Data Augmentation

Instructor: Tom Ko

Challenges in Speech recognition

- Speech signals are easily affected by a lot of factors:
  - Different Speakers
  - Channels
  - Speech type: read or conversational
  - Emotions -> affect speaking rate and volume
  - Environment: noisy or clean
  - Distance to the microphone
  - Sampling rate

- No existing technique can perfectly remove all of the above variations.
- Try to reduce the variations by carefully select the training data.

Solutions

- Reduce the mismatch between training data and test data.
- Carefully select training data according to the application
  - Voice search -> Read speech training data
  - Automatic meeting transcription -> Conversational speech
  - ASR system in a big room -> Data with reverberation
- Nowadays, people is trying to solve the problem with the use of BIG data in order to cover all the variations.

Amount of training data

- (numbers reported at 2015)

System	hours
Baidu	7,000
IFLYTEK	~100,000
Microsoft	>20,000

Microsoft Cortana

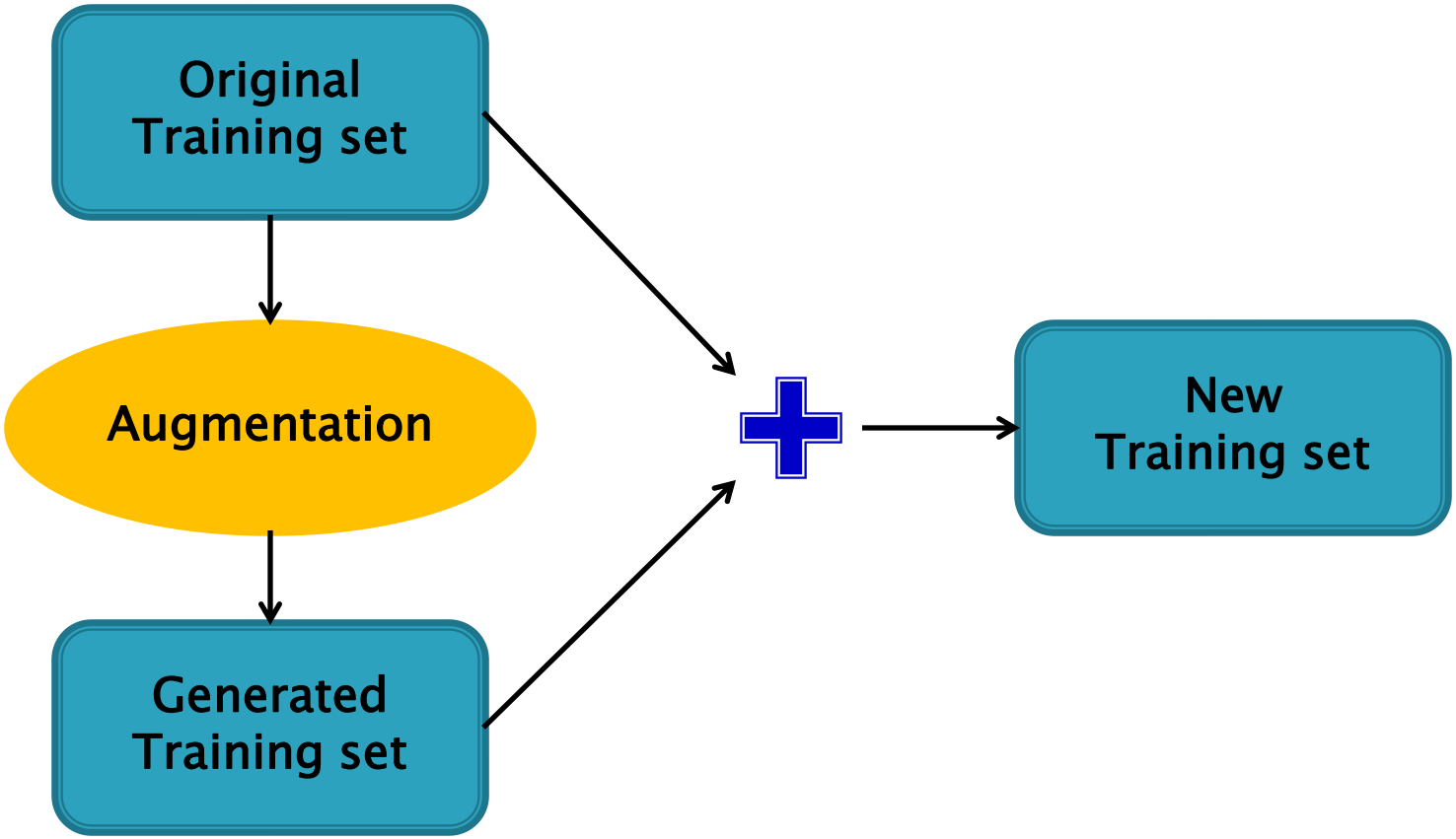
- Initialize an ASR system with 2000 hours training data.
- Put the system online and collect live data.
- Transcribe the live data and retrain the model.
- Update the system.

- Nothing is better than having live data from users.

