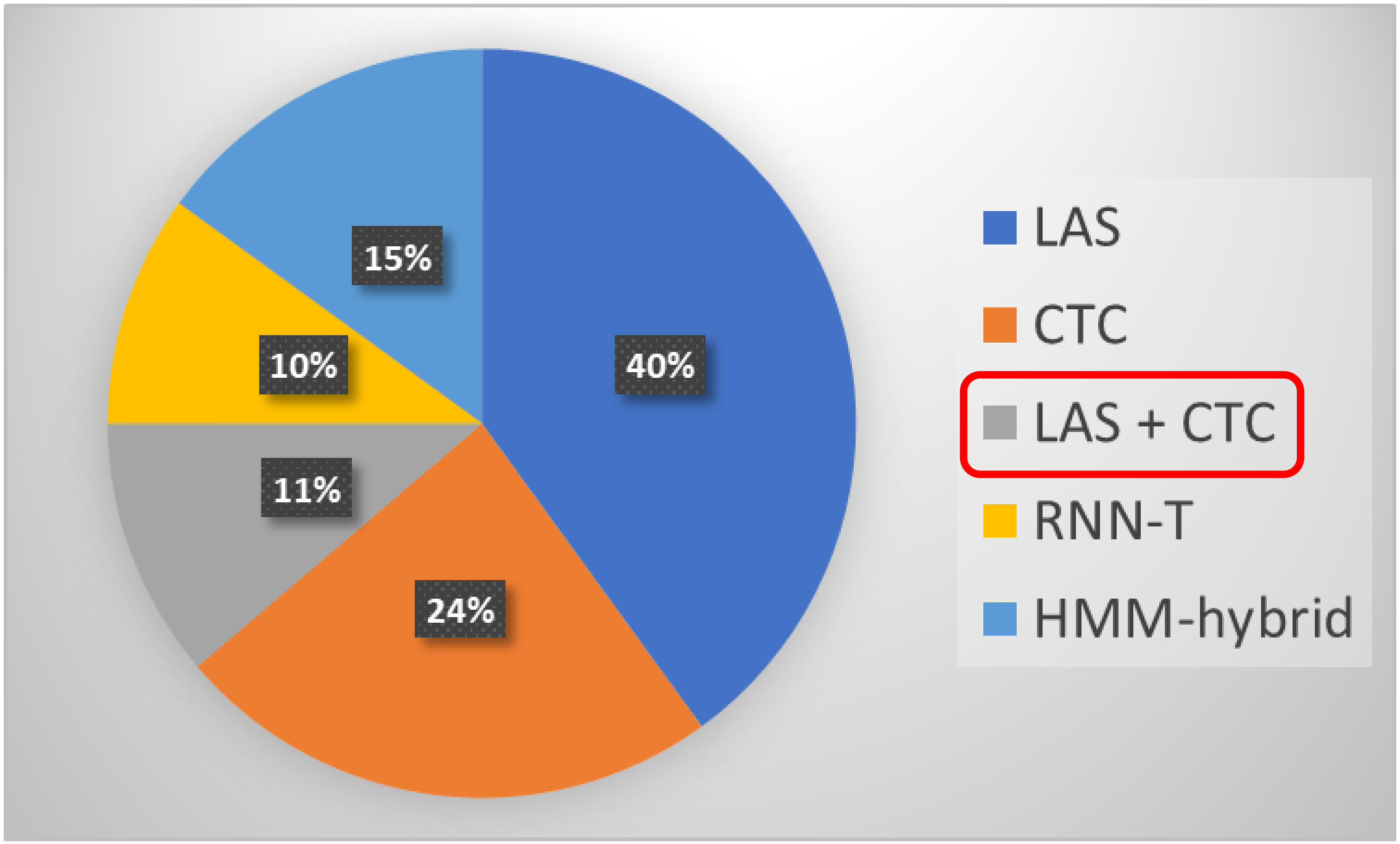
80 hours

[Bahdanau. et al., ICASSP'16]

Models

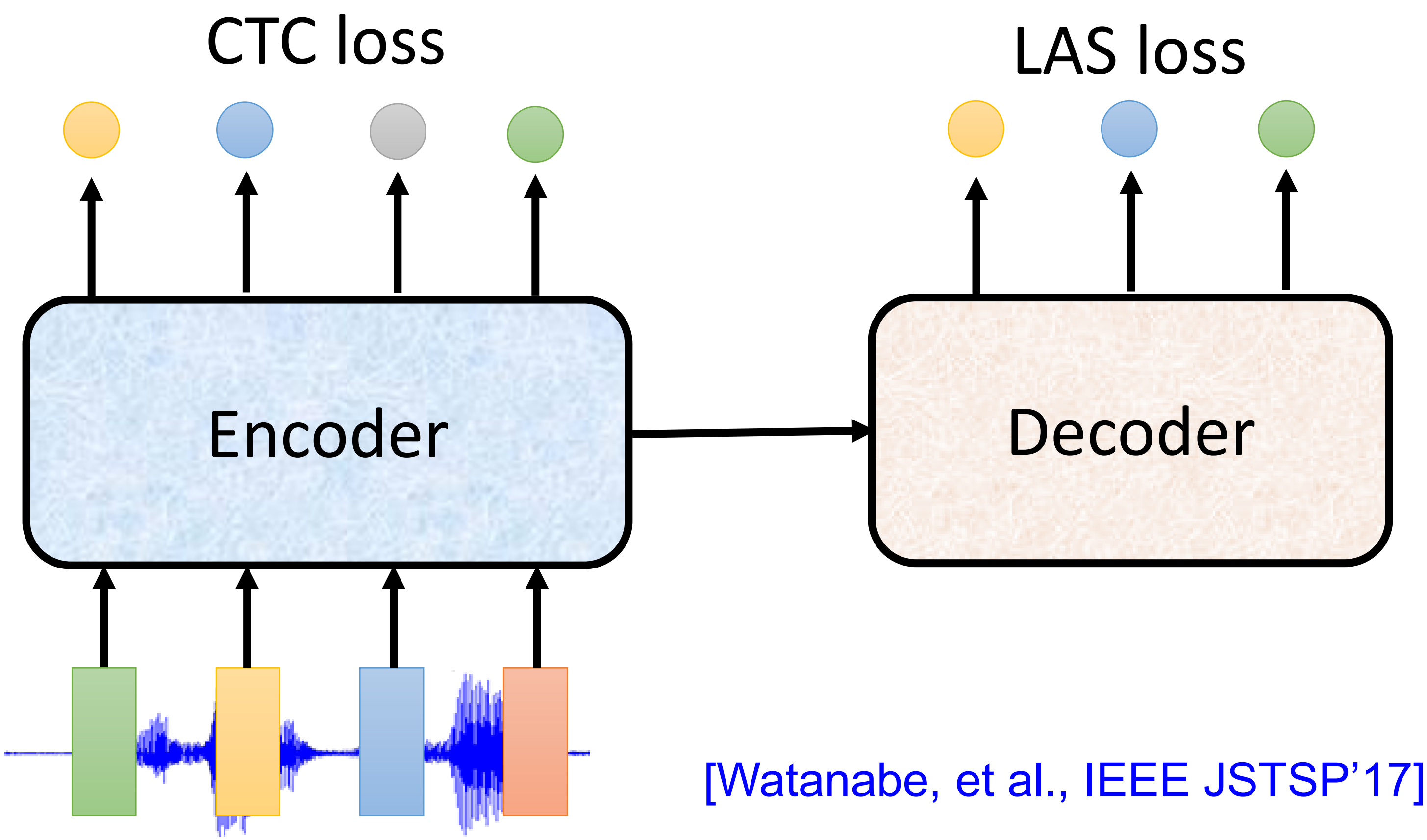
Go through more than 100 papers in INTERSPEECH'19, ICASSP'19, ASRU'19

(Thanks to the TAs of DLHLP 2020)



LAS + CTC

Make convergence faster

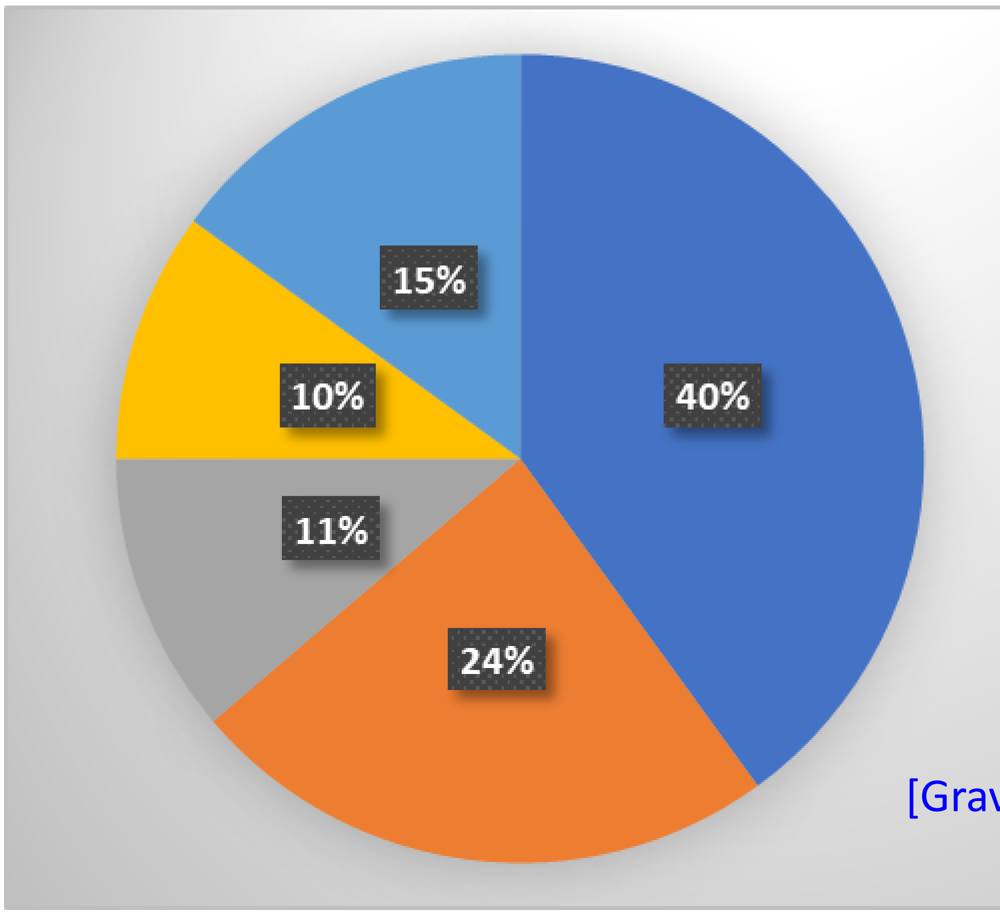


[Watanabe, et al., IEEE JSTSP'17]

Models

Go through more than 100 papers in INTERSPEECH’19, ICASSP’19, ASRU’19

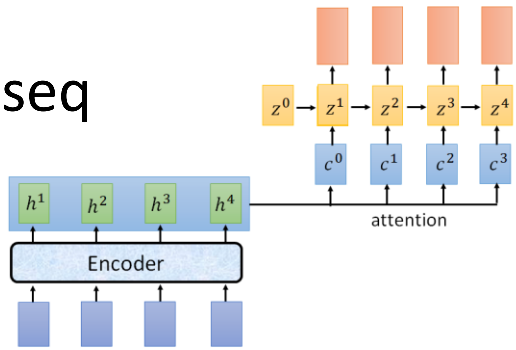
(Tha



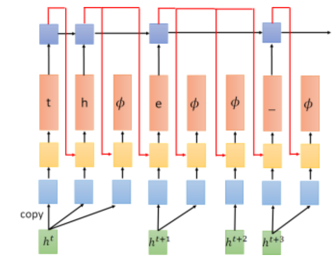
[Graves, ICML workshop’12]

More models …

LAS: aka seq2seq

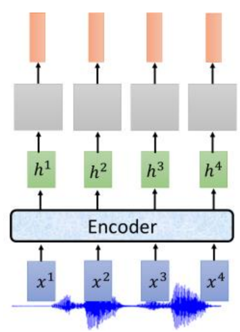


RNN-T: input one vector, output multiple tokens



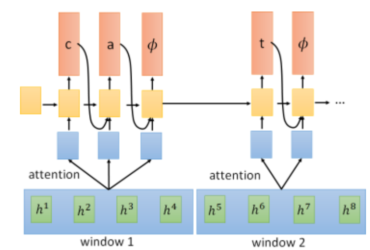
CTC: input one vector, output one token

- decoder is a linear classifier



Neural Transducer: RNN-T that takes a small segment as input

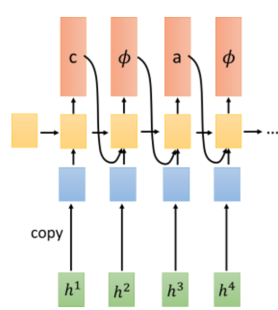
[Jaitly, et al., NIPS'16]



RNA: input one vector, output one token

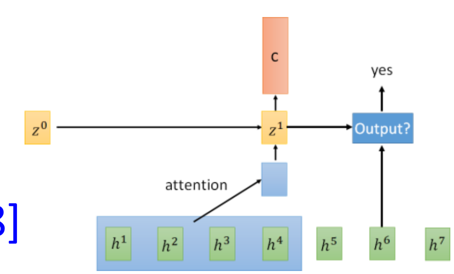
- decoder is an RNN

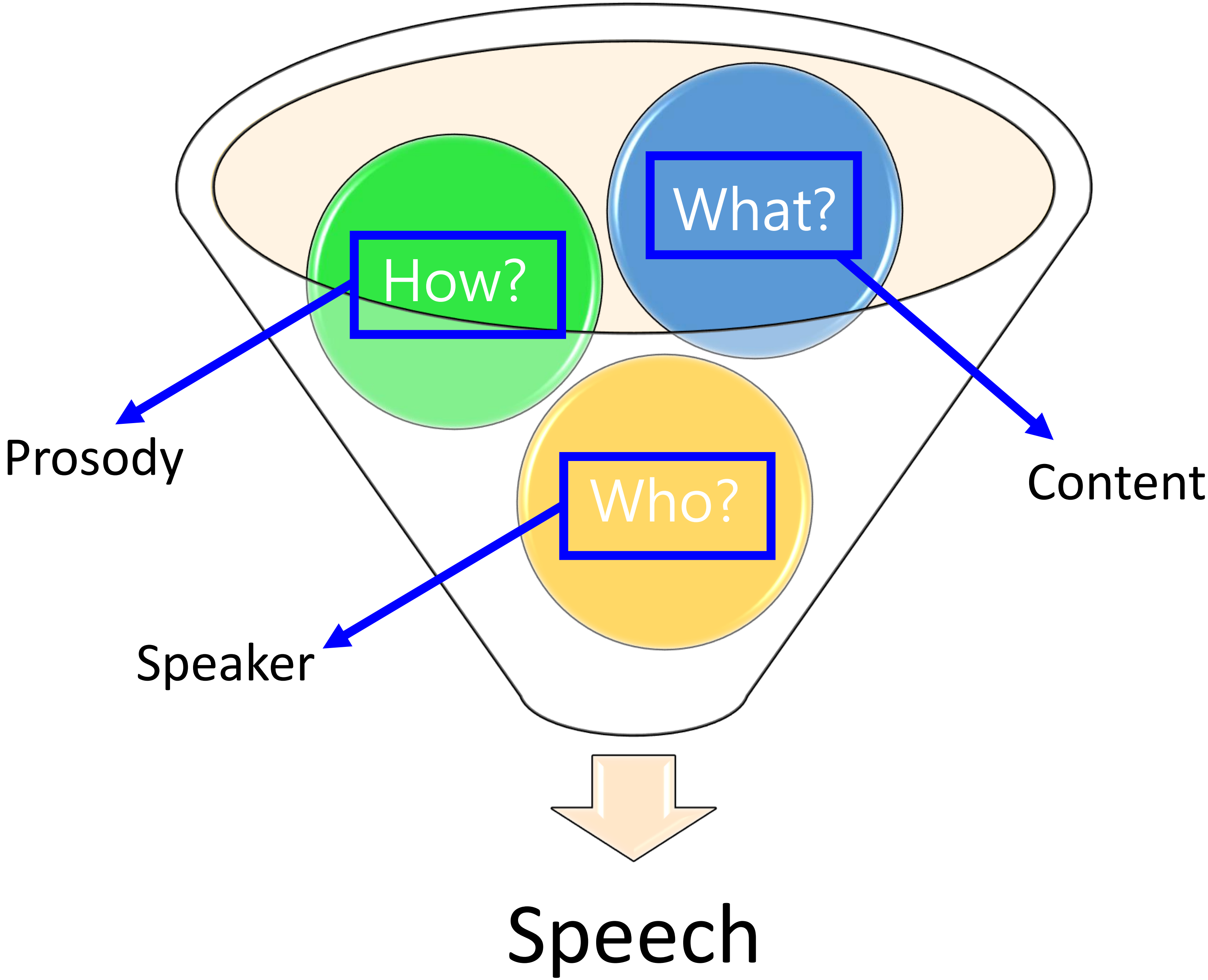
[Sak, et al., INTERSPEECH'17]



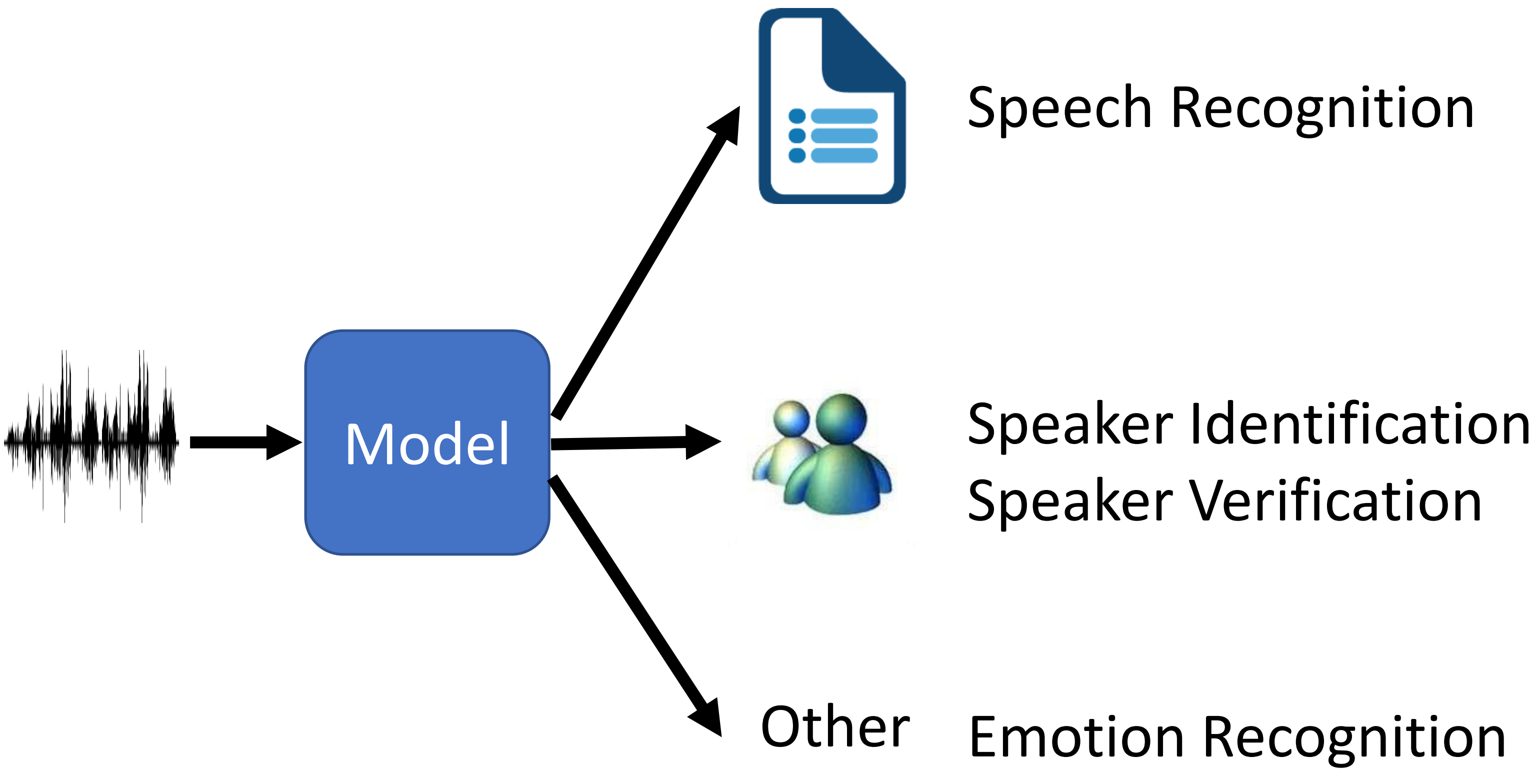
MoCha: Neural Transducer decides the sizes of small segments

[Chiu, et al., ICLR'18]




How?
What?
Who?
Prosody
Content
Speaker
Speech

More than Speech Recognition



Week 3

Week 3					
Date	2021/8/16	2021/8/17	2021/8/18	2021/8/19	2021/8/20
Weekday	Mon	Tue	Wed	Thur	Fri
09:00-09:30 (GMT+8)		Poster Session 3 Poster		Speaker: Shou De Lin Title: Machine Learning for Dynamic Environment Lecture Info Course Link	
09:30-11:00 (GMT+8)			Panel Discussion Panelists: * Cho-Jui Hsieh Bio * Pin-Yu Chen Bio * Soheil Feizi Bio * Sijia Liu Bio Title: Trustworthy Machine Learning: Challenges and Opportunities Course Link		
11:00-12:00 (GMT+8)					
12:00-20:00 (GMT+8)	Break				
20:00-20:45 (GMT+8)		Speaker: Karteek Alahari Title: Continual Visual Learning Lecture Info Course Link	Poster Session 4 Poster	Speaker: Michael Bronstein Title: Geometric Deep Learning Lecture Info Course Link	
20:45-21:00 (GMT+8)					Closing Speaker: Program Committee
21:00-22:00 (GMT+8)	Speaker: Prateek Mittal Title: ML privacy Lecture Info Course Link				Panel Discussion Panelists: * Shinji Watanabe Bio * Shang-Wen Li Bio * Mirco Ravanelli Bio * Titouan Parcollet Bio Title: Self supervised learning for speech Course Link
22:00-23:00 (GMT+8)					

Speech Recognition

Learning with less supervision

There are around 7,000 languages in the world.

(Most languages do not have a large amount of paired data.)

Image: https://acutrans.com/top-10-most-commonly-spoken-languages-in-the-world/

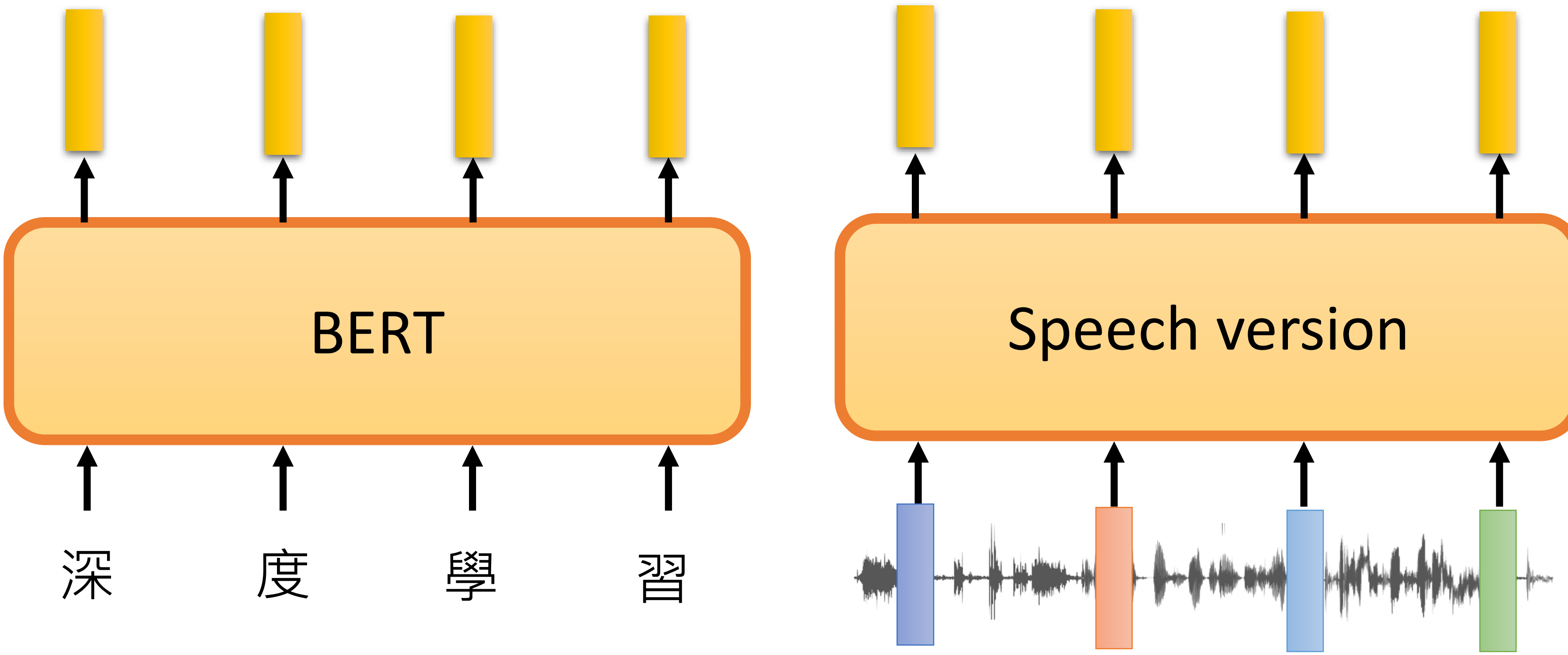
Learning with less supervision

Self-supervised Learning

Learning from Unpaired Data

Meta Learning

Speech BERT



Week 3

Week 3					
Date	2021/8/16	2021/8/17	2021/8/18	2021/8/19	2021/8/20
Weekday	Mon	Tue	Wed	Thur	Fri
09:00-09:30 (GMT+8)		**Poster Session 3** Poster		**Speaker:** Shou De Lin **Title:** Machine Learning for Dynamic Environment Lecture Info Course Link	
09:30-11:00 (GMT+8)			Panel Discussion **Panelists:** * Cho-Jui Hsieh Bio * Pin-Yu Chen Bio * Soheil Feizi Bio * Sijia Liu Bio **Title:** Trustworthy Machine Learning: Challenges and Opportunities Course Link		
11:00-12:00 (GMT+8)					
12:00-20:00 (GMT+8)	Break				
20:00-20:45 (GMT+8)		**Speaker:** Karteek Alahari **Title:** Continual Visual Learning Lecture Info Course Link	**Poster Session 4** Poster	**Speaker:** Michael Bronstein **Title:** Geometric Deep Learning Lecture Info Course Link	
20:45-21:00 (GMT+8)					Closing **Speaker:** Program Committee
21:00-22:00 (GMT+8)	**Speaker:** Prateek Mittal **Title:** ML privacy Lecture Info Course Link				Panel Discussion **Panelists:** * Shinji Watanabe Bio * Shang-Wen Li Bio * Mirco Ravanelli Bio * Titouan Parcollet Bio **Title:** Self supervised learning for speech Course Link
22:00-23:00 (GMT+8)					

