Machine Speech Chain:
A Machine that Learned to Listen, Speak, and

while

LISTEN SPEAKING

Dr.-Ing. Sakriani Sakti
Research Associate Professor of
Nara Institute of Science and Technology (NAIST), Japan
Research Scientist of RIKEN Center for
Advanced Intelligence Project AIP (RIKEN AIP), Japan

Co-Authors: Andros Tjandra, Johanes Effendi, Sahoko Nakayama, Sashi Novitasari, Satoshi Nakamura

Human Communication

Human-to-Human Communication

- **Speech in Human Communication**

  → The most natural modality to express & share their ideas, experiences, and knowledge

Business

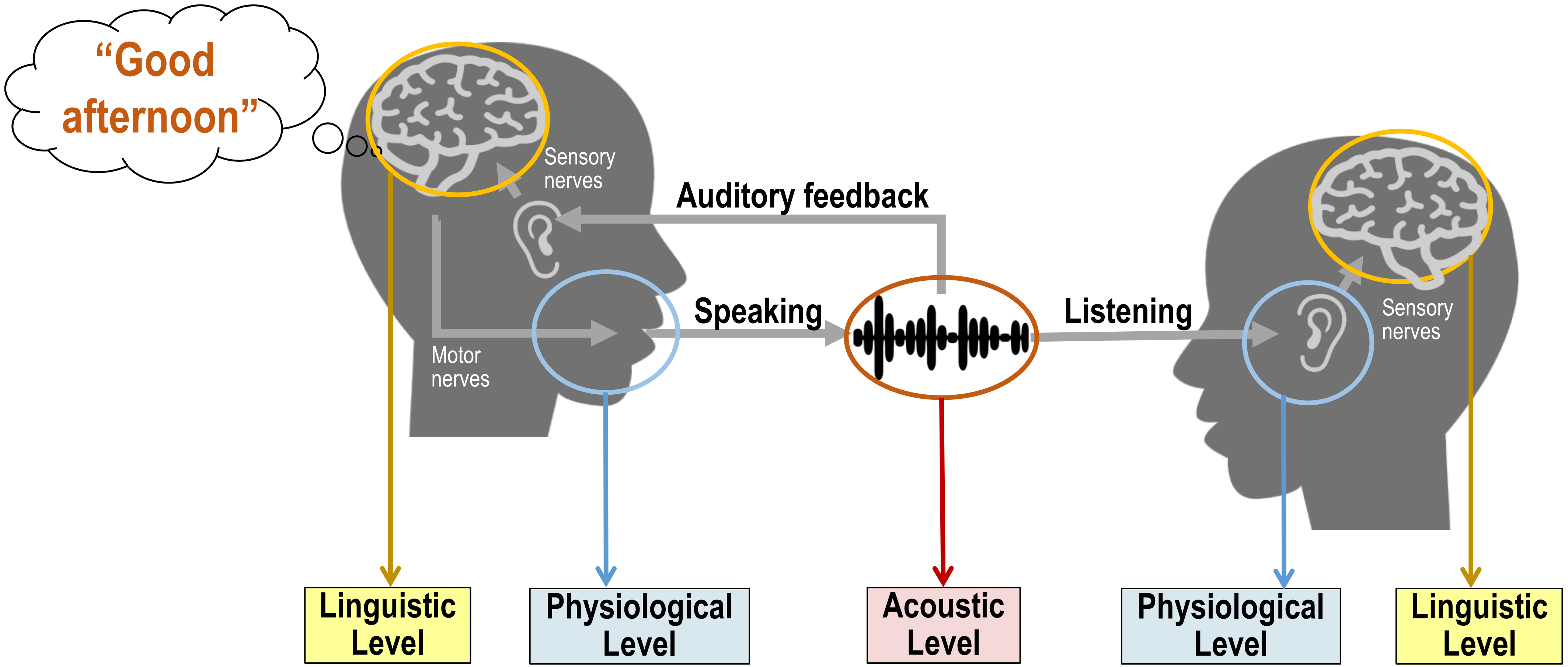
Meeting

Conversations

Lecture

How do We Communicate?

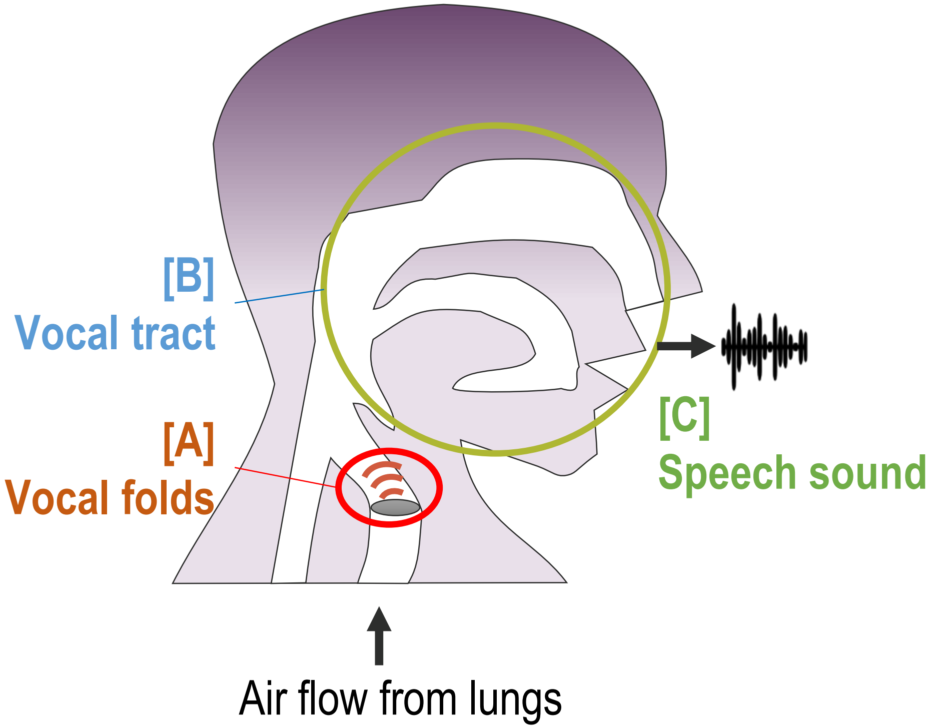
- **Speech Chain** [Denes & Pinson, 1993]



How do We Speak?

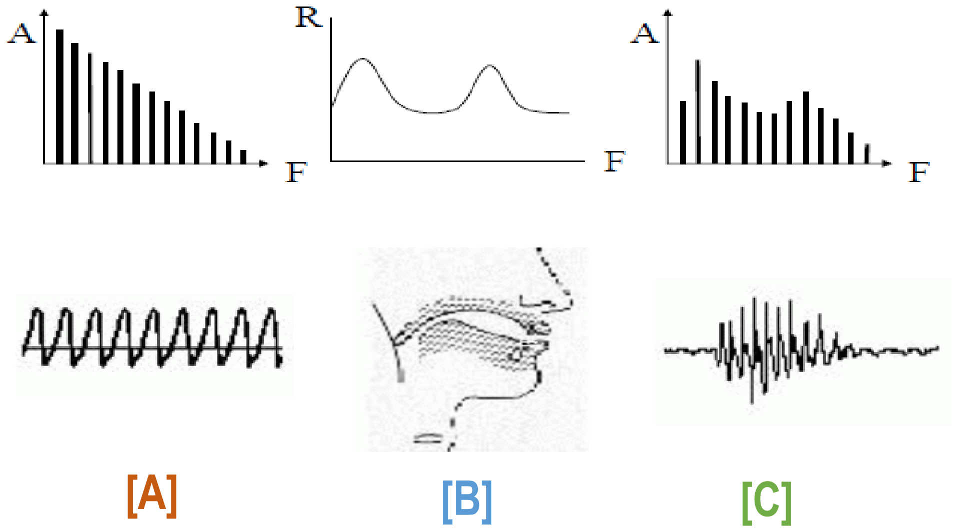
- **Speech Production Model**

  → By expelling air from the lungs through the trachea, and passed through the larynx then out the mouth or nose (vocal tract)

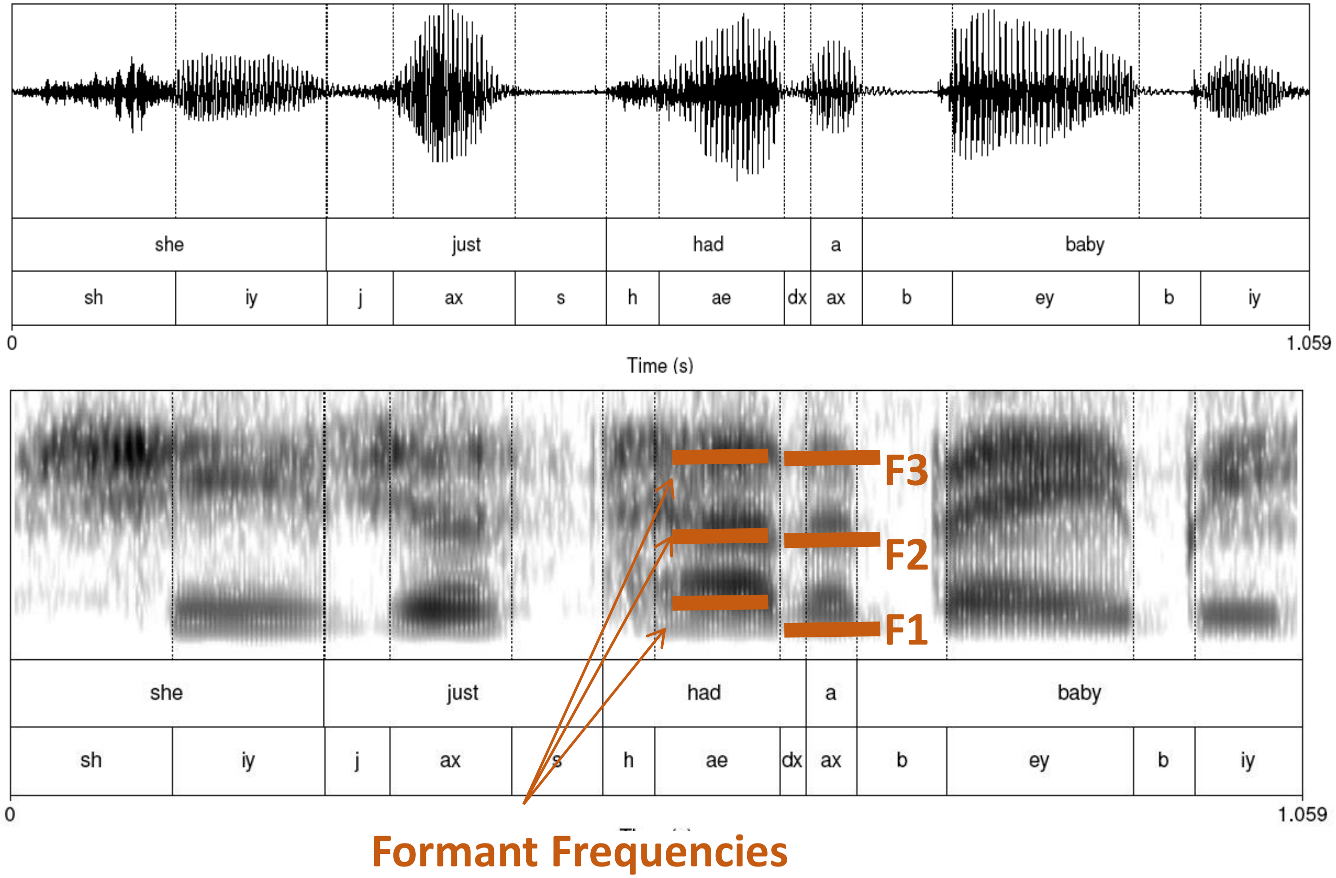


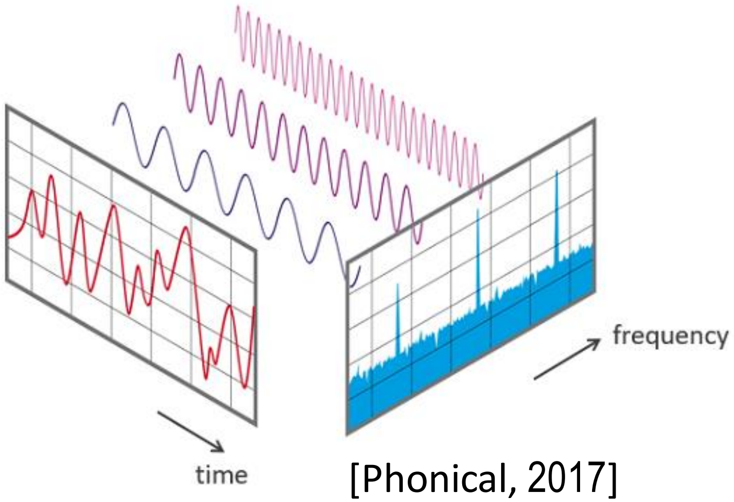
$h(t)$

Source (Glotal Pulse) $x(t)$ → **Vocal tract filter** → **Speech Sound** $y(t) = x(t) * h(t)$

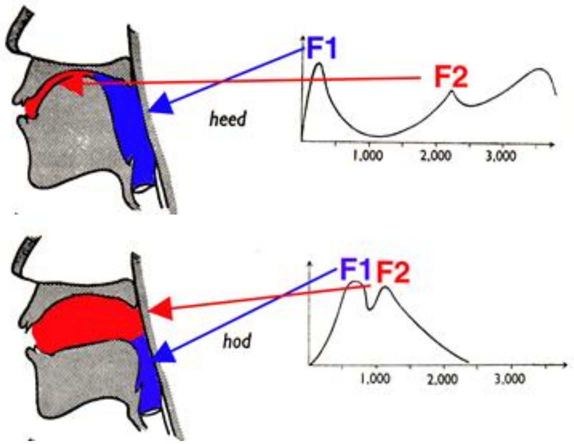


Speech Utterances





[Phonical, 2017]



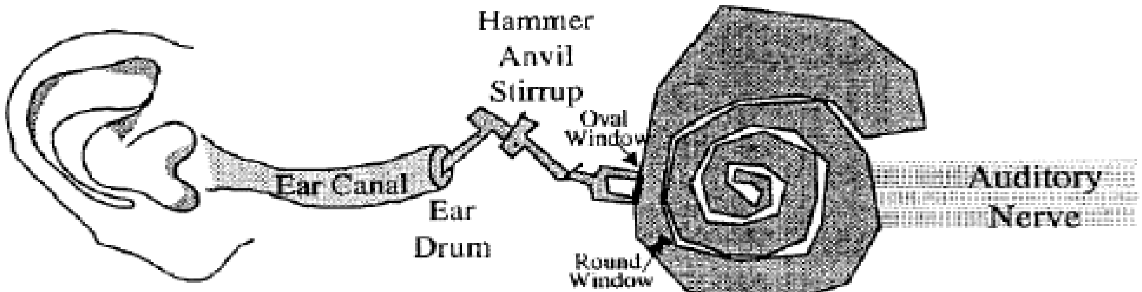
"She just had a baby"

[Source: https://web.stanford.edu/class/cs224s/lectures/224s.17.lec2.pdf]

How do We Hear?

- **Human Ear**

  → Receive the acoustic waves, amplify the intensity, & analyze the frequency



[Bosi & Goldberg, 2003]

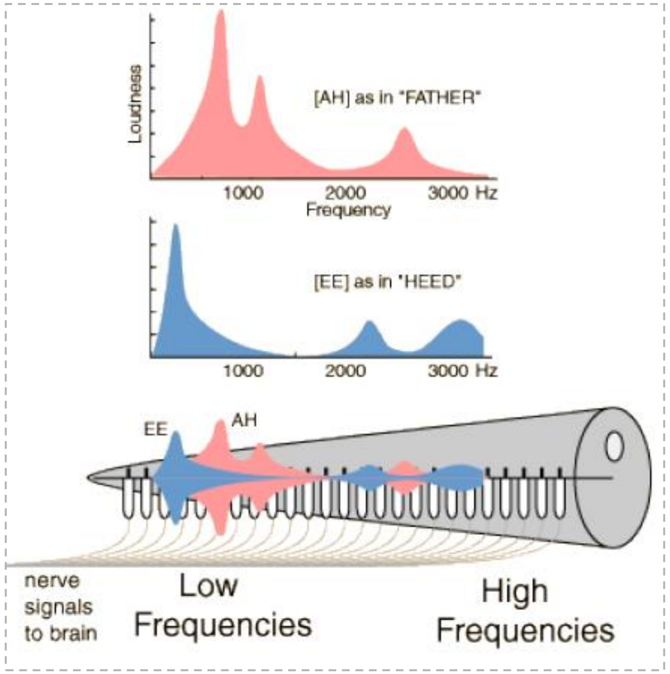
Middle Ear
Converts air movement in ear canal to fluid movement in cochlea.

Cochlea

Outer Ear
Collects sound and funnels it down to ear drum. Physical size tuned to sounds around 4 kHz.

Inner Ear
Cochlea separates sounds by frequency. Hair cells convert fluid motion into electrical impulses in auditory nerve.

