BIO-MOTIVATED FEATURES AND DEEP LEARNING FOR ROBUST SPEECH RECOGNITION

Fernando de la Calle Silos

September 29, 2017

Department of Signal Theory and Communications
Universidad Carlos III de Madrid

Table of contents

Introduction

Robust Speech Recognition

- Speech Recognition systems are affected by external influences not related with what has been spoken.

Robust Speech Recognition

- Speech Recognition systems are affected by external influences not related with what has been spoken.
- Not robust as humans.

Robust Speech Recognition

- Speech Recognition systems are affected by external influences not related with what has been spoken.
- Not robust as humans.
- Recognition accuracy degrades:

Robust Speech Recognition

- Speech Recognition systems are affected by external influences not related with what has been spoken.
- Not robust as humans.
- Recognition accuracy degrades:
  - Additive noise.

Robust Speech Recognition

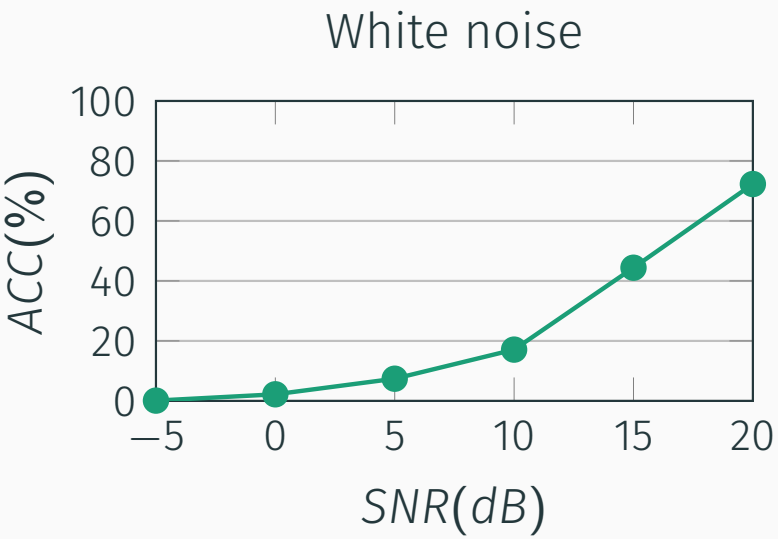
- Speech Recognition systems are affected by external influences not related with what has been spoken.
- Not robust as humans.
- Recognition accuracy degrades:
  - Additive noise.
  - Reverberant conditions.

Robust Speech Recognition

- Speech Recognition systems are affected by external influences not related with what has been spoken.
- Not robust as humans.
- Recognition accuracy degrades:
  - Additive noise.
  - Reverberant conditions.

Robust Speech Recognition

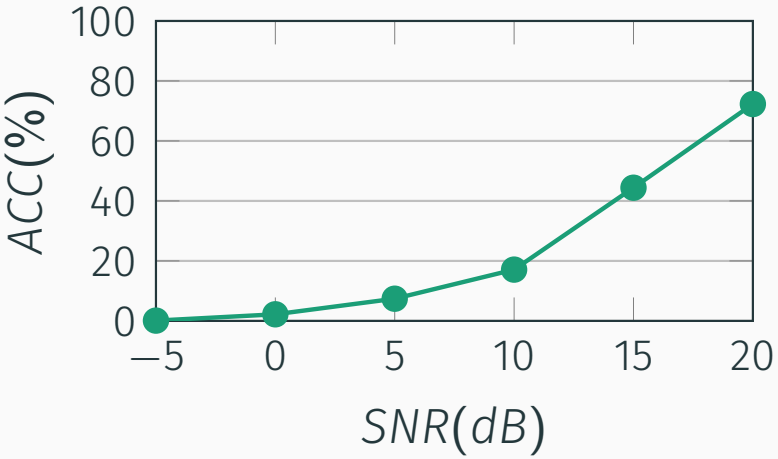
- Speech Recognition systems are affected by external influences not related with what has been spoken.
- Not robust as humans.
- Recognition accuracy degrades:
  - Additive noise.
  - Reverberant conditions.



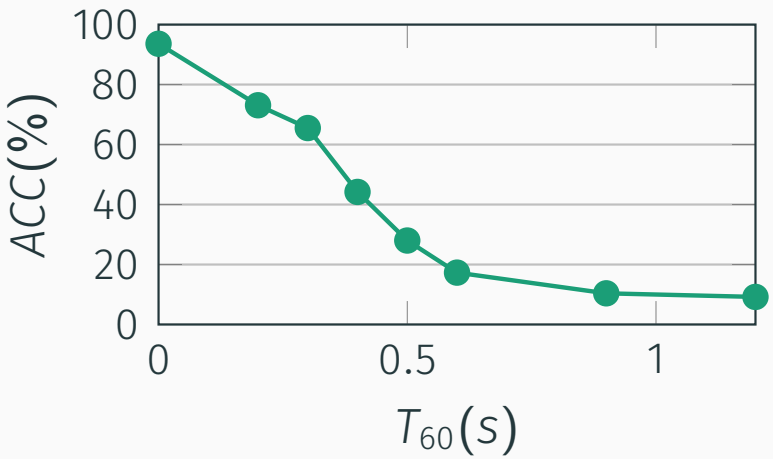
Robust Speech Recognition

- Speech Recognition systems are affected by external influences not related with what has been spoken.
- Not robust as humans.
- Recognition accuracy degrades:
  - Additive noise.
  - Reverberant conditions.

White noise



Reverberation



Techniques to improve the robustness

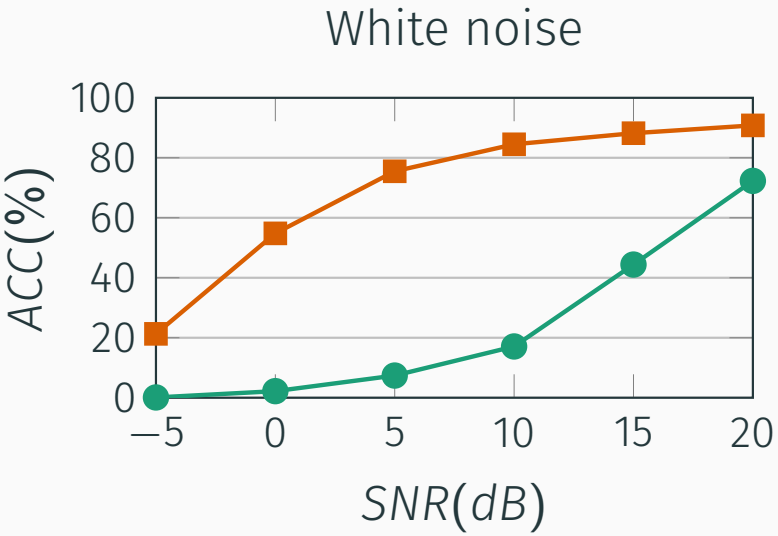
- Training with contaminated data.

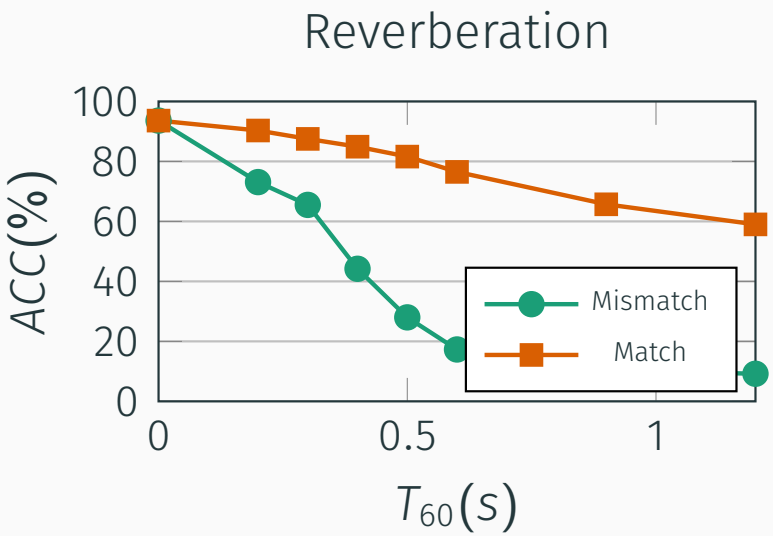
Techniques to improve the robustness

- Training with contaminated data.

Techniques to improve the robustness

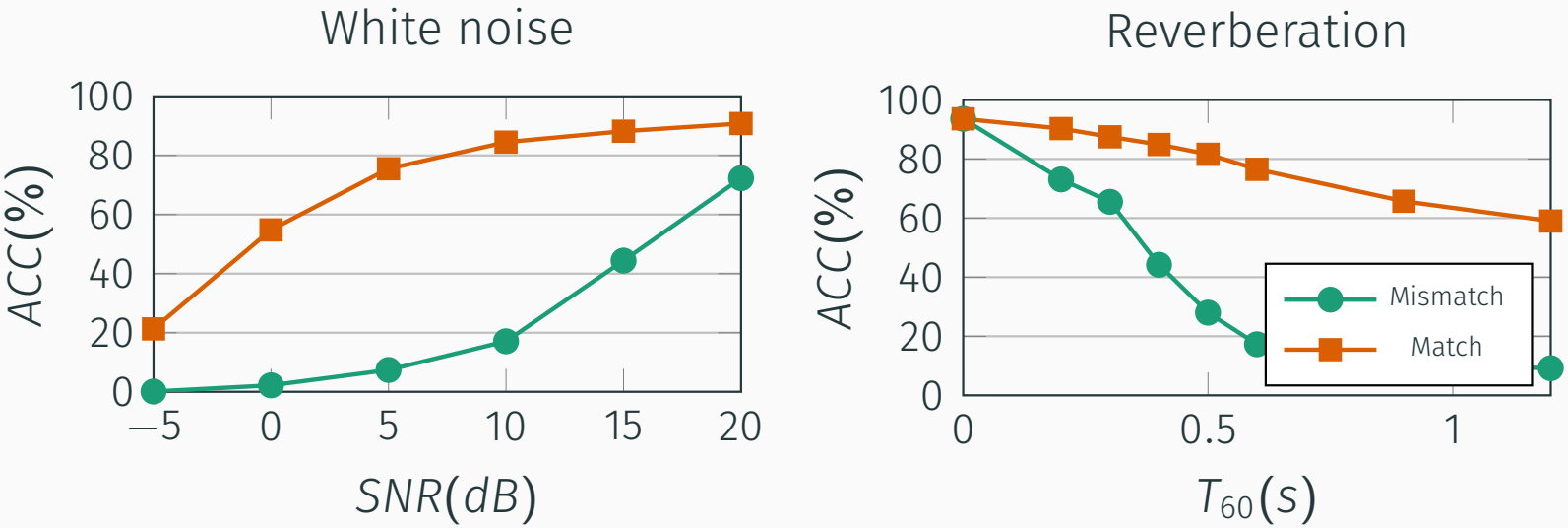
- Training with contaminated data.





Techniques to improve the robustness

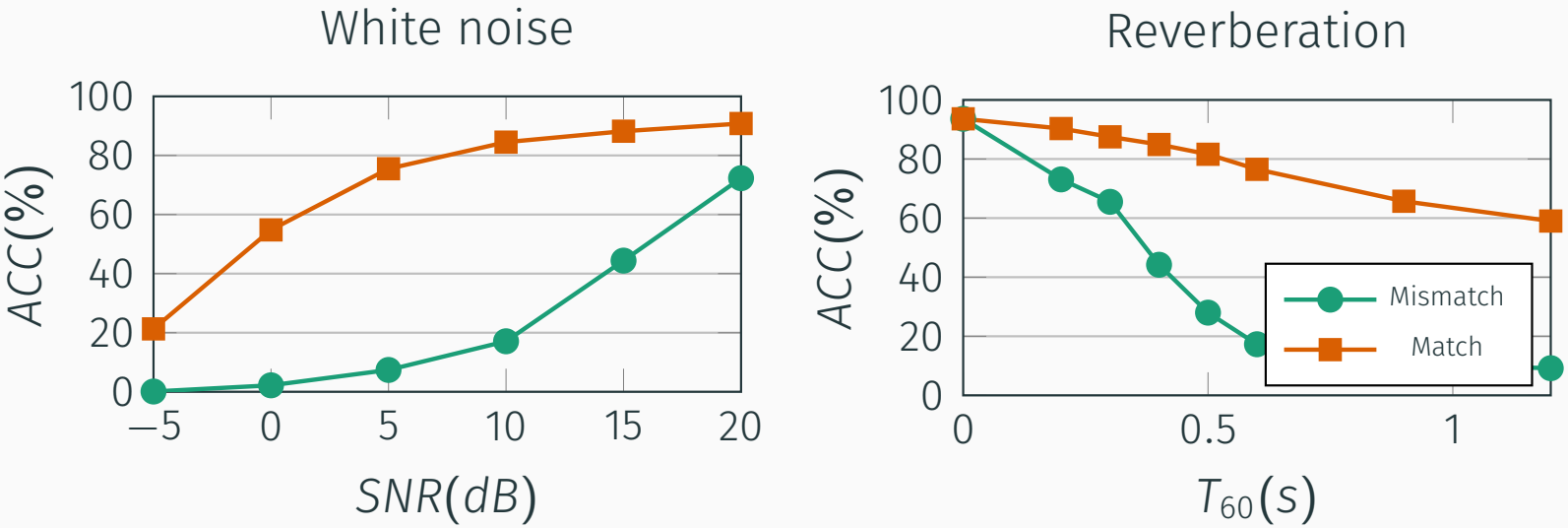
- Training with contaminated data.



- Model domain techniques: VTS (Moreno et al., 1996).

Techniques to improve the robustness

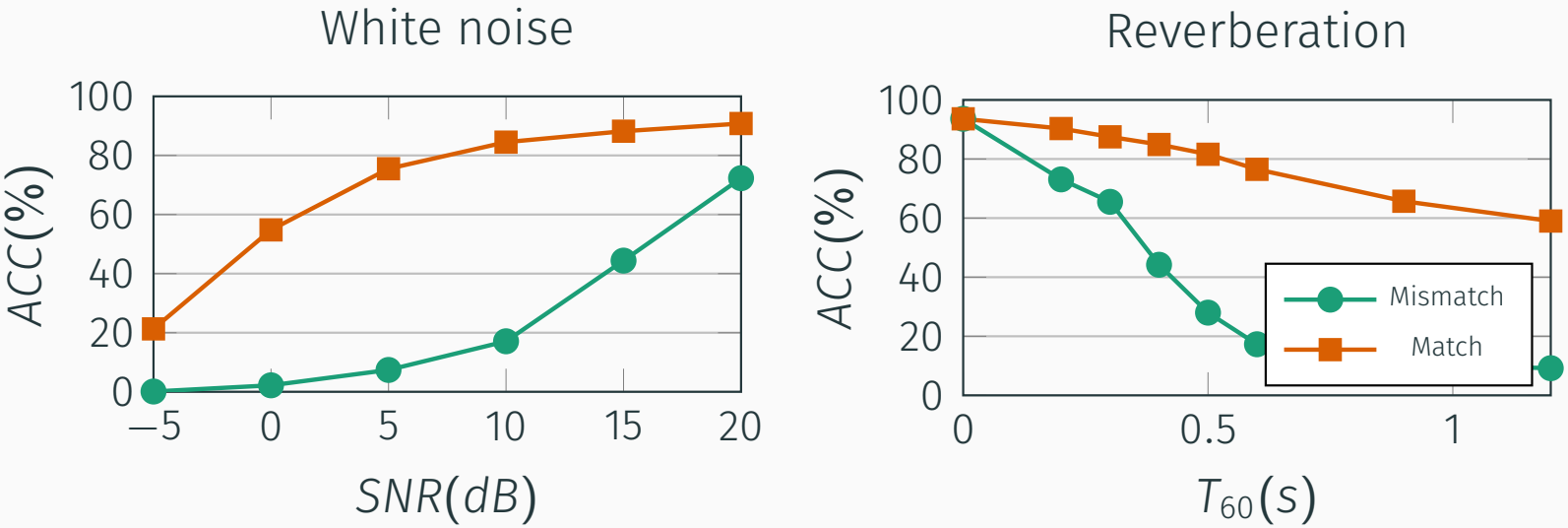
- Training with contaminated data.



- Model domain techniques: VTS (Moreno et al., 1996).
- Feature domain techniques:

Techniques to improve the robustness

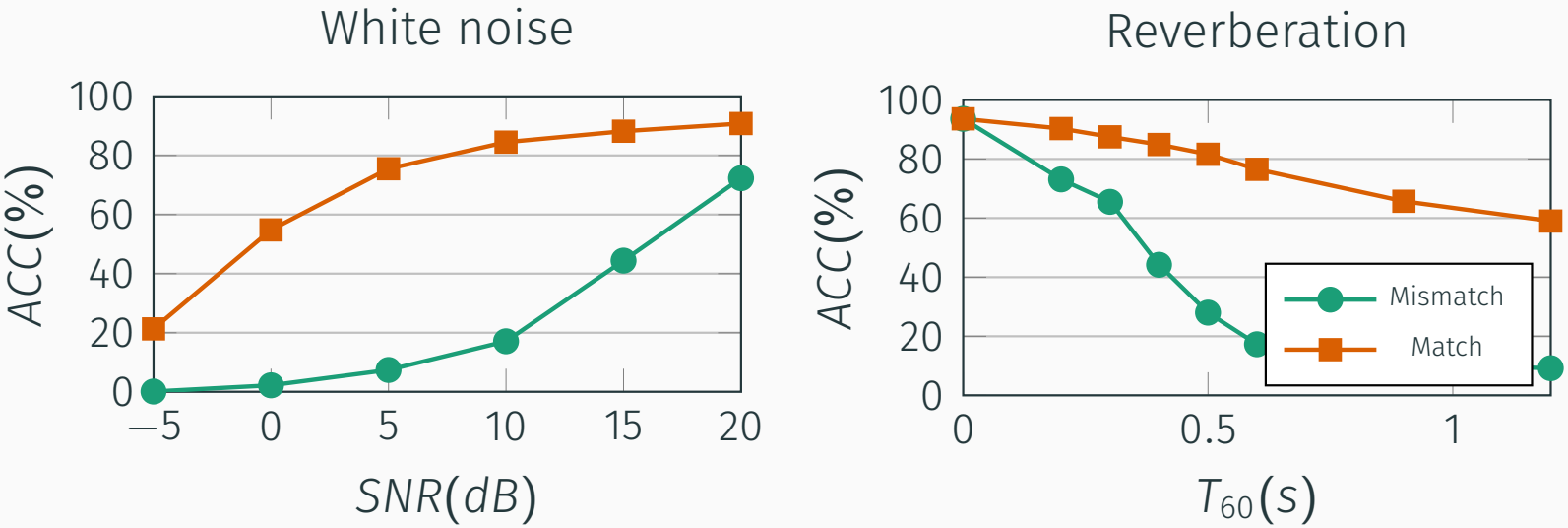
- Training with contaminated data.



- Model domain techniques: VTS (Moreno et al., 1996).
- Feature domain techniques:
  - Feature normalization methods: CMVN.

Techniques to improve the robustness

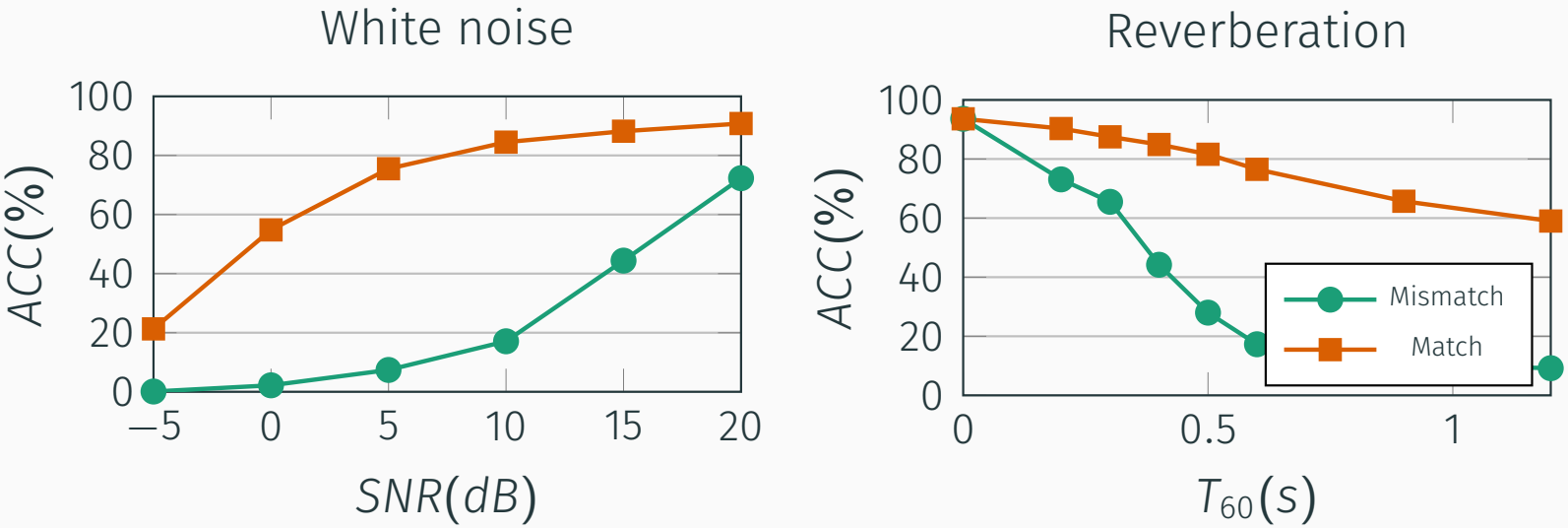
- Training with contaminated data.