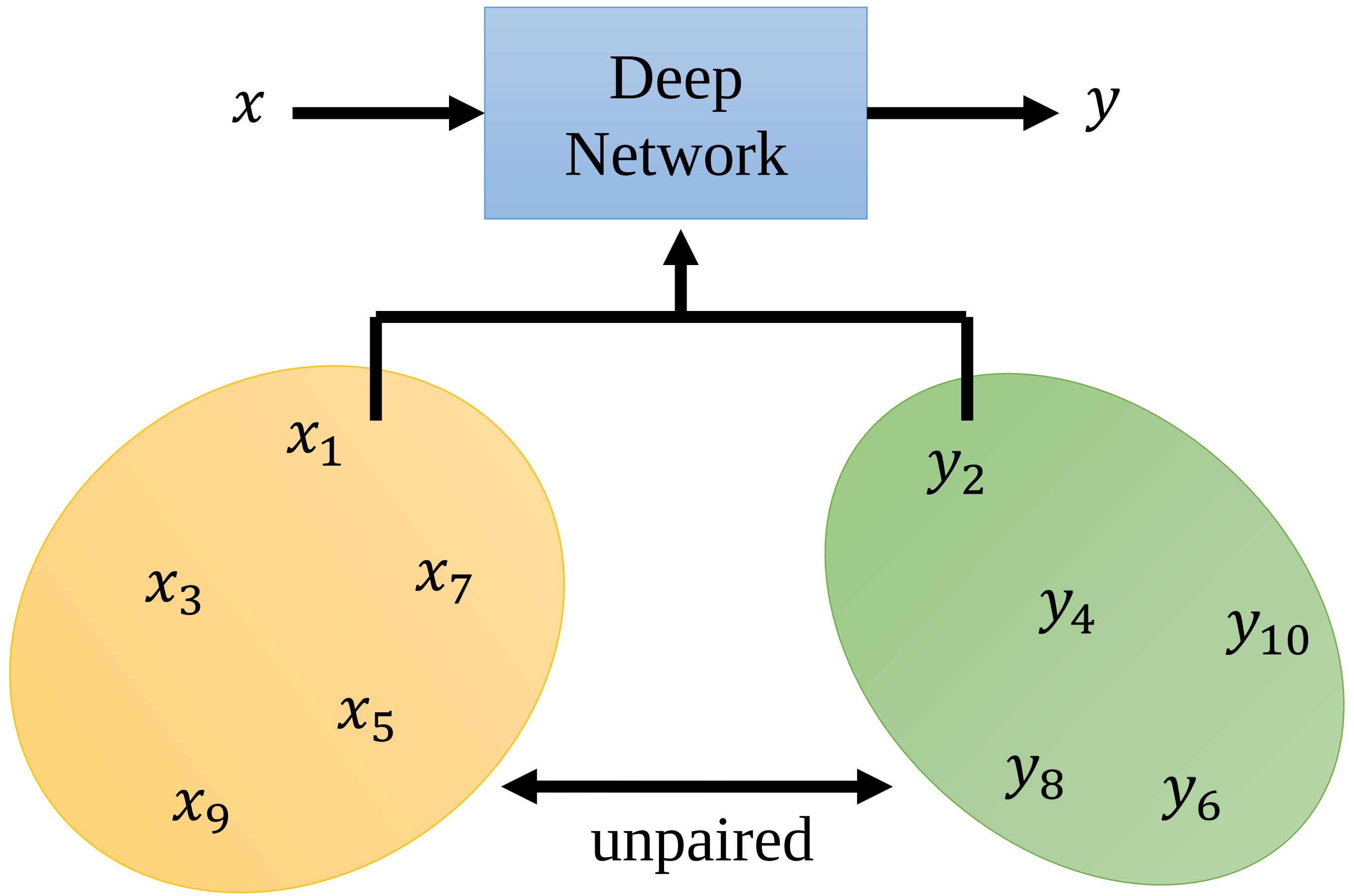
Learning with less supervision

Self-supervised Learning

Learning from Unpaired Data

Meta Learning

Learning from Unpaired Data

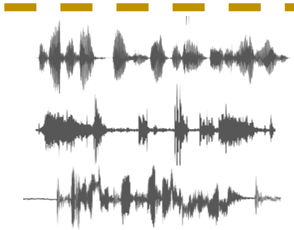


Basic Idea - GAN

How are you.
hh aw aa r y uw
He think it's…
hh ih hh ih ng kcl …
Thanks for…
th ae ng k s f ao r …

Data (Text)

Data (Audio)

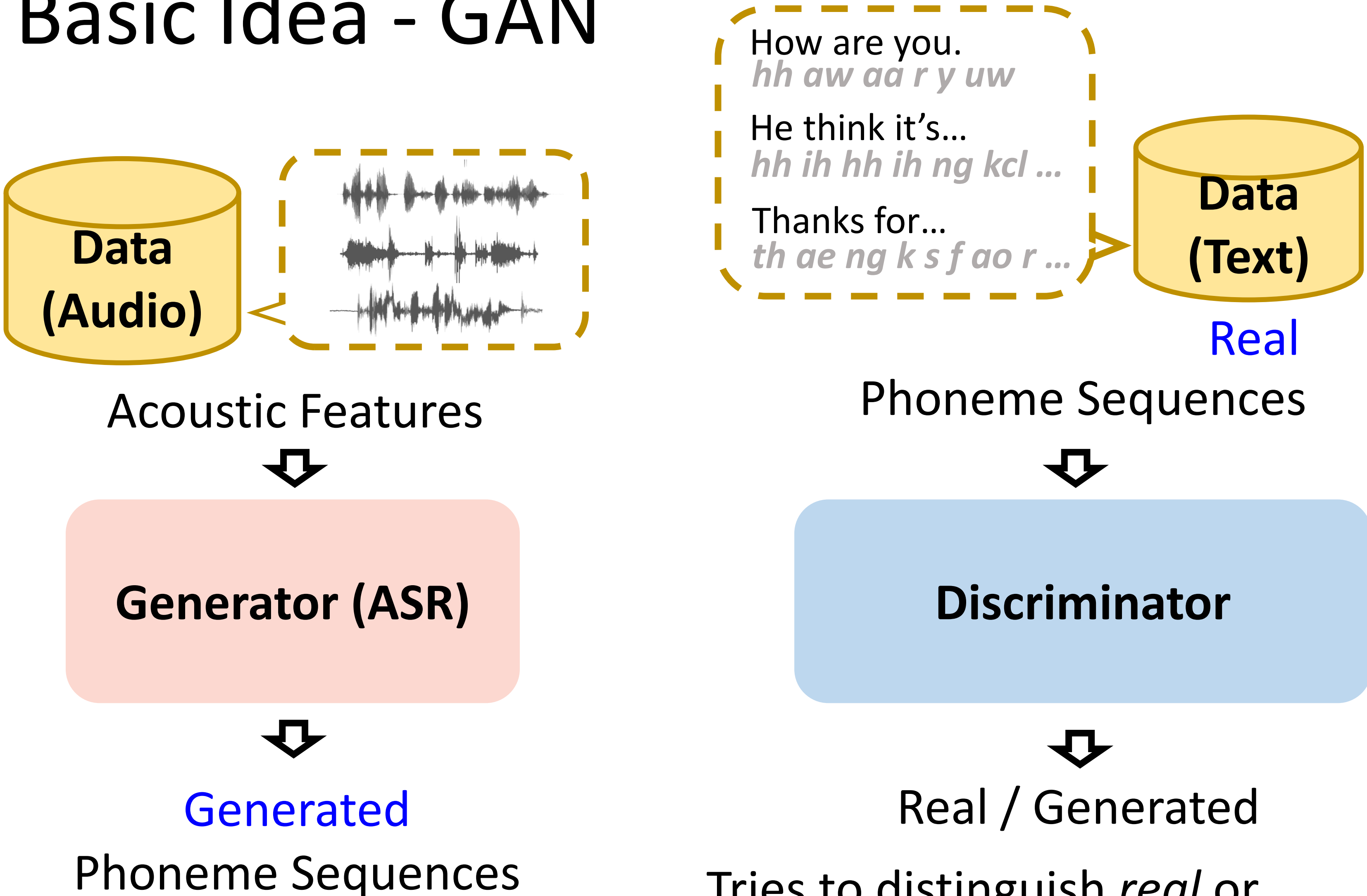


Acoustic Features

Generator (ASR)

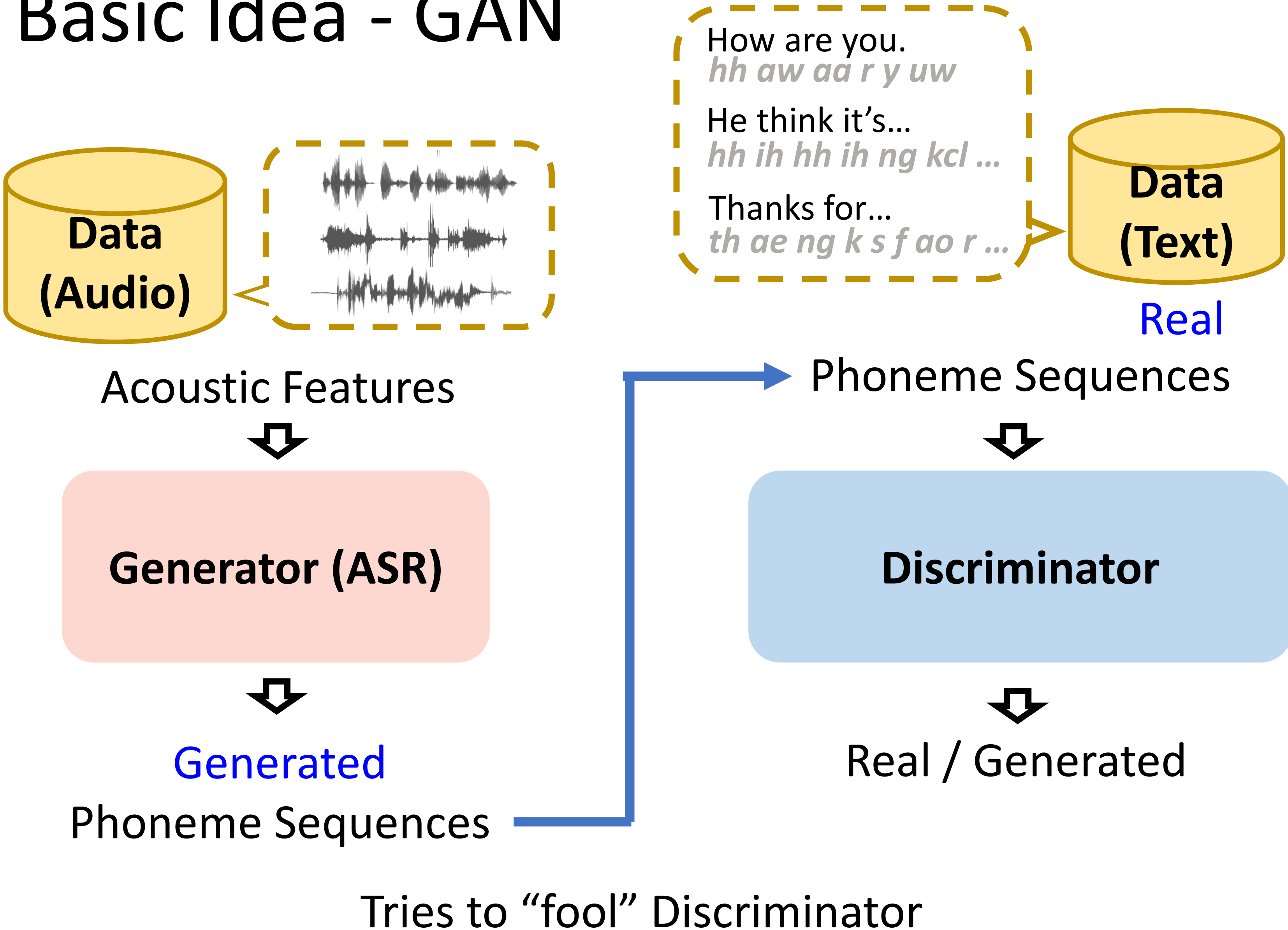
Generated
Phoneme Sequences

Basic Idea - GAN

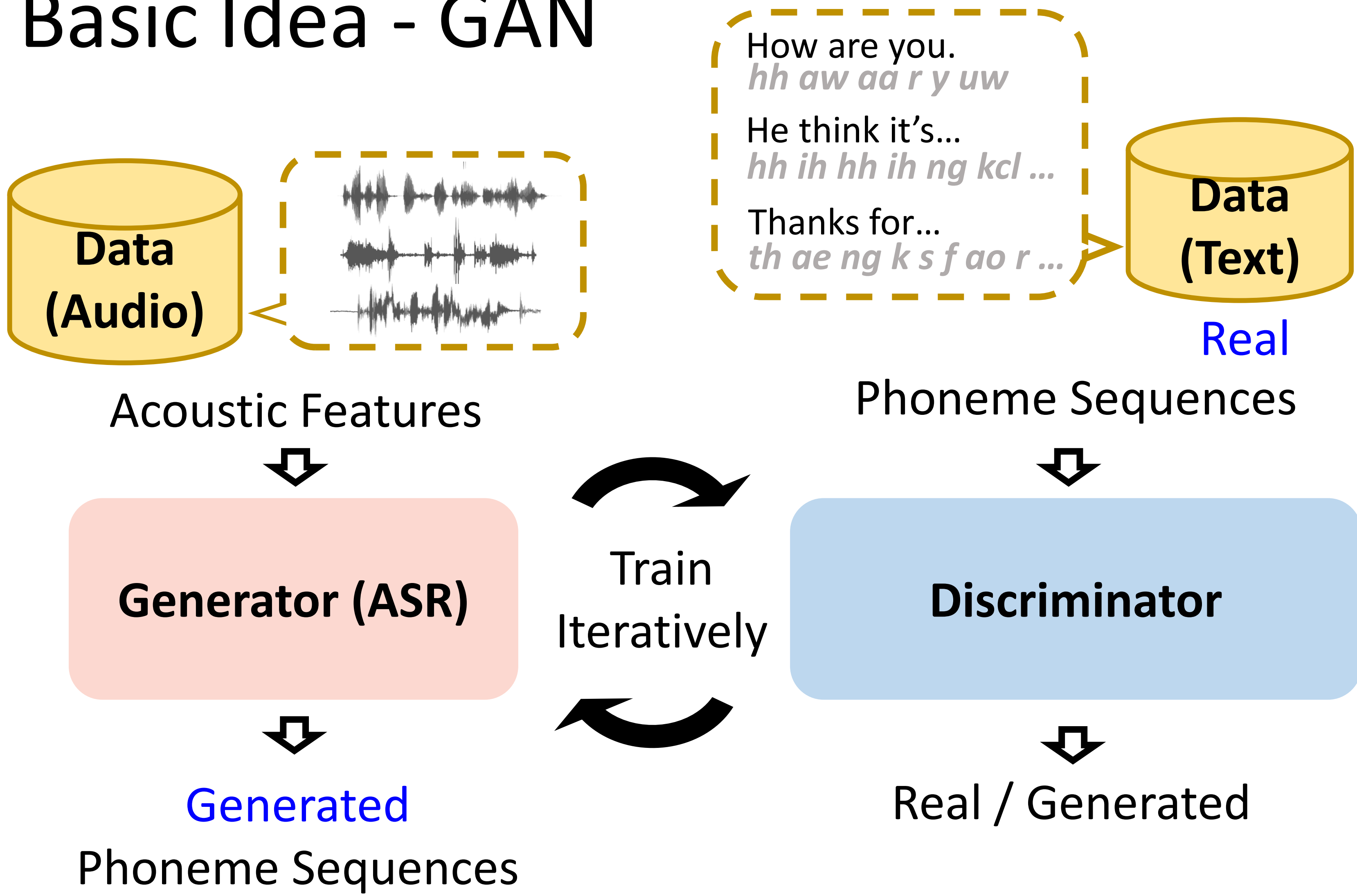


Tries to distinguish *real* or *generated* phoneme sequence.

Basic Idea - GAN

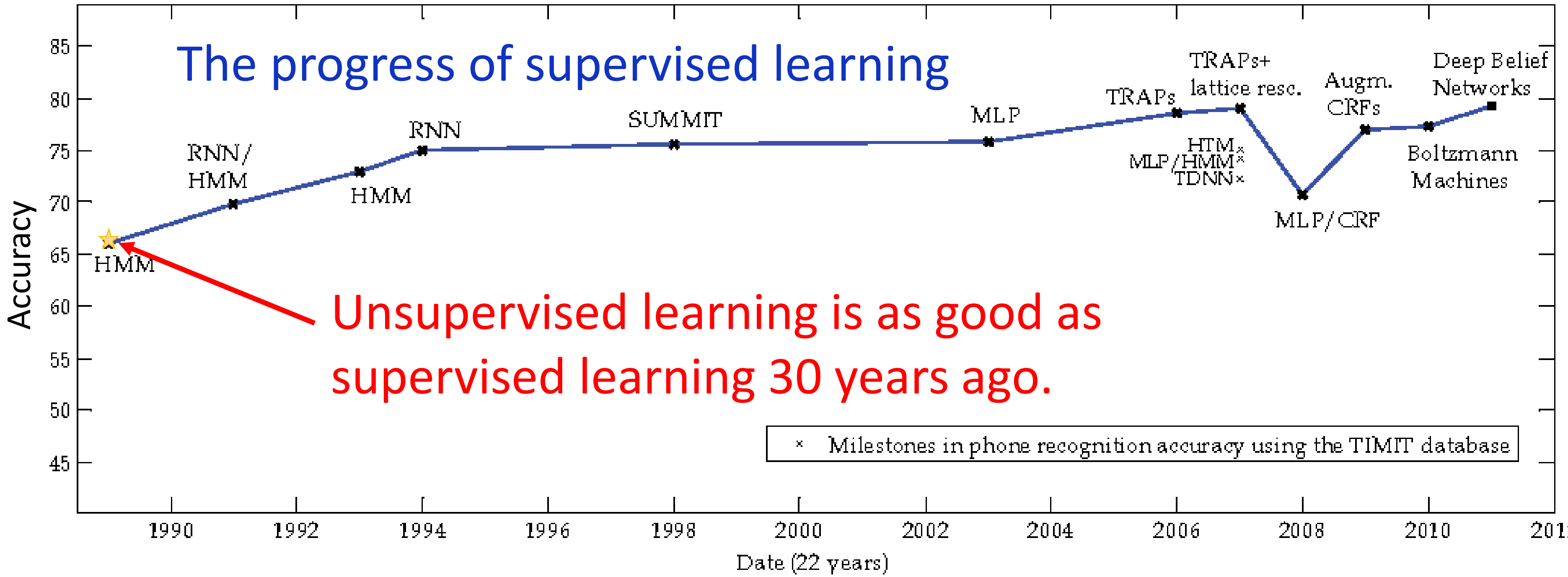


Basic Idea - GAN



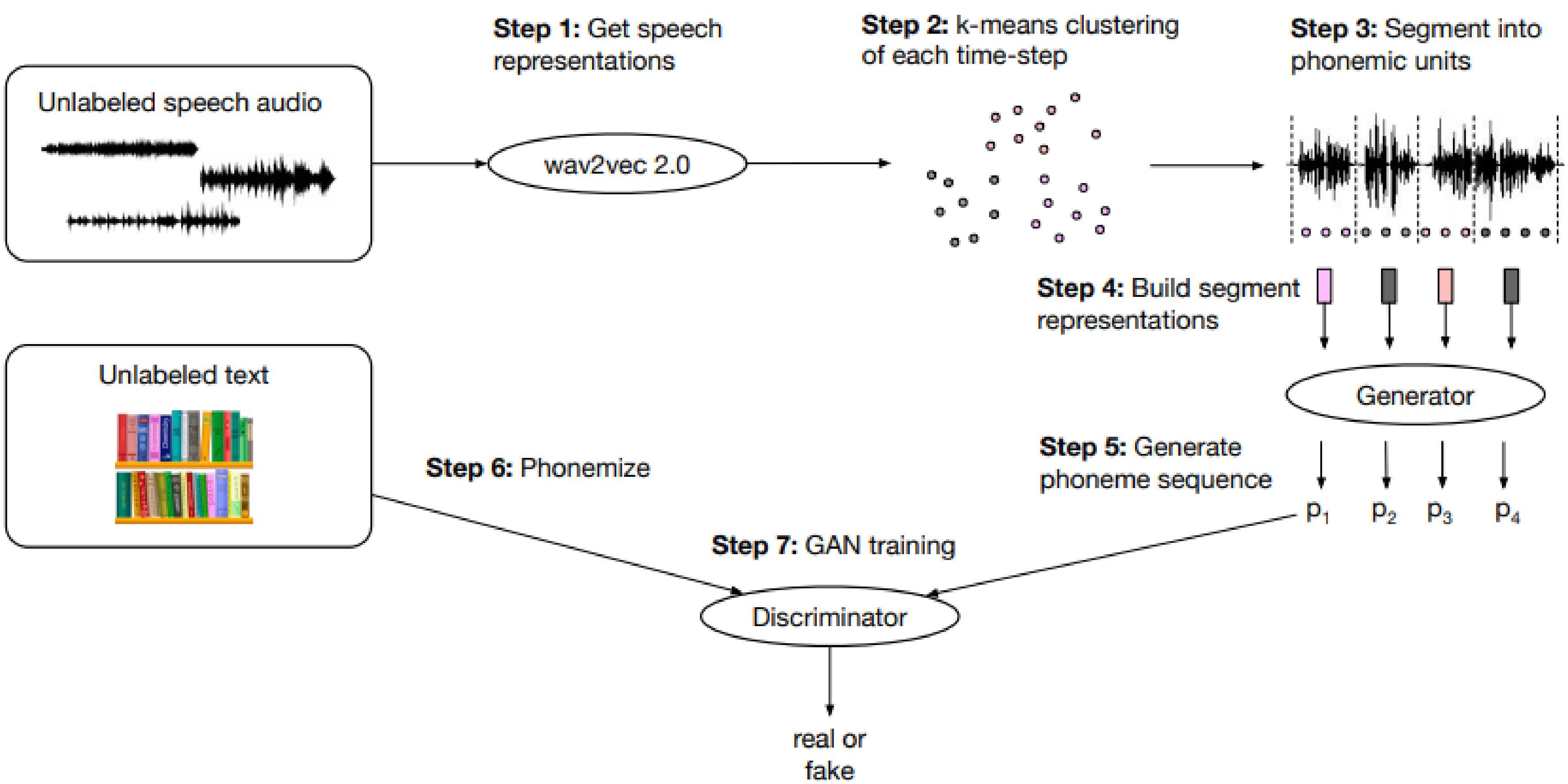
How is the results?

- Unsupervised setting on TIMIT (text and audio are unpair, text is not the transcription of audio)
  - 63.6% PER (oracle boundaries) [Liu, et al., INTERSPEECH 2018]
  - 41.6% PER (automatic segmentation) [Yeh, et al., ICLR 2019]
  - 33.1% PER (automatic segmentation) [Chen, et al., INTERSPEECH 2019]



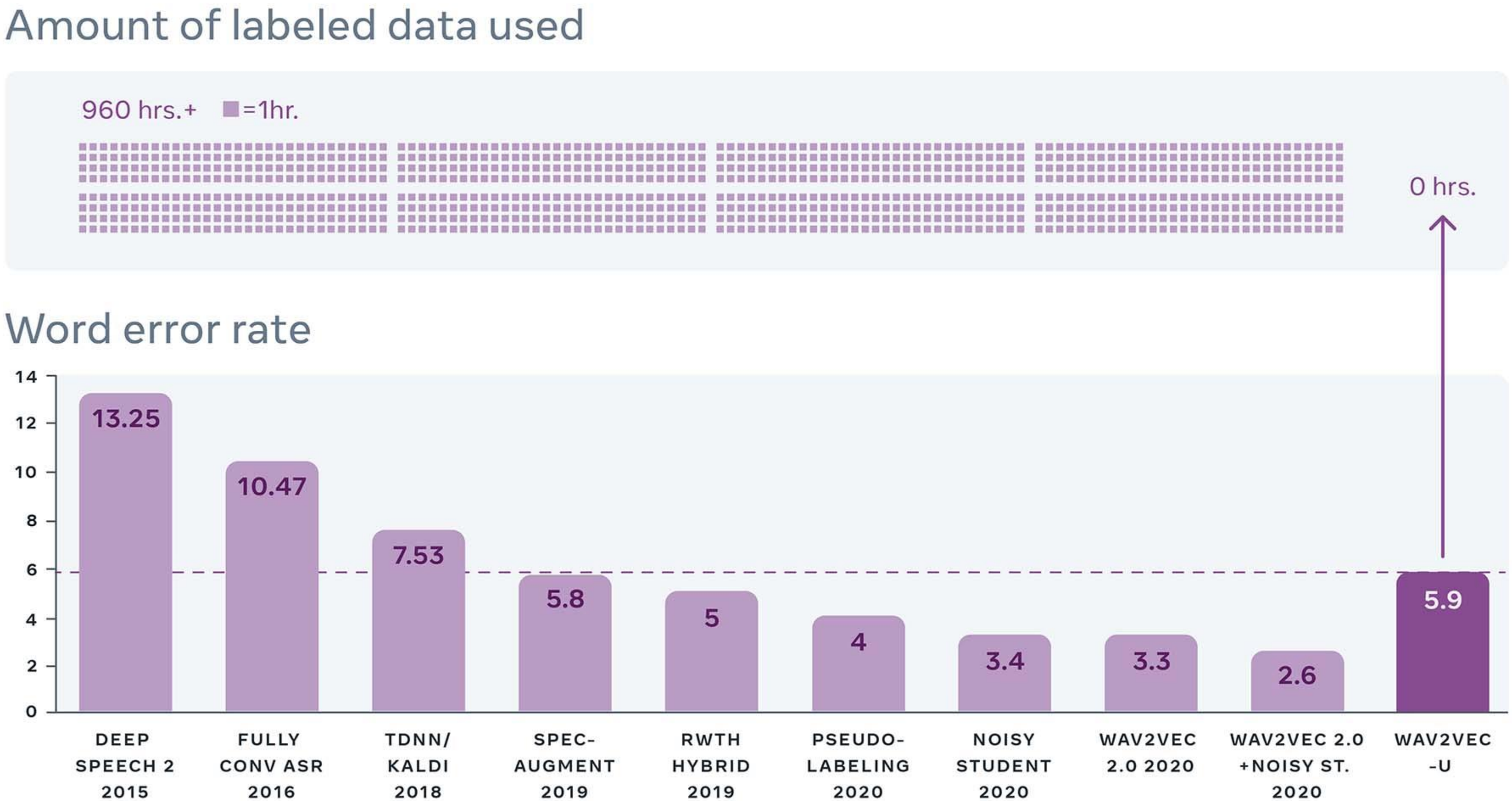
The image is modified from: Phone recognition on the TIMIT database Lopes, C. and Perdigão, F., 2011. Speech Technologies, Vol 1, pp. 285--302.

