

Speech Technology: Frontiers and Applications

Hot Topics in Speech Recognition

Xiangang Li, Guoguo Chen



Outline

- Far-field speech recognition
- Mix-lingual speech recognition
- “Low resource” speech recognition
- Homework

- Far-field speech recognition
- Mix-lingual speech recognition
- “Low resource” speech recognition
- Homework

Far-field: the task



Siri released in 2011
near-field



Echo released in 2014
far-field

Far-field: the difficulties

- Low SNR
 - Background noise
 - Speaker playback
 - Microphone sensitivity
- High reverberation
 - Reflections
 - Diffusions



Reverberation Example. Clean signal, followed by different versions of reverberation (with longer and longer decay times).



Close talk example.



Far-field example.

Far-field: data augmentation

- The idea
 - Augments the training data as if it were collected in a far-field setting
- Methods
 - Reverb augmentation 收集冲击函数, 卷积噪声
 - Noise augmentation 加性噪声
 - Volume perturbation
 - Speed perturbation 1. 改速度不改音调(sox tempo); 2. 改速度改音调(sox speed)
 - Frequency masking
 - Time masking
 - ...

Far-field: data augmentation

- The ASPIRE Challenge
 - IARPA's challenge for far-field ASR
 - Training data: English portion of the Fisher database
 - Test data: collected in noisy and reverberant environments
- The challenges
 - Training/test mis-match
 - Unknown devices
 - Unknown acoustic space

Far-field: data augmentation

- The ASPIRE Challenge
 - IARPA's challenge for far-field ASR
 - Training data: English portion of the Fisher database
 - Test data: collected in noisy and reverberant environments
- The challenges
 - Training/test mis-match
 - Unknown devices
 - Unknown acoustic space



Far-field: data augmentation

- The ASPIRE Challenge
 - IARPA's challenge for far-field ASR
 - Training data: English portion of the Fisher database
 - Test data: collected in noisy and reverberant environments
- The challenges
 - Training/test mis-match
 - Unknown devices
 - Unknown acoustic space

JHU ASPIRE SYSTEM : ROBUST LVCSR WITH TDNNS, IVECTOR ADAPTATION AND RNN-LMS

Vijayaditya Peddinti¹, Guoguo Chen¹, Vimal Manohar¹, Tom Ko³ Daniel Povey^{1,2}, Sanjeev Khudanpur^{1,2}

¹Center for language and speech processing &
²Human Language Technology Center of Excellence
The Johns Hopkins University, Baltimore, MD 21218, USA
³Huawei Noah's Ark Research Lab, Hong Kong, China

{vijay.p, guoguo, khudanpur}@jhu.edu, {vimal.manohar91, tomkocse, dpovey}@gmail.com

Reverberation robust acoustic modeling using i-vectors with time delay neural networks

Vijayaditya Peddinti¹, Guoguo Chen¹, Daniel Povey^{1,2}, Sanjeev Khudanpur^{1,2}

¹Center for language and speech processing &
²Human Language Technology Center of Excellence
The Johns Hopkins University, Baltimore, MD 21218, USA
{vijay.p, guoguo, khudanpur}@jhu.edu, dpovey@gmail.com

Abstract

In reverberant environments there are long term interactions between speech and corrupting sources. In this paper a time delay neural network (TDNN) architecture, capable of learning long term temporal relationships and translation invariant representations, is used for reverberation robust acoustic modeling. Further, i-vectors are used as an input to the neural network to perform instantaneous speaker and environment adaptation, providing 10% relative improvement in word error rate. By subsampling the outputs at TDNN layers across time steps, training time is reduced. Using a parallel training algorithm we show that the TDNN can be trained on ~ 5500 hours of speech data in 3 days using up to 32 GPUs. The TDNN is shown to provide results competitive with state of the art systems in the IARPA ASPIRE challenge, with 27.7% WER on the dev test set.

Index Terms: far field speech recognition, time delay neural networks, reverberation

the input context up to 280 milliseconds. The ability to process such a wide temporal context enables the network to deal with late reverberations.

i-vectors which capture both speaker and environment specific information have been shown to be useful for rapid adaptation of the neural network [7, 8, 9]. i-vector based adaptation has also been shown to be effective in reverberant environments [10]. In this paper we use this adaptation technique.

We show experimental results on the ASPIRE far-field speech recognition challenge held by IARPA [11]. This challenge uses the English portion of the Fisher database [12] for acoustic and language model training. We show that in this large data scenario the proposed network architecture, combined with a parallel training technique [13], can train on multi-condition training data of ~ 5500 hours, using up to 32 GPUs, in 3 days.