- Model domain techniques: VTS (Moreno et al., 1996).
- Feature domain techniques:
  - Feature normalization methods: CMVN.
  - Feature compensation methods: Spectral Substraction (Berouti et al., 1979).

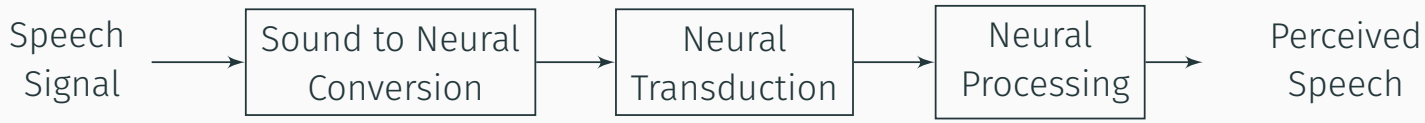
Techniques to improve the robustness

- Training with contaminated data.

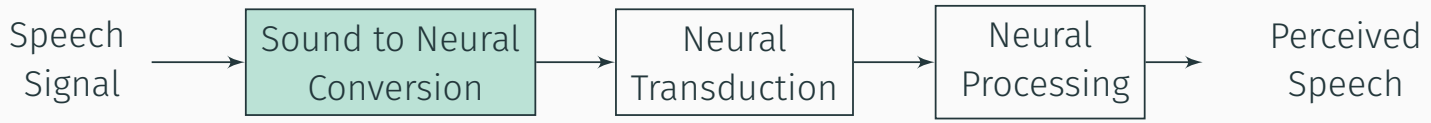


- Model domain techniques: VTS (Moreno et al., 1996).
- Feature domain techniques:
  - Feature normalization methods: CMVN.
  - Feature compensation methods: Spectral Substraction (Berouti et al., 1979).
  - Noise-resistant features: Auditory based.

Model of the Human Auditory System (HAS)

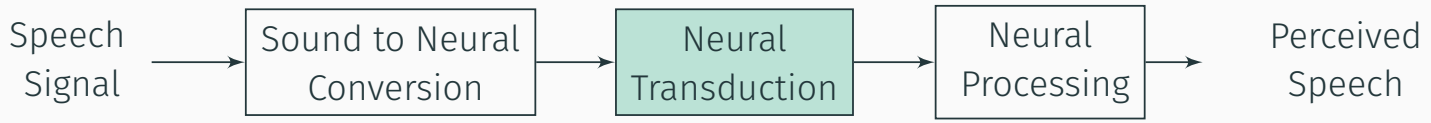


Model of the Human Auditory System (HAS)



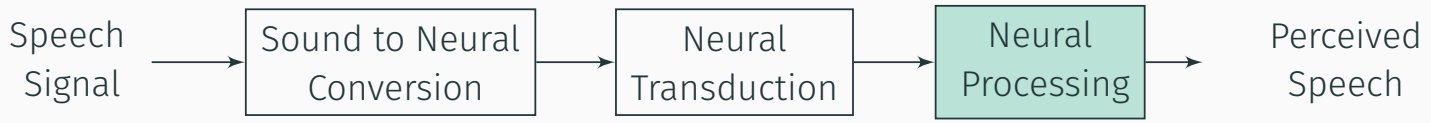
- Frequency Selectivity.
- Auditory Masking

Model of the Human Auditory System (HAS)



- Neural Coding.
- Synchrony Effect.

Model of the Human Auditory System (HAS)



- Not yet well understood.

Modeling the HAS for ASR

- Classic auditory models: Seneff (1990) or Zhang et al. (2001).

Modeling the HAS for ASR

- Classic auditory models: Seneff (1990) or Zhang et al. (2001).
  - High computational cost.

Modeling the HAS for ASR

- Classic auditory models: Seneff (1990) or Zhang et al. (2001).
  - High computational cost.
  - Complexity.

Modeling the HAS for ASR

- Classic auditory models: Seneff (1990) or Zhang et al. (2001).
  - High computational cost.
  - Complexity.
  - Unstable.

Modeling the HAS for ASR

- Classic auditory models: Seneff (1990) or Zhang et al. (2001).
  - High computational cost.
  - Complexity.
  - Unstable.
- Auditory based features:

Modeling the HAS for ASR

- Classic auditory models: Seneff (1990) or Zhang et al. (2001).
  - High computational cost.
  - Complexity.
  - Unstable.
- Auditory based features:
  - Mel-Frequency Cepstral Coefficients (MFCC) (Davis and Mermelstein, 1980).

Modeling the HAS for ASR

- Classic auditory models: Seneff (1990) or Zhang et al. (2001).
  - High computational cost.
  - Complexity.
  - Unstable.
- Auditory based features:
  - Mel-Frequency Cepstral Coefficients (MFCC) (Davis and Mermelstein, 1980).
  - Perceptually-based Linear Prediction (PLP) (Hermansky et al., 1985).

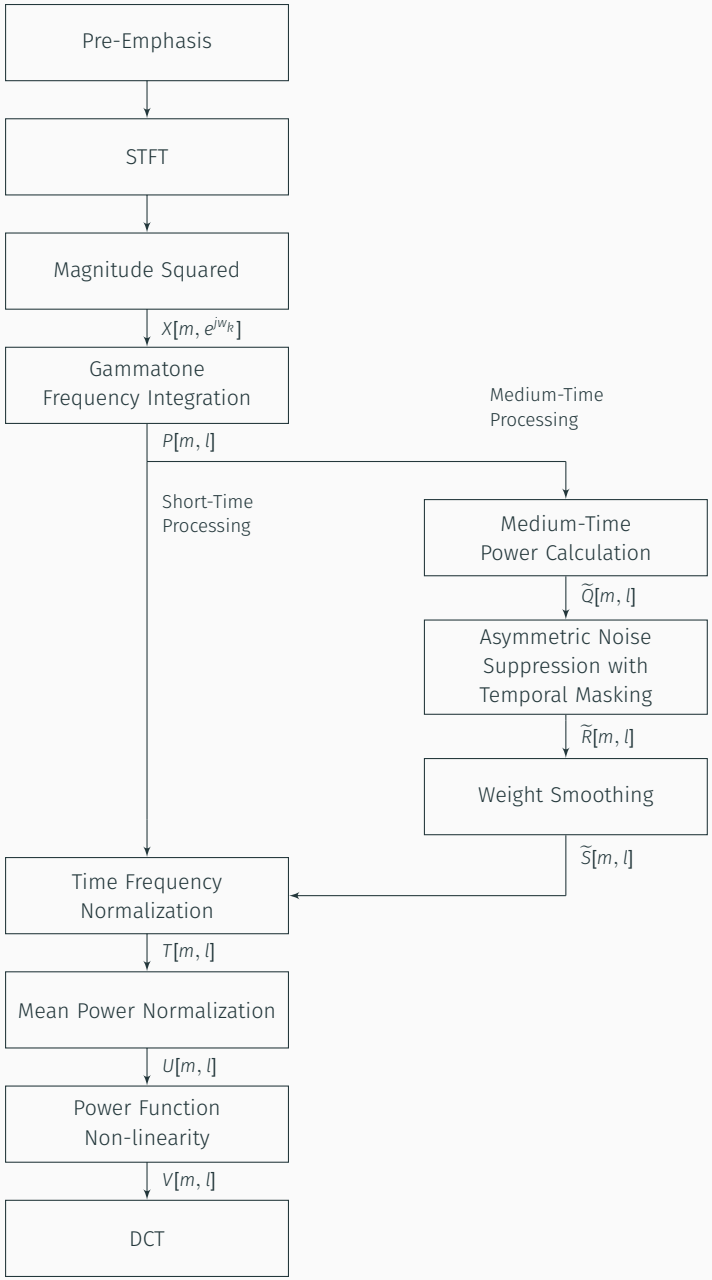
Modeling the HAS for ASR

- Classic auditory models: Seneff (1990) or Zhang et al. (2001).
  - High computational cost.
  - Complexity.
  - Unstable.
- Auditory based features:
  - Mel-Frequency Cepstral Coefficients (MFCC) (Davis and Mermelstein, 1980).
  - Perceptually-based Linear Prediction (PLP) (Hermansky et al., 1985).
  - Relative Spectral processing (RASTA) (Hermansky and Morgan, 1994).

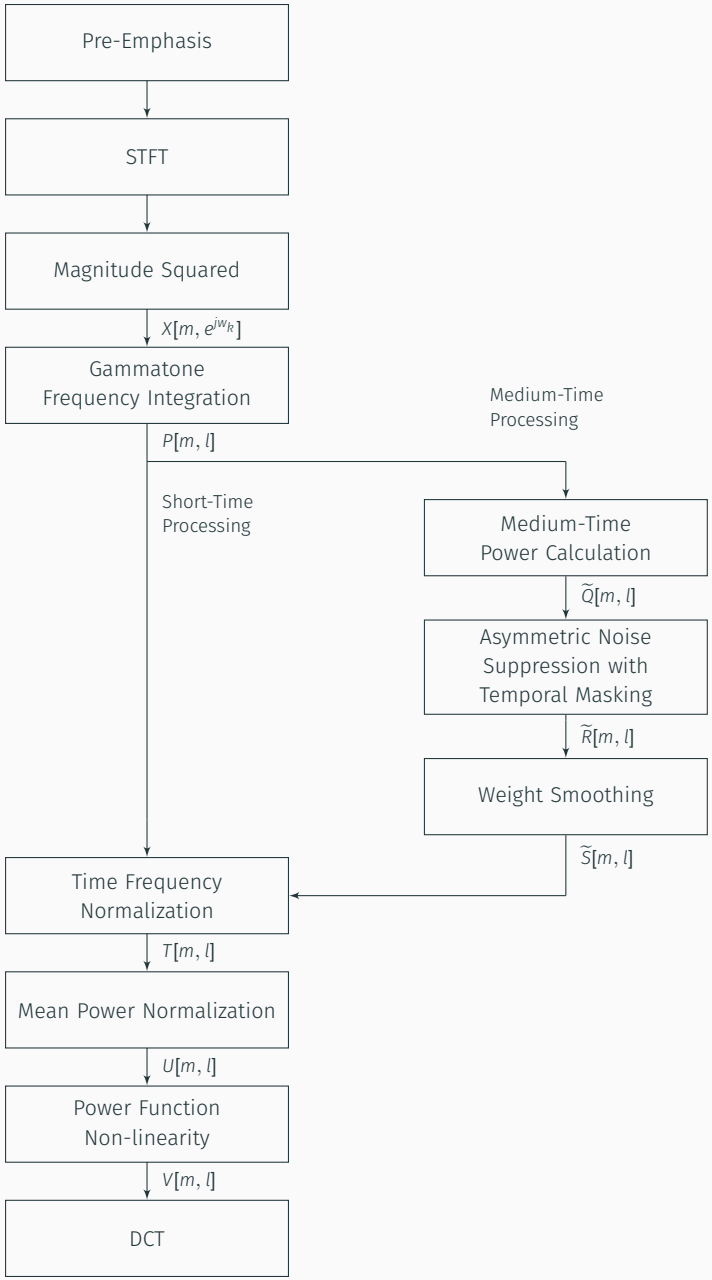
Modeling the HAS for ASR

- Classic auditory models: Seneff (1990) or Zhang et al. (2001).
  - High computational cost.
  - Complexity.
  - Unstable.
- Auditory based features:
  - Mel-Frequency Cepstral Coefficients (MFCC) (Davis and Mermelstein, 1980).
  - Perceptually-based Linear Prediction (PLP) (Hermansky et al., 1985).
  - Relative Spectral processing (RASTA) (Hermansky and Morgan, 1994).
  - Power Normalized Cepstral Coefficients (PNCC) (Kim and Stern, 2016).

Power Normalized Cepstral Coefficients (PNCC)

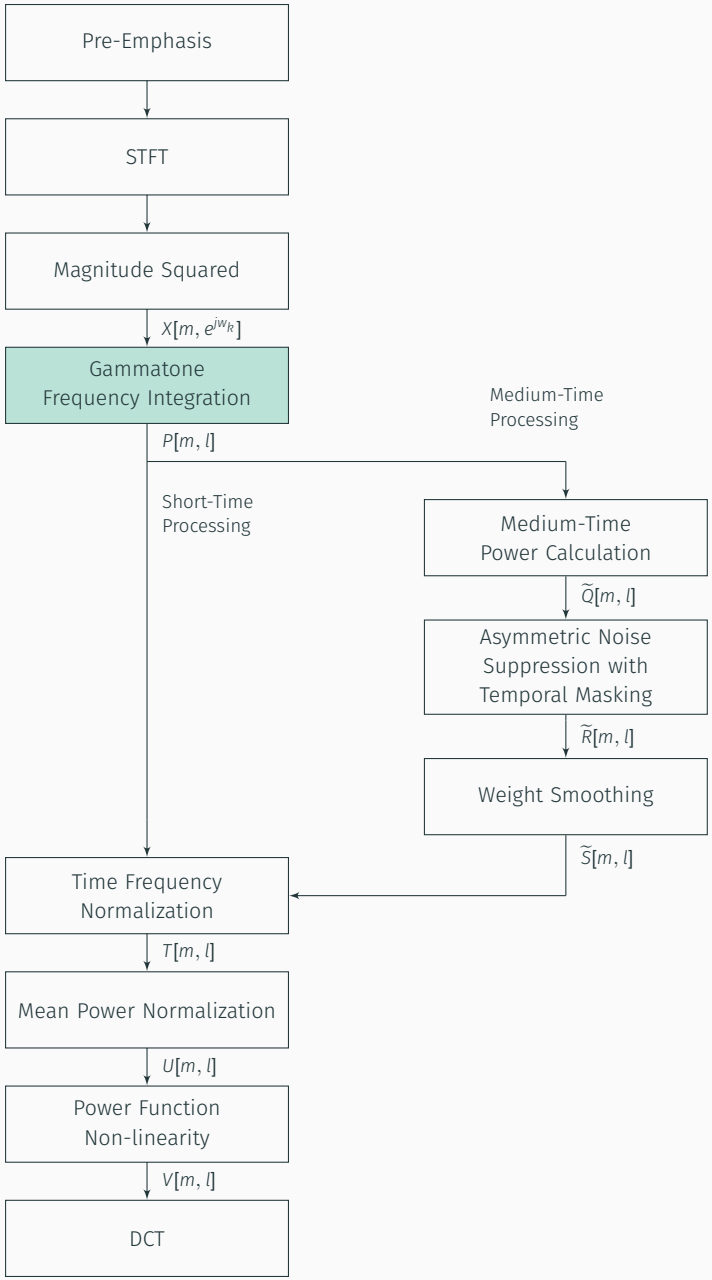


Power Normalized Cepstral Coefficients (PNCC)



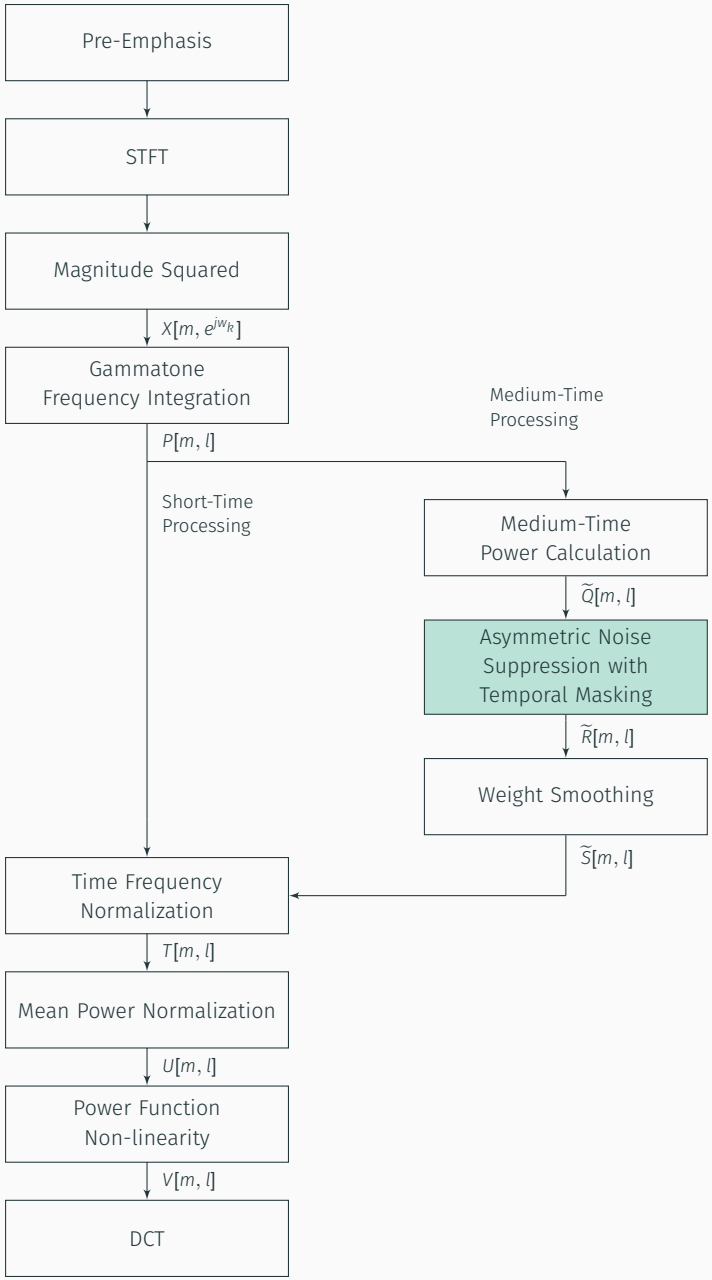
- Pragmatic implementation of auditory principles:

Power Normalized Cepstral Coefficients (PNCC)



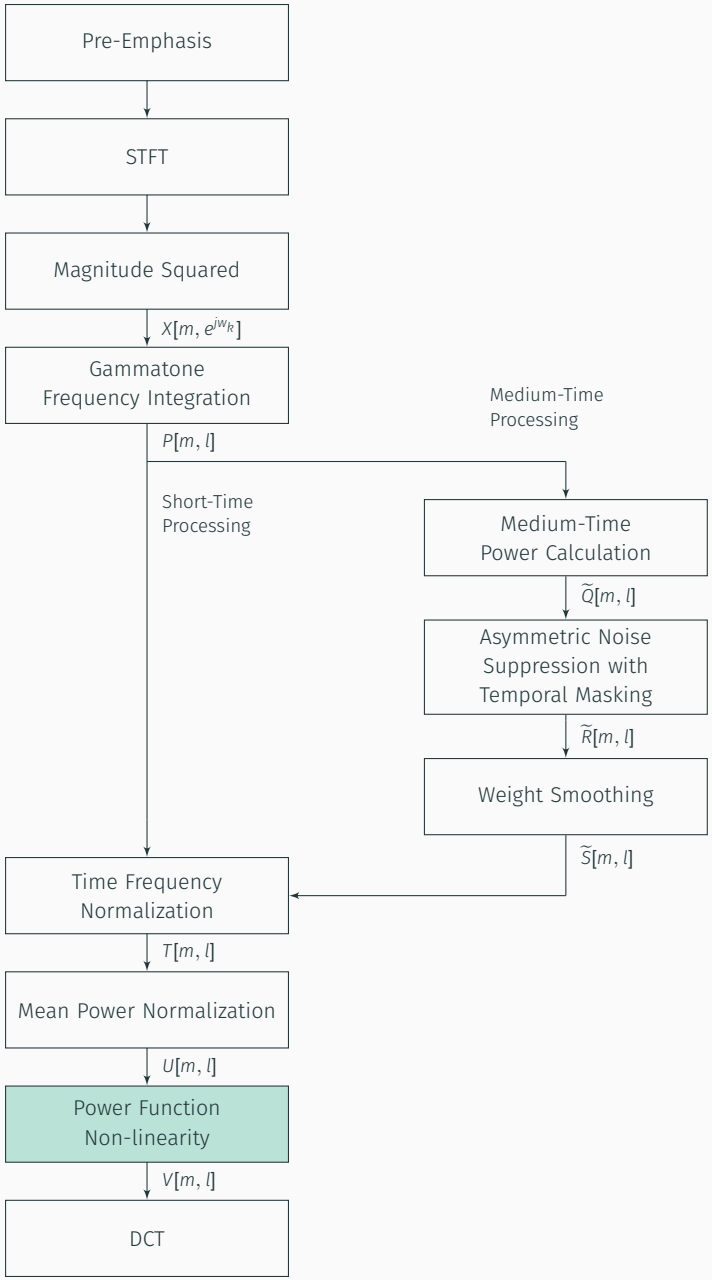
- Pragmatic implementation of auditory principles:
  - Gammatone filter-banks (ERBs).