The cochlea in inner ear



Separate Sound by Frequency

directional microphone

impedance matching, overload protection

neural encoding, frequency analysis

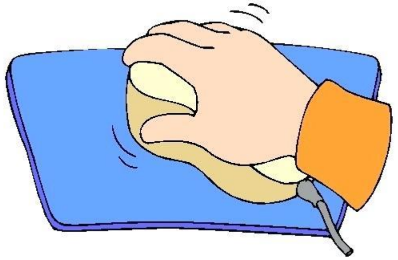
[Source: http://hyperphysics.phy-astr.gsu.edu/hbase/Sound/cochimp.html]

Human-Machine Interaction

Human-Machine Interaction

- **Modality in Human-Machine Interaction**

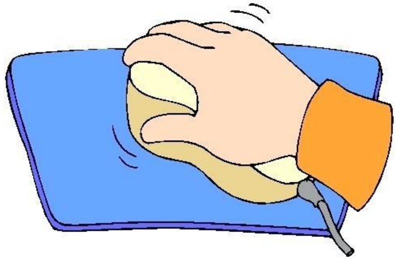
Communication	Channel
Input	
- Keyboard, mouse, touch screen	tactile
- Microphone	audio
- Scanner	visual
- Camera, Eye tracking, Gaze tracking	visual
Output	
- Display	visual
- Loudspeaker	audio

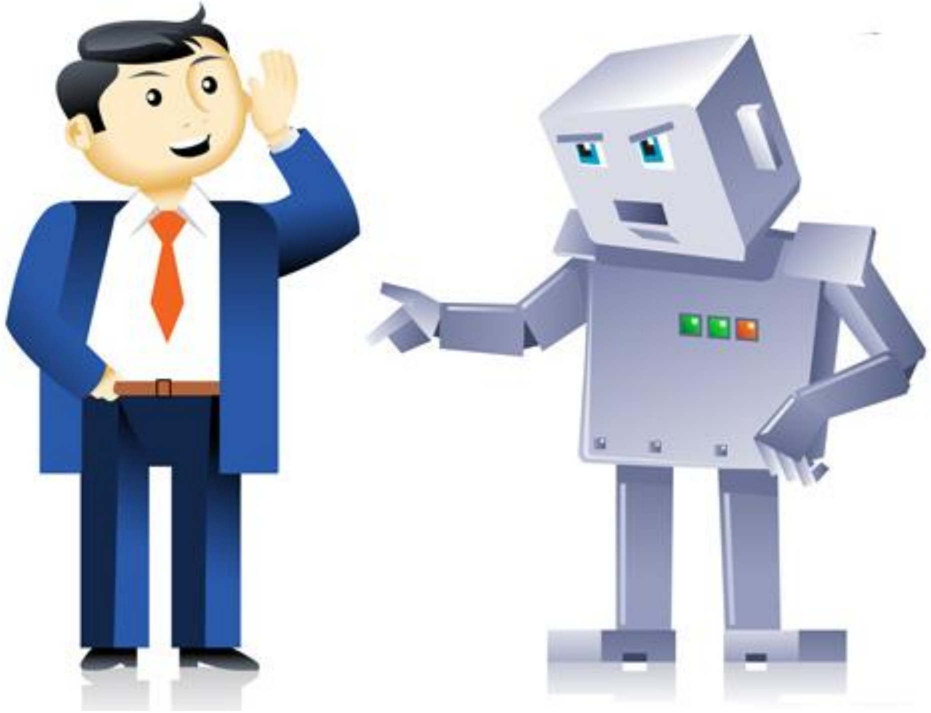


Human-Machine Interaction

- **Modality in Human-Machine Interaction**

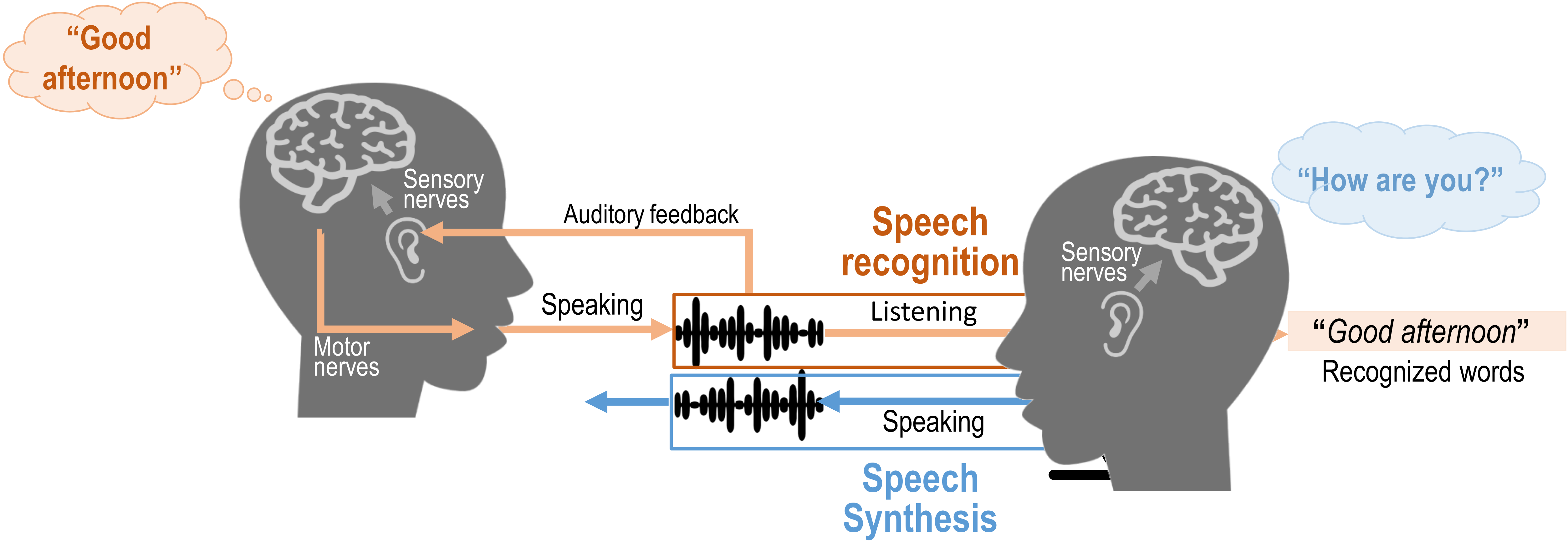
→ One of the earliest objectives in artificial intelligence (AI) has been to realize a technology or a machine that can communicate with the human





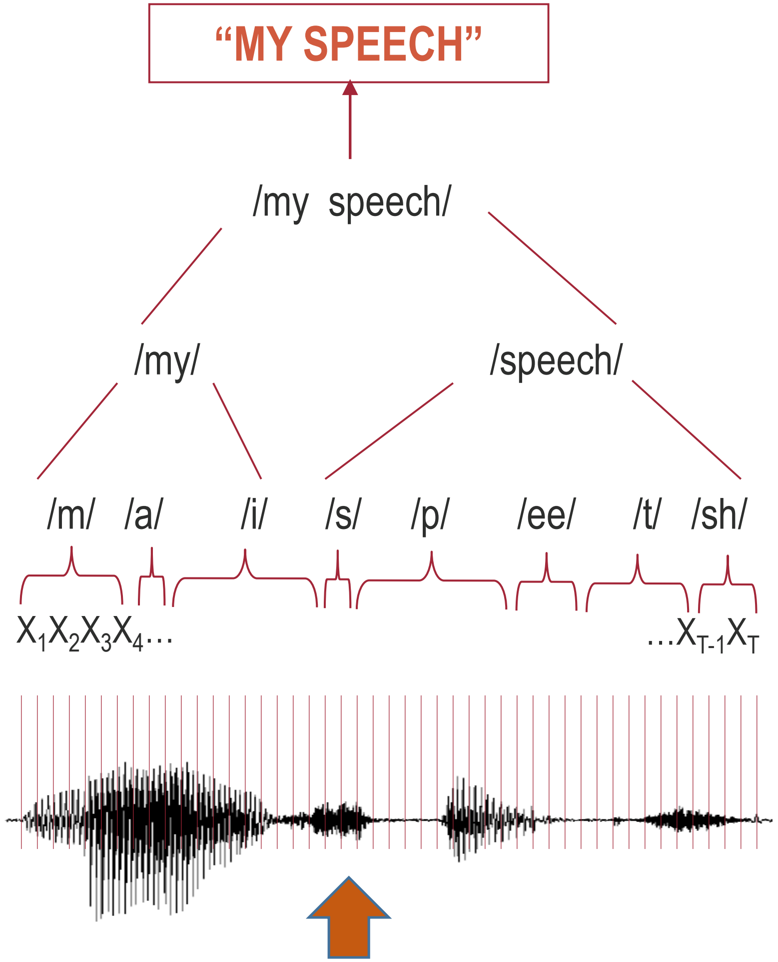
Human-Machine Interaction

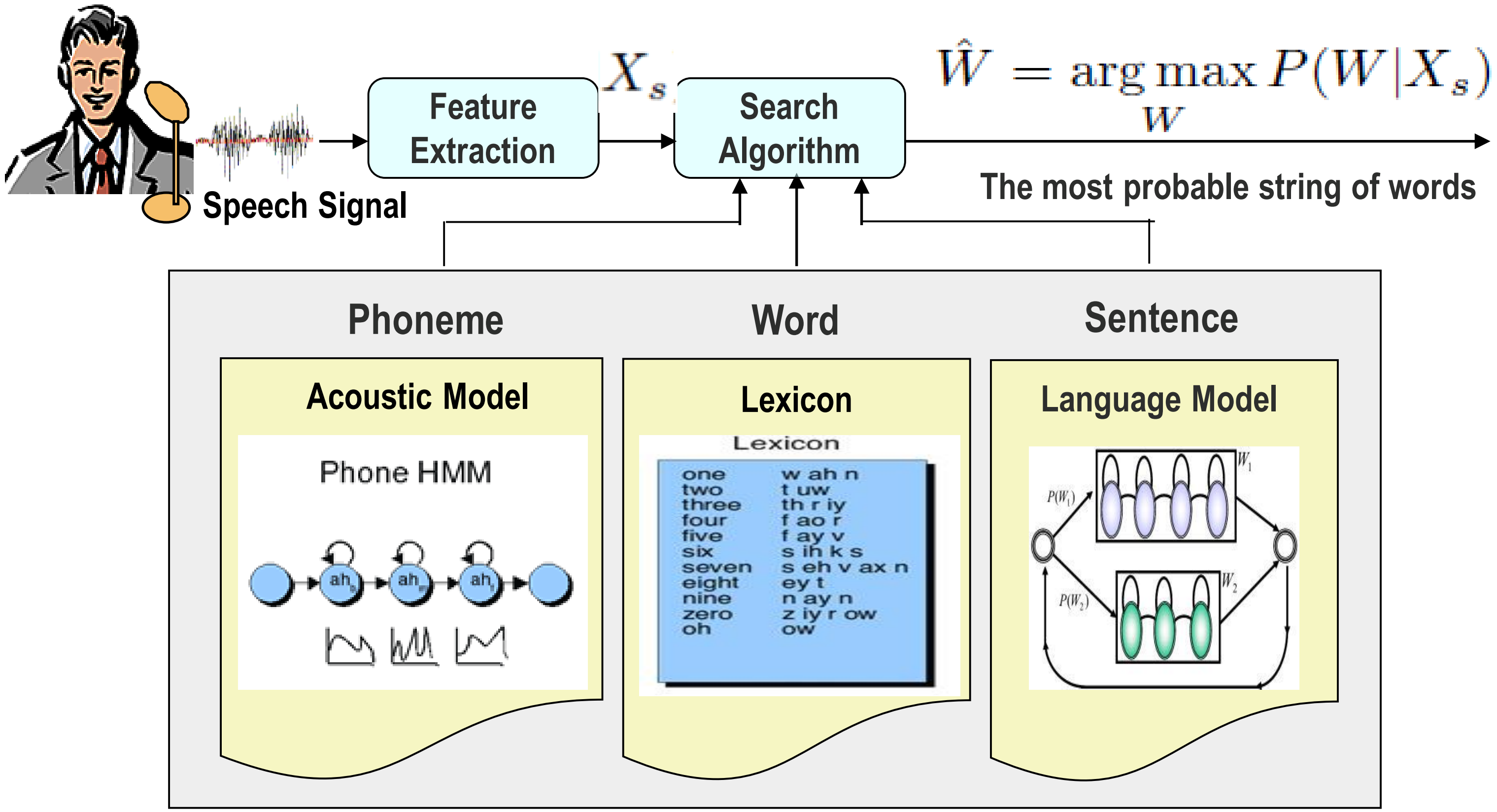
- **Modality in Human-Machine Interaction**
  → Providing a technology with ability to listen and speak



Automatic Speech Recognition (ASR)

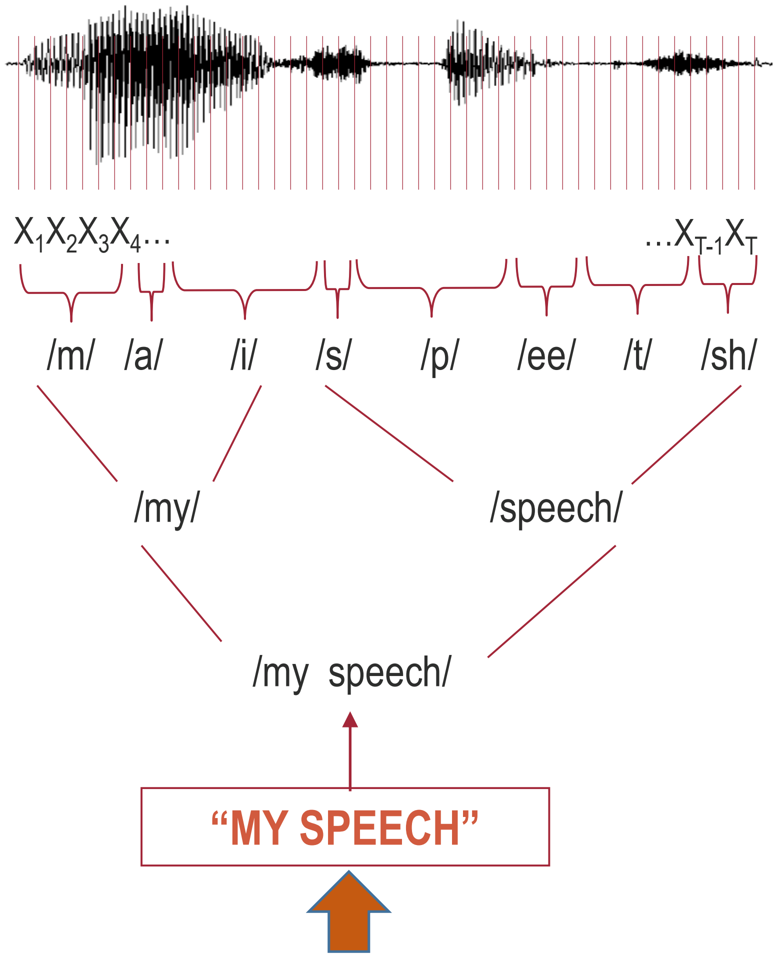
- **Traditional ASR based on Hidden Markov Model (HMM)**

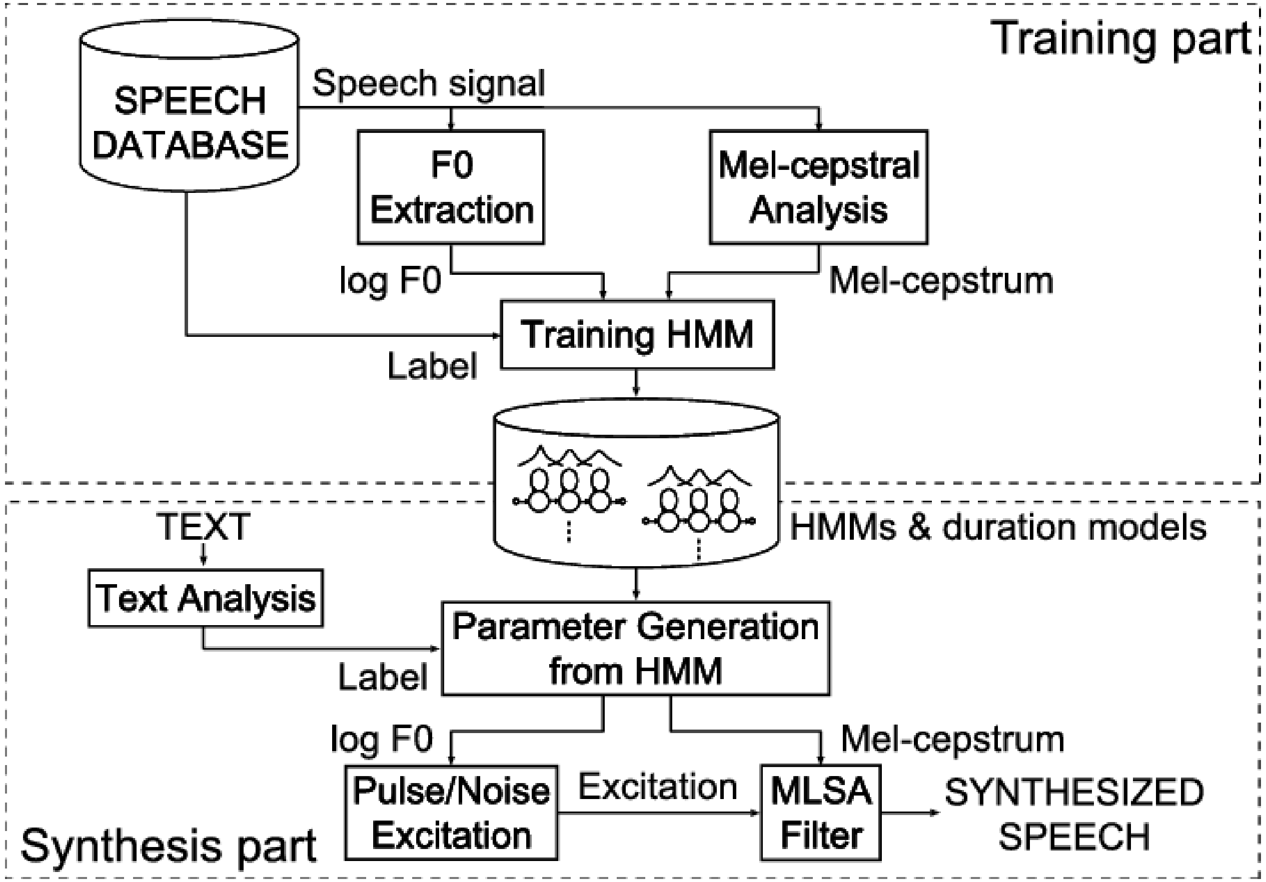




Text-to-Speech Synthesis (TTS)

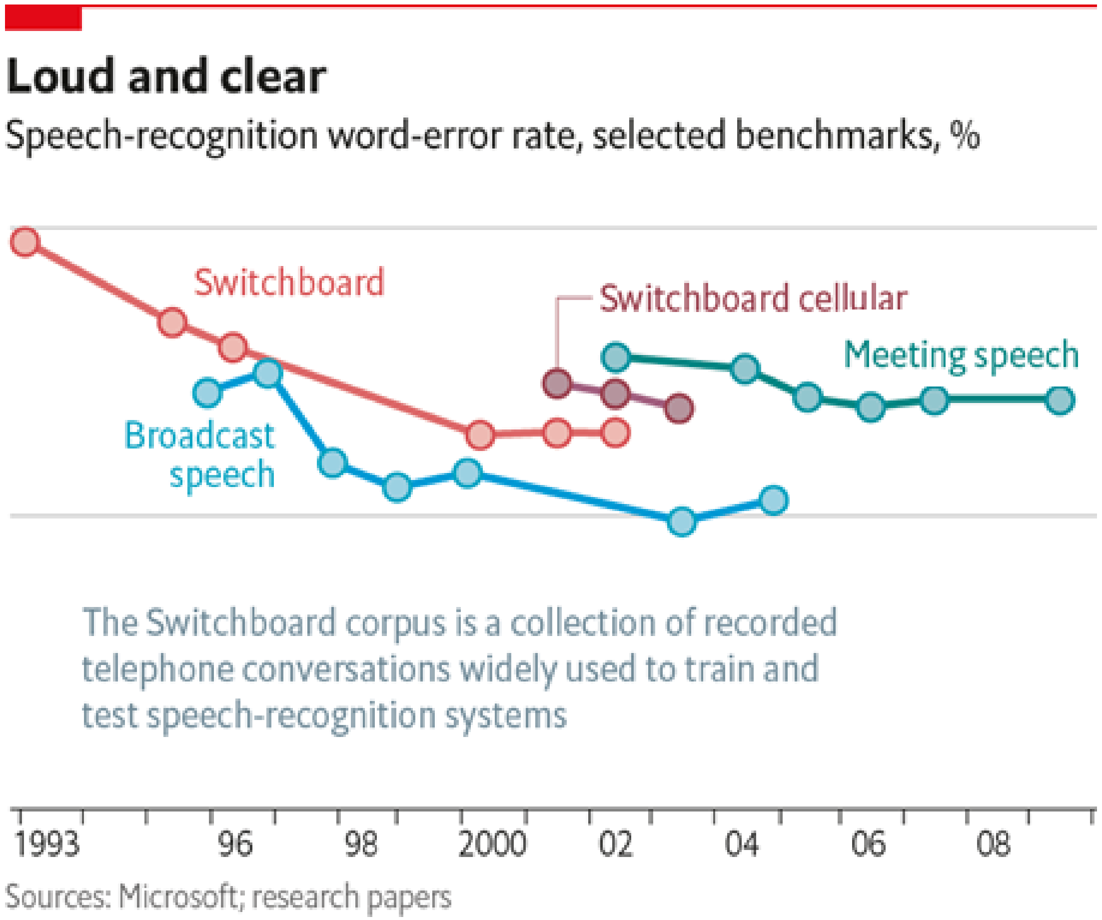
- Traditional TTS based on Hidden Markov Model (HMM)





[Zen et al., 2009]