Using the TDNN architecture helps us to achieve results close to those of the best combined system submitted to the AS-

1. JHU ASPIRE system: robust LVCSR with TDNNs, ivector adaptation and RNN-LMs
2. Reverberation robust acoustic modeling using i-vectors with time delay neural networks

Far-field: data augmentation

Table 1: Comparison of input contexts and training data augmentation, used for training the TDNNs

Acoustic Model	context	training data	dev WER
TDNN A	[-13, 9]	clean	47.6
TDNN A	[-13, 9]	rvb	31.7
TDNN B	[-16, 12]	rvb	30.8
TDNN B	[-16, 12]	rvb + sp	31.0
TDNN C	[-22, 12]	rvb + sp	31.1
DNN	[-16, 12]	rvb	33.1

rvb : reverberation of training data using real world RIRs

sp : speed perturbation of data prior to reverberation

JHU ASPIRE SYSTEM : ROBUST LVCSR WITH TDNNS, IVECTOR ADAPTATION AND RNN-LMS

Vijayaditya Peddinti¹, Guoguo Chen¹, Vimal Manohar¹, Tom Ko³ Daniel Povey^{1,2}, Sanjeev Khudanpur^{1,2}

¹Center for language and speech processing &
²Human Language Technology Center of Excellence
The Johns Hopkins University, Baltimore, MD 21218, USA
³Huawei Noah's Ark Research Lab, Hong Kong, China
{vijay.p, guoguo, khudanpur}@jhu.edu, {vimal.manohar91, tomkocse, dpovey}@gmail.com

Reverberation robust acoustic modeling using i-vectors with time delay neural networks

Vijayaditya Peddinti¹, Guoguo Chen¹, Daniel Povey^{1,2}, Sanjeev Khudanpur^{1,2}

¹Center for language and speech processing &
²Human Language Technology Center of Excellence
The Johns Hopkins University, Baltimore, MD 21218, USA
{vijay.p, guoguo, khudanpur}@jhu.edu, dpovey@gmail.com

Abstract

In reverberant environments there are long term interactions between speech and corrupting sources. In this paper a time delay neural network (TDNN) architecture, capable of learning long term temporal relationships and translation invariant representations, is used for reverberation robust acoustic modeling. Further, i-vectors are used as an input to the neural network to perform instantaneous speaker and environment adaptation, providing 10% relative improvement in word error rate. By subsampling the outputs at TDNN layers across time steps, training time is reduced. Using a parallel training algorithm we show that the TDNN can be trained on ~ 5500 hours of speech data in 3 days using up to 32 GPUs. The TDNN is shown to provide results competitive with state of the art systems in the IARPA ASPIRE challenge, with 27.7% WER on the dev test set.

Index Terms: far field speech recognition, time delay neural networks, reverberation

the input context up to 280 milliseconds. The ability to process such a wide temporal context enables the network to deal with late reverberations.

i-vectors which capture both speaker and environment specific information have been shown to be useful for rapid adaptation of the neural network [7, 8, 9]. i-vector based adaptation has also been shown to be effective in reverberant environments [10]. In this paper we use this adaptation technique.

We show experimental results on the ASPIRE far-field speech recognition challenge held by IARPA [11]. This challenge uses the English portion of the Fisher database [12] for acoustic and language model training. We show that in this large data scenario the proposed network architecture, combined with a parallel training technique [13], can train on multi-condition training data of ~ 5500 hours, using up to 32 GPUs, in 3 days.

Using the TDNN architecture helps us to achieve results close to those of the best combined system submitted to the AS-

1. JHU ASPIRE system: robust LVCSR with TDNNs, ivector adaptation and RNN-LMs
2. Reverberation robust acoustic modeling using i-vectors with time delay neural networks

Far-field: data augmentation

Table 2: Comparison of systems w/ & w/o volume perturbed training data and w/ & w/o volume normalized test data

Acoustic Model	Training Data	Test Data	dev WER
TDNN B	rvb		38.3
TDNN B	rvb	vol. norm.	30.8
TDNN B	rvb +vp		33.3
TDNN B	rvb +vp	vol. norm.	30.9

rvb : reverberation of training data using real world RIRs

vp : volume perturbation of data after reverberation

JHU ASPIRE SYSTEM : ROBUST LVCSR WITH TDNNS, IVECTOR ADAPTATION AND RNN-LMS

Vijayaditya Peddinti¹, Guoguo Chen¹, Vimal Manohar¹, Tom Ko³ Daniel Povey^{1,2}, Sanjeev Khudanpur^{1,2}