Data Augmentation

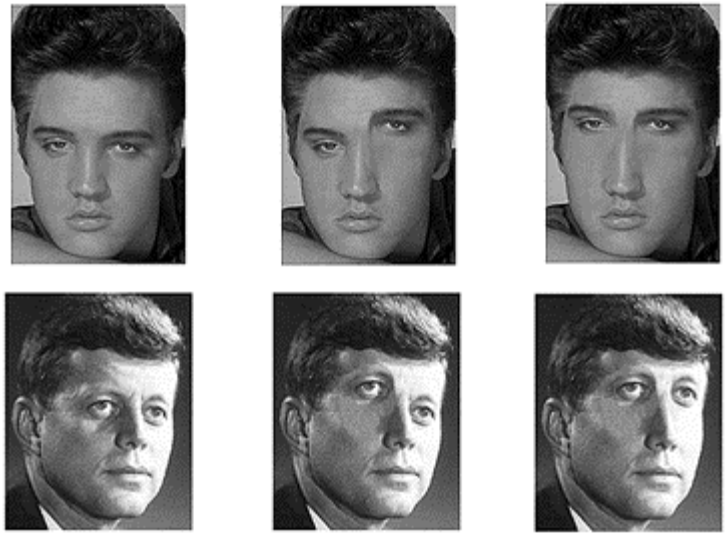
- Some speech variations can be overcome by BIG data.
- It is difficult to collect so much data. (definition of BIG is also changing)
- Motivation: to simulate / generate BIG data from existing limited data.
- Make the model generalize better
- Simulate the target scenario
- Commonly used in computer vision (CV)

Mixing with the original data



Data Augmentation for CV

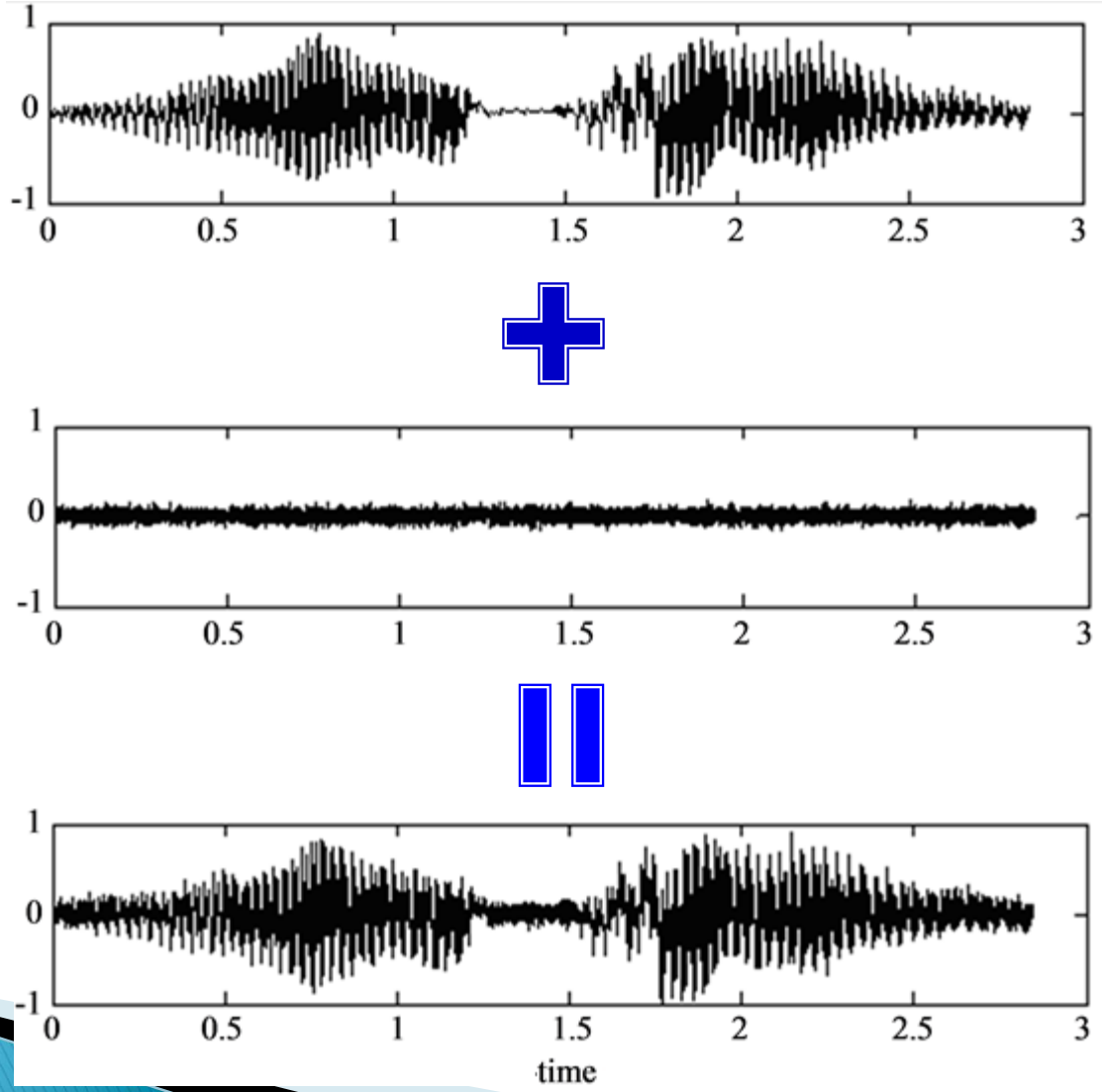
- In CV, augmentation methods involve:
  - Flipping
  - Zooming
  - Rotation
  - Cropping
  - Noise injection