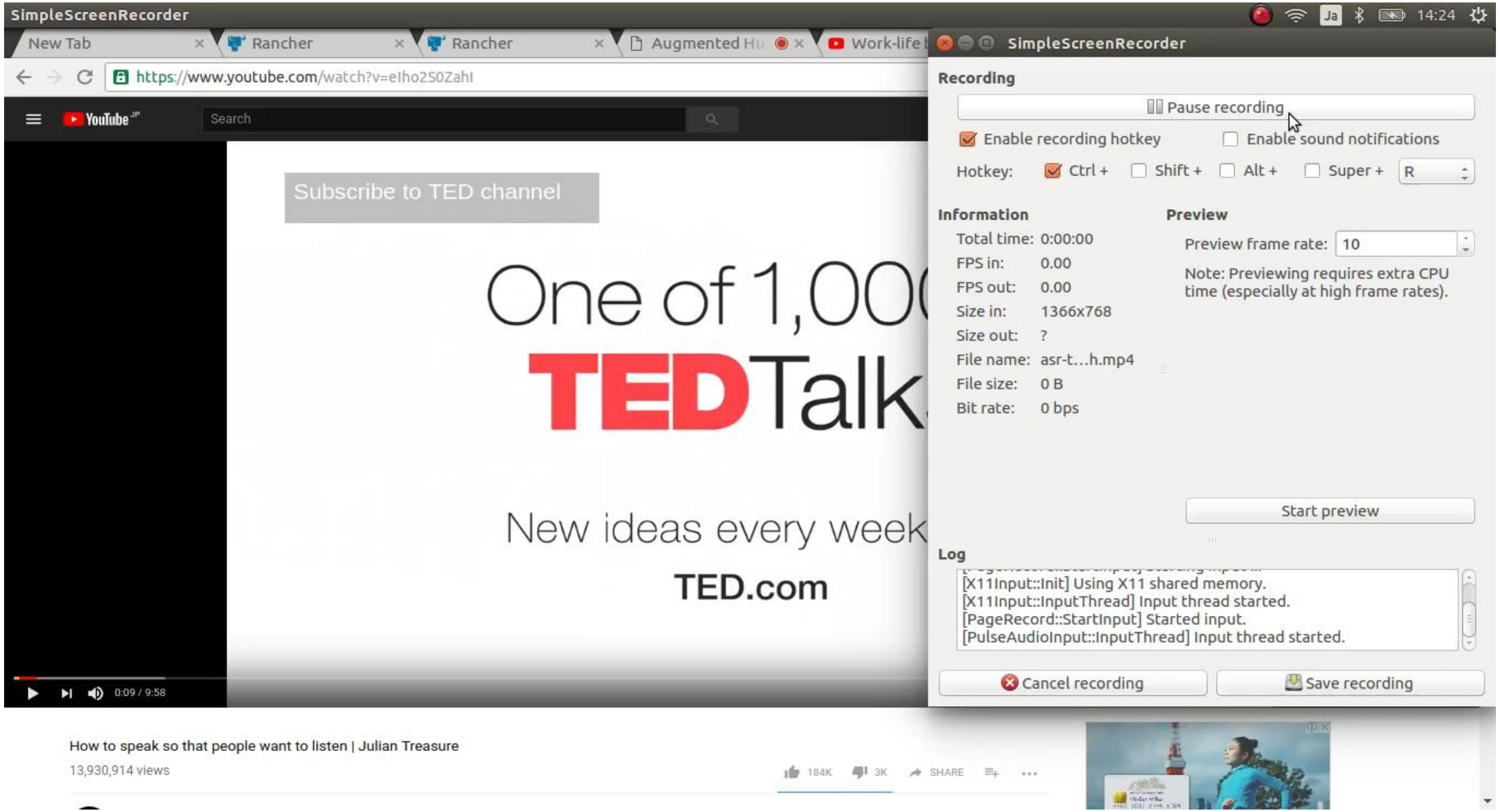
ASR and TTS Performance



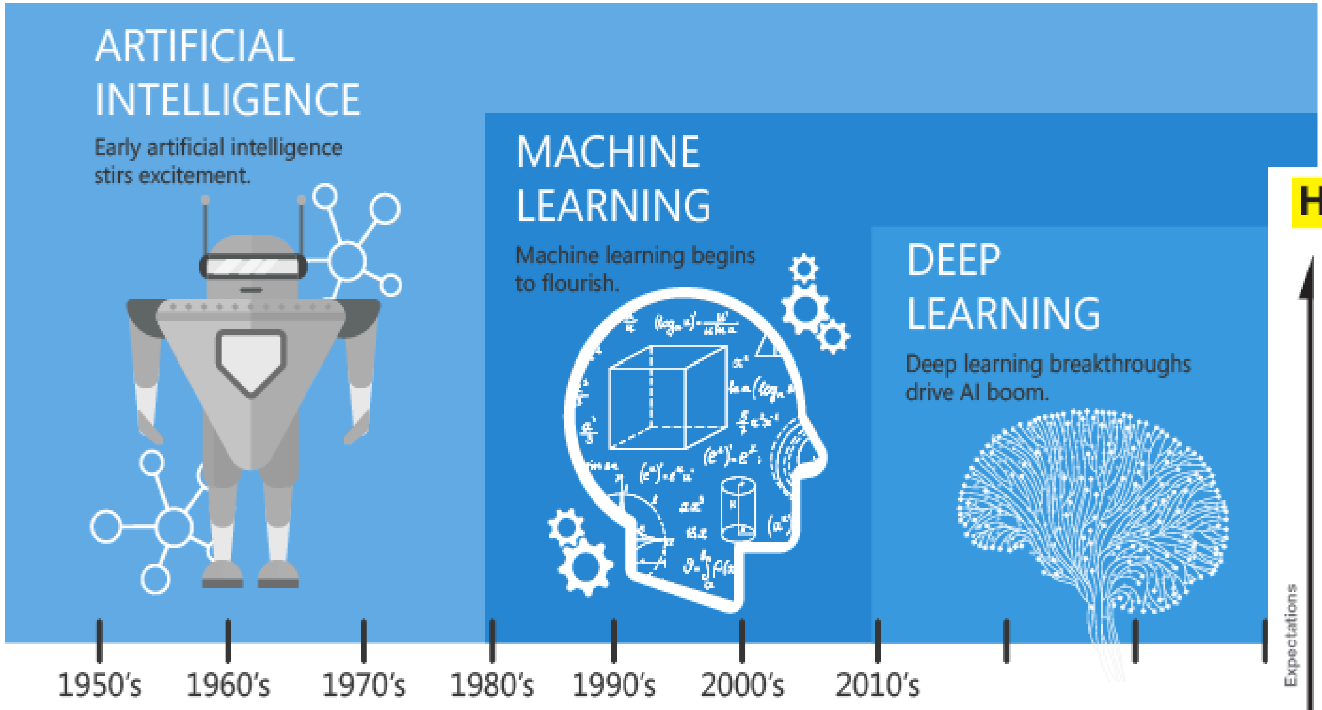
TTS: From robot voice to human-like voice

[Source: https://www.economist.com/technology-quarterly/2017-05-01/language]

ASR and TTS Performance

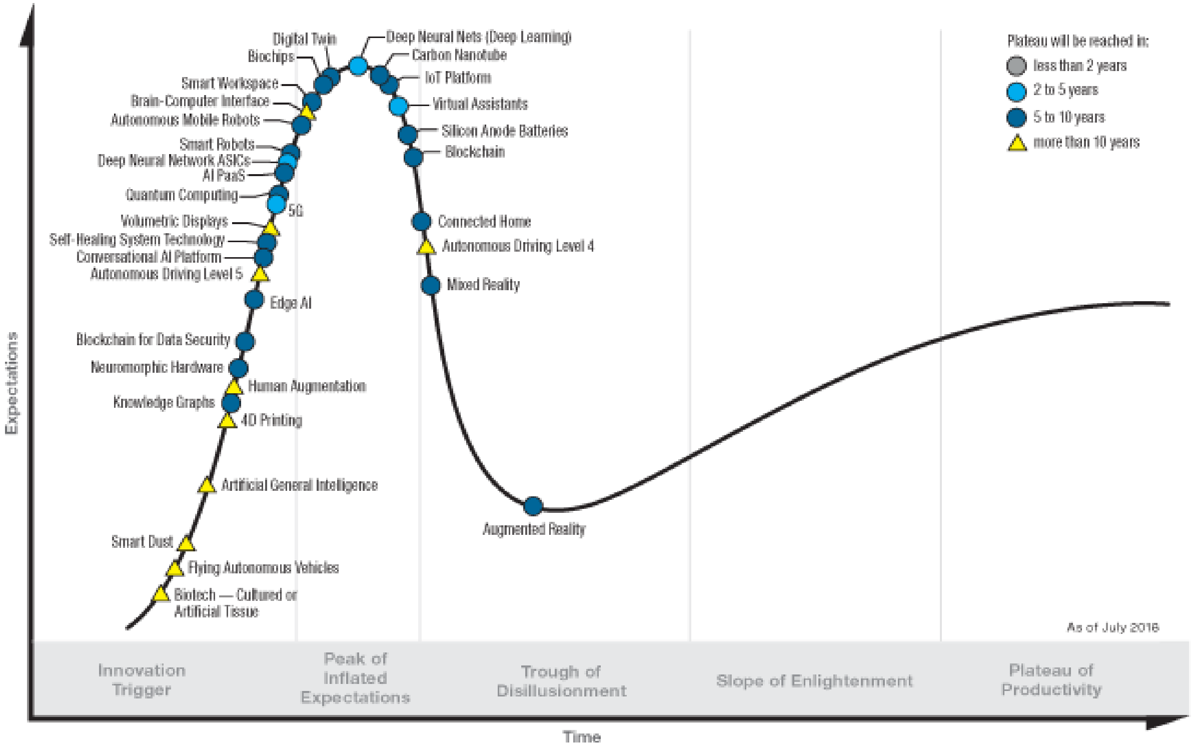


Paradigm Shift: Deep Learning Hype



[Source: Linked IN | Machine Learning vs Deep Learning]

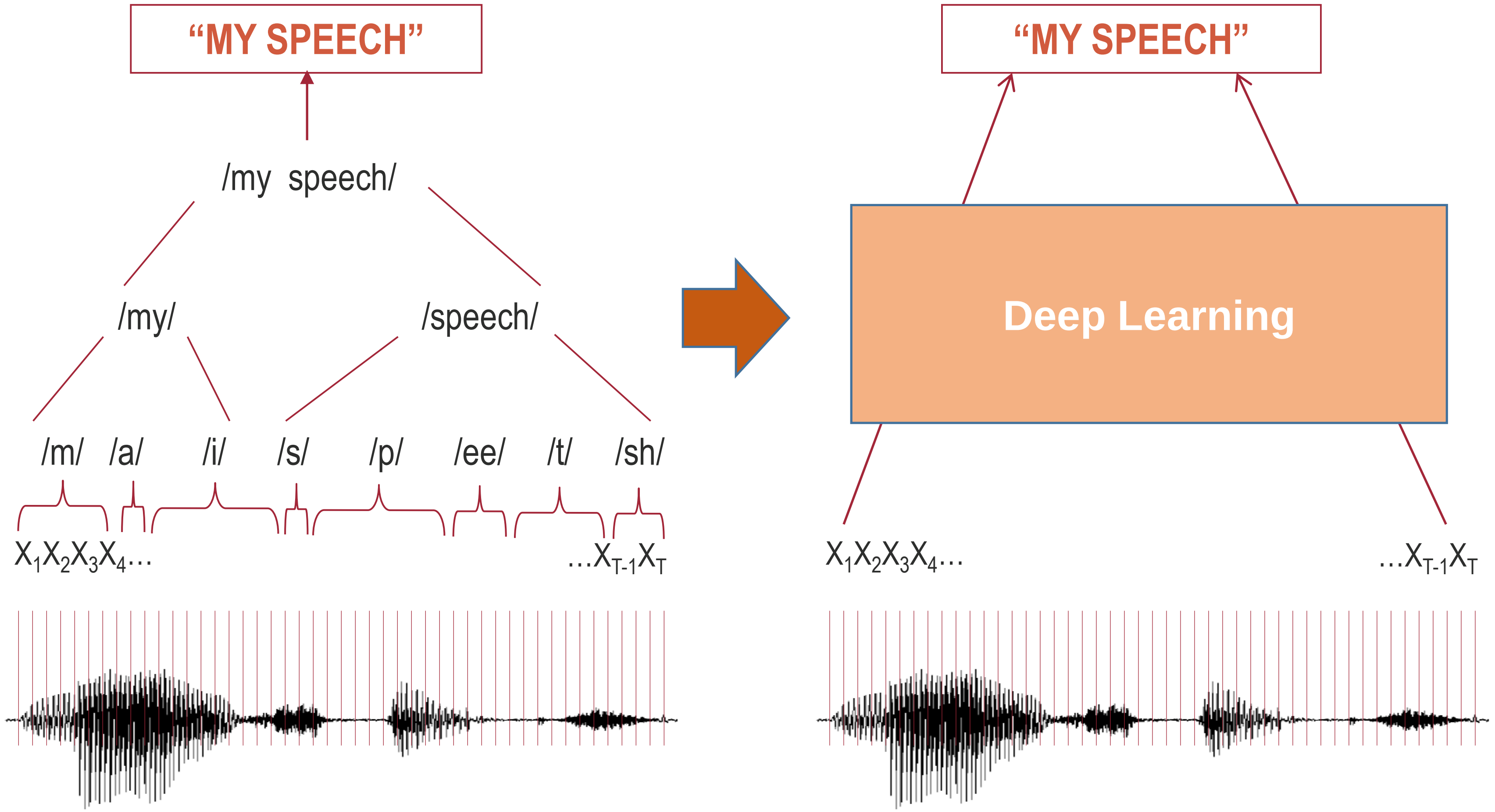
Hype Cycle for Emerging Technologies, 2018



[Source: https://www.gartner.com/en/newsroom/press-releases/2017-08-15-gartner-identifies-three-megatrends-that-will-drive-digital-business-into-the-next-decade]

Recent ASR Technology

- ## ASR based on Deep Learning



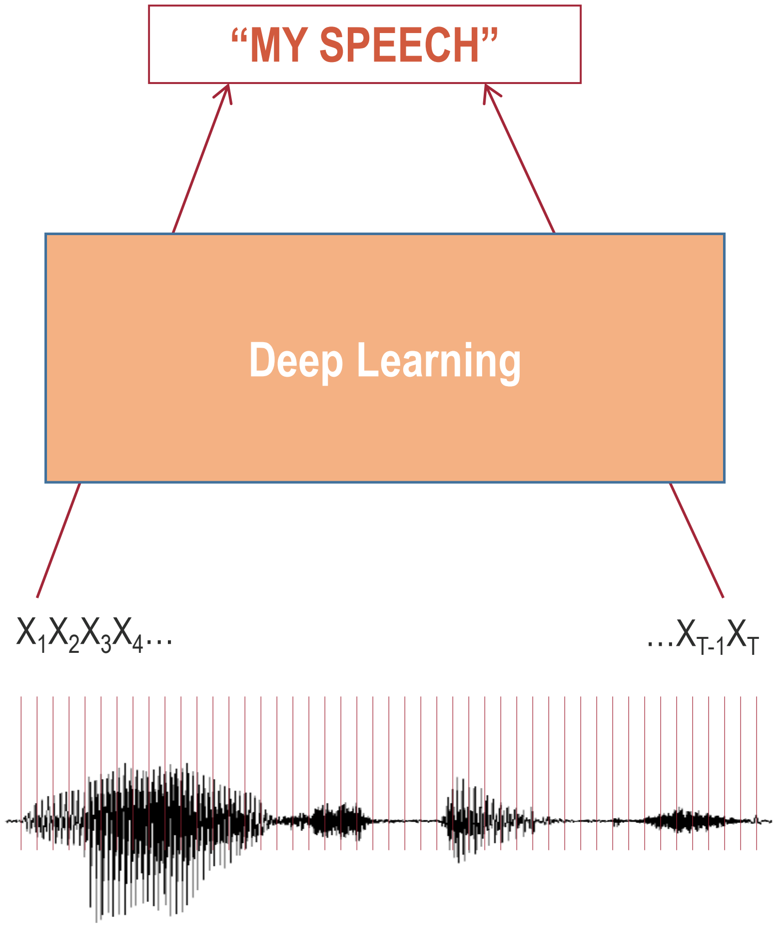
Important factors of Deep Learning:

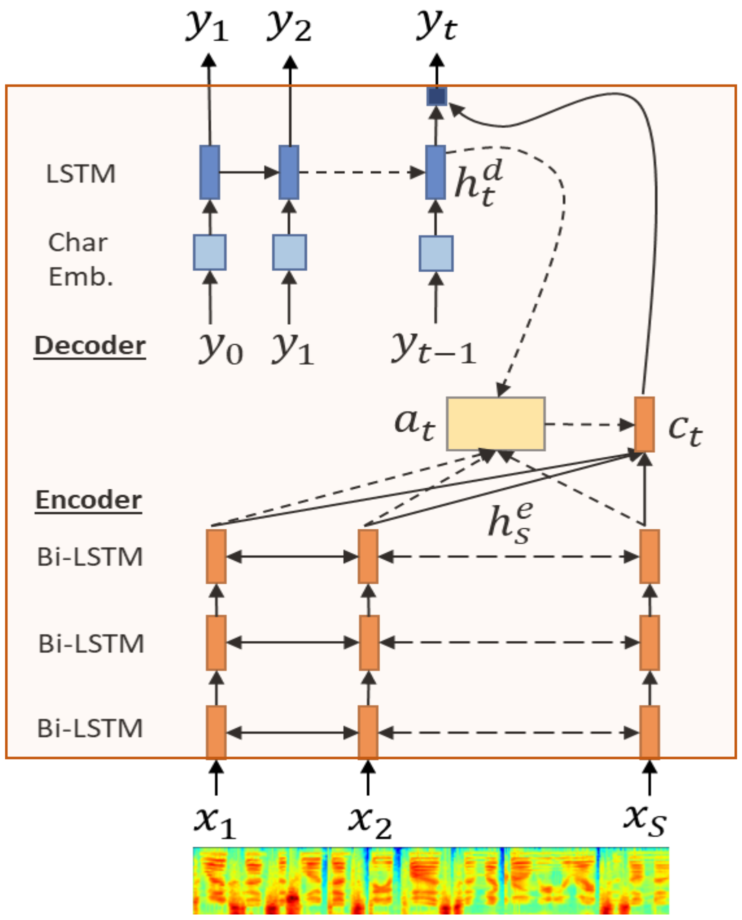
→ Simplify many complicated hand-engineered models

→ Let the networks find the way that map from speech to text

Recent ASR Technology

- ASR based on Deep Learning



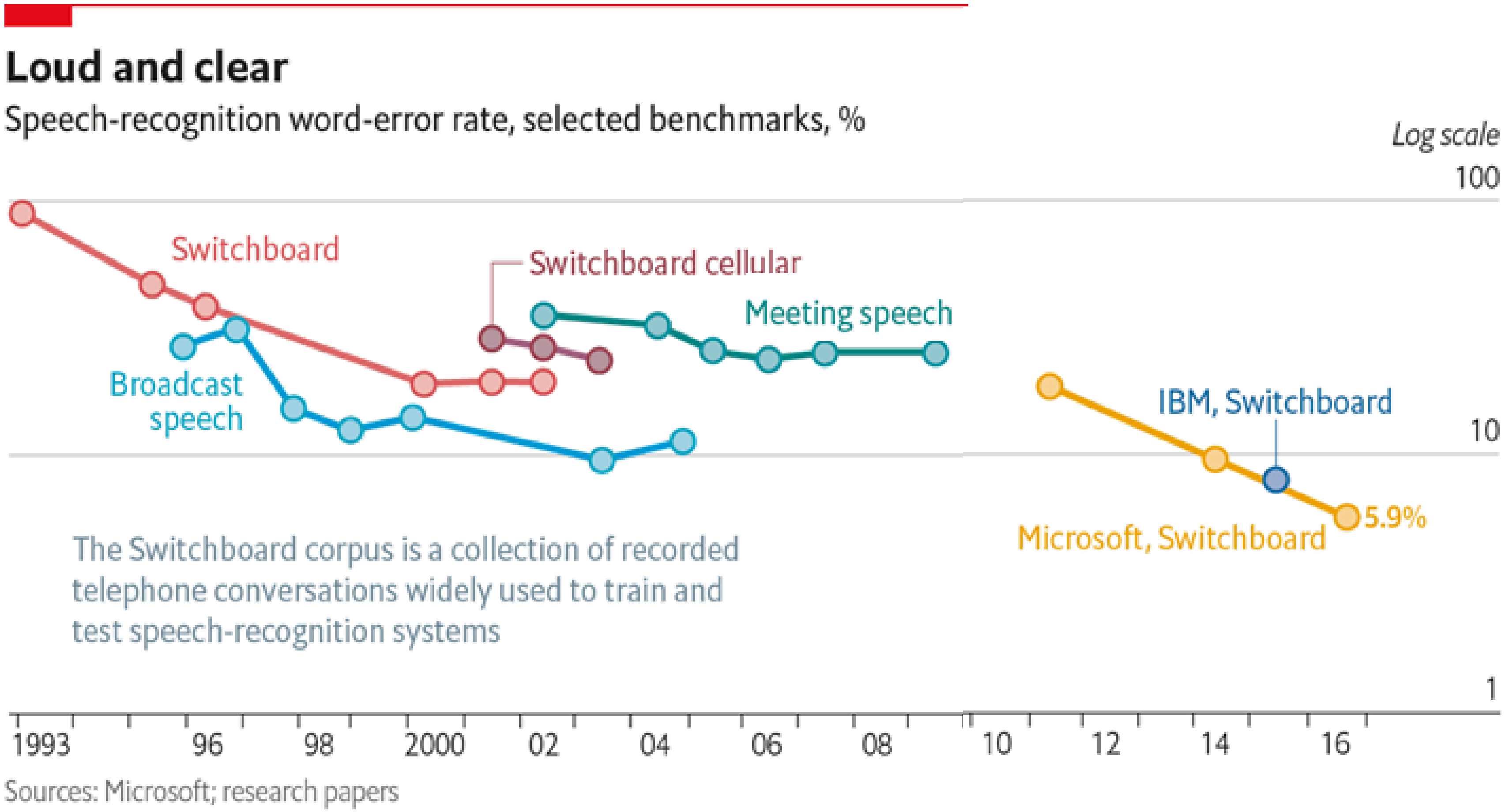


Input and Output
- $\boldsymbol{x} = [x_1, \ldots, x_S]$ (Speech features)
- $\boldsymbol{y} = [y_1, \ldots, y_T]$ (Text)