¹Center for language and speech processing &

²Human Language Technology Center of Excellence
The Johns Hopkins University, Baltimore, MD 21218, USA

³Huawei Noah's Ark Research Lab, Hong Kong, China
{vijay.p, guoguo, khudanpur}@jhu.edu, {vimal.manohar91, tomkocse, dpovey}@gmail.com

Reverberation robust acoustic modeling using i-vectors with time delay neural networks

Vijayaditya Peddinti¹, Guoguo Chen¹, Daniel Povey^{1,2}, Sanjeev Khudanpur^{1,2}

¹Center for language and speech processing &

²Human Language Technology Center of Excellence
The Johns Hopkins University, Baltimore, MD 21218, USA
{vijay.p, guoguo, khudanpur}@jhu.edu, dpovey@gmail.com

Abstract

In reverberant environments there are long term interactions between speech and corrupting sources. In this paper a time delay neural network (TDNN) architecture, capable of learning long term temporal relationships and translation invariant representations, is used for reverberation robust acoustic modeling. Further, i-vectors are used as an input to the neural network to perform instantaneous speaker and environment adaptation, providing 10% relative improvement in word error rate. By subsampling the outputs at TDNN layers across time steps, training time is reduced. Using a parallel training algorithm we show that the TDNN can be trained on ~ 5500 hours of speech data in 3 days using up to 32 GPUs. The TDNN is shown to provide results competitive with state of the art systems in the IARPA ASPIRE challenge, with 27.7% WER on the dev test set.

Index Terms: far field speech recognition, time delay neural networks, reverberation

the input context up to 280 milliseconds. The ability to process such a wide temporal context enables the network to deal with late reverberations.

i-vectors which capture both speaker and environment specific information have been shown to be useful for rapid adaptation of the neural network [7, 8, 9]. i-vector based adaptation has also been shown to be effective in reverberant environments [10]. In this paper we use this adaptation technique.

We show experimental results on the ASPIRE far-field speech recognition challenge held by IARPA [11]. This challenge uses the English portion of the Fisher database [12] for acoustic and language model training. We show that in this large data scenario the proposed network architecture, combined with a parallel training technique [13], can train on multi-condition training data of ~ 5500 hours, using up to 32 GPUs, in 3 days.

Using the TDNN architecture helps us to achieve results close to those of the best combined system submitted to the AS-

1. JHU ASPIRE system: robust LVCSR with TDNNs, ivector adaptation and RNN-LMs
2. Reverberation robust acoustic modeling using i-vectors with time delay neural networks

Far-field: data augmentation

- Other tricks

- i-vector
- RNNLM
- Pronunciation modeling
- Sequence training
-

JHU ASPIRE SYSTEM : ROBUST LVCSR WITH TDNNS, IVECTOR ADAPTATION AND RNN-LMS

Vijayaditya Poddinti¹, Guoguo Chen¹, Vimal Manohar¹, Tom Ko³, Daniel Povey^{1,2}, Sanjeev Khudanpur^{1,2}

¹Center for language and speech processing &
²Human Language Technology Center of Excellence
The Johns Hopkins University, Baltimore, MD 21218, USA
³Huawei Noah's Ark Research Lab, Hong Kong, China
{vijay.p, guoguo, khudanpur}@jhu.edu, {vimal.manohar91, tomkocse, dpovey}@gmail.com

Reverberation robust acoustic modeling using i-vectors with time delay neural networks

Vijayaditya Poddinti¹, Guoguo Chen¹, Daniel Povey^{1,2}, Sanjeev Khudanpur^{1,2}

¹Center for language and speech processing &
²Human Language Technology Center of Excellence
The Johns Hopkins University, Baltimore, MD 21218, USA
{vijay.p, guoguo, khudanpur}@jhu.edu, dpovey@gmail.com

Abstract

In reverberant environments there are long term interactions between speech and corrupting sources. In this paper a time delay neural network (TDNN) architecture, capable of learning representations, is used for reverberation robust acoustic modeling. Further, i-vectors are used as an input to the neural network to perform instantaneous speaker and environment adaptation, providing 10% relative improvement in word error rate. By sub-sampling the outputs at TDNN layers across time steps, training time is reduced. Using a parallel training algorithm we show that the TDNN can be trained on ~ 5500 hours of speech data in 3 days using up to 32 GPUs. The TDNN is shown to provide results competitive with state of the art systems in the IARPA ASPIRE challenge, with 27.7% WER on the dev test set.