Recent Progress of Unsupervised ASR



https://ai.facebook.com/blog/wav2vec-unsupervised-speech-recognition-without-supervision/

More Application: Unsupervised ASR



https://ai.facebook.com/blog/wav2vec-unsupervised-speech-recognition-without-supervision/

Learning with less supervision

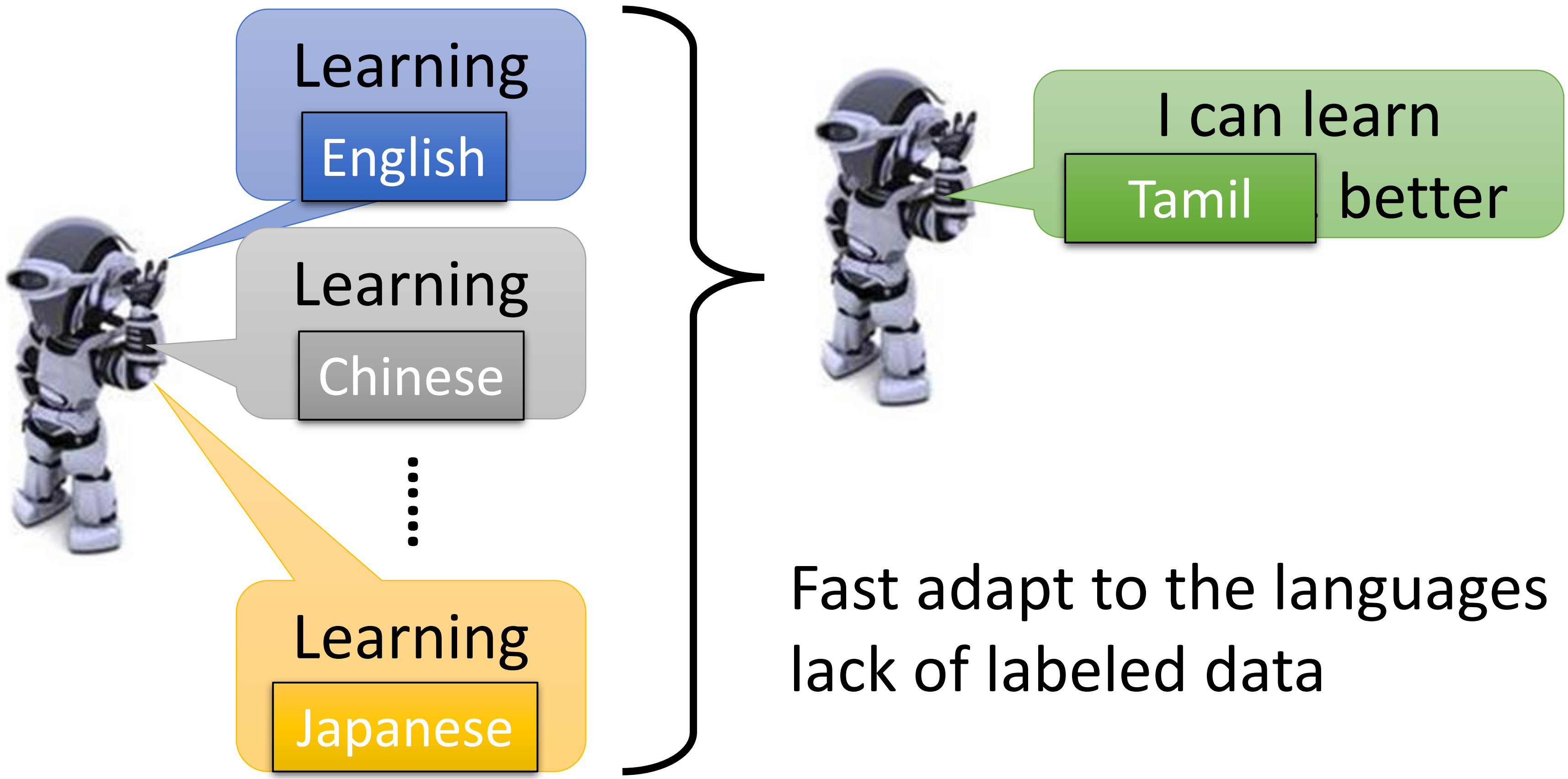
Self-supervised Learning

Learning from Unpaired Data

Meta Learning

Meta Learning

Training Tasks ***Testing Tasks***



Fast adapt to the languages lack of labeled data

Week 2

Week 2					
Date	2021/8/9	2021/8/10	2021/8/11	2021/8/12	2021/8/13
Weekday	Mon	Tue	Wed	Thur	Fri
09:00-10:00 (GMT+8)		**Speaker:** Ming-Wei Chang **Title:** Pre-training for Natural Language Processing Lecture Info Course Link	**Speaker:** Philipp Krähenbühl **Title:** Computer Vision Lecture Info Course Link	**Poster Session 2** Poster	
10:00-10:30 (GMT+8)					
10:30-11:00 (GMT+8)					**Speaker:** Been Kim **Title:** Interpretable machine learning Lecture Info Course Link
11:00-12:00 (GMT+8)					
12:00-20:00 (GMT+8)	Break				
20:00-21:00 (GMT+8)	**Speaker:** John Shawe-Taylor **Title:** An introduction to Statistical Learning Theory and PAC-Bayes Analysis Lecture Info Course Link			**Speakers:** Thang Vu, Shang-Wen Li **Title:** Meta Learning for Human Language Processing Lecture Info1 Lecture Info2 Course Link	
21:00-22:00 (GMT+8)		**Speaker:** Hung-Yi Lee **Title:** Deep Learning for Speech Processing Lecture Info Course Link	**Speaker:** Song Han **Title:** TinyML and Efficient Deep Learning Lecture Info Course Link		**Speaker:** Srinivasan Arunachalam **Title:** Overview of learning quantum states Lecture Info Course Link
22:00-23:00 (GMT+8)					

To learn more:

https://jeffeuxmartin.github.io/meta-learning-hlp/

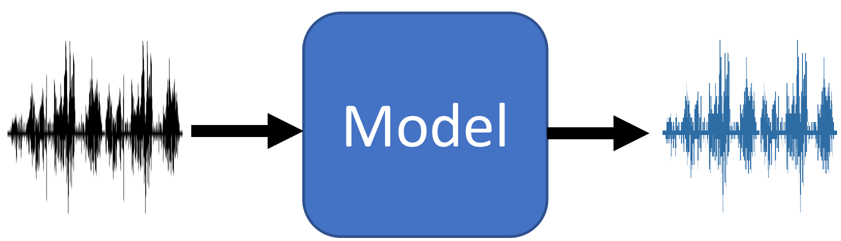
Learning with less supervision

Self-supervised Learning

Learning from Unpaired Data

Meta Learning

One slide for this course



Speech and text can be represented as sequence.

Training a seq-to-seq network

Speech Synthesis

Hung-yi Lee

Before End-to-end

Source of video: https://www.youtube.com/watch?v=0rAyrmm7vv0

VODER (1939): New York World's Fair

Before End-to-end

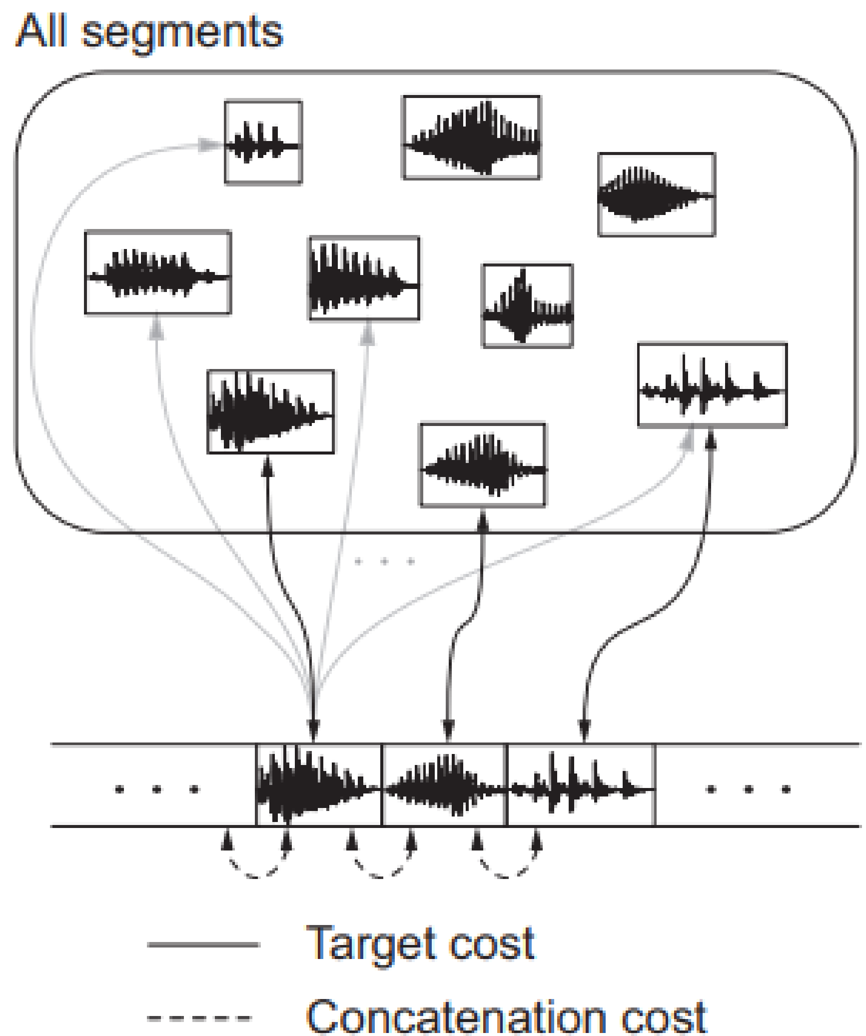
- IBM computer (1960s): John Larry Kelly Jr. using an IBM computer to synthesize speech at Bell lab.

Source of video and audio: https://youtu.be/UGsfwhb4-bQ
https://www.vintagecomputermusic.com/mp3/s2t9_Computer_Speech_Demonstration.mp3

Before End-to-end

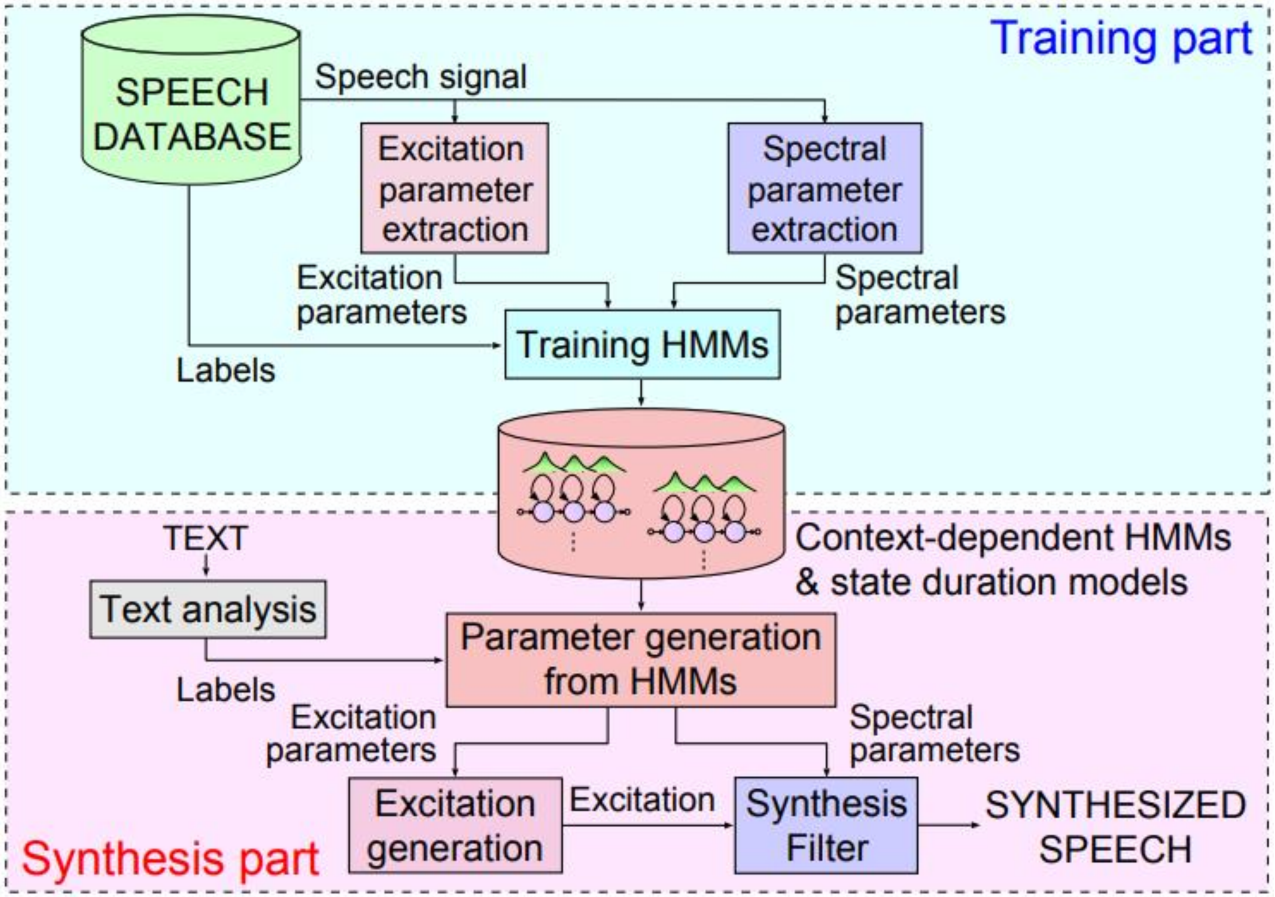
speeches from a large database

Concatenative Approach



Source of image:
https://www.cs.cmu.edu/~pmuthuku/mlsp_page/lectures/spss_specom.pdf

Before End-to-end



<u>Parametric Approach</u>

Source of image: http://hts.sp.nitech.ac.jp/?Tutorial

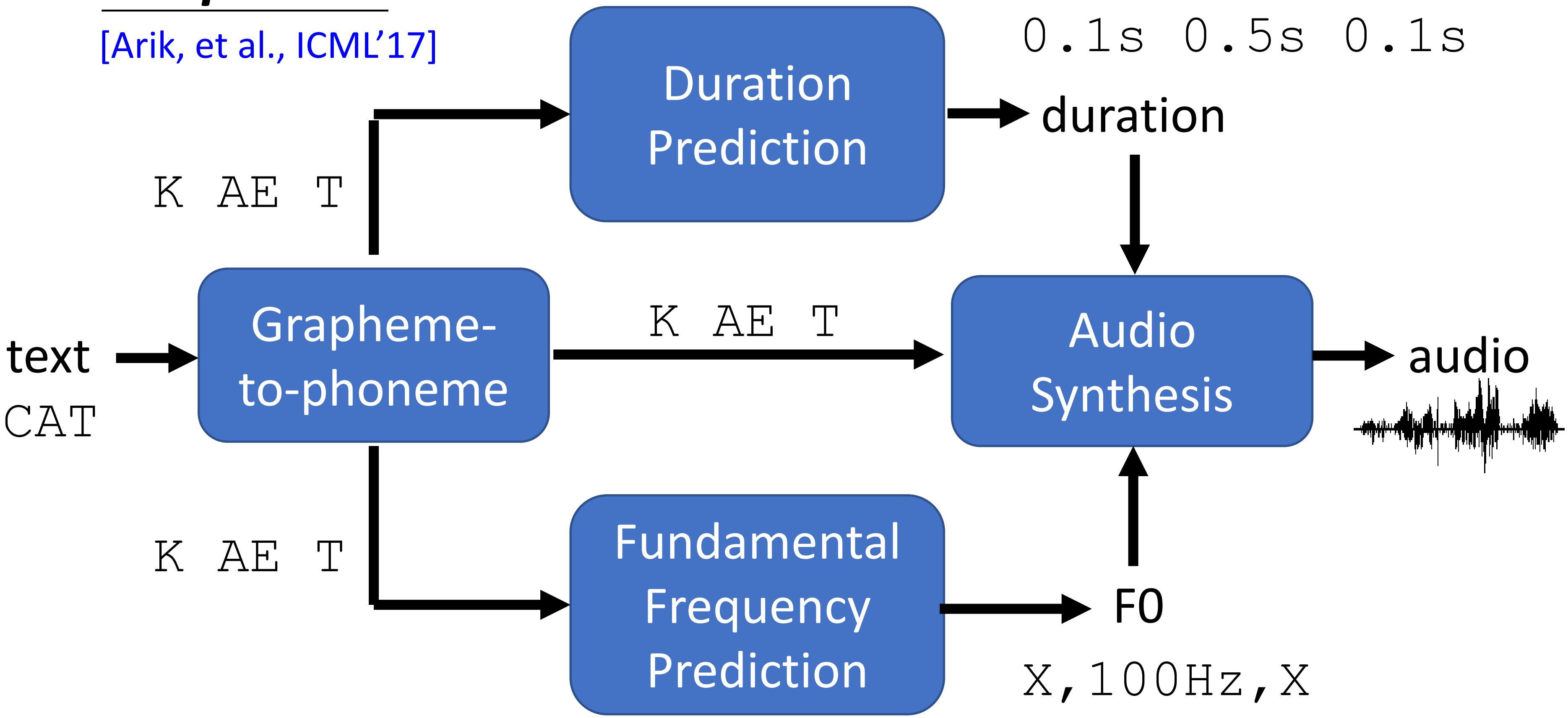
Before End-to-end

Deep Voice 3 is end-to-end.

[Ping, et al., ICLR’18]

<u>Deep Voice</u>

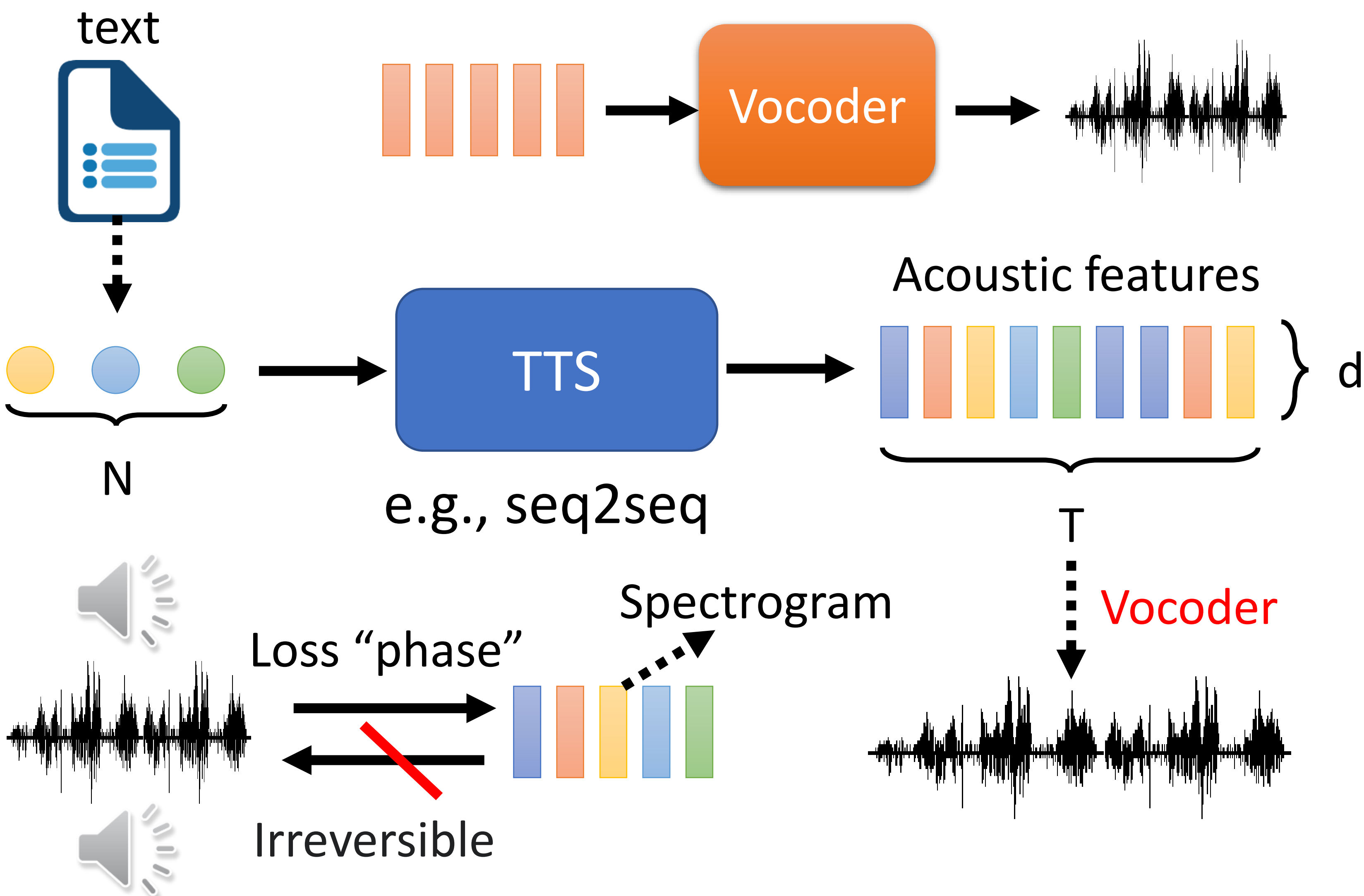
[Arik, et al., ICML’17]



All the components are deep learning based.

End-to-end

- Rule-based
- Deep Learning: WaveNet



Tacotron

[Wang, et al., INTERSPEECH'17]
[Shen, et al., ICASSP'18]

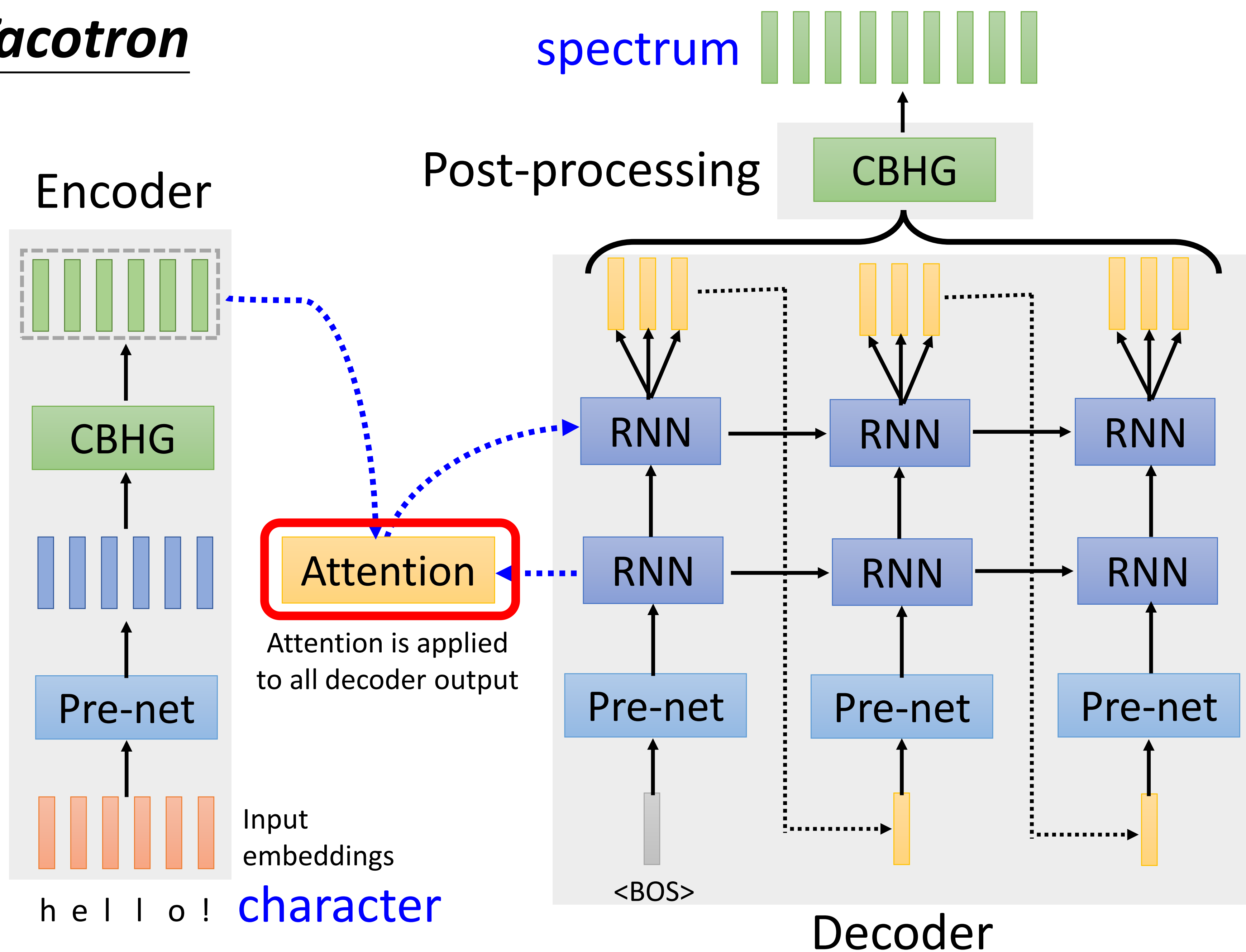
TACOTRON: TOWARDS END-TO-END SPEECH SYNTHESIS

Yuxuan Wang*, RJ Skerry-Ryan*, Daisy Stanton, Yonghui Wu, Ron J. Weiss†, Navdeep Jaitly, Zongheng Yang, Ying Xiao*, Zhifeng Chen, Samy Bengio†, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, Rif A. Saurous*

Google, Inc.
{yxwang,rjryan,rif}@google.com

*These authors really like tacos.
†These authors would prefer sushi.

Tacotron


spectrum
Post-processing
CBHG
Encoder
CBHG
Pre-net
Input
embeddings
h e l l o !
character
Attention
Attention is applied
to all decoder output
RNN
RNN
RNN
RNN
RNN
RNN
Pre-net
Pre-net
Pre-net
<BOS>
Decoder

How good is Tacotron?