Data Augmentation for speech

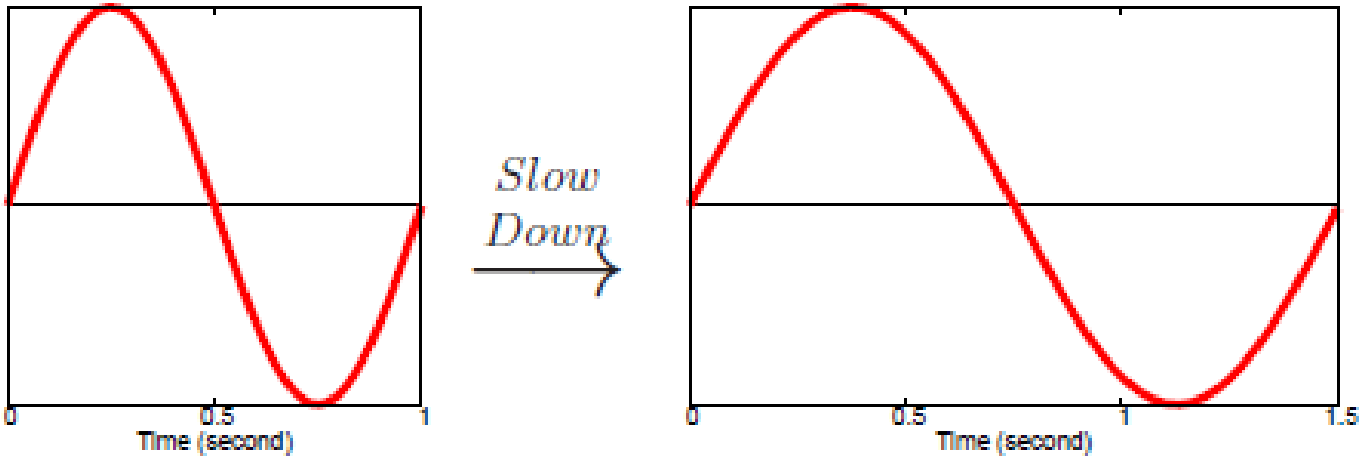
- Noise addition
- Speech perturbation
- Reverberation simulation
- Spectrum augmentation

Noise addition



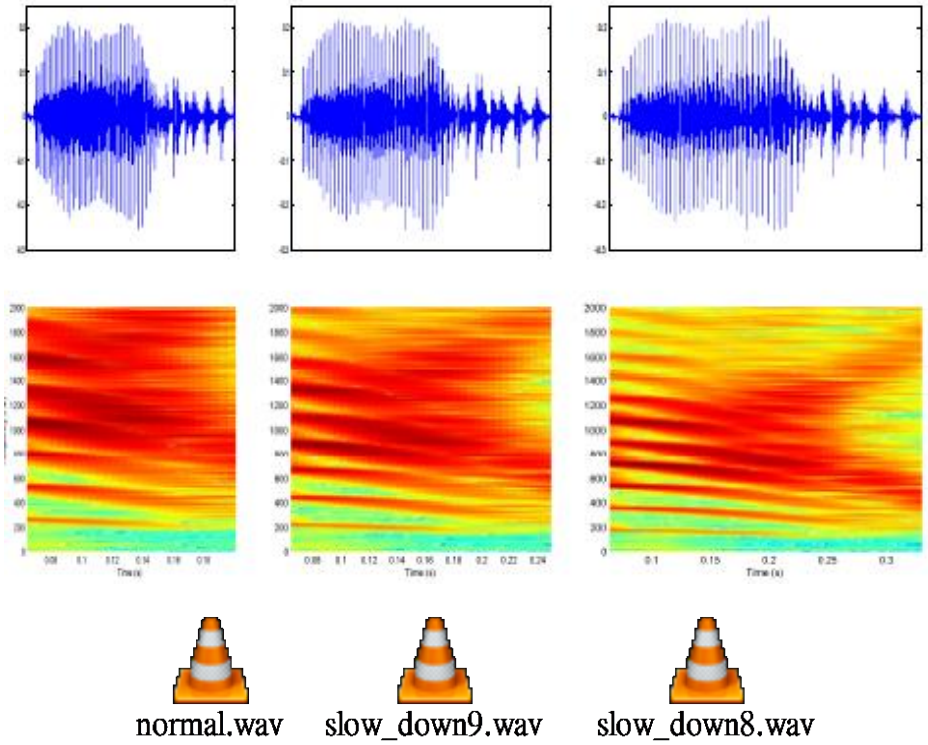
Speed perturbation

- Change the speed of the speech signal.



- It is equivalent to changing pitch and speaking rate.