Index Terms: far field speech recognition, time delay neural networks, reverberation

the input context up to 280 milliseconds. The ability to process such a wide temporal context enables the network to deal with late reverberations.

i-vectors which capture both speaker and environment specific information have been shown to be useful for rapid adaptation of the neural network [7, 8, 9]. i-vector based adaptation has also been shown to be effective in reverberant environments [10]. In this paper we use this adaptation technique.

We show experimental results on the ASPIRE far-field speech recognition challenge held by IARPA [11]. This challenge uses the English portion of the Fisher database [12] for acoustic and language model training. We show that in this large data scenario the proposed network architecture, combined with a parallel training technique [13], can train on multi-condition training data of ~ 5500 hours, using up to 32 GPUs, in 3 days.

Using the TDNN architecture helps us to achieve results close to those of the best combined system submitted to the AS-

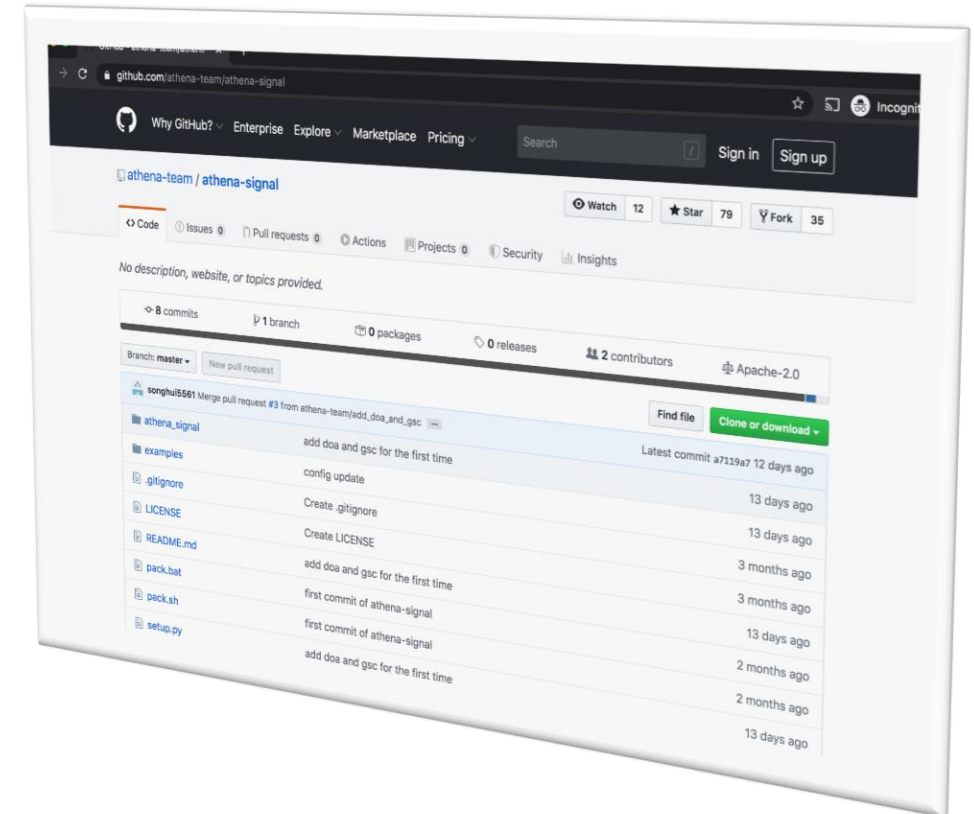
1. JHU ASPIRE system: robust LVCSR with TDNNs, ivector adaptation and RNN-LMs
2. Reverberation robust acoustic modeling using i-vectors with time delay neural networks

Far-field: signal processing

- The idea
 - Processes the input audio and improves the audio quality
- Methods
 - Acoustic echo cancellation (AEC)
 - Automatic gain control (AGC)
 - Beamforming (BF)
 - Noise suppression (NS)
 -