Model states
- $h^e_{[1..S]}$ = encoder states
- h_t^d = decoder state at time t
- a_t = attention probability

NN types
- LSTM (Long short-term memory)
- Bi-LSTM (Bidirectional LSTM)

ASR Progress



[Source: https://www.economist.com/technology-quarterly/2017-05-01/language]

ASR Progress

- **IBM vs Microsoft:"Human parity"speech recognition record**

  → Makes the same / fewer errors than professional transcriptionists

Model	N-gram LM		Neural net LM	
	CH	SWB	CH	SWB
Povey et al. [54] LSTM	15.3	8.5	-	-
Saon et al. [51] LSTM	15.1	9.0	-	-
Saon et al. [51] system	13.7	7.6	12.2	6.6
2016 Microsoft system	13.3	7.4	11.0	5.8
Human transcription			11.3	5.9

[Xiaong et al., 2017]

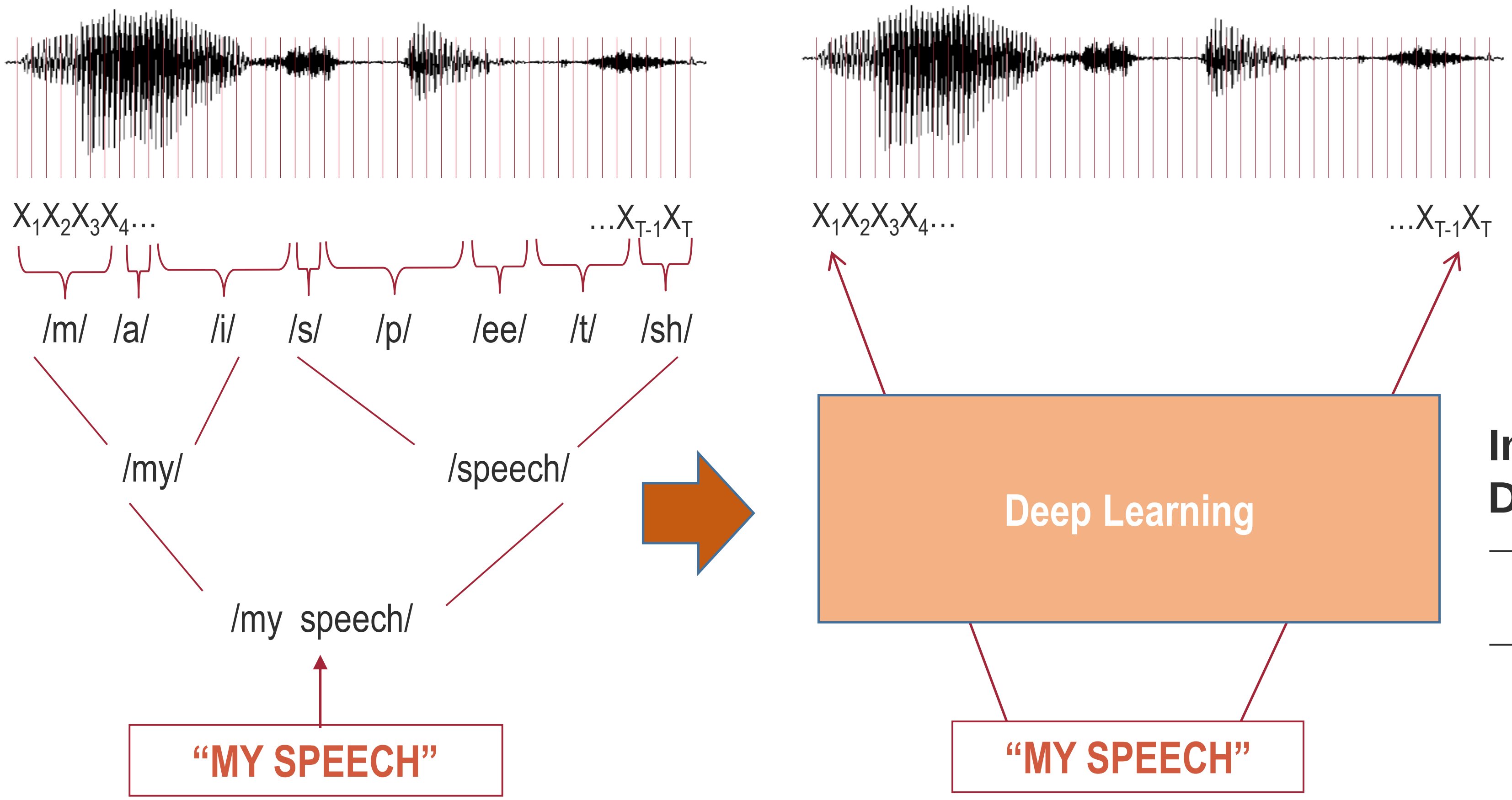
New IBM System

	WER [%]	
	SWB	CH
n-gram	6.7	12.1
n-gram + model-M	6.1	11.2
n-gram + model-M + Word-LSTM	5.6	10.4
n-gram + model-M + Char-LSTM	5.7	10.6
n-gram + model-M + Word-LSTM-MTL	5.6	10.3
n-gram + model-M + Char-LSTM-MTL	5.6	10.4
n-gram + model-M + Word-DCC	5.8	10.8
n-gram + model-M + 4 LSTMs + DCC	**5.5**	**10.3**

[Saon et al., 2017]

Recent TTS Technology

▪ TTS based on Deep Learning



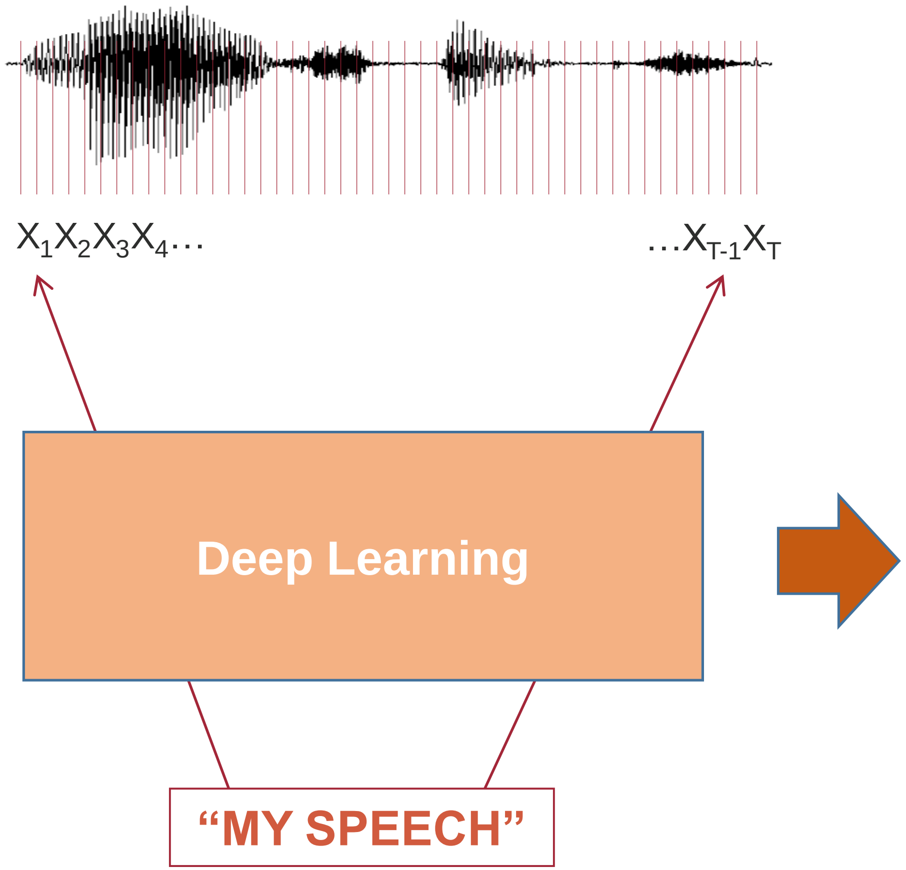
Important factors of Deep Learning:

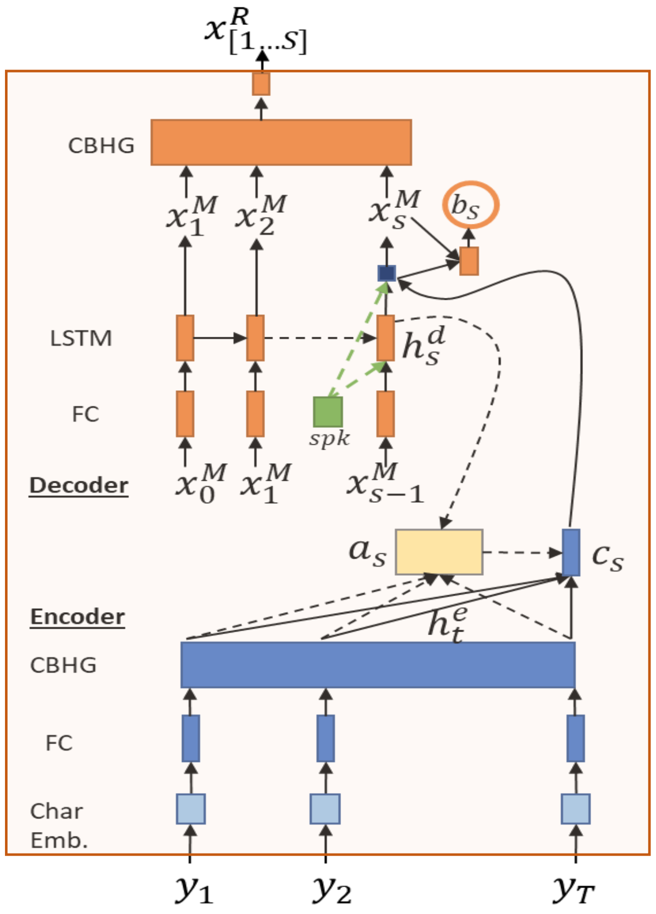
→ Simplify many complicated hand-engineered models

→ Let the networks find the way that map from text to speech

Recent TTS Technology

- TTS based on Deep Learning





Input and Output

- $\boldsymbol{x^R} = [x_1, \ldots, x_S]$ (linear spect. Feat.)
- $\boldsymbol{x^M} = [x_1, \ldots, x_S]$ (mel spect. feat)
- $\boldsymbol{y} = [y_1, \ldots, y_T]$ (text)

Model states

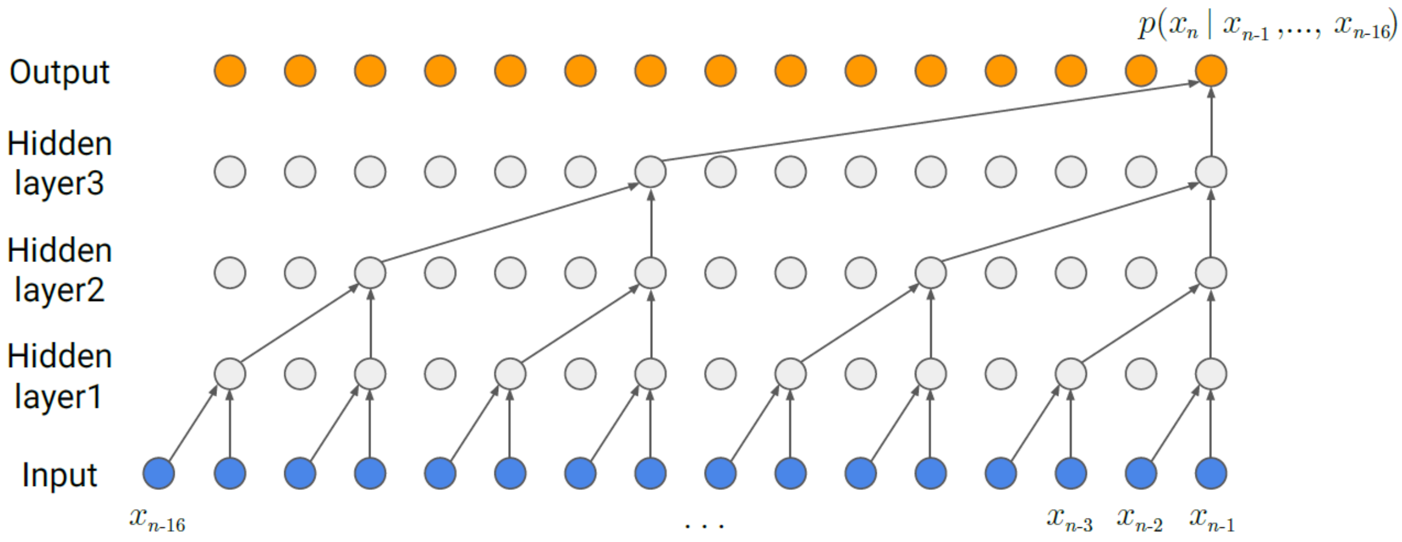
- $h^e_{[1..S]}$ = encoder states
- h^d_S = decoder state at time t
- a_S = attention probability

NN types

- FC (Full-connected)
- LSTM (Long short-term memory)
- Bi-LSTM (Bidirectional LSTM)
- CBHG (Conv bank + highway net + bidirectional GRU)

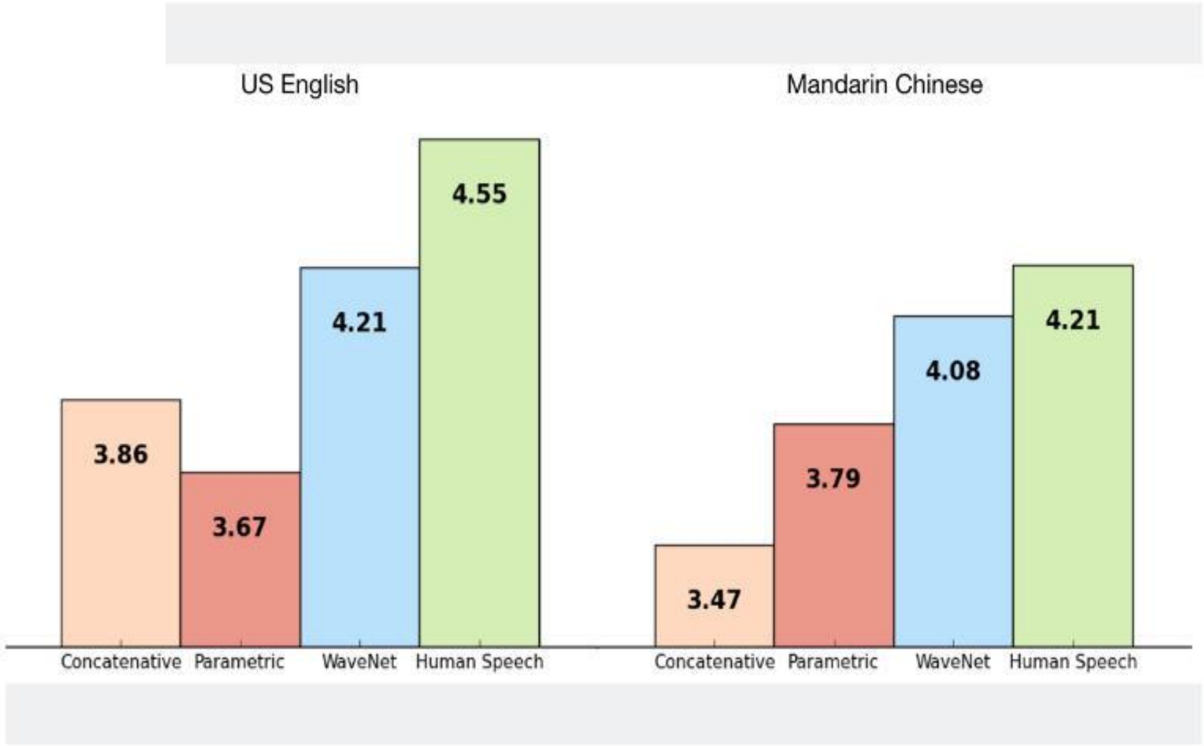
TTS Progress

- **Google's DeepMind:**
  Major milestone in making machines talk like humans



WaveNet: Generative Model for Raw Audio

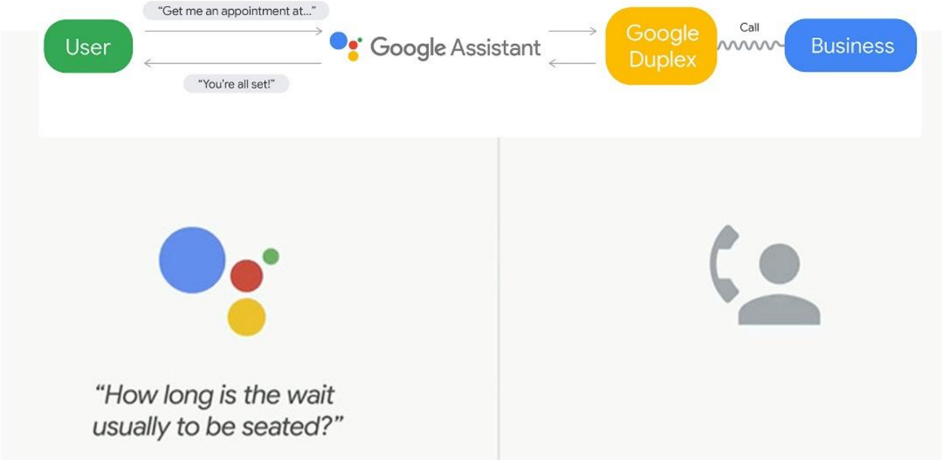
[Source: https://www.zdnet.com/article/googles-deepmind-claims-major-milestone-in-making-machines-talk-like-humans/]



TTS Progress

- **Google Duplex:**
  AI System for Accomplishing Real-World Tasks Over the Phone

Google AI



Duplex scheduling a hair salon appointment:

Duplex calling a restaurant:

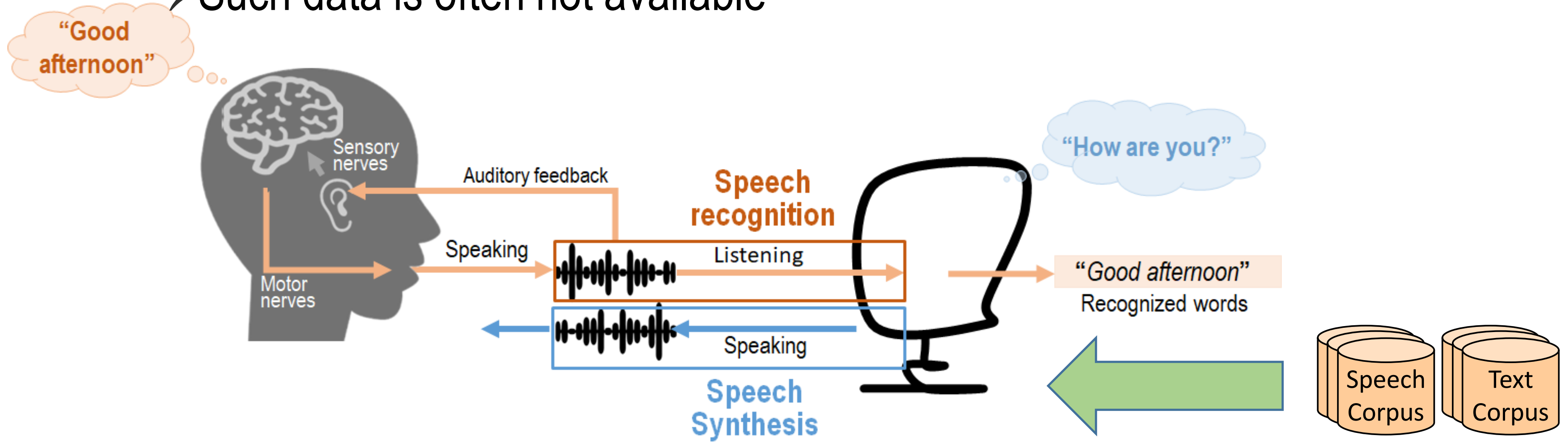
[Source: https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html]

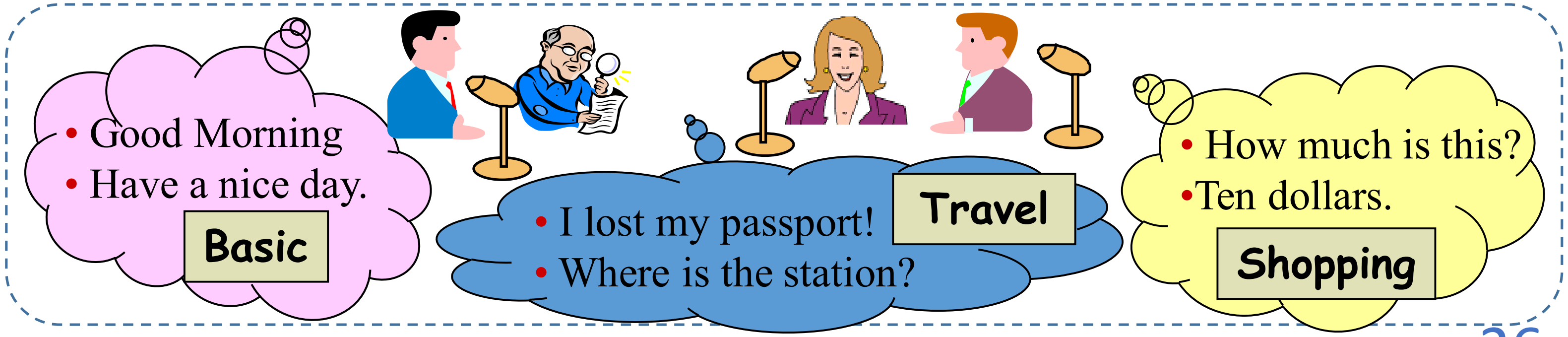
What is left?
Are all problems solved?

Machine Learning vs Human Learning

▪ Learning Issues

→ It requires a lot of parallel speech and text, more than human need
→ Such data is often not available





Machine Learning vs Human Learning

- **Learning Issues**
  - → It requires a lot of parallel speech and text, more than human need
  - → Such data is often not available

Lewis, M. Paul (ed.), 2009. Ethnologue: Languages of the World, Sixteenth edition. Dallas, Tex.: SIL International. Online version: http://www.ethnologue.com/.

Area	Living Languages		Number of Speakers	
	Count	Percent	Count	Percent
Africa	2,110	30.5	726,453,403	12.2
Americas	993	14.4	50,496,321	0.8
Asia	**2,322**	**33.6**	**3,622,771,264**	**60.8**
Europe	234	3.4	1,553,360,941	26.1
Pacific	1,250	18.1	6,429,788	0.1
Totals	6,909	100	5,959,511,717	100

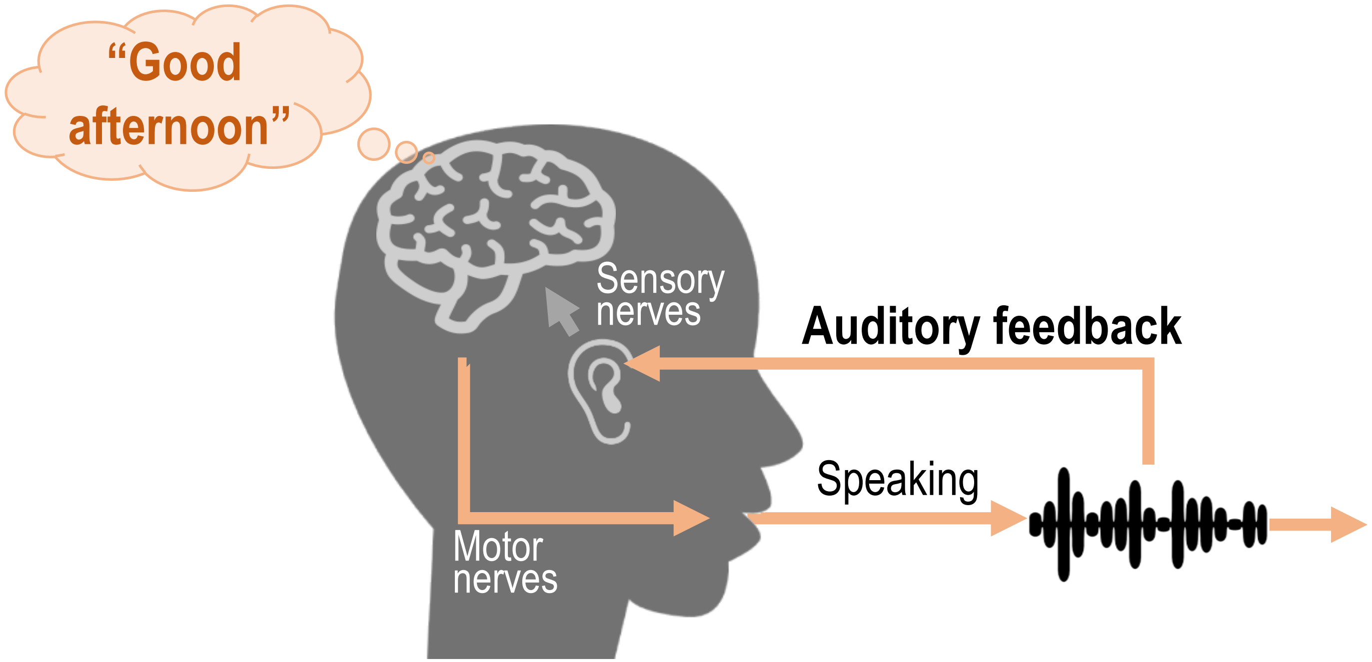
Only up to ~100 languages are covered by language technologies.

Nearly 7000 living languages (spoken by 350 million people) have not yet been covered.

Machine Learning vs Human Learning

- ## Human Learning

→ Humans learn how to talk by constantly repeating their articulations & listening to sounds produced
→ A closed-loop speech chain mechanism has a critical auditory feedback mechanism



Children who lose their hearing often have difficulty to produce clear speech

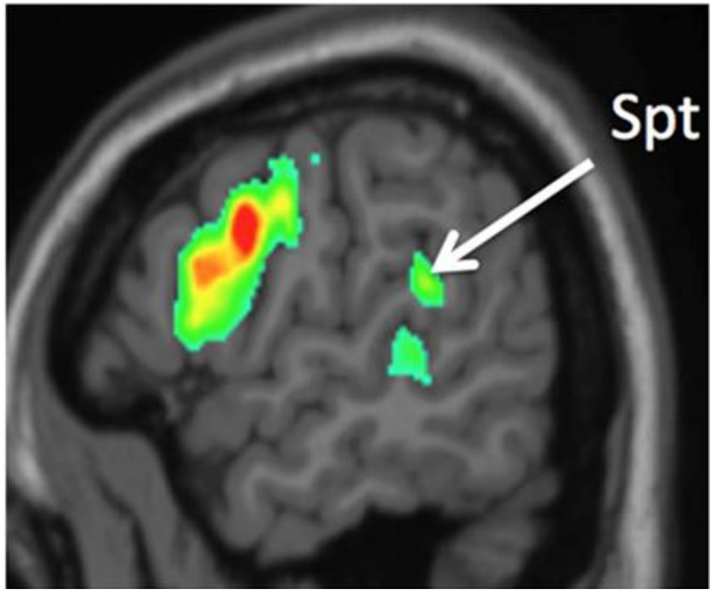
Adults who become deaf after becoming proficient with a language nonetheless suffer speech articulation declines as a result of the lack of auditory feedback

[Waldstein, 1990]

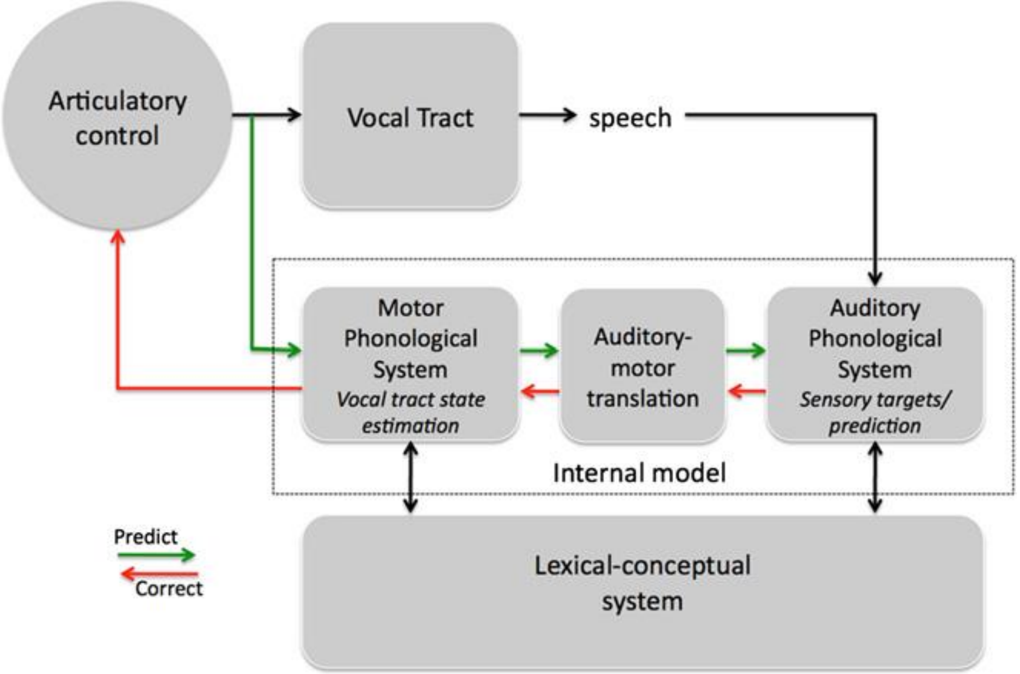
Machine Learning vs Human Learning

- **Human Brain: Sensorimotor Integration in Speech Processing**

(1) the auditory system is critically involved in the production of speech
(2) the motor system is critically involved in the perception of speech



Spt exhibits sensorimotor response properties, activating both during the passive perception of speech and during covert (subvocal) speech articulation [Hickok et al, 2003]

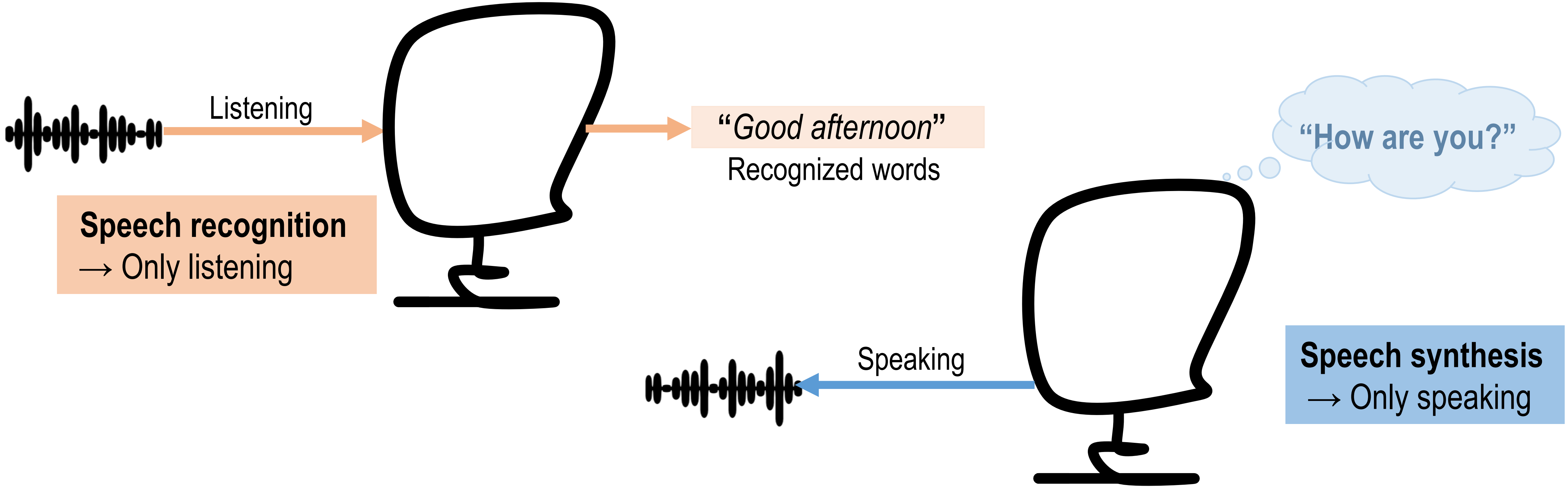


An Integrated State Feedback Control (SFC) Model of Speech Production [Hickok et al. 2011]

Machine Learning vs Human Learning

- ## Machine Learning

→ Computers are able to learn how to listen or learn how to speak
→ But, computers cannot hear their own voice



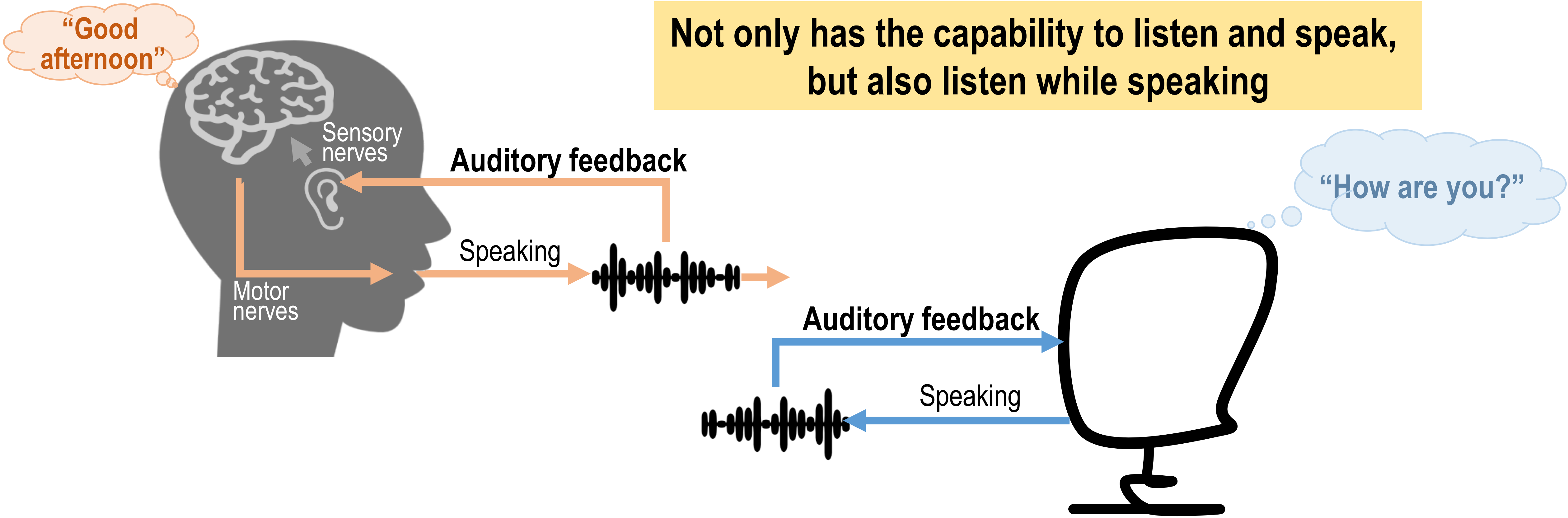
Part I
Basic Machine Speech Chain

[A. Tjandra, S. Sakti, S. Nakamura, "Listening while Speaking: Speech Chain by Deep Learning", in Proc. ASRU, 2017]

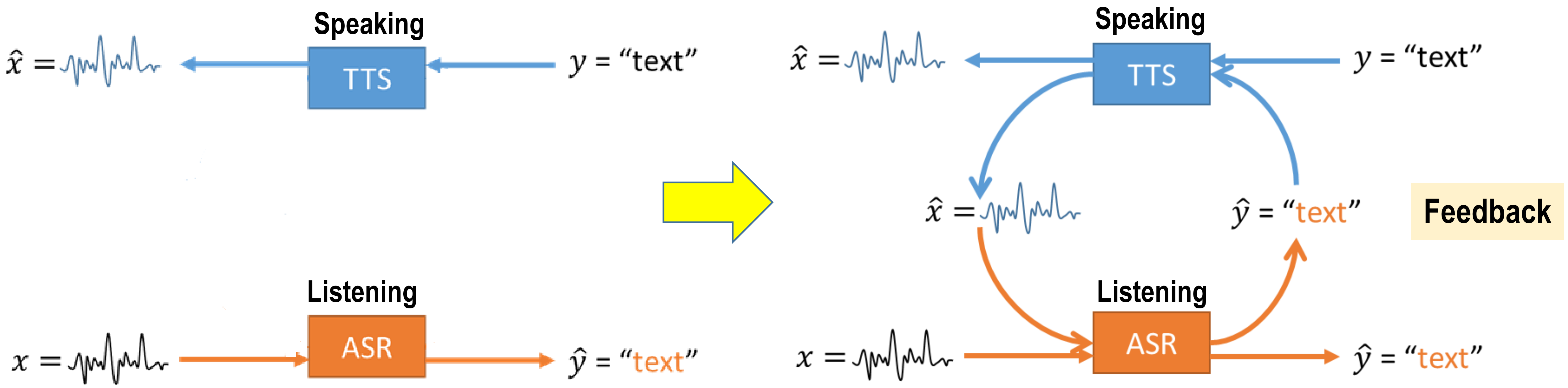
Machine Speech Chain

- **Proposed Method**
  - → Develop a closed-loop speech chain model based on deep learning
  - → The first deep learning model that integrates human speech perception & production behaviors