Version 1

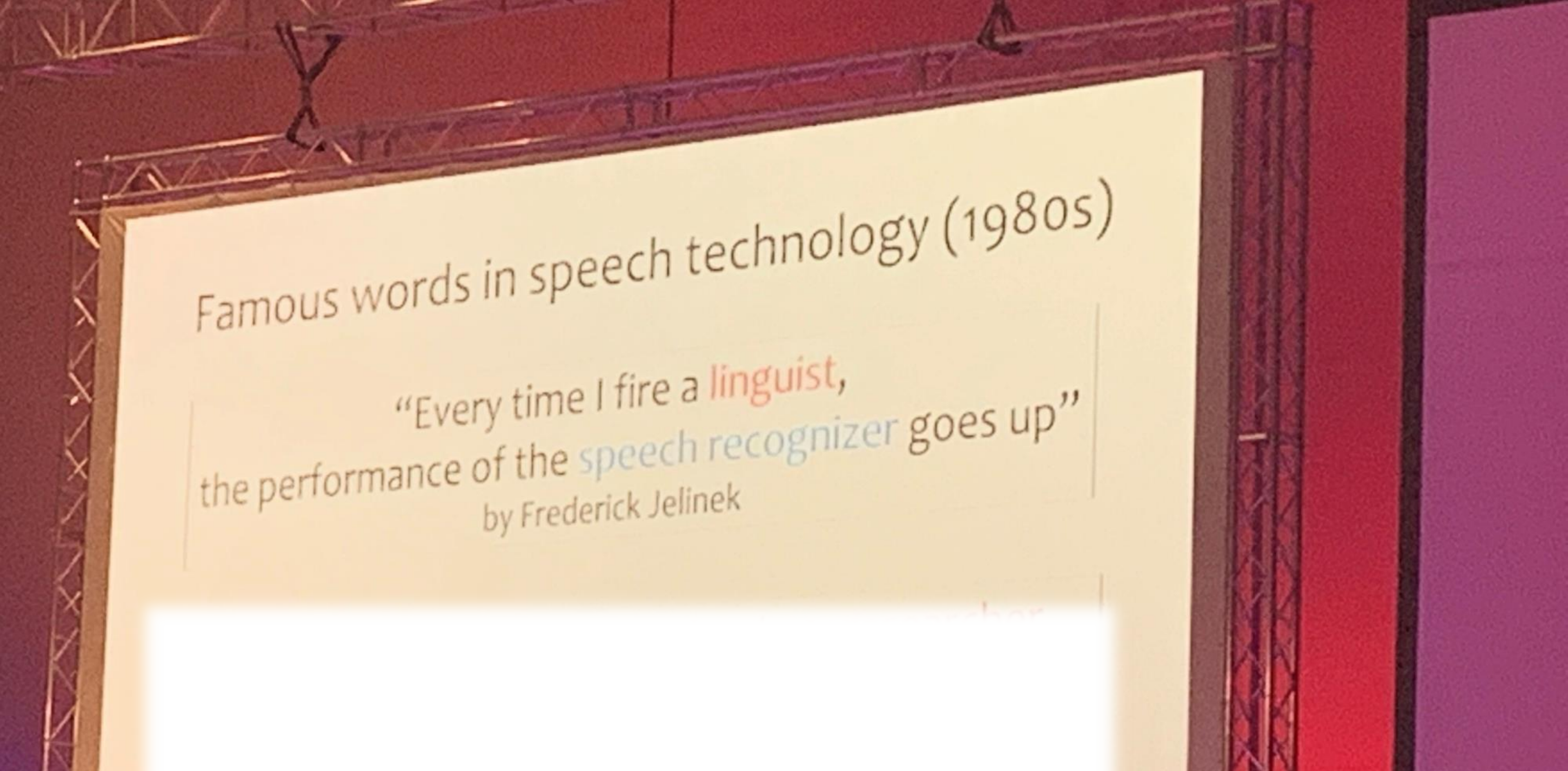
[Wang, et al., INTERSPEECH'17]

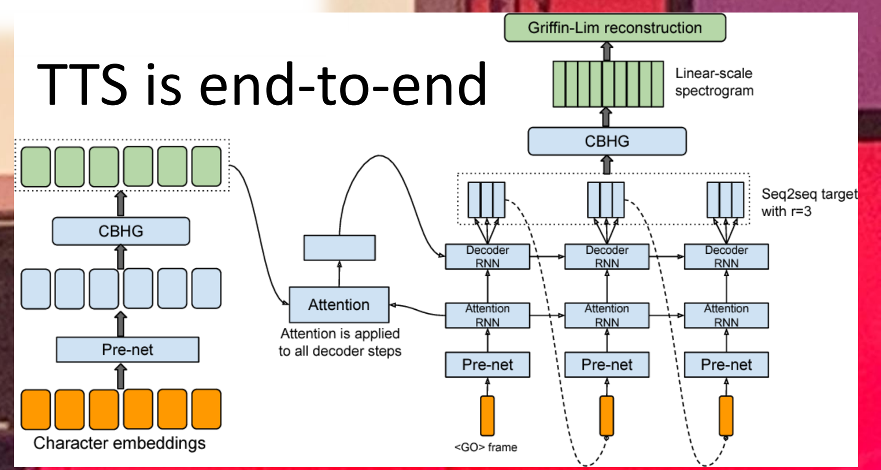
	mean opinion score
Tacotron	3.82 ± 0.085
Parametric	3.69 ± 0.109
Concatenative	4.09 ± 0.119

Version 2

[Shen, et al., ICASSP'18]

System	MOS
Parametric	3.492 ± 0.096
Tacotron (Griffin-Lim)	4.001 ± 0.087
Concatenative	4.166 ± 0.091
WaveNet (Linguistic)	4.341 ± 0.051
Ground truth	4.582 ± 0.053
Tacotron 2 (this paper)	$\mathbf{4.526 \pm 0.066}$

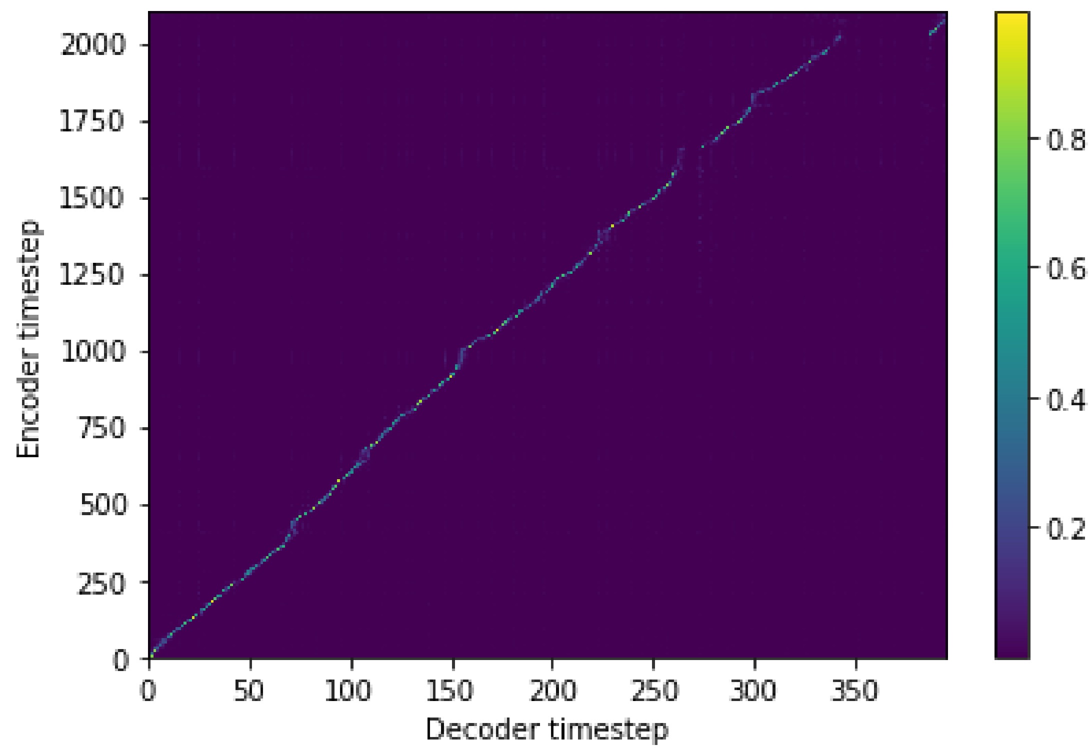




(Keiichi Tokuda, keynote, INTERSPEECH'19)

"A text-to-speech synthesis system typically consists of multiple stages, such as a text analysis frontend, an acoustic model and an audio synthesis module. Building these components often requires extensive domain expertise and may contain brittle design choices. In this paper, we present Tacotron, an end-to-end generative text-to-speech model that synthesizes speech directly from characters."

Provided by
Po-chun Hsu

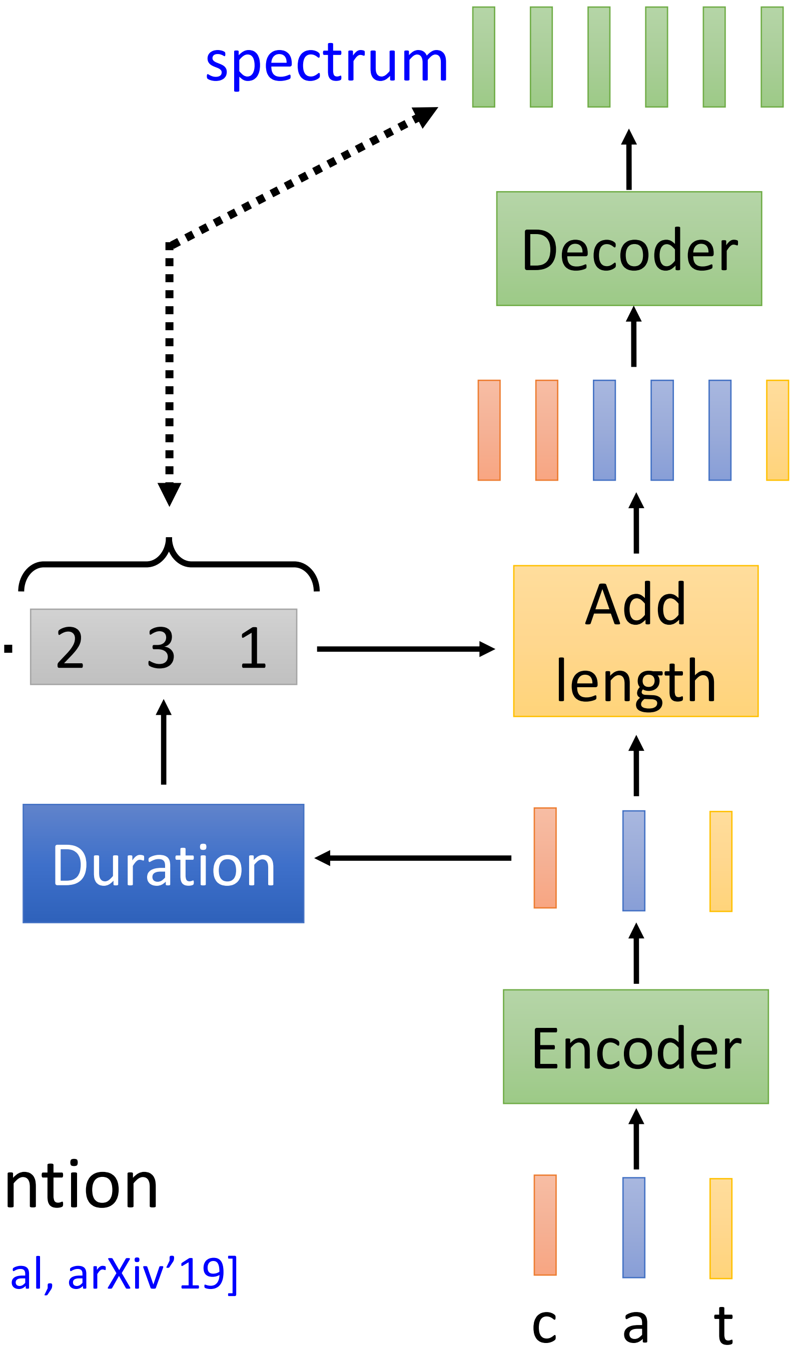


Fast Speech

[Ren, et al., NeurIPS'19]

The renaissance of duration modeling.

Easy to control



How to train the model end-to-end?

Duration Informed Attention Network (DurIAN) [Yu, et al, arXiv'19]

Fast Speech

Source of results:
https://arxiv.org/pdf/1905.09263.pdf

In 50 sentences:

Method	Repeats	Skips	Error Sentences	Error Rate
Tacotron 2	4	11	12	24%
Transformer TTS	7	15	17	34%
FastSpeech	0	0	0	0%

zero zero zero zero zero zero zero zero two seven nine eight F three forty zero zero zero zero zero six four two eight zero one eight

c five eight zero three three nine a zero bf eight FALSE zero zero zero bba3add2 - c229 - 4cdb -

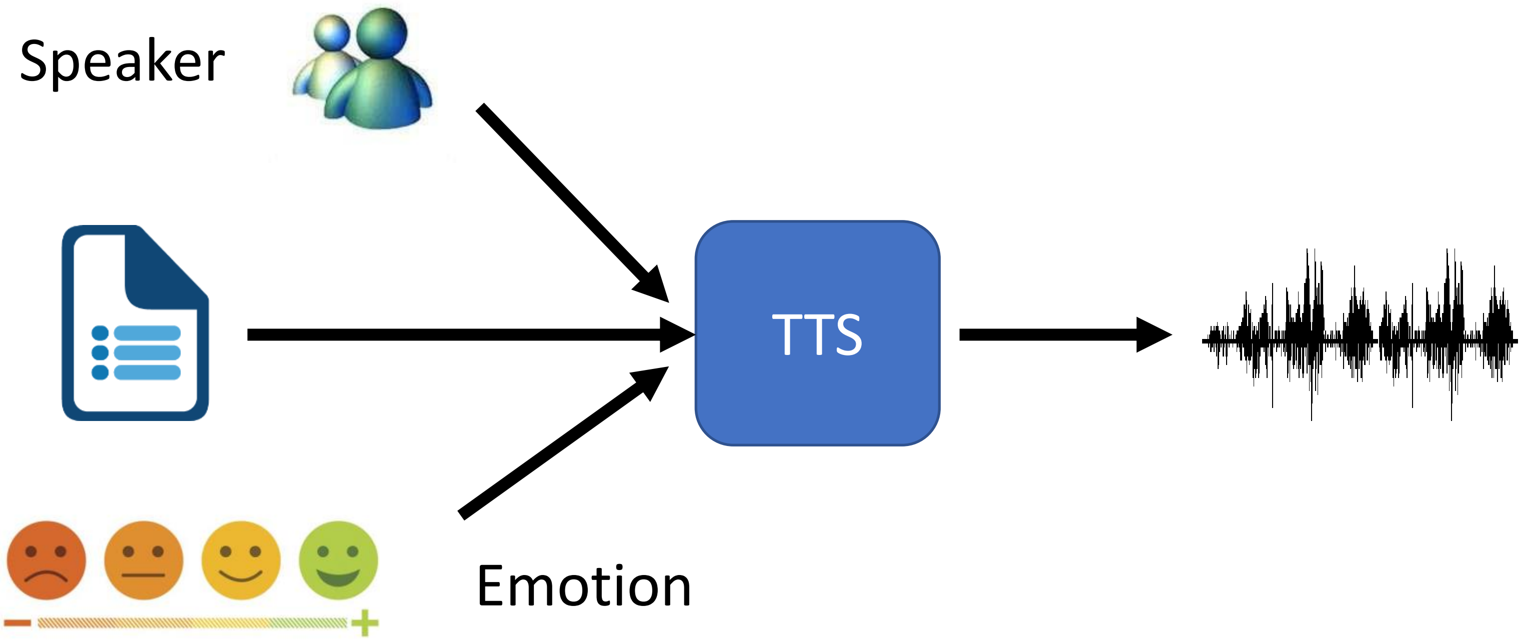
Calendaring agent failed with error code 0x80070005 while saving appointment .

Exit process - break ld - Load module - output ud - Unload module - ignore ser - System error - ignore ibp - Initial breakpoint -

h t t p colon slash slash teams slash sites slash T A G slash default dot aspx As always , any feedback , comments ,

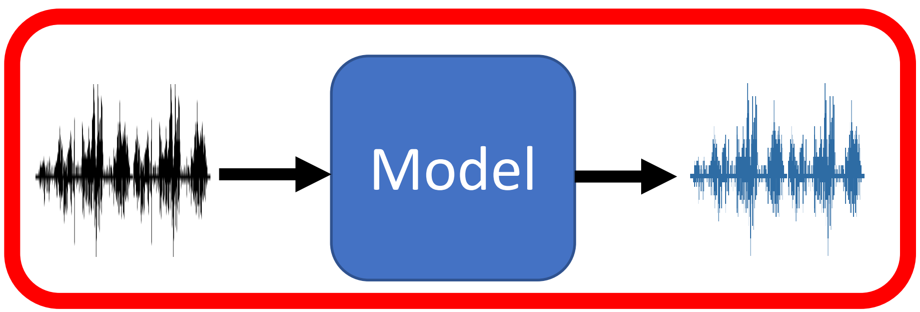
two thousand and five h t t p colon slash slash news dot com dot com slash i slash n e slash f d slash two zero zero three slash f d

Controllable TTS



To learn more:
Xu Tan, Tao Qin, Frank Soong, Tie-Yan Liu, A Survey on Neural Speech Synthesis, 2021, https://arxiv.org/abs/2106.15561

One slide for this course



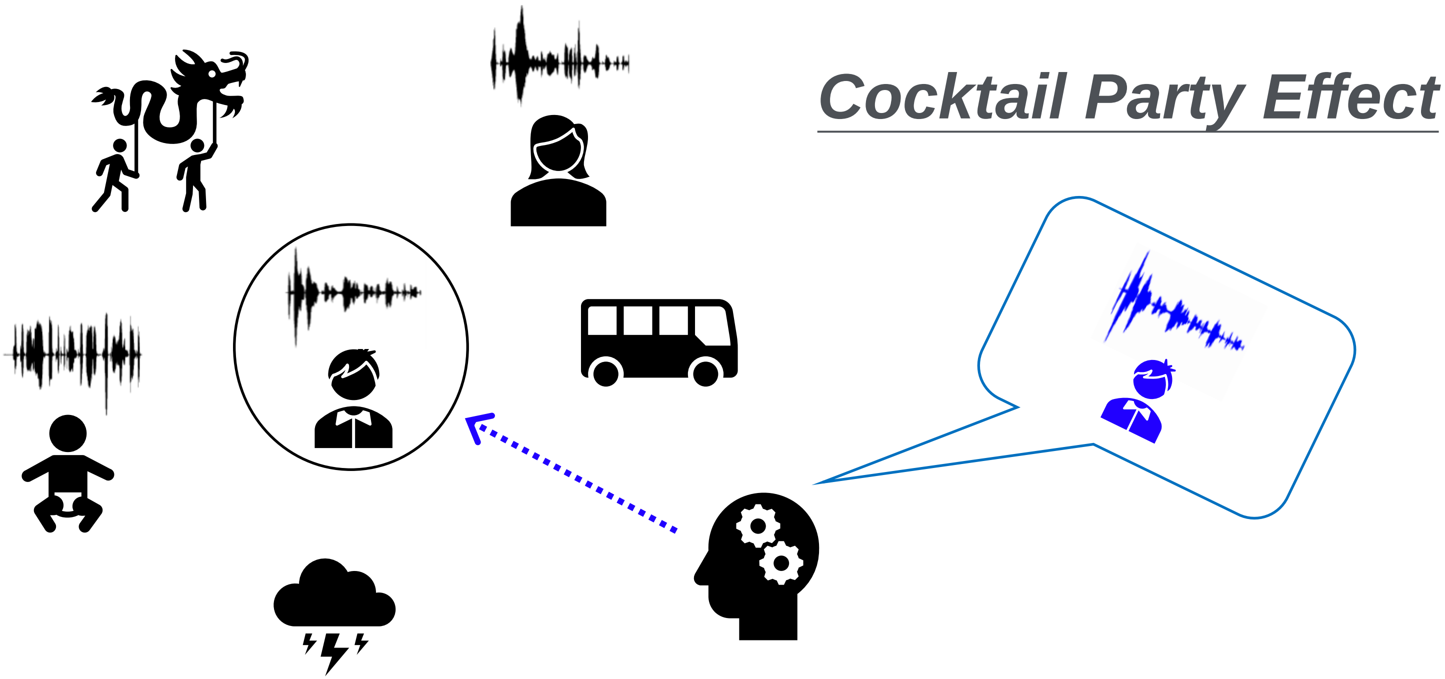
Speech and text can be represented as sequence.

Training a seq-to-seq network

Speech Separation

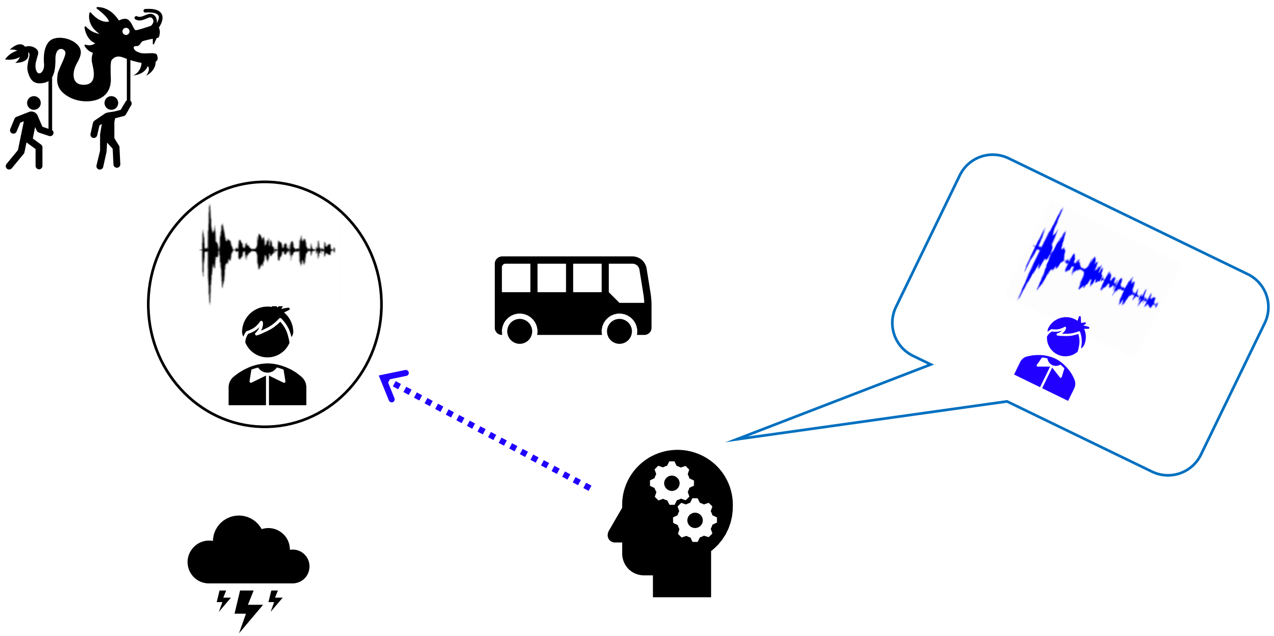
Speech Separation

- Humans can focus on the voice produced by a single speaker in a crowded and noisy environments.



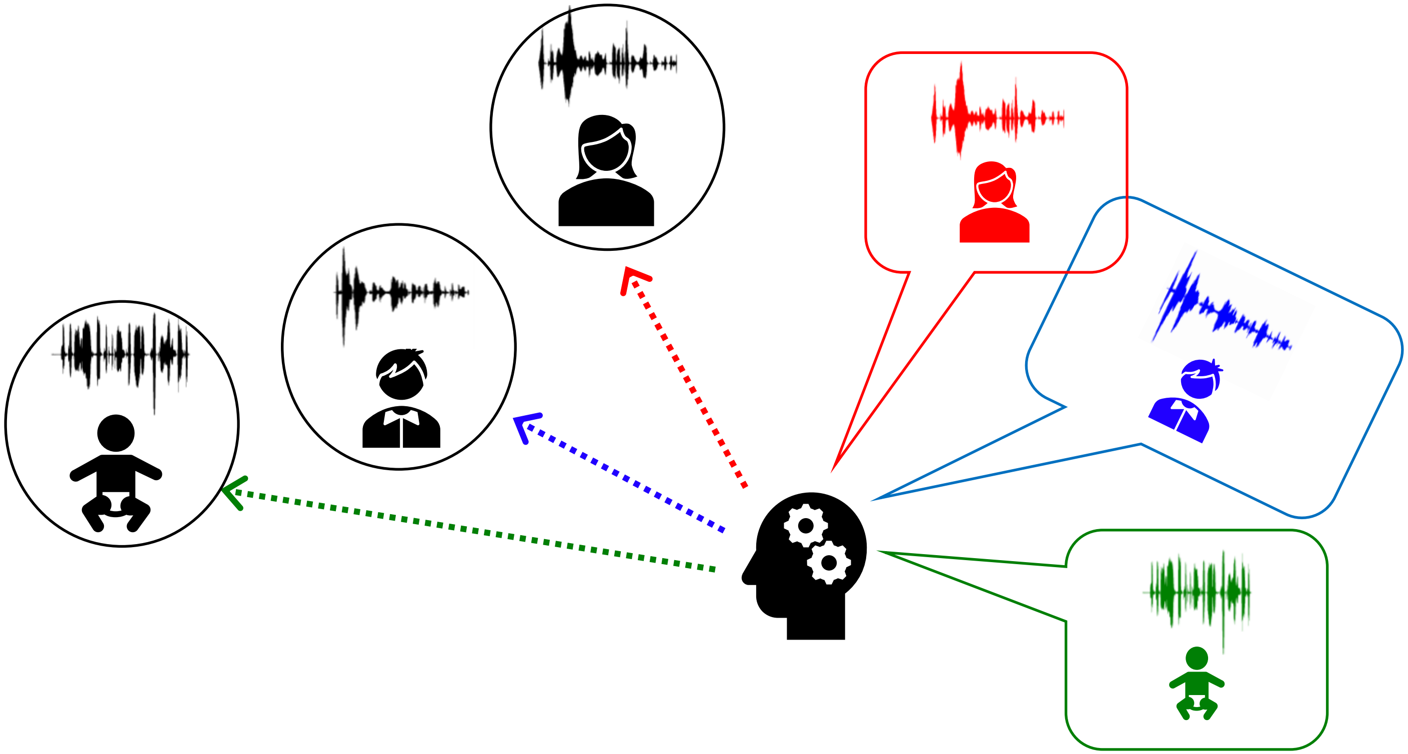
Speech Separation

- **Speech Enhancement**: speech-nonspeech separation (de-noising)



Speech Separation

- **Speaker Separation**: multi-speaker talking

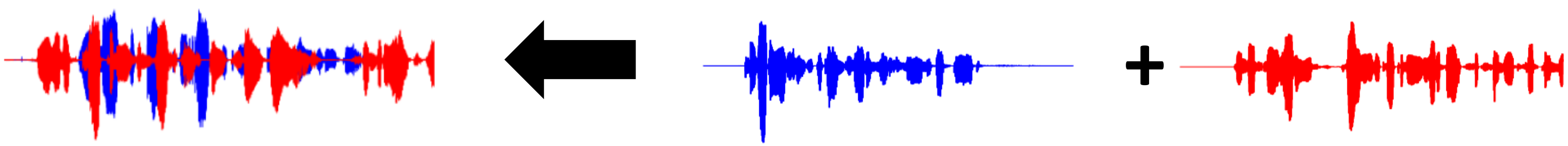


Speaker Separation

Seq2seq with two decoders (?)