Far-field: signal processing

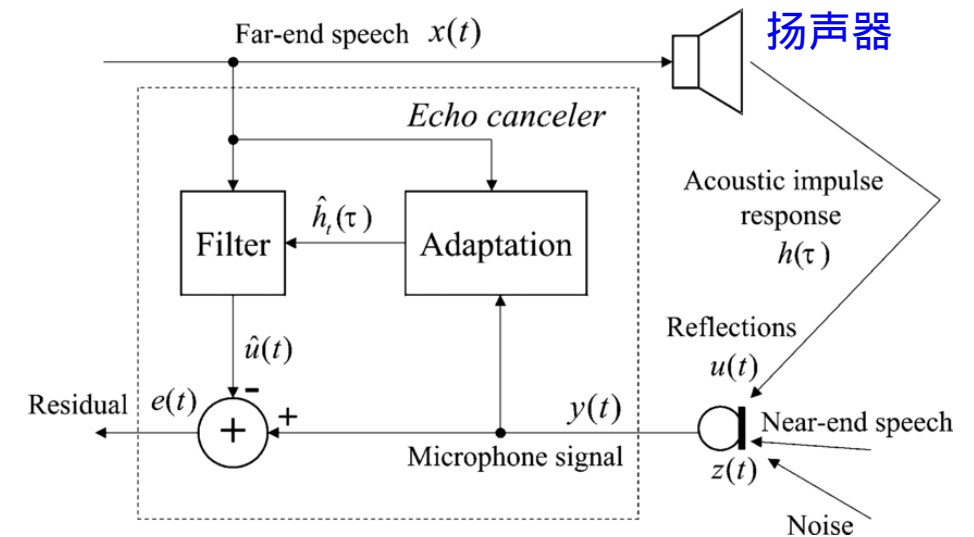
- The idea
 - Processes the input audio and improves the audio quality
- Methods
 - Acoustic echo cancellation (AEC)
 - Automatic gain control (AGC)
 - Beamforming (BF)
 - Noise suppression (NS)
 -



Athena signal: <https://github.com/athena-team/athena-signal>

Far-field: signal processing

- **Acoustic echo cancellation**
 - Cancels acoustic feedback between a speaker and a microphone
 - Widely used in telecommunication
 - Necessary in smart speakers
- Technical details
 - Time delay estimation
 - Double-talk detection
 - Residual echo suppression
 -

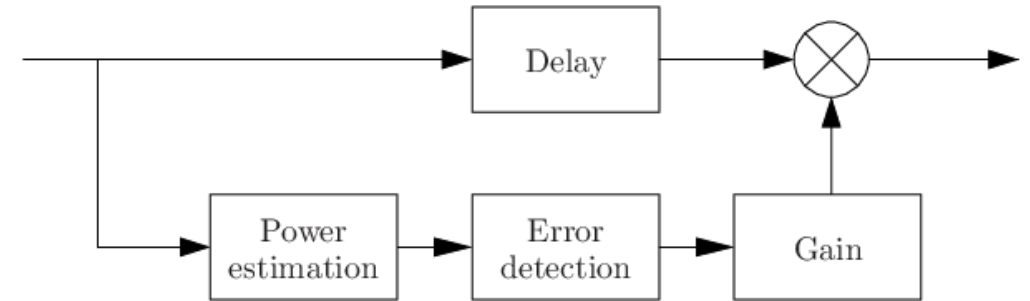


A typical AEC framework.

Photo credit: <https://www.researchgate.net/>

Far-field: signal processing

- Automatic gain control
 - Maintains a suitable signal amplitude at its output
 - Radio/Radar/Telephone
- Technical details
 - Gain factor calculation
 - Affects both signal and noise 不会改变信噪比

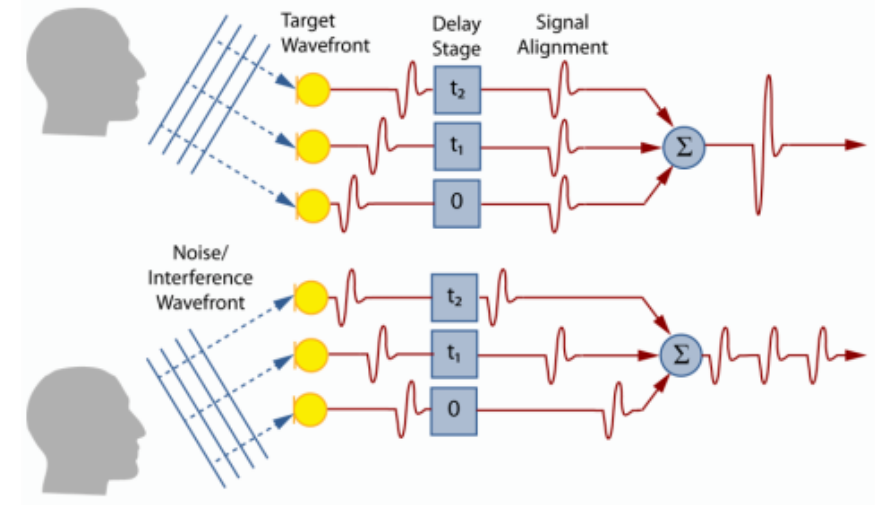


A typical AGC framework.

