A Machine Learning Framework for Enhancement and Recognition of Microphone Array Speech

Chin-Hui Lee
School of ECE, Georgia Tech
chl@ece.gatech.edu

In collaboration with GT and USTC teams

Outline and Talk Agenda

- DSP based on learning nonlinear spectral regression
  - Paradigm shift: spectral mapping with deep learning & big data
- Three classical single-channel DSP problems (Part 1)
  - DNN-based speech enhancement (SE)
  - DNN-based source or speech separation (SS)
  - DNN-based speech dereverberation
- Extension to far-field microphone array speech (Part 2)
  - Two-stage architecture for SE/SS and robust speech recognition
  - Multiple sources of interferences in reverberant conditions
  - Speaker-dependent enhancement (only five-minute training)
  - Black-box LVCSR (already clean- or multi-condition trained)
  - Comparing multi-channel DNN architectures & performances
- Summary, supplements, references and recent efforts

Speech in Noisy Environment

1. Additive noise (mathematical mixing):

$$\left.\begin{array}{l} y(t) = x(t) + n(t) \xrightarrow{\text{STFT}} Y(l,k) = X(l,k) + N(l,k) \end{array}\right\}\text{Focused in first part}$$

2. Convolutional noise:

$$y(t) = x(t) * h(t)$$

Often solved with mixing signal assumptions by mathematical optimization in conventional approaches!!

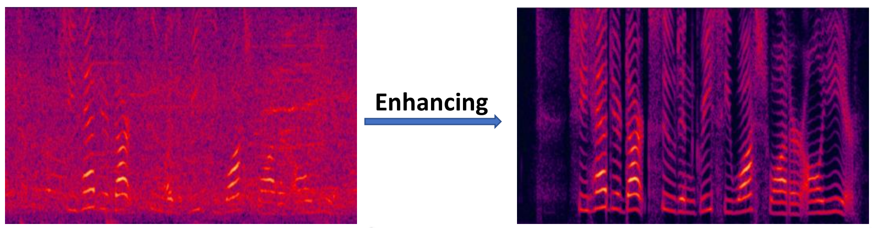
3. Mixed noise:

$$\left.\begin{array}{l} y(t) = x(t) * h(t) + n(t) \\ y(t) = [x(t) + v(t)] * h(t) + n(t) \end{array}\right\}\text{Focused in second part}$$

Part 1:
Single-Channel Speech Enhancement, Separation and Dereverberation
-- Speactral Mapping with DNN Regression

Topic 1: Speech Enhancement (SE)

- Speech enhancement: improving the intelligibility and/or overall perceptual quality of degraded speech signals using digital signal processing (DSP) techniques
- One of the most addressed classical SP problems
  - ➢ Issues: musical noise and non-stationary backgrounds



Noisy speech,
Exhibition, SNR=5dB

Clean speech

Conventional Speech Enhancement

- **Classified by the number of signal channels**
  1. Single channel speech enhancement
     - ❑ Time and frequency information

     Focused in Part 1

  2. Array based speech enhancement
     - ❑ Time, frequency and spatial information

     Focused in Part 2

- **Conventional Techniques: math and physics**
  - ➢ Spectral subtraction, Wiener filtering, masking
  - ➢ MMSE log spectral amplitude (MMSE-LSA)
  - ➢ Optimally modified LSA (OM-LSA)
  - ➢ Many others for single- and multi-channels SE

$$\boxed{\hat{X}(l,k) = Y(l,k) - \hat{N}(l,k)}$$

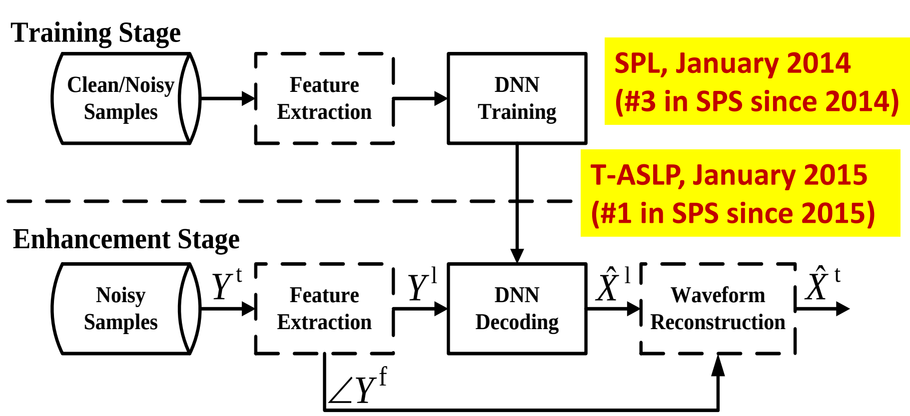
Learning-Based Speech Enhancement

- Early: HMM-based speech estimation (Erphraim & Malah, 1984)
- Deep denoising autoencoder (Lu, Tsao, Matsuda, Hori, 2013)
- Classification-based separation (Wang & Wang, 2013)
- Nonlinear regression function *F(.) for spectral mapping in 2012 – most previous DNN efforts were for classification-based learning*

$$X(l,k) = F\big(Y(l,k)\big) + E(l,k)$$

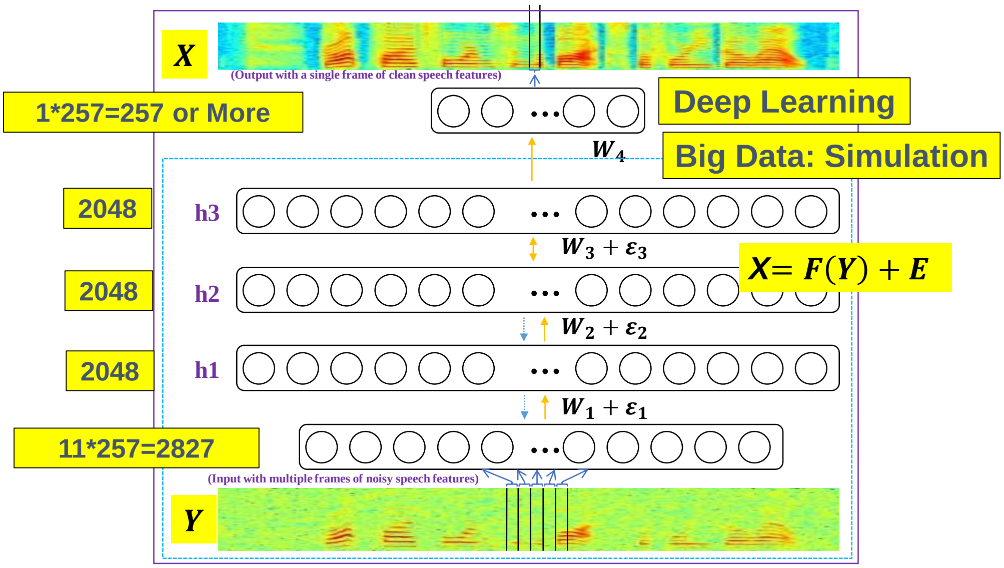
- What is *F(.)*? *What parameters? How many?*
- How to obtain a lot of the training pairs, *(Y, X)?*
- Any special assumptions? Generalization issues?
- How to estimate the parameters?
- How to handle mismatched spectral magnitude & phase

DNN-Based SE System Overview



1. Feature extraction: log-power spectra (LPS)
2. Waveform reconstruction: overlap-add (OLA) algorithm
3. Training: RBM pre-training + back-propagation fine-tuning
4. Phase (later)

DNN Based Spectral Mapping: A Paradigm Shift



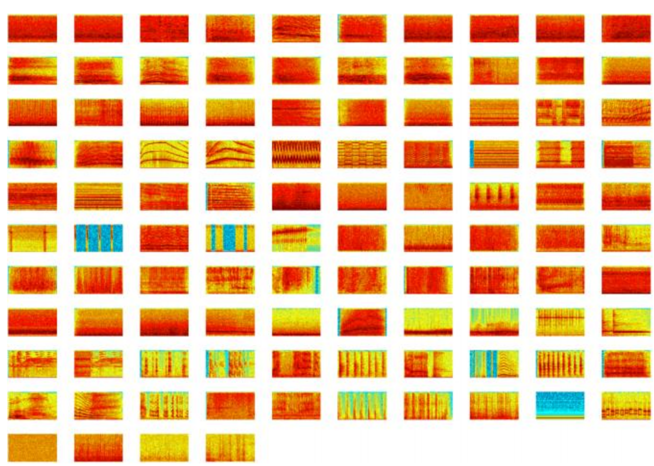
High-dim Vector-to-vector nonlinear regression: 20 million parameters

Noise-Universal SE-DNN – A Hope

- **DNN to learn the characteristics of many noise types**

➢ Classifications: Crowding、machine、transportation、animal、nature、human, etc.

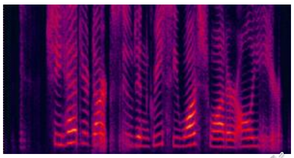
alarm cry



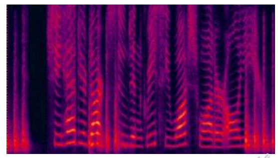
G. Hu, 100 non-speech environmental sounds, 2004. http://www.cse.ohiostate.edu/pnl/corpus/HuCorpus.html.

Enhanced Results: Non-stationary Noise

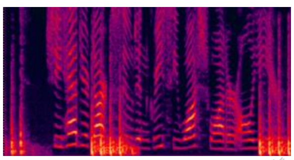
- An utterance with machine gun noise at SNR= -5dB: with 104-noise enhanced (upper left, PESQ=2.78), MMSE enhanced (lower left, PESQ=1.86), 4-noise enhanced (upper right, PESQ=2.14), and noisy speech (lower right, PESQ=1.85):



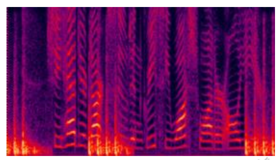
104NT-DNN enhanced **PESQ=2.78**



4NT-DNN enhanced PESQ=2.14



Log-MMSE enhanced PESQ=1.86



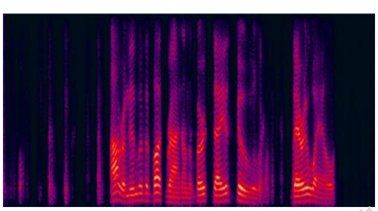
noisy, machine gun, SNR=-5dB PESQ=1.85

Even the 4NT-DNN is much than Log-MMSE, SE-DNN is capable of suppress highly non-stationary noise. Why?

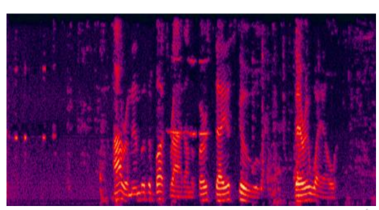
More enhancement examples can be found at: home.ustc.edu.cn/~xuyong62/demo/SE_DNN.html

Enhanced Results: Real-World Speech

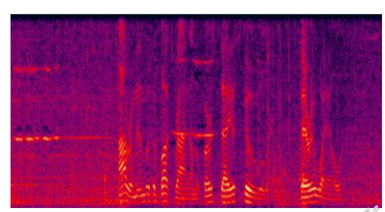
- Spectrograms of an utterance extracted from the movie ***Forrest Gump***: DNN (left), Log-MMSE (right), and noisy (middle) with unseen noise



Universal SE-DNN enhanced



Log-MMSE enhanced



Noisy

- Good generalization capacity to real-world noisy speech
- Publicly available tool packages
  - GPU C++ version: https://github.com/yongxuUSTC/DNN-for-speech-enhancement
  - Python version: https://github.com/yongxuUSTC/sednn

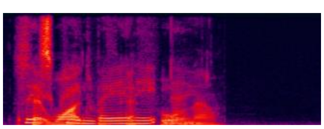
Topic 2: Source & Speech Separation (SS)

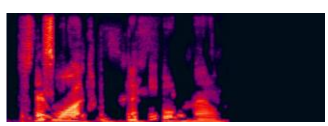
- Source separation aims at separating a target speaker' speech from mixed speech with interfering speakers (one dominating speaker) to improve the intelligibility and overall perceptual quality of separated speech and possibly for ASR/SID/LID using acoustic signal processing techniques
- Ideal DNN-based speech enhancement / source separation
  - More than one-hour training speech from the target speaker

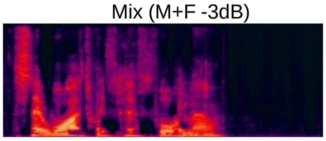
Mixed speech SIR=0dB → Separation → Separated Target

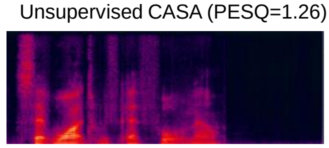
- Log-MMSE based speech enhancement

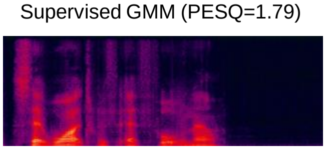
Comparisons: M-F Mixture (5-min Target, Non-ideal)

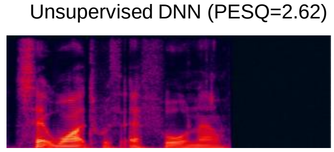

Mix (M+F -3dB)


Unsupervised CASA (PESQ=1.26)


Supervised GMM (PESQ=1.79)


Unsupervised DNN (PESQ=2.62)


Semi-supervised DNN (PESQ=2.68)


Target(M)

Single-Channel SS
- Supervised: both speakers known
- Semi-supervised: only target known
- Unsupervised: both speaker unknown
- Multi-channel BSS: not compared

T-ASLP, August 2016
T-ASLP, July 2017

- Semi-supervised DNN: better than supervised GMM
- Unsupervised DNN: better than state-of-the-art CASA

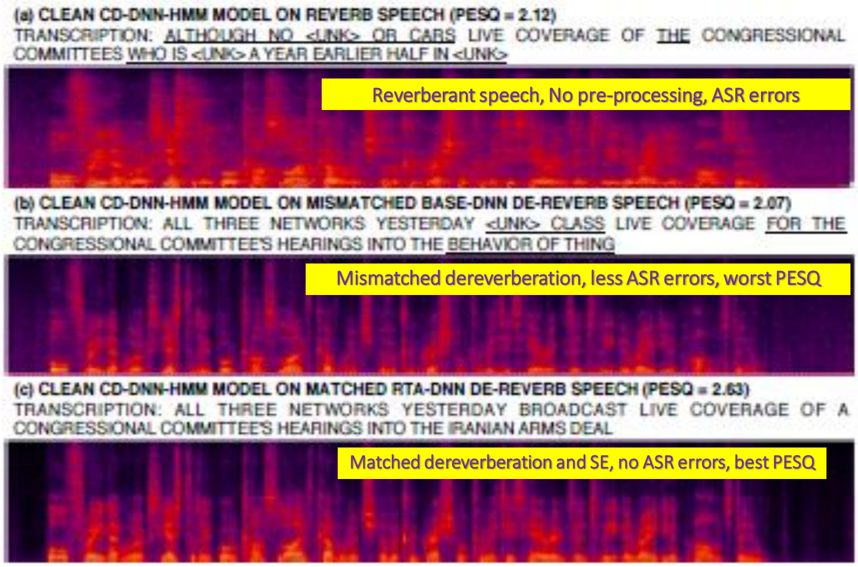
Topic 3: Speech Dereverberation

Reverb Enhanced

Issues:
- A lot
- RIR
- RT60

T-ASLP, Jan 2017
J-STLP, Dec 2017

Good DSP will Lead to Accurate ASR



Summary: So Far

1. Spectral mapping with deep learning & big data: a paradigm shift
2. Large training set: learning rich regression structure
   - Even simulation data can be very useful if properly generated
3. For DNN-based speech enhancement, separation, and dereverberation the results are amazingly good so far
   - Multiple sources of interferences: next target
   - Array-based enhancement (DNN too big?) and ASR

   Focused in Part 2
4. 50+ papers, SPL: #3, T-SALP: #1 top cited paper in IEEE SPS
5. Need to combine with conventional techniques, e.g., IRM, IME
6. A New Hope: with proper pre-processing followed by integrated post-processing ➡ leading to robust ASR! But for black-box ASR?
   - Lowest errors in CHiME-2, -4, & REVERB Challenges

Part 2:
Single- and Multi-Channel Master-Voice Separation and Recognition of Array Speech -- Preliminaries with Two-Stage Enhancement

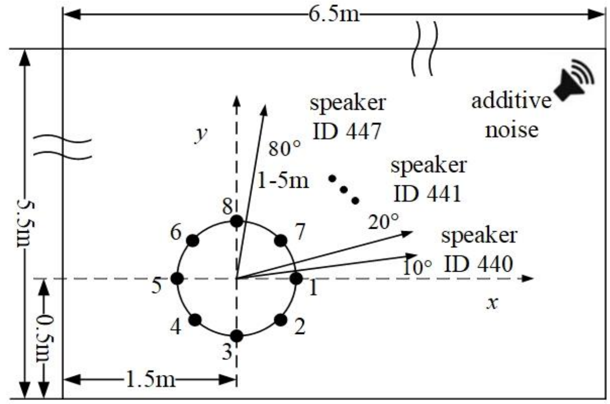
(**Balancing Temporal and Spatial Information**)

Issues: from Single-Channel to Multi-Channel Speech

- Unknown or imprecise array configurations or microphone types
- Robustness to room types, target, interference and array positions
- Room-specific or general room conditions: RIR and RT_{60}
- Whither temporal or spatial information: input vectors could be too long?
- Backend ASR is a black box, often multi-condition trained. What to optimize?
- This talk: living-room or in-vehicle application scenarios
  - Condition-specific but with more complex training data generation
  - Multiple sources of interferences and additive noises with room reverberation
  - Speaker-independent (SI) pre-processing could not deliver satisfactory performances
  - SD master voice pre-processing performs better but how much training data?
  - In this talk: less than 5-minute master training voice, and mostly clean-trained ASR

Acoustic Environment for Array Speech Simulation

- Room Size: 6m*5.5m*3m
- Clean speech: WSJ
- Noise: OSU-100, NOISEX
- RIR: ISM, RT60: 0.2-0.3 sec
- Training SD: 30 hours (from 40 to 20K utterances)
- Testing SD: 1800 utterances: unseen speakers and noises
- Talking distance: 1-5 meters
- SINR: 5, 10 and 15 dB at the receiving microphone
- Reference: Microphone #1



- ➢ SINR > 5dB: or WER may exceed 100%
- ➢ DoA assumed known by wake-up or cameras

An example

Single-Channel Speaker-Dependent Speech Separation

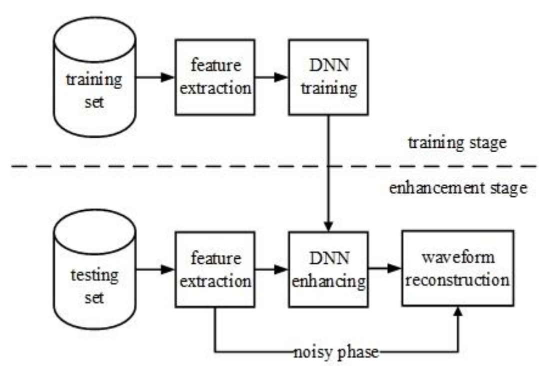
Baseline DNN configuration ('single')

- 3 hidden layer: 2048 units each
- Input: 257-dim LPS features
- Temporal context size: 11-frame input
- The dual outputs: estimated 257-dim LPS and 257-dim IRM (ideal ratio mask or IRM for post-processing) features
- Multi-target learning

$$E_r = \frac{1}{N}\sum_{n=1}^{N}\left\|\hat{X}_n(\bar{Y}_{n\pm\tau}, W, b) - \bar{X}_n\right\|_2^2 + \beta * \frac{1}{N}\sum_{n=1}^{N}\left\|IRM_n(\bar{Y}_{n\pm\tau}, W, b) - IRM_n\right\|_2^2.$$

$$\alpha = 0.\beta = 0.05, \quad \delta = 0.7, \gamma = 1.25$$

$$IRM_n = \sqrt{\frac{e^{X_n}}{e^{X_n} + e^{N_n}}}$$



$$\hat{X}'_n(d) = \begin{cases} Y_n(d), & IRM_n(d) \geq \gamma \\ (Y_n(d) + \hat{X}_n(d))/2, & \delta \leq IRM_n(d) < \gamma \\ \hat{X}_n(d), & IRM < \delta \end{cases}$$

Overall Single-Channel ASR Result Summary

- Task: 230K-word WSJ speaker-independent recognition with trigram LM perplex of 141
- AM: CD-DNN-HMM trained with 70 hours of WSJ clean data, 6 hidden layers with 2048 units each, input is 40-dim FMLLR, 11-frame expansion, output is 3455 shared states
- Test: Nov92, 8 speakers, 4 males and 4 females, 300 utterances, clean WER: ~3%
- SD Separator training: with 8-speaker clean adaptation data, 40 utterances each

WER (in %) and WERR (in parentheses in %) at 1m range

	SINR = 5dB	SINR = 10dB	SINR = 15dB
Noisy reverberant	73.95	41.88	20.08
DNN processed	21.95 (70.32)	14.98 (64.23)	11.89 (40.79)

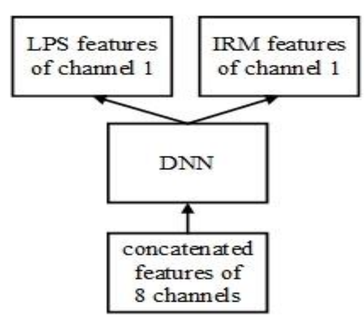
WER (in %) and WERR (in parentheses in %) at 3m range

	SINR = 5dB	SINR = 10dB	SINR = 15dB
Noisy reverberant	79.28	50.10	24.46
DNN processed	24.76 (68.77)	18.31 (63.45)	15.61 (36.18)

Proposed Multi-Channel DNN-Based Speech Enhancement – DNN Architecture 1 (DNN1)

DNN architecture '1×8×1'

- Enhancement with 1 DNN
  - Training: randomly selected 90 out of 240 hours
  - 3 hidden layers, 2048 units in each
  - 8-frame Input: 8-channel LPS concatenated
  - Temporal context size in each channel: 1
  - Dual outputs: 257-dim LPS + 257-dim IRM of channel 1
- $\alpha = 0, \beta = 0.05, \delta = 0.7, \gamma = 1.25$



IRM-based post-processing ('pp')

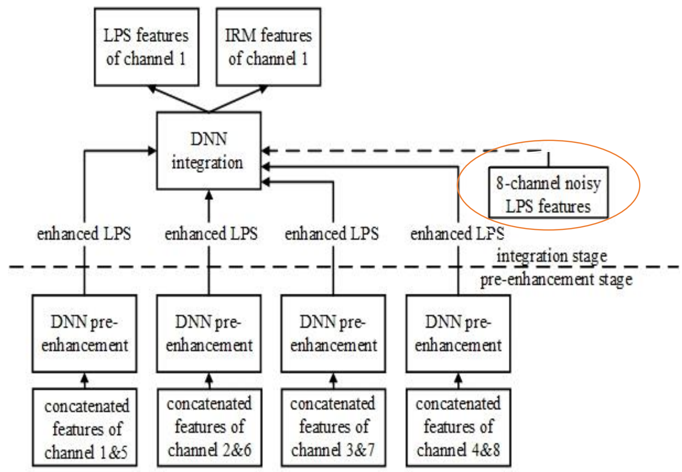
$$E_r = \frac{1}{N}\sum_{n=1}^{N}\left\|\hat{X}_n(\bar{Y}_{n\pm\tau}, W, b) - \bar{X}_n\right\|_2^2 + \beta * \frac{1}{N}\sum_{n=1}^{N}\left\|IRM_n(\bar{Y}_{n\pm\tau}, W, b) - IRM_n\right\|_2^2.$$

$$\hat{X}_n'(d) = \begin{cases} Y_n(d), & IRM_n(d) \ge \gamma \\ (Y_n(d) + \hat{X}_n(d))/2, & \delta \le IRM_n(d) < \gamma \\ \hat{X}_n(d), & IRM < \delta \end{cases}$$

Proposed Two-stage Multi-Channel DNN-Based Speech Enhancement – DNN Architecture 2 (DNN2)

DNN architecture ‘4×2×5+int08’

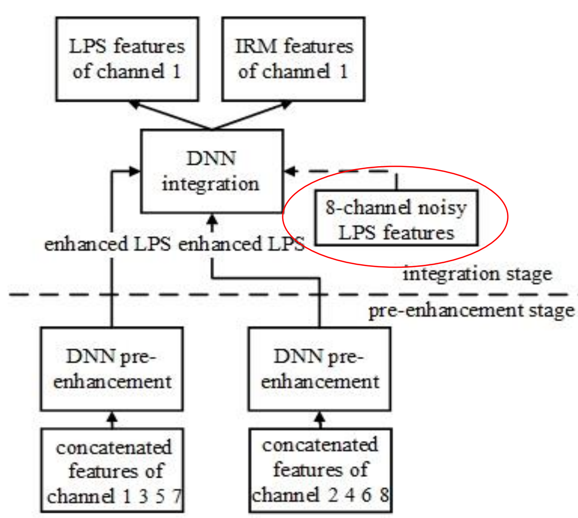
- Stage 1: pre-enhancement with 4 DNNs
  - ‘pre2×5-1-5’ ‘pre2×5-2-6’ ‘pre2×5-3-7’ and ‘pre2×5-4-8’
  - 2 hidden layers, 2048 units in each
  - Reference channel: channel 1
  - 10-frame Input : 2-channel LPS concatenated
  - Temporal context size in each channel: 5
  - Outputs: enhanced 257-dim LPS of channel 1
- Stage 2: integration with 1 DNN
  - 3 hidden layers, 2048 units in each
  - Input with 4 or 12 frames: 4 enhanced LPS
  - concatenated with 8-channel noisy LPS
  - Dual outputs: 257-dim LPS + 257-dim IRM



Proposed Two-stage Multi-Channel DNN-Based Speech Enhancement – DNN Architecture 3 (DNN3)

DNN architecture ‘2×4×3+int0/8’

- Stage 1: pre-enhancement with 2 DNNs
  - ‘pre4×3-1-3-5-7’ and ‘pre4×3-2-4-6-8’
  - 2 hidden layers, 2048 units in each
  - 12-frame Input : 4-channel LPS concatenated
  - Temporal context size in each channel: 3
  - Outputs: enhanced 257-dim LPS of channel 1
- Stage 2: integration with 1 DNN
  - 3 hidden layer with 2048 units in each
  - Input with 2 or 10 frames: 2 enhanced LPS concatenated with 8-channel noisy LPS
  - Dual outputs: 257-dim LPS + 257-dim IRM



Preliminary Results for Architecture Comparisons

WER and PESQ results at 3m for Speaker 447

	SINR/dB RT60/s	5 0.2-0.3	10 0.2-0.3	15 0.2-0.3	Average WER (%)	Average PESQ
	noisy ch1	89.60	64.63	28.47	60.90	2.16
Baseline	single	28.60	20.32	16.42	21.78	2.40
DNN1	1×8×1	9.89	7.26	5.98	7.71	2.80
DNN2	pre2×5-3-7	14.71	10.20	8.42	11.03	2.74
DNN2	4×2×5+int0	11.34	8.33	6.98	8.88	2.90
DNN2	4×2×5+int8	10.66	8.07	6.58	8.44	2.90
DNN3	pre4×3-1-3-5-7	11.09	7.96	6.94	8.66	2.82
DNN3	2×4×3+int0	9.65	7.25	5.93	7.61	2.92
DNN3	2×4×3+int8	9.19	7.01	6.01	7.40	2.93
Post-processing	2×4×3+int8+pp	9.17	6.75	5.85	**7.25**	**2.94**
	anechoic				3.16	

LISTEN Workshop, 07/17/18 25

Overall PESQ and ASR Result Summary

- Task: 230K-word WSJ continuous speech recognition with trigram LM perplex of 141
- AM: CD-DNN-HMM trained with 70 hours of **WSJ1 clean data**, 6 hidden layers with 2048 units each, input is 40-dim FMLLR transformed MFCC, 11-frame expansion, output is 3455 senones
- Test: Nov92, 8 speakers, 4 males and 4 females, about 1800 utterances for each speaker
- Separator training: with 8-speaker clean adaptation data, about five minutes each

Average PESQ, WER and WERR (in parentheses in %) over all 8 speakers for systems at SINR 5-15dB, RT60 0.2-0.3s, and 1-5m

1-5m	Average PESQ	Average WER % (WERR)
noisy ch1	2.15	48.47
baseline: single	2.43	17.89 (63.10%)
proposed: 2×4×3+int8	2.92	6.78 (62.04%)
proposed: 2×4×3+int8+pp	2.95	6.56 (3.24%)
Anechoic	---	3.15

Discussion on Clean- vs. Multi-Condition Training

- Same single-channel and multi-channel speech pre-processing
- AM: 41-dim fbank features, 11 frames expansion (per-utt cmvn), 4 hidden layers, 1024 hidden nodes

WER(%) Spk 447 3m	clean	noisy	8-channel enhanced
MC RIR interfere fbank	3.43	11.79	6.31
MC RIR fbank	3.69	36.31	6.56
MC fbank	3.16	41.93	6.95
Clean fbank	2.9	66.96	7.94

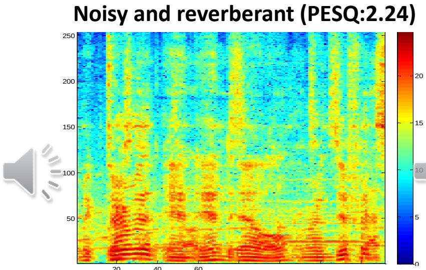
+18.27% -46.48% -20.53%

- Clean ASR: 70 hours clean data
- MC fbank ASR: add 90 kinds of noise, SNR=0dB 5dB 10dB, 280 hours noisy data
- MC RIR fbank ASR: convolve clean data with the RIR of 80 degree(RT60=0.2&0.3s), add 90 kinds of noise, SNR= 0, 5, 10 dB; 280 hours (contain 70 hours anechoic data)
- MC RIR interfere ASR: convolve clean data with RIR (RT_{60}=0.2&0.3s), add 74 interferers, add 90 kinds of noise, SINR= 0, 5, 10 dB, still 280 h (contain 70 hours anechoic data)

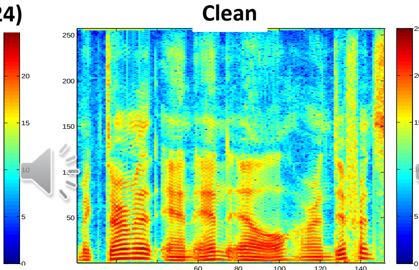
Enhanced speech works better for multi-condition model. Better enhancement put clean model on top?

An Example Test Utterance *– 3m 447 SINR=10dB RT60=0.2s*

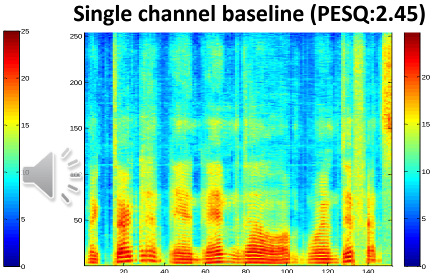
Noisy and reverberant (PESQ:2.24)



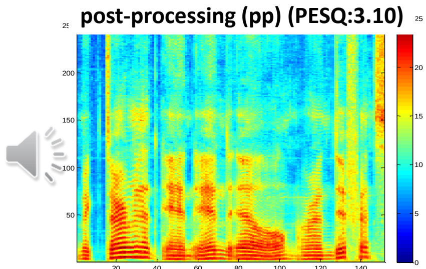
Clean



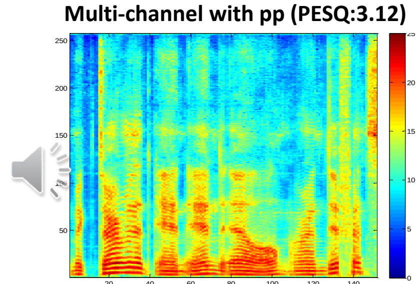
Single channel baseline (PESQ:2.45)



Multi-channel without post-processing (pp) (PESQ:3.10)



Multi-channel with pp (PESQ:3.12)



Summary: Two-Stage DNN Architecture (for SD Separation and SI Recognition)

- Achieve significant PESQ improvement and WER reduction for multi-channel DNN, also effective in handling single-channel speech (5-min SD training)
- Assume little on array configurations; not sensitive to array geometry
- Propose a new two-stage enhancement strategy: pre-enhancement and integration combining both temporal and spatial information in spectra
- Need to replace known with estimated power for real-world applications
    - Power equalization caused about 10% degradation: how to reduce it?
- Explore other techniques, e.g., SE for black-box clean- or multi-condition ASR?
- Research further into DNN architectures for array-based processing
- Investigate robustness issues in varying rooms, positions and array conditions

Acknowledgment



Thank You

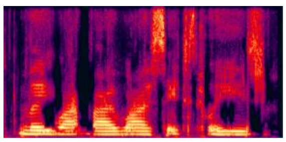
Part 3:
Supplementary Slides
Simulation, Recent Efforts and References

Single-Channel SS: Definitions

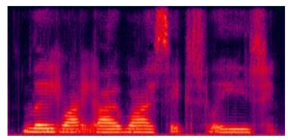
- Assuming mixed speech from one target speaker and one interfering speaker (more speakers later)
- Supervised separation: both speakers known
  - GMM-based joint and conditional distributions
- Semi-supervised separation: only target known
  - Most reasonable scenario (more speakers later)
  - This talk: DNN-based
- Unsupervised separation: both speakers unknown
  - Computational auditory scene analysis (CASA)
  - This talk: DNN-based with gender mixture detection
- Blind source separation (BSS)
  - Usually for multi-channel source separation

Source Separation: Real vs. Ideal

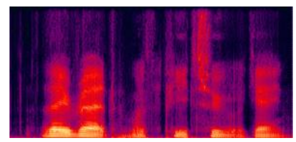
- **A few minutes versus hours of target speech for training: Source Separation Challenge (SSC)**



target of GMM output



target of DNN-2 output



Interference of DNN-2 output

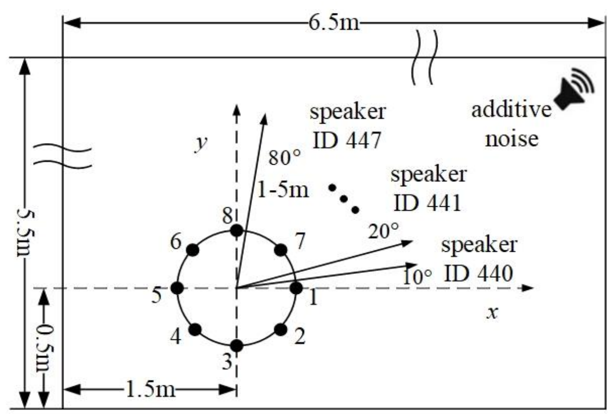
- **Target: "lay white by Y 6 please"**
- **Interference : "lay red with P 2 again"**

Acoustic Environment for Array Speech Simulation

Room Size: 6m*5.5m*3m

Speaker ID (gender)	Target Direction(°)	Interference Direction(°)
440 (M)	10	50
441 (F)	20	60
442 (F)	30	70
443 (M)	40	80
444 (F)	50	10
445 (F)	60	20
446 (M)	70	30
447 (M)	80	40

An example



- For single-channel, the single mic is Ref 1
- Horizontal range to the center is 1m, 3 m, 5m
- DoA assumed known by wake-up or cameras

Data Simulation – Single Channel (1/2)

- Data Source: WSJ Corpus; OSU 100-type noise set
- Training Data Generation (for speaker-dependent master voice separation)
  - Target: **4** males, **4** females; ID number 440-447 from WSJ0 corpus; 40 clean utterances for each
  - Interfering speakers: **72** speakers from WSJ0 corpus
  - Noise: random **90** types from OSU 100-type noise Corpus
  - One simulated training utterance: **1** target utterance ⊗ target Room Impulse Response(RIR) + **1** random interfering utterance ⊗ interfering RIR + **1** random noise
  - Aligned utterance: target utterance ⊗ Direct path of RIR;
  - Room environments: range 1m, 3m, 5m; each with 2 kinds RT60s of 0.2s and 0.3s;
  - SINR configurations (received at the microphone): SINR = 5dB, SNR = 10, 15dB; SINR = 10dB, SNR = 15dB; SINR = 15dB, SNR = 20dB (when SNR was too low the ASR WER often exceeded 100%)
  - Normalization: For each training and testing target utterance, the reverberant utterance power = clean reference power = noisy reverberant power (can be relaxed later with estimation of TR_{60})
  - Training data: SINR=5dB : SINR=10dB : SINR=15dB = 1 : 1 : 1; also include 40 SINR30dB SNR 31dB utterances without reverberation; about 20000 simulated utterances making a 30-hour training set

Data Simulation– Single Channel (2/2)

- Testing Data Generation
  - Target: the same 8 target with training, about 40 unseen utterances for each from WSJ0 Corpus
  - Interfering speakers: unseen **10** speakers from WSJ0 Corpus
  - Noise: unseen **10** kinds from OSU noise Corpus
  - One simulated testing utterance: 1 target utterance ⊗ target RIR + 1 random interfering utterance ⊗ interfering RIR + 1 random noise
  - Same SINR pairs: SINR = 5dB, SNR = 10, 15dB; SINR = 10dB, SNR = 15dB; SINR = 15dB, SNR = 20dB
  - 1m 3m 5m range, both with RT60s of 0.2s and 0.3s
  - Testing data: about 1800 simulated utterances (generated Nov92 clean test utterances)
  - 300 utterances for each situation (too low SINR often caused over 100% WER)