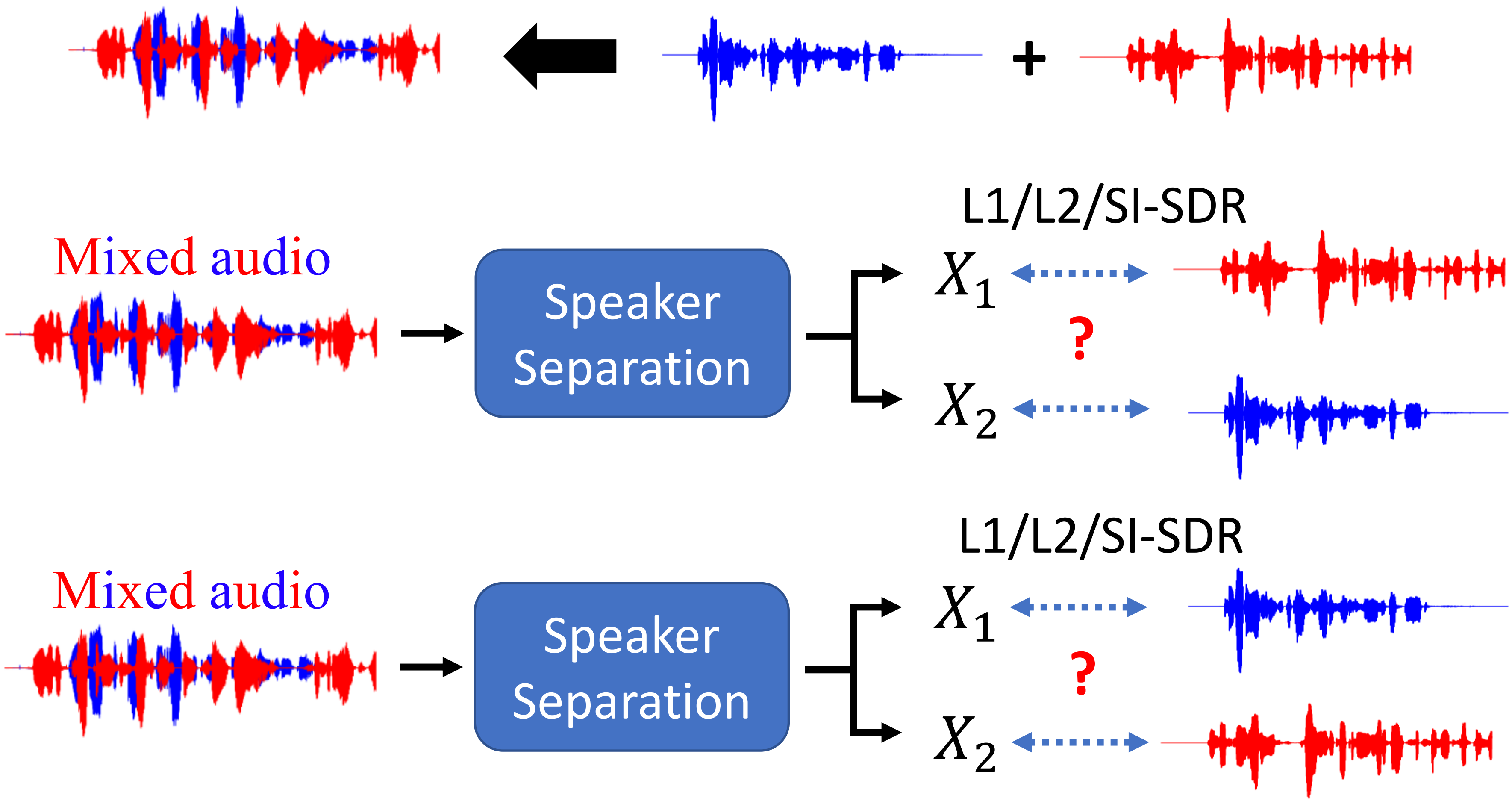
- Focusing on two speakers
- Focusing on single microphone
- **Speaker independent**: training and testing speakers are completely different

Training Data:



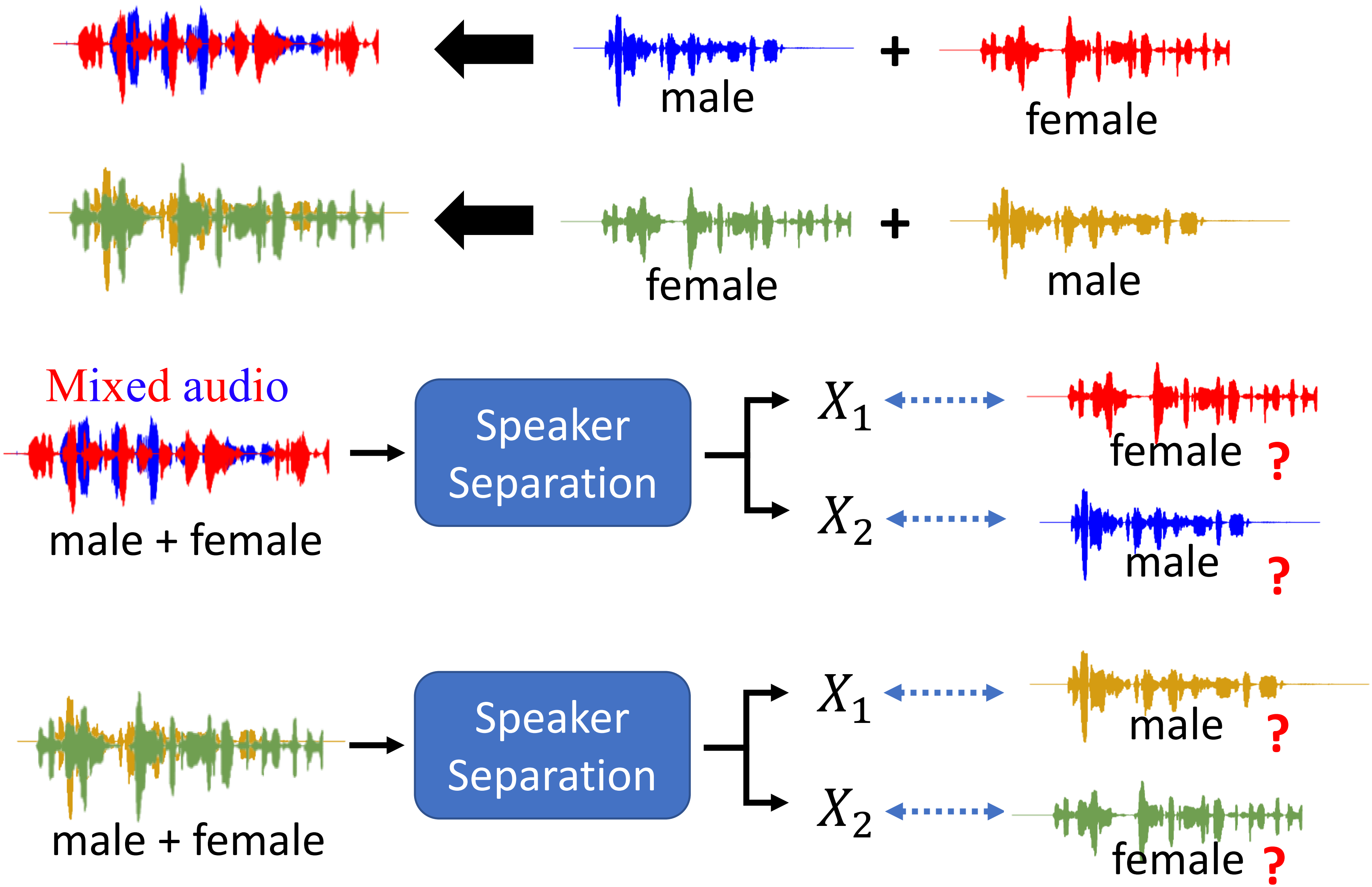
It is easy to generate training data.

Permutation Issue

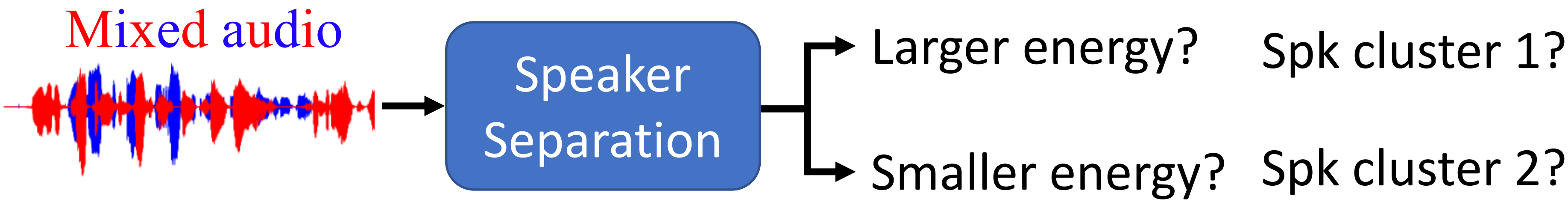


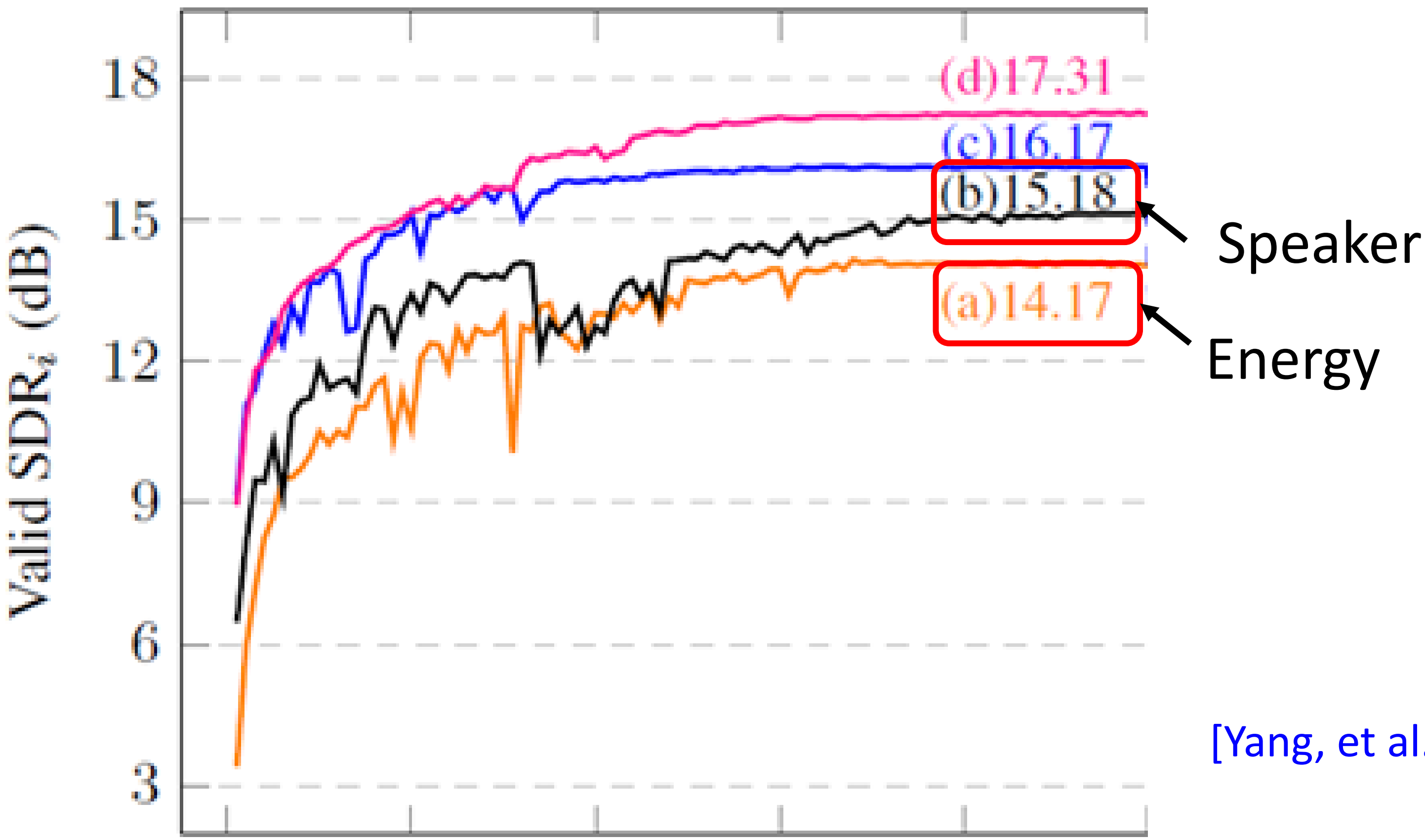
Permutation Issue

Cluster by Gender? Pitch? Energy?



Clustered by … ?

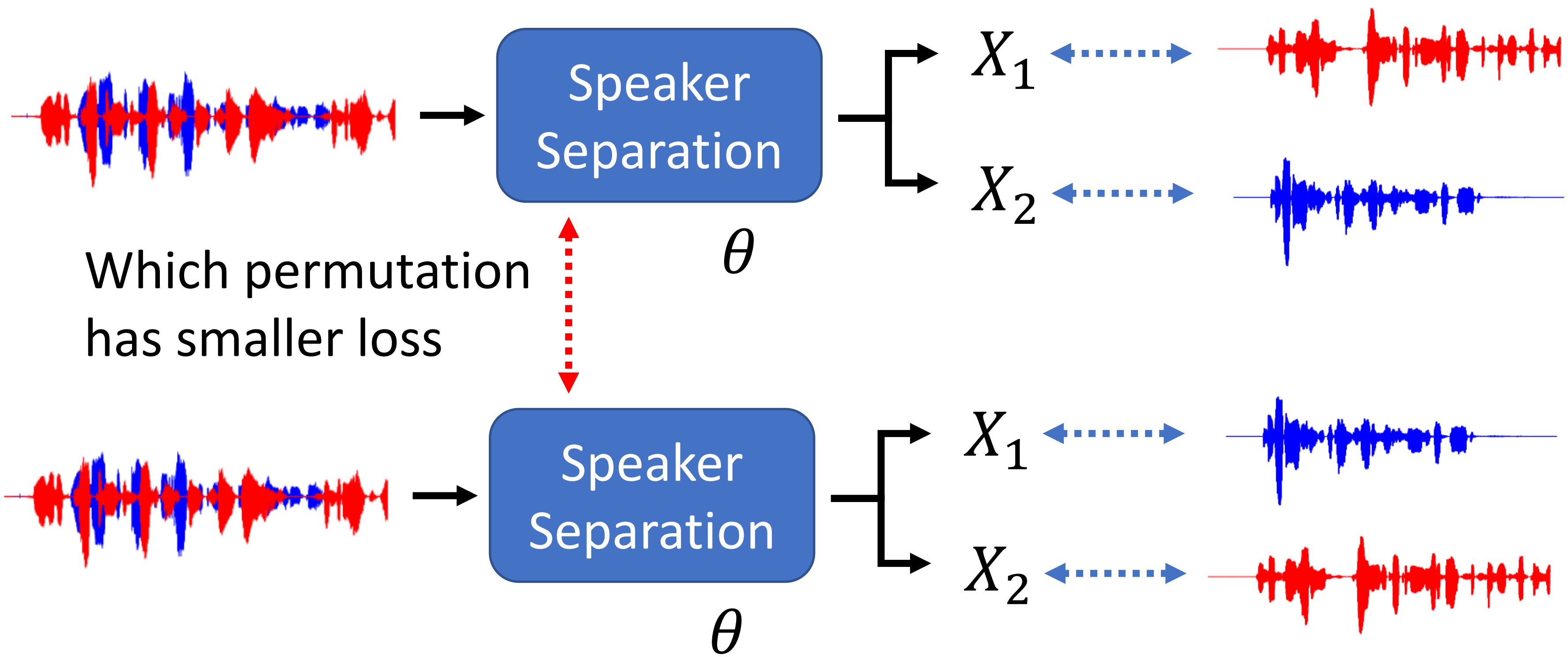




[Yang, et al., ICASSP'20]

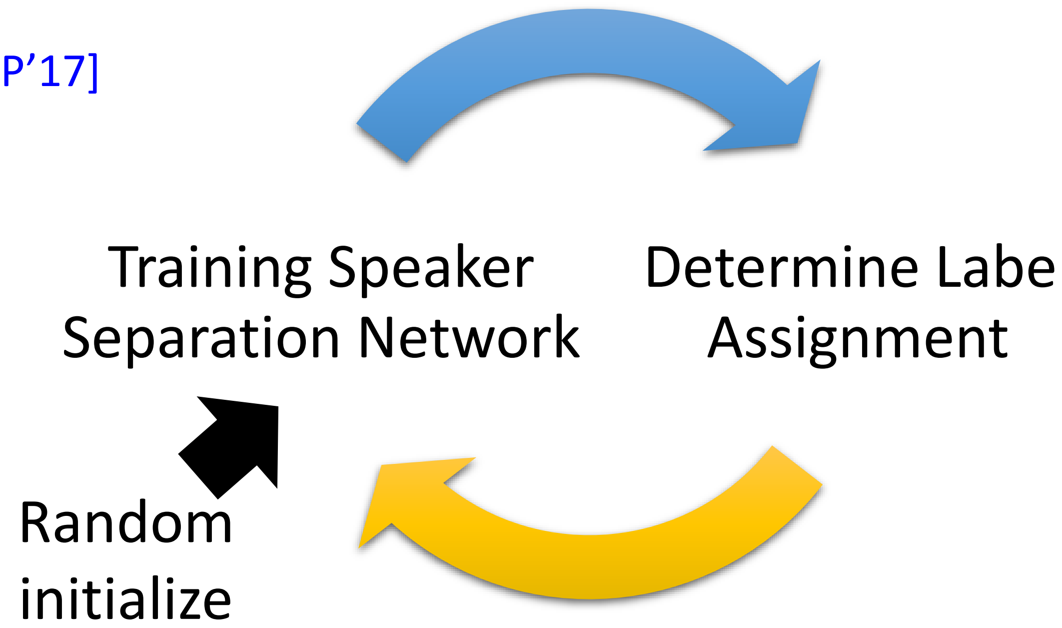
Permutation Invariant Training (PIT) [Kolbæk, et al., TASLP'17]

Given a speaker separation model θ, we can determine the permutation

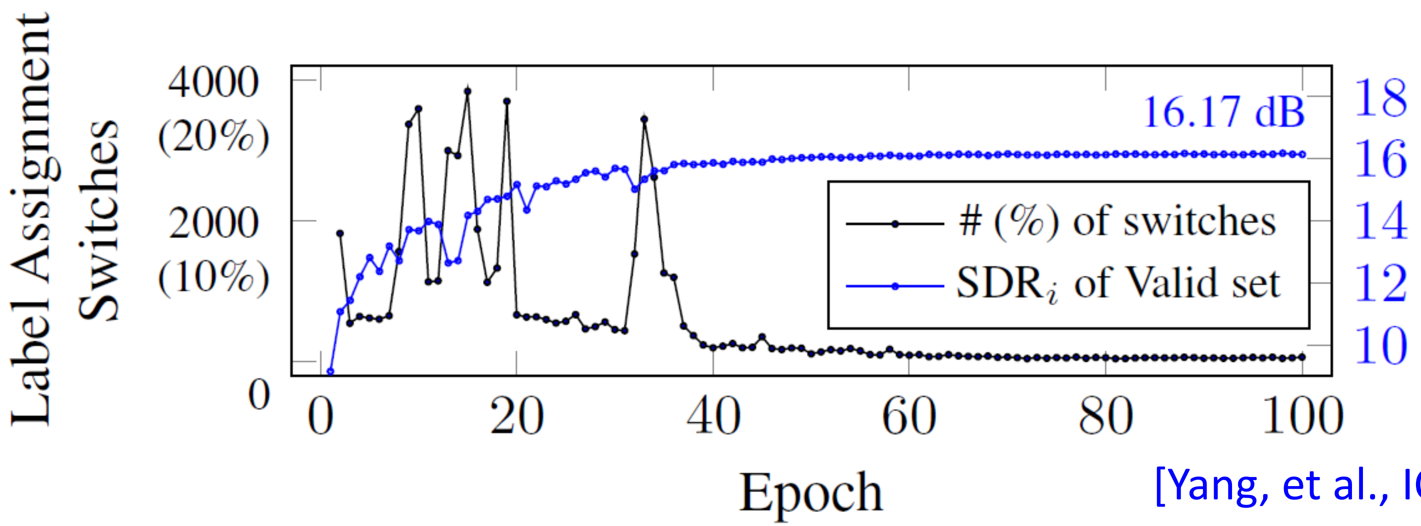


But we need permutation to train speaker separation model …

PIT [Kolbæk, et al., TASLP'17]

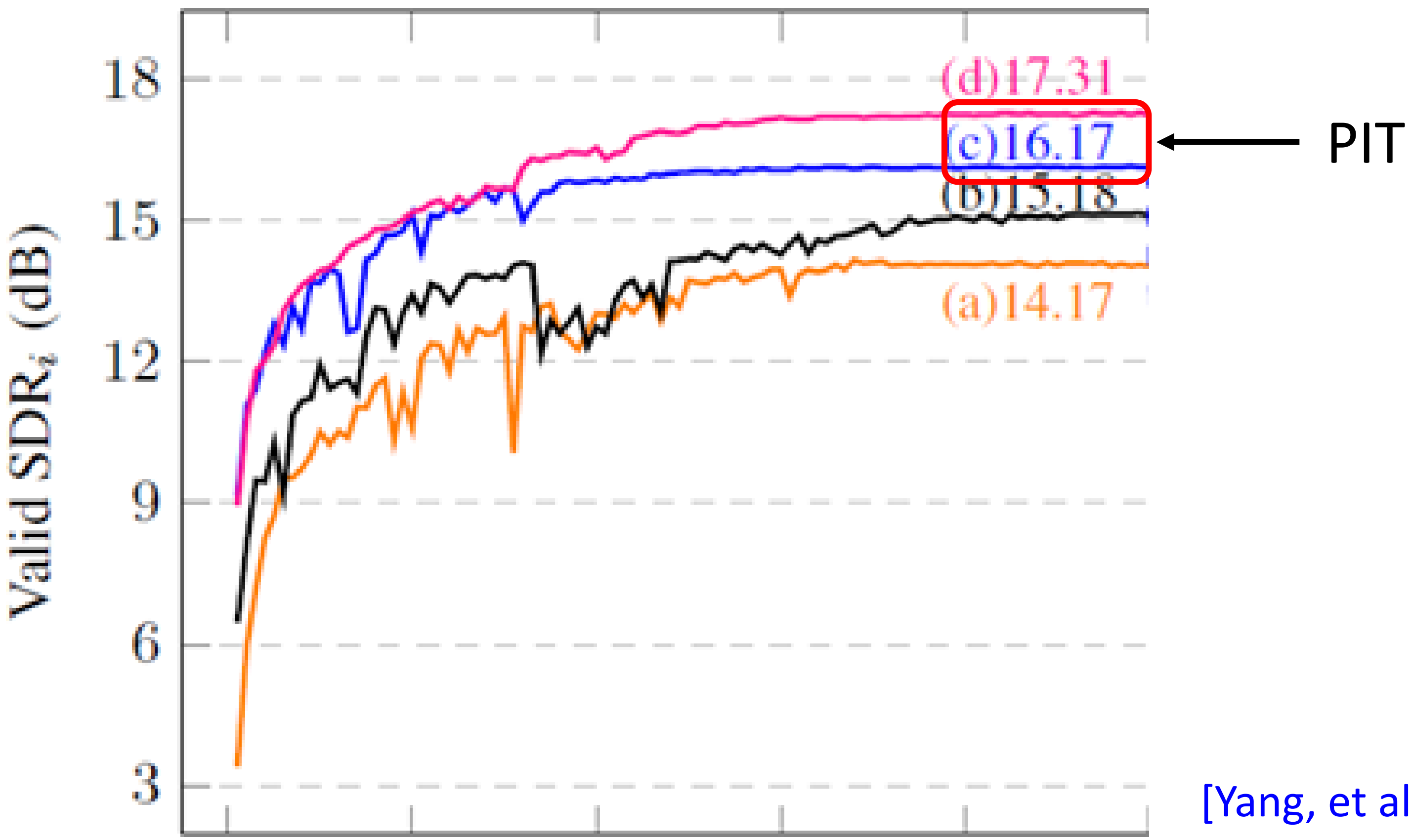


At the beginning, the assignment is not stable.