Photo credit: <https://www.researchgate.net/>

Far-field: signal processing

- **Beamforming**
 - Boosts SNR for a certain direction or source
 - Radio/Radar/Sonar.....
- Technical details
 - A lot of different algorithms
 - Direction of arrival (DOA) estimation
 -

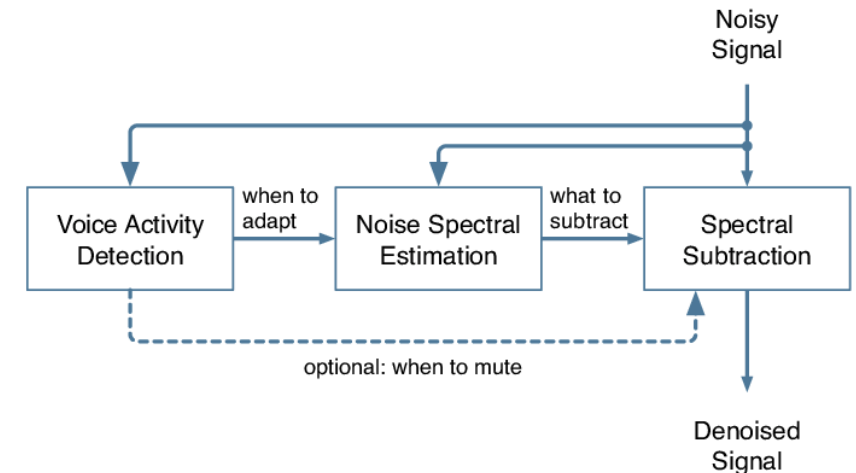


A typical delay-and-sum framework for beamforming.

Photo credit: <https://beamforming-noise-cancellation.weebly.com/beamforming.html>

Far-field: signal processing

- Noise suppression
 - Boosts speech SNR
- Technical details
 - A lot of different algorithms
 - Voice activities detection
 - Introduces distortion (bad impact on model) 适配
 -



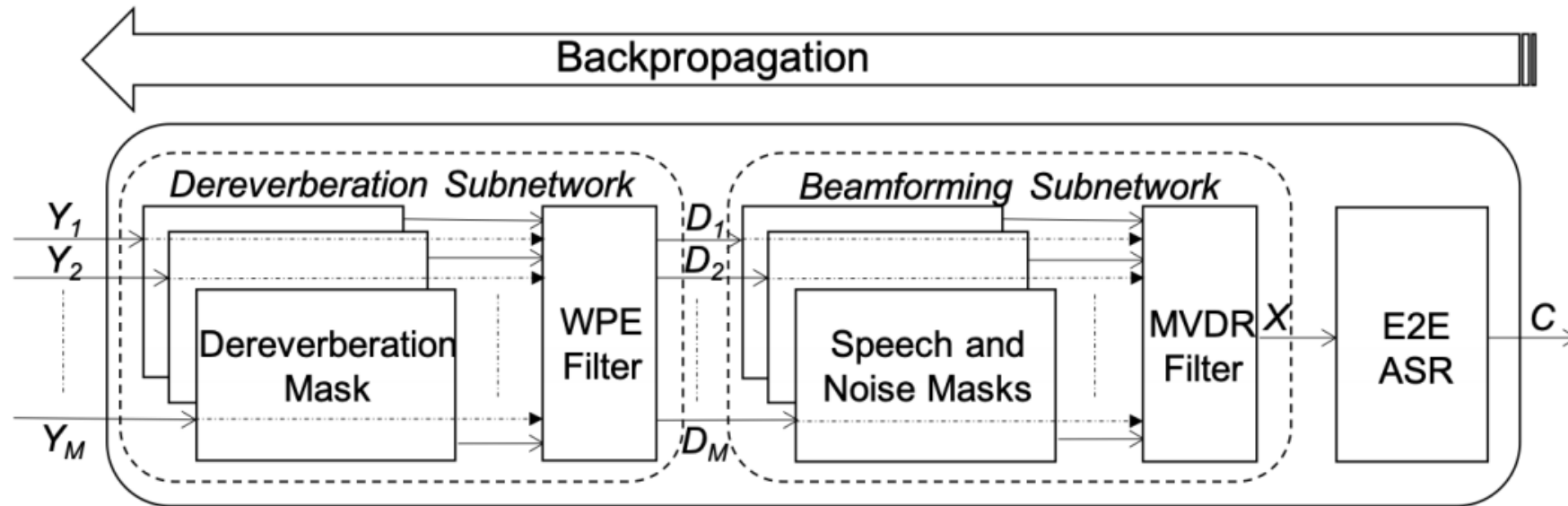
A typical noise suppression framework.