Speed perturbation



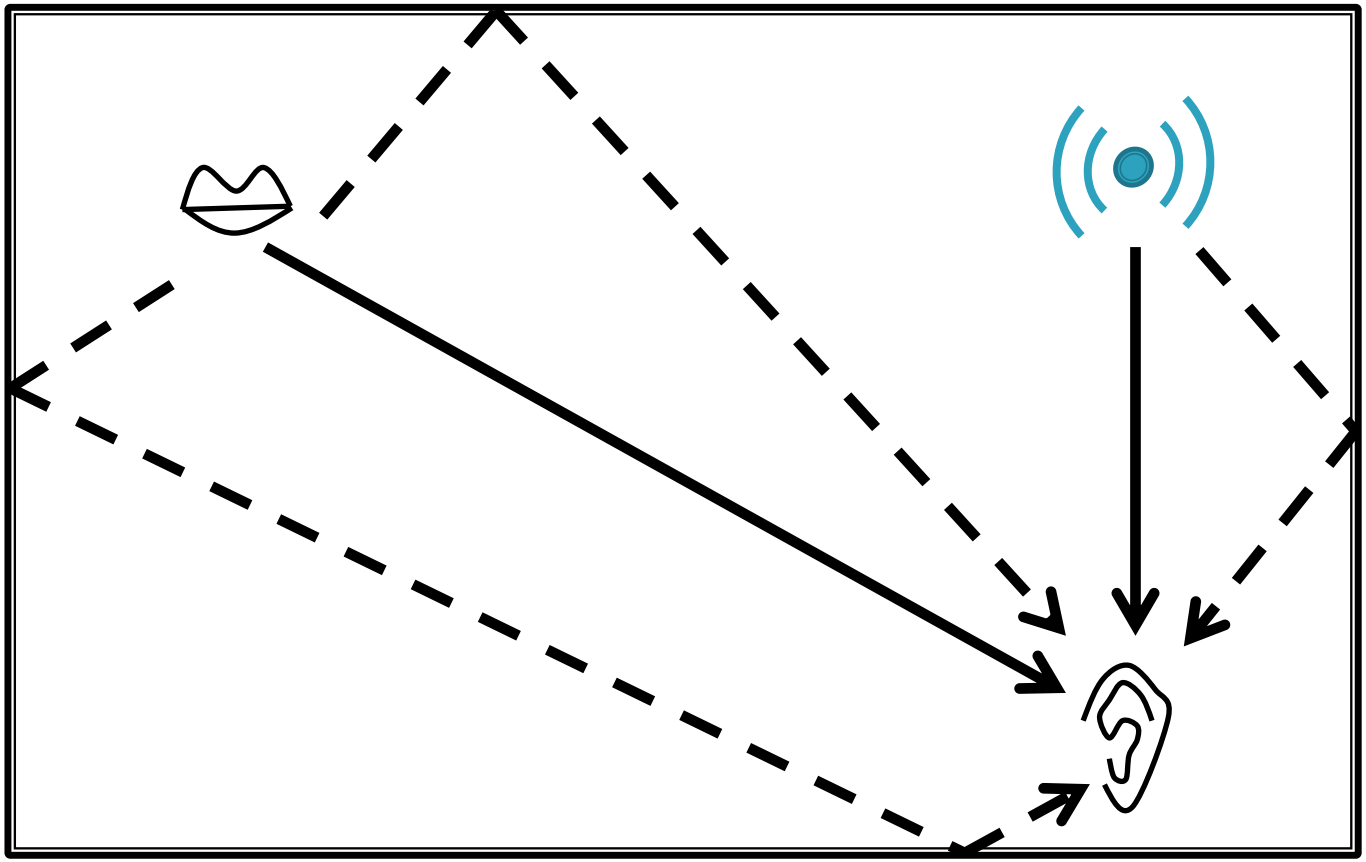
- In practice, the original data is augmented to make a 3-fold data with 10% faster speech and 10% slower speech