Not only has the capability to listen and speak, but also listen while speaking



Machine Speech Chain

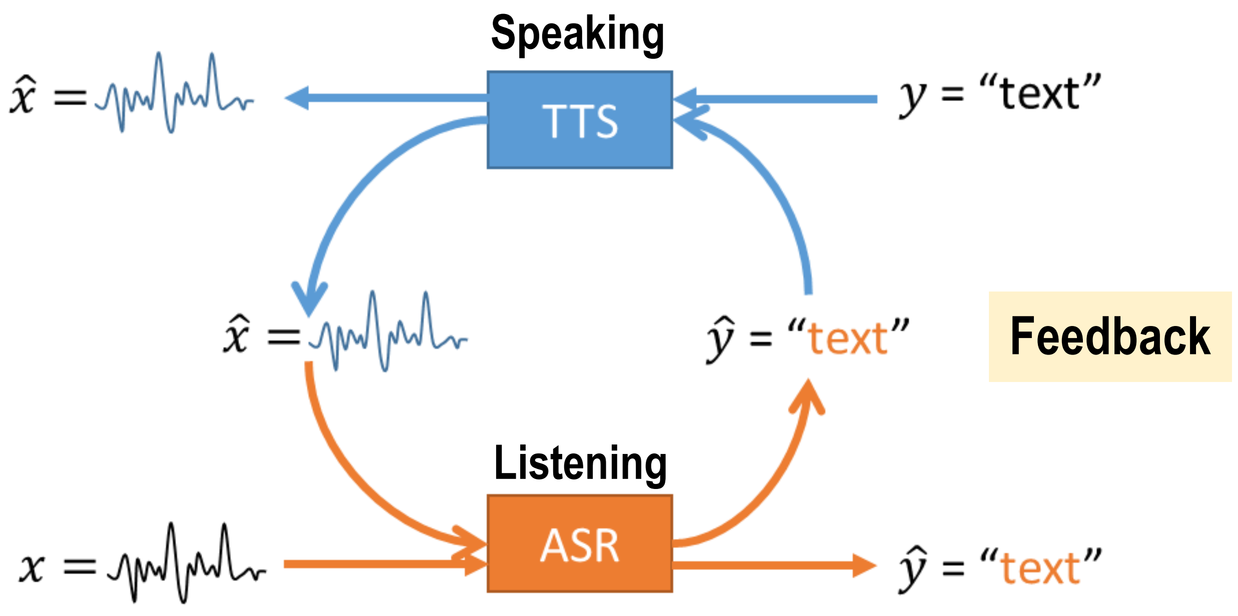


A closed-loop architecture:

→ **In training stage:**
- Allow to train with labeled and unlabeled data (semi-supervised learning)
- Allow ASR and TTS to teach each other using unlabel data and generate useful feedback

→ **In Inference stage:** Possible to use ASR & TTS module independently

Overall Architecture



Definition:

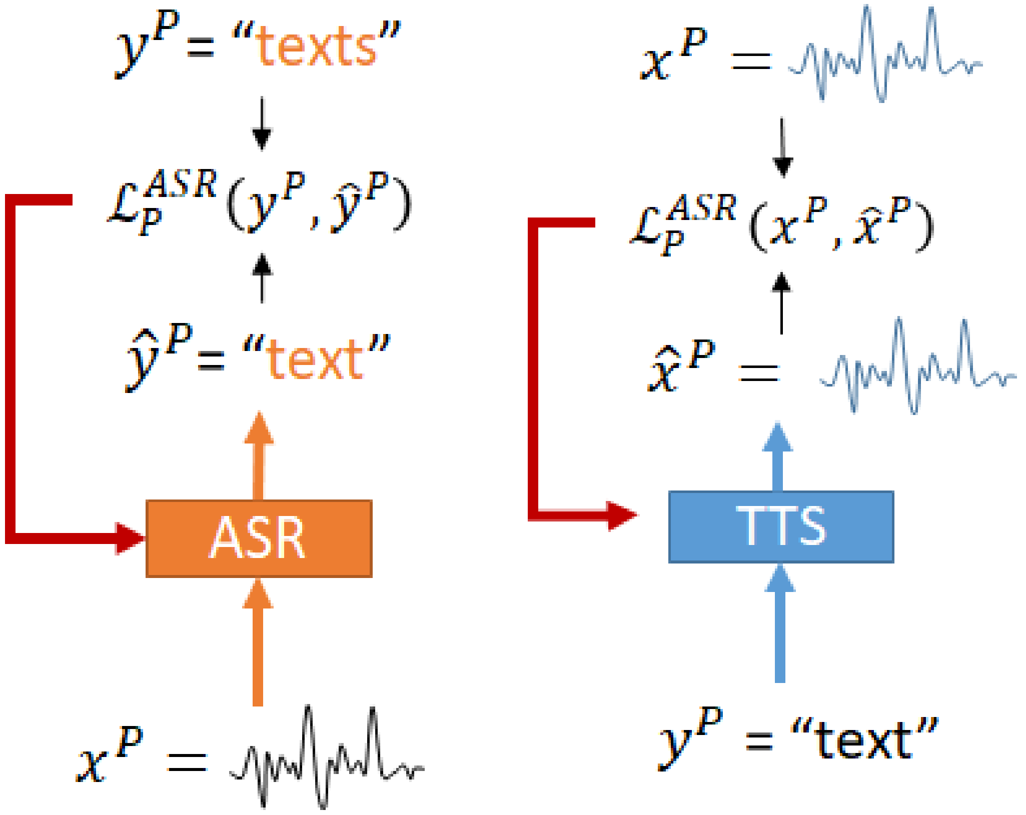
- x = original speech, y = original text
- $\hat{x}$ = predicted speech, $\hat{y}$ = predicted text
- $ASR(x): x \rightarrow \hat{y}$ (seq2seq model transform speech to text)
- $TTS(y): y \rightarrow \hat{x}$ (seq2seq model transform text to speech)

Learning in Machine Speech Chain

Case #1: Supervised Learning with Speech-Text Data

Given a pair speech-text $(\boldsymbol{x}^P, \boldsymbol{y}^P)$

- Train ASR and TTS in supervised learning
- Directly optimized:
  $\rightarrow ASR$ by minimizing $\mathcal{L}_P^{ASR}(y^P, \hat{y}^P)$
  $\rightarrow TTS$ by minimizing $\mathcal{L}_P^{TTS}(x^P, \hat{x}^P)$
- Update both ASR and TTS independently



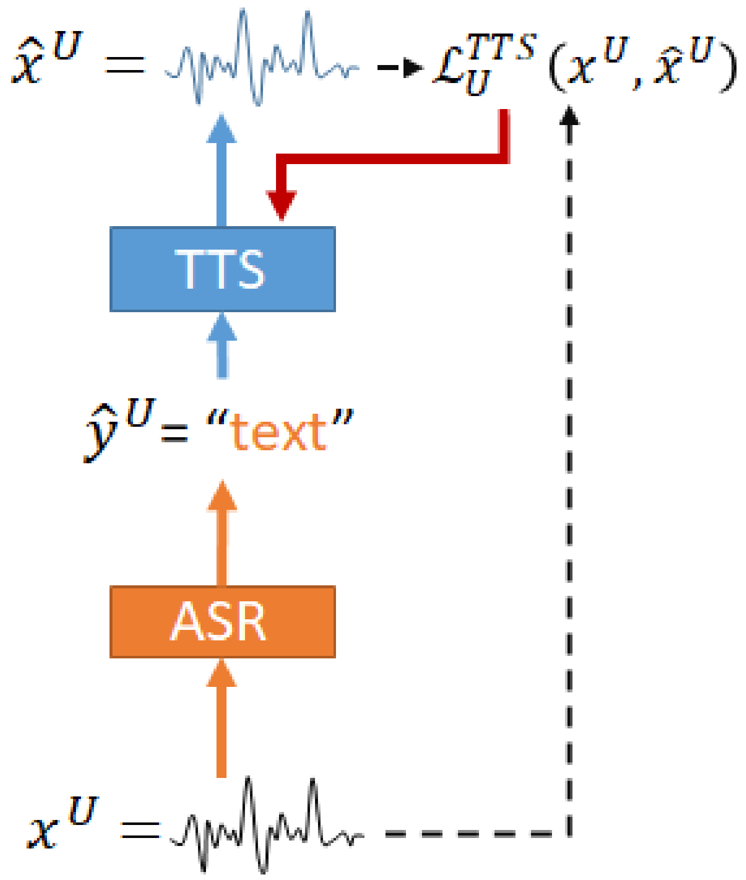
Learning in Machine Speech Chain

Case #2: Unsupervised Learning with Speech Only

Given the unlabeled speech features x^U

1. ASR predicts the transcription $\hat{y}^U$
2. Based on $\hat{y}^U$, TTS tries to reconstruct speech features $\hat{x}^U$
3. Calculate $\mathcal{L}_U^{TTS}(x^U, \hat{x}^U)$ between original speech features x^U and the predicted $\hat{x}^U$

Possible to improve TTS with speech only by the support of ASR



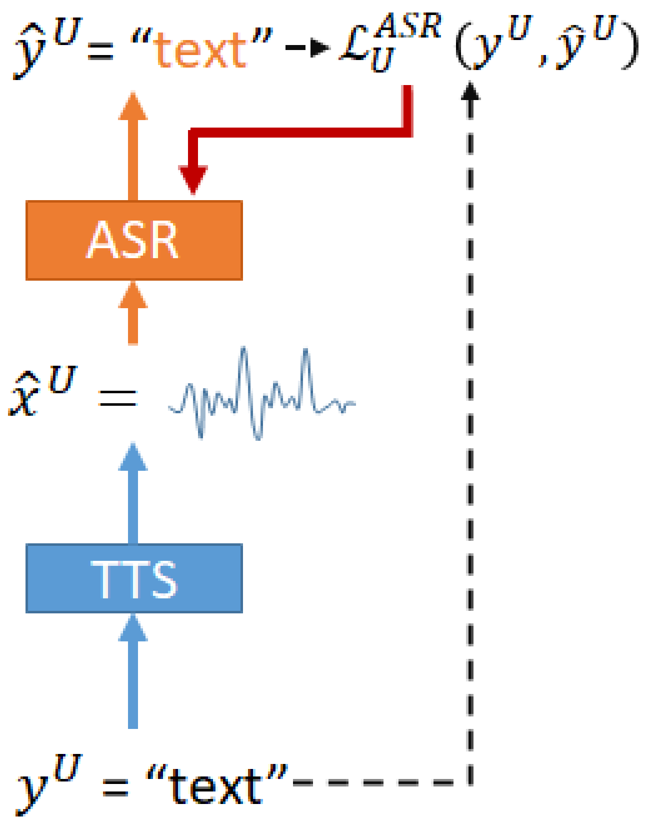
Learning in Machine Speech Chain

Case #3: Unsupervised Learning with Text Only

Given the unlabeled text features y^U

1. TTS generates speech features $\hat{x}^U$
2. Based on $\hat{x}^U$, ASR tries to reconstruct text features $\hat{y}^U$
3. Calculate $\mathcal{L}_U^{ASR}(y^U, \hat{y}^U)$between original text features $y^U$and the predicted $\hat{y}^U$

Possible to improve ASR with text only by the support of TTS



Learning in Machine Speech Chain

- **Training Objective**

$$\mathcal{L} = \alpha * \left(\mathcal{L}_P^{ASR} + \mathcal{L}_P^{TTS}\right) + \beta * \left(\mathcal{L}_U^{ASR} + \mathcal{L}_U^{TTS}\right)$$

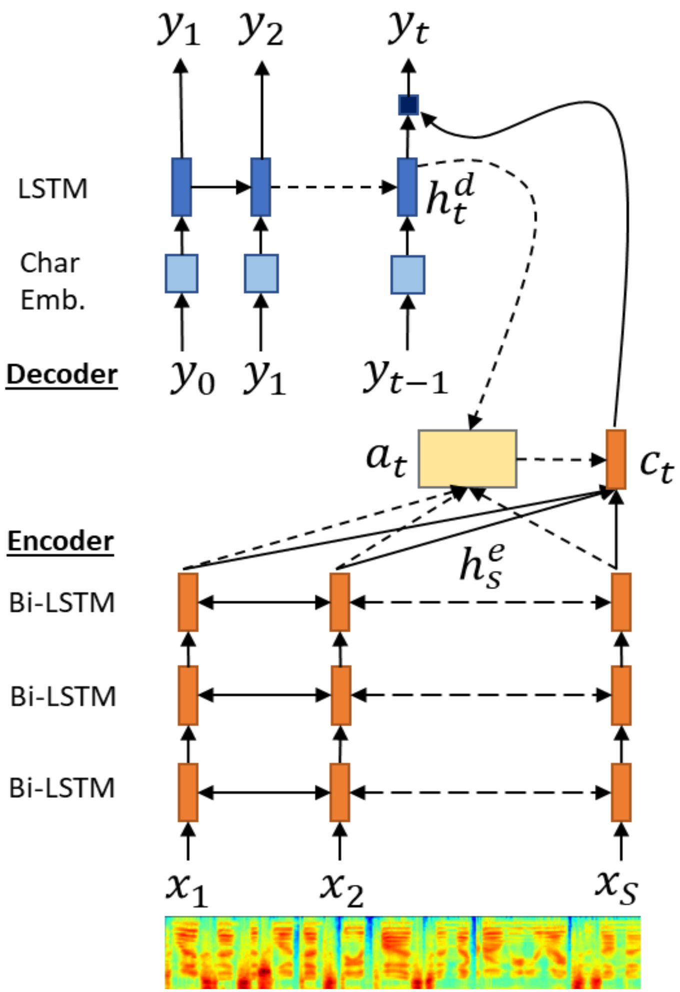
- **Basic Idea**

$\longrightarrow$ Possible to train the new matters without forgetting the old one

$\longrightarrow$ $\alpha > 0$: keep use some portions of the loss and the gradient provided by the paired training set

$\longrightarrow$ $\alpha = 0$: completely learn new matters with only speech or only text

Sequence-to-Sequence ASR



Similar to [LAS, Chan et al. 2015]

Input & output

- $\boldsymbol{x} = [x_1, \ldots, x_S] \rightarrow$ speech feature
- $\boldsymbol{y} = [y_1, \ldots, y_T] \rightarrow$ text

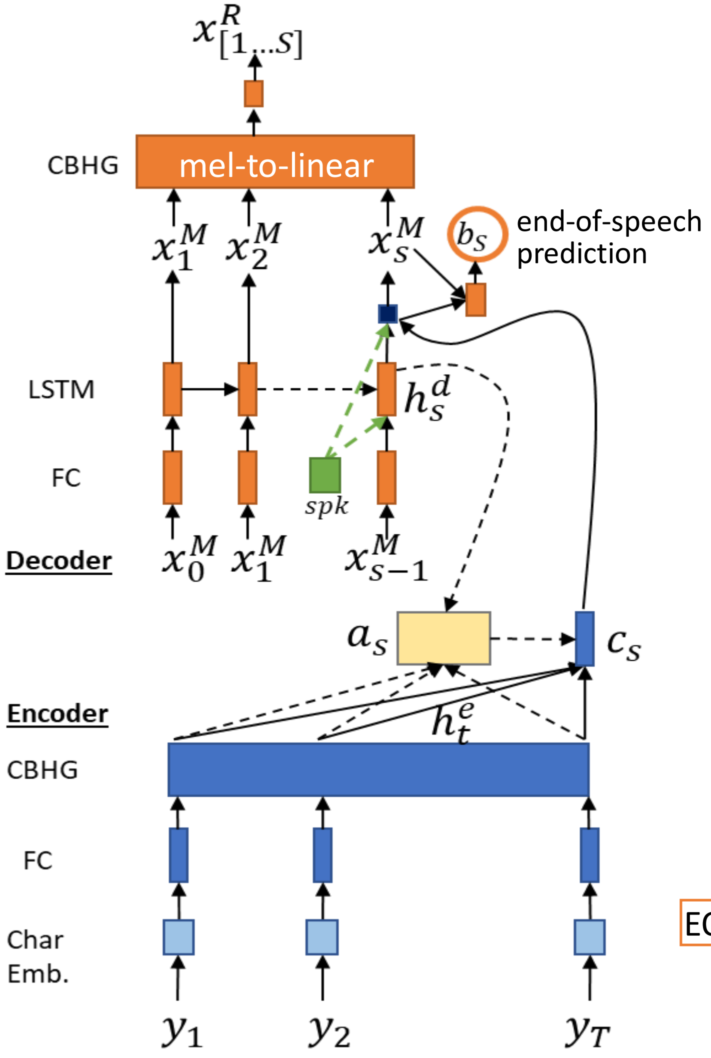
Model states

- $h^e_{[1..S]}$ = encoder states
- h^d_t = decoder state at time t
- a_t = attention probability at time t
  - $a_t(s) = Align\left(h^e_s, h^d_t\right)$
  - $a_t(s) = \dfrac{\exp\left(Score\left(h^e_s, h^d_t\right)\right)}{\sum_{s=1}^{S} \exp\left(Score\left(h^e_s, h^d_t\right)\right)}$
- $c_t = \sum_{s=1}^{S} a_t(s) * h^e_s$ (expected context)

Loss function

$$\mathcal{L}_{ASR}\left(y, p_y\right) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{c\in[1..C]} 1(y_t = c) * \log p_{y_t}[c]$$

Sequence-to-Sequence TTS



Similar to [Tacotron: Wang et al., 2017]

Input & output

- $\boldsymbol{x^R} = [x_1, \dots, x_S]$ (linear spectrogram feature)
- $\boldsymbol{x^M} = [x_1, \dots, x_S]$ (mel spectrogram feature)
- $\boldsymbol{y} = [y_1, \dots, y_T]$ (text)

Model states

- $h^e_{[1..S]}$ = encoder states
- h^d_S = decoder state at time t
- a_S = attention probability at time t
- $c_S = \sum_{s=1}^{S} a_s(t) * h_t^e$ (expected context)

Loss function

Reconst. MSE $\mathcal{L}_{TTS1}(x, \hat{x}) = \frac{1}{S}\sum_{s=1}^{S}(x_s^M - \hat{x}_s^M)^2 + (x_s^R - \hat{x}_s^R)^2$

EOS cross entropy $\mathcal{L}_{TTS2}(b, \hat{b}) = -\frac{1}{S}\sum_{s=1}^{S}\left(b_s \log(\hat{b}_s) + (1 - b_s)\log(1 - \hat{b}_s)\right)$

$$\mathcal{L}_{TTS}(x, \hat{x}, b, \hat{b}) = \mathcal{L}_{TTS1}(x, \hat{x}) + \mathcal{L}_{TTS2}(b, \hat{b})$$

Experiments on Single-Speaker Speech Chain

▪ Features

Speech:

- 80 Mel-spectrogram (used by ASR & TTS)
- 1024-dim linear magnitude spectrogram (SFFT) (used by TTS)
- TTS reconstruct speech waveform by using Griffin-Lim
  to predict the phase & inverse STFT

Text:

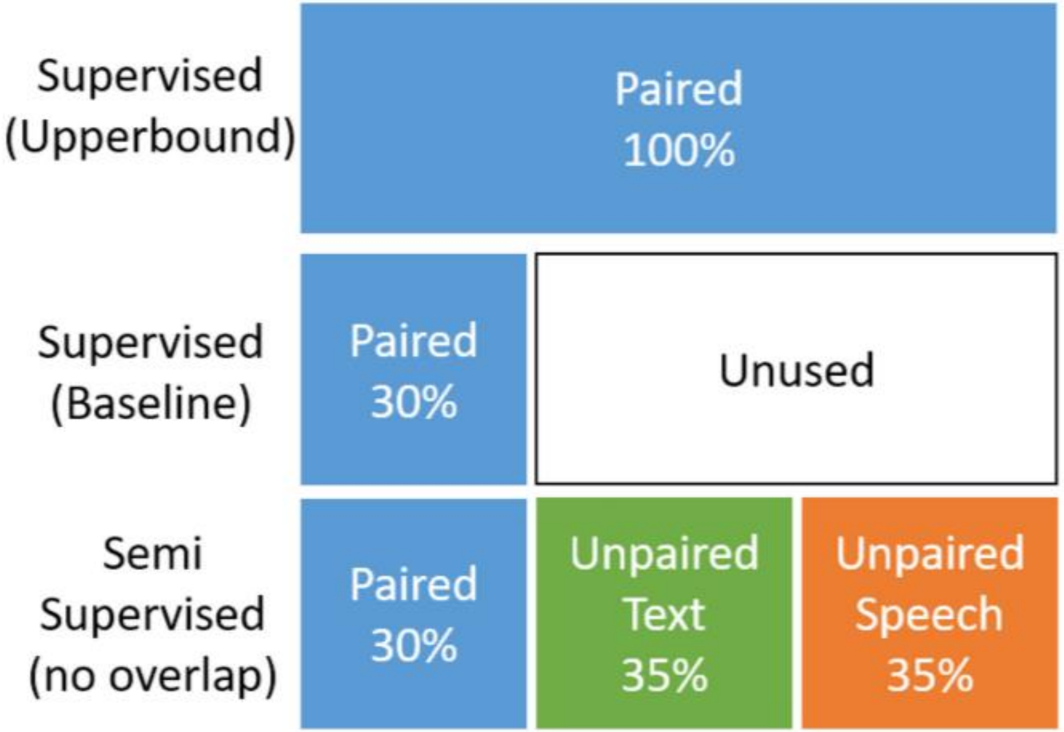
Character-based prediction

- a-z (26 alphabet)
- 6 punctuation mark (,:'?.-)
- 3 special tags <s> </s> <spc> (start, end, space)