Photo credit: <https://www.researchgate.net/>

- Combine signal processing and modeling
 - Directly connecting signal processing and the model usually leads to worse performance
 - Fine-tune the model parameters with the signal processing frontend 微调
 - Collect new data with the signal processing frontend
 - Play and record existing data through the signal processing frontend

- The idea
 - We have to fine-tune the ASR model with the signal processing frontend
 - Why don't we train them jointly?

Far-field: end-to-end



End-to-end multichannel ASR architecture

Photo credit: [An Investigation of End-to-End Multichannel Speech Recognition for Reverberant and Mismatch Conditions](#)

Far-field: end-to-end

Method	Dereverberation	Beamformer			REVERB Simulated						REVERB Real		DIRHA LA
		Method	Reference	Mask	Room 1		Room 2		Room 3		Room 1		
				Type	Near	Far	Near	Far	Near	Far	Near	Far	
Challenge baseline [6]	-	-	-	-	16.2	18.7	20.5	32.5	24.8	38.9	50.1	47.6	-
E2E Baseline	-	-	-	-	5.4	7.1	7.6	12.9	9.7	16.1	23.9	26.8	55.3
Pipeline	WPE	-	-	-	6.0	6.6	7.1	9.8	8.0	11.2	17.7	18.4	42.3
	DNN-WPE	-	-	-	5.7	6.0	7.5	9.3	7.8	10.1	16.4	18.5	41.3
	-	BeamformIt	X-Corr	-	5.8	6.1	5.8	8.5	6.9	10.2	14.6	16.1	39.2
	WPE	BeamformIt	X-Corr	-	6.6	5.9	6.1	7.0	6.8	8.2	11.3	11.9	30.7
	DNN-WPE	BeamformIt	X-Corr	-	6.3	5.8	6.4	6.8	6.6	7.7	11.0	10.8*	31.3
E2E	WPE	-	-	-	6.3	6.7	6.7	8.9	7.4	10.6	17.0	19.8	42.3
	-	MVDR	Ch 2	TF	5.7	6.1	5.6	8.2	6.2	10.2	12.6	17.3	42.3
	-	MVDR	Ch 2	SAD	7.2	7.2	6.4	8.6	7.1	12.1	16.0	20.5	45.3
	WPE	MVDR	Ch 2	TF	5.5	5.7*	5.3*	6.6*	6.5	7.6*	10.7	13.7	35.4
	WPE	MVDR	Ch 2	SAD	8.3	7.8	6.9	7.0	7.6	8.6	10.8	13.9	31.6
	WPE	MVDR	Attention	SAD	6.4	6.3	5.9	6.8	6.3*	7.6*	8.7*	12.4	29.1*
Tachioka et. al. [31]	Spectral subtraction	Delay-sum	-	-	5.0	5.6	5.6	8.2	5.7	10.5	16.9	20.3	-
Alam et. al. [32]	Iterative deconvolution	-	-	-	6.7	7.3	8.0	11.1	8.1	12.1	21.4	22.0	-
Wang et. al. [10]	-	-	-	-	-	-	-	-	-	-	-	-	35.1

beamformIt

WER (%) on REVERB and DIRHA-WSJ (LA array) evaluation sets. Pipeline represents the traditional system where you have separate components for signal processing frontend and model, while E2E represents the end-to-end system.