[Yang, et al., ICASSP'20]

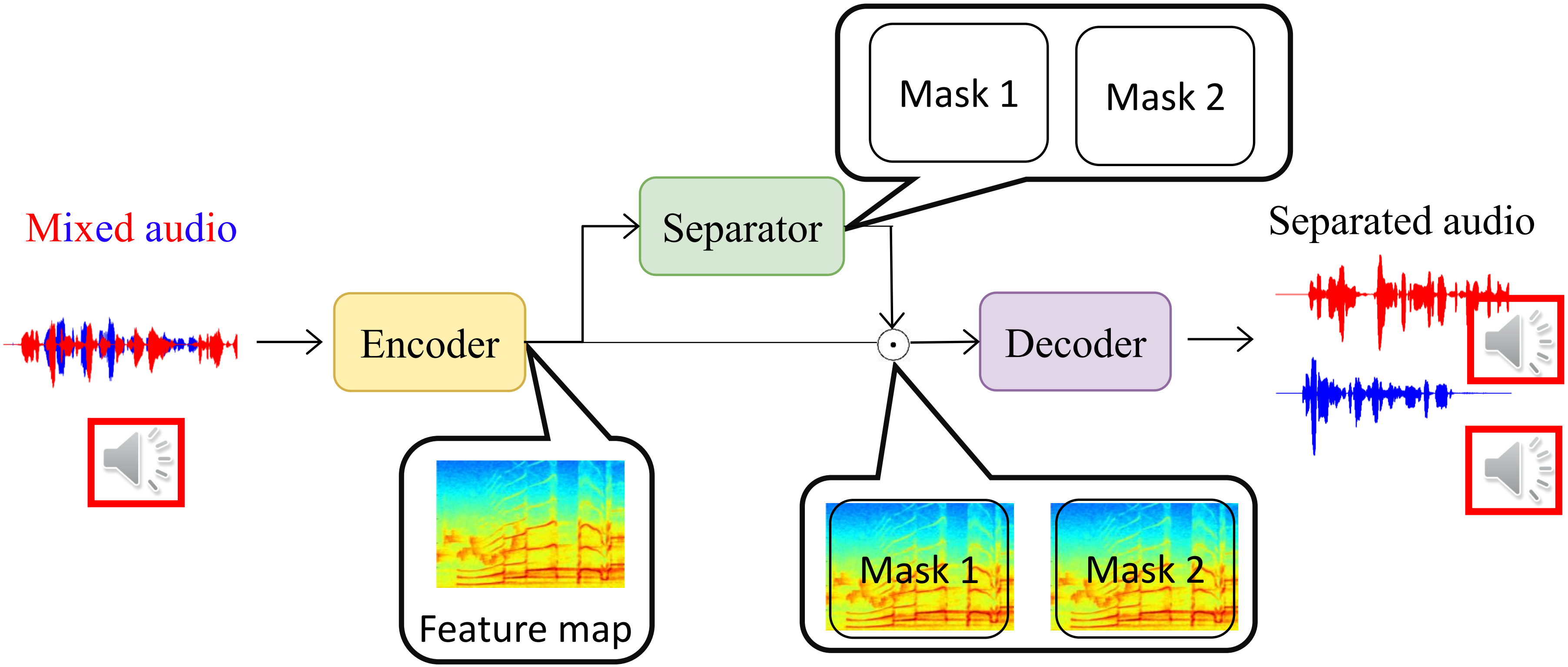
PIT [Kolbæk, et al., TASLP'17]



[Yang, et al., ICASSP'20]

TasNet – Time-domain Audio Separation Network

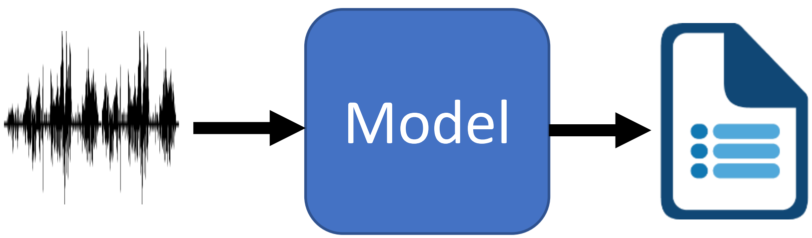
[Luo, et al., TASLP'19]

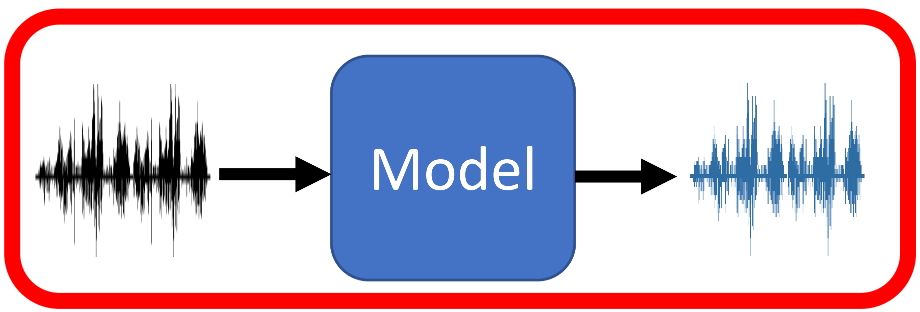


To learn more ……

- Denoise Wavnet [Rethage, et al., ICASSP’18]
- Chimera++ [Wang, et al., ICASSP’18]
- Phase Reconstruction Model [Wang, et al., ICASSP’19]
- Deep Complex U-Net: Complex masking [Choi, et al., ICLR’19]
- Deep CASA: Make CASA great again! [Liu, et al., TASLP’19]
- Wavesplit: state-of-the-art on benchmark corpus WSJ0-2mix [Zeghidour, et al., arXiv’20]

One slide for this course

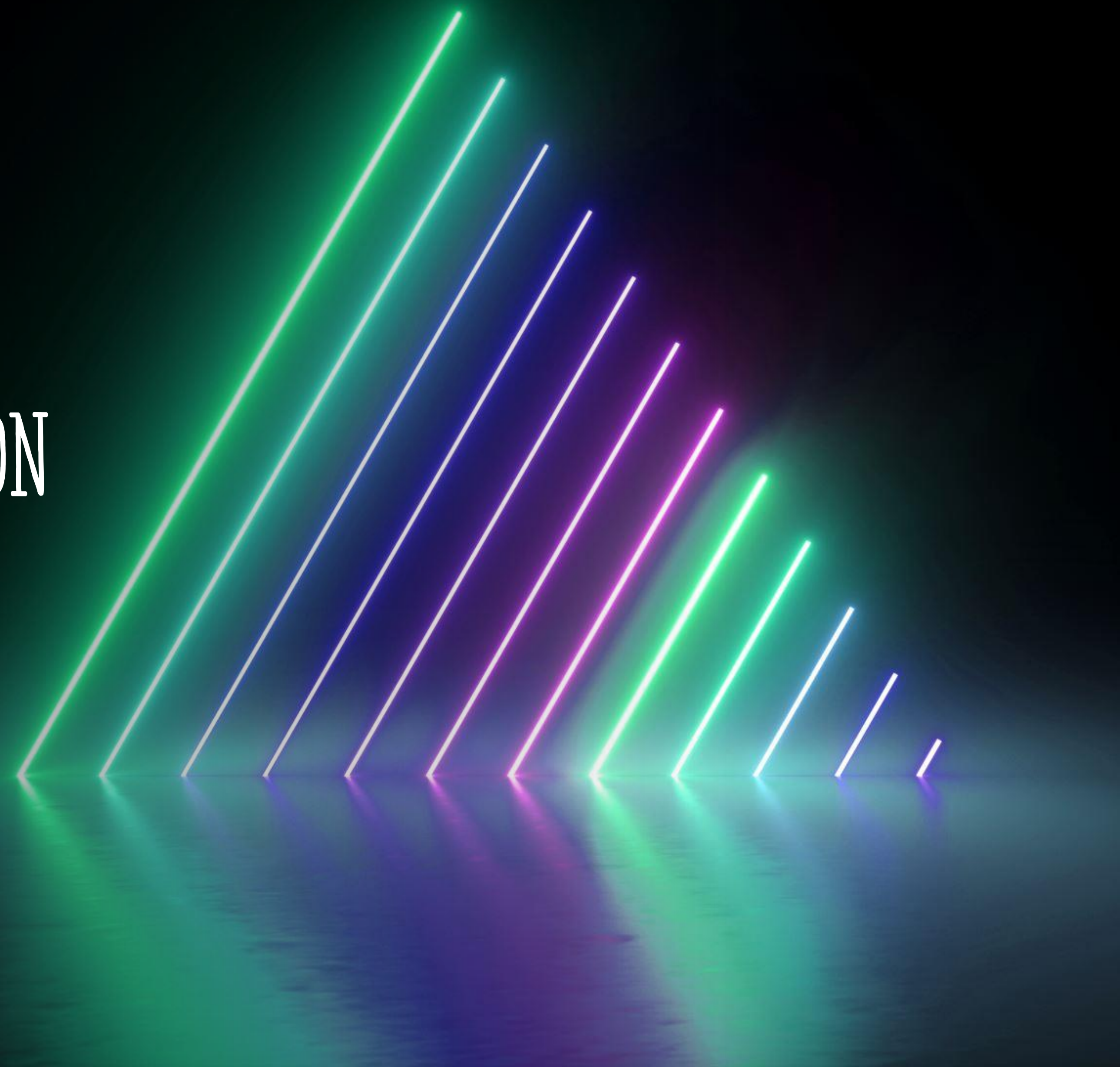




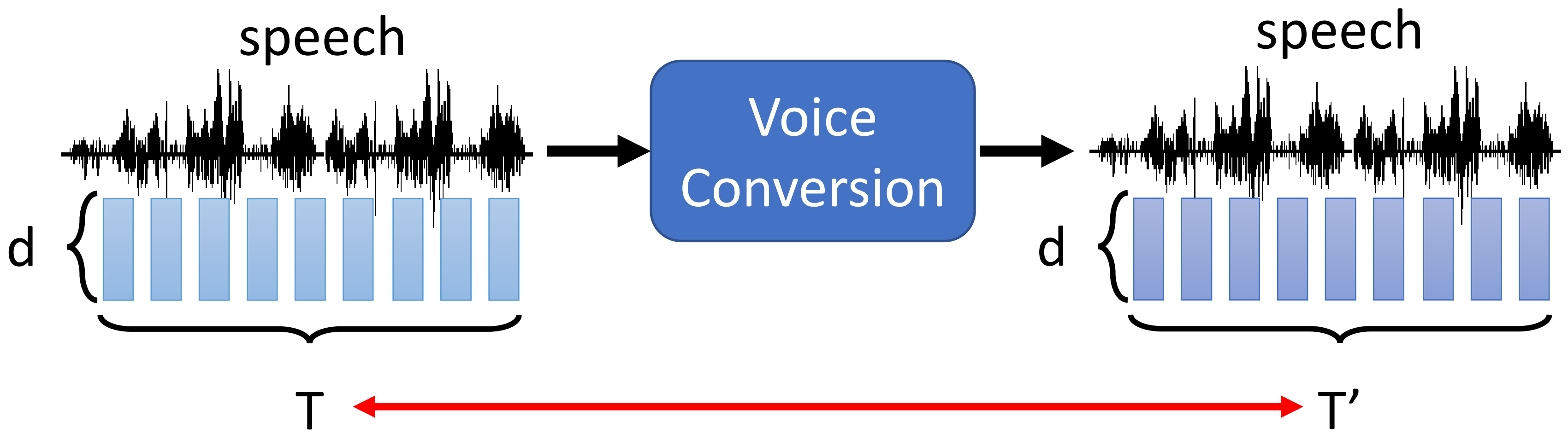
Speech and text can be represented as sequence.

Training a seq-to-seq network

VOICE CONVERSION



What is Voice Conversion (VC)?

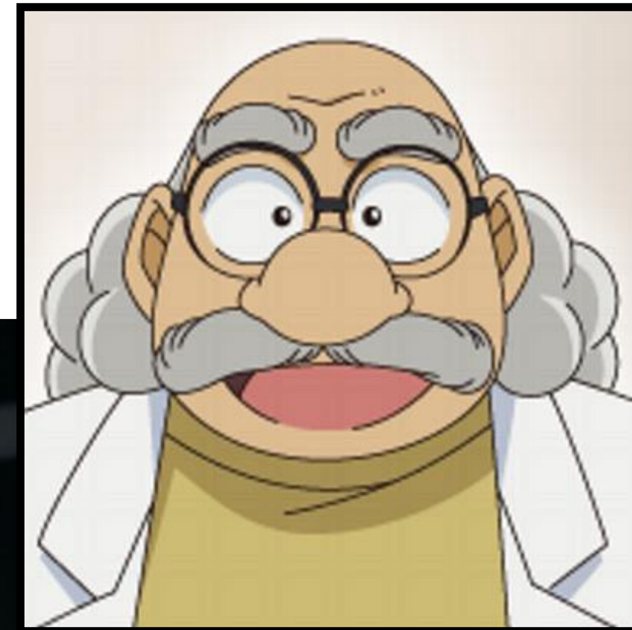


What is preserved? Content

What is changed? Many different aspects …

Speaker

Agasa
Hiroshi



Speaking Style

- Emotion
  [Gao, et al., INTERSPEECH'19]
- Normal-to-Lombard
  [Seshadri, et al., ICASSP'19]
- Whisper-to-Normal
  [Patel, et al., SSW'19]
- Singers vocal technique conversion
  [Luo, et al., ICASSP'20]

Normal Lombard

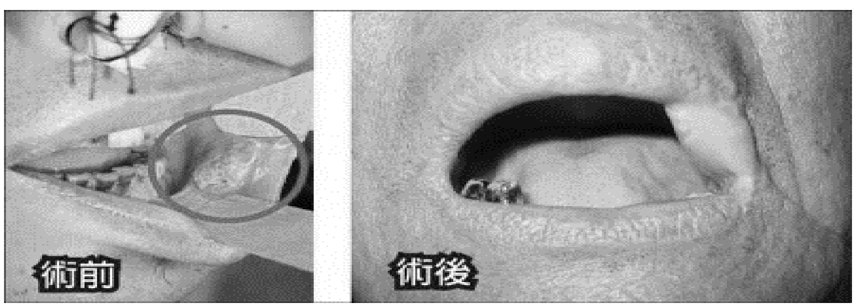
Source of audio: https://shreyas253.github.io/SpStyleConv_CycleGAN/

'lip thrill' or 'vibrato'

Improving Intelligibility

- Improving the speech intelligibility
  - oral cancer (top five cancer for male in Taiwan)
  - surgical patients who have had parts of their articulators removed

[Biadsy, et al., INTERSPEECH’19][Chen et al., INTERSPEECH’19]

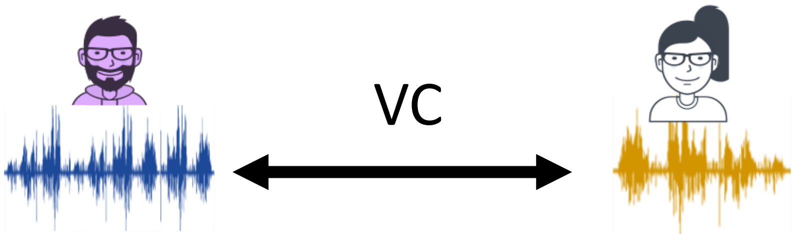


Before After

Before

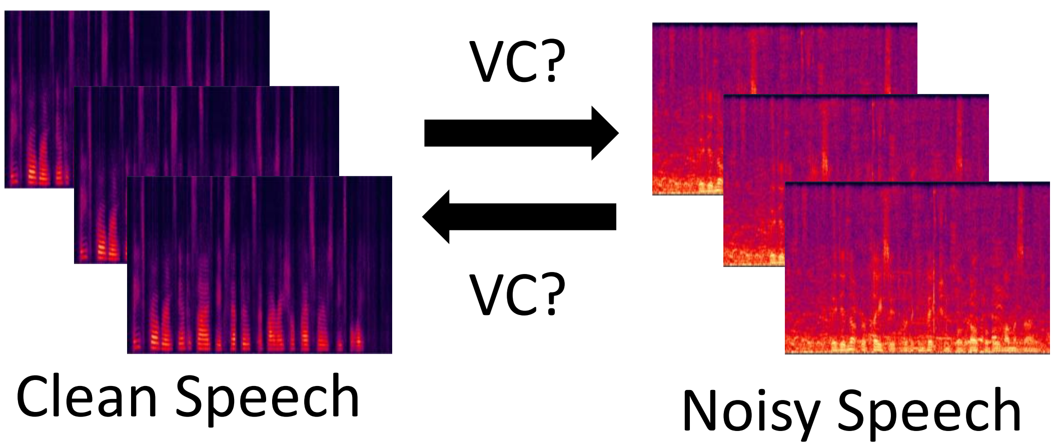
After

Data Augmentation



Training
Data x 2

[Keskin, et al., ICML workshop’19]



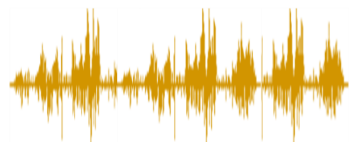
[Mimura, et al., ASRU 2017]

Categories

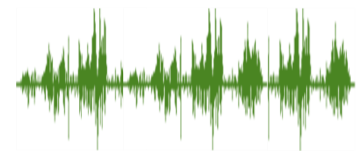
Lack of training data:

- Model Pre-training [Huang, et al., arXiv'19]
- Synthesized data! [Biadsy, et al., INTERSPEECH'19]

Parallel Data



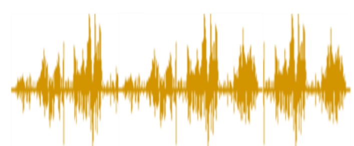
How are you?



How are you?

- This is "*audio style transfer*"
- Borrowing techniques from image style transfer

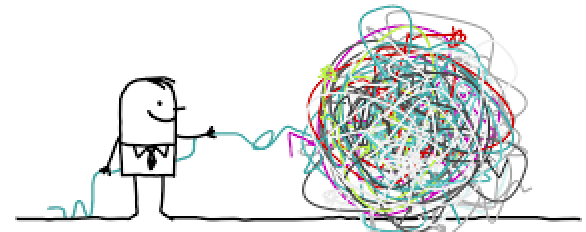
Unparallel Data



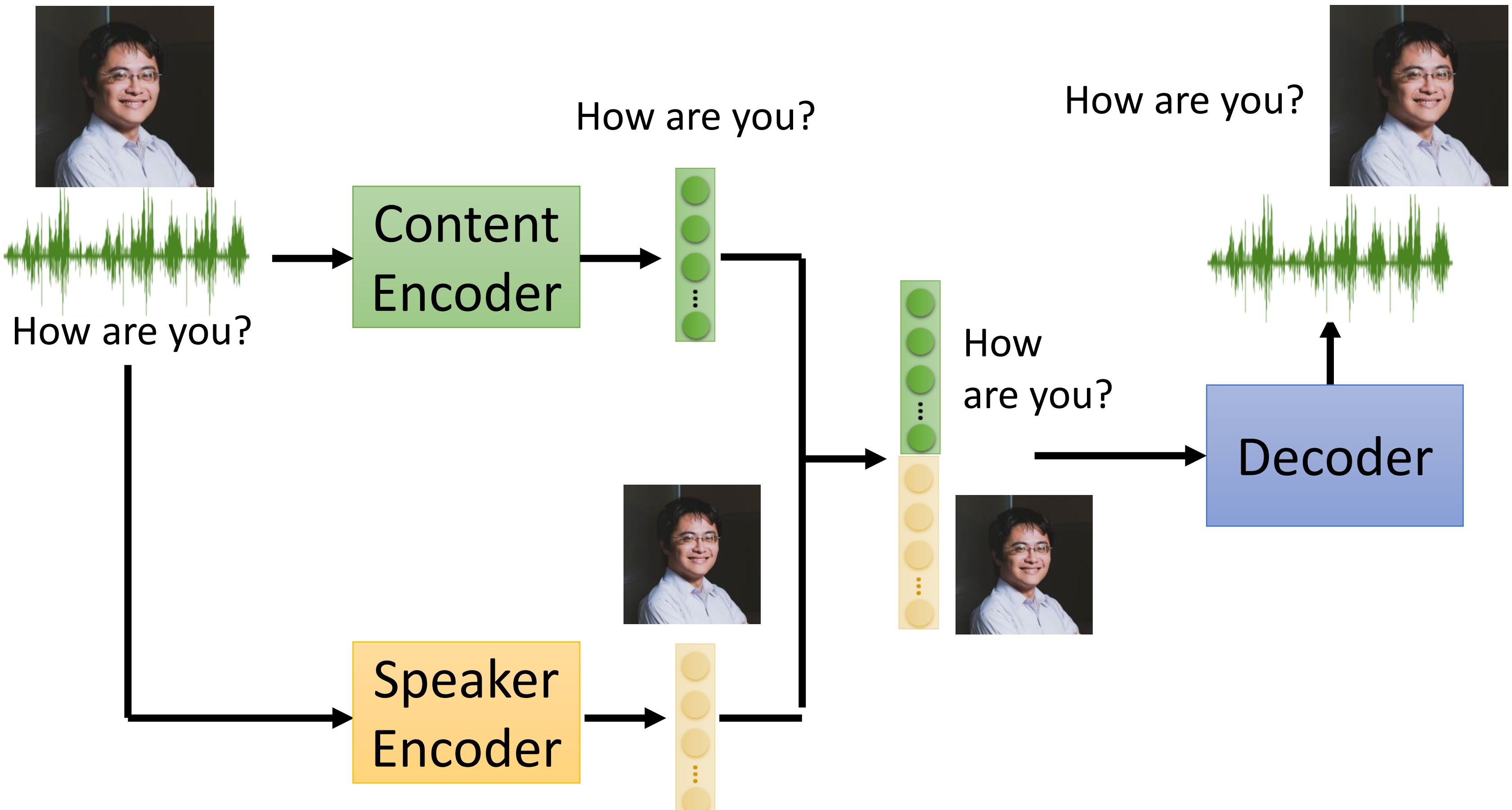
天氣真好

How are you?

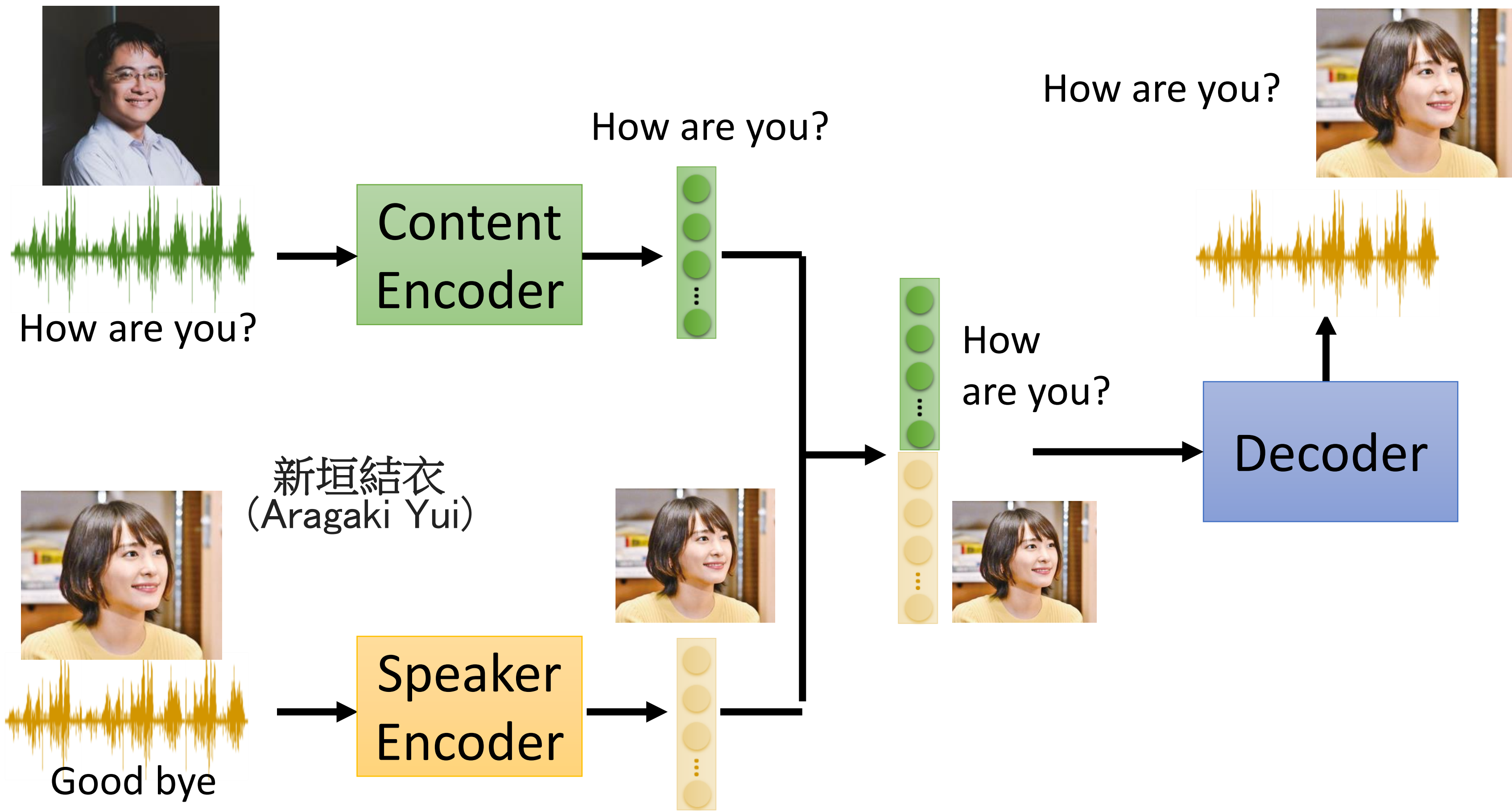
Feature Disentangle



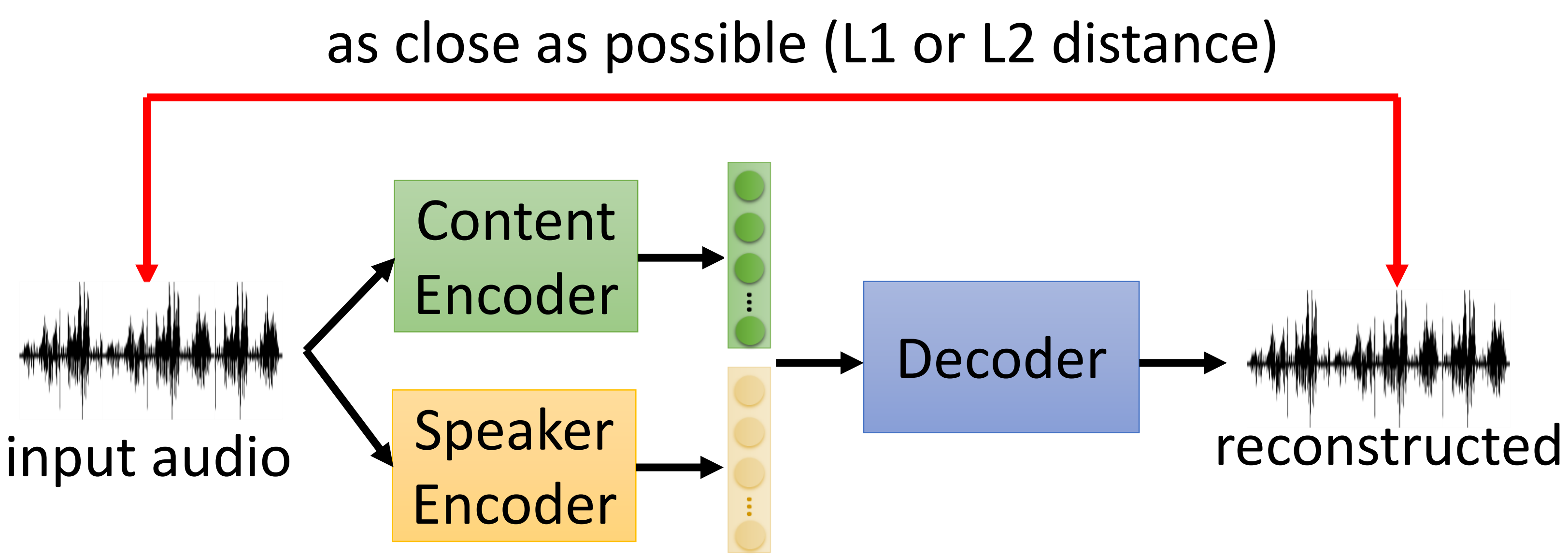
Source: https://www.dreamstime.com/illustration/disentangle.html



Feature Disentangle

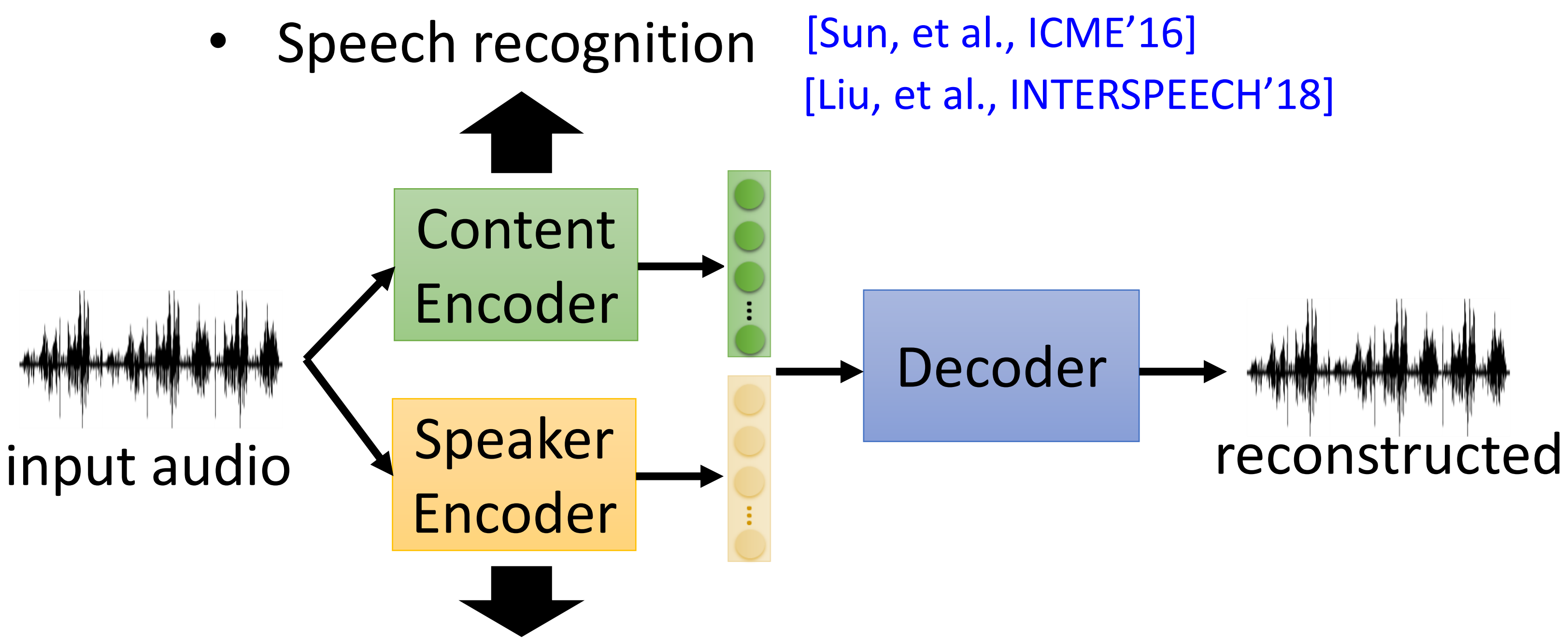


Feature Disentangle



How can you make one encoder for content and one for speaker?