Experiments on Single-Speaker Speech Chain

- **Data set**
  → Single speaker LJSpeech (13,100 utterances)
  → Randomly select 94% (total 12,314 utts) for training
  3% (total 393 utts) for dev set
  3% (total 393 utts) for test set

- **Evaluation**
  → ASR: Character error rate (CER)
  → TTS: L2-norm squared between the predicted and ground truth log Mel-spectrogram



ASR and TTS Results

- **ASR**

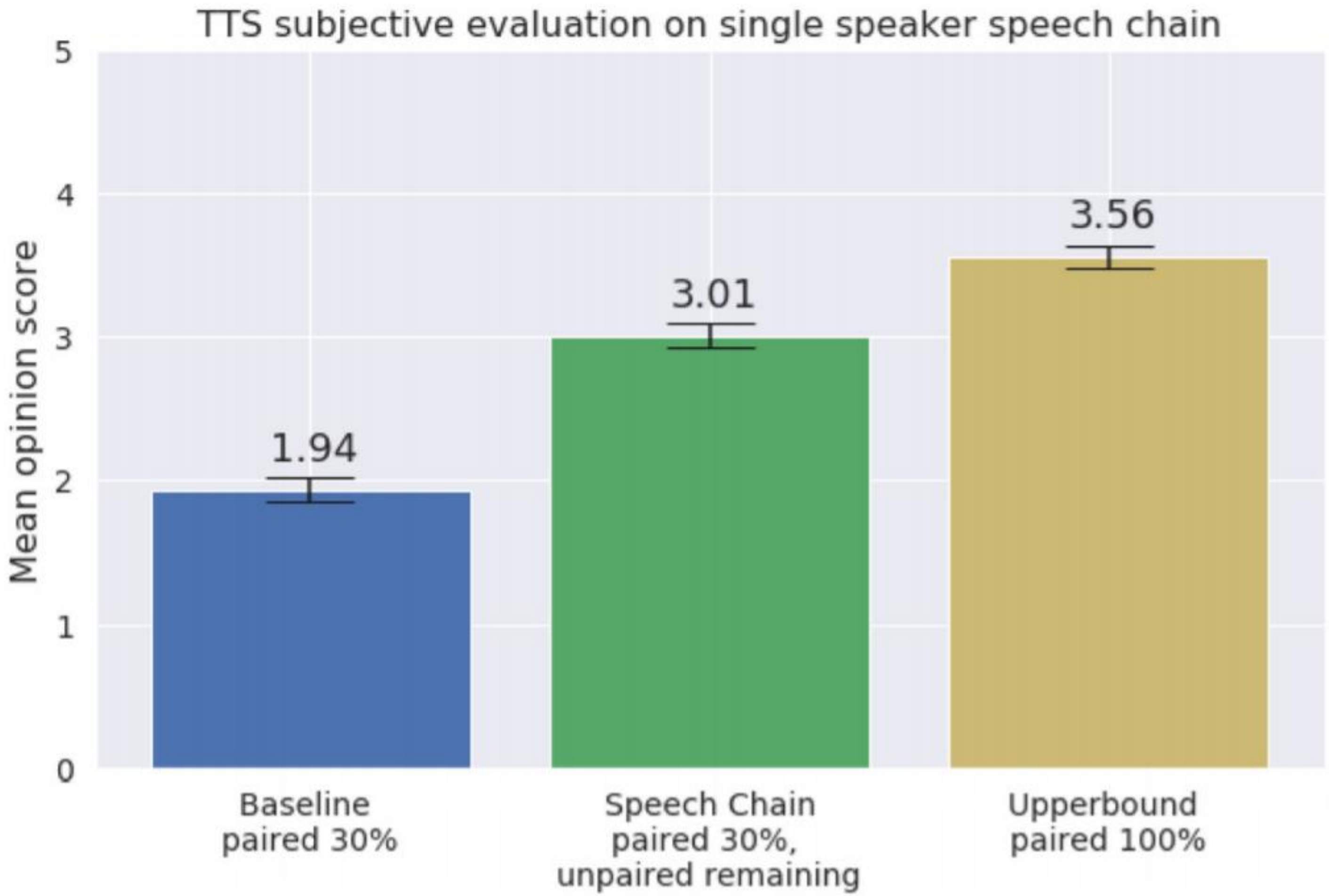
Model	Paired	Unpaired		CER (%)
		Text	Speech	
Supervised (Baseline)				
Enc-Dec Att	10%	-	-	31.7
Enc-Dec Att	20%	-	-	9.9
Enc-Dec Att	30%	-	-	6.8
Enc-Dec Att	40%	-	-	4.9
Enc-Dec Att	50%	-	-	4.1
Semi-supervised (Speech Chain)				
Enc-Dec Att	10%	45%	45%	12.3
Enc-Dec Att	20%	40%	40%	5.6
Enc-Dec Att	30%	35%	35%	4.7
Enc-Dec Att	40%	30%	30%	3.8
Enc-Dec Att	50%	25%	25%	3.5
Supervised (Upperbound)				
Enc-Dec Att	100%	-	-	3.1

- **TTS**

Model	Paired	Unpaired		$L2\text{-norm}^2$
		Text	Speech	
Supervised (Baseline)				
Enc-Dec Att	10%	-	-	1.05
Enc-Dec Att	20%	-	-	0.91
Enc-Dec Att	30%	-	-	0.71
Enc-Dec Att	40%	-	-	0.69
Enc-Dec Att	50%	-	-	0.66
Semi-supervised (Speech Chain)				
Enc-Dec Att	10%	45%	45%	0.87
Enc-Dec Att	20%	40%	40%	0.73
Enc-Dec Att	30%	35%	35%	0.66
Enc-Dec Att	40%	30%	30%	0.65
Enc-Dec Att	50%	25%	25%	0.64
Supervised (Upperbound)				
Enc-Dec Att	100%	-	-	0.606

TTS Subjective Evaluation

- **Mean Opinion Score (MOS)**



Discussion

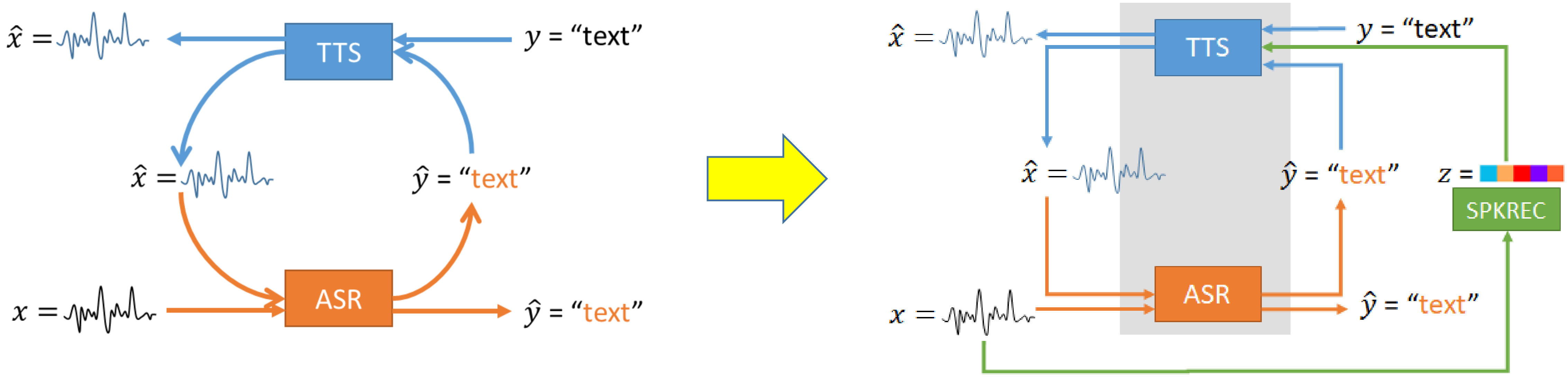
- **Summary:**
  - Inspired by human speech chain, we proposed machine speech chain to achieve semi-supervised learning
  - Enables ASR & TTS to assist each other when they receive unpaired data
  - Allows ASR & TTS to infer the missing pair and optimize the models with reconstruction loss

Part II
Multi-speaker Machine Speech Chain

[A. Tjandra, S. Sakti, S. Nakamura, "Machine Speech Chain with One-shot Speaker Adaptation", in Proc. INTERSPEECH, 2018]

Multi-Speaker Machine Speech Chain

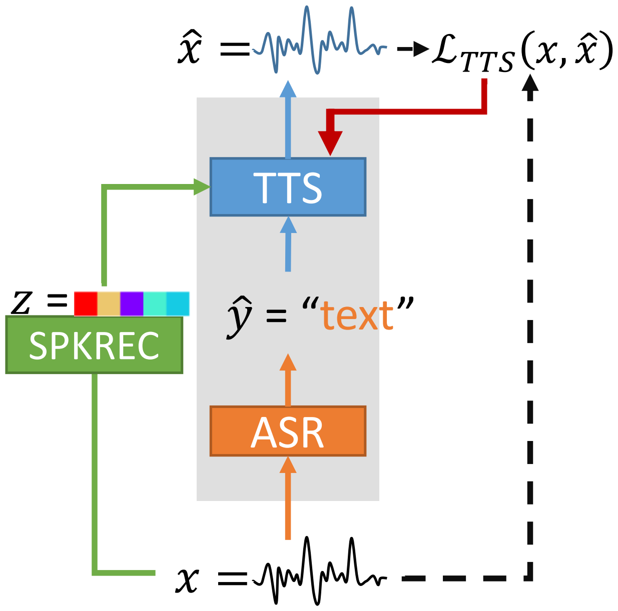
- **Motivation**
  → Basic Machine Speech Chain was able to improve single-speaker result significantly
  → Limitation: couldn't perform on unseen speaker
- **Proposed Approach: Handle voice characteristics from unknown speakers**
  → Integrate a speaker recognition system into the speech chain loop
  → Extend the capability of TTS to handle the unseen speaker using one-shot speaker adaptation



Utilizing [Deep speaker; Li et al., 2017]

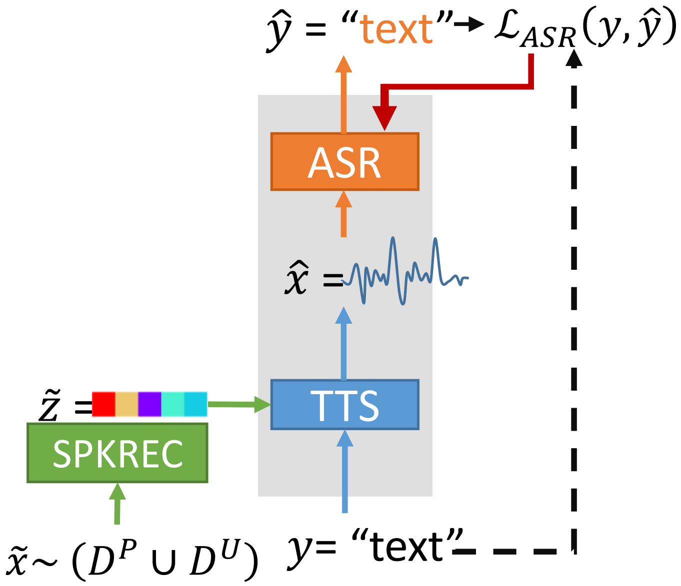
Multi-Speaker Machine Speech Chain

- **Train with Speech only: ASR→TTS**



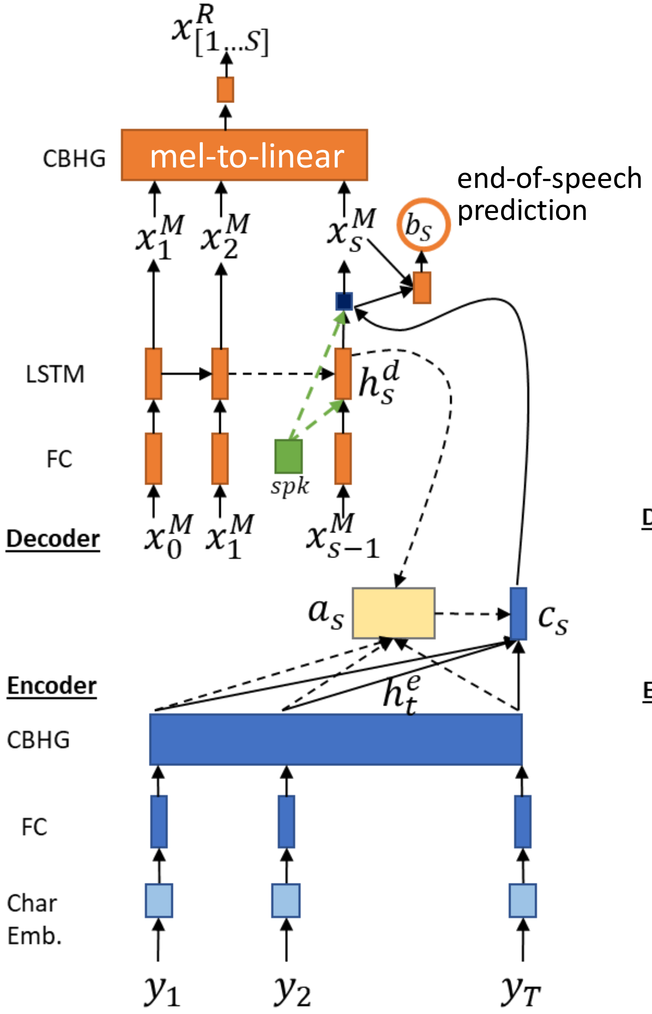
→ ASR predicts most possible transcription $\hat{y}$
→ SPKREC provide a speaker embedding z
→ TTS based on $[\hat{y}, z]$ tries to reconstruct speech $\hat{x}$

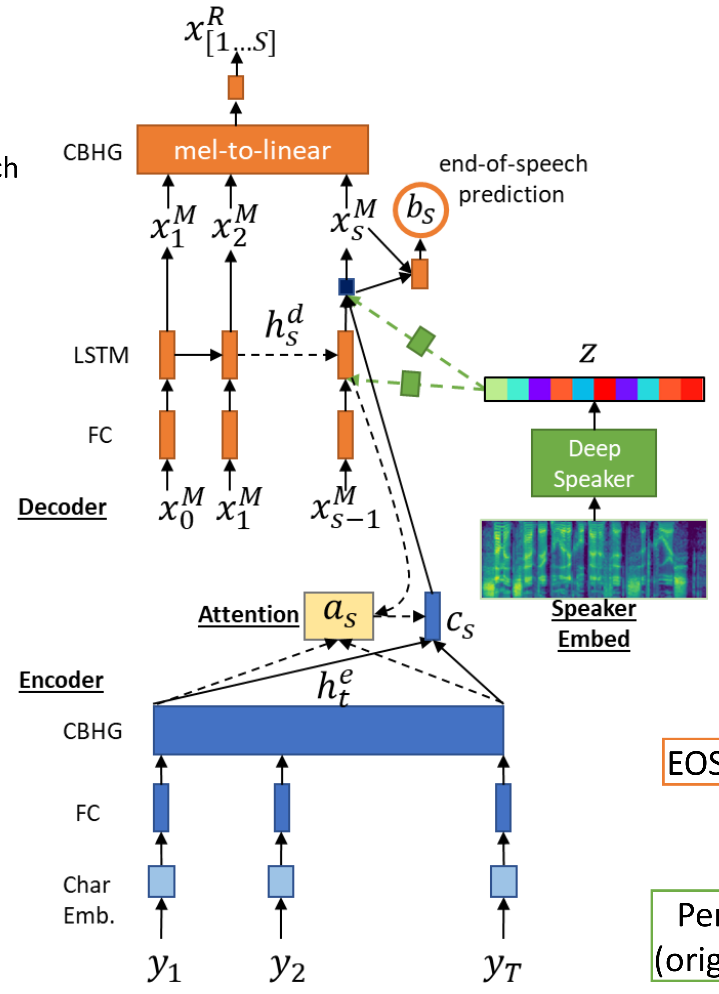
- **Train with Text only: TTS→ASR**



→ Sample a speaker vector $\tilde{z}$ from available speech
→ TTS generates speech features $\hat{x}$ based on $[y, \tilde{z}]$
→ ASR given $\hat{x}$ tries to reconstruct text $\hat{y}$

Sequence-to-Sequence TTS





Input & output

- $\boldsymbol{x^R} = [x_1, \dots, x_S]$ → linear spectrogram
- $\boldsymbol{x^M} = [x_1, \dots, x_S]$ → mel spectrogram
- $\boldsymbol{y} = [y_1, \dots, y_T]$ → text
- $\boldsymbol{z}$ → speaker embedding vector

Model states

- $h^e_{[1..S]}$ = encoder states
- h^d_s = decoder state at time t
- a_s = attention probability at time t
- $c_s = \sum_{s=1}^{S} a_s(t) * h^e_t$ (expected context)

Loss function

Reconst. MSE $\mathcal{L}_{TTS1}(x,\hat{x}) = \frac{1}{S}\sum_{s=1}^{S}(x^M_s - \hat{x}^M_s)^2 + (x^R_s - \hat{x}^R_s)^2$

EOS cross entropy $\mathcal{L}_{TTS2}(b,\hat{b}) = -\frac{1}{S}\sum_{s=1}^{S}\left(b_s \log(\hat{b}_s) + (1-b_s)\log(1-\hat{b}_s)\right)$

Perceptual loss (original vs gen sp) $\mathcal{L}_{TTS3}(z,\hat{z}) = 1 - \frac{\langle z,\hat{z}\rangle}{\|z\|_2 + \|\hat{z}\|_2}$

$$\mathcal{L}_{TTS}(x,\hat{x},b,\hat{b}) = \mathcal{L}_{TTS1}(x,\hat{x}) + \mathcal{L}_{TTS2}(b,\hat{b}) + \mathcal{L}_{TTS3}(z,\hat{z})$$

Experiments on Multi-Speakers

- **Data set**
  - **Training set: Supervised (paired text & speech)**
    - WSJ SI-84 dataset (baseline)
      (7138 utterances, ~16 h, 84 speakers)
    - WSJ SI-284 dataset (upperbound)
      (37318 utterances, ~81 h, 284 speakers)
  - **Training set: Unsupervised (unpaired text & speech)**
    - WSJ SI-200 dataset
      (30180 utterances, ~66 hours, 200 speakers)
    - Notes: SI-200 doesn't overlap with SI-84
  - **Development set:** dev93
  - **Evaluation set:** eval92

ASR and TTS Results

ASR

Model	CER (%)
Supervised training: WSJ *train_si84* (paired) → Baseline	
Att Enc-Dec [58]	17.01
Att Enc-Dec [59]	17.68
Att Enc-Dec (ours)	17.35
Supervised training: WSJ *train_si284* (paired) → Upperbound	
Att Enc-Dec [58]	8.17
Att Enc-Dec [59]	7.69
Att Enc-Dec (ours)	7.12
Semi-supervised training: WSJ *train_si84* (paired) + *train_si200* (unpaired)	
Label propagation (greedy)	17.52
Label propagation (beam=5)	14.58
Proposed speech chain (Sec. IV)	**9.86**