SINR/dB	5	5	10	10	15	15
RT60/s	0.2	0.3	0.2	0.3	0.2	0.3

Data Simulation – Multi-Channel

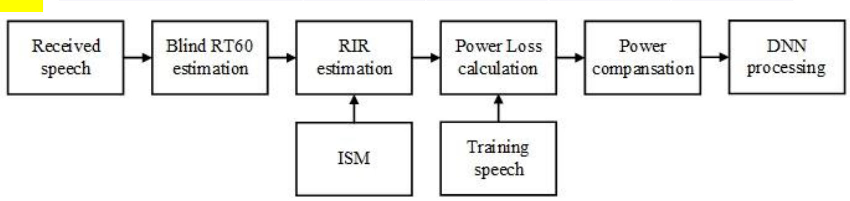
- Training Data Generation (for speaker-dependent master voice separation)
  - Using the same data generation strategy
  - 8 channels could have 8 times (240 hours) as much data as baseline (20K utterances)
  - But only 3 times (90 hours, 75000 utterances) training data of all channels were used
- Testing Data Generation
  - About 1800 simulated utterances are received by all channels
- Data Generation Assumptions
  - SINR was measured at the receiving microphone
  - With SINR at a lower level, the WER could exceed 100%
- DNN-Based Enhancement: assuming known power of desired anechoic outputs
  - Estimated power (from estimated RT_{60}) gave similar PESQ and ASR results

Discussion on Utterance Power Equalization

PESQ and WER results w/o power equalization at 1-3 m

Relaxing known power assumption using data from Speaker 447 ('pest'): giving about 10% degradation

ID 447	1m PESQ	1m WER	3m PESQ	3m WER
2×4×3+int8	3.07	6.76	2.93	7.40
2×4×3+int8+pest	3.07	6.81	2.88	8.15
2×4×3+int8+pp	3.11	**6.59**	**2.94**	**7.25**
2×4×3+int8+pp+pest	**3.12**	6.69	2.89	8.05



Selected Journal Publications

1. Y. Xu, J. Du, L.-R. Dai and C.-H. Lee, "An Experimental Study on Speech Enhancement Based on Deep Neural Networks," *IEEE Signal Processing Letters*, Vol. 21, No. 1, pp. 65-68, January 2014.
2. Y. Xu, J. Du, L.-R. Dai and C.-H. Lee, "A Regression Approach to Speech Enhancement Based on Deep Neural Networks," *IEEE/ACM T-ASLP.*, Vol. 23, No. 1, pp. 7-19, January 2015.
3. J. Du, Y. Tu, L.-R. Dai, C.-H. Lee, "A Regression Approach to Single-Channel Speech Separation via High-Resolution Deep Neural Networks," *IEEE/ACM T-ASLP.*, Vol. 24, No. 8, pp. 1424-1436, 2016.
4. B. Wu, K. Li, M. Yang, C.-H. Lee, "A Reverberant-Time-Aware Approach to Speech Dereverberation Based on Deep Neural Networks," *IEEE/ACM T-ASLP.*, Vol. 25, No. 1, pp. 98-107, January 2017.
5. Y. Wang, J. Du, L.-R. Dai, C.-H. Lee, "A Gender Mixture Detection Approach to Unsupervised Single-Channel Speech Separation Based on Deep Neural Networks," *IEEE/ACM T-ASLP.*, Vol. 25, No. 7, pp. 1535-1546, July 2017.
6. Y.-H. Tu, J. Du, Q. Wang, X. Bao, L.-R. Dai and C.-H. Lee, "An Information Fusion Framework with Multi-Channel Feature Concatenation and Multi-Perspective System Combination for Deep Learning Based Robust Recognition of Microphone Array Speech," *Computer Speech & Language*, Vol. 46, pp. 517-534, 2017.
7. B. Wu, K. Li, F. Ge, Z. Huang, M. Yang, S. M. Siniscalchi, and C.-H. Lee, "An End-to-End Deep Learning Approach to Simultaneous Dereverberation and Acoustic Modeling for Robust Speech Recognition," *IEEE J-STLP*, Vol. 11, Issue 8, pp. 1932-1300, December 2017.
8. T. Gao, J. Du, L.-R. Dai and C.-H. Lee, "A unified DNN approach to speaker-dependent simultaneous speech enhancement and speech separation in low SNR environments," Vol. 95, pp. 28-39, *Speech Com.*, Dec. 2017.
9. B. Wu, M. Yang, K. Li, Z. Huang, M. Siniscalchi, T. Wang and C.-H. Lee, "A Reverberation-Time-Aware Approach Leveraging Spatial Info for Microphone Array Dereverberation," *EURASIP J. on Advances in Signal Proc*, 2018.
10. Y.-H. Lai, Y. Tsao, X. Lu, F. Chen, Y.-T. Su, J. K.-C. Chen, M.-J. Lien, H.-Y. Chen, L. P.-H. Li and C.-H. Lee, "A Noise Classification Based Deep Learning Noise Reduction Approach to Improving Speech Intelligibility for Cochlear Implant Recipients," *to appear in Ear and Hearing*, 2018.

Recent Journal and Conference Efforts

- Speaker-dependent enhancement and separation (Speech Comm. 12/17)
- Array-based dereverberation (EURASIP JASP, 12/17)
- Joint SS and AM for multi-talker speech (SPS, 2018)
- Multi-objective learning and ensembling for Compact SE (T-ASLP, 07/18)
- Generalized Gaussian densities for regression error modeling (T-ASLP and IS2018)

- Multi-task learning of LPS and IRM for SE (Interspeech2015)
- DNN-based VAD: SE followed by speech detection (Interspeech2015)
- SNR-progressive learning for SE (Interspeech2016)
- ML approach to DNN parameter estimation for SE (Interspeech2017)
- Generating mixing noises with noise basis functions for SE (Interspeech2017)
- Iterative mask estimation and post-processing for Array SE (for CHiME-4, ASPSIPA2017)
- Combining conventional and DNN techniques for SE and ASR (ICASSP2018)
- SE for speaker diarization (ICASSP2018 and Interspeech2018)
- Two-stage enhancement of microphone array speech for ASR (ISCSLP2018)