Photo credit: [An Investigation of End-to-End Multichannel Speech Recognition for Reverberant and Mismatch Conditions](#)

- **Remarks**
 - Neural frontend, and end-to-end approaches are giving very promising results
 - Traditional signal processing frontend is still heavily used in industry production systems
 - Trend of end-to-end solutions in some industry production systems

Outline

- Far-field speech recognition
- **Mix-lingual speech recognition**
- “Low resource” speech recognition
- Homework

Mix-lingual: definition

- **Multi-lingual**
 - Combines data from multiple languages to improve ASR performance
- **Cross-lingual**
 - Uses data from one or multiple languages to create ASR system for a different language
- **Mix-lingual**
 - Mixed language for a single ASR system, e.g., “我经常阅读paper”

Mix-lingual: a practical issue

- Code-mixing v.s. code-switching
 - Code-mixing: mixing the lexicons of two or more languages together
 - Code-switching: completely switching from one language's lexicon to another
- Code-mixing
 - 我想听 yesterday once more
 - 我朋友在 Google 工作
 - 我收到的验证码是 ABC123
- Important for Chinese ASR in production systems

Mix-lingual: a typical ZH/EN mix system

- **Mixed phone set**
 - English phones
 - Vowels, English-specific consonants
 - Chinese phones
 - Tonal vowels, Chinese-specific consonants
 - Shared phones
 - Shared consonants
 - Non-speech phones
 - SIL, SPN, NSN, etc.
- **Valuable resource**
 - CMU dictionary

Mix-lingual: a typical ZH/EN mix system

- **Mixed lexicon**
 - Pronunciation for English words
 - CMU dictionary
 - Spelled words (APP, API, etc)
 - Pronunciation generation (G2P) [word->phone](#)
 - Pronunciation for Chinese words
 - Existing dictionary (e.g., Pinyin based)
 - Words/characters lookup
- **Valuable resource**
 - Phonetisaurus: a grapheme-to-phoneme (G2P) toolkit

Mix-lingual: ASRU Code-mix Eval

- Organized by Datatang
- Mandarin Chinese and English, 16kHz, cellphone speech
- Training data
 - 500h Mandarin Chinese speech
 - 200h code-mix speech
 - 960h Librispeech 960h
- Official trigram LM (1M vocab)
- Two dev sets, each 20h
- One test set, 20h

Mix-lingual: ASRU Code-mix Eval

- Track1: (500h+200h+960h) AM training, fixed official LM, traditional speech recognition
- Track2: (500h+200h+960h) AM training, additional LM training data allowed, traditional speech recognition
- Track3: (500h+200h+960h) AM training, fixed official LM, end-to-end speech recognition
- MER (mixed error rate) = CER for Chinese, WER for English

Mix-lingual: ASRU Code-mix Eval

MER	Track1 (Fixed LM)	Track2 (Additional LM)	E2E (Fixed LM)
Team1	4.94%	4.72%	N/A
Team2	5.05%	5.64%	10.96%
Team3	5.28%	N/A	N/A
Team4	5.74%	5.80%	N/A
Team5	N/A	N/A	5.91%
Team6	N/A	N/A	8.82%
Team7	6.61%	5.88%	9.00%
Team8	N/A	N/A	9.37%

ASRU 2019 code-mix evaluation results

Outline

- Far-field speech recognition
- Mix-lingual speech recognition
- “Low resource” speech recognition
- Homework

Low-resource: definition

- Low computation resource?
 - Training?
 - Decoding?
- Low data resource?
 - Hours of data?
 - Tens of hours of data?
 - Hundreds of hours of data?

Low-resource: a practical view

- A new language or a new domain
- Manually transcribed data: < 500 hours
- How can we improve our system performance?

Low-resource: data augmentation

- Noise augmentation
- Volume perturbation
- Speed perturbation
- Frequency masking
- Time masking
-

Low-resource: data augmentation

System	Fold	Epochs	LM	SWB	CHE	Total
Baseline	1	6	fg	13.7	27.7	20.7
VTLP	3	2	fg	13.1	26.5	19.9
VTLP	5	2	fg	13.2	26.7	20.0
VTLP + time-warp	3	2	fg	13.3	26.8	20.1
Tempo-perturbed	3	2	fg	13.5	27.0	20.3
Speed-perturbed	3	2	fg	13.1	26.1	19.7
Speed-perturbed	3	6	fg	12.9	25.7	19.3

Results (% WER) for the baseline and speed-perturbed DNN systems on the subsets of the Hub5 00 evaluation set

Photo credit: [Audio Augmentation for Speech Recognition](#)

Low-resource: transfer learning

- Multi-lingual
- Domain adaptation

Low-resource: unsupervised learning

- Traditional unsupervised learning
- Masked Predictive Coding (MPC) based unsupervised learning

用模型对下载到的数据做预测, 模型集成做数据清洗

Outline

- Far-field speech recognition
- Mix-lingual speech recognition
- “Low resource” speech recognition
- Homework

Thanks!