Experimental Results

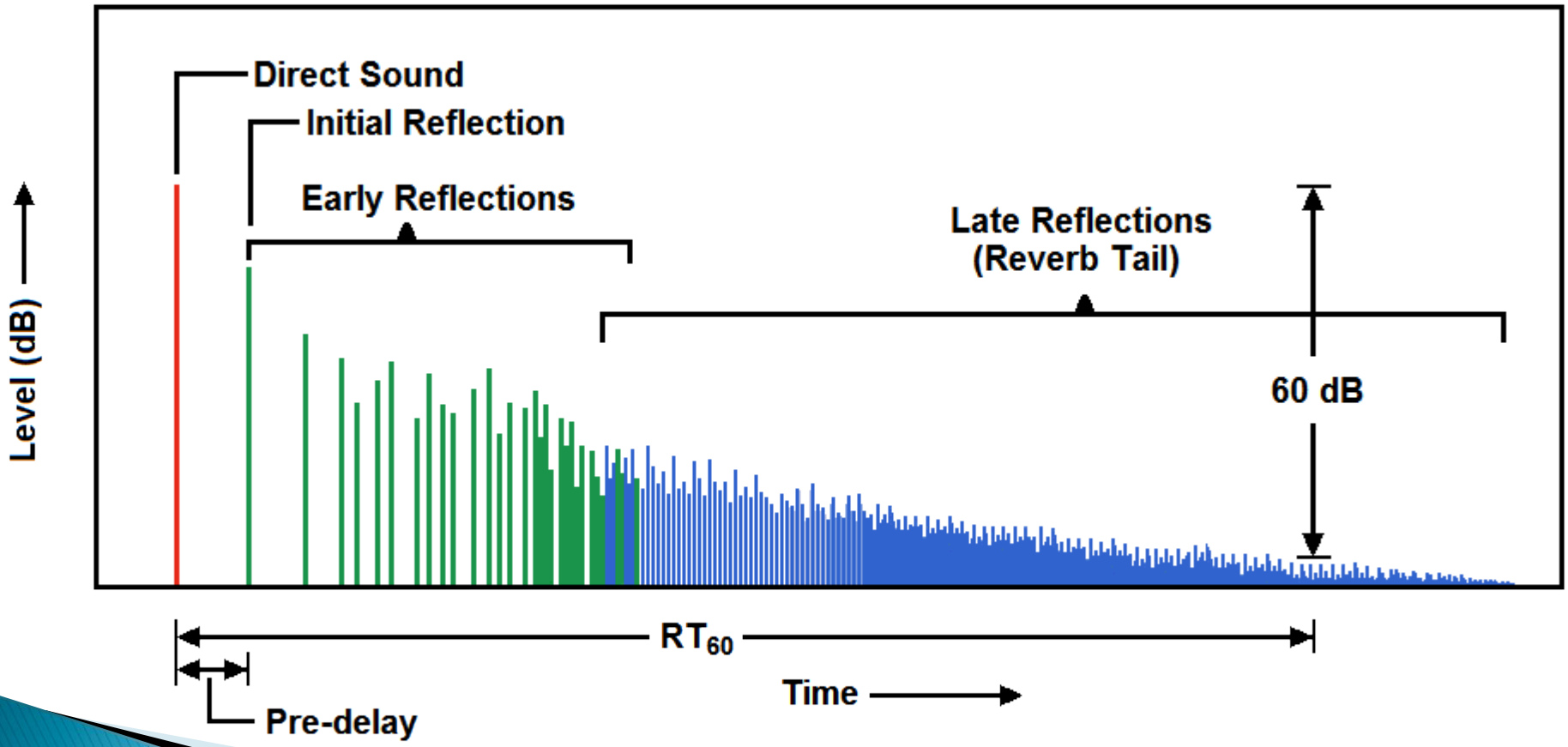
LVCSR task	Hours	Baseline	Speed-perturbed	Rel. Improvement
GALE Mandarin	100	17.51	17.16	**2.0**
Tedlium	118	17.9	17.2	**3.9**
Switchboard	300	20.7	19.3	**6.7**
Librispeech	960	12.93	12.51	**3.2**

" Audio Augmentation for Speech Recognition", **Tom Ko**, Vijayaditya Peddinti, Daniel Povey, Sanjeev Khudanpur in *Proceedings of Interspeech, September, 2015, Dresden, Germany*

Reverberation



Room Impulse Response (RIR)



Objectives

- Using clean training data will have 20-30% accuracy drop in reverberant environment
- Using reverberant training data can improve robustness of DNN-based acoustic models.
- Problem: Real reverberant data is difficult to collect
- Solution: Collect the RIRs and reverberate the data by myself

Measuring real RIRs

Measuring real RIRs

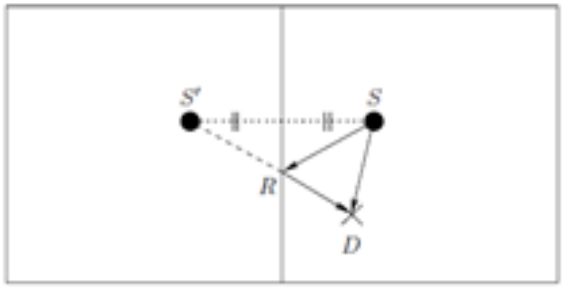
Measuring real RIRs

Motivation

- Real room impulse responses (RIRs) are difficult to acquire.
- To examine the impact of using simulated room impulse responses (RIRs)