1. Pre-training Encoders

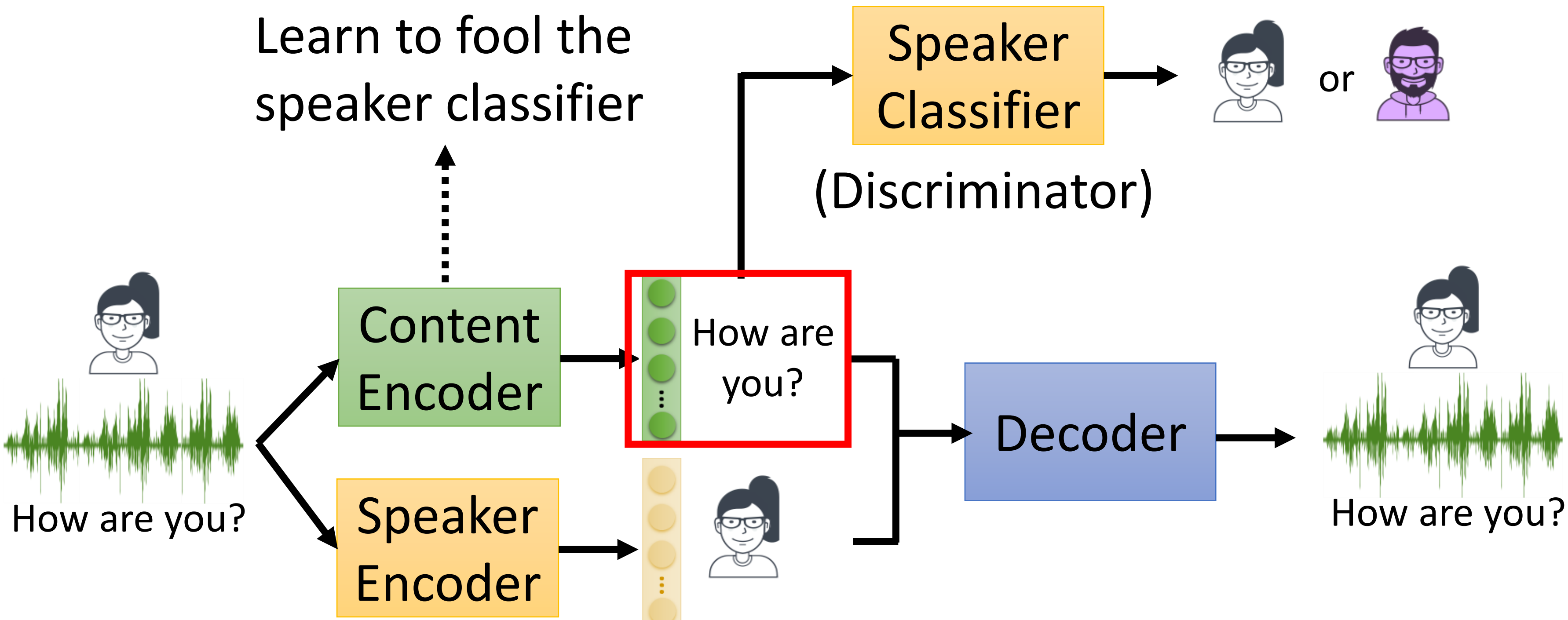


• Speaker embedding (i-vector, d-vector, x-vector …)

e.g., AutoVC [Qian, et al., ICML'19]

[Chou, et al., INTERSPEECH'18]

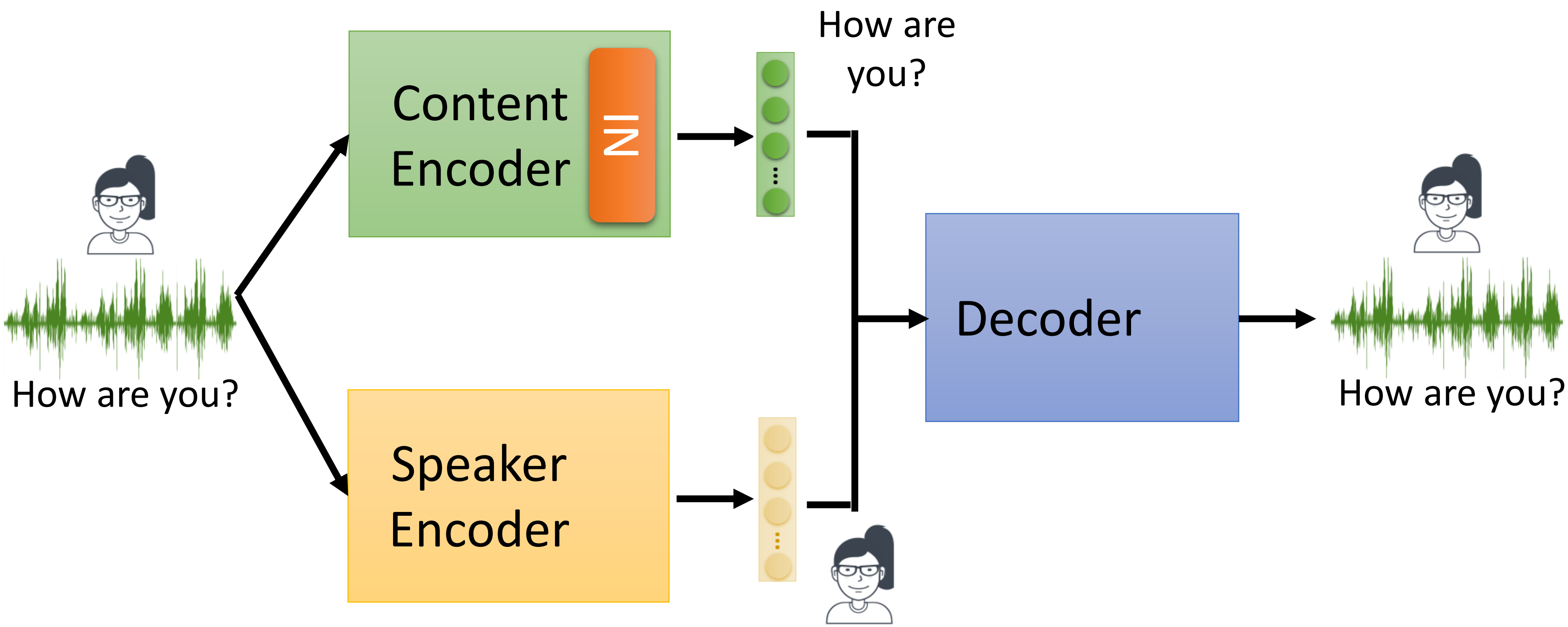
2. Adversarial Training



Speaker classifier and encoder are learned iteratively

Just as Generative Adversarial Network (GAN)

3. Designing network architecture

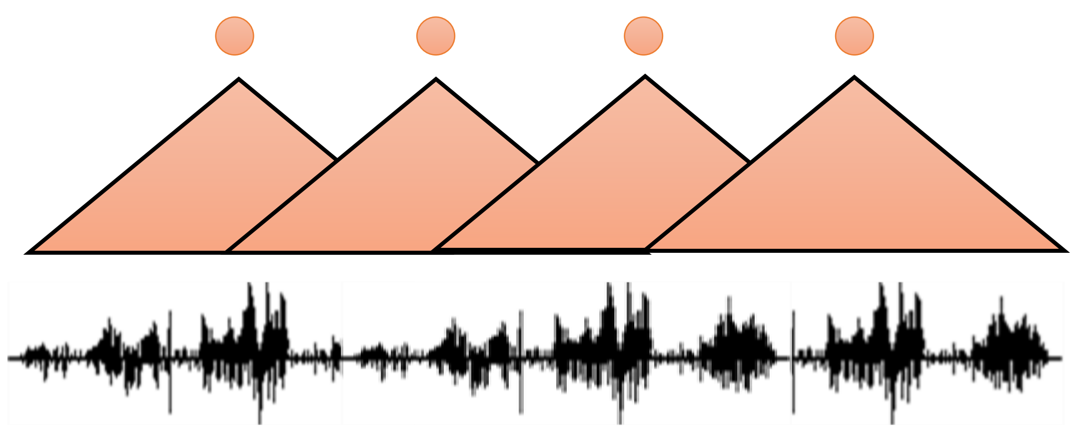


IN = instance normalization (remove speaker information)

3. Designing network architecture

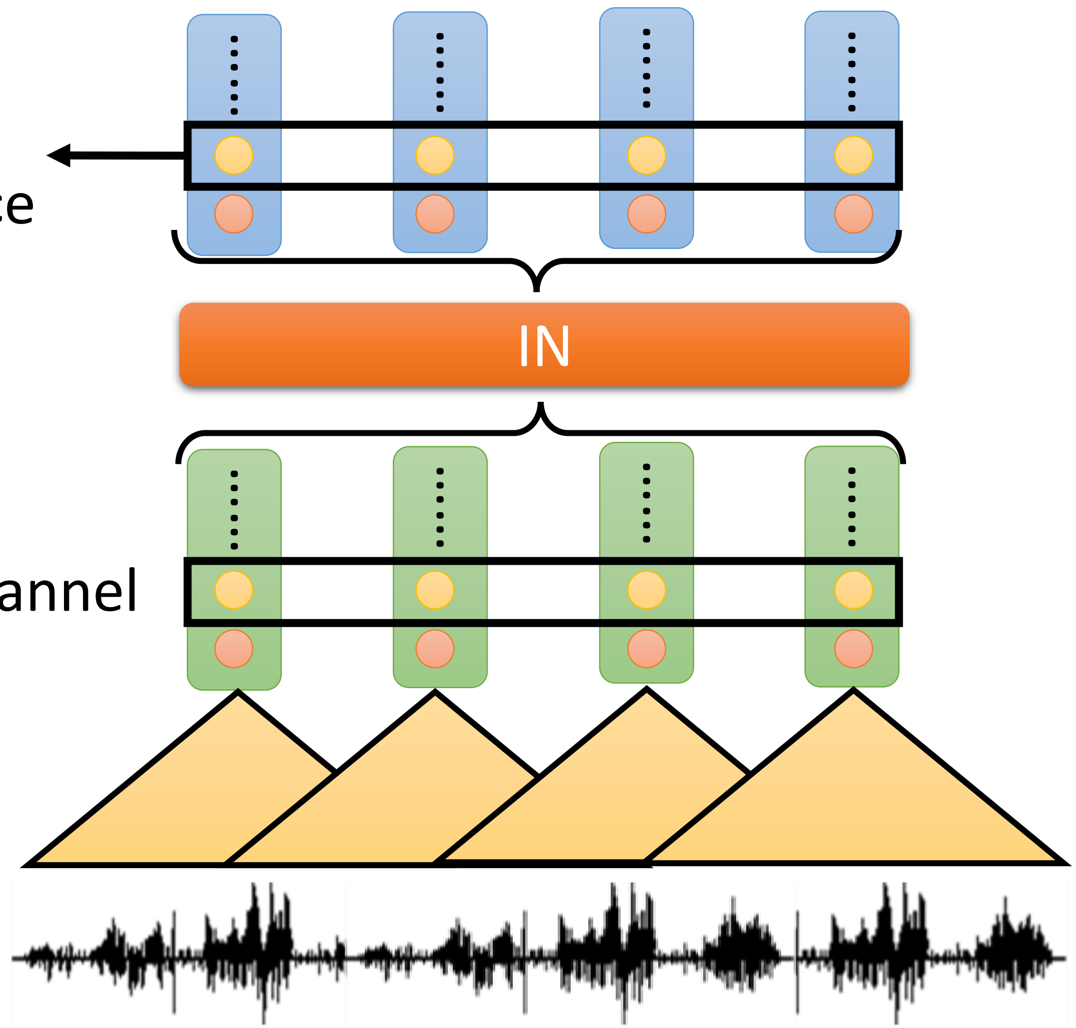
IN = instance normalization (remove speaker information)

Content Encoder



3. Designing network architecture

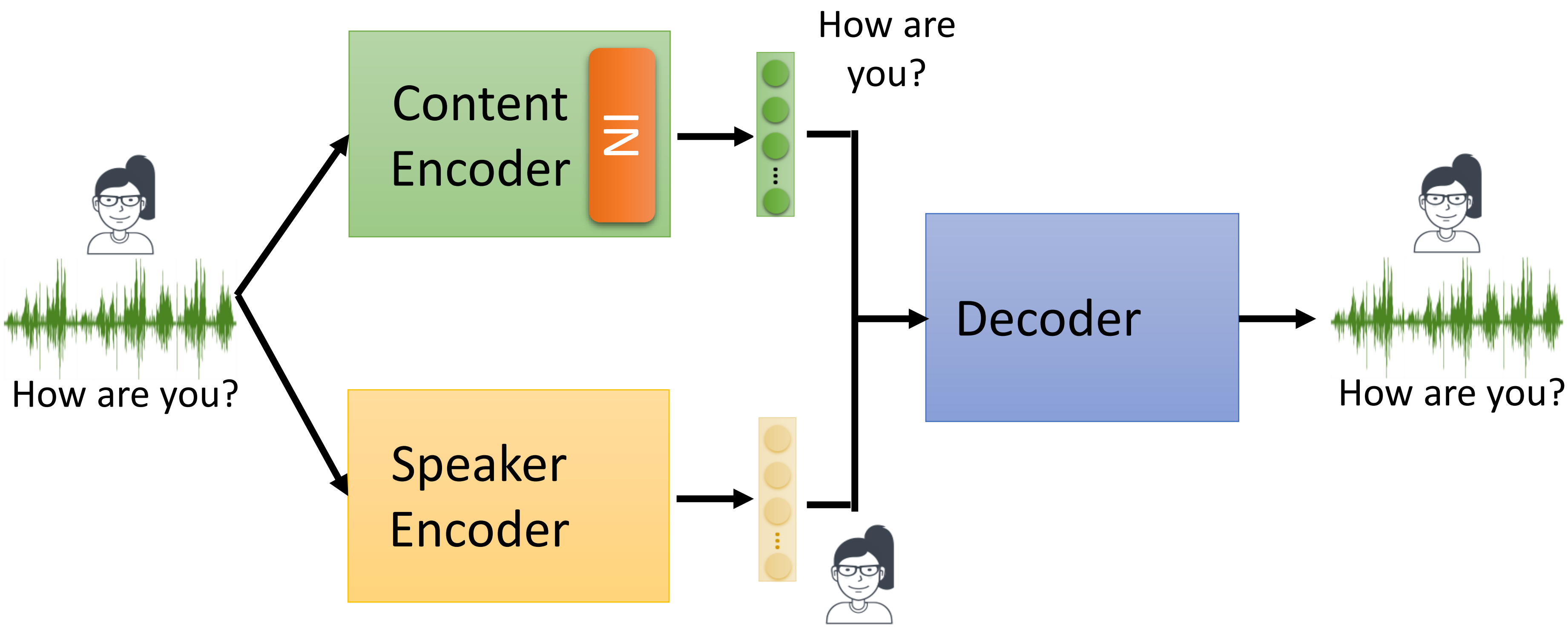
Each channel has zero mean and unit variance



Normalize for each channel

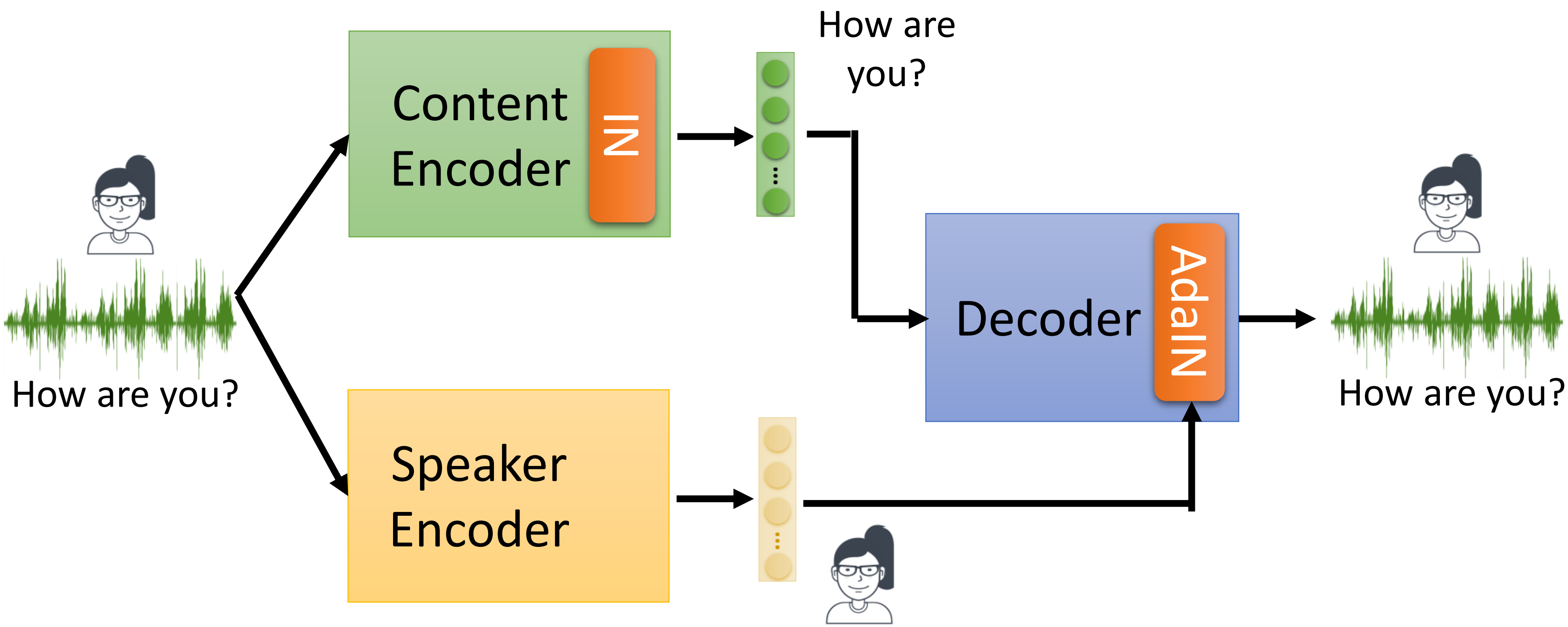
Content Encoder

3. Designing network architecture



IN = instance normalization (remove speaker information)

3. Designing network architecture



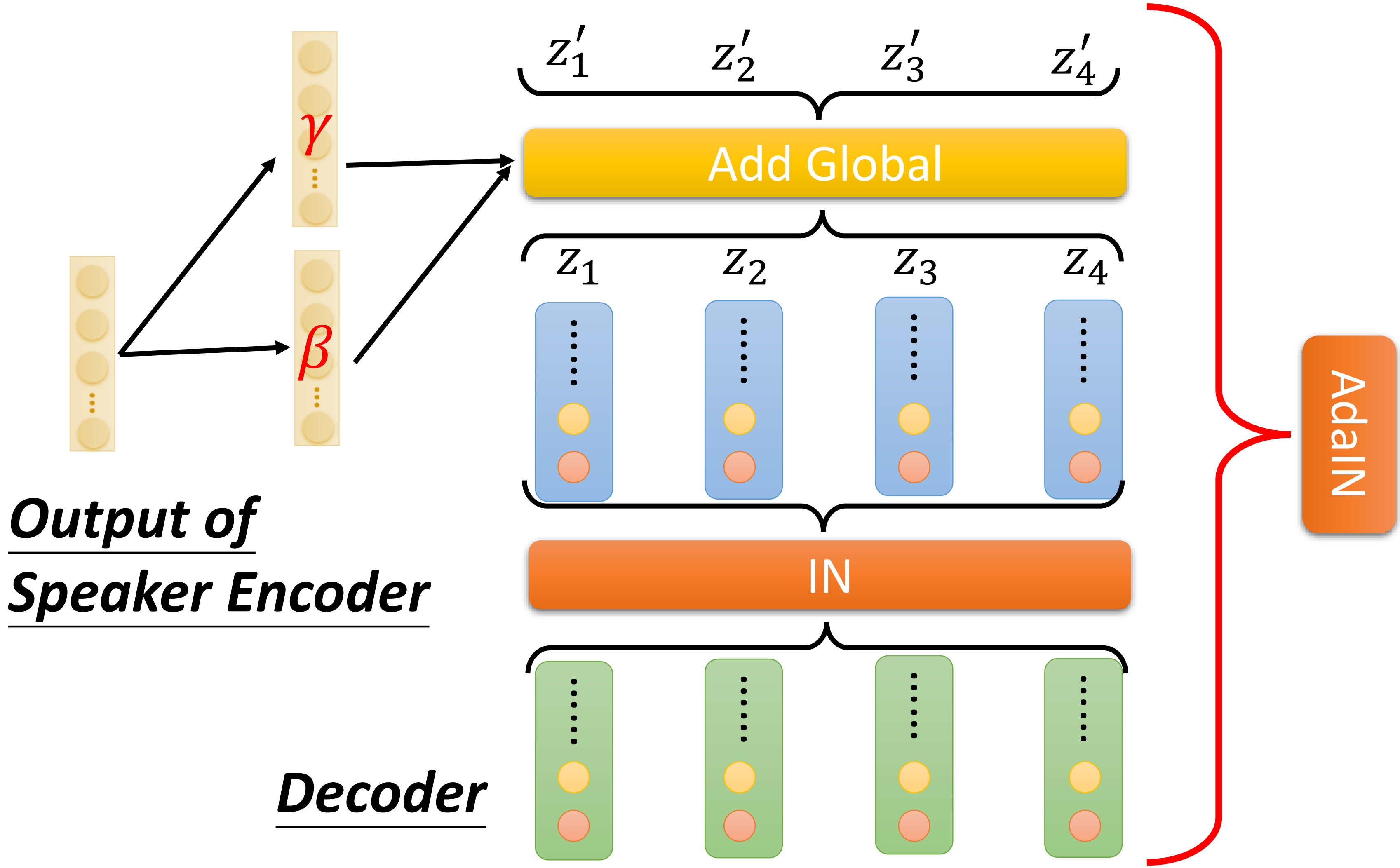
IN = instance normalization (remove speaker information)

AdaIN = adaptive instance normalization (only influence speaker information)