TTS

Model	L2-norm^2
Supervised training: WSJ *train_si84* (paired) → Baseline	
Proposed Tacotron (Sec. IV-C) (ours)	1.036
Supervised training: WSJ *train_si284* (paired) → Upperbound	
Proposed Tacotron (Sec. IV-C) (ours)	0.836
Semi-supervised training: WSJ *train_si84* (paired) + *train_si200* (unpaired)	
Proposed speech chain (Sec. IV + Sec. IV-C)	**0.886**

TTS Speech Output

- **Text:** "the busses aren't the problem, they actually provide a solution"
  - Single Speaker (LJSpeech) (p = paired, u = unpaired)

Baseline (P 30%)	Sp-Chain (S 30% + U 70%)	Full (P 100%)

  - Multispeaker (WSJ)

Speaker	Baseline (P SI84)	Sp-Chain (P si84 + U si200)	Full (P si284)
Female A			
Male B			

Discussion

- **Summary:**
  - **Improved machine speech chain to handle voice characteristics from unknown speakers**
    $\rightarrow$ TTS can generate speech with similar voice characteristic only with one-shot speaker example
    $\rightarrow$ ASR also get new data from the combination between a text sentence and an arbitrary voice characteristic
  - **By combining both models, we could train with auxiliary feedback loss**

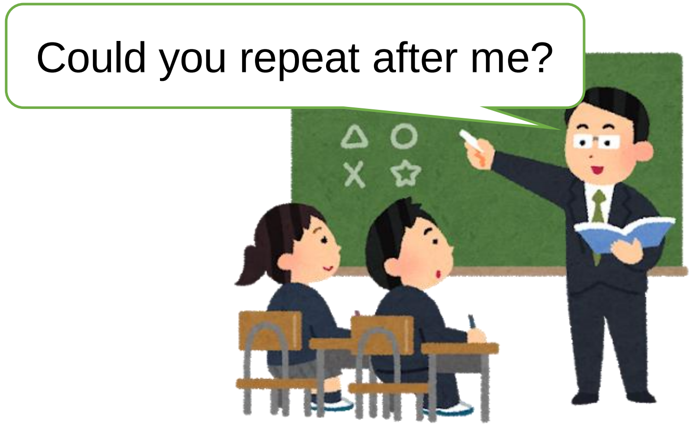
Part III
Cross-Lingual Machine Speech Chain

[S. Novitasari, A. Tjandra, S. Sakti, S. Nakamura, "Cross-Lingual Machine Speech Chain for Javanese, Sundanese, Balinese, and Bataks Speech Recognition and Synthesis", in Proc. SLTU, 2020]

Cross-Lingual Machine Speech Chain

- ## Motivation

→ Development of ASR and TTS for under-resourced languages are difficult
→ A large amount of parallel speech-text data is often unavailable
→ The human can learn a new language directly (without textbook) by listening and speaking



- ## Proposed Approach: **Learn new languages with Machine Speech Chain**

→ Listening while speaking on new languages
→ Enable to perform cross-lingual semi-supervised learning
→ No need parallel speech & text of the new language

Proposed Approach

- **Application:** **Cross-Lingual Machine Speech Chain for Javanese, Sundanese, Balinese, and Bataks ASR and TTS**

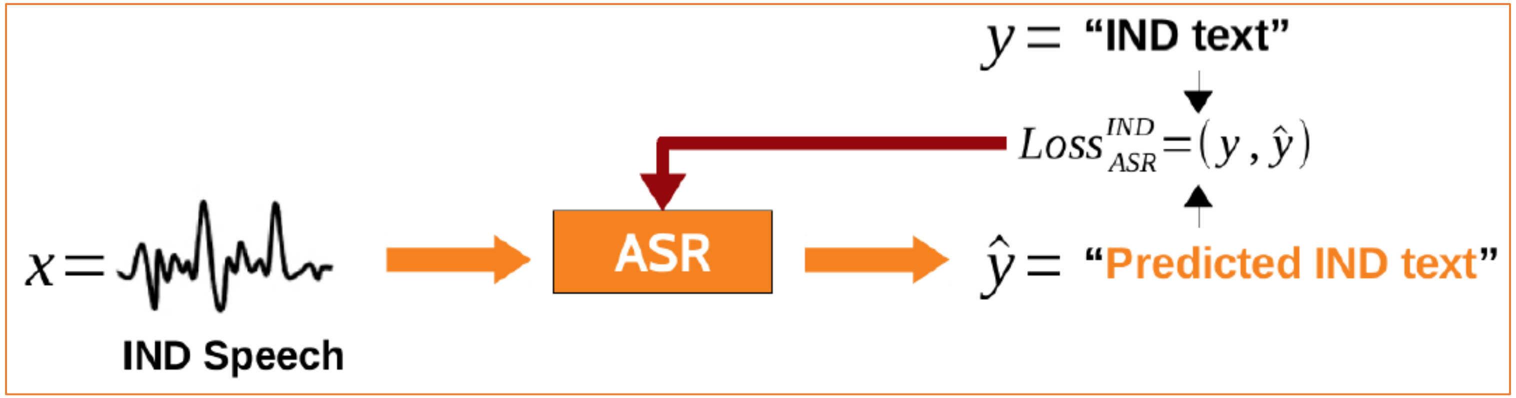
→ Indonesia is an archipelago comprising approximately **17500 islands**

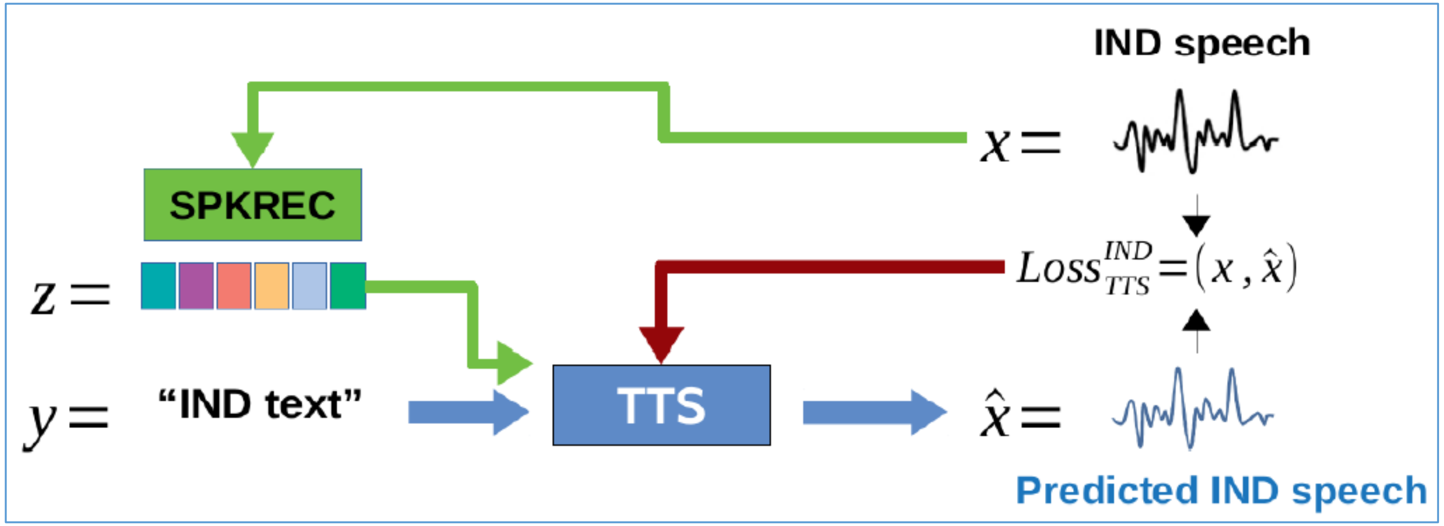
→ Approximately, there are **300 ethnic groups**, that speak **726 native languages**

→ Most of them are **under-resourced languages**

Learning Process

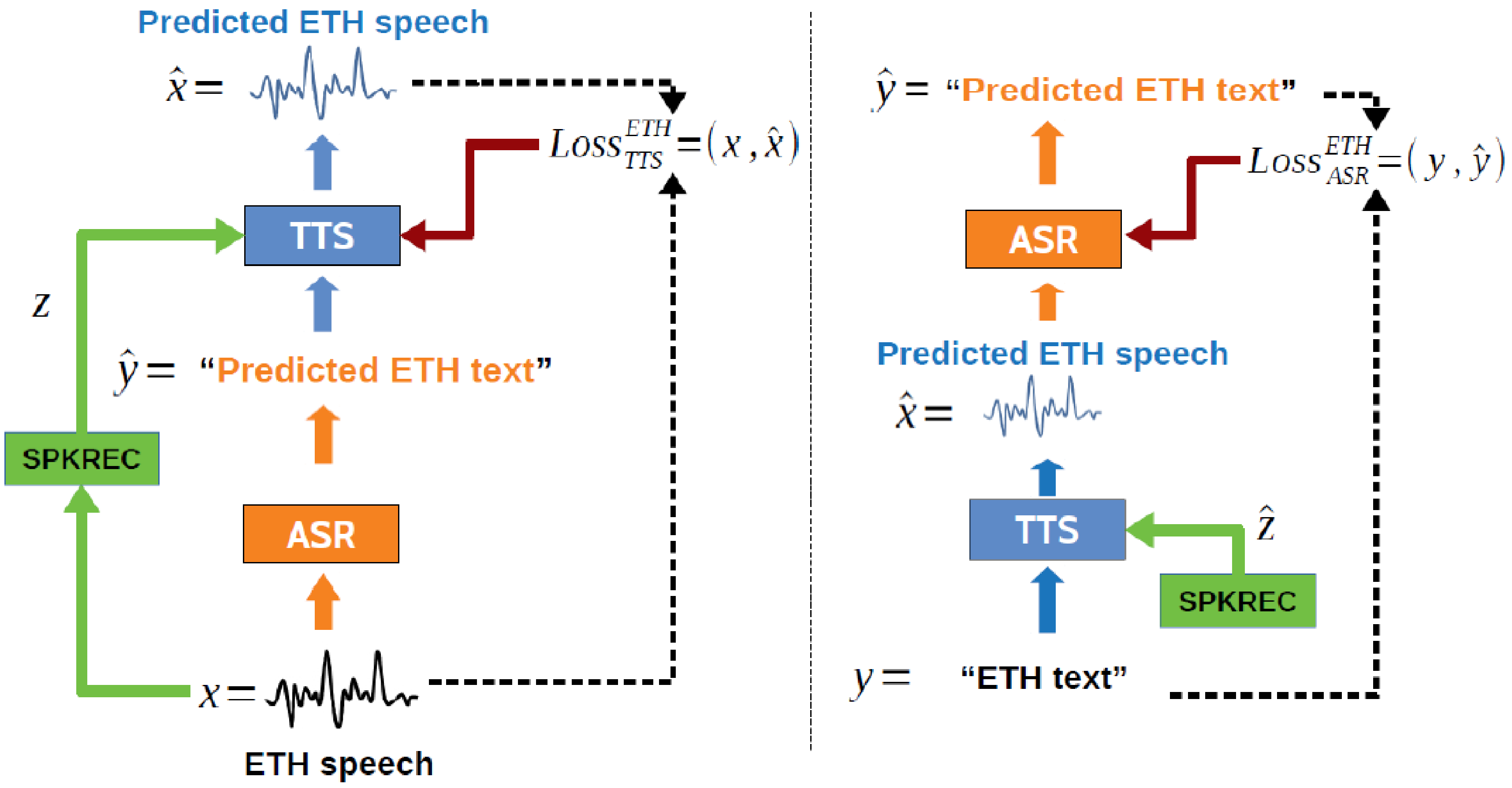
Step 1: ASR and TTS supervised training using paired speech and text of rich-resourced language (Indonesian)





Learning Process

Step 2: ASR and TTS unsupervised training using unpaired data of under-resourced languages (Indonesian ethnic languages: Javanese, Sundanese, Balinese, Bataks)



Experiments on Cross-lingual Speech Chain

- **Data set**
  - **Rich-resourced language (Indonesian language)**

    Supervised (paired text & speech)

    ⟶ Full set: 400 spkrs, 84k utterances (~80 hours of speech)

    ⟶ Test set: 10% of data (40 spkrs)

    Remaining data with 360 spkrs (20% dev set; 80% training set)

  - **Under-resourced language (Ethnics language)**

    Unsupervised (unpaired data: only text / only speech)

    ⟶ Full set: 40 spkrs (10 spkrs/languange), 325 utterances/language

    ⟶ Test set: 10% of data -- 16 spkrs (4 spkrs/language), 50 utterances/language

    Remaining data with 36 spkrs (4 spkrs/language), 225 utterances/language (10% dev set; 90% training set)

ASR and TTS Results

ASR

Training		Testing				
ASR System	**Data**	**Javanese**	**Sundanese**	**Balinese**	**Bataks**	**Avr**
Baseline IND	Sup IND (Sp+Txt)	107.26	90.70	97.98	109.85	101.45
Proposed1 IND+ETH	Sup IND (Sp+Txt) + Unsup ETH (Txt Only)	63.73	63.04	70.80	72.79	67.59
Proposed2 IND+ETH	Sup IND (Sp+Txt) + Unsup ETH (Sp+Txt)	31.96	31.97	27.00	37.37	32.08
Topline IND+ETH	Sup IND (Sp+Txt) + Sup ETH (Sp+Txt)	20.20	17.89	15.41	26.69	20.05

TTS

Training		Testing				
TTS System	**Data**	**Javanese**	**Sundanese**	**Balinese**	**Bataks**	**Avr**
Baseline IND	Sup IND (Sp+Txt)	1.016	1.247	1.129	1.254	1.162
Proposed IND+ETH	Sup IND (Sp+Txt) + Unsup ETH (Sp+Txt)	0.547	0.531	0.560	0.510	0.537
Topline IND+ETH	Sup IND (Sp+Txt) + Sup ETH (Sp+Txt)	0.415	0.470	0.478	0.399	0.441

Discussion

- **Summary:**
  - Construct ASR and TTS for ethnic languages (Javanese, Sundanese, Balinese, and Bataks, when no paired speech or text data were available.
  - Pre-trained on Indonesian with parallel speech-text in a supervised manner
  - Performed speech chain mechanism with only limited text or speech of ethnic languages (unsupervised learning)
  - Enables ASR and TTS to teach each other even without any paired data
  - The framework can be applied to any cross-lingual tasks without significant modification

Machine Speech Chain Publications

General Machine Speech Chain Framework

- A. Tjandra, S. Sakti, S. Nakamura, "Listening while Speaking: Speech Chain by Deep Learning", in Proc. IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop, 2017
- A. Tjandra, S. Sakti, S. Nakamura, "Machine Speech Chain with One-shot Speaker Adaptation", in Proc. INTERSPEECH, 2018
- A. Tjandra, S. Sakti, S. Nakamura, "End-to-end Feedback Loss in Speech Chain Framework via Straight-through Estimator", in Proc. IEEE ICASSP, 2019
- A. Tjandra, S. Sakti, S. Nakamura, "Machine Speech Chain," IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP), Vol. 28, pp. 976-989, 2020

Multilingual Machine Speech Chain

- S. Nakayama, A. Tjandra, S. Sakti, S. Nakamura, "Speech Chain for Semi-Supervised Learning of Japanese-English Code-Switching ASR and TTS", in Proc. SLT, 2018
- S. Nakayama, A. Tjandra, S. Sakti, S. Nakamura, "Zero-shot Code-switching ASR and TTS with Multilingual Machine Speech Chain,“ in Proc. IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop, 2019
- S. Novitasari, A. Tjandra, S. Sakti, S. Nakamura, "Cross-Lingual Machine Speech Chain for Javanese, Sundanese, Balinese, and Bataks Speech Recognition and Synthesis", in Proc. SLTU, 2020

Multimodal Machine Speech Chain

- J. Effendi, A. Tjandra, S. Sakti, S. Nakamura, “Listening while Speaking and Visualizing: Improving ASR through Multimodal Chain,” in Proc. IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop, 2019
- J. Effendi, A. Tjandra, S. Sakti, S. Nakamura, "Augmenting Images for ASR and TTS through Single-loop and Dual-loop Multimodal Chain Framework," in Proc. of INTERSPEECH, pp. to appear, 2020

Incremental (Real-time) Machine Speech Chain

- S. Novitasari, A. Tjandra, T. Yanagita, S. Sakti, S. Nakamura, "Incremental Machine Speech Chain for Enabling Listening while Speaking in Real-time," in Proc. of INTERSPEECH, pp. to appear, 2020

Citations

- **[Denes & Pinson, 1993]** -- P. Denes and E. Pinson, “The Speech Chain”, ser. Anchor books. Worth Publishers, 1993. [Online]. Available: https://books.google.co.jp/books? id=ZMTm3nlDfroC
- **[Bosi & Goldberg, 2003]** – M. Bosi, and R.E. Goldberg, “Introduction to digital audio coding and standards”, Boston: Kluwer Academic Pub., 2003
- **[Zen et al., 2009]** -- H. Zen, K. Tokuda, and A. Black, “Statistical parametric speech synthesis,” Speech Comm., vol. 51, no. 11, pp. 1039–1064, 2009
- **[Sakti et al., 2008]** -- Sakti, S., Kelana, E., Riza, H., Sakai, S., Markov, K., Nakamura, S., 2008. Development of Indonesian LVCSR system within A-STAR project. In: Proc. Workshop on Technologies and Corpora for Asia-Pacific Speech Translation, Hyderabad, India, pp. 19–24.
- **[Sakti et al., 2013]** -- S. Sakti, M. Paul, A. Finch, S. Sakai, T.-T. Vu, N. Kimura, C. Hori, E. Sumita, S. Nakamura, J. Park, C. Wutiwiwatchai, B. Xu, H. Riza, K. Arora, C.-M. Luong, H. Li, **"A-STAR: Toward Translating Asian Spoken Languages"**, Special issue on S2ST, Computer Speech and Language Journal (Elsevier), vol. 27, Issue 2, pp. 509-527, February 2013
- **[Sakti et al., 2015]** -- S. Sakti, O. Shagdar, F. Nashashibi, S. Nakamura, "Context awareness and priority control for ITS based on automatic speech recognition", in Proc. 14th International Conference on ITS Telecommunications, 2015
- **[Xiong et al., 2017]** -- W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu and G. Zweig, "Achieving Human Parity in Conversational Speech Recognition“, Microsoft Research Technical Report MSR-TR-2016-71, 2017
- **[Saon et al., 2017]** -- G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, D. Dimitriadis, X. Cui, B. Ramabhadran, M. Picheny, L. Lim, B. Roomi, P. Hall, "English Conversational Telephone Speech Recognition by Humans and Machines“, ASRU 2017
- **[Waldstein, 1990]** – R.S. Waldstein, “Effects of postlingual deafness on speech production: Implications for the role of auditory feedback. J. Acoust. Soc. Am. 88, 2099–2114, 1990
- **[Hickok, 2003]** – G. Hickok and B. Buchsbaum, “Temporal lobe speech perception systems are part of the verbal working memory circuit: Evidence from two recent fMRI studies. Behav. Brain Sci. 26, 740–741, 2003
- [**Hickok, 2011]** – G. Hickok, J. Houde, F. Rong, "Sensorimotor Integration in Speech Processing: Computational Basis and Neural Organization",
- Neuron Perspective, Vol. 69, Issue 3, pp. 407-422, 2011
- **[Tjandra, Sakti, and Nakamura, ASRU 2017a]** -- A. Tjandra, S. Sakti, S. Nakamura, "Attention-based Wav2Text with Feature Transfer Learning", in Proc. ASRU, 2017
- **[Tjandra, Sakti, and Nakamura, ASRU 2017b]** -- A. Tjandra, S. Sakti, S. Nakamura, "Listening while Speaking: Speech Chain by Deep Learning", in Proc. ASRU, 2017
- **[Tjandra, Sakti, and Nakamura, INTERSPEECH 2018]** -- A. Tjandra, S. Sakti, S. Nakamura, "Machine Speech Chain with One-shot Speaker Adaptation", in Proc. INTERSPEECH, 2018

Thank you