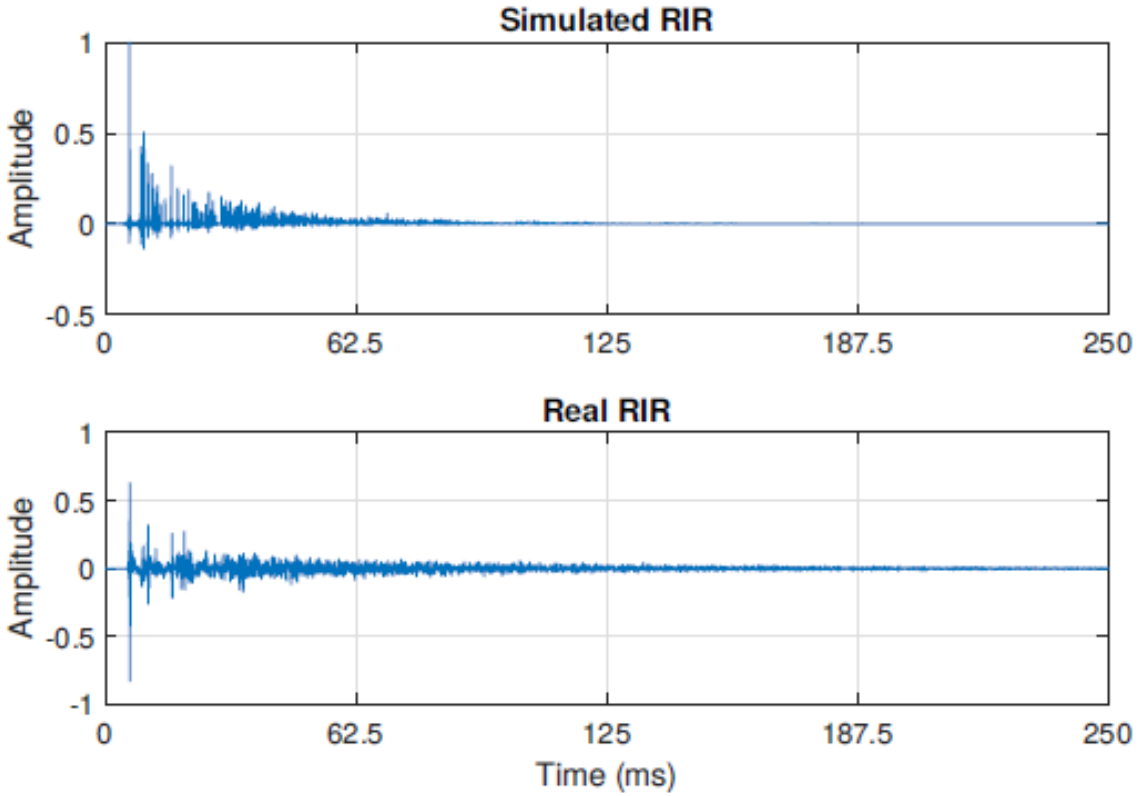
Simulation of RIRs

- The Image Method



- Room parameters and receiver position are uniformly sampled from the following ranges.
- Room width & length:
  - (small room set): 1 – 10m.
  - (medium room set): 10 – 30m.
  - (large room set): 30 – 50m.
- Room height: 2 – 5m
- Absorption coefficient of walls: 0.2 – 0.8

Comparison of simulated and real RIRs



Reverberation of training data

- For each speech recording,
  - A room is sampled based on P(r)
  - An RIR from the room is sampled based on P(h|r) to convolve with the speech recording
  - A noise recording is randomly selected
    - A separate RIR is sampled to convolve the noise.
  - Interpolation

RIRs and noises

- 325 real RIRs
  - RWCP, REVERB and Aachen
- 20,000 simulated RIRs from 200 rooms
- 843 point source noises from MUSAN

Experimental setup

- Evaluate on several LVCSR tasks:
  - SWBD (close-talking)
  - ASpIRE (far-field)
- Acoustic model:
  - time-delay neural network (TDNN)
  - bi-directional long-short term memory (BLSTM)

Results on Switchboard (SWBD)

- SWBD 300-hour training data.

Training data	Hours	Epoch	ASpIRE	**SWBD**
clean only (Baseline)†	900	4	56.3	15.4
Mixing reverberated and clean data:				
sim-rvb ($\mathcal{S}_{small}$)	1800	2	39.4	15.2
sim-rvb ($\mathcal{S}_{med}$)	1800	2	40.4	14.9
sim-rvb ($\mathcal{S}_{large}$)	1800	2	41.3	15.0
real-rvb	1800	2	38.6	14.9
With addition of noises:				
real-rvb + point-source	1800	2	34.7	15.1
sim-rvb ($\mathcal{S}_{med}$) + point-source	1800	2	34.9	15.0
sim-rvb ($\mathcal{S}_{med}$) + real-rvb + point-source	1800	2	34.3	15.0

† : We perform a 3-fold augmentation of the 300-hour SWBD data to create a total of 900 hours of training set using *Speed Perturbation*