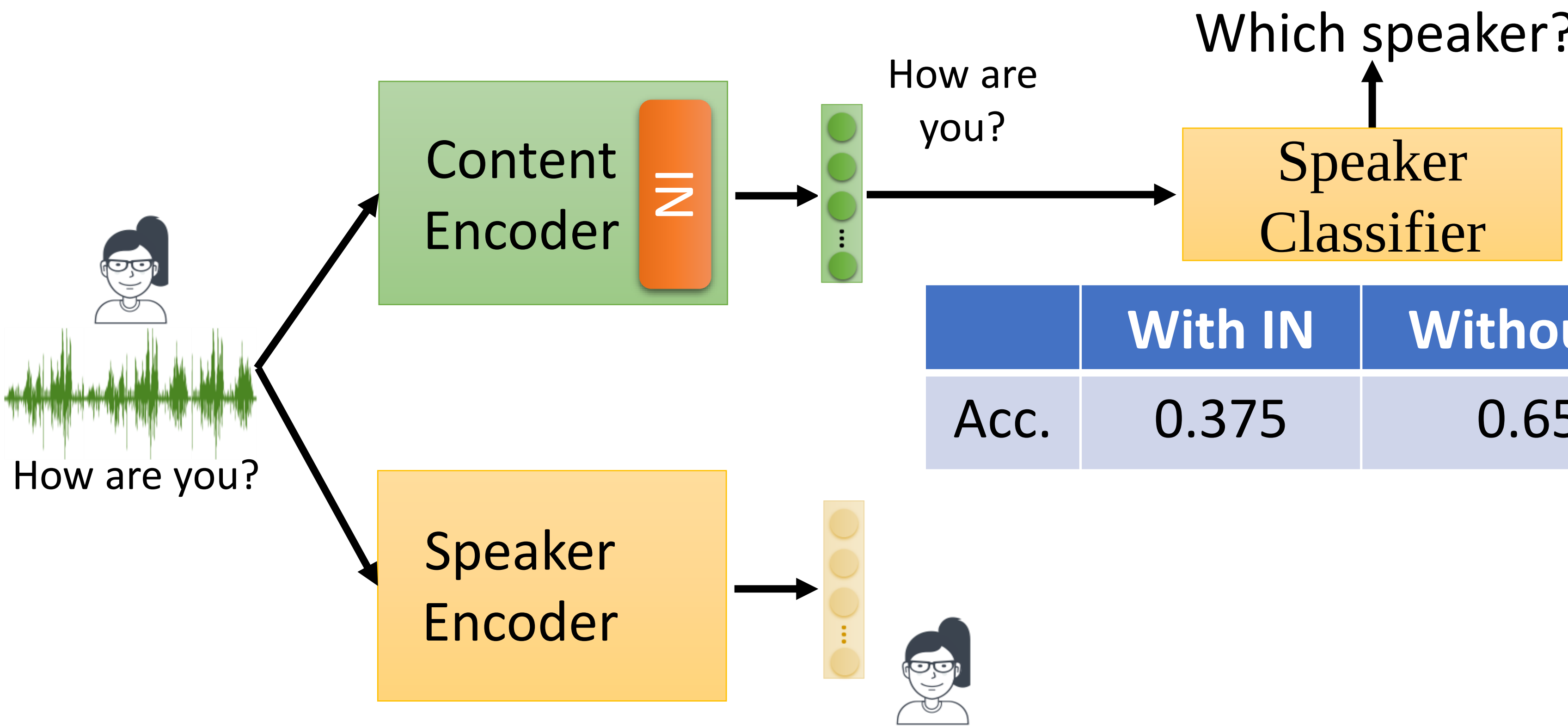
AdaIN = adaptive instance normalization

(only influence speaker information)

$$z_i' = \gamma \odot z_i + \beta$$



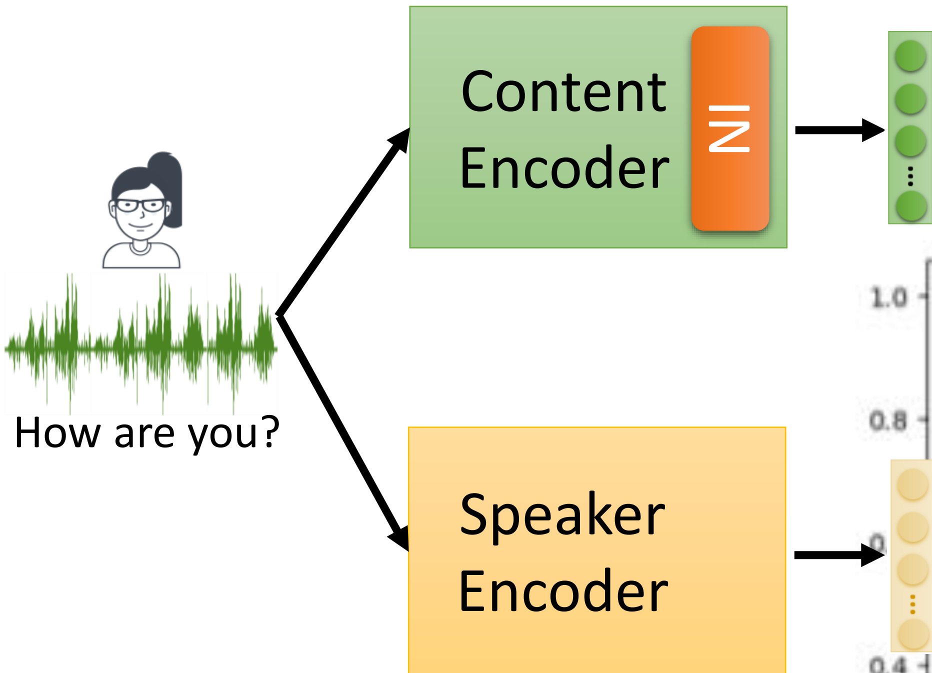
3. Designing network architecture



	With IN	Without IN
Acc.	0.375	0.658

Training from VCTK

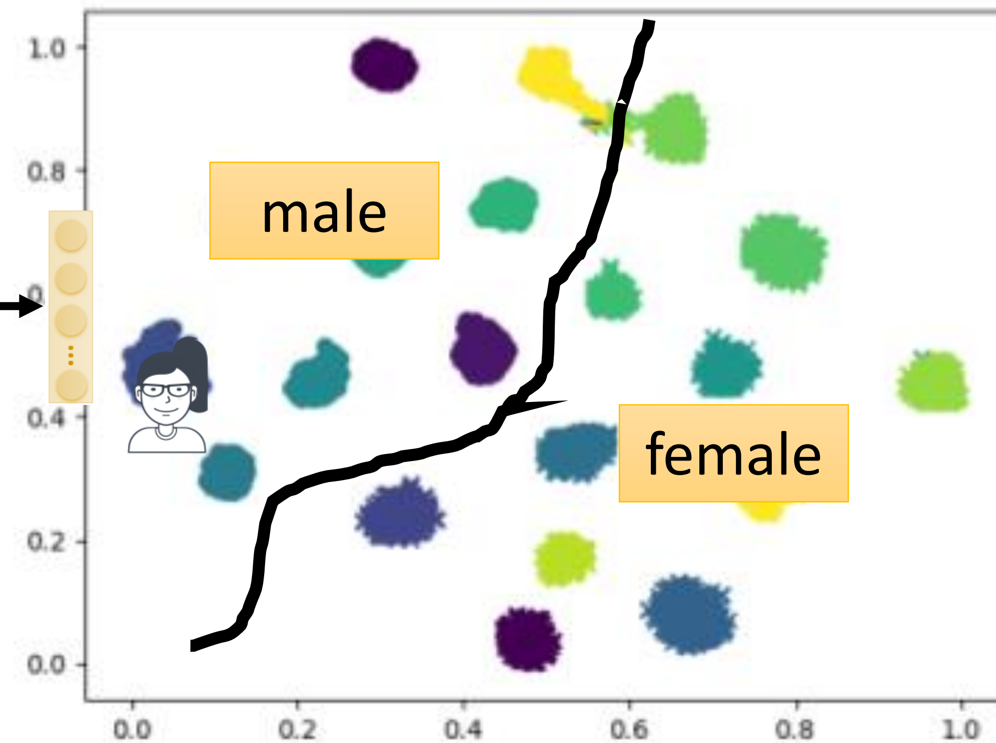
3. Designing network architecture



For more results

[Chou, et al., INTERSPEECH 2019]

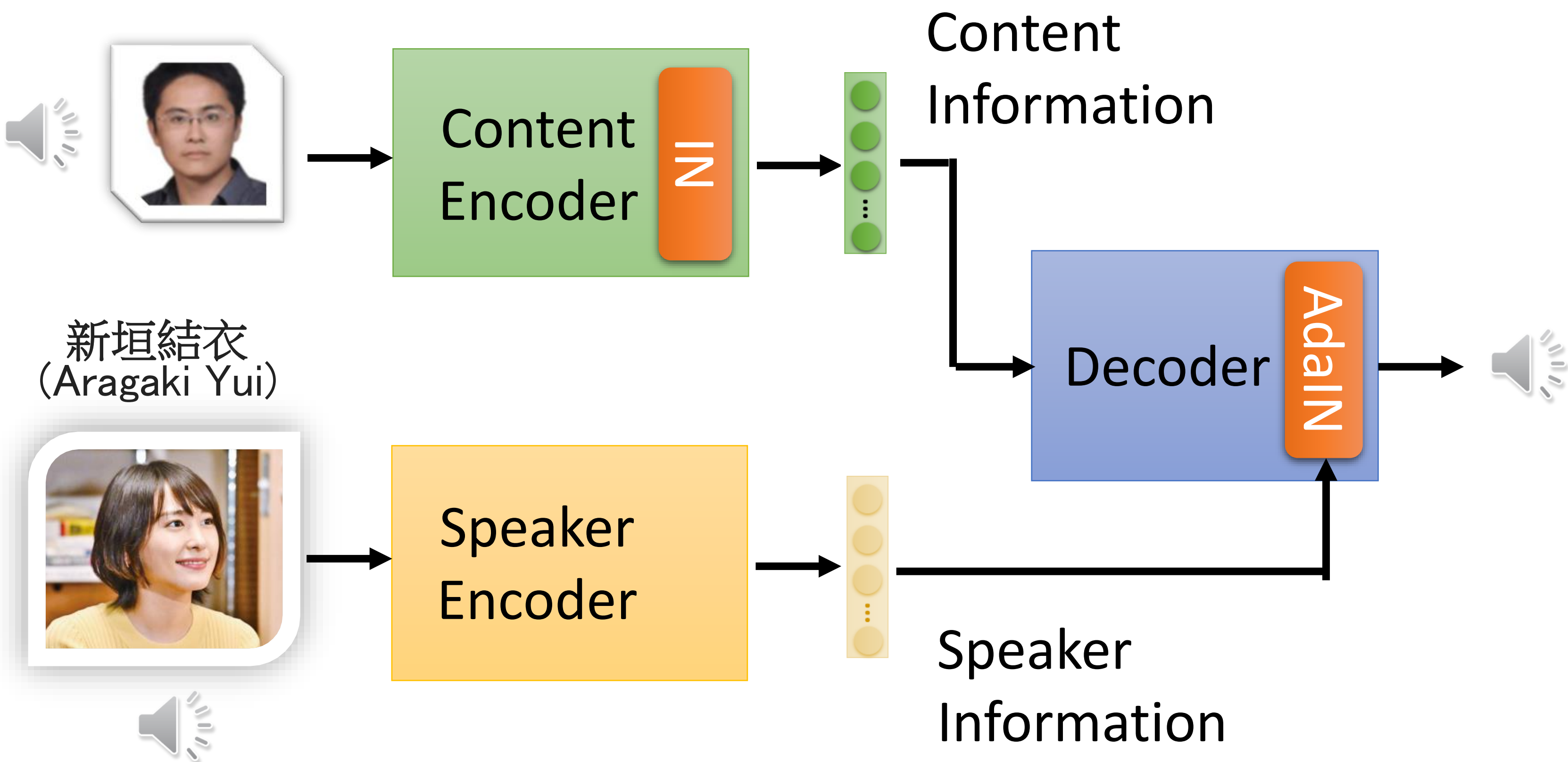
Unseen Speaker Utterances



Training from VCTK

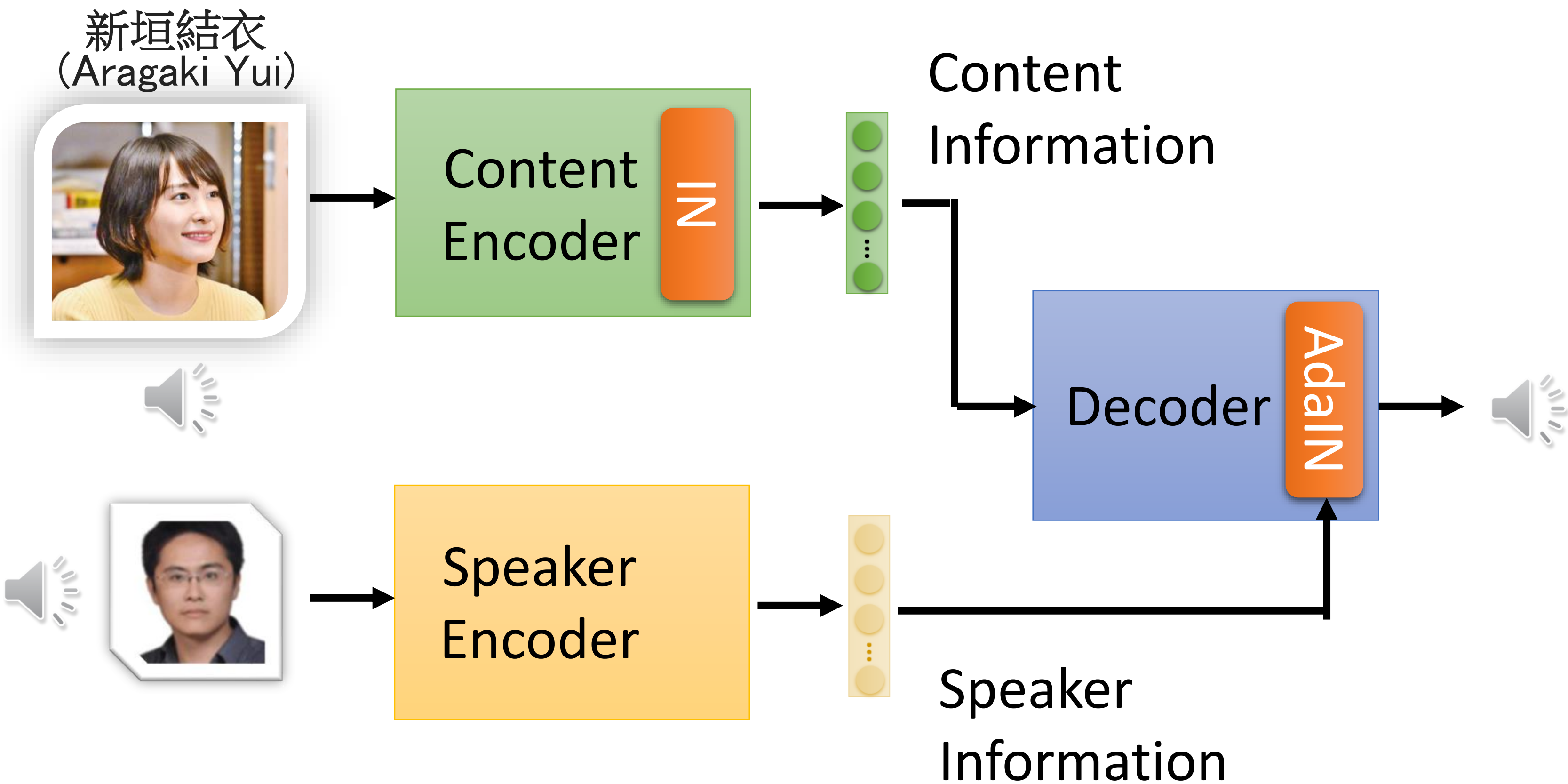
3. Designing network architecture

The speakers are **unseen** during training (**one-shot VC**).

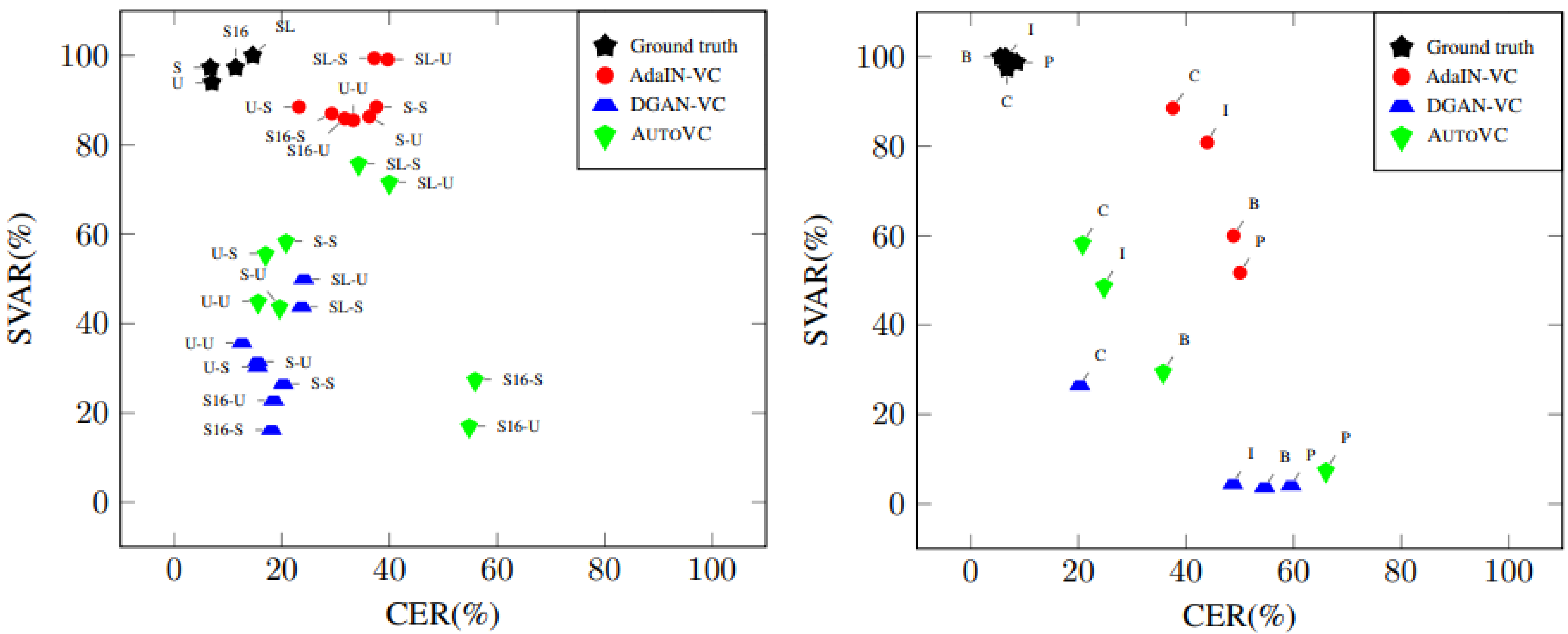


3. Designing network architecture

The speakers are **unseen** during training (**one-shot VC**).



How Far are we from Robust VC?



[Huang, et al., SLT’21]

One slide for this course

Speech and text can be represented as sequence.

Training a seq-to-seq network

If you are familiar with seq2seq, then you are ready to engage in speech technology.

(To be the top in the field, you need to understand more than seq2seq.)

Reference

- [Arik, et al., ICML’17] Sercan O. Arik, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, Shubho Sengupta, Mohammad Shoeybi, Deep Voice: Real-time Neural Text-to-Speech, ICML, 2017
- [Bahdanau. et al., ICASSP’16] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, Yoshua Bengio, End-to-End Attention-based Large Vocabulary Speech Recognition, ICASSP, 2016
- [Biadsy, et al., INTERSPEECH’19] Fadi Biadsy, Ron J. Weiss, Pedro J. Moreno, Dimitri Kanevsky, Ye Jia, Parrotron: An End-to-End Speech-to-Speech Conversion Model and its Applications to Hearing-Impaired Speech and Speech Separation, INTERSPEECH, 2019
- [Chan, et al., ICASSP’16] William Chan, Navdeep Jaitly, Quoc V. Le, Oriol Vinyals, Listen, Attend and Spell, ICASSP, 2016
- [Chen et al., INTERSPEECH’19] Li-Wei Chen, Hung-Yi Lee, Yu Tsao, Generative adversarial networks for unpaired voice transformation on impaired speech, INTERSPEECH, 2019

Reference

- [Chen, et al., INTERSPEECH'19] Kuan-yu Chen, Che-ping Tsai, Da-Rong Liu, Hung-yi Lee and Lin-shan Lee, "Completely Unsupervised Phoneme Recognition By A Generative Adversarial Network Harmonized With Iteratively Refined Hidden Markov Models", INTERSPEECH, 2019
- [Chiu, et al., ICLR'18] Chung-Cheng Chiu, Colin Raffel, Monotonic Chunkwise Attention, ICLR, 2018
- [Chiu, et al., ICASSP'18] Chung-Cheng Chiu, Tara N. Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J. Weiss, Kanishka Rao, Ekaterina Gonina, Navdeep Jaitly, Bo Li, Jan Chorowski, Michiel Bacchiani, State-of-the-art Speech Recognition With Sequence-to-Sequence Models, ICASSP, 2018
- [Choi, et al., ICLR'19] Hyeong-Seok Choi, Jang-Hyun Kim, Jaesung Huh, Adrian Kim, Jung-Woo Ha, Kyogu Lee, Phase-aware Speech Enhancement with Deep Complex U-Net, ICLR, 2019
- [Chorowski. et al., NIPS'15] Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, Yoshua Bengio, Attention-Based Models for Speech Recognition, NIPS, 15

Reference

- [Chou, et al., INTERSPEECH’18] Ju-chieh Chou, Cheng-chieh Yeh, Hung-yi Lee, Lin-shan Lee, "Multi-target Voice Conversion without Parallel Data by Adversarially Learning Disentangled Audio Representations", INTERSPEECH, 2018
- [Gao, et al., INTERSPEECH’19] Jian Gao, Deep Chakraborty, Hamidou Tembine, Olaitan Olaleye, Nonparallel Emotional Speech Conversion, INTERSPEECH, 2019
- [Graves, ICML workshop’12] Alex Graves, Sequence Transduction with Recurrent Neural Networks, ICML workshop, 2012
- [Graves, et al., ICML’14] Alex Graves, Navdeep Jaitly, Towards end-to-end speech recognition with recurrent neural networks, ICML, 2014
- [Huang, et al., arXiv’19] Wen-Chin Huang,Tomoki Hayashi,Yi-Chiao Wu,Hirokazu Kameoka,Tomoki Toda, Voice Transformer Network: Sequence-to-Sequence Voice Conversion Using Transformer with Text-to-Speech Pretraining, arXiv, 2019
- [Huang, et al., SLT’21] Tzu-hsien Huang, Jheng-hao Lin, Chien-yu Huang, Hung-yi Lee, How Far Are We from Robust Voice Conversion: A Survey, SLT, 2021
- [Jaitly, et al., NIPS’16] Navdeep Jaitly, Quoc V. Le, Oriol Vinyals, Ilya Sutskever, David Sussillo, Samy Bengio, An Online Sequence-to-Sequence Model Using Partial Conditioning, NIPS, 2016

Reference

- [Keskin, et al., ICML workshop'19] Gokce Keskin, Tyler Lee, Cory Stephenson, Oguz H. Elibol, Measuring the Effectiveness of Voice Conversion on Speaker Identification and Automatic Speech Recognition Systems, ICML workshop, 2019
- [Kolbæk, et al., TASLP'17] Morten Kolbæk, Dong Yu, Zheng-Hua Tan, Jesper Jensen, Multi-talker Speech Separation with Utterance-level Permutation Invariant Training of Deep Recurrent Neural Networks, TASLP, 2017
- [Li, et al., ICASSP'19] Bo Li, Yu Zhang, Tara Sainath, Yonghui Wu, William Chan, Bytes are All You Need: End-to-End Multilingual Speech Recognition and Synthesis with Bytes, ICASSP 2019
- [Liu, et al., INTERSPEECH'18] Songxiang Liu, Jinghua Zhong, Lifa Sun, Xixin Wu, Xunying Liu, Helen Meng, Voice Conversion Across Arbitrary Speakers Based on a Single Target-Speaker Utterance, INTERSPEECH, 2018
- [Liu, et al., INTERSPEECH'18] Da-Rong Liu, Kuan-Yu Chen, Hung-Yi Lee, Lin-shan Lee, Completely Unsupervised Phoneme Recognition by Adversarially Learning Mapping Relationships from Audio Embeddings, INTERSPEECH, 2018

Reference

- [Liu, et al., TASLP’19] Yuzhou Liu, DeLiang Wang, Divide and Conquer: A Deep CASA Approach to Talker-Independent Monaural Speaker Separation, in *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* 2019
- [Lu, et al., INTERSPEECH’15] Liang Lu, Xingxing Zhang, Kyunghyun Cho, Steve Renals, A Study of the Recurrent Neural Network Encoder-Decoder for Large Vocabulary Speech Recognition, INTERSPEECH, 2015
- [Luo, et al., TASLP’19] Yi Luo, Nima Mesgarani, Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation, TASLP, 2019
- [Luo, et al., ICASSP‘20] Yin-Jyun Luo, Chin-Chen Hsu, Kat Agres, Dorien Herremans, Singing Voice Conversion with Disentangled Representations of Singer and Vocal Technique Using Variational Autoencoders, ICASSP, 2020
- [Mimura, et al., ASRU 2017]Masato Mimura, Shinsuke Sakai, and Tatsuya Kawahara, Cross-domain Speech Recognition Using Nonparallel Corpora with Cycle-consistent Adversarial Networks, ASRU, 2017

Reference

- [Patel, et al., SSW'19] Maitreya Patel, Mihir Parmar, Savan Doshi, Nirmesh Shah and Hemant A. Patil, Novel Inception-GAN for Whisper-to-Normal Speech Conversion, ISCA Speech Synthesis Workshop, 2019
- [Ping, et al., ICLR'18] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O. Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, John Miller, Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning, ICLR, 2018
- [Qian, et al., ICML'19] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, Mark Hasegawa-Johnson, AUTOVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss, ICML, 2019
- [Ren, et al., NeurIPS'19] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, Tie-Yan Liu, FastSpeech: Fast, Robust and Controllable Text to Speech, NeurIPS, 2019
- [Rethage, et al., ICASSP'18] Dario Rethage, Jordi Pons, Xavier Serra, A Wavenet for Speech Denoising, ICASSP, 2018

Reference

- [Sak, et al., INTERSPEECH’15] Haşim Sak, Andrew Senior, Kanishka Rao, Françoise Beaufays, Fast and Accurate Recurrent Neural Network Acoustic Models for Speech Recognition, INTERSPEECH, 2015
- [Sak, et al., INTERSPEECH’17] Haşim Sak, Matt Shannon, Kanishka Rao, Françoise Beaufays, Recurrent Neural Aligner: An Encoder-Decoder Neural Network Model for Sequence to Sequence Mapping, INTERSPEECH, 2017
- [Seshadri, et al., ICASSP’19] Shreyas Seshadri, Lauri Juvela, Junichi Yamagishi, Okko Räsänen, Paavo Alku,Cycle-consistent Adversarial Networks for Non-parallel Vocal Effort Based Speaking Style Conversion, *ICASSP, 2019*
- [Shen, et al., ICASSP’18] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, Yonghui Wu, Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions, ICASSP, 2018
- [Soltau, et al., ICASSP’14] Hagen Soltau, George Saon, Tara N. Sainath, Joint training of convolutional and non-convolutional neural networks, ICASSP, 2014

Reference

- [Sun, et al., ICME’16] Lifa Sun, Kun Li, Hao Wang, Shiyin Kang, Helen Meng, Phonetic posteriorgrams for many-to-one voice conversion without parallel data training, ICME, 2016
- [Wang, et al., INTERSPEECH’17] Yuxuan Wang, R.J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, Rif A. Saurous, Tacotron: Towards End-to-End Speech Synthesis, INTERSPEECH, 2017
- [Wang, et al., ICASSP’18] Zhong-Qiu Wang, Jonathan Le Roux, John R. Hershey, Alternative Objective Functions for Deep Clustering, ICASSP, 2018
- [Wang, et al., ICASSP’19] Zhong-Qiu Wang, Ke Tan, DeLiang Wang, Deep Learning Based Phase Reconstruction for Speaker Separation: A Trigonometric Perspective, ICASSP, 2019
- [Watanabe, et al., IEEE JSTSP’17] S. Watanabe, T. Hori, S. Kim, J. R. Hershey and T. Hayashi, "Hybrid CTC/Attention Architecture for End-to-End Speech Recognition," in IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 8, pp. 1240-1253, Dec. 2017

Reference

- [Yang, et al., ICASSP’20] Gene-Ping Yang, Szu-Lin Wu, Yao-Wen Mao, Hung-yi Lee, Lin-shan Lee, Interrupted and cascaded permutation invariant training for speech separation, ICASSP, 2020
- [Yeh, et al., ICLR’19] Chih-Kuan Yeh, Jianshu Chen, Chengzhu Yu, Dong Yu, Unsupervised Speech Recognition via Segmental Empirical Output Distribution Matching, ICLR, 2019
- [Yu, et al, arXiv’19] Chengzhu Yu, Heng Lu, Na Hu, Meng Yu, Chao Weng, Kun Xu, Peng Liu, Deyi Tuo, Shiyin Kang, Guangzhi Lei, Dan Su, Dong Yu, DurIAN: Duration Informed Attention Network For Multimodal Synthesis, arXiv, 2019
- [Zeghidour, et al., arXiv’20] Neil Zeghidour, David Grangier, Wavesplit: End-to-End Speech Separation by Speaker Clustering, arXiv, 2020

To learn more …

YouTube Channel teaching Deep Learning/Human Language Processing
(in Mandarin, more than 7M Total Views and 80k Subscribers)