Single-Microphone Speech Enhancement and Separation Using Deep Learning

Kolbæk, Morten

DOI (link to publication from Publisher):
10.54337/aau300036831

Publication date:
2018

Document Version
Other version

Single-Microphone Speech Enhancement and Separation Using Deep Learning

November 30, 2018

Morten Kolbæk

PhD Fellow
Department of Electronic Systems
Aalborg University
Denmark

AALBORG UNIVERSITY
DENMARK

Supervisors: Prof. Jesper Jensen, AAU, Oticon
Prof. Zheng-Hua Tan, AAU

Stay Abroad: Dr. Dong Yu, Tencent AI Lab / Microsoft Research

Agenda

Introduction:

- Cocktail Party Problem
- Speech Enhancement and Separation
- Deep Learning

Scientific Contributions:

- Generalization of Deep Learning based Speech Enhancement
  - Human Receivers - Speech Intelligibility
  - Machine Receivers - Speaker Verification
- On STOI Optimal Deep Learning based Speech Enhancement
- Permutation Invariant Training for Deep Learning based Speech Separation
- Summary and Conclusion

Part I

Introduction

The Cocktail Party Problem

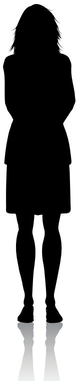
- Cocktail Party Problem
- Speech Enhancement and Separation
- Deep Learning

The Cocktail Party Problem

*How do we recognize what one person is saying when others are speaking at the same time (the "**cocktail party problem**")? On what logical basis could one **design a machine** ("filter") for carrying out such an operation?*

– Colin Cherry, 1953.

The Cocktail Party Problem
The Vision: Solve the Problem



The Cocktail Party Problem

The Vision: Solve the Problem

The Cocktail Party Problem

The Vision: Solve the Problem

The Cocktail Party Problem

The Vision: Solve the Problem

The Cocktail Party Problem

The Vision: Solve the Problem

The Cocktail Party Problem

The Vision: Solve the Problem

The Cocktail Party Problem

The Vision: Solve the Problem

The Cocktail Party Problem

The Vision: Solve the Problem

The Cocktail Party Problem

The Vision: Solve the Problem

The Cocktail Party Problem

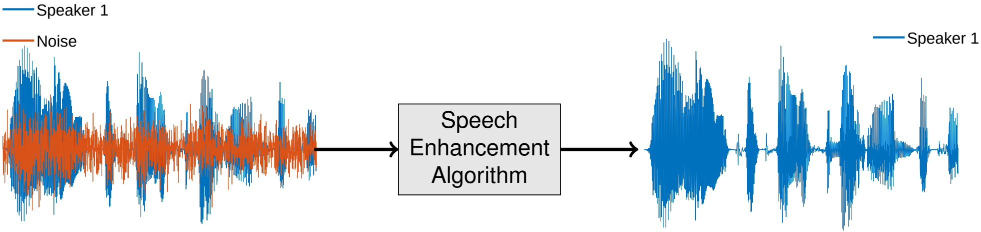
The Vision: Solve the Problem

Speech Enhancement and Separation

- Cocktail Party Problem
- Speech Enhancement and Separation
- Deep Learning

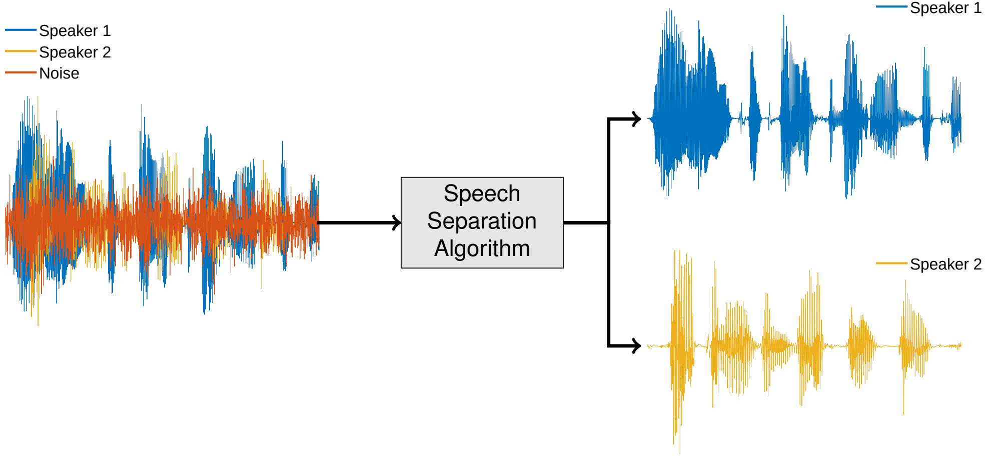
Single-Microphone Speech Enhancement

First Task of the Thesis



Single-Microphone Speech Separation

Second Task of the Thesis



Speech Enhancement and Separation
Two Motivating Applications

Why Is Solving the Cocktail Party Problem Important?

Human Receivers

- **Potential:** Hundreds of millions of people worldwide have a hearing loss.
- **Challenge:** Hearing impaired often struggle in "cocktail party" situations.
- **Solution:** Algorithms that can enhance the speech signal of interest.
- **Application:** Hearing Assistive Devices e.g. hearing aids or cochlear implants.

Machine Receivers

- **Potential:** Millions of people vocally interact with smartphones.
- **Challenge:** These devices operate in complex acoustic environments.
- **Solution:** Noise-robust human-machine interface.
- **Application:** Social robots or digital assistants e.g. Google Asst., Siri, etc.

Speech Enhancement and Separation

Two Motivating Applications

Why Is Solving the Cocktail Party Problem Important?

Human Receivers

- **Potential:** Hundreds of millions of people worldwide have a hearing loss.
- **Challenge:** Hearing impaired often struggle in "cocktail party" situations.
- **Solution:** Algorithms that can enhance the speech signal of interest.
- **Application:** Hearing Assistive Devices e.g. hearing aids or cochlear implants.

Machine Receivers

- **Potential:** Millions of people vocally interact with smartphones.
- **Challenge:** These devices operate in complex acoustic environments.
- **Solution:** Noise-robust human-machine interface.
- **Application:** Social robots or digital assistants e.g. Google Asst., Siri, etc.

Speech Enhancement and Separation
Old Problem: Whats new?

Whats new? – A paradigm shift!

Classical Paradigm

- Derive the solution using specific mathematical models that approximate speech and noise.
- Simplifying assumptions for mathematical tractability.
- Generally not data-driven.
- Good performance when assumptions are valid (sometimes they are not).

Deep Learning Paradigm

- Learn the solution using general mathematical models that have "observed" speech and noise.
- No explicit assumptions.
- Data-driven.
- State-of-the-art performance given enough data and computational resources.

Speech Enhancement and Separation

Old Problem: Whats new?

Whats new? – A paradigm shift!

Classical Paradigm

- Derive the solution using specific mathematical models that approximate speech and noise.
- Simplifying assumptions for mathematical tractability.
- Generally not data-driven.
- Good performance when assumptions are valid (sometimes they are not).

Deep Learning Paradigm

- Learn the solution using general mathematical models that have "observed" speech and noise.
- No explicit assumptions.
- Data-driven.
- State-of-the-art performance given enough data and computational resources.

Speech Enhancement and Separation
Old Problem: Whats new?

Whats new? – A paradigm shift!

Classical Paradigm

- Derive the solution using specific mathematical models that approximate speech and noise.
- Simplifying assumptions for mathematical tractability.
- Generally not data-driven.
- Good performance when assumptions are valid (sometimes they are not).

Deep Learning Paradigm

- Learn the solution using general mathematical models that have "observed" speech and noise.
- No explicit assumptions.
- Data-driven.
- State-of-the-art performance given enough data and computational resources.

Speech Enhancement and Separation

Old Problem: Whats new?

Whats new? – A paradigm shift!

Classical Paradigm

- Derive the solution using specific mathematical models that approximate speech and noise.
- Simplifying assumptions for mathematical tractability.
- Generally not data-driven.
- Good performance when assumptions are valid (sometimes they are not).

Deep Learning Paradigm

- Learn the solution using general mathematical models that have "observed" speech and noise.
- No explicit assumptions.
- Data-driven.
- State-of-the-art performance given enough data and computational resources.

Speech Enhancement and Separation

Old Problem: Whats new?

Whats new? – A paradigm shift!

Classical Paradigm

- Derive the solution using specific mathematical models that approximate speech and noise.
- Simplifying assumptions for mathematical tractability.
- Generally not data-driven.
- Good performance when assumptions are valid (sometimes they are not).

Deep Learning Paradigm

- Learn the solution using general mathematical models that have "observed" speech and noise.
- No explicit assumptions.
- Data-driven.
- State-of-the-art performance given enough data and computational resources.

Speech Enhancement and Separation

Old Problem: Whats new?

Whats new? – A paradigm shift!

Classical Paradigm

- Derive the solution using specific mathematical models that approximate speech and noise.
- Simplifying assumptions for mathematical tractability.
- Generally not data-driven.
- Good performance when assumptions are valid (sometimes they are not).

Deep Learning Paradigm

- Learn the solution using general mathematical models that have "observed" speech and noise.
- No explicit assumptions.
- Data-driven.
- State-of-the-art performance given enough data and computational resources.

Speech Enhancement and Separation

Old Problem: Whats new?

Whats new? – A paradigm shift!

Classical Paradigm

- Derive the solution using specific mathematical models that approximate speech and noise.
- Simplifying assumptions for mathematical tractability.
- Generally not data-driven.
- Good performance when assumptions are valid (sometimes they are not).

Deep Learning Paradigm

- Learn the solution using general mathematical models that have "observed" speech and noise.
- No explicit assumptions.
- Data-driven.
- State-of-the-art performance given enough data and computational resources.

Speech Enhancement and Separation
Old Problem: Whats new?

Whats new? – A paradigm shift!

Classical Paradigm

- Derive the solution using specific mathematical models that approximate speech and noise.
- Simplifying assumptions for mathematical tractability.
- Generally not data-driven.
- Good performance when assumptions are valid (sometimes they are not).

Deep Learning Paradigm

- Learn the solution using general mathematical models that have "observed" speech and noise.
- No explicit assumptions.
- Data-driven.
- State-of-the-art performance given enough data and computational resources.

Speech Enhancement and Separation

Old Problem: Whats new?

Whats new? – A paradigm shift!

Classical Paradigm

- Derive the solution using specific mathematical models that approximate speech and noise.
- Simplifying assumptions for mathematical tractability.
- Generally not data-driven.
- Good performance when assumptions are valid (sometimes they are not).

Deep Learning Paradigm

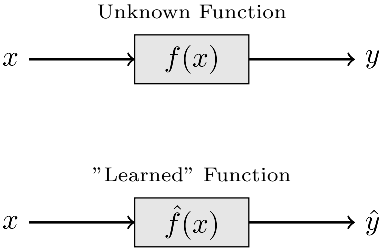
- Learn the solution using general mathematical models that have "observed" speech and noise.
- No explicit assumptions.
- Data-driven.
- State-of-the-art performance given enough data and computational resources.

Deep Learning

- Cocktail Party Problem
- Speech Enhancement and Separation
- Deep Learning

Deep Learning
What is it?

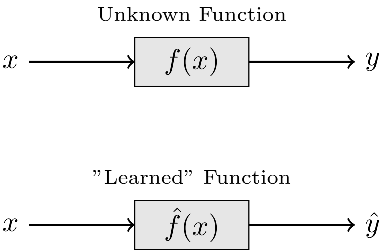
- **Deep Learning:** Subfield of Machine Learning.
- **Machine Learning:** Use data to "learn" or approximate unknown functions $f(x)$ that can be used to make predictions.



$$\hat{y} \approx y$$

Deep Learning
What is it?

- **Deep Learning:** Subfield of Machine Learning.
- **Machine Learning:** Use data to "learn" or approximate unknown functions $f(x)$ that can be used to make predictions.



$$\hat{y} \approx y$$

Deep Learning

What is it? – Classical Regression Example

- Estimate Happiness from income
  - Hypothesis: Happiness is associated with income.
  - Data: Perceived happiness and income from people.
  - Candidate Models:
    - 7-params. (Big Capacity)

$$\hat{f}_1(x) = ax^6 + bx^5 + cx^4 + dx^3 + ex^2 + fx + g$$

    - 4-params. (Small Capacity)

$$\hat{f}_2(x) = ax^3 + bx^2 + cx + d$$

  - **Goal:** Find parameters of $\hat{f}_1(x)$ and $\hat{f}_2(x)$ that best explain the observations.

Deep Learning

What is it? – Classical Regression Example

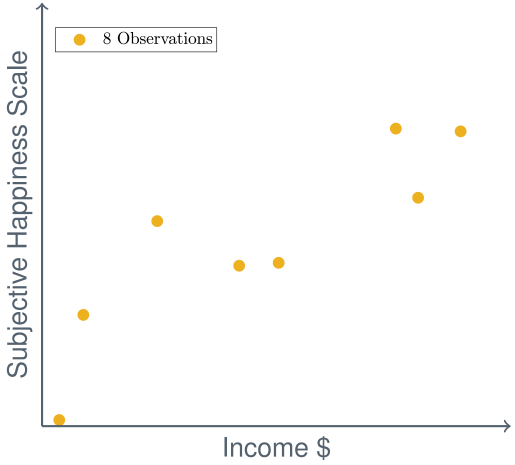
- Estimate Happiness from income
  - Hypothesis: Happiness is associated with income.
  - Data: Perceived happiness and income from people.
  - Candidate Models:
    - 7-params. (Big Capacity)

$$\hat{f}_1(x) = ax^6 + bx^5 + cx^4 + dx^3 + ex^2 + fx + g$$

    - 4-params. (Small Capacity)

$$\hat{f}_2(x) = ax^3 + bx^2 + cx + d$$

  - **Goal:** Find parameters of $\hat{f}_1(x)$ and $\hat{f}_2(x)$ that best explain the observations.



Deep Learning

What is it? – Classical Regression Example

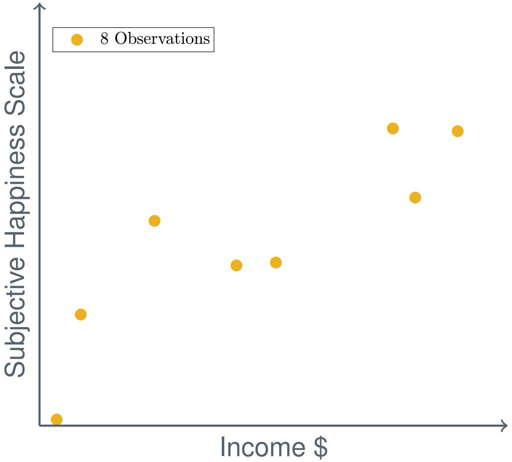
- Estimate Happiness from income
  - Hypothesis: Happiness is associated with income.
  - Data: Perceived happiness and income from people.
  - Candidate Models:
    - 7-params. (Big Capacity)

$$\hat{f}_1(x) = ax^6 + bx^5 + cx^4 + dx^3 + ex^2 + fx + g$$

    - 4-params. (Small Capacity)

$$\hat{f}_2(x) = ax^3 + bx^2 + cx + d$$

  - **Goal:** Find parameters of $\hat{f}_1(x)$ and $\hat{f}_2(x)$ that best explain the observations.



Deep Learning

What is it? – Classical Regression Example

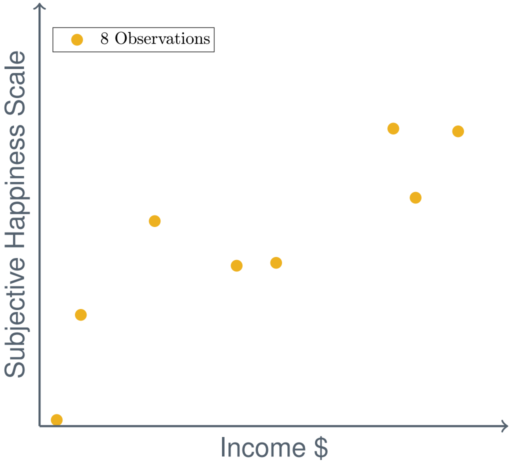
- Estimate Happiness from income
  - Hypothesis: Happiness is associated with income.
  - Data: Perceived happiness and income from people.
  - Candidate Models:
    - 7-params. (Big Capacity)

$$\hat{f}_1(x) = ax^6 + bx^5 + cx^4 + dx^3 + ex^2 + fx + g$$

    - 4-params. (Small Capacity)

$$\hat{f}_2(x) = ax^3 + bx^2 + cx + d$$

  - **Goal:** Find parameters of $\hat{f}_1(x)$ and $\hat{f}_2(x)$ that best explain the observations.



Deep Learning

What is it? – Classical Regression Example

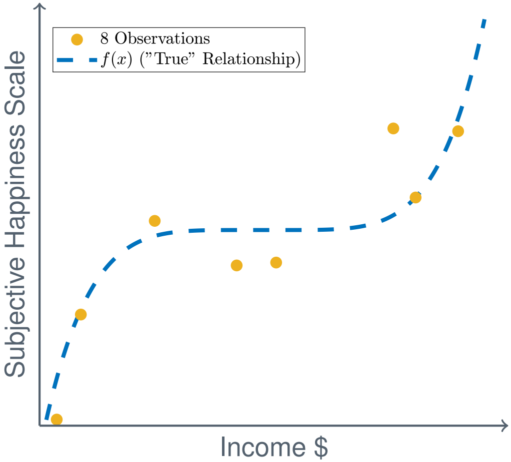
- Estimate Happiness from income
  - Hypothesis: Happiness is associated with income.
  - Data: Perceived happiness and income from people.
  - Candidate Models:
    - 7-params. (Big Capacity)

$$\hat{f}_1(x) = ax^6 + bx^5 + cx^4 + dx^3 + ex^2 + fx + g$$

    - 4-params. (Small Capacity)

$$\hat{f}_2(x) = ax^3 + bx^2 + cx + d$$

  - **Goal:** Find parameters of $\hat{f}_1(x)$ and $\hat{f}_2(x)$ that best explain the observations.



Deep Learning

What is it? – Classical Regression Example

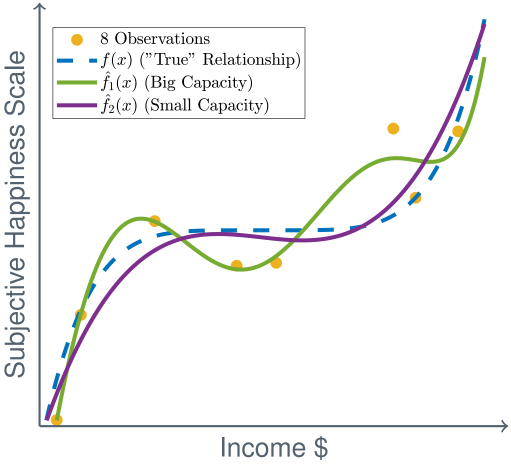
- Estimate Happiness from income
  - Hypothesis: Happiness is associated with income.
  - Data: Perceived happiness and income from people.
  - Candidate Models:
    - 7-params. (Big Capacity)

$$\hat{f}_1(x) = -0.2x^6 + 2.5x^5 - 8.1x^4 + 10.3x^3 - 5.4x^2 + 1.2x + 0.3$$

    - 4-params. (Small Capacity)

$$\hat{f}_2(x) = -22.2x^3 + 2.6x^2 + 3.8x - 0.6$$

  - **Goal:** Find parameters of $\hat{f}_1(x)$ and $\hat{f}_2(x)$ that best explain the observations.



Deep Learning

What is it? – Classical Regression Example

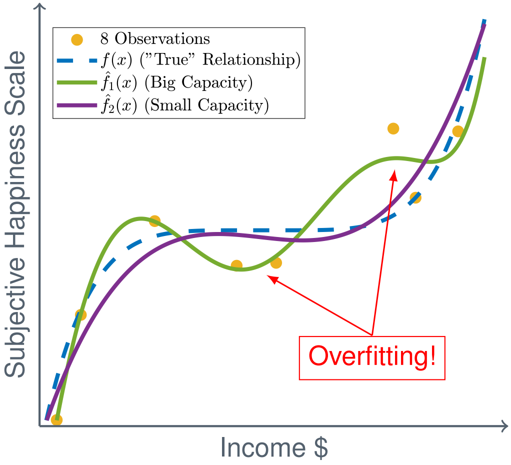
- Estimate Happiness from income
  - Hypothesis: Happiness is associated with income.
  - Data: Perceived happiness and income from people.
  - Candidate Models:
    - 7-params. (Big Capacity)

$$\hat{f}_1(x) = -0.2x^6 + 2.5x^5 - 8.1x^4 + 10.3x^3 - 5.4x^2 + 1.2x + 0.3$$

    - 4-params. (Small Capacity)

$$\hat{f}_2(x) = -22.2x^3 + 2.6x^2 + 3.8x - 0.6$$

  - **Goal:** Find parameters of $\hat{f}_1(x)$ and $\hat{f}_2(x)$ that best explain the observations.



Deep Learning

What is it? – Classical Regression Example

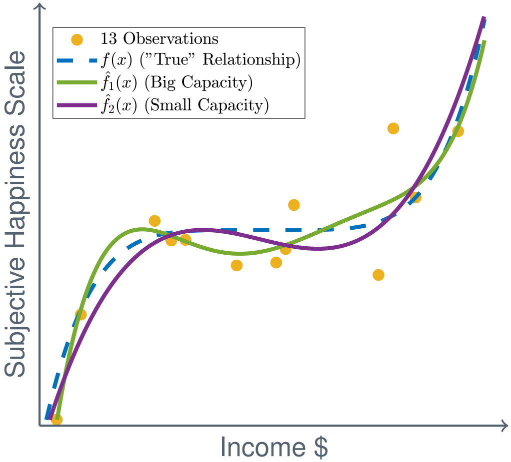
- Estimate Happiness from income
  - Hypothesis: Happiness is associated with income.
  - Data: Perceived happiness and income from people.
  - Candidate Models:
    - 7-params. (Big Capacity)

$$\hat{f}_1(x) = 1.1x^6 - 6.5x^5 + 15.1x^4 - 18.0x^3 + 11.0x^2 - 2.7x + 0.6$$

    - 4-params. (Small Capacity)

$$\hat{f}_2(x) = 18.2x^3 - 19.4x^2 + 9.3x - 1.2$$

  - **Goal:** Find parameters of $\hat{f}_1(x)$ and $\hat{f}_2(x)$ that best explain the observations.



Deep Learning

What is it? – Classical Regression Example

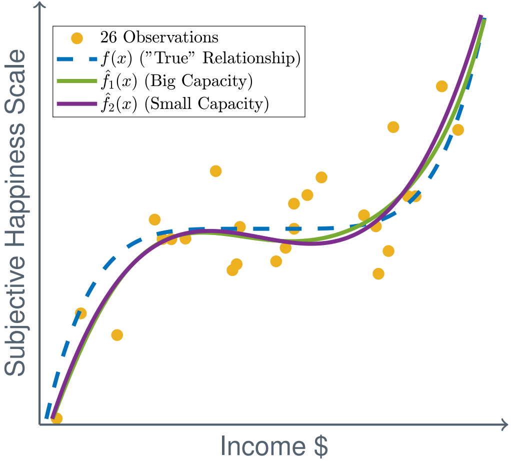
- Estimate Happiness from income
  - Hypothesis: Happiness is associated with income.
  - Data: Perceived happiness and income from people.
  - Candidate Models:
    - 7-params. (Big Capacity)

$$\hat{f}_1(x) = -0.3x^6 + 3.2x^5 + 11.1x^4 + 17.3x^3 - 13.6x^2 + 5.6x - 0.5$$

    - 4-params. (Small Capacity)

$$\hat{f}_2(x) = -9.2x^3 + 2.9x^2 + 1.1x - 0.2$$

  - **Goal:** Find parameters of $\hat{f}_1(x)$ and $\hat{f}_2(x)$ that best explain the observations.



Deep Learning

What is it? – Classical Regression Example

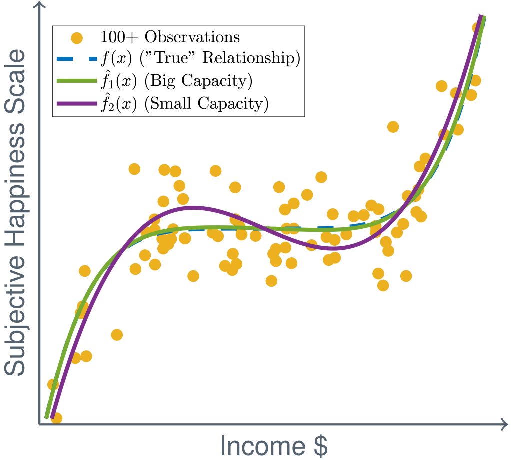
- Estimate Happiness from income
  - Hypothesis: Happiness is associated with income.
  - Data: Perceived happiness and income from people.
  - Candidate Models:
    - 7-params. (Big Capacity)

$$\hat{f}_1(x) = -0.1x^6 + 2.0x^5 - 7.7x^4 + 13.3x^3 - 11.7x^2 + 5.3x - 0.5$$

    - 4-params. (Small Capacity)

$$\hat{f}_2(x) = 10.9x^3 - 10.4x^2 + 5.1x - 0.5$$

  - **Goal:** Find parameters of $\hat{f}_1(x)$ and $\hat{f}_2(x)$ that best explain the observations.



Deep Learning

What is it? – Classical Regression Example

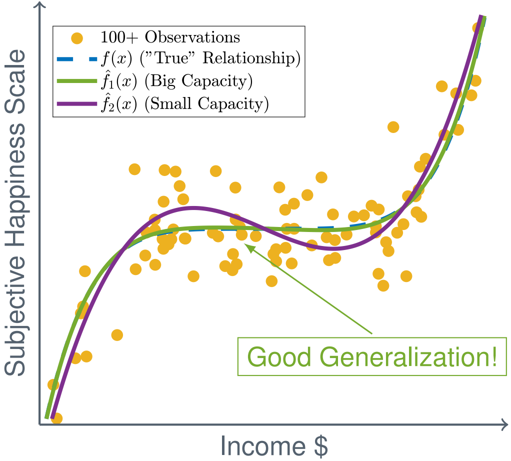
- Estimate Happiness from income
  - Hypothesis: Happiness is associated with income.
  - Data: Perceived happiness and income from people.
  - Candidate Models:
    - 7-params. (Big Capacity)

$$\hat{f}_1(x) = -0.1x^6 + 2.0x^5 - 7.7x^4 + 13.3x^3 - 11.7x^2 + 5.3x - 0.5$$

    - 4-params. (Small Capacity)

$$\hat{f}_2(x) = 10.9x^3 - 10.4x^2 + 5.1x - 0.5$$

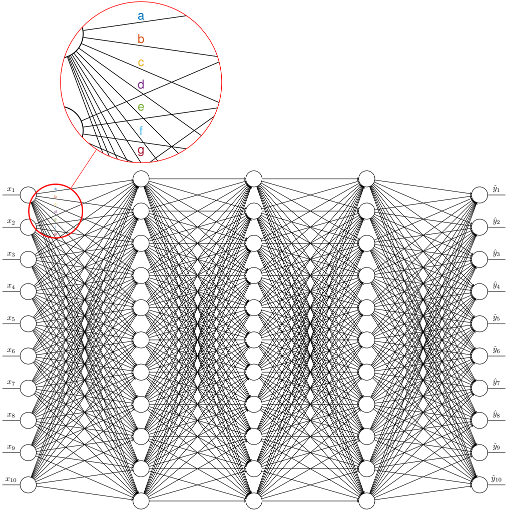
  - **Goal:** Find parameters of $\hat{f}_1(x)$ and $\hat{f}_2(x)$ that best explain the observations.



Deep Learning

What is it? – Essentially Regression with Deep Neural Networks

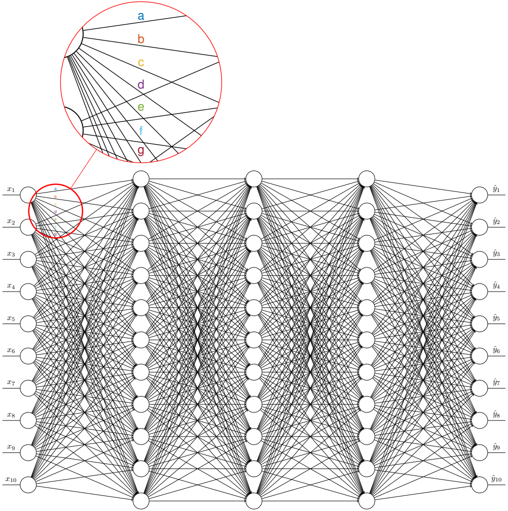
- Deep Learning
  - "Regression" using Deep Neural Networks.
- Deep Neural Network
  - Non-linear function with potentially MANY (millions) parameters.
  - If big enough, they can approximate any function.
  - With enough data, they can learn complex mappings.



Deep Learning

What is it? – Essentially Regression with Deep Neural Networks

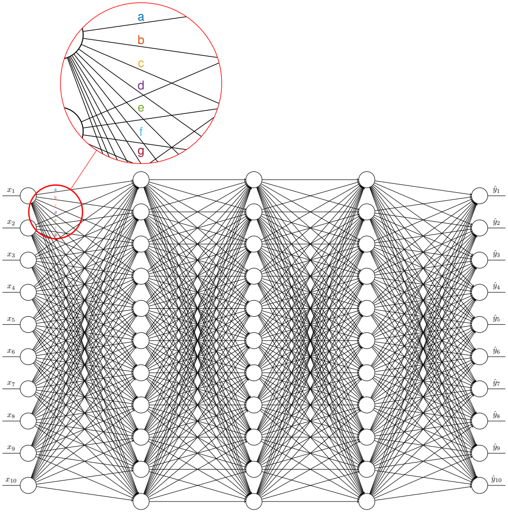
- Deep Learning
  - "Regression" using Deep Neural Networks.
- Deep Neural Network
  - Non-linear function with potentially MANY (millions) parameters.
  - If big enough, they can approximate any function.
  - With enough data, they can learn complex mappings.



Deep Learning

What is it? – Essentially Regression with Deep Neural Networks

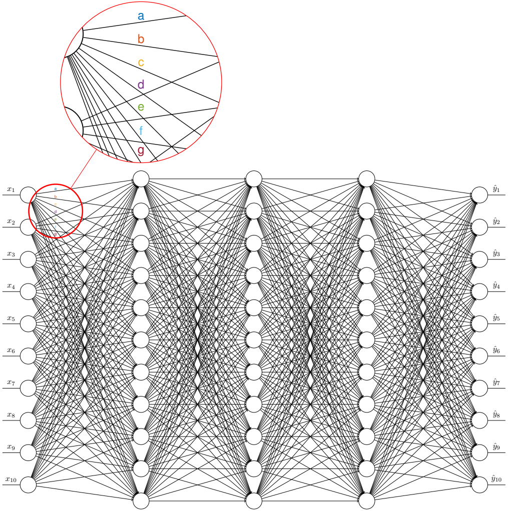
- Deep Learning
  - "Regression" using Deep Neural Networks.
- Deep Neural Network
  - Non-linear function with potentially MANY (millions) parameters.
  - If big enough, they can approximate any function.
  - With enough data, they can learn complex mappings.



Deep Learning

What is it? – Essentially Regression with Deep Neural Networks

- Deep Learning
  - "Regression" using Deep Neural Networks.
- Deep Neural Network
  - Non-linear function with potentially MANY (millions) parameters.
  - If big enough, they can approximate any function.
  - With enough data, they can learn complex mappings.



Deep Learning
What Can It Do?

Deep Learning
What Can It Do?

Deep Learning
What Can It Do?

Deep Learning
What Can It Do?

Deep Learning

What Can It Do?

Google

Deep Learning

What Can It Do?

Google

facebook

Deep Learning
What Can It Do?

Deep Learning

What Can It Do?

Google

facebook

Microsoft

Deep Learning
What Can It Do?

Google
facebook
Microsoft
amazon

Deep Learning
What Can It Do?

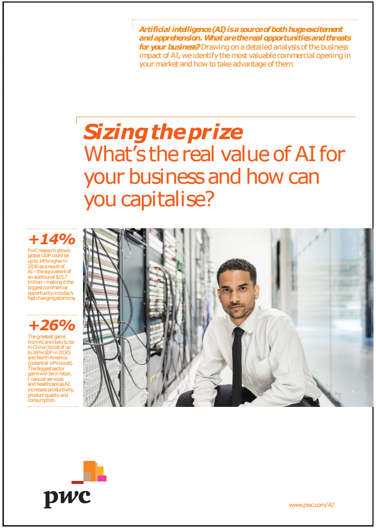
Google
facebook
Microsoft
amazon
Alibaba Group

Deep Learning
What Can It Do?

$15.7 trillion

Game changer

AI could contribute up to $15.7 trillion to the global economy in 2030, more than the current output of China and India combined.



Artificial intelligence (AI) is a source of both huge excitement and apprehension. What are the real opportunities and threats for your business? Drawing on a detailed analysis of the business impact of AI, we identify the most valuable commercial opening in your market and how to take advantage of them.

Sizing the prize
What's the real value of AI for your business and how can you capitalise?

+14%
PwC research shows global GDP could be up to 14% higher in 2030 as a result of AI – the equivalent of an additional $15.7 trillion – making it the biggest commercial opportunity in today's fast changing economy.

+26%
The greatest gains from AI are likely to be in China (boost of up to 26% GDP in 2030) and North America (potential 14% boost). The biggest sector gains will be in retail, financial services and healthcare as AI increases productivity, product quality and consumption.

pwc

www.pwc.com/AI

Part II

Scientific Contributions

Generalization of DNN based Speech Enhancement

Human Receivers - Speech Intelligibility

Generalization of DNN based Speech Enhancement

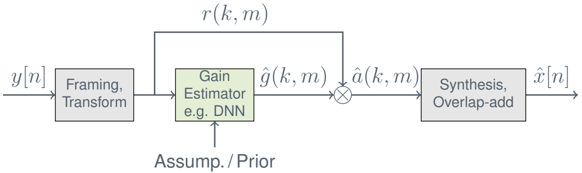
Human Receivers - Motivation and Research Gap

Promising Results

- Recent studies show that speech enhancement algorithms based on deep learning outperform classical techniques.
- DNNs trained and tested in **"narrow"** conditions.

Research Gap

- **Unknown** how these algorithms perform in general **"broader"** conditions and in conditions with a mismatch between training and test.



$y[n]$: Noisy speech (time-domain)
$r(k,m)$: Noisy speech (transform-domain)
$\hat{g}(k,m)$: Estimated gain
$\hat{a}(k,m)$: Enhanced speech (transform-domain)
$\hat{x}[n]$: Enhanced speech (time-domain)

Generalization of DNN based Speech Enhancement

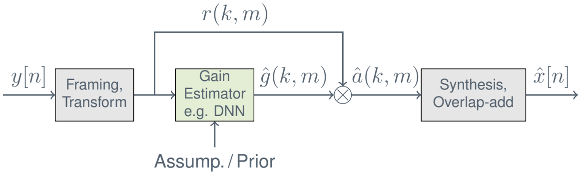
Human Receivers - Motivation and Research Gap

Promising Results

- Recent studies show that speech enhancement algorithms based on deep learning outperform classical techniques.
- DNNs trained and tested in **"narrow"** conditions.

Research Gap

- **Unknown** how these algorithms perform in general **"broader"** conditions and in conditions with a mismatch between training and test.



$y[n]$: Noisy speech (time-domain)
$r(k,m)$: Noisy speech (transform-domain)
$\hat{g}(k,m)$: Estimated gain
$\hat{a}(k,m)$: Enhanced speech (transform-domain)
$\hat{x}[n]$: Enhanced speech (time-domain)

Generalization of DNN based Speech Enhancement

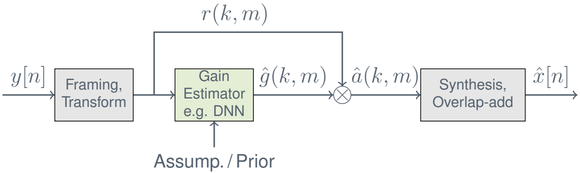
Human Receivers - Motivation and Research Gap

Promising Results

- Recent studies show that speech enhancement algorithms based on deep learning outperform classical techniques.
- DNNs trained and tested in **"narrow"** conditions.

Research Gap

- **Unknown** how these algorithms perform in general **"broader"** conditions and in conditions with a mismatch between training and test.



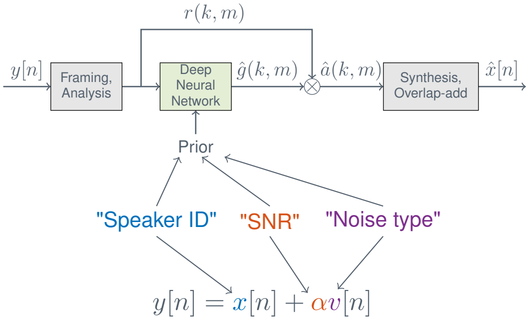
$y[n]$: Noisy speech (time-domain)
$r(k,m)$: Noisy speech (transform-domain)
$\hat{g}(k,m)$: Estimated gain
$\hat{a}(k,m)$: Enhanced speech (transform-domain)
$\hat{x}[n]$: Enhanced speech (time-domain)

Generalization of DNN based Speech Enhancement

Human Receivers - Contribution

Contribution

- We studied generalizability capability of deep neural network-based speech enhancement algorithms for additive-noise corrupted speech [1].
- Specifically, our goal was to study the generalization error w.r.t. three dimensions:
  - Speaker Identity
  - Signal-to-Noise Ratio
  - Noise type
- We trained multiple DNNs with various priors.
- Generalization was evaluated using PESQ and STOI, which are speech quality and intelligibility estimators, respectively.



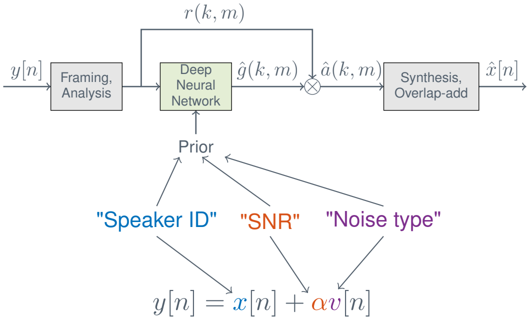
[1] M. Kolbæk, *et al.*, *IEEE TASLP*, 2017

Generalization of DNN based Speech Enhancement

Human Receivers - Contribution

Contribution

- We studied generalizability capability of deep neural network-based speech enhancement algorithms for additive-noise corrupted speech [1].
- Specifically, our goal was to study the generalization error w.r.t. three dimensions:
  - Speaker Identity
  - Signal-to-Noise Ratio
  - Noise type
- We trained multiple DNNs with various priors.
- Generalization was evaluated using PESQ and STOI, which are speech quality and intelligibility estimators, respectively.



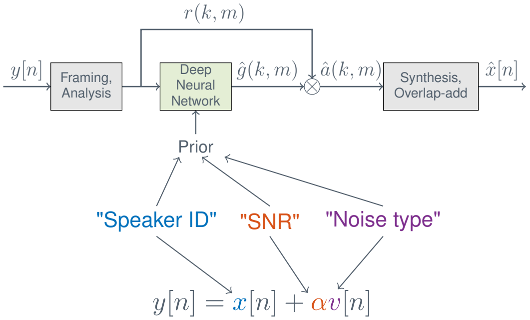
[1] M. Kolbæk, *et al.*, *IEEE TASLP*, 2017

Generalization of DNN based Speech Enhancement

Human Receivers - Contribution

Contribution

- We studied generalizability capability of deep neural network-based speech enhancement algorithms for additive-noise corrupted speech [1].
- Specifically, our goal was to study the generalization error w.r.t. three dimensions:
  - Speaker Identity
  - Signal-to-Noise Ratio
  - Noise type
- We trained multiple DNNs with various priors.
- Generalization was evaluated using PESQ and STOI, which are speech quality and intelligibility estimators, respectively.



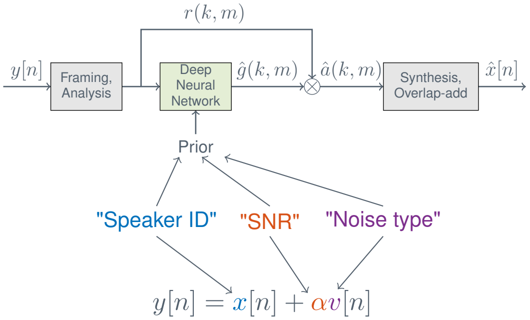
[1] M. Kolbæk, *et al.*, *IEEE TASLP*, 2017

Generalization of DNN based Speech Enhancement

Human Receivers - Contribution

Contribution

- We studied generalizability capability of deep neural network-based speech enhancement algorithms for additive-noise corrupted speech [1].
- Specifically, our goal was to study the generalization error w.r.t. three dimensions:
  - Speaker Identity
  - Signal-to-Noise Ratio
  - Noise type
- We trained multiple DNNs with various priors.
- Generalization was evaluated using PESQ and STOI, which are speech quality and intelligibility estimators, respectively.



[1] M. Kolbæk, *et al.*, *IEEE TASLP*, 2017

Generalization of DNN based Speech Enhancement

Human Receivers - Results and Conclusion

Results and Conclusion

- Performance (PESQ and STOI) is generally reduced when a "narrow" system is tested in a more general scenario.
- Performance is comparable or exceeding performance of a classical technique.
- Matching the noise type is the most critical, whereas matching the speaker and SNR is less critical.
- Listening tests show small improvement in speech intelligibility relative to previously published results.
- Both PESQ and informal listening tests indicate that DNN systems improve speech quality.

Generalization of DNN based Speech Enhancement
Human Receivers - Results and Conclusion

Results and Conclusion

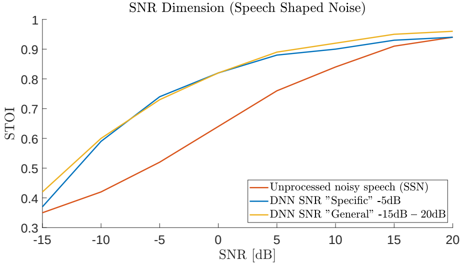
- Performance (PESQ and STOI) is generally reduced when a "narrow" system is tested in a more general scenario.
- Performance is comparable or exceeding performance of a classical technique.
- Matching the noise type is the most critical, whereas matching the speaker and SNR is less critical.
- Listening tests show small improvement in speech intelligibility relative to previously published results.
- Both PESQ and informal listening tests indicate that DNN systems improve speech quality.

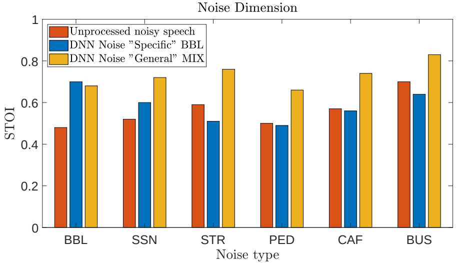
Generalization of DNN based Speech Enhancement

Human Receivers - Results and Conclusion

Results and Conclusion

- Performance (PESQ and STOI) is generally reduced when a "narrow" system is tested in a more general scenario.
- Performance is comparable or exceeding performance of a classical technique.
- Matching the noise type is the most critical, whereas matching the speaker and SNR is less critical.
- Listening tests show small improvement in speech intelligibility relative to previously published results.
- Both PESQ and informal listening tests indicate that DNN systems improve speech quality.



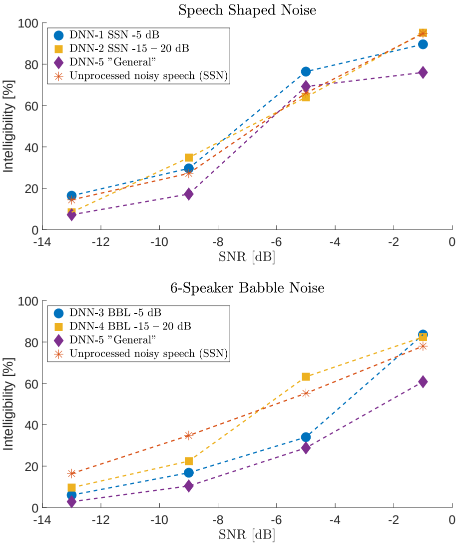


Generalization of DNN based Speech Enhancement

Human Receivers - Results and Conclusion

Results and Conclusion

- Performance (PESQ and STOI) is generally reduced when a "narrow" system is tested in a more general scenario.
- Performance is comparable or exceeding performance of a classical technique.
- Matching the noise type is the most critical, whereas matching the speaker and SNR is less critical.
- Listening tests show small improvement in speech intelligibility relative to previously published results.
- Both PESQ and informal listening tests indicate that DNN systems improve speech quality.

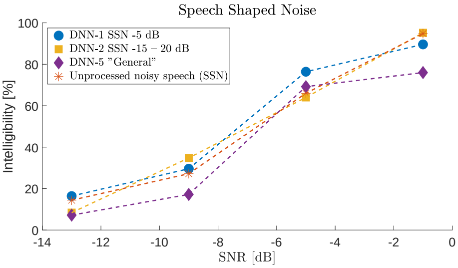


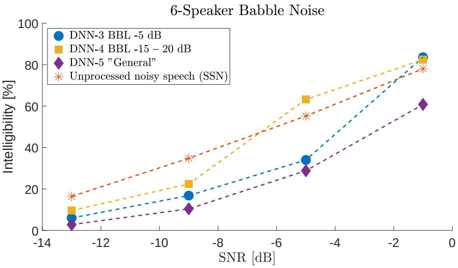
Generalization of DNN based Speech Enhancement

Human Receivers - Results and Conclusion

Results and Conclusion

- Performance (PESQ and STOI) is generally reduced when a "narrow" system is tested in a more general scenario.
- Performance is comparable or exceeding performance of a classical technique.
- Matching the noise type is the most critical, whereas matching the speaker and SNR is less critical.
- Listening tests show small improvement in speech intelligibility relative to previously published results.
- Both PESQ and informal listening tests indicate that DNN systems improve speech quality.





Generalization of DNN based Speech Enhancement

Human Receivers - Speech Intelligibility

Generalization of DNN based Speech Enhancement

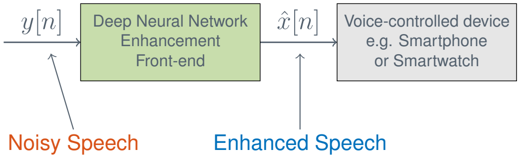
Machine Receivers - Speaker Verification

Motivation

- Digital devices with voice-user interfaces struggle in "cocktail-party" conditions.
- Such devices can benefit from denoising front-ends.
- A State-of-the-art noise-robust speaker verification system relies on speaker dependent non-negative matrix factorization (Thomsen *et al.* 2016).

Research Gap

- It is unknown how well DNN based speech enhancement algorithms work as denoising front-ends for speaker verification systems.



Generalization of DNN based Speech Enhancement

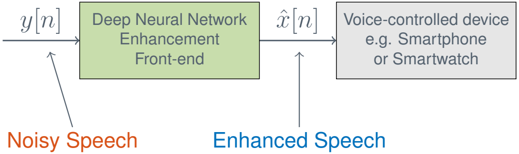
Machine Receivers - Speaker Verification

Motivation

- Digital devices with voice-user interfaces struggle in "cocktail-party" conditions.
- Such devices can benefit from denoising front-ends.
- A State-of-the-art noise-robust speaker verification system relies on speaker dependent non-negative matrix factorization (Thomsen *et al.* 2016).

Research Gap

- It is unknown how well DNN based speech enhancement algorithms work as denoising front-ends for speaker verification systems.

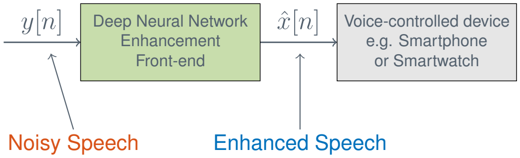


Generalization of DNN based Speech Enhancement

Machine Receivers - Speaker Verification

Motivation

- Digital devices with voice-user interfaces struggle in "cocktail-party" conditions.
- Such devices can benefit from denoising front-ends.
- A State-of-the-art noise-robust speaker verification system relies on speaker dependent non-negative matrix factorization (Thomsen *et al.* 2016).



Research Gap

- It is unknown how well DNN based speech enhancement algorithms work as denoising front-ends for speaker verification systems.

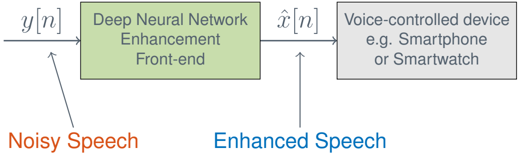
Generalization of DNN based Speech Enhancement
Machine Receivers - Speaker Verification

Motivation

- Digital devices with voice-user interfaces struggle in "cocktail-party" conditions.
- Such devices can benefit from denoising front-ends.
- A State-of-the-art noise-robust speaker verification system relies on speaker dependent non-negative matrix factorization (Thomsen *et al.* 2016).

Research Gap

- It is unknown how well DNN based speech enhancement algorithms work as denoising front-ends for speaker verification systems.



Generalization of DNN based Speech Enhancement

Machine Receivers - Contribution

Contribution

- We designed a DNN based speech enhancement front-end for a speaker verification system [2].
- Goal was to study the generalization error w.r.t. three dimensions:
  - Speaker Identity
  - Signal-to-Noise Ratio
  - Noise type
- Generalization was evaluated using equal error rates and the results were compared to existing enhancement techniques.

[2] M. Kolbæk, *et al.*, *IEEE SLT*, 2016

Generalization of DNN based Speech Enhancement

Machine Receivers - Contribution

Contribution

- We designed a DNN based speech enhancement front-end for a speaker verification system [2].
- Goal was to study the generalization error w.r.t. three dimensions:
  - Speaker Identity
  - Signal-to-Noise Ratio
  - Noise type
- Generalization was evaluated using equal error rates and the results were compared to existing enhancement techniques.

[2] M. Kolbæk, *et al.*, *IEEE SLT*, 2016

Generalization of DNN based Speech Enhancement

Machine Receivers - Contribution

Contribution

- We designed a DNN based speech enhancement front-end for a speaker verification system [2].
- Goal was to study the generalization error w.r.t. three dimensions:
  - Speaker Identity
  - Signal-to-Noise Ratio
  - Noise type
- Generalization was evaluated using equal error rates and the results were compared to existing enhancement techniques.

[2] M. Kolbæk, *et al.*, *IEEE SLT*, 2016

Generalization of DNN based Speech Enhancement

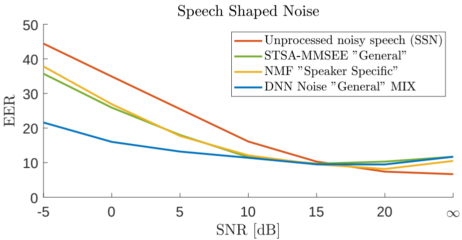
Machine Receivers - Results and Conclusion

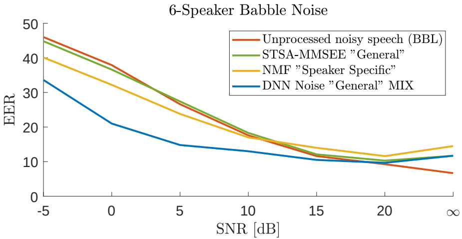
Results

- Male-speaker "general" DNN-based speech enhancement front-end generally leads to lower EER compared to classical techniques.
- Even NMF which is "narrow", i.e. speaker, text, and noise type dependent.

Conclusion

- DNN based speech enhancement front-end improves state-of-the-art noise-robust speaker verification.
- Eliminating the need for noise type and speaker dependent front-ends.





Generalization of DNN based Speech Enhancement

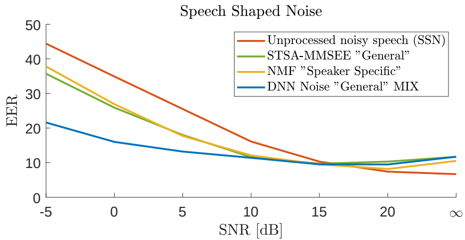
Machine Receivers - Results and Conclusion

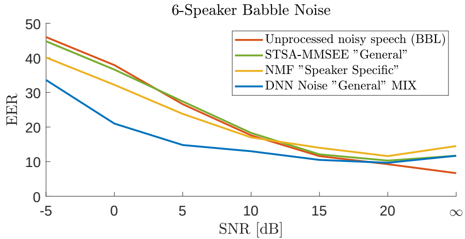
Results

- Male-speaker "general" DNN-based speech enhancement front-end generally leads to lower EER compared to classical techniques.
- Even NMF which is "narrow", i.e. speaker, text, and noise type dependent.

Conclusion

- DNN based speech enhancement front-end improves state-of-the-art noise-robust speaker verification.
- Eliminating the need for noise type and speaker dependent front-ends.





Generalization of DNN based Speech Enhancement

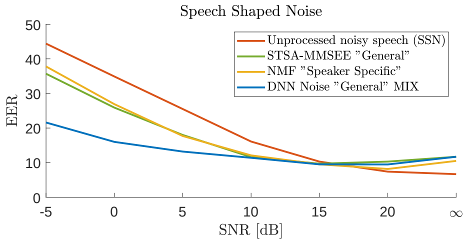
Machine Receivers - Results and Conclusion

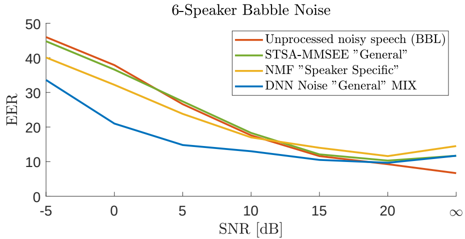
Results

- Male-speaker "general" DNN-based speech enhancement front-end generally leads to lower EER compared to classical techniques.
- Even NMF which is "narrow", i.e. speaker, text, and noise type dependent.

Conclusion

- DNN based speech enhancement front-end improves state-of-the-art noise-robust speaker verification.
- Eliminating the need for noise type and speaker dependent front-ends.



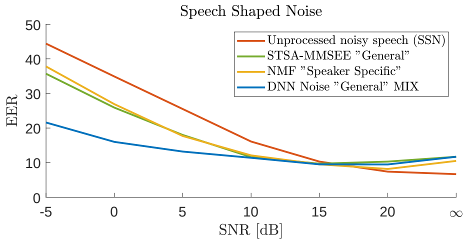


Generalization of DNN based Speech Enhancement

Machine Receivers - Results and Conclusion

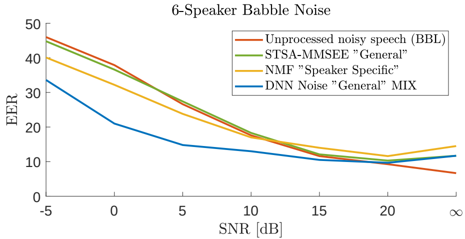
Results

- Male-speaker "general" DNN-based speech enhancement front-end generally leads to lower EER compared to classical techniques.
- Even NMF which is "narrow", i.e. speaker, text, and noise type dependent.



Conclusion

- DNN based speech enhancement front-end improves state-of-the-art noise-robust speaker verification.
- Eliminating the need for noise type and speaker dependent front-ends.



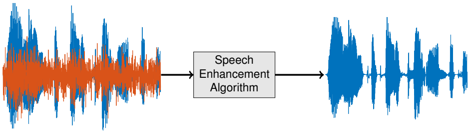
On STOI Optimal DNN based Speech Enhancement
Optimality

On STOI Optimal DNN based Speech Enhancement

Motivation, Research Gap, and Contribution

Motivation

- Goal of speech enhancement algorithms is often to improve speech intelligibility (SI).
- Often these algorithms are designed to minimize the short-time spectral amplitude (STSA) mean-square error (MSE) with no link to SI.
- Can we use a function with a stronger link to SI? – e.g. the STOI SI estimator.



Mean-Square Error:

$$J_{MSE} = \frac{1}{K}\sum_{k=1}^{K}(a(k,m) - \hat{a}(k,m))^2$$

$a(k,m)$: Clean Speech STFT Amplitudes

$\hat{a}(k,m)$: Enhanced Clean Speech STFT Amplitudes

Research Gap

- No DNN-based speech enhancement algorithm exists that maximize STOI.

Contribution

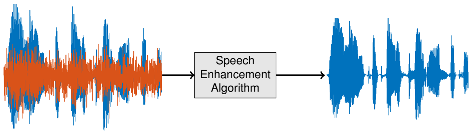
- We propose such an algorithm [3,4].

[3] M. Kolbæk, *et al.*, *IEEE ICASSP*, 2018
[4] M. Kolbæk, *et al.*, *IEEE TASLP*, 2018

On STOI Optimal DNN based Speech Enhancement
Motivation, Research Gap, and Contribution

Motivation

- Goal of speech enhancement algorithms is often to improve speech intelligibility (SI).
- Often these algorithms are designed to minimize the short-time spectral amplitude (STSA) mean-square error (MSE) with no link to SI.
- Can we use a function with a stronger link to SI? – e.g. the STOI SI estimator.



Mean-Square Error:

$$J_{MSE} = \frac{1}{K}\sum_{k=1}^{K}(a(k,m) - \hat{a}(k,m))^2$$

$a(k,m)$: Clean Speech STFT Amplitudes
$\hat{a}(k,m)$: Enhanced Clean Speech STFT Amplitudes

Research Gap

- No DNN-based speech enhancement algorithm exists that maximize STOI.

Contribution

- We propose such an algorithm [3,4].

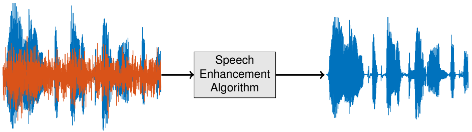
[3] M. Kolbæk, *et al.*, *IEEE ICASSP*, 2018
[4] M. Kolbæk, *et al.*, *IEEE TASLP*, 2018

On STOI Optimal DNN based Speech Enhancement

Motivation, Research Gap, and Contribution

Motivation

- Goal of speech enhancement algorithms is often to improve speech intelligibility (SI).
- Often these algorithms are designed to minimize the short-time spectral amplitude (STSA) mean-square error (MSE) with no link to SI.
- Can we use a function with a stronger link to SI? – e.g. the STOI SI estimator.



Mean-Square Error:

$$J_{MSE} = \frac{1}{K}\sum_{k=1}^{K}(a(k,m) - \hat{a}(k,m))^2$$

$a(k,m)$: Clean Speech STFT Amplitudes

$\hat{a}(k,m)$: Enhanced Clean Speech STFT Amplitudes

Research Gap

- No DNN-based speech enhancement algorithm exists that maximize STOI.

Contribution

- We propose such an algorithm [3,4].

[3] M. Kolbæk, *et al.*, *IEEE ICASSP*, 2018
[4] M. Kolbæk, *et al.*, *IEEE TASLP*, 2018

On STOI Optimal DNN based Speech Enhancement

Motivation, Research Gap, and Contribution

Motivation

- Goal of speech enhancement algorithms is often to improve speech intelligibility (SI).
- Often these algorithms are designed to minimize the short-time spectral amplitude (STSA) mean-square error (MSE) with no link to SI.
- Can we use a function with a stronger link to SI? – e.g. the STOI SI estimator.

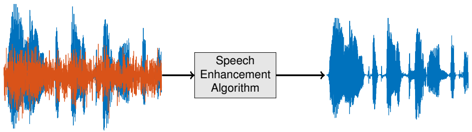
Research Gap

- No DNN-based speech enhancement algorithm exists that maximize STOI.

Contribution

- We propose such an algorithm [3,4].

[3] M. Kolbæk, *et al.*, *IEEE ICASSP*, 2018
[4] M. Kolbæk, *et al.*, *IEEE TASLP*, 2018



Mean-Square Error:

$$J_{MSE} = \frac{1}{K}\sum_{k=1}^{K}(a(k,m) - \hat{a}(k,m))^2$$

$a(k,m)$: Clean Speech STFT Amplitudes

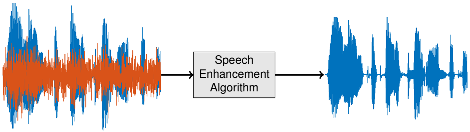
$\hat{a}(k,m)$: Enhanced Clean Speech STFT Amplitudes

On STOI Optimal DNN based Speech Enhancement

Motivation, Research Gap, and Contribution

Motivation

- Goal of speech enhancement algorithms is often to improve speech intelligibility (SI).
- Often these algorithms are designed to minimize the short-time spectral amplitude (STSA) mean-square error (MSE) with no link to SI.
- Can we use a function with a stronger link to SI? – e.g. the STOI SI estimator.



Mean-Square Error:

$$J_{MSE} = \frac{1}{K}\sum_{k=1}^{K}(a(k,m) - \hat{a}(k,m))^2$$

$a(k,m)$: Clean Speech STFT Amplitudes

$\hat{a}(k,m)$: Enhanced Clean Speech STFT Amplitudes

Research Gap

- No DNN-based speech enhancement algorithm exists that maximize STOI.

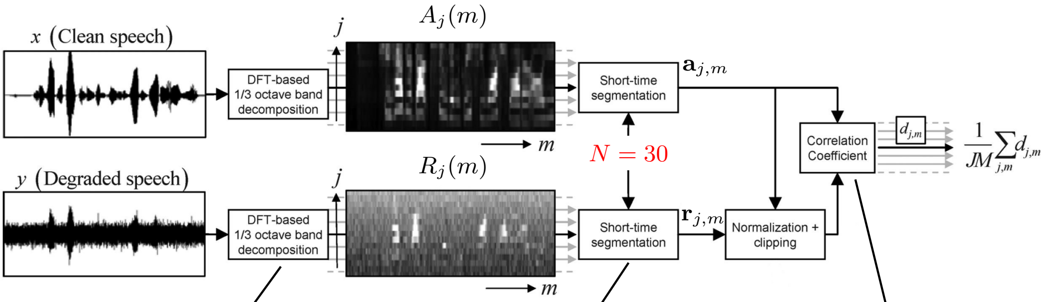
Contribution

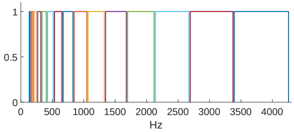
- We propose such an algorithm [3,4].

[3] M. Kolbæk, *et al.*, *IEEE ICASSP*, 2018
[4] M. Kolbæk, *et al.*, *IEEE TASLP*, 2018

On STOI Optimal DNN based Speech Enhancement

Short-Time Objective Intelligibility (STOI) - Architecture





$$\mathbf{a}_{j,m} = [A_j(m-N+1), \ldots, A_j(m)]^T$$

$$\mathbf{r}_{j,m} = [R_j(m-N+1), \ldots, R_j(m)]^T$$

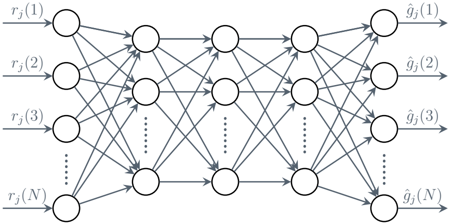
$$\frac{\left(\mathbf{a}_{j,m} - \mu_{\mathbf{a}_{j,m}}\right)^T \left(\hat{\mathbf{a}}_{j,m} - \mu_{\hat{\mathbf{a}}_{j,m}}\right)}{\left\|\mathbf{a}_{j,m} - \mu_{\mathbf{a}_{j,m}}\right\| \left\|\hat{\mathbf{a}}_{j,m} - \mu_{\hat{\mathbf{a}}_{j,m}}\right\|}$$

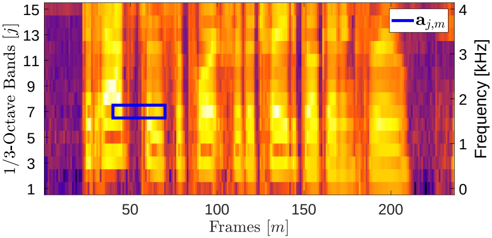
Figure reprinted from C. H. Taal *et al.*, 2011

On STOI Optimal DNN based Speech Enhancement

Proposed STOI-based Approach

- **STOI**-based Speech Enhancement Model





$$\hat{\mathbf{a}}_{j,m} = \hat{\mathbf{g}}_{j,m} \circ \mathbf{r}_{j,m}$$

$\hat{\mathbf{g}}_{j,m}$: Estimated Gains

$\mathbf{r}_{j,m}$: Noisy Speech 1/3-Octave band

$\mathbf{a}_{j,m}$: Est. Clean Speech 1/3-Octave band

$$\mathcal{L}_{ELC} = \frac{\left(\mathbf{a}_{j,m} - \mu_{\mathbf{a}_{j,m}}\right)^T \left(\hat{\mathbf{a}}_{j,m} - \mu_{\hat{\mathbf{a}}_{j,m}}\right)}{\left\|\mathbf{a}_{j,m} - \mu_{\mathbf{a}_{j,m}}\right\| \left\|\hat{\mathbf{a}}_{j,m} - \mu_{\hat{\mathbf{a}}_{j,m}}\right\|}$$

$$\mathcal{L}_{EMSE} = \frac{1}{N} \left\|\mathbf{a}_{j,m} - \hat{\mathbf{a}}_{j,m}\right\|^2$$

$\mathcal{L}$: Loss for sample m in band j

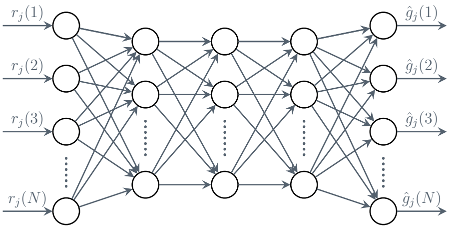
ELC : Envelope Linear Correlation

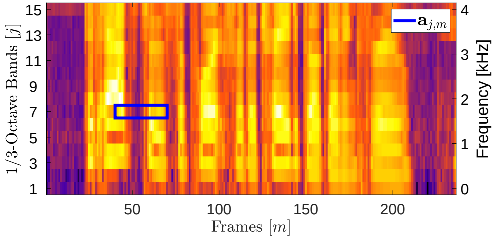
$EMSE$: Envelope Mean-Square Error

On STOI Optimal DNN based Speech Enhancement

Proposed STOI-based Approach

- **STOI**-based Speech Enhancement Model





$$\hat{\mathbf{a}}_{j,m} = \hat{\mathbf{g}}_{j,m} \circ \mathbf{r}_{j,m}$$

$\hat{\mathbf{g}}_{j,m}$: Estimated Gains

$\mathbf{r}_{j,m}$: Noisy Speech 1/3-Octave band

$\mathbf{a}_{j,m}$: Est. Clean Speech 1/3-Octave band

$$\mathcal{L}_{ELC} = \frac{\left(\mathbf{a}_{j,m} - \mu_{\mathbf{a}_{j,m}}\right)^T \left(\hat{\mathbf{a}}_{j,m} - \mu_{\hat{\mathbf{a}}_{j,m}}\right)}{\left\|\mathbf{a}_{j,m} - \mu_{\mathbf{a}_{j,m}}\right\| \left\|\hat{\mathbf{a}}_{j,m} - \mu_{\hat{\mathbf{a}}_{j,m}}\right\|}$$

$$\mathcal{L}_{EMSE} = \frac{1}{N} \left\|\mathbf{a}_{j,m} - \hat{\mathbf{a}}_{j,m}\right\|^2$$

$\mathcal{L}$: Loss for sample m in band j

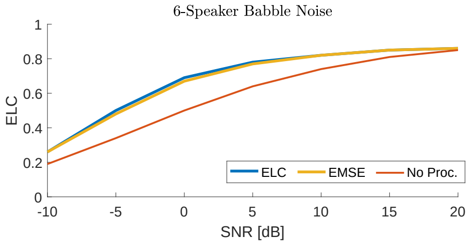
ELC : Envelope Linear Correlation

$EMSE$: Envelope Mean-Square Error

On STOI Optimal DNN based Speech Enhancement

Experimental Results

- DNNs designed to maximize approximate-STOI, improves ELC at various SNRs (and noise types).



- Similar conclusions can be drawn for DNNs that minimize EMSE.
- Same conclusions hold when the same DNNs are evaluated using STOI.
- Apparently, nothing to gain in terms of STOI, when maximizing ELC compared to minimizing MSE.

New Hypothesis

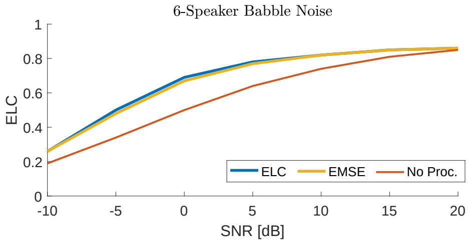
- Are the solutions in fact the same?

On STOI Optimal DNN based Speech Enhancement

Experimental Results

Experimental Results

- DNNs designed to maximize approximate-STOI, improves ELC at various SNRs (and noise types).
- Similar conclusions can be drawn for DNNs that minimize EMSE.
- Same conclusions hold when the same DNNs are evaluated using STOI.
- Apparently, nothing to gain in terms of STOI, when maximizing ELC compared to minimizing MSE.



New Hypothesis

- Are the solutions in fact the same?

On STOI Optimal DNN based Speech Enhancement

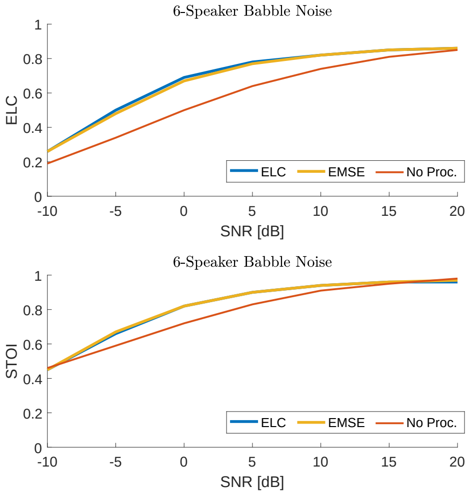
Experimental Results

Experimental Results

- DNNs designed to maximize approximate-STOI, improves ELC at various SNRs (and noise types).
- Similar conclusions can be drawn for DNNs that minimize EMSE.
- Same conclusions hold when the same DNNs are evaluated using STOI.
- Apparently, nothing to gain in terms of STOI, when maximizing ELC compared to minimizing MSE.

New Hypothesis

- Are the solutions in fact the same?



On STOI Optimal DNN based Speech Enhancement

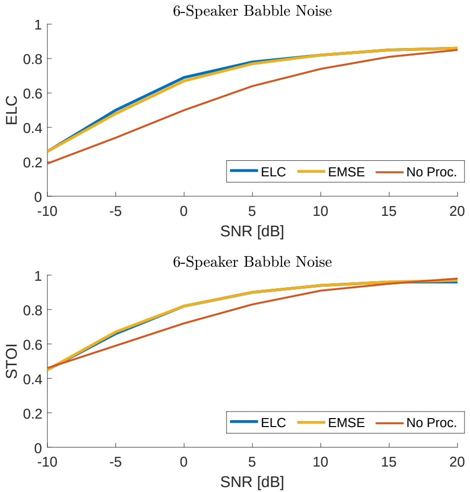
Experimental Results

Experimental Results

- DNNs designed to maximize approximate-STOI, improves ELC at various SNRs (and noise types).
- Similar conclusions can be drawn for DNNs that minimize EMSE.
- Same conclusions hold when the same DNNs are evaluated using STOI.
- Apparently, nothing to gain in terms of STOI, when maximizing ELC compared to minimizing MSE.

New Hypothesis

- Are the solutions in fact the same?



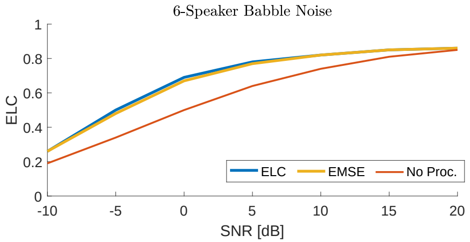
On STOI Optimal DNN based Speech Enhancement
Experimental Results

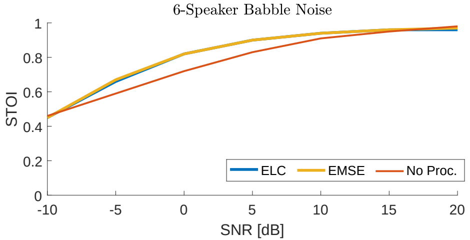
Experimental Results

- DNNs designed to maximize approximate-STOI, improves ELC at various SNRs (and noise types).
- Similar conclusions can be drawn for DNNs that minimize EMSE.
- Same conclusions hold when the same DNNs are evaluated using STOI.
- Apparently, nothing to gain in terms of STOI, when maximizing ELC compared to minimizing MSE.

New Hypothesis

- Are the solutions in fact the same?





On STOI Optimal DNN based Speech Enhancement
Theoretical Results and Conclusion

Method

- Using Bayesian statistics we derive the maximum mean ELC (MMELC) estimator.

$$\hat{\underline{a}}_{MMELC} = \arg\max_{\hat{\underline{a}}} \int \mathcal{L}_{ELC}\left(\underline{a}, \hat{\underline{a}}\right) f_{\underline{A}|\underline{R}}\left(\underline{a}|\underline{r}\right)\, d\underline{a} = \frac{\mathbb{E}_{\underline{A}|\underline{r}}\left[\underline{e}(\underline{A}|\underline{r})\right]}{\|\mathbb{E}_{\underline{A}|\underline{r}}\left[\underline{e}(\underline{A}|\underline{r})\right]\|}$$

Result

- We show, under certain general conditions, that the MMELC estimator is asymptotically (in N) equivalent to the classical STSA-MMSE estimator.

$$\hat{\underline{a}}_{MMSE} = \arg\min_{\hat{\underline{a}}} \int \left(\underline{a} - \hat{\underline{a}}\right)^2 f_{\underline{A}|\underline{R}}\left(\underline{a}|\underline{r}\right)\, d\underline{a} = \mathbb{E}_{\underline{A}|\underline{r}}\left[\underline{A}|\underline{r}\right]$$

Conclusion

- The STSA-MMSE estimator leads to the same approximate-STOI value as the MMELC estimator.
- For practical DNN based speech enhancement algorithms this is valid already at $N > 15$.
- No reason to optimize for ELC if the goal is to perform optimally w.r.t. STOI. STSA-MSE is near optimal.

$$\lim_{N \to \infty} \hat{\underline{a}}_{MMELC} = \hat{\underline{a}}_{MMSE} - \mu_{\hat{\underline{a}}_{MMSE}}$$

On STOI Optimal DNN based Speech Enhancement

Theoretical Results and Conclusion

Method

- Using Bayesian statistics we derive the maximum mean ELC (MMELC) estimator.

$$\hat{\underline{a}}_{MMELC} = \arg\max_{\hat{\underline{a}}} \int \mathcal{L}_{ELC}\left(\underline{a}, \hat{\underline{a}}\right) f_{\underline{A}|\underline{R}}\left(\underline{a}|\underline{r}\right)\, d\underline{a}$$

$$= \frac{\mathbb{E}_{\underline{A}|\underline{r}}\left[\underline{e}(\underline{A}|\underline{r})\right]}{\|\mathbb{E}_{\underline{A}|\underline{r}}\left[\underline{e}(\underline{A}|\underline{r})\right]\|}$$

Result

- We show, under certain general conditions, that the MMELC estimator is asymptotically (in N) equivalent to the classical STSA-MMSE estimator.

$$\hat{\underline{a}}_{MMSE} = \arg\min_{\hat{\underline{a}}} \int \left(\underline{a} - \hat{\underline{a}}\right)^2 f_{\underline{A}|\underline{R}}\left(\underline{a}|\underline{r}\right)\, d\underline{a}$$

$$= \mathbb{E}_{\underline{A}|\underline{r}}\left[\underline{A}|\underline{r}\right]$$

Conclusion

- The STSA-MMSE estimator leads to the same approximate-STOI value as the MMELC estimator.
- For practical DNN based speech enhancement algorithms this is valid already at $N > 15$.
- No reason to optimize for ELC if the goal is to perform optimally w.r.t. STOI. STSA-MSE is near optimal.

$$\lim_{N\to\infty} \hat{\underline{a}}_{MMELC} = \hat{\underline{a}}_{MMSE} - \mu_{\hat{\underline{a}}_{MMSE}}$$

On STOI Optimal DNN based Speech Enhancement

Theoretical Results and Conclusion

Method

- Using Bayesian statistics we derive the maximum mean ELC (MMELC) estimator.

$$\hat{\underline{a}}_{MMELC} = \arg\max_{\hat{\underline{a}}} \int \mathcal{L}_{ELC}\left(\underline{a}, \hat{\underline{a}}\right) f_{\underline{A}|\underline{R}}\left(\underline{a}|\underline{r}\right)\, d\underline{a}$$

$$= \frac{\mathbb{E}_{\underline{A}|\underline{r}}\left[\underline{e}(\underline{A}|\underline{r})\right]}{\|\mathbb{E}_{\underline{A}|\underline{r}}\left[\underline{e}(\underline{A}|\underline{r})\right]\|}$$

Result

- We show, under certain general conditions, that the MMELC estimator is asymptotically (in N) equivalent to the classical STSA-MMSE estimator.

$$\hat{\underline{a}}_{MMSE} = \arg\min_{\hat{\underline{a}}} \int \left(\underline{a} - \hat{\underline{a}}\right)^2 f_{\underline{A}|\underline{R}}\left(\underline{a}|\underline{r}\right)\, d\underline{a}$$

$$= \mathbb{E}_{\underline{A}|\underline{r}}\left[\underline{A}|\underline{r}\right]$$

Conclusion

- The STSA-MMSE estimator leads to the same approximate-STOI value as the MMELC estimator.
- For practical DNN based speech enhancement algorithms this is valid already at $N > 15$.
- No reason to optimize for ELC if the goal is to perform optimally w.r.t. STOI. STSA-MSE is near optimal.

$$\lim_{N\to\infty} \hat{\underline{a}}_{MMELC} = \hat{\underline{a}}_{MMSE} - \mu_{\hat{\underline{a}}_{MMSE}}$$

On STOI Optimal DNN based Speech Enhancement

Theoretical Results and Conclusion

Method

- Using Bayesian statistics we derive the maximum mean ELC (MMELC) estimator.

$$\hat{\underline{a}}_{MMELC} = \arg\max_{\hat{\underline{a}}} \int \mathcal{L}_{ELC}\left(\underline{a}, \hat{\underline{a}}\right) f_{\underline{A}|\underline{R}}\left(\underline{a}|\underline{r}\right)\, d\underline{a}$$

$$= \frac{\mathbb{E}_{\underline{A}|\underline{r}}\left[\underline{e}(\underline{A}|\underline{r})\right]}{\|\mathbb{E}_{\underline{A}|\underline{r}}\left[\underline{e}(\underline{A}|\underline{r})\right]\|}$$

Result

- We show, under certain general conditions, that the MMELC estimator is asymptotically (in N) equivalent to the classical STSA-MMSE estimator.

$$\hat{\underline{a}}_{MMSE} = \arg\min_{\hat{\underline{a}}} \int \left(\underline{a} - \hat{\underline{a}}\right)^2 f_{\underline{A}|\underline{R}}\left(\underline{a}|\underline{r}\right)\, d\underline{a}$$

$$= \mathbb{E}_{\underline{A}|\underline{r}}\left[\underline{A}|\underline{r}\right]$$

Conclusion

- The STSA-MMSE estimator leads to the same approximate-STOI value as the MMELC estimator.
- For practical DNN based speech enhancement algorithms this is valid already at $N > 15$.
- No reason to optimize for ELC if the goal is to perform optimally w.r.t. STOI. STSA-MSE is near optimal.

$$\lim_{N\to\infty} \hat{\underline{a}}_{MMELC} = \hat{\underline{a}}_{MMSE} - \mu_{\hat{\underline{a}}_{MMSE}}$$

On STOI Optimal DNN based Speech Enhancement

Theoretical Results and Conclusion

Method

- Using Bayesian statistics we derive the maximum mean ELC (MMELC) estimator.

$$\hat{\underline{a}}_{MMELC} = \arg\max_{\hat{\underline{a}}} \int \mathcal{L}_{ELC}\left(\underline{a}, \hat{\underline{a}}\right) f_{\underline{A}|\underline{R}}\left(\underline{a}|\underline{r}\right)\, d\underline{a}$$

$$= \frac{\mathbb{E}_{\underline{A}|\underline{r}}\left[\underline{e}(\underline{A}|\underline{r})\right]}{||\mathbb{E}_{\underline{A}|\underline{r}}\left[\underline{e}(\underline{A}|\underline{r})\right]||}$$

Result

- We show, under certain general conditions, that the MMELC estimator is asymptotically (in N) equivalent to the classical STSA-MMSE estimator.

$$\hat{\underline{a}}_{MMSE} = \arg\min_{\hat{\underline{a}}} \int \left(\underline{a} - \hat{\underline{a}}\right)^2 f_{\underline{A}|\underline{R}}\left(\underline{a}|\underline{r}\right)\, d\underline{a}$$

$$= \mathbb{E}_{\underline{A}|\underline{r}}\left[\underline{A}|\underline{r}\right]$$

Conclusion

- The STSA-MMSE estimator leads to the same approximate-STOI value as the MMELC estimator.
- For practical DNN based speech enhancement algorithms this is valid already at $N > 15$.
- No reason to optimize for ELC if the goal is to perform optimally w.r.t. STOI. STSA-MSE is near optimal.

$$\lim_{N\to\infty} \hat{\underline{a}}_{MMELC} = \hat{\underline{a}}_{MMSE} - \mu_{\hat{\underline{a}}_{MMSE}}$$

Permutation Invariant Training for Speech Separation

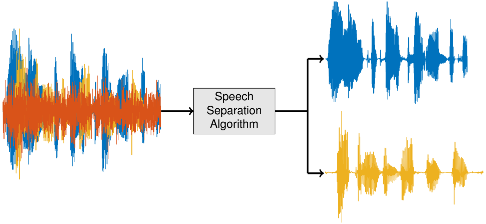
Permutation Invariant Training

Permutation Invariant Training for Speech Separation

Motivation, Research Gap, and Contribution

Motivation

- Speech separation algorithms are useful for various applications.
- E.g. "Cocktail party" situations.
- Existing solutions are complicated or limited.



Research Gap

- No DNN-only solution exists for speaker independent multi-talker speech separation.

Contribution

- We propose such algorithms [5,6,7].

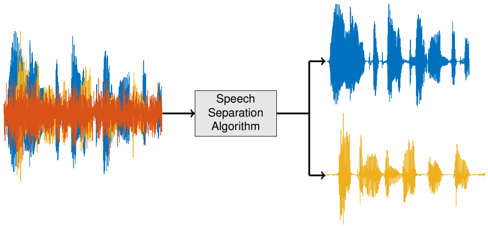
[5] D. Yu, *et al.*, *IEEE ICASSP*, 2017
[6] M. Kolbæk, *et al.*, *IEEE TASLP*, 2017
[7] M. Kolbæk, *et al.*, *IEEE MLSP*, 2017

Permutation Invariant Training for Speech Separation

Motivation, Research Gap, and Contribution

Motivation

- Speech separation algorithms are useful for various applications.
- E.g. "Cocktail party" situations.
- Existing solutions are complicated or limited.



Research Gap

- No DNN-only solution exists for speaker independent multi-talker speech separation.

Contribution

- We propose such algorithms [5,6,7].

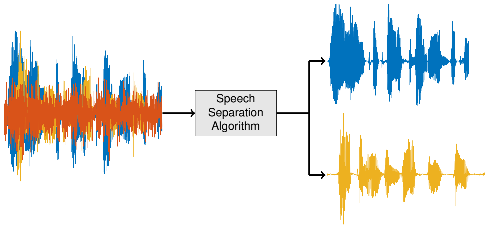
[5] D. Yu, *et al.*, *IEEE ICASSP*, 2017
[6] M. Kolbæk, *et al.*, *IEEE TASLP*, 2017
[7] M. Kolbæk, *et al.*, *IEEE MLSP*, 2017

Permutation Invariant Training for Speech Separation

Motivation, Research Gap, and Contribution

Motivation

- Speech separation algorithms are useful for various applications.
- E.g. "Cocktail party" situations.
- Existing solutions are complicated or limited.



Research Gap

- No DNN-only solution exists for speaker independent multi-talker speech separation.

Contribution

- We propose such algorithms [5,6,7].

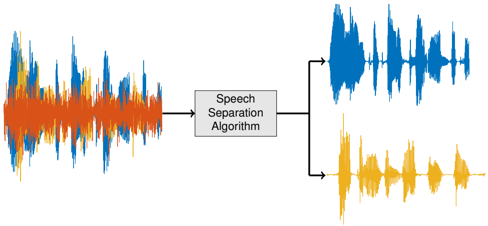
[5] D. Yu, *et al.*, *IEEE ICASSP*, 2017
[6] M. Kolbæk, *et al.*, *IEEE TASLP*, 2017
[7] M. Kolbæk, *et al.*, *IEEE MLSP*, 2017

Permutation Invariant Training for Speech Separation

Motivation, Research Gap, and Contribution

Motivation

- Speech separation algorithms are useful for various applications.
- E.g. "Cocktail party" situations.
- Existing solutions are complicated or limited.



Research Gap

- No DNN-only solution exists for speaker independent multi-talker speech separation.

Contribution

- We propose such algorithms [5,6,7].

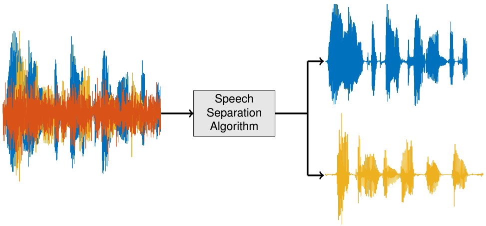
[5] D. Yu, *et al.*, *IEEE ICASSP*, 2017
[6] M. Kolbæk, *et al.*, *IEEE TASLP*, 2017
[7] M. Kolbæk, *et al.*, *IEEE MLSP*, 2017

Permutation Invariant Training for Speech Separation

Motivation, Research Gap, and Contribution

Motivation

- Speech separation algorithms are useful for various applications.
- E.g. "Cocktail party" situations.
- Existing solutions are complicated or limited.



Research Gap

- No DNN-only solution exists for speaker independent multi-talker speech separation.

Contribution

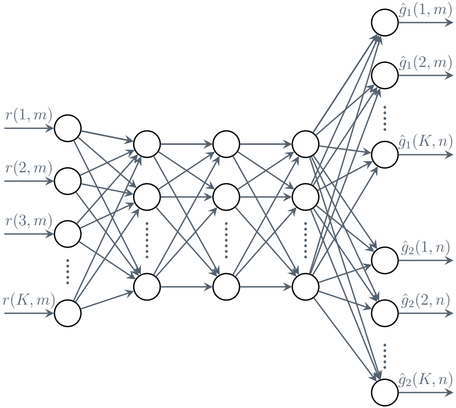
- We propose such algorithms [5,6,7].

[5] D. Yu, *et al.*, *IEEE ICASSP*, 2017
[6] M. Kolbæk, *et al.*, *IEEE TASLP*, 2017
[7] M. Kolbæk, *et al.*, *IEEE MLSP*, 2017

Permutation Invariant Training for Speech Separation

Label Permutation Problem

- 2-Speaker Separation Model ($S = 2$)



- MSE Cost Function

$$J_{MSE} = \frac{1}{SK}\sum_{s=1}^{S}\sum_{k=1}^{K}(a_s(k,m) - \hat{g}_s(k,m)r(k,m))^2$$

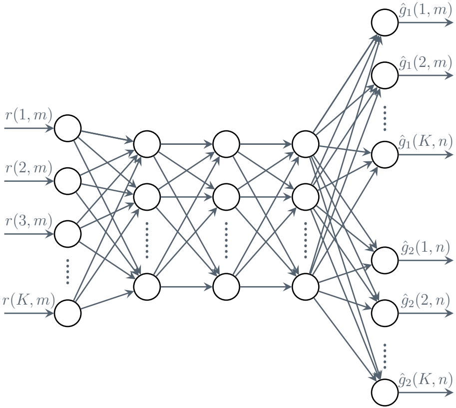
$$= \frac{1}{SK}\sum_{s=1}^{S}\sum_{k=1}^{K}(a_s(k,m) - \hat{a}_s(k,m))^2$$

- Training Progress for Speaker "Independent" Data

Permutation Invariant Training for Speech Separation

Label Permutation Problem

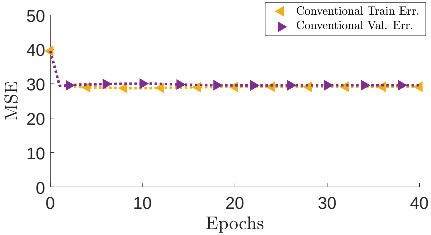
- 2-Speaker Separation Model ($S = 2$)



- MSE Cost Function

$$J_{MSE} = \frac{1}{SK}\sum_{s=1}^{S}\sum_{k=1}^{K}(a_s(k,m) - \hat{g}_s(k,m)r(k,m))^2$$
$$= \frac{1}{SK}\sum_{s=1}^{S}\sum_{k=1}^{K}(a_s(k,m) - \hat{a}_s(k,m))^2$$

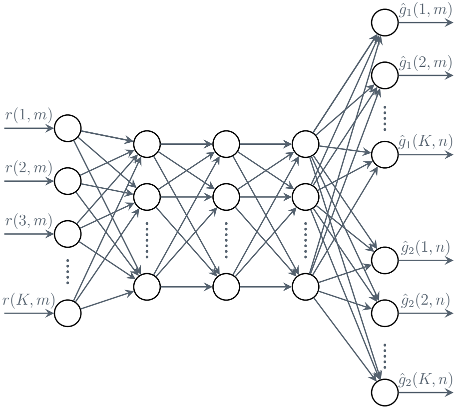
- Training Progress for Speaker "Independent" Data



Permutation Invariant Training for Speech Separation

Label Permutation Problem

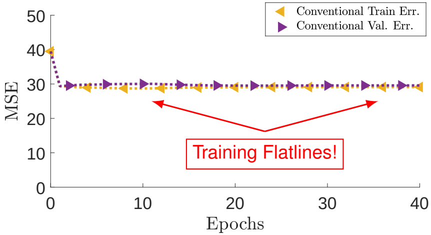
- 2-Speaker Separation Model ($S = 2$)



- MSE Cost Function

$$J_{MSE} = \frac{1}{SK}\sum_{s=1}^{S}\sum_{k=1}^{K}(a_s(k,m) - \hat{g}_s(k,m)r(k,m))^2$$
$$= \frac{1}{SK}\sum_{s=1}^{S}\sum_{k=1}^{K}(a_s(k,m) - \hat{a}_s(k,m))^2$$

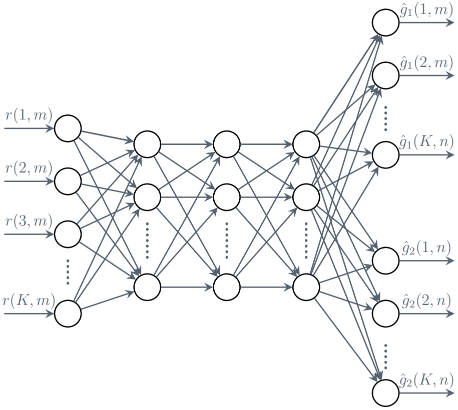
- Training Progress for Speaker "Independent" Data



Permutation Invariant Training for Speech Separation

Label Permutation Problem

- 2-Speaker Separation Model ($S = 2$)



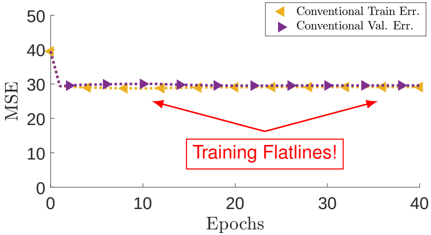
- MSE Cost Function

$$J_{MSE} = \frac{1}{SK}\sum_{s=1}^{S}\sum_{k=1}^{K}(a_s(k,m) - \hat{g}_s(k,m)r(k,m))^2$$

$$= \frac{1}{SK}\sum_{s=1}^{S}\sum_{k=1}^{K}(a_s(k,m) - \hat{a}_s(k,m))^2$$

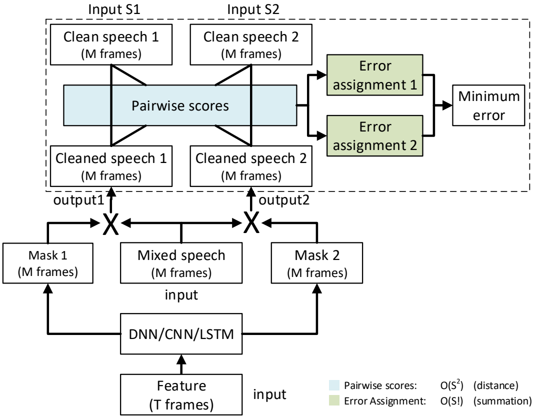
Permutation problem!

- Training Progress for Speaker "Independent" Data



Permutation Invariant Training for Speech Separation
Frame-level Permutation Invariant Training

- 2-Speaker Frame-level PIT Technique



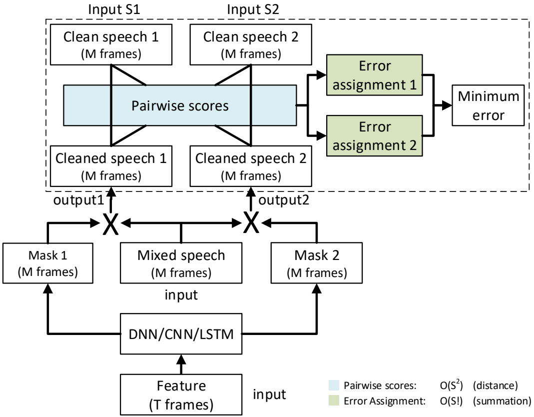
- PIT MSE Cost Function

$$J_{PIT} = \min_{\theta \in \mathcal{P}} \frac{1}{SK} \sum_{s=1}^{S} \sum_{k=1}^{K} (a_s(k,m) - \hat{a}_{\theta(s)}(k,m))^2$$

- PIT Training Progress (SGD)

Permutation Invariant Training for Speech Separation
Frame-level Permutation Invariant Training

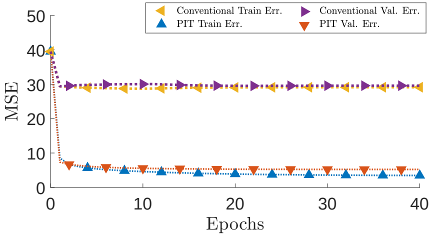
- 2-Speaker Frame-level PIT Technique



- PIT MSE Cost Function

$$J_{PIT} = \min_{\theta \in \mathcal{P}} \frac{1}{SK} \sum_{s=1}^{S} \sum_{k=1}^{K} (a_s(k,m) - \hat{a}_{\theta(s)}(k,m))^2$$

- PIT Training Progress (SGD)



Permutation Invariant Training for Speech Separation

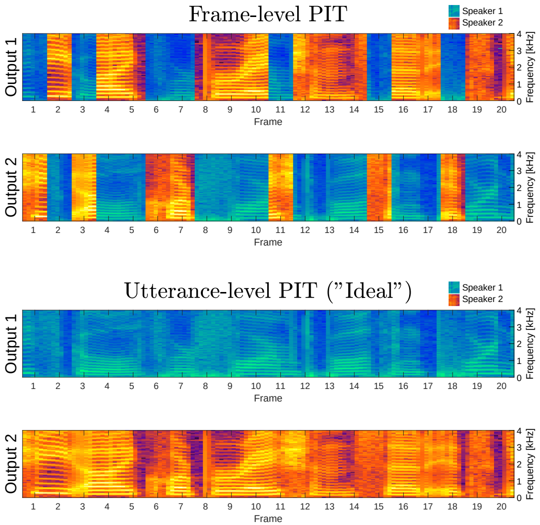
Utterance-level Permutation Invariant Training

- **Problem:** With Frame-level PIT permutation is unknown during inference.
- **Solution:** Train with permutation corresponding to minimum utterance-level error (*for all* m).

$$\theta^* = \underset{\theta \in \mathcal{P}}{\text{argmin}} \frac{1}{SMK} \sum_{s=1}^{S} \sum_{m=1}^{M} \sum_{k=1}^{K} (a_s(k,m) - \hat{a}_{\theta(s)}(k,m))^2$$

$$J_{uPIT} = \frac{1}{SK} \sum_{s=1}^{S} \sum_{k=1}^{K} (a_s(k,m) - \hat{a}_{\theta^*(s)}(k,m))^2$$

- Utterance-level PIT minimizes the utterance-level error, hence reducing context switch.
- **Note:** No extra computations during inference.



Permutation Invariant Training for Speech Separation

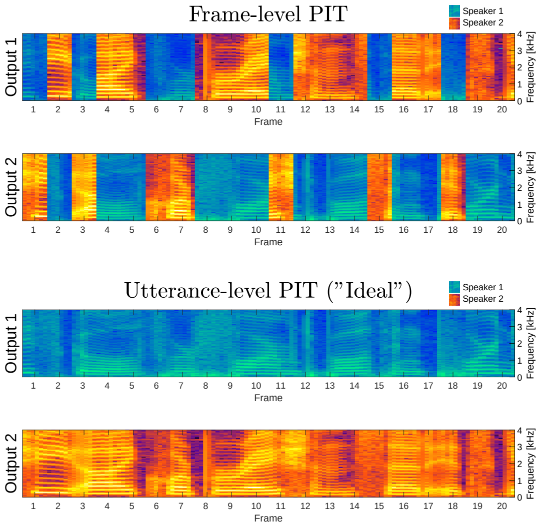
Utterance-level Permutation Invariant Training

- **Problem:** With Frame-level PIT permutation is unknown during inference.
- **Solution:** Train with permutation corresponding to minimum utterance-level error (*for all* m).

$$\theta^* = \underset{\theta \in \mathcal{P}}{\operatorname{argmin}} \frac{1}{SMK} \sum_{s=1}^{S} \sum_{m=1}^{M} \sum_{k=1}^{K} \left(a_s(k,m) - \hat{a}_{\theta(s)}(k,m)\right)^2$$

$$J_{uPIT} = \frac{1}{SK} \sum_{s=1}^{S} \sum_{k=1}^{K} \left(a_s(k,m) - \hat{a}_{\theta^*(s)}(k,m)\right)^2$$

- Utterance-level PIT minimizes the utterance-level error, hence reducing context switch.
- **Note:** No extra computations during inference.



Permutation Invariant Training for Speech Separation

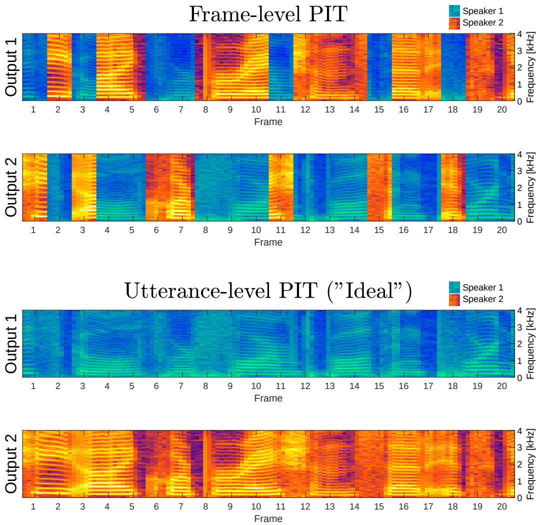
Utterance-level Permutation Invariant Training

- **Problem:** With Frame-level PIT permutation is unknown during inference.
- **Solution:** Train with permutation corresponding to minimum utterance-level error (*for all* m).

$$\theta^* = \underset{\theta \in \mathcal{P}}{\text{argmin}} \frac{1}{SMK} \sum_{s=1}^{S} \sum_{m=1}^{M} \sum_{k=1}^{K} (a_s(k,m) - \hat{a}_{\theta(s)}(k,m))^2$$

$$J_{uPIT} = \frac{1}{SK} \sum_{s=1}^{S} \sum_{k=1}^{K} (a_s(k,m) - \hat{a}_{\theta^*(s)}(k,m))^2$$

- Utterance-level PIT minimizes the utterance-level error, hence reducing context switch.
- **Note:** No extra computations during inference.



Permutation Invariant Training for Speech Separation

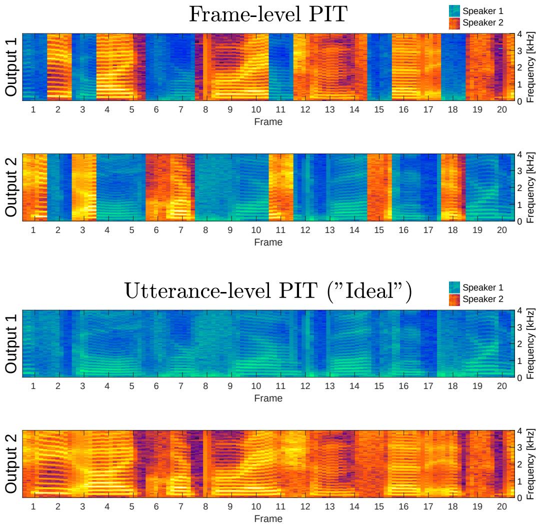
Utterance-level Permutation Invariant Training

- **Problem:** With Frame-level PIT permutation is unknown during inference.
- **Solution:** Train with permutation corresponding to minimum utterance-level error (*for all* m).

$$\theta^* = \underset{\theta \in \mathcal{P}}{\text{argmin}} \frac{1}{SMK} \sum_{s=1}^{S} \sum_{m=1}^{M} \sum_{k=1}^{K} \left(a_s(k,m) - \hat{a}_{\theta(s)}(k,m)\right)^2$$

$$J_{uPIT} = \frac{1}{SK} \sum_{s=1}^{S} \sum_{k=1}^{K} \left(a_s(k,m) - \hat{a}_{\theta^*(s)}(k,m)\right)^2$$

- Utterance-level PIT minimizes the utterance-level error, hence reducing context switch.
- **Note:** No extra computations during inference.



Permutation Invariant Training for Speech Separation
Results and Conclusion

Result

- State-of-the-art on 2-talker and 3-talker speaker-independent speech separation tasks.
- DNNs trained with uPIT works well for speech separation and enhancement jointly.
- More interestingly, works well without prior knowledge about the number of speakers.

Conclusion

- uPIT is a DNN training technique that enable DNN-only algorithms for **speaker-independent** multi-talker **speech separation** and **enhancement**.

Permutation Invariant Training for Speech Separation

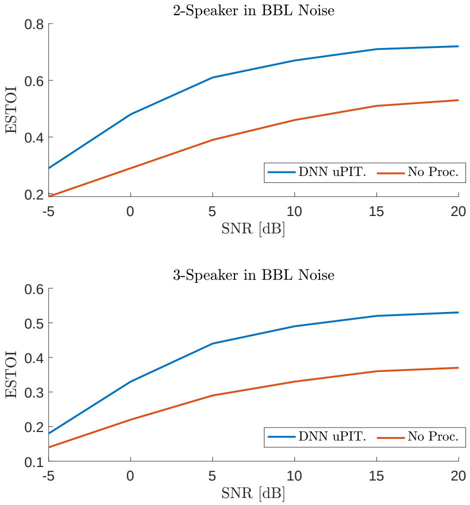
Results and Conclusion

Result

- State-of-the-art on 2-talker and 3-talker speaker-independent speech separation tasks.
- DNNs trained with uPIT works well for speech separation and enhancement jointly.
- More interestingly, works well without prior knowledge about the number of speakers.

Conclusion

- uPIT is a DNN training technique that enable DNN-only algorithms for **speaker-independent** multi-talker **speech separation** and **enhancement**.



Permutation Invariant Training for Speech Separation

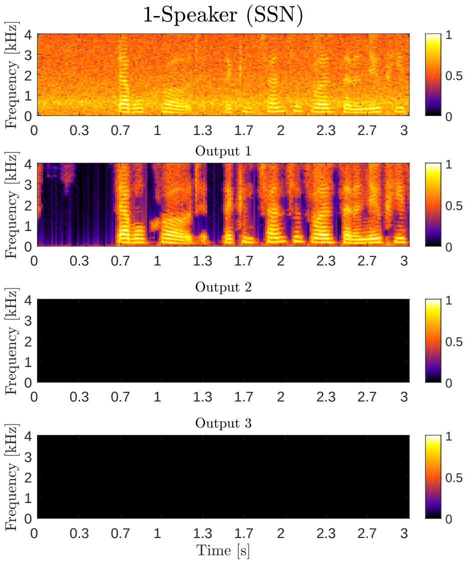
Results and Conclusion

Result

- State-of-the-art on 2-talker and 3-talker speaker-independent speech separation tasks.
- DNNs trained with uPIT works well for speech separation and enhancement jointly.
- More interestingly, works well without prior knowledge about the number of speakers.

Conclusion

- uPIT is a DNN training technique that enable DNN-only algorithms for **speaker-independent** multi-talker **speech separation** and **enhancement**.



Permutation Invariant Training for Speech Separation

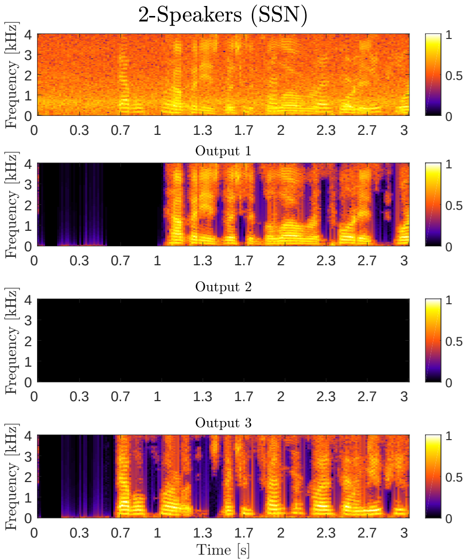
Results and Conclusion

Result

- State-of-the-art on 2-talker and 3-talker speaker-independent speech separation tasks.
- DNNs trained with uPIT works well for speech separation and enhancement jointly.
- More interestingly, works well without prior knowledge about the number of speakers.

Conclusion

- uPIT is a DNN training technique that enable DNN-only algorithms for **speaker-independent** multi-talker **speech separation** and **enhancement**.



Permutation Invariant Training for Speech Separation

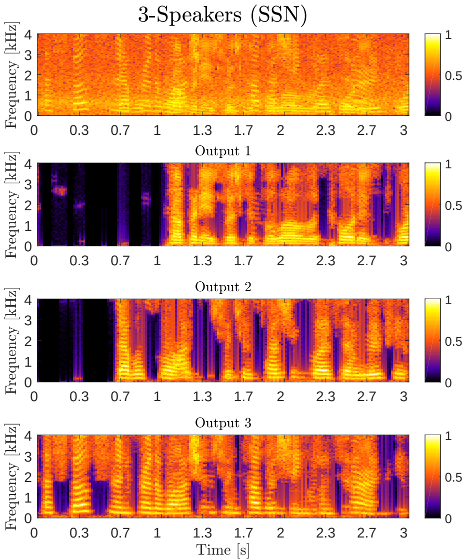
Results and Conclusion

Result

- State-of-the-art on 2-talker and 3-talker speaker-independent speech separation tasks.
- DNNs trained with uPIT works well for speech separation and enhancement jointly.
- More interestingly, works well without prior knowledge about the number of speakers.

Conclusion

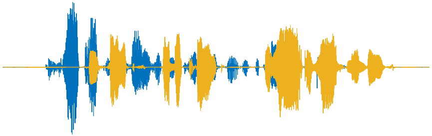
- uPIT is a DNN training technique that enable DNN-only algorithms for **speaker-independent** multi-talker **speech separation** and **enhancement**.



Permutation Invariant Training for Speech Separation

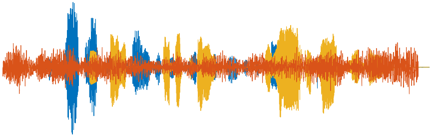
Demo - 2-Speaker Separation and Enhancement

Play Male + Female



The swap offer requires at least eighty percent of the total be tendered
He cites double-quote the law of large numbers

Play Male + Female + Noise



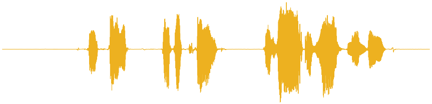
Play Separated Male

The swap offer requires at least eighty percent of the total be tendered

Play Separated and Enhanced Male

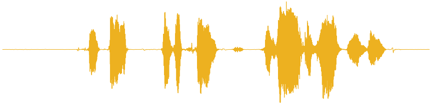
The swap offer requires at least eighty percent of the total be tendered

Play Separated Female



He cites double-quote the law of large numbers

Play Separated and Enhanced Female



He cites double-quote the law of large numbers

Summary and Conclusion

- Generalization of Deep Learning based Speech Enhancement
  - Human Receivers - Speech Intelligibility
  - Machine Receivers - Speaker Verification
- On STOI Optimal Deep Learning based Speech Enhancement
- Permutation Invariant Training for Deep Learning based Speech Separation
- Summary and Conclusion

Summary and Conclusion

Academic Output

Academic Output: 3 Journal papers and 4 Conference papers

[1] M. Kolbæk, Z. H. Tan, and J. Jensen, "Speech Intelligibility Potential of General and Specialized Deep Neural Network Based Speech Enhancement Systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 153–167, 2017.

[2] M. Kolbœk, Z. H. Tan, and J. Jensen, "Speech Enhancement using Long Short-Term Memory based Recurrent Neural Networks for Noise Robust Speaker Verification," in *Proc. SLT*, 2016, pp. 305–311.

[3] M. Kolbæk, Z.-H. Tan, and J. Jensen, "Monaural Speech Enhancement using Deep Neural Networks by Maximizing a Short-Time Objective Intelligibility Measure," in *Proc. ICASSP*, 2018, pp. 5059 – 5063.

[4] M. Kolbæk, Z. H. Tan, and J. Jensen, "On the Relationship Between Short-Time Objective Intelligibility and Short-Time Spectral-Amplitude Mean-Square Error for Speech Enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 283–295, 2018.

[5] D. Yu, M. Kolbæk, Z. H. Tan, and J. Jensen, "Permutation Invariant Training of Deep Models for Speaker-independent Multi-talker Speech Separation," in *Proc. ICASSP*, 2017, pp. 241–245.

[6] M. Kolbæk, D. Yu, Z. H. Tan, and J. Jensen, "Multi-talker Speech Separation With Utterance-Level Permutation Invariant Training of Deep Recurrent Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[7] M. Kolbæk, D. Yu, Z. H. Tan, and J. Jensen, "Joint separation and denoising of noisy multi-talker speech using recurrent neural networks and permutation invariant training," in *Proc. MLSP*, 2017, pp. 1–6.

Summary and Conclusion
Deep Learning based Speech Enhancement and Separation

Concluding Remarks

- **Generalizability** [1, 2]
  - Matching the noise type is the most critical, whereas matching the speaker and SNR is less critical if a modest amount of speakers are included in the training set.
  - A male-speaker "general" DNN based speech enhancement front-end achieves state-of-the-art performance on a speaker verification task.
- **Optimality** [3, 4]
  - The STSA-MMSE estimator is asymptotically equivalent to the MMELC estimator.
  - The STSA-MSE cost function leads to enhanced speech signals which are essentially optimal in terms of STOI. In other words, there is no benefit from optimizing for STOI.
- **Permutation Invariant Training** [5, 6, 7]
  - A training criterion that enable DNNs to work well on single-microphone speaker-independent multi-talker speech separation and enhancement.
  - Simple solution to the label permutation problem.
  - Achieves state-of-the-art performance.

Summary and Conclusion

Deep Learning based Speech Enhancement and Separation

Concluding Remarks

- **Generalizability** [1, 2]
  - Matching the noise type is the most critical, whereas matching the speaker and SNR is less critical if a modest amount of speakers are included in the training set.
  - A male-speaker "general" DNN based speech enhancement front-end achieves state-of-the-art performance on a speaker verification task.
- **Optimality** [3, 4]
  - The STSA-MMSE estimator is asymptotically equivalent to the MMELC estimator.
  - The STSA-MSE cost function leads to enhanced speech signals which are essentially optimal in terms of STOI. In other words, there is no benefit from optimizing for STOI.
- **Permutation Invariant Training** [5, 6, 7]
  - A training criterion that enable DNNs to work well on single-microphone speaker-independent multi-talker speech separation and enhancement.
  - Simple solution to the label permutation problem.
  - Achieves state-of-the-art performance.

Summary and Conclusion

Deep Learning based Speech Enhancement and Separation

Concluding Remarks

- **Generalizability** [1, 2]
  - Matching the noise type is the most critical, whereas matching the speaker and SNR is less critical if a modest amount of speakers are included in the training set.
  - A male-speaker "general" DNN based speech enhancement front-end achieves state-of-the-art performance on a speaker verification task.
- **Optimality** [3, 4]
  - The STSA-MMSE estimator is asymptotically equivalent to the MMELC estimator.
  - The STSA-MSE cost function leads to enhanced speech signals which are essentially optimal in terms of STOI. In other words, there is no benefit from optimizing for STOI.
- **Permutation Invariant Training** [5, 6, 7]
  - A training criterion that enable DNNs to work well on single-microphone speaker-independent multi-talker speech separation and enhancement.
  - Simple solution to the label permutation problem.
  - Achieves state-of-the-art performance.

Summary and Conclusion

Deep Learning based Speech Enhancement and Separation

Concluding Remarks

- **Generalizability** [1, 2]
  - Matching the noise type is the most critical, whereas matching the speaker and SNR is less critical if a modest amount of speakers are included in the training set.
  - A male-speaker "general" DNN based speech enhancement front-end achieves state-of-the-art performance on a speaker verification task.
- **Optimality** [3, 4]
  - The STSA-MMSE estimator is asymptotically equivalent to the MMELC estimator.
  - The STSA-MSE cost function leads to enhanced speech signals which are essentially optimal in terms of STOI. In other words, there is no benefit from optimizing for STOI.
- **Permutation Invariant Training** [5, 6, 7]
  - A training criterion that enable DNNs to work well on single-microphone speaker-independent multi-talker speech separation and enhancement.
  - Simple solution to the label permutation problem.
  - Achieves state-of-the-art performance.

Summary and Conclusion

Deep Learning based Speech Enhancement and Separation

Concluding Remarks

- **Generalizability** [1, 2]
  - Matching the noise type is the most critical, whereas matching the speaker and SNR is less critical if a modest amount of speakers are included in the training set.
  - A male-speaker "general" DNN based speech enhancement front-end achieves state-of-the-art performance on a speaker verification task.
- **Optimality** [3, 4]
  - The STSA-MMSE estimator is asymptotically equivalent to the MMELC estimator.
  - The STSA-MSE cost function leads to enhanced speech signals which are essentially optimal in terms of STOI. In other words, there is no benefit from optimizing for STOI.
- **Permutation Invariant Training** [5, 6, 7]
  - A training criterion that enable DNNs to work well on single-microphone speaker-independent multi-talker speech separation and enhancement.
  - Simple solution to the label permutation problem.
  - Achieves state-of-the-art performance.

Summary and Conclusion

Deep Learning based Speech Enhancement and Separation

Concluding Remarks

- **Generalizability** [1, 2]
  - Matching the noise type is the most critical, whereas matching the speaker and SNR is less critical if a modest amount of speakers are included in the training set.
  - A male-speaker "general" DNN based speech enhancement front-end achieves state-of-the-art performance on a speaker verification task.

- **Optimality** [3, 4]
  - The STSA-MMSE estimator is asymptotically equivalent to the MMELC estimator.
  - The STSA-MSE cost function leads to enhanced speech signals which are essentially optimal in terms of STOI. In other words, there is no benefit from optimizing for STOI.

- **Permutation Invariant Training** [5, 6, 7]
  - A training criterion that enable DNNs to work well on single-microphone speaker-independent multi-talker speech separation and enhancement.
  - Simple solution to the label permutation problem.
  - Achieves state-of-the-art performance.

Summary and Conclusion

Deep Learning based Speech Enhancement and Separation

Concluding Remarks

- **Generalizability** [1, 2]
  - Matching the noise type is the most critical, whereas matching the speaker and SNR is less critical if a modest amount of speakers are included in the training set.
  - A male-speaker "general" DNN based speech enhancement front-end achieves state-of-the-art performance on a speaker verification task.

- **Optimality** [3, 4]
  - The STSA-MMSE estimator is asymptotically equivalent to the MMELC estimator.
  - The STSA-MSE cost function leads to enhanced speech signals which are essentially optimal in terms of STOI. In other words, there is no benefit from optimizing for STOI.

- **Permutation Invariant Training** [5, 6, 7]
  - A training criterion that enable DNNs to work well on single-microphone speaker-independent multi-talker speech separation and enhancement.
  - Simple solution to the label permutation problem.
  - Achieves state-of-the-art performance.

Summary and Conclusion

Deep Learning based Speech Enhancement and Separation

Concluding Remarks

- **Generalizability** [1, 2]
  - Matching the noise type is the most critical, whereas matching the speaker and SNR is less critical if a modest amount of speakers are included in the training set.
  - A male-speaker "general" DNN based speech enhancement front-end achieves state-of-the-art performance on a speaker verification task.

- **Optimality** [3, 4]
  - The STSA-MMSE estimator is asymptotically equivalent to the MMELC estimator.
  - The STSA-MSE cost function leads to enhanced speech signals which are essentially optimal in terms of STOI. In other words, there is no benefit from optimizing for STOI.

- **Permutation Invariant Training** [5, 6, 7]
  - A training criterion that enable DNNs to work well on single-microphone speaker-independent multi-talker speech separation and enhancement.
  - Simple solution to the label permutation problem.
  - Achieves state-of-the-art performance.

Summary and Conclusion

Deep Learning based Speech Enhancement and Separation

Concluding Remarks

- **Generalizability** [1, 2]
  - Matching the noise type is the most critical, whereas matching the speaker and SNR is less critical if a modest amount of speakers are included in the training set.
  - A male-speaker "general" DNN based speech enhancement front-end achieves state-of-the-art performance on a speaker verification task.

- **Optimality** [3, 4]
  - The STSA-MMSE estimator is asymptotically equivalent to the MMELC estimator.
  - The STSA-MSE cost function leads to enhanced speech signals which are essentially optimal in terms of STOI. In other words, there is no benefit from optimizing for STOI.

- **Permutation Invariant Training** [5, 6, 7]
  - A training criterion that enable DNNs to work well on single-microphone speaker-independent multi-talker speech separation and enhancement.
  - Simple solution to the label permutation problem.
  - Achieves state-of-the-art performance.

Summary and Conclusion

Deep Learning based Speech Enhancement and Separation

Concluding Remarks

- **Generalizability** [1, 2]
  - Matching the noise type is the most critical, whereas matching the speaker and SNR is less critical if a modest amount of speakers are included in the training set.
  - A male-speaker "general" DNN based speech enhancement front-end achieves state-of-the-art performance on a speaker verification task.

- **Optimality** [3, 4]
  - The STSA-MMSE estimator is asymptotically equivalent to the MMELC estimator.
  - The STSA-MSE cost function leads to enhanced speech signals which are essentially optimal in terms of STOI. In other words, there is no benefit from optimizing for STOI.

- **Permutation Invariant Training** [5, 6, 7]
  - A training criterion that enable DNNs to work well on single-microphone speaker-independent multi-talker speech separation and enhancement.
  - Simple solution to the label permutation problem.
  - Achieves state-of-the-art performance.

Summary and Conclusion

Not there yet, but a small step closer.

Thank you.

AALBORG UNIVERSITY
DENMARK