Results on ASpIRE

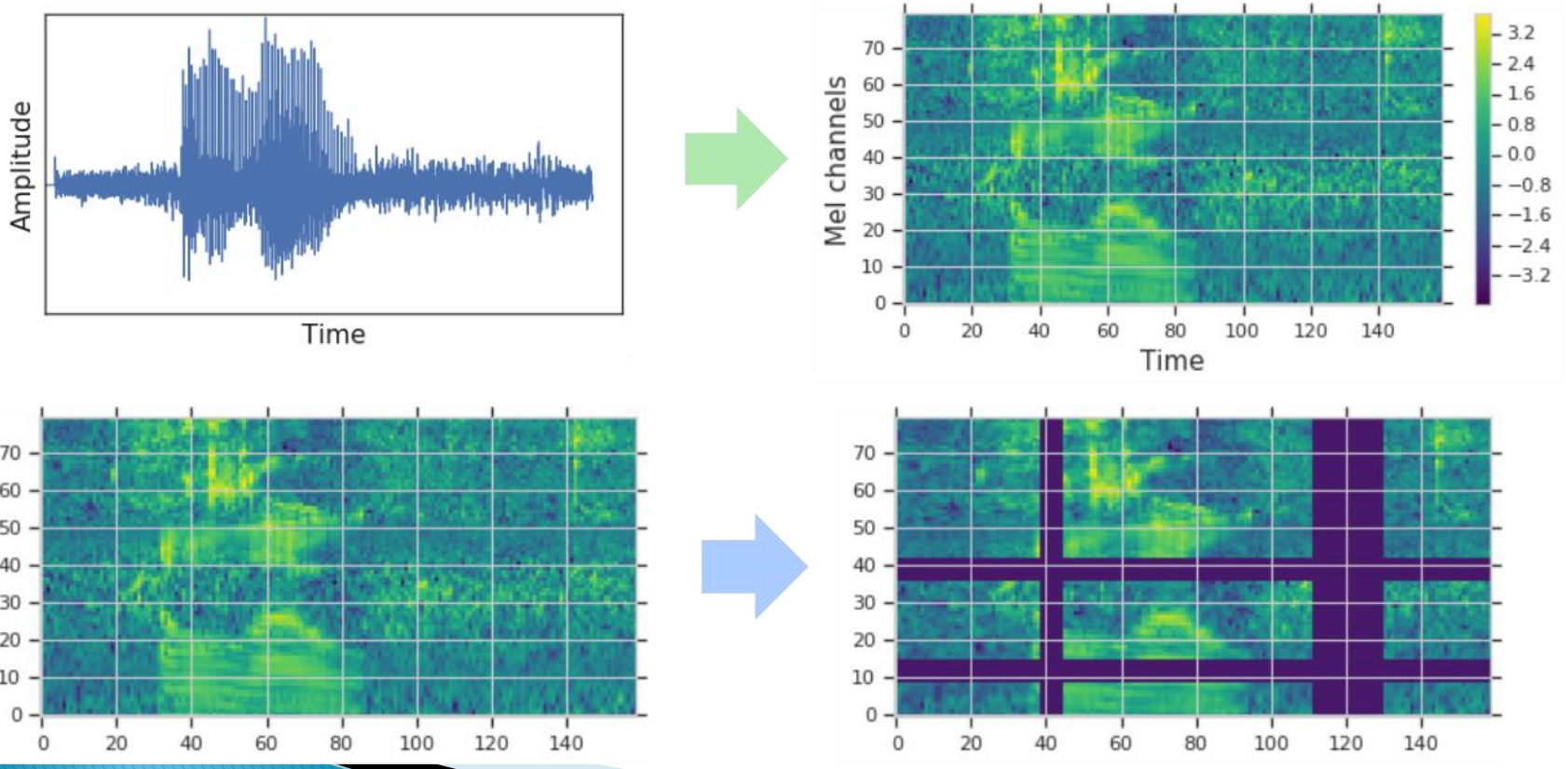
Table 2. Results on ASpIRE *dev* set with $\sim$ 5100 hours of training data.

Training data	Model	Objective	WER
clean only	TDNN	CE	45
real-rvb + isotropic	TDNN	CE	31.0
clean only	TDNN	LF-MMI	40.9
real-rvb + isotropic	TDNN	LF-MMI	27.8
sim-rvb ($\mathcal{S}_{med}$) + point-source	TDNN	LF-MMI	27.0
real-rvb + isotropic	BLSTM	LF-MMI	25.7
sim-rvb ($\mathcal{S}_{med}$) + point-source	BLSTM	LF-MMI	**24.6**

" A study on data augmentation of reverberant speech for robust speech recognition", **Tom Ko**, Vijayaditya Peddinti, Daniel Povey, Michael L. Seltzer, Sanjeev Khudanpur in *ICASSP 2017*

Spectrogram augmentation

- SpecAugment applies an augmentation policy directly to the audio spectrogram



Performance of SpecAugment

	LibriSpeech 960h		Switchboard 300h	
	test-clean	test-other	Switchboard	CallHome
Previous SOTA	2.95	7.50	8.3	17.3
Our Results	**2.5**	**5.8**	**6.8**	**14.1**

" SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition", **Daniel et.al** in *Interspeech 2019*

Data Augmentation in NLP

- In natural language processing (NLP) field, it is hard to augment text due to high complexity of language.
- Consider you are working on a text classification task to classify if the sentence is a question or statement.
  - 你知不知道怎樣去學校
  - 可以教我怎樣踢波嗎
  - 我該做什么呢
  - 你懂不懂做這題
  - 有什么值得高興
  - 他怎么樣
  - 為何他不去
  - 為什么我過不了
  - 這次作業我多少分
  - 有沒有拿到學位
  - 從哪來的
  - 有沒有錢
  - 可否幫我檢查一下
  - 你會么

Data Augmentation in NLP

- Imagine how people express when they want to check their bank account:
  - 我的銀行戶口有多少錢
  - 我要查一下帳戶餘額
  - 我要看看帳目
  - 告訴我我的總金額數目
  - 我有多少錢剩下
  - 我還有多少銀兩
  - 我要知道我的貨幣總值
  - 我的金子怎樣了
  - 秀一下我的財富
  - 顯示我的財寶

Data Augmentation in NLP

- There is never a perfect augmentation approach in NLP.
- However, the followings are approaches which people find useful in particular scenarios:
  - Word level augmentation
  - Character level augmentation
  - Back translation

Word level replacement

- Generate new sentences by replacing the original word with any of its similar words.
  - The thesaurus approach
  - Word embedding approach

The thesaurus approach

- The **thesaurus** is a specific **dictionary** that presents synonyms (words that have similar meaning) for every word listed.