Power Normalized Cepstral Coefficients (PNCC)



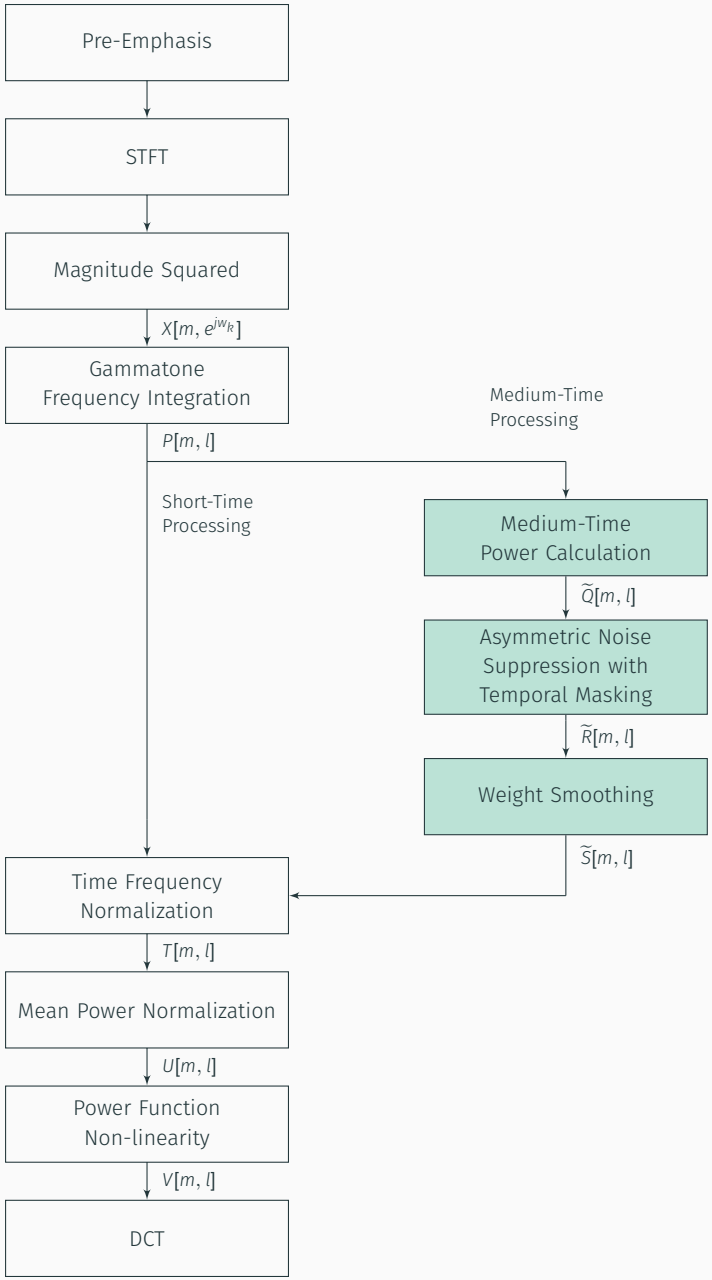
- Pragmatic implementation of auditory principles:
  - Gammatone filter-banks (ERBs).
  - Elementary temporal masking model.

Power Normalized Cepstral Coefficients (PNCC)



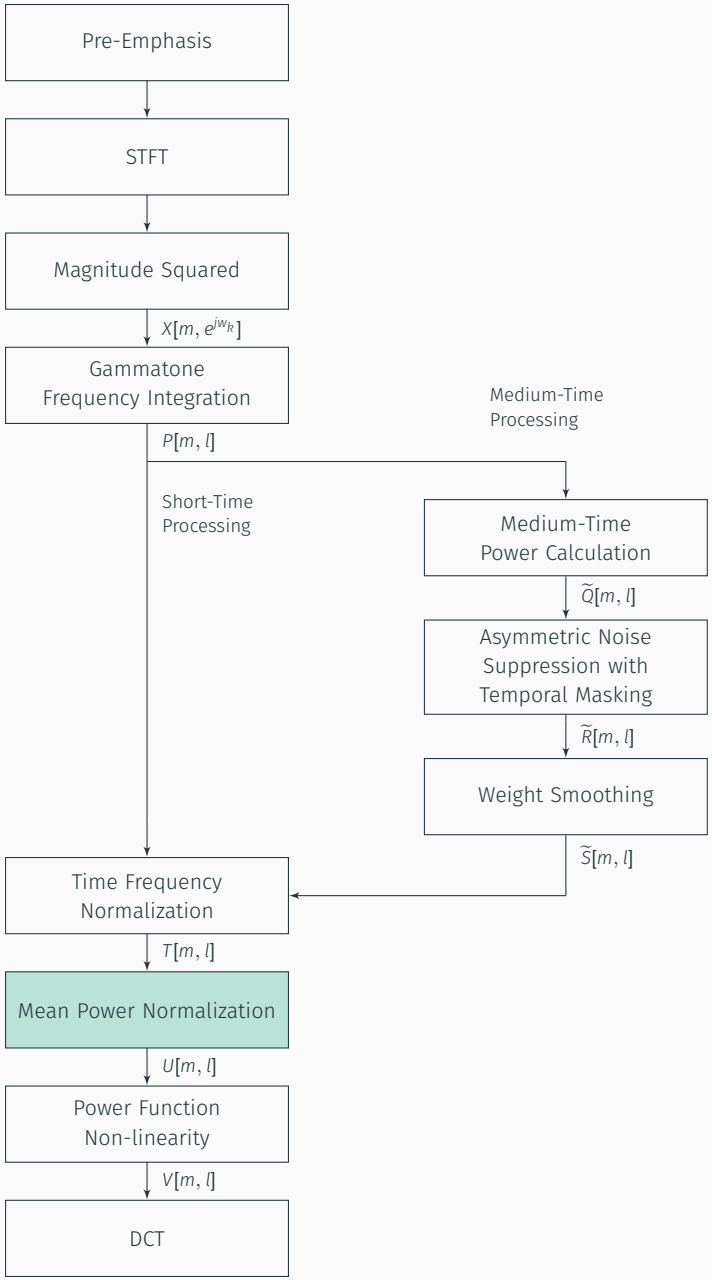
- Pragmatic implementation of auditory principles:
  - Gammatone filter-banks (ERBs).
  - Elementary temporal masking model.
  - Power law nonlinearity.

Power Normalized Cepstral Coefficients (PNCC)



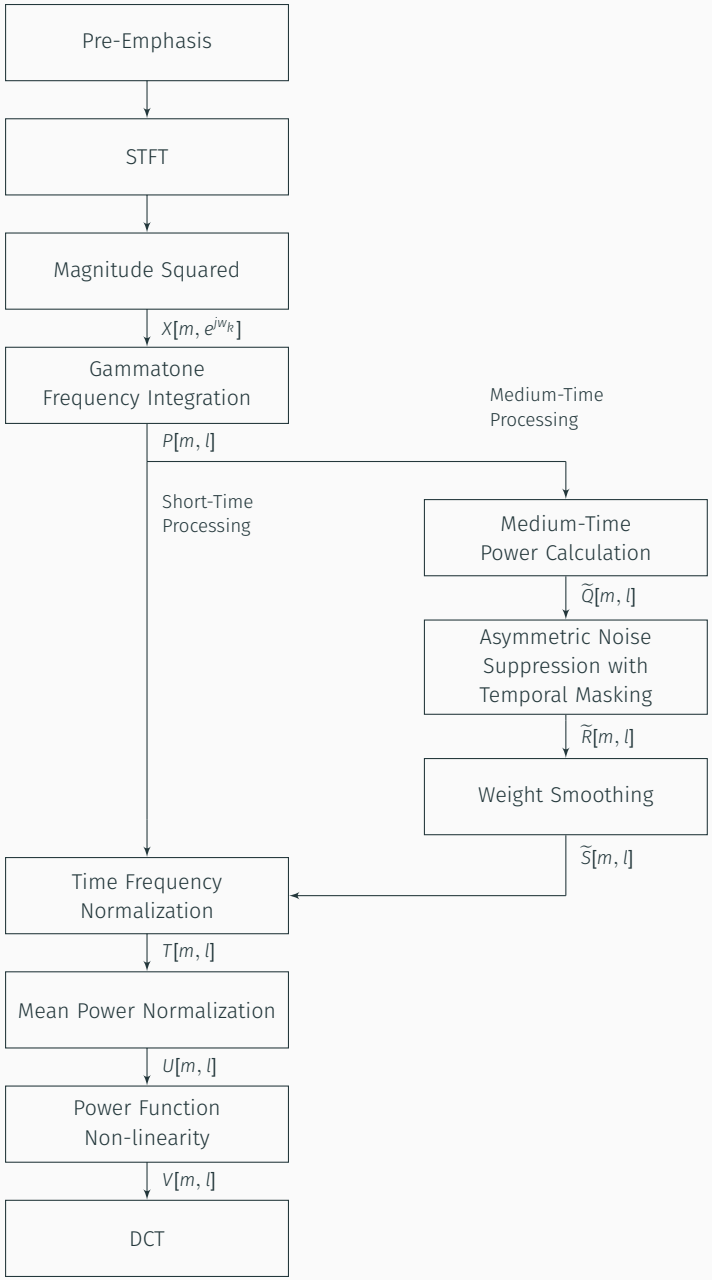
- Pragmatic implementation of auditory principles:
  - Gammatone filter-banks (ERBs).
  - Elementary temporal masking model.
  - Power law nonlinearity.
- Medium-time environmental compensation.

Power Normalized Cepstral Coefficients (PNCC)



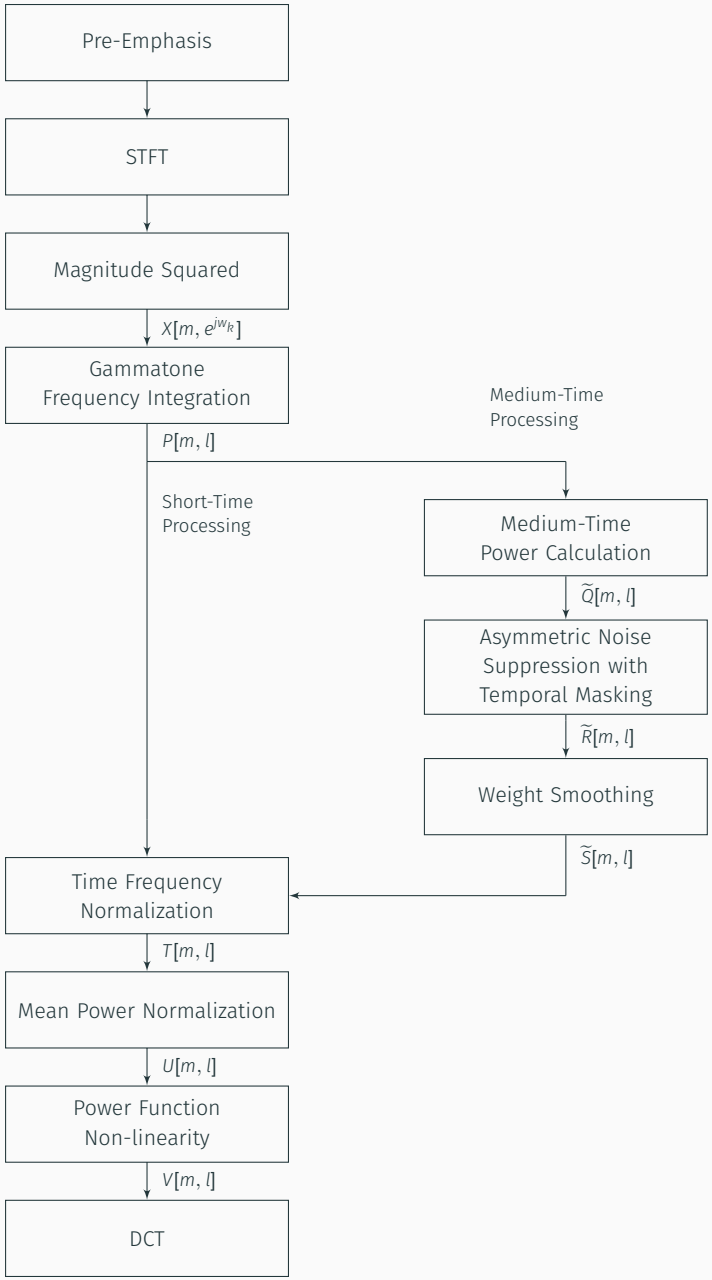
- Pragmatic implementation of auditory principles:
  - Gammatone filter-banks (ERBs).
  - Elementary temporal masking model.
  - Power law nonlinearity.
- Medium-time environmental compensation.
- Mean Power Normalization.

Power Normalized Cepstral Coefficients (PNCC)



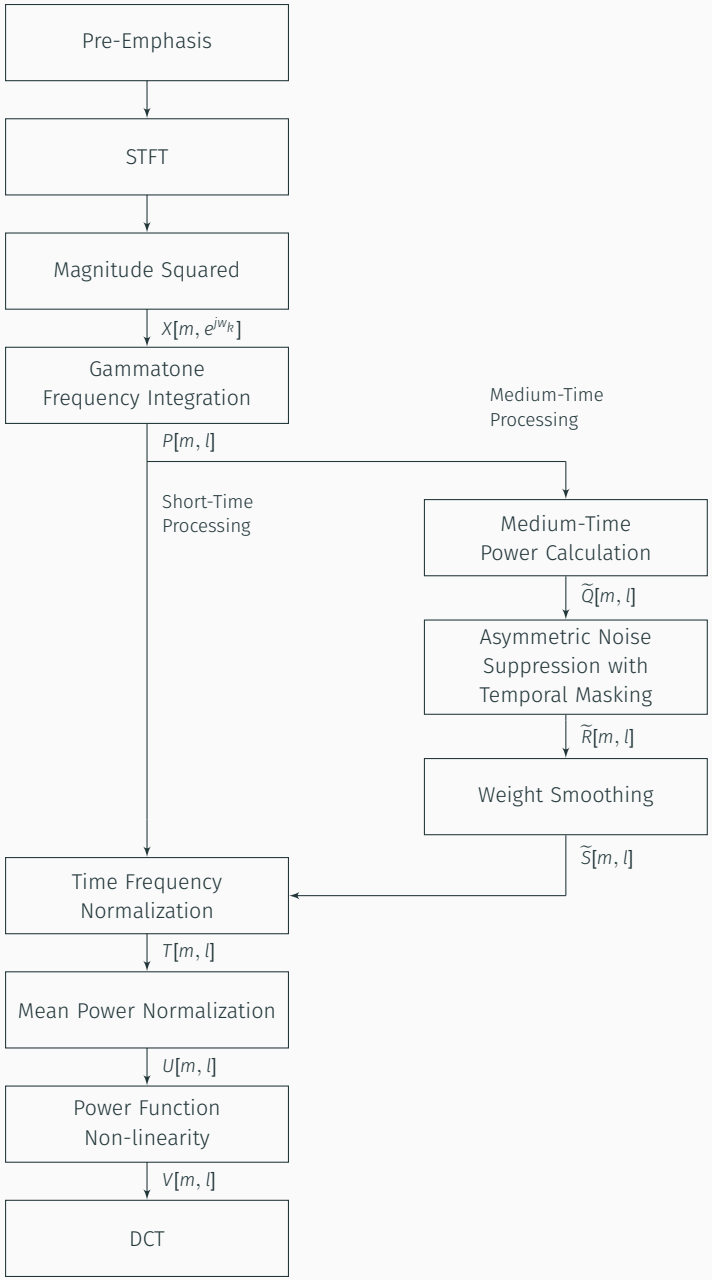
- Pragmatic implementation of auditory principles:
  - Gammatone filter-banks (ERBs).
  - Elementary temporal masking model.
  - Power law nonlinearity.
- Medium-time environmental compensation.
- Mean Power Normalization.
- Not represented:

Power Normalized Cepstral Coefficients (PNCC)



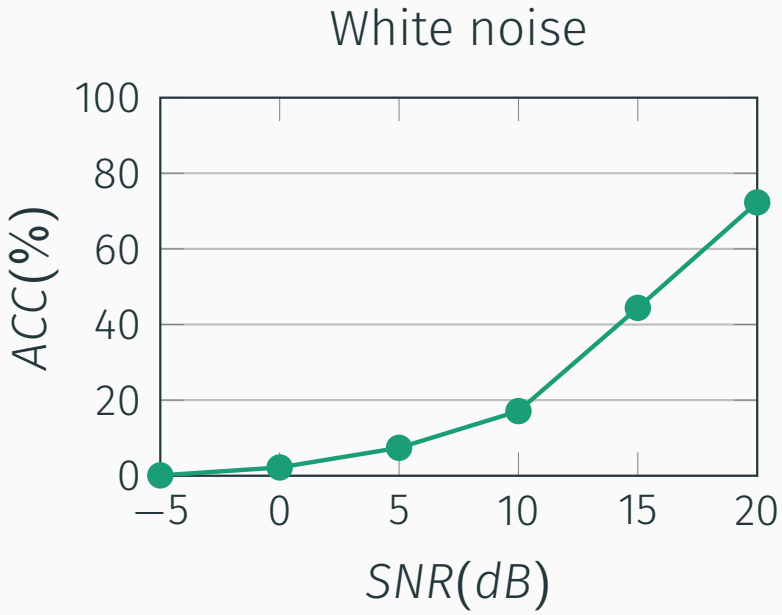
- Pragmatic implementation of auditory principles:
  - Gammatone filter-banks (ERBs).
  - Elementary temporal masking model.
  - Power law nonlinearity.
- Medium-time environmental compensation.
- Mean Power Normalization.
- Not represented:
  - Auditory Masking.

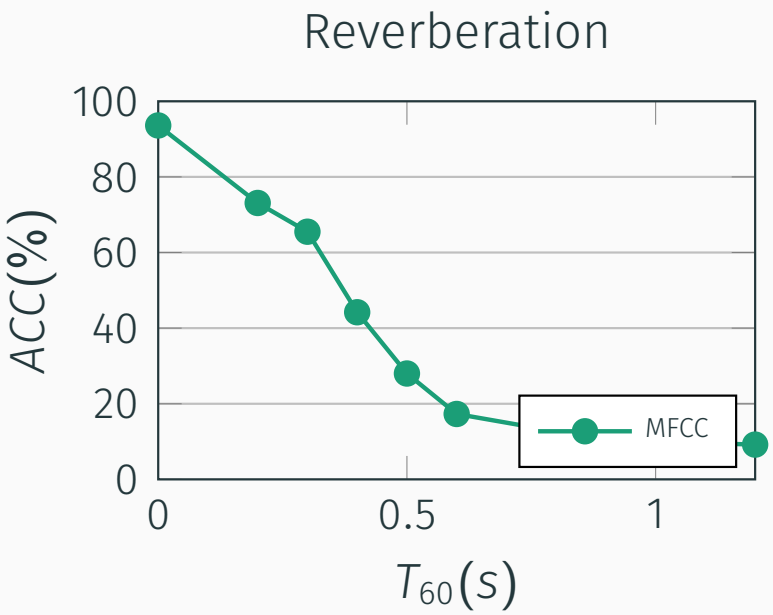
Power Normalized Cepstral Coefficients (PNCC)



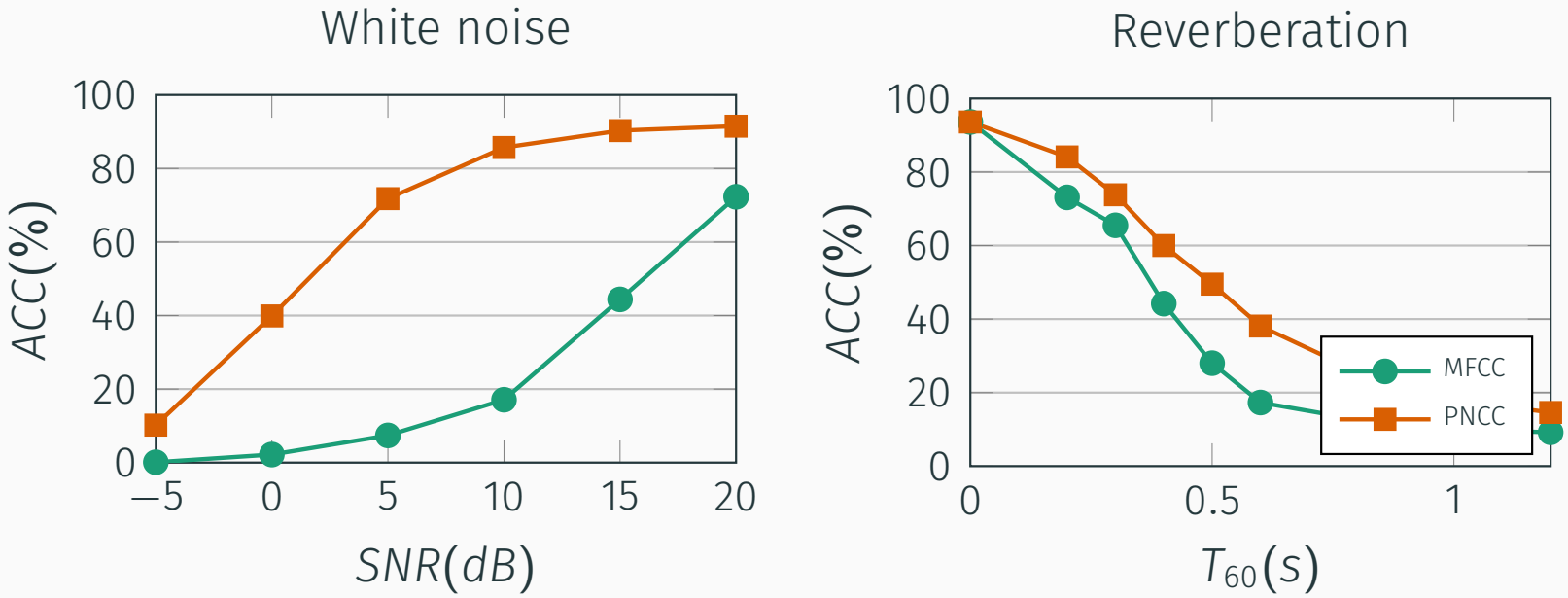
- Pragmatic implementation of auditory principles:
  - Gammatone filter-banks (ERBs).
  - Elementary temporal masking model.
  - Power law nonlinearity.
- Medium-time environmental compensation.
- Mean Power Normalization.
- Not represented:
  - Auditory Masking.
  - Detailed timing structure.

PNCC Results





PNCC Results



It is our main baseline.

Power-Normalized Cochleograms

Introduction

- Modeling of the masking phenomena in the cochlea.

Introduction

- Modeling of the masking phenomena in the cochlea.
- Masking smoothes away noise without deteriorating the signal quality.

Introduction

- Modeling of the masking phenomena in the cochlea.
- Masking smoothes away noise without deteriorating the signal quality.
- By filtering a spectro-temporal representation of speech as if it were an image, based on *mathematical morphology*.

Introduction

- Modeling of the masking phenomena in the cochlea.
- Masking smoothes away noise without deteriorating the signal quality.
- By filtering a spectro-temporal representation of speech as if it were an image, based on *mathematical morphology*.
- Structuring Element (SE) design based on the masking properties of the HAS.

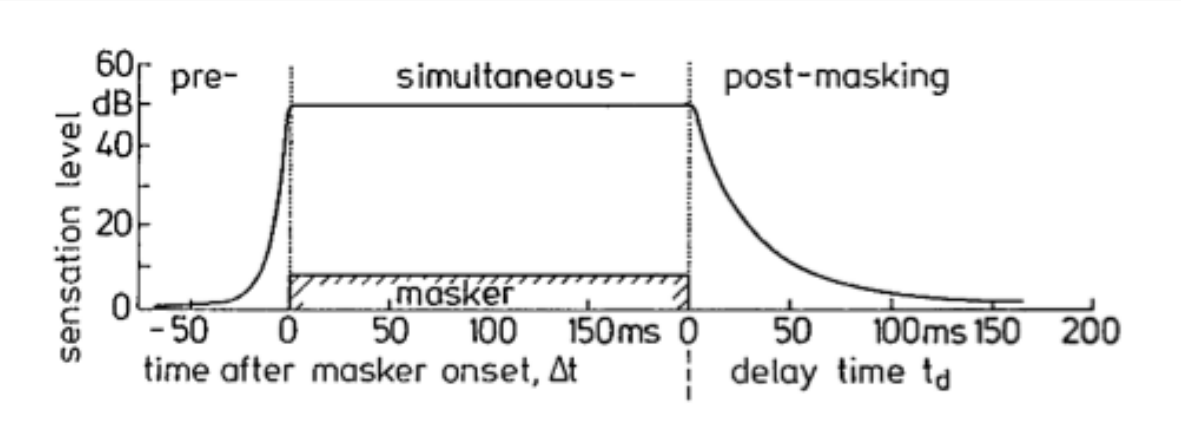
Auditory masking

The perception of one sound is affected by the presence of another sound.

From Fastl and Zwicker (2007).

Auditory masking

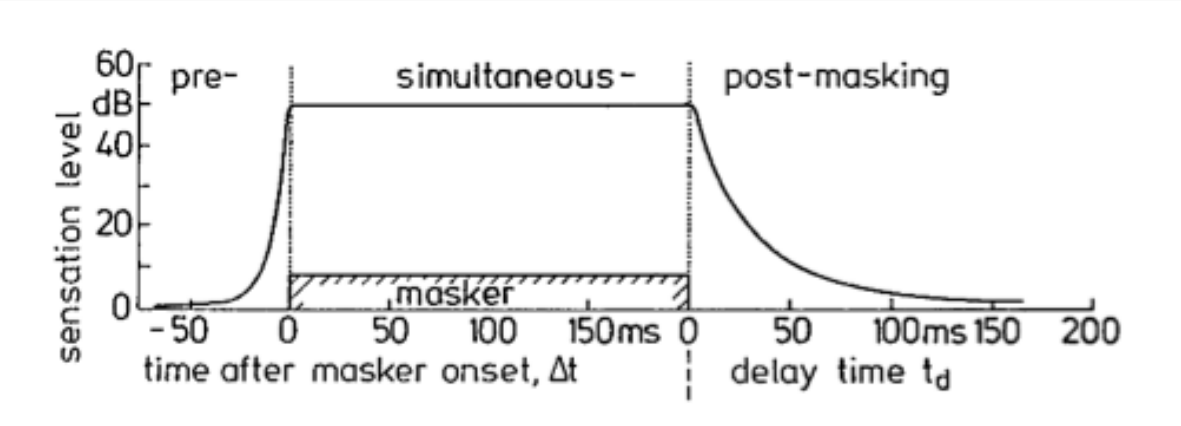
The perception of one sound is affected by the presence of another sound.



From Fastl and Zwicker (2007).

Auditory masking

The perception of one sound is affected by the presence of another sound.

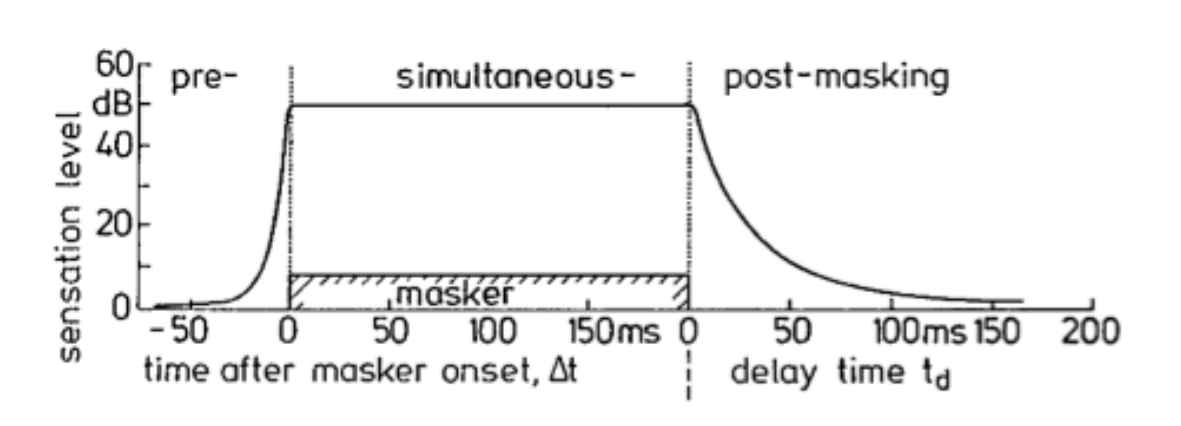


- Simultaneous or spectral masking.

From Fastl and Zwicker (2007).

Auditory masking

The perception of one sound is affected by the presence of another sound.

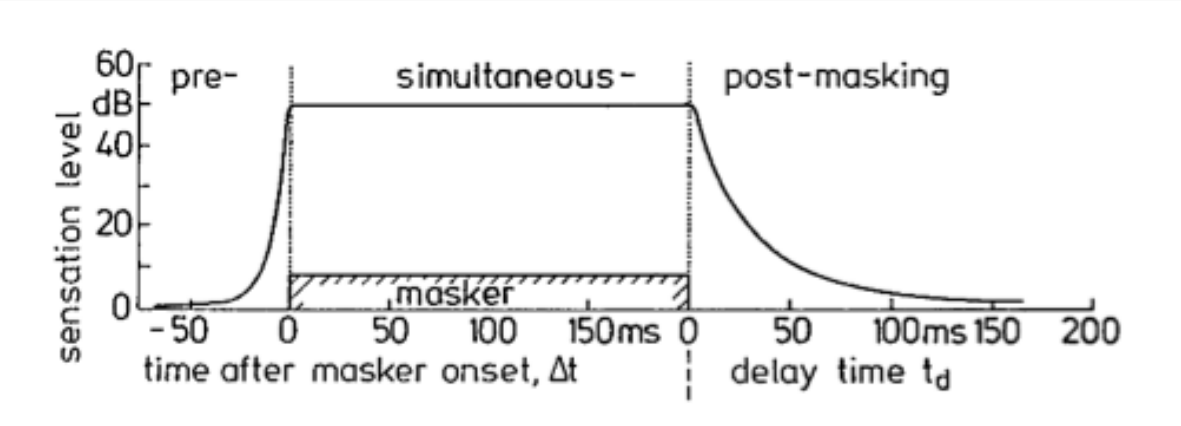


- Simultaneous or spectral masking.
- Temporal masking.

From Fastl and Zwicker (2007).

Auditory masking

The perception of one sound is affected by the presence of another sound.

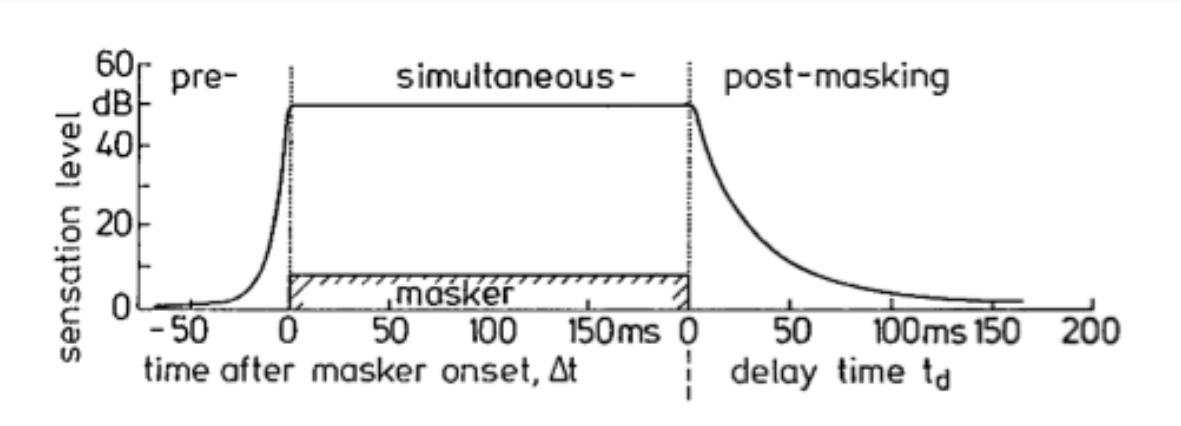


- Simultaneous or spectral masking.
- Temporal masking.
  - Premasking.

From Fastl and Zwicker (2007).

Auditory masking

The perception of one sound is affected by the presence of another sound.

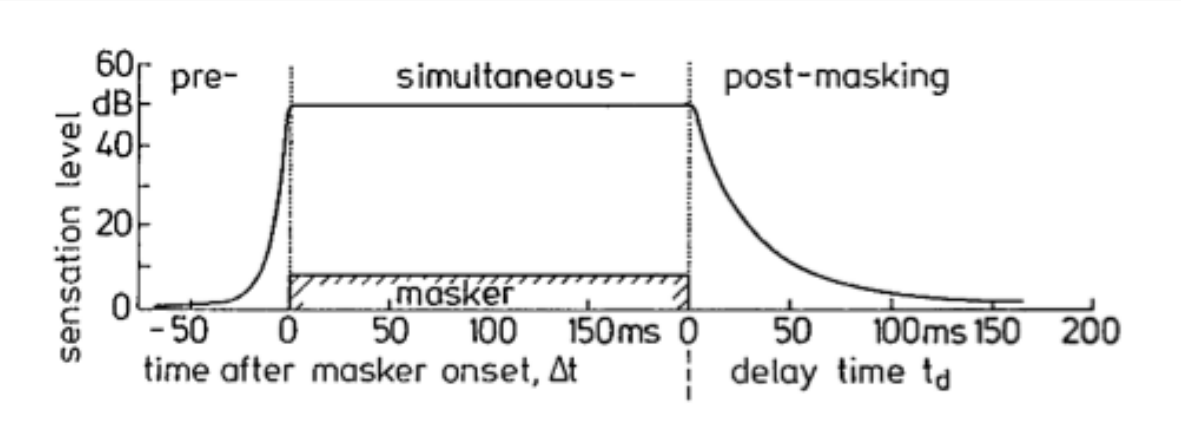


- Simultaneous or spectral masking.
- Temporal masking.
  - Premasking.
  - Postmasking.

From Fastl and Zwicker (2007).

Auditory masking

The perception of one sound is affected by the presence of another sound.



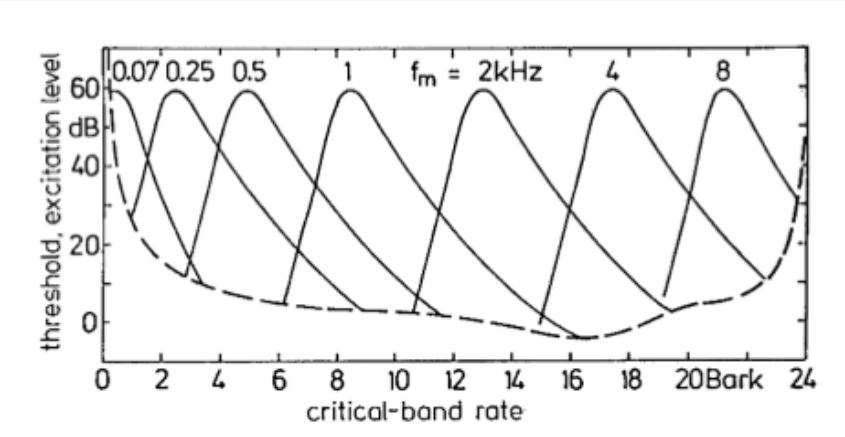
- Simultaneous or spectral masking.
- Temporal masking.
  - Premasking.
  - Postmasking.
- Masking models based from empirical measurements.

From Fastl and Zwicker (2007).

Simultaneous Masking Model

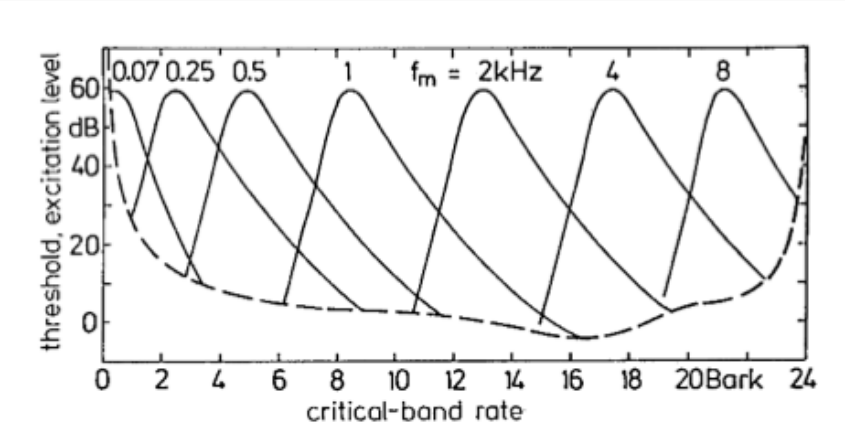
Simultaneous Masking Model

Masking curves presented by Fastl and Zwicker (2007).



Simultaneous Masking Model

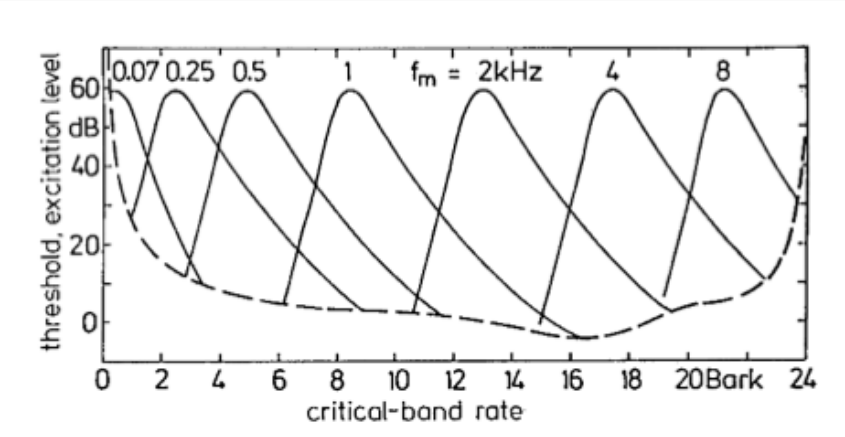
Masking curves presented by Fastl and Zwicker (2007).



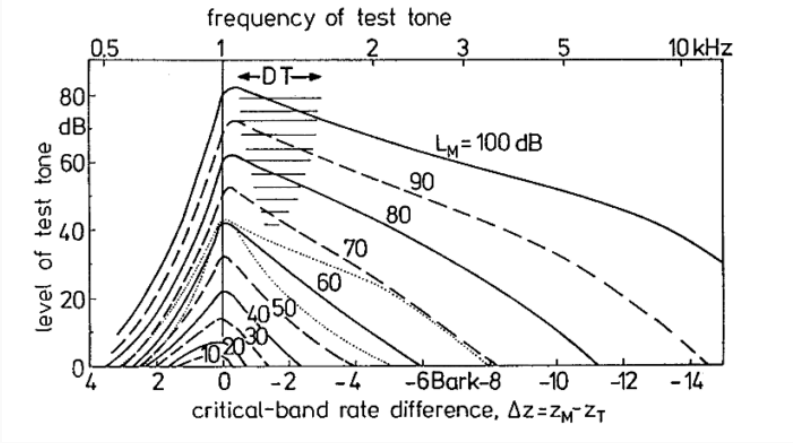
- Represented in bark scale.

Simultaneous Masking Model

Masking curves presented by Fastl and Zwicker (2007).



- Represented in bark scale.
- Varying intensity levels can also have an effect on masking.



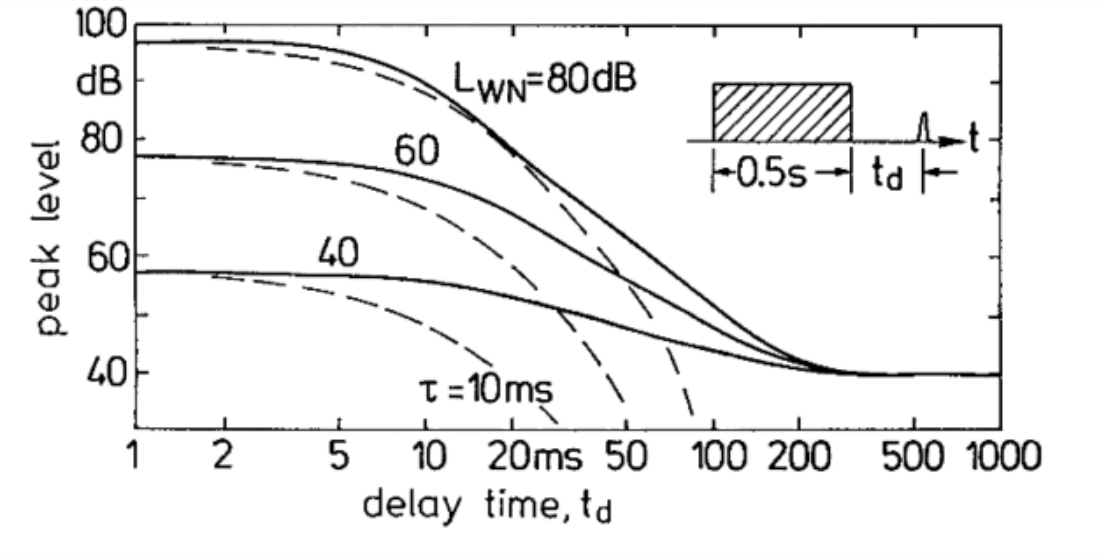
Temporal Masking Model

- Linear model for premasking.

From Fastl and Zwicker (2007).

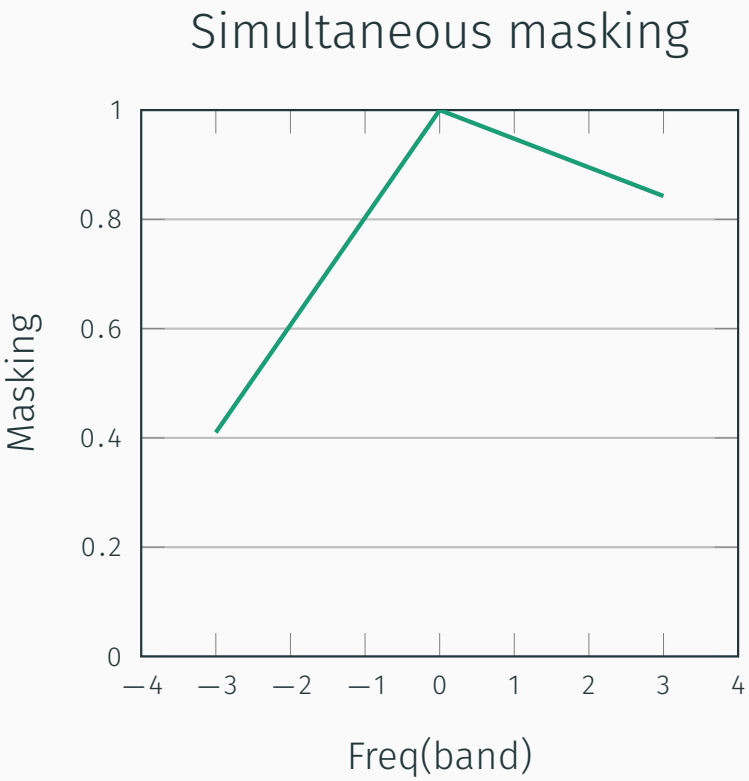
Temporal Masking Model

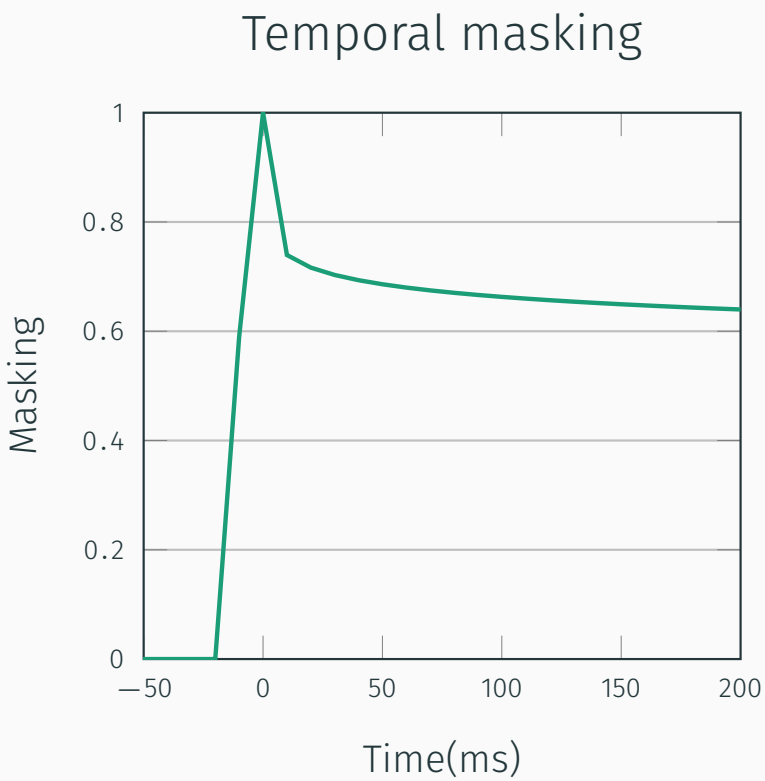
- Linear model for premasking.
- Logarithmic model for postmasking (Jesteadt et al., 1982).



From Fastl and Zwicker (2007).

Masking Models





Morphological Filtering

- Most prominent or salient elements of the cochleogram mask their surroundings.

Morphological Filtering

- Most prominent or salient elements of the cochleogram mask their surroundings.
- Smoothing of the spectro-temporal envelope.

Morphological Filtering

- Most prominent or salient elements of the cochleogram mask their surroundings.
- Smoothing of the spectro-temporal envelope.
- *Structuring Element* (SE) based on the masking models.

Morphological Filtering

- Most prominent or salient elements of the cochleogram mask their surroundings.
- Smoothing of the spectro-temporal envelope.
- *Structuring Element* (SE) based on the masking models.
- Closing operator preserves the regions with a similar shape.

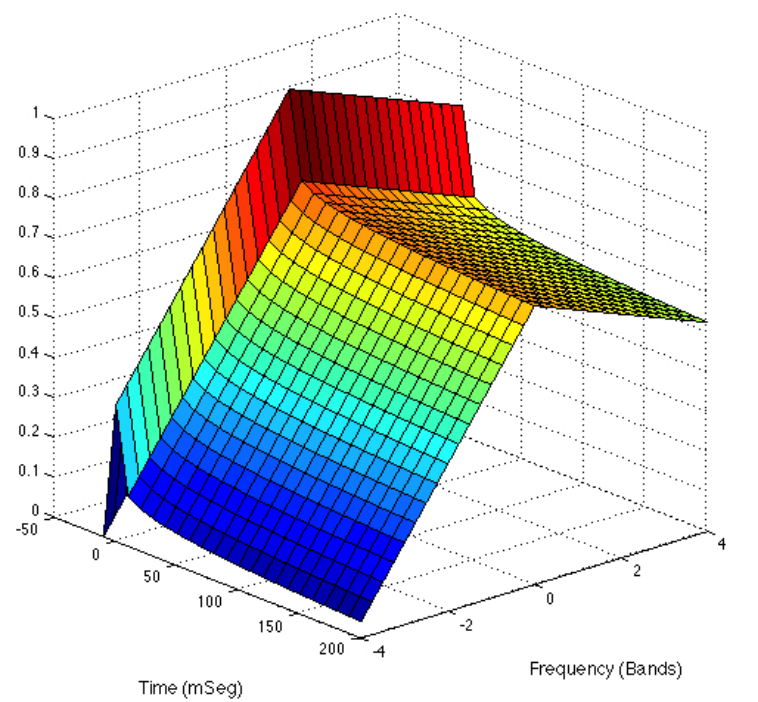
Morphological Filtering

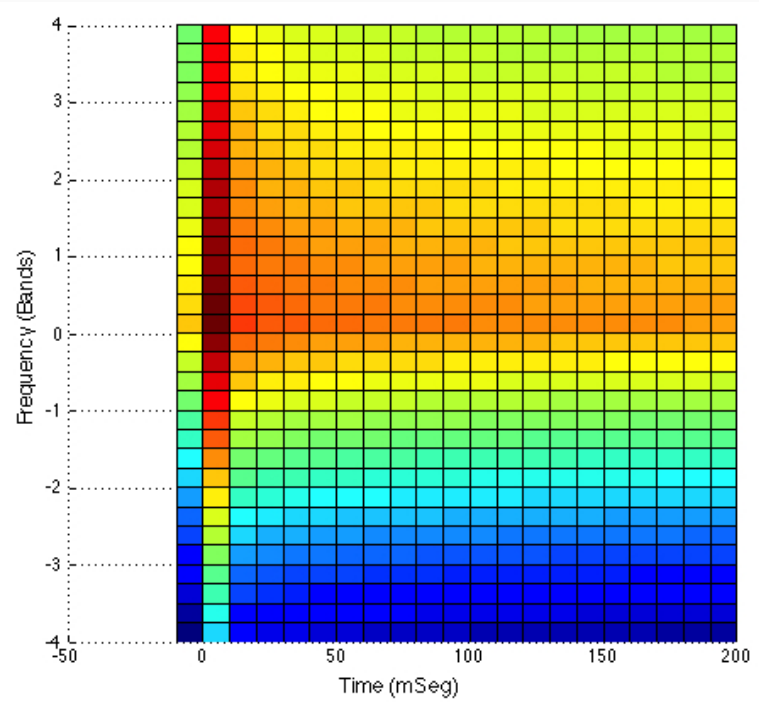
- Most prominent or salient elements of the cochleogram mask their surroundings.
- Smoothing of the spectro-temporal envelope.
- *Structuring Element* (SE) based on the masking models.
- Closing operator preserves the regions with a similar shape.

Morphological Filtering

- Most prominent or salient elements of the cochleogram mask their surroundings.
- Smoothing of the spectro-temporal envelope.
- *Structuring Element* (SE) based on the masking models.
- Closing operator preserves the regions with a similar shape.

Piecewise-lineal SE





Piecewise-convex SE

- SE for a single frequency-time point is very sharp.

Piecewise-convex SE

- SE for a single frequency-time point is very sharp.
- Smoothness restriction.

Piecewise-convex SE

- SE for a single frequency-time point is very sharp.
- Smoothness restriction.
- Masking response of a particular time and frequency is the aggregation of many single-point responses (Fastl and Zwicker, 2007).

Piecewise-convex SE

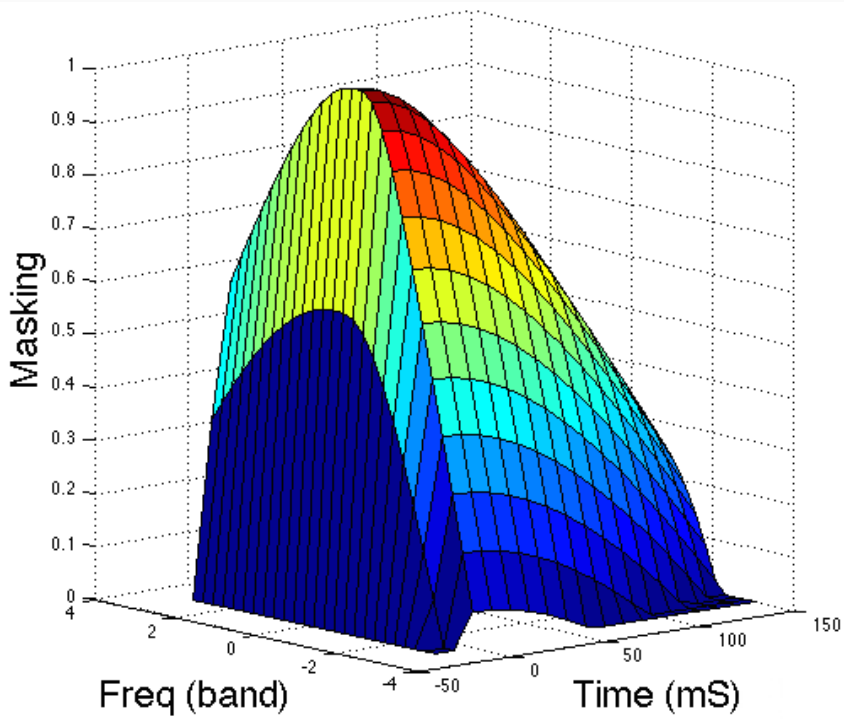
- SE for a single frequency-time point is very sharp.
- Smoothness restriction.
- Masking response of a particular time and frequency is the aggregation of many single-point responses (Fastl and Zwicker, 2007).
- A piecewise-convex SE is built by fitting 4 hyperboloid quadrants to contour of the linear SE.

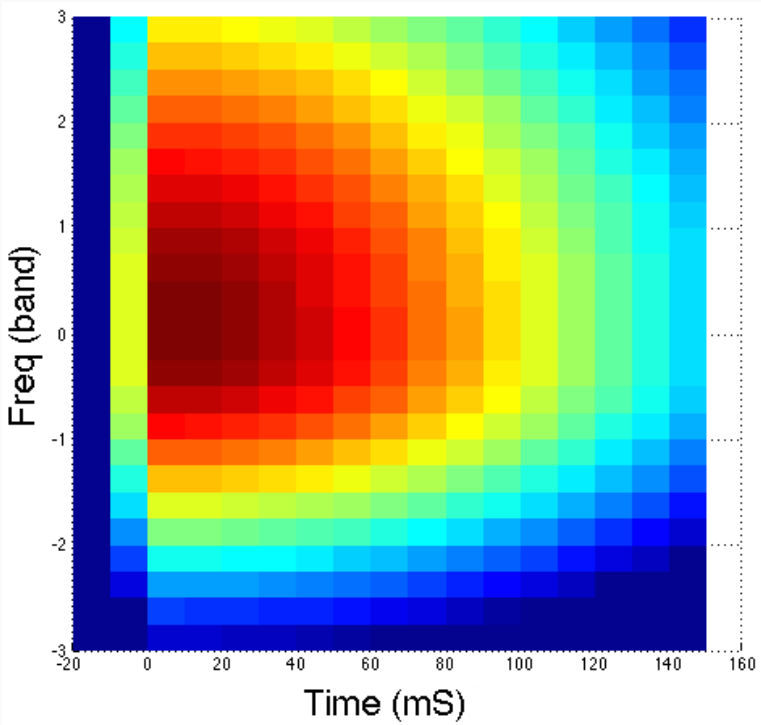
Piecewise-convex SE

- SE for a single frequency-time point is very sharp.
- Smoothness restriction.
- Masking response of a particular time and frequency is the aggregation of many single-point responses (Fastl and Zwicker, 2007).
- A piecewise-convex SE is built by fitting 4 hyperboloid quadrants to contour of the linear SE.

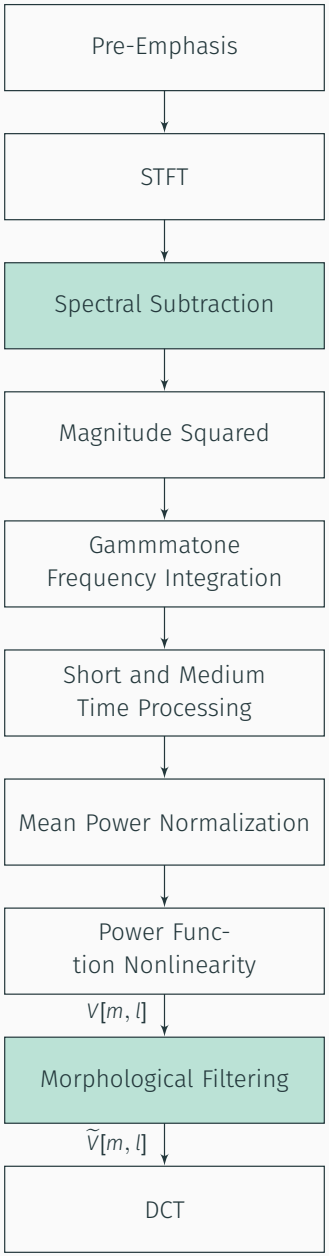
Piecewise-convex SE

- SE for a single frequency-time point is very sharp.
- Smoothness restriction.
- Masking response of a particular time and frequency is the aggregation of many single-point responses (Fastl and Zwicker, 2007).
- A piecewise-convex SE is built by fitting 4 hyperboloid quadrants to contour of the linear SE.

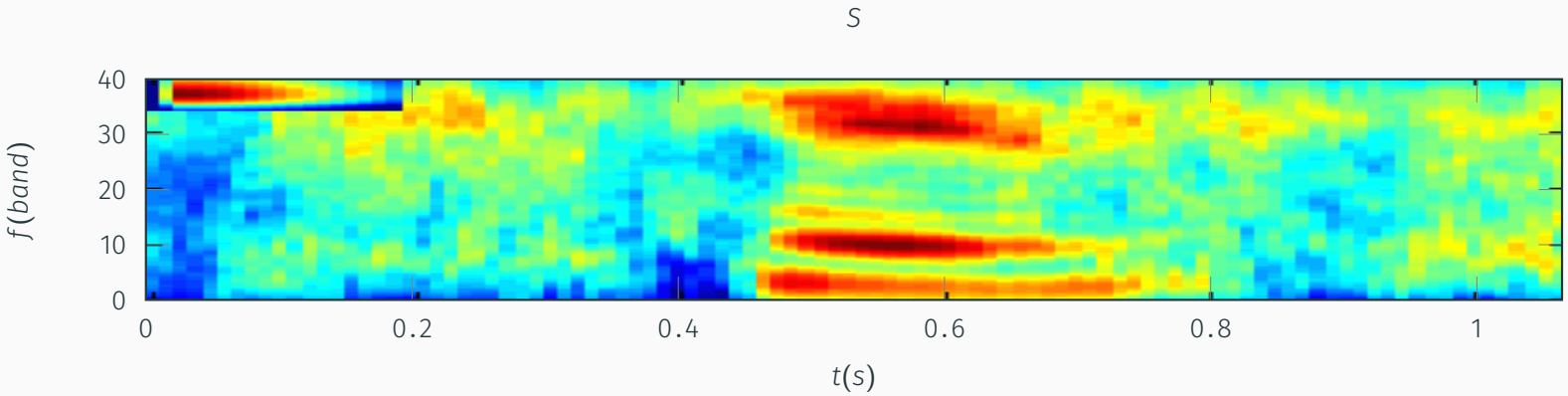




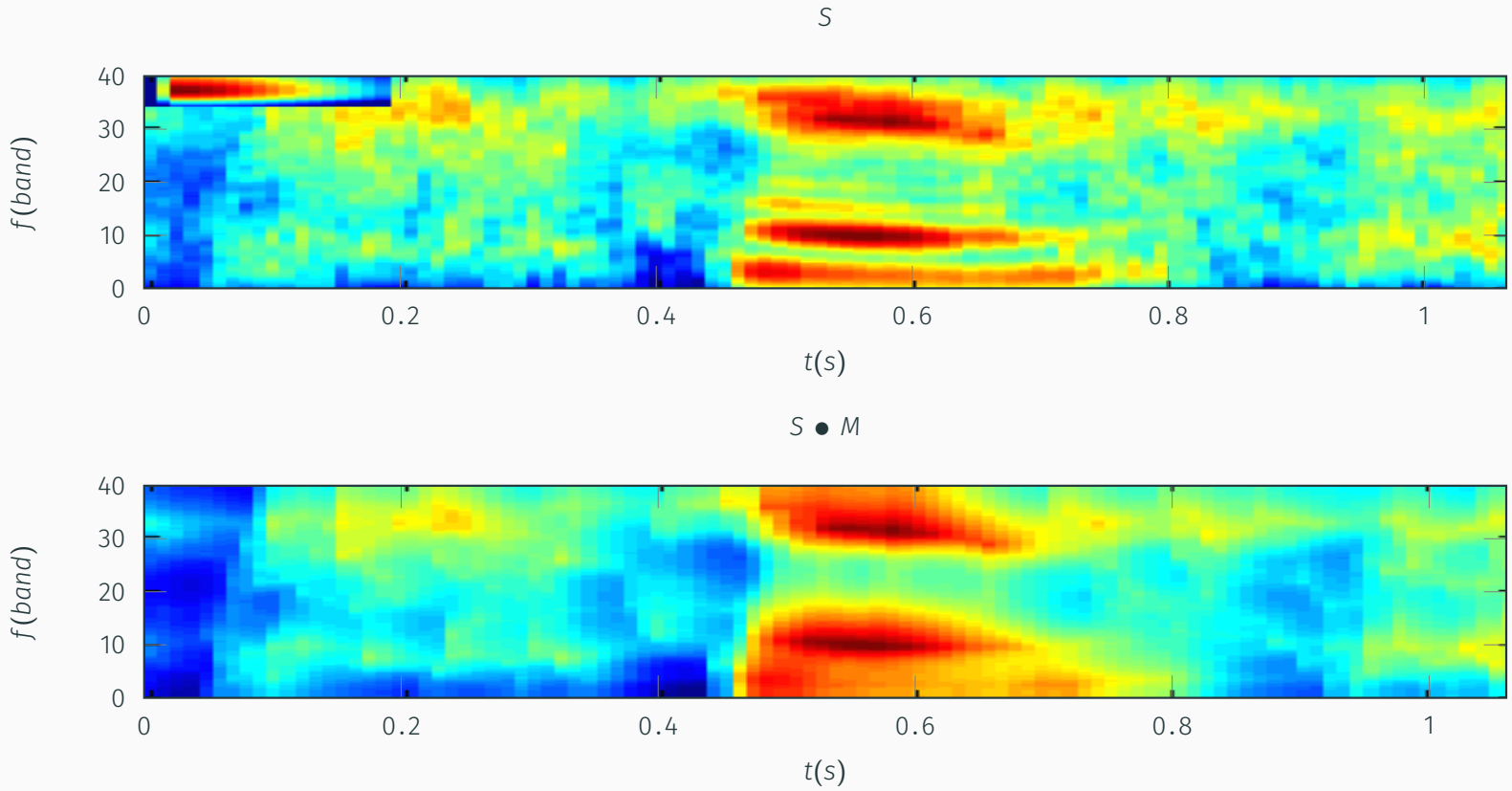
PNCC Spectro-temporal Representation



Filtering Process

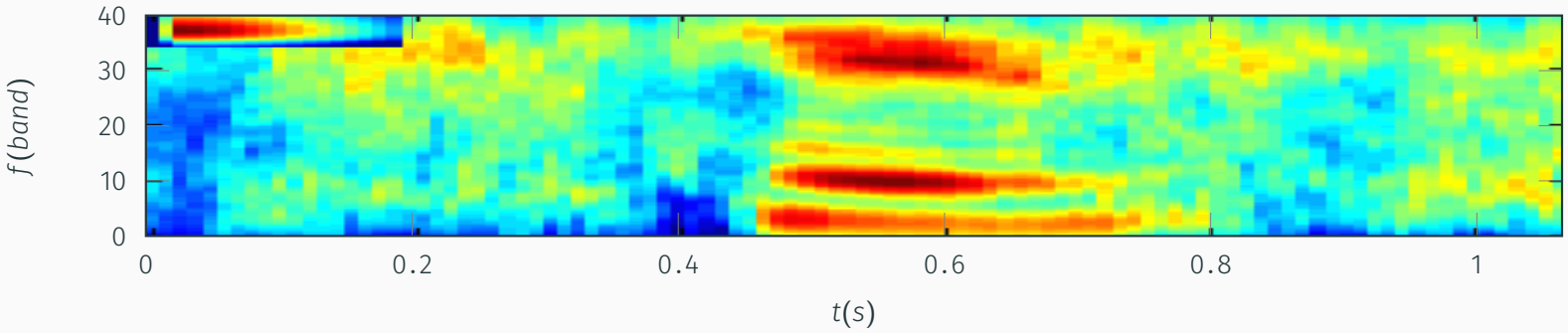


Filtering Process

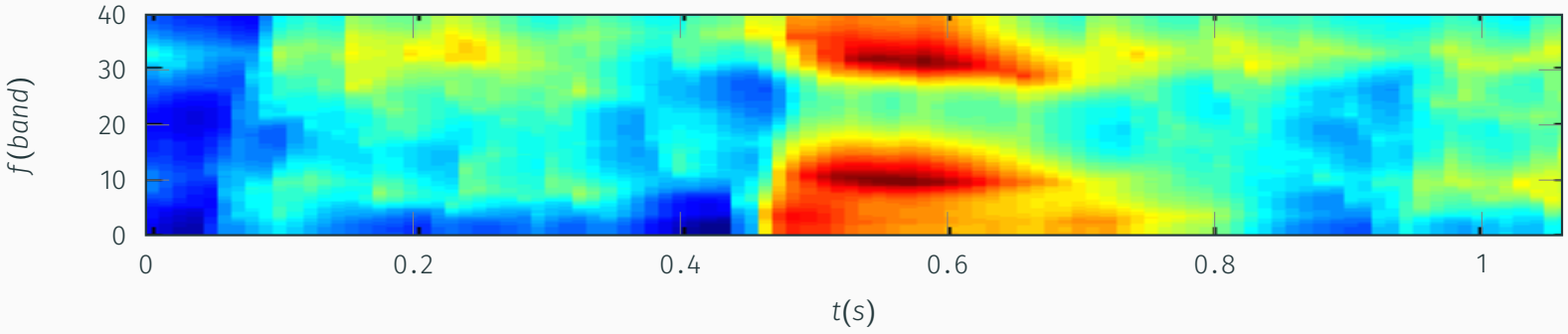


Filtering Process

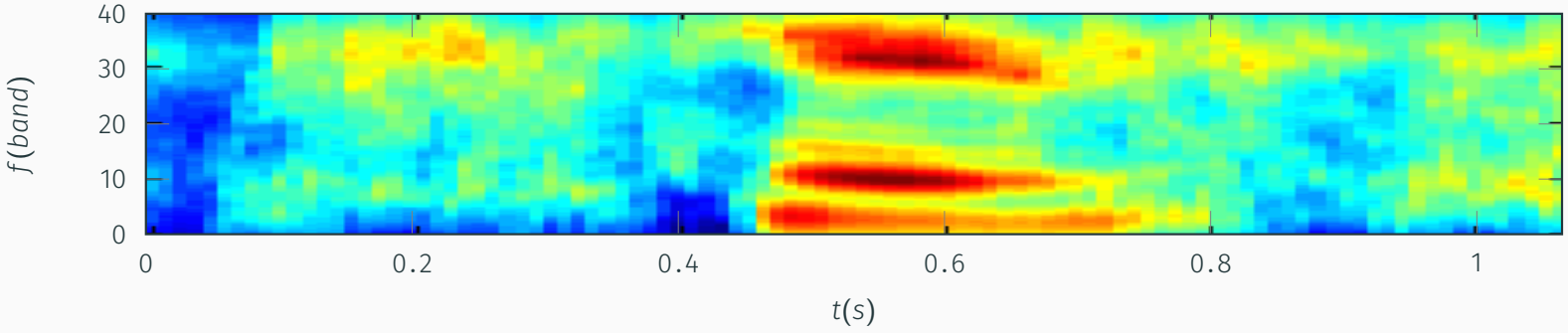
S



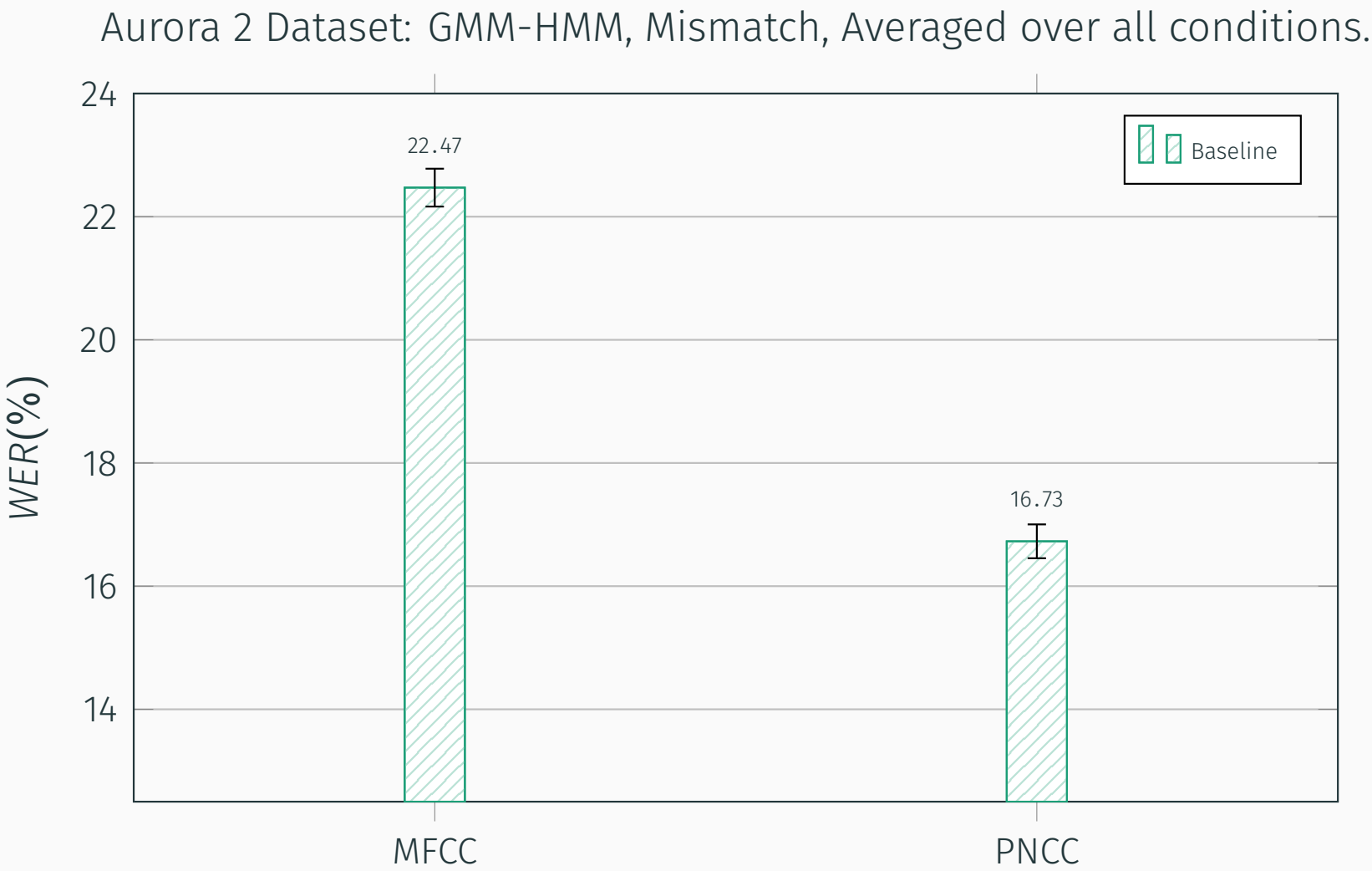
$S \bullet M$



$S' = \lambda S + (1 - \lambda) S \bullet M$ with $\lambda = 0.5$

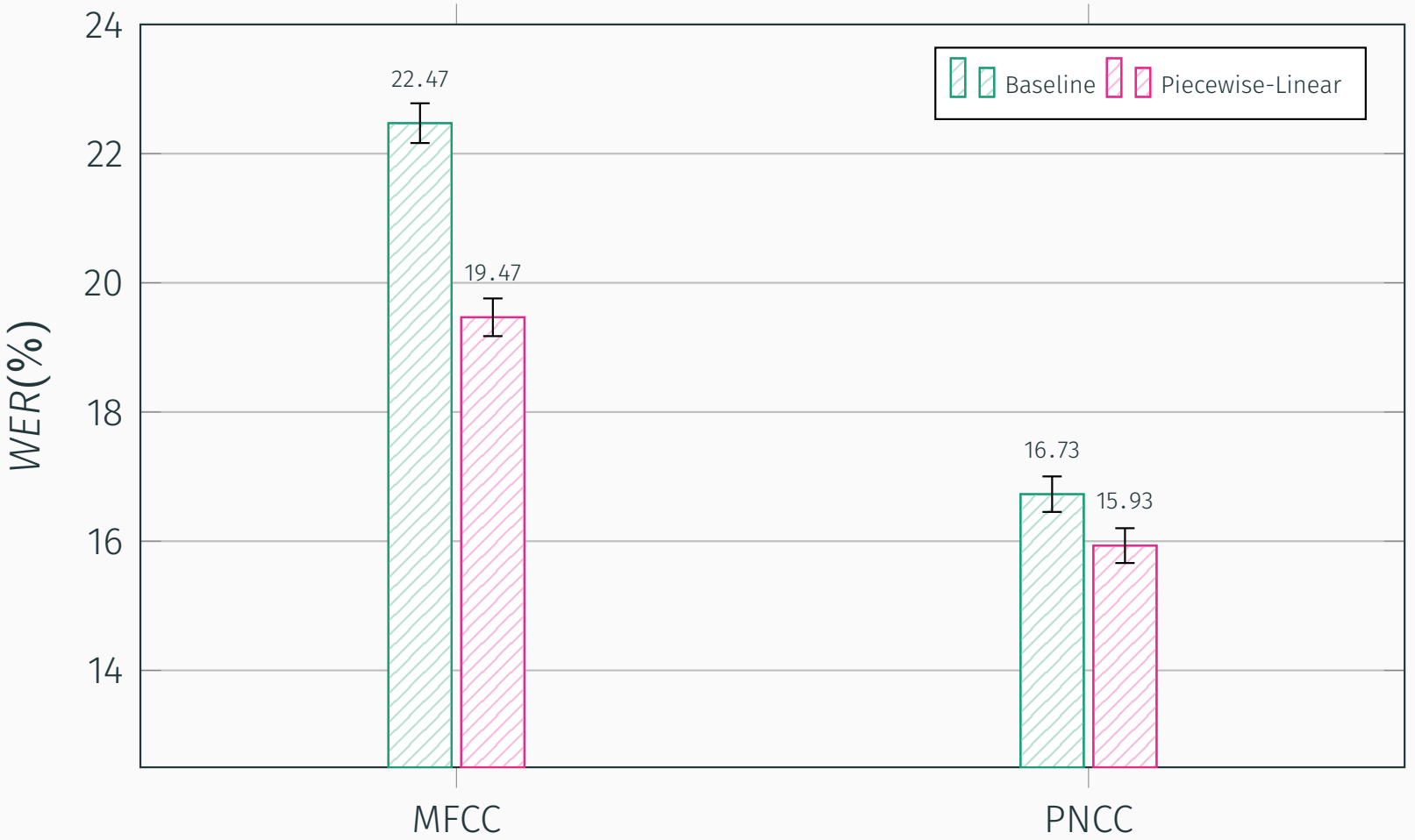


Results: Aurora 2

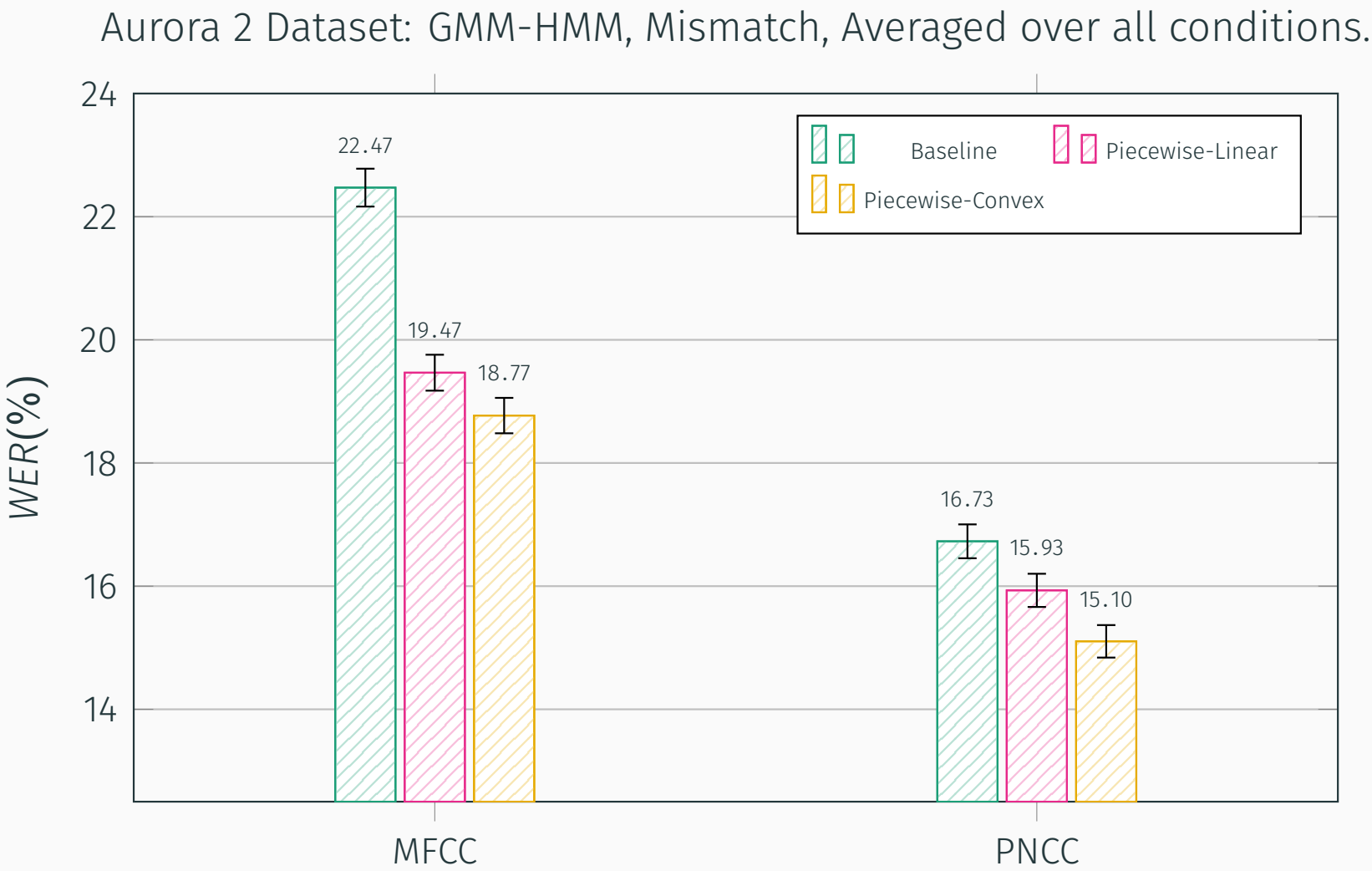


Results: Aurora 2

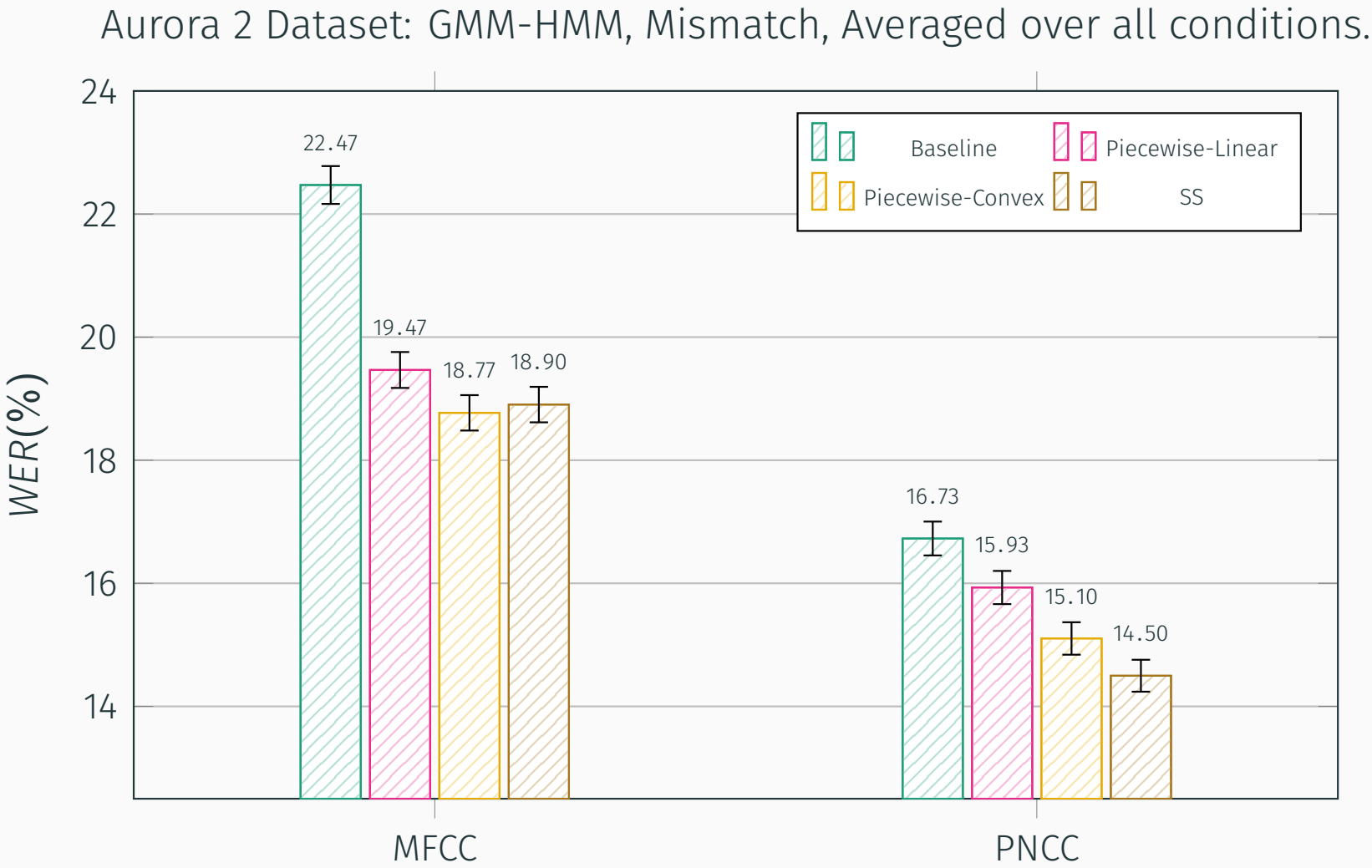
Aurora 2 Dataset: GMM-HMM, Mismatch, Averaged over all conditions.



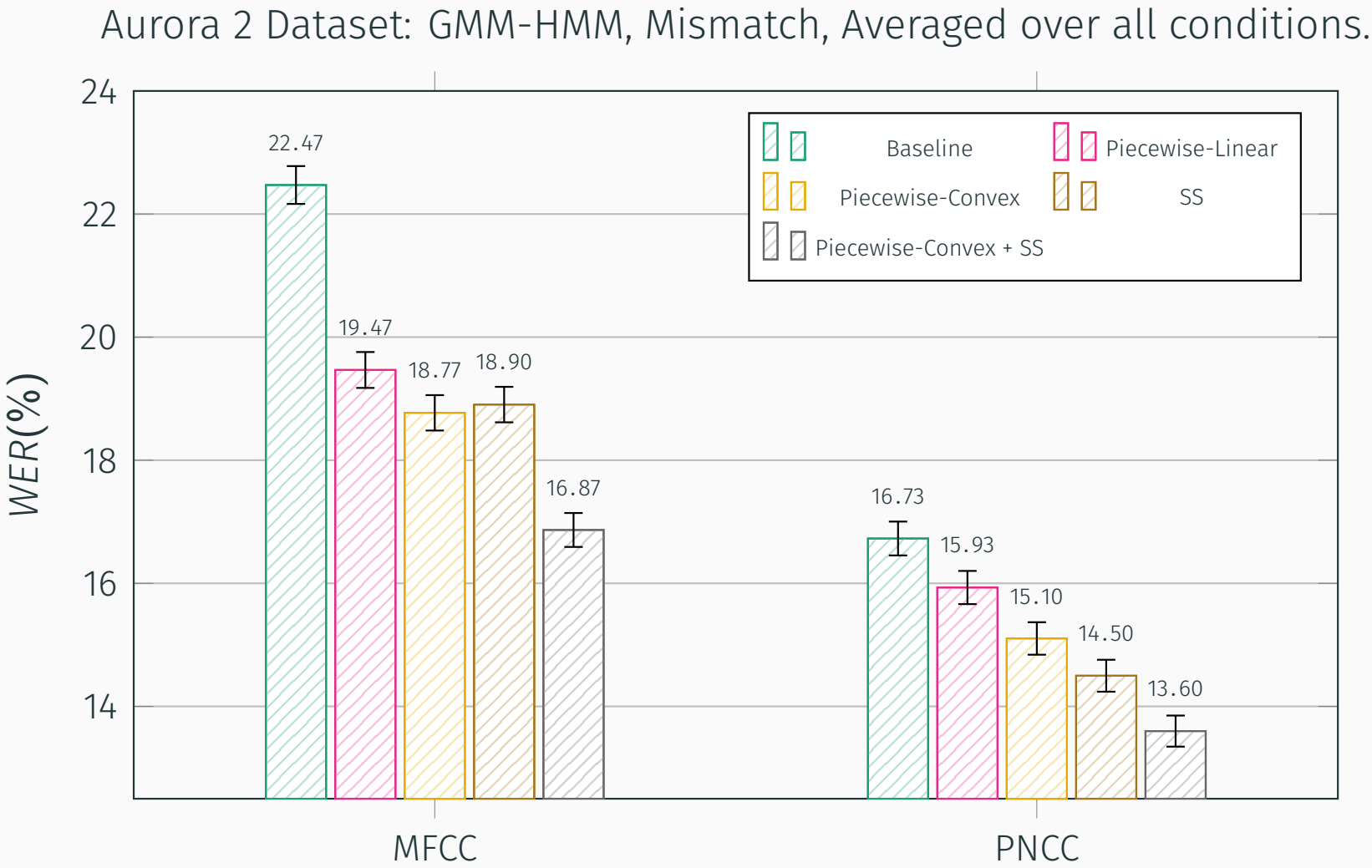
Results: Aurora 2



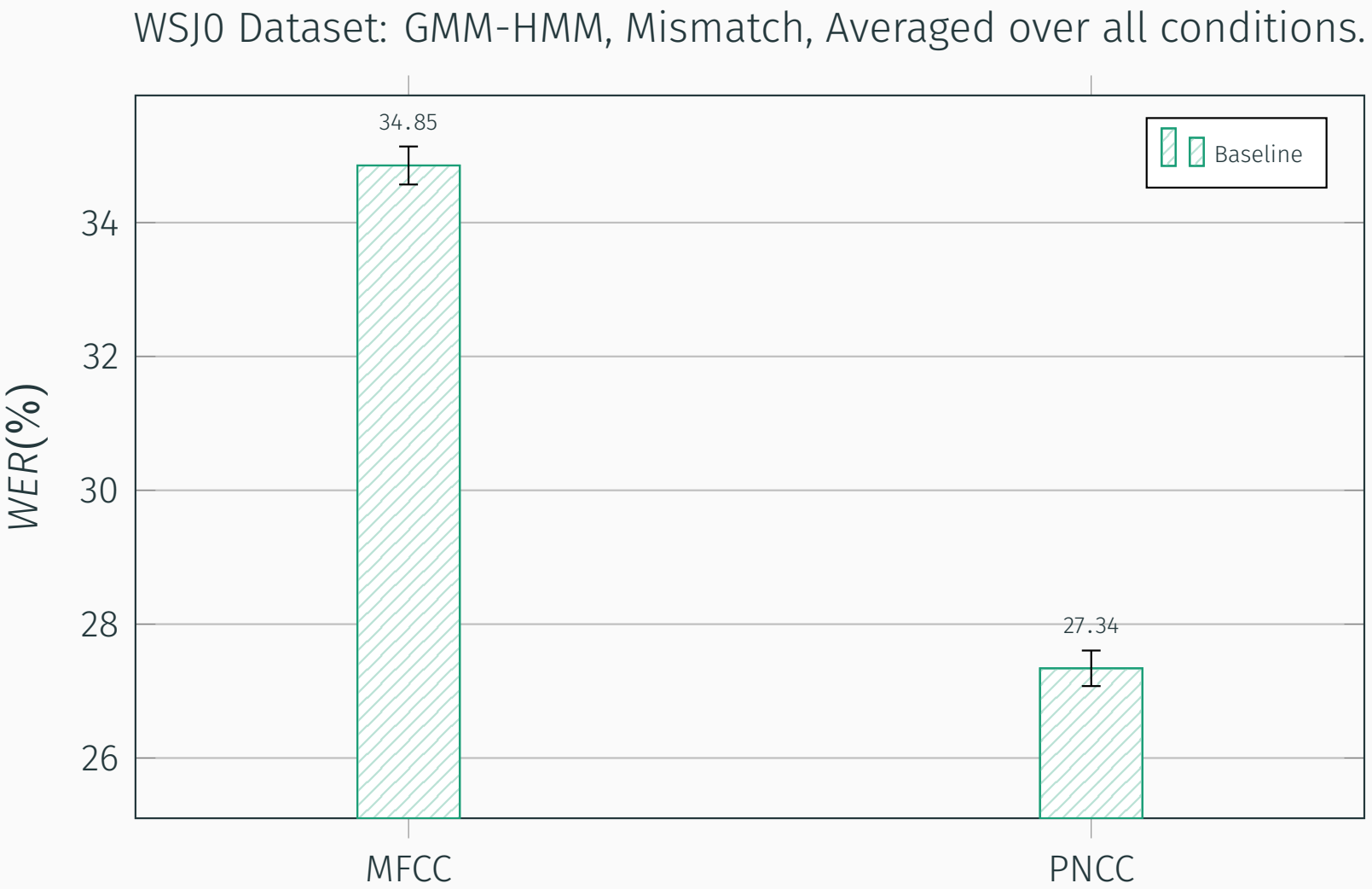
Results: Aurora 2



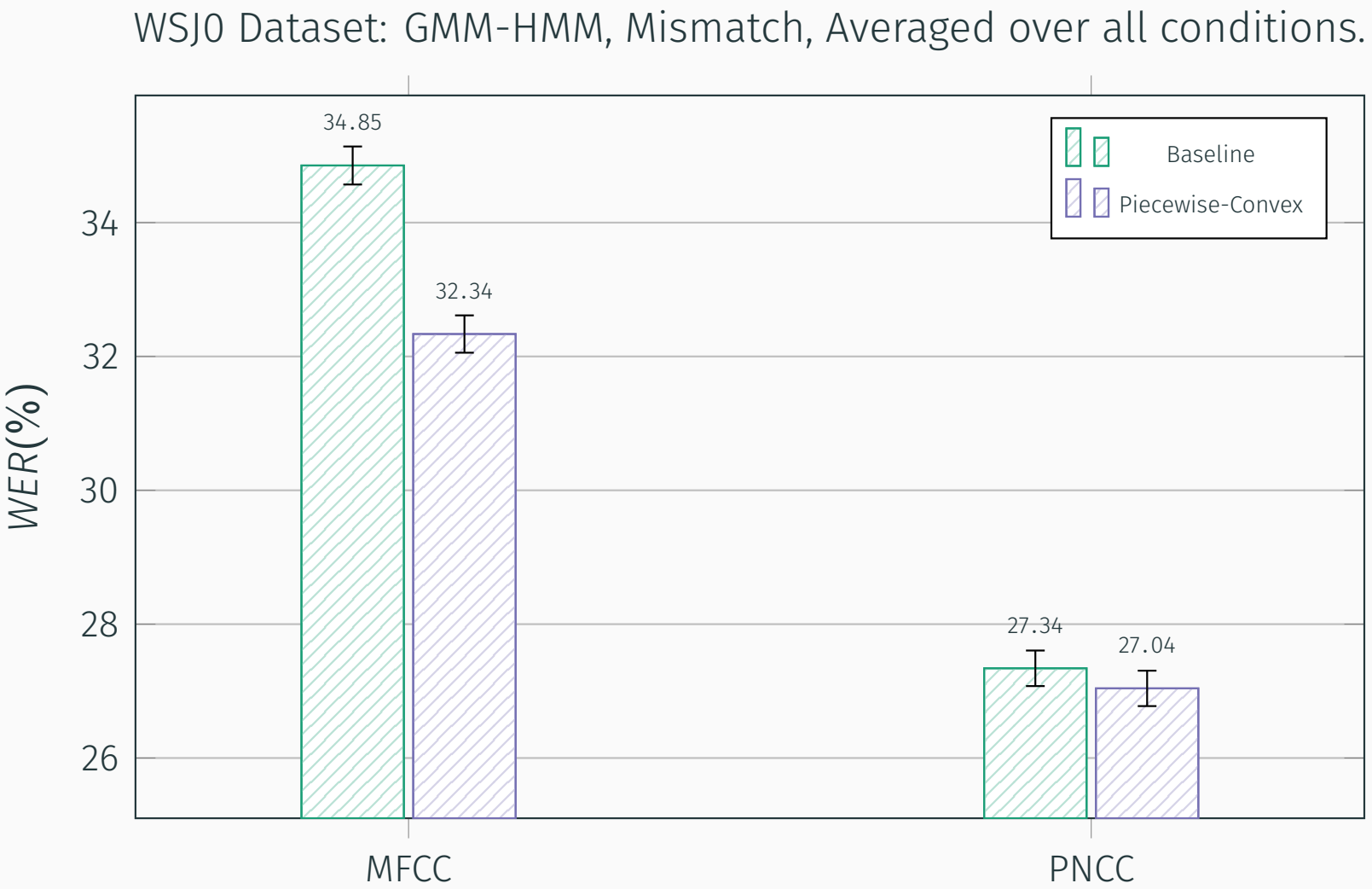
Results: Aurora 2



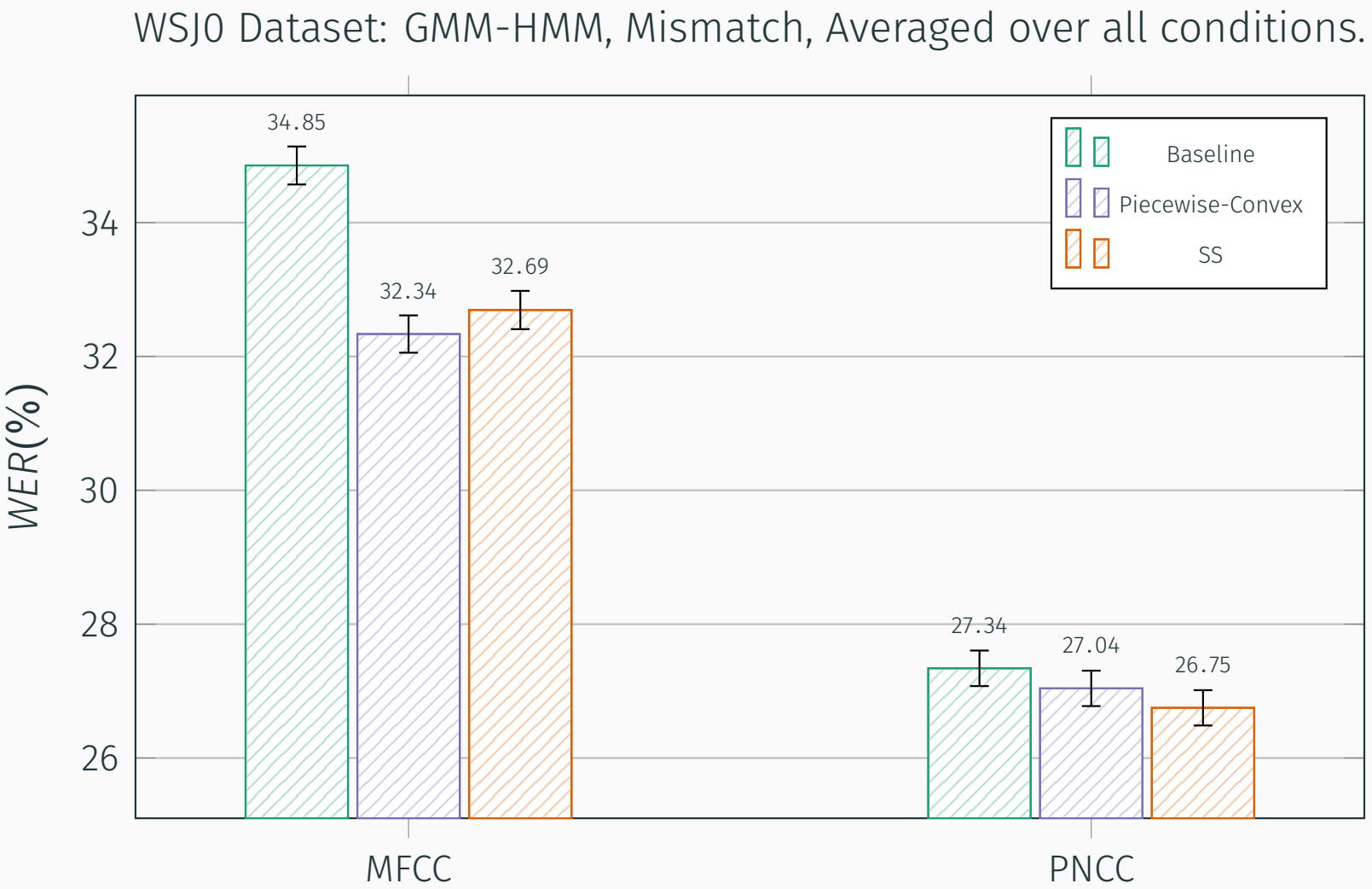
Results: WSJ0



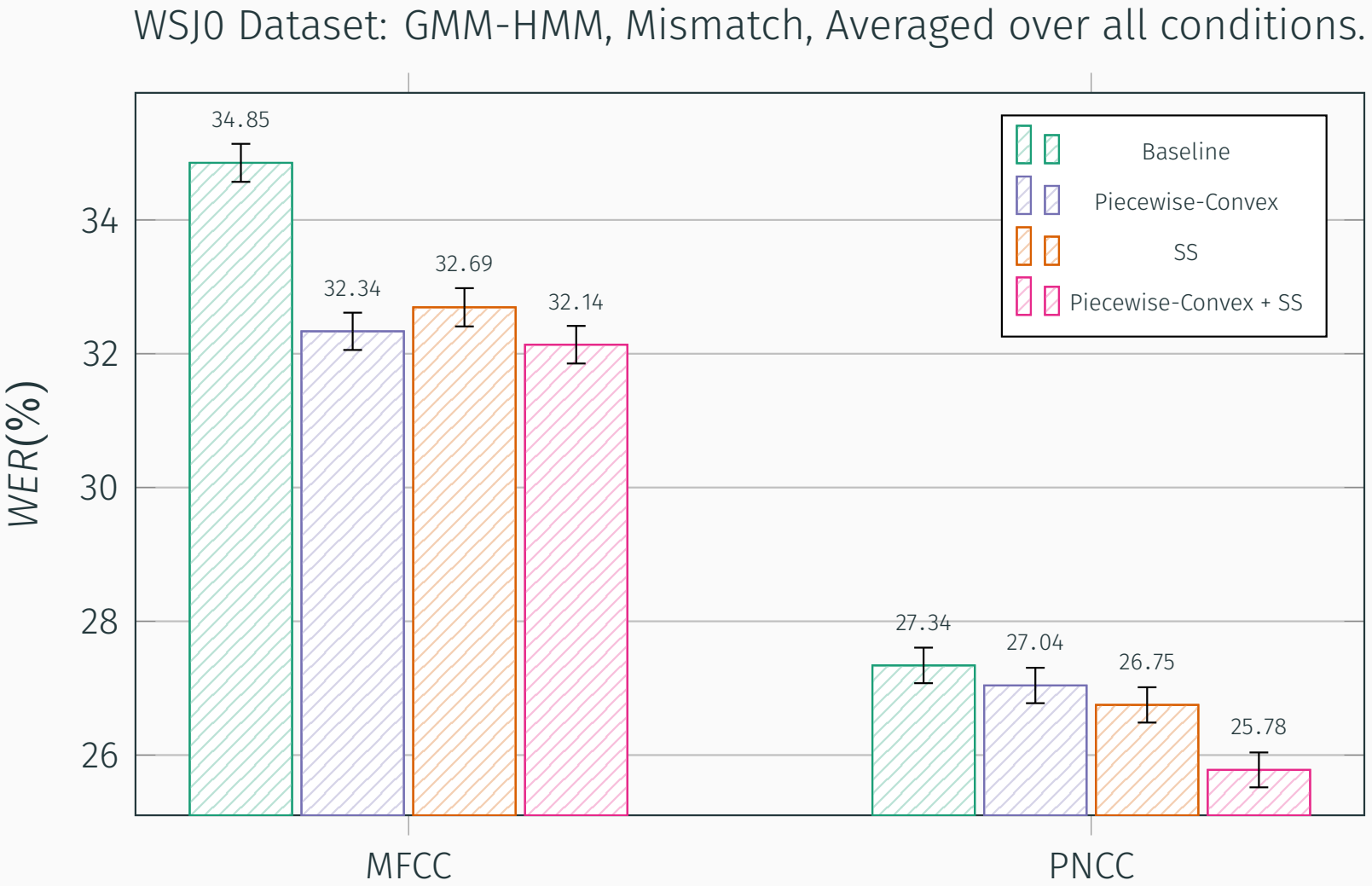
Results: WSJ0



Results: WSJ0



Results: WSJ0



Synchrony-Based Features

Synchrony Effect

- The response of an auditory-nerve fiber roughly follows the shape of the input signal.

From Rose et al. (1971).

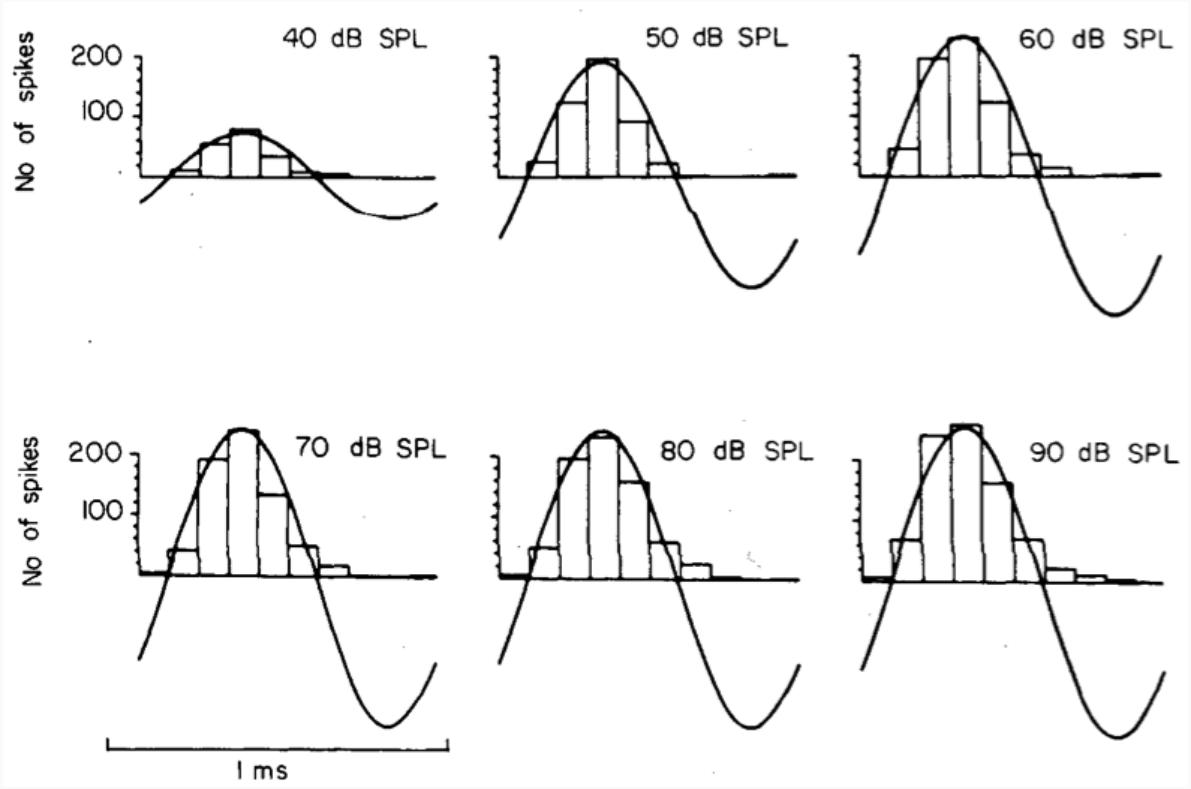
Synchrony Effect

- The response of an auditory-nerve fiber roughly follows the shape of the input signal.

From Rose et al. (1971).

Synchrony Effect

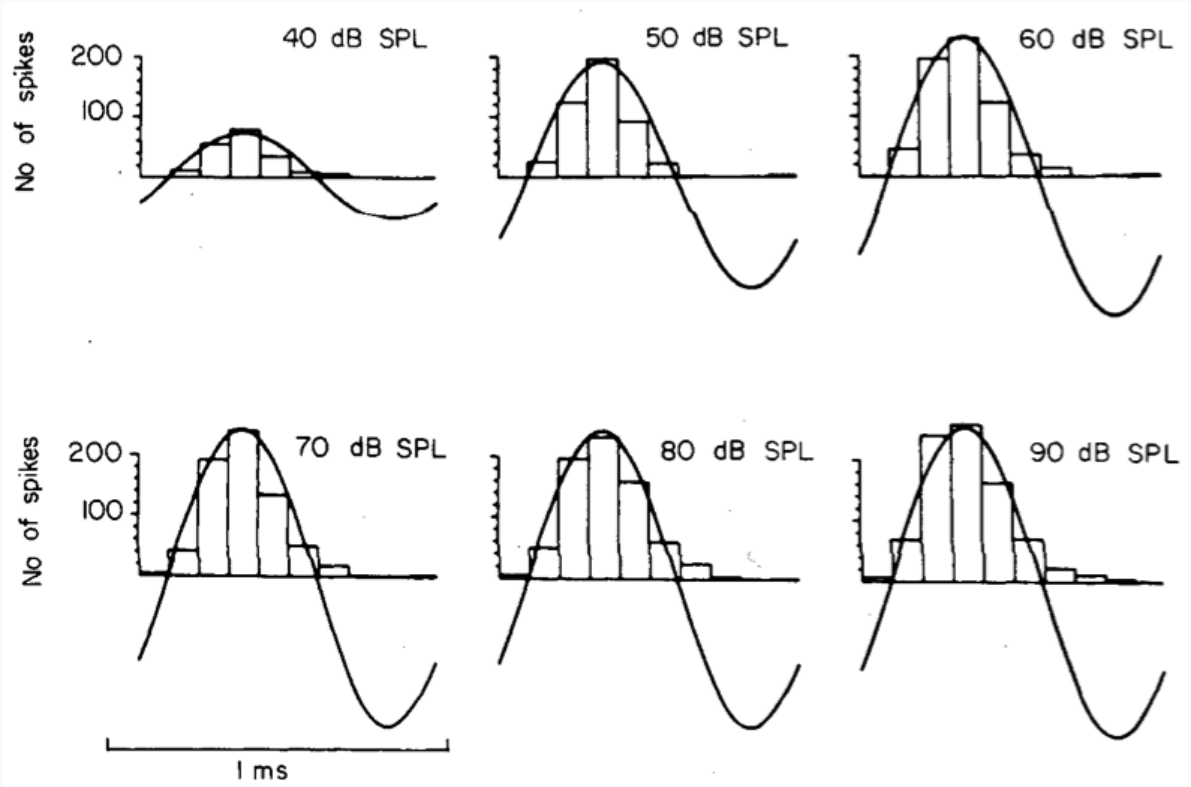
- The response of an auditory-nerve fiber roughly follows the shape of the input signal.



From Rose et al. (1971).

Synchrony Effect

- The response of an auditory-nerve fiber roughly follows the shape of the input signal.

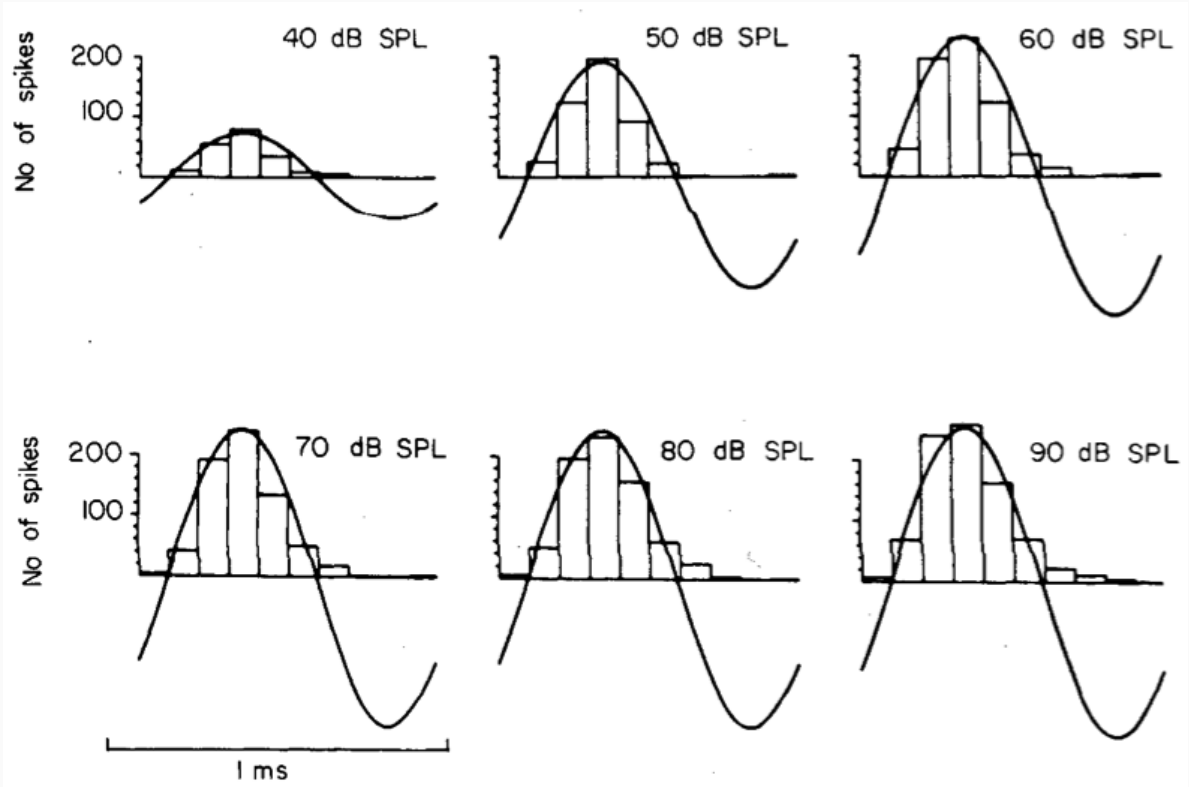


- Compare arrival times of signals to the ears

From Rose et al. (1971).

Synchrony Effect

- The response of an auditory-nerve fiber roughly follows the shape of the input signal.



- Compare arrival times of signals to the ears
- Role in the robust interpretation.

From Rose et al. (1971).

Motivation

- Motivated by physiological findings by Young and Sachs (1979).

Motivation

- Motivated by physiological findings by Young and Sachs (1979).
- Measure the auditory nerve activity using different vowels as stimulus.

Motivation

- Motivated by physiological findings by Young and Sachs (1979).
- Measure the auditory nerve activity using different vowels as stimulus.
- Knowing the presented vowel and the Characteristic Frequency (CF) of the measurement fibers, two responses can be computed.

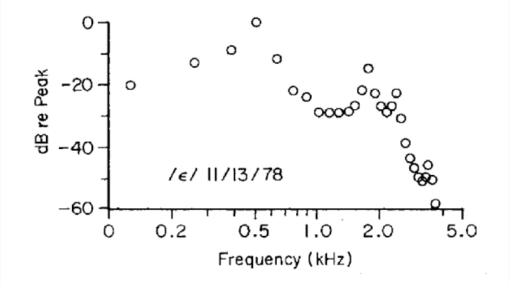
Mean Rate (MR)

MR is simply the mean activity for each fiber over time.

From Young and Sachs (1979).

Mean Rate (MR)

MR is simply the mean activity for each fiber over time.

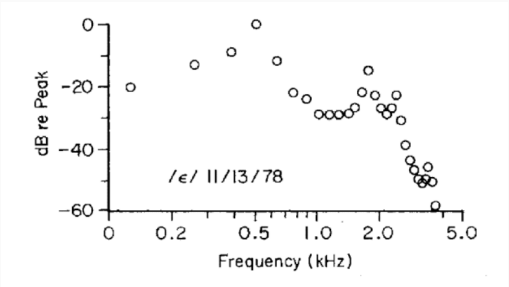


Stimulus

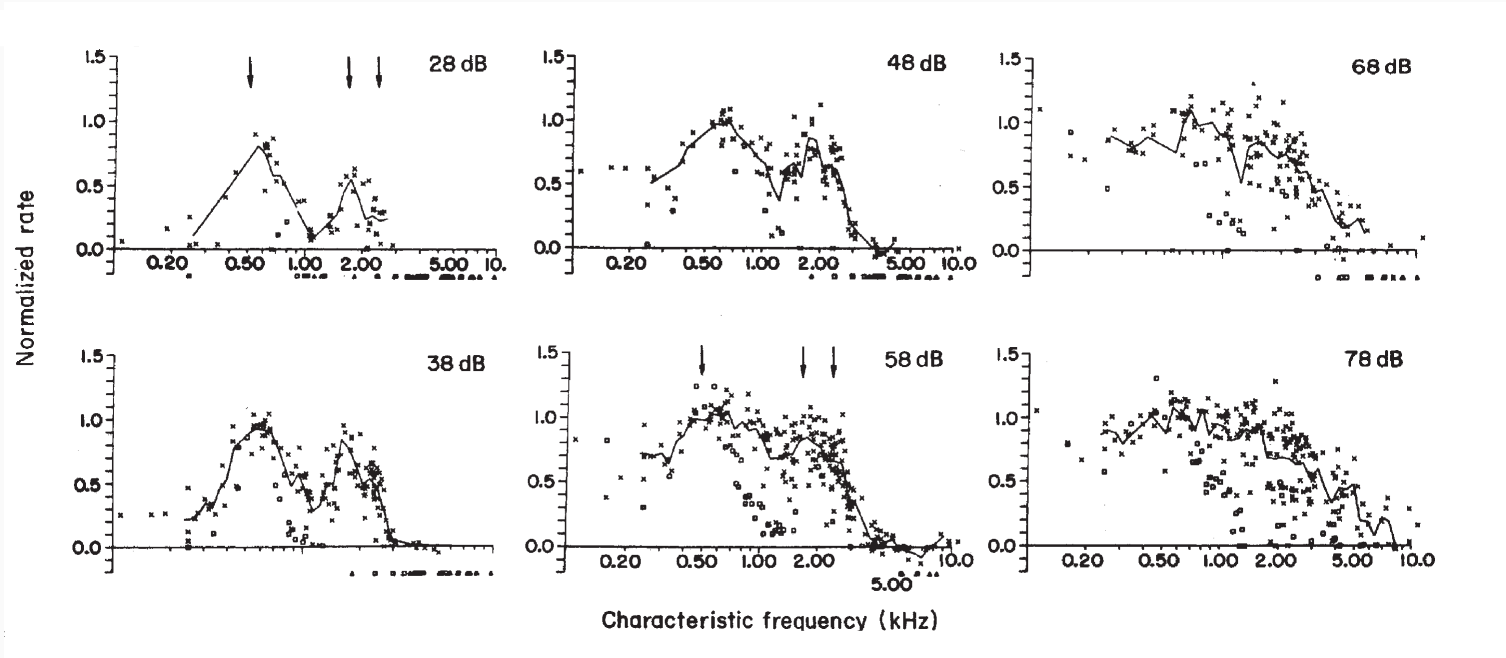
From Young and Sachs (1979).

Mean Rate (MR)

MR is simply the mean activity for each fiber over time.



Stimulus



From Young and Sachs (1979).

MR

Average Localized Synchrony Rate (ALSR)

1. Compute the period histogram, $r_l[n]$, for each fiber (through two fundamental period of the vowel).

Average Localized Synchrony Rate (ALSR)

1. Compute the period histogram, $r_l[n]$, for each fiber (through two fundamental period of the vowel).
2. Over the period histogram apply Fourier transform.

$$r_l[n] = R_0 + 2 \sum_{k=1}^{N/2-1} \mathrm{R_{k,l}} \cos\left(\frac{2\pi k}{N} n + \theta_{k,l}\right)$$

Average Localized Synchrony Rate (ALSR)

1. Compute the period histogram, $r_l[n]$, for each fiber (through two fundamental period of the vowel).
2. Over the period histogram apply Fourier transform.

$$r_l[n] = R_0 + 2 \sum_{k=1}^{N/2-1} \mathrm{R_{k,l}} \cos\left(\frac{2\pi k}{N} n + \theta_{k,l}\right)$$

3. Compute the ALSR:

$$ALSR_k = \frac{1}{M_k} \sum_{l \in C_k} R_{k,l}$$

Average Localized Synchrony Rate (ALSR)

1. Compute the period histogram, $r_l[n]$, for each fiber (through two fundamental period of the vowel).
2. Over the period histogram apply Fourier transform.

$$r_l[n] = R_0 + 2 \sum_{k=1}^{N/2-1} \mathrm{R_{k,l}} \cos\left(\frac{2\pi k}{N} n + \theta_{k,l}\right)$$

3. Compute the ALSR:

$$ALSR_k = \frac{1}{M_k} \sum_{l \in C_k} R_{k,l}$$

Average Localized Synchrony Rate (ALSR)

1. Compute the period histogram, $r_l[n]$, for each fiber (through two fundamental period of the vowel).
2. Over the period histogram apply Fourier transform.

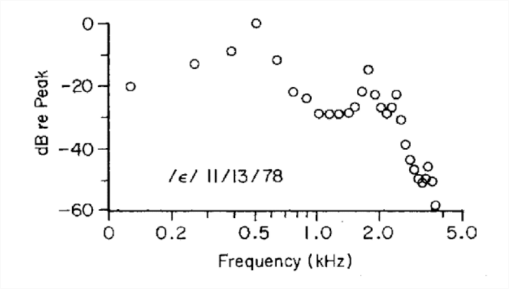
$$r_l[n] = R_0 + 2 \sum_{k=1}^{N/2-1} \mathrm{R_{k,l}} \cos\left(\frac{2\pi k}{N} n + \theta_{k,l}\right)$$

3. Compute the ALSR:

$$ALSR_k = \frac{1}{M_k} \sum_{l \in C_k} R_{k,l}$$

Describes how the auditory nerve activity is synchronized with the nearest harmonic of the fundamental frequency of the vowel.

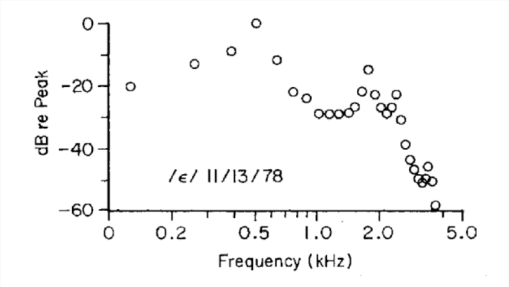
Average Localized Synchrony Rate (ALSR)



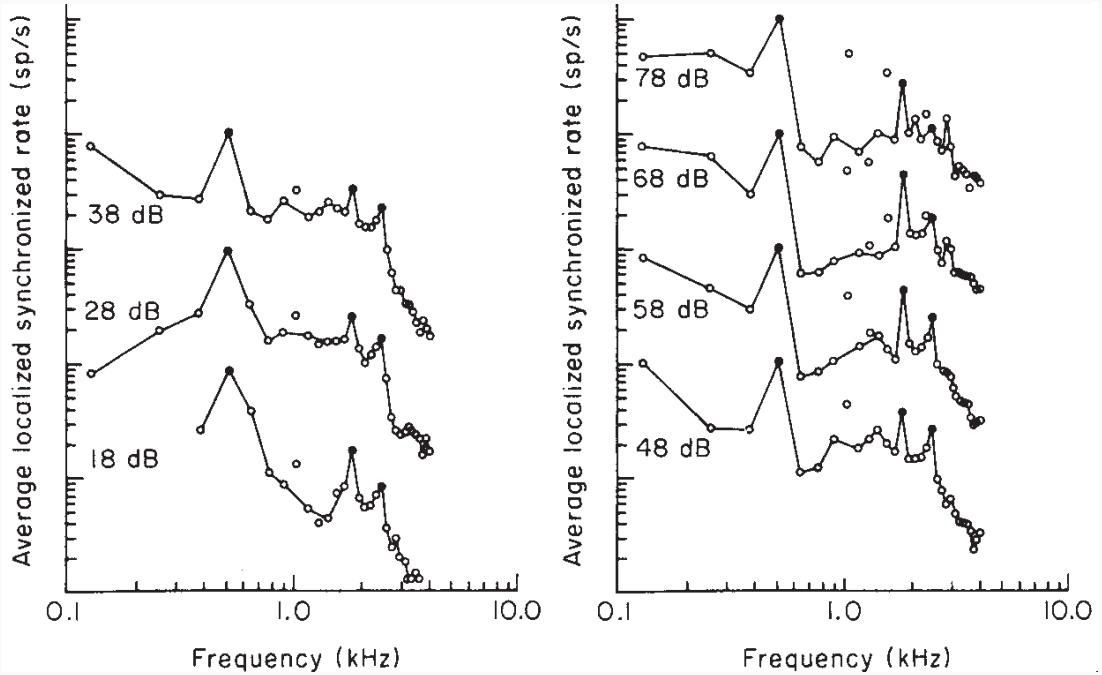
Stimulus

From Young and Sachs (1979).

Average Localized Synchrony Rate (ALSR)



Stimulus



ALSR

From Young and Sachs (1979).

Synchrony Features Motivation

- ALSR is much more robust to changes in intensity than MR.

Synchrony Features Motivation

- ALSR is much more robust to changes in intensity than MR.
- Potentially robust to other types of signal variability.

Synchrony Features Motivation

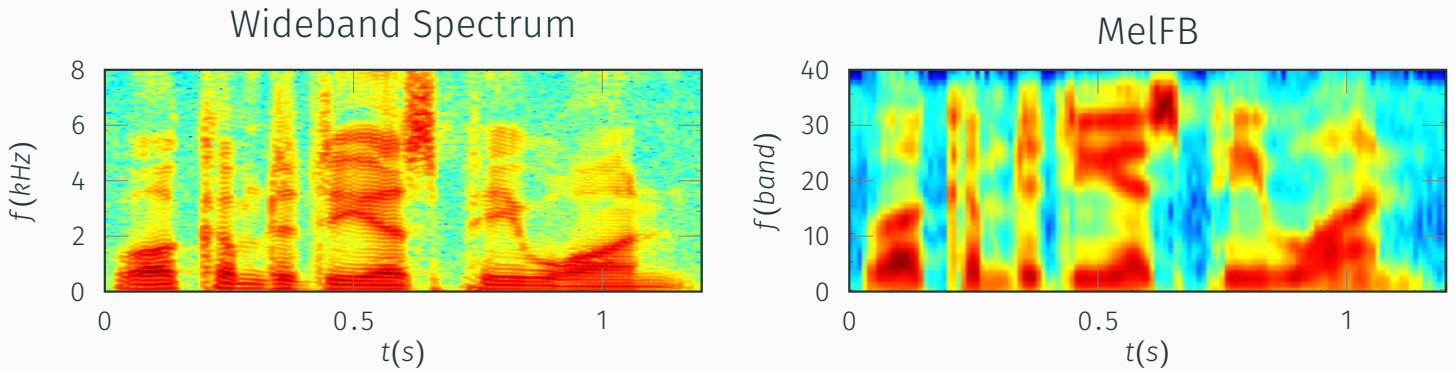
- ALSR is much more robust to changes in intensity than MR.
- Potentially robust to other types of signal variability.
- Conventional feature extraction is based on short-time energy in each frequency band.

Synchrony Features Motivation

- ALSR is much more robust to changes in intensity than MR.
- Potentially robust to other types of signal variability.
- Conventional feature extraction is based on short-time energy in each frequency band.

Synchrony Features Motivation

- ALSR is much more robust to changes in intensity than MR.
- Potentially robust to other types of signal variability.
- Conventional feature extraction is based on short-time energy in each frequency band.



Synchrony Features

- Two approaches:

Synchrony Features

- Two approaches:
    1. Modify the ALSR.

Synchrony Features

- Two approaches:
  1. Modify the ALSR.
  2. Generalized Synchrony Detector (GSD).

Synchrony Features

- Two approaches:
  1. Modify the ALSR.
  2. Generalized Synchrony Detector (GSD).
- Problem: we need an auditory nerve model.

Synchrony Features

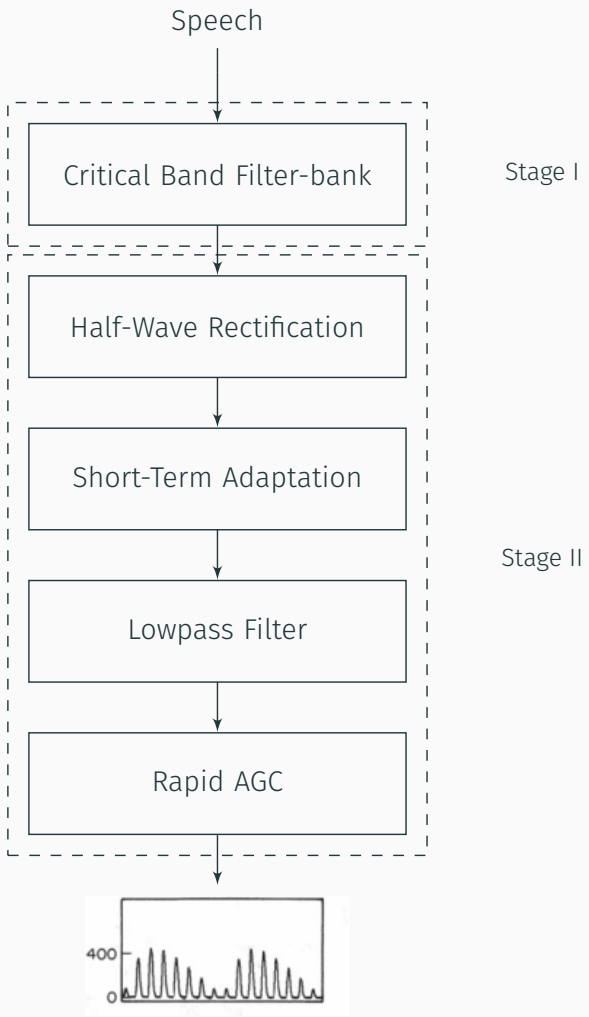
- Two approaches:
  1. Modify the ALSR.
  2. Generalized Synchrony Detector (GSD).
- Problem: we need an auditory nerve model.
- Simplification of the Seneff (1990) model.

Synchrony Features

- Two approaches:
  1. Modify the ALSR.
  2. Generalized Synchrony Detector (GSD).
- Problem: we need an auditory nerve model.
- Simplification of the Seneff (1990) model.

Synchrony Features

- Two approaches:
  1. Modify the ALSR.
  2. Generalized Synchrony Detector (GSD).
- Problem: we need an auditory nerve model.
- Simplification of the Seneff (1990) model.



Modified ALSR (MALSR)

- We take the fundamental frequency of the vowel as the central frequency of the filter for which we are computing the ALSR.

Modified ALSR (MALSR)

- We take the fundamental frequency of the vowel as the central frequency of the filter for which we are computing the ALSR.
- Short-time Fourier transform of the outputs of the auditory filters.

Modified ALSR (MALSR)

- We take the fundamental frequency of the vowel as the central frequency of the filter for which we are computing the ALSR.
- Short-time Fourier transform of the outputs of the auditory filters.
- Averaged across channels.

Modified ALSR (MALSR)

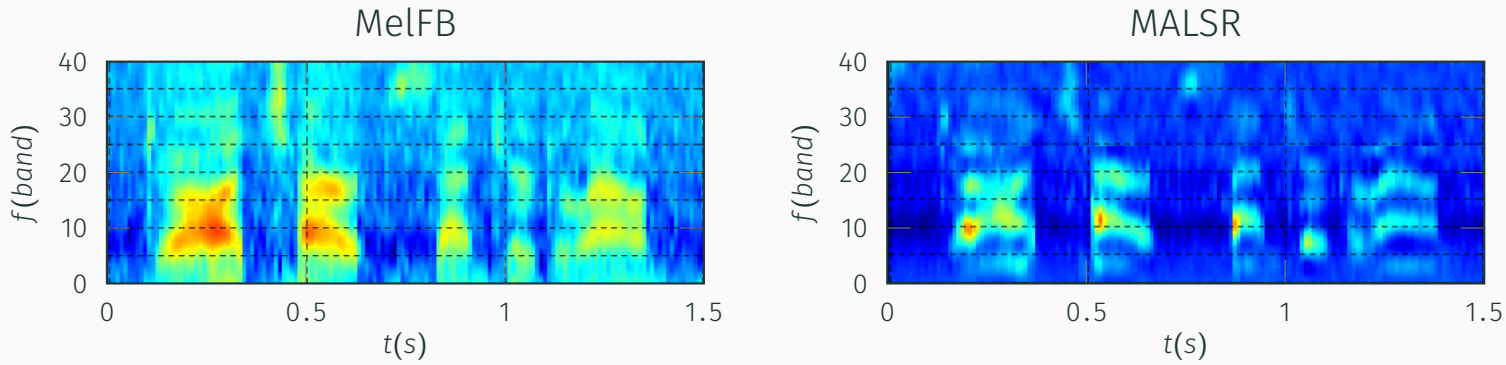
- We take the fundamental frequency of the vowel as the central frequency of the filter for which we are computing the ALSR.
- Short-time Fourier transform of the outputs of the auditory filters.
- Averaged across channels.
- Synchrony response disappears below 1000 Hz, linear transition between MR and ALSR over 300 to 1200 Hz.

Modified ALSR (MALSR)

- We take the fundamental frequency of the vowel as the central frequency of the filter for which we are computing the ALSR.
- Short-time Fourier transform of the outputs of the auditory filters.
- Averaged across channels.
- Synchrony response disappears below 1000 Hz, linear transition between MR and ALSR over 300 to 1200 Hz.

Modified ALSR (MALSR)

- We take the fundamental frequency of the vowel as the central frequency of the filter for which we are computing the ALSR.
- Short-time Fourier transform of the outputs of the auditory filters.
- Averaged across channels.
- Synchrony response disappears below 1000 Hz, linear transition between MR and ALSR over 300 to 1200 Hz.



White Noise, 10 dB SNR.

Modified GDS (MGSD)