- Original:
  - This **article** will focus on summarizing data augmentation **techniques** in NLP.

- Augmented Text:
  - This **write-up** will focus on summarizing data augmentation **methods** in NLP.

The word embedding approach

- Look for the similar words in the word embedding space.

- Original:
  - The quick **brown** fox jumps over the lazy dog.

- Augmented Text:
  - The quick **gray** fox jumps over the lazy dog.

Non-contextual word embeddings

- For word embeddings estimated from word2vec, they are non-contextual word embeddings.

- Original:
  - Do you like playing **football**?

- Augmented Text:
  - Do you like playing **soccer**?

- Original:
  - How much is a **football**?

- Augmented Text:
  - How much is a **soccer**?

Contextualized word embeddings

- The word “apple” has a different semantic meaning according to the context.
- With a contextualized language model, the word “apple” would have a different vector representation according to its context.

- Original:
  - I like **apple** pies.
- Augmented Text:
  - I like **orange** pies.
- Original:
  - I like **Apple** notebooks.
- Augmented Text:
  - I like **Lenovo** notebooks.

Character level augmentation

- Simulate the real-world text, e.g. typo

- Original:
  - The quick brown fox jumps over the lazy dog
- Augmented Text:
  - The quick brown fox jumps over the lazy **d0g**

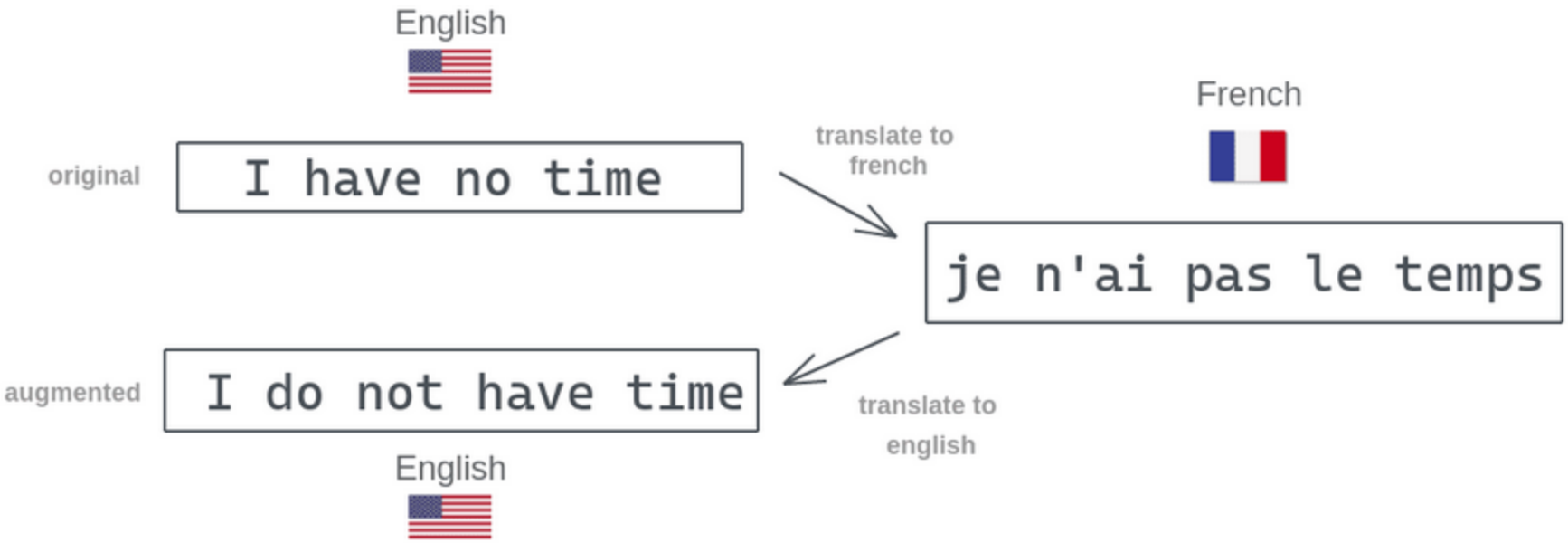
- Original:
  - The quick brown fox jumps over the lazy dog
- Augmented Text:
  - The **2uick hrown gox** jumps over the lazy dog

Character level replacement

Character	Possible Replacement
0	o (zero), O
1	l (lower case of L), I (upper case of i)

e	2, @, 3, #, 4, $, w, r, s, d, f
h	t, y, u, g, j, b, n, m

Back translation



- This works for application like text classification.

Reading list

" Audio Augmentation for Speech Recognition", **Tom Ko**, Vijayaditya Peddinti, Daniel Povey, Sanjeev Khudanpur in *Proceedings of Interspeech, September, 2015, Dresden, Germany*

- " A study on data augmentation of reverberant speech for robust speech recognition", **Tom Ko**, Vijayaditya Peddinti, Daniel Povey, Michael L. Seltzer, Sanjeev Khudanpur in *ICASSP 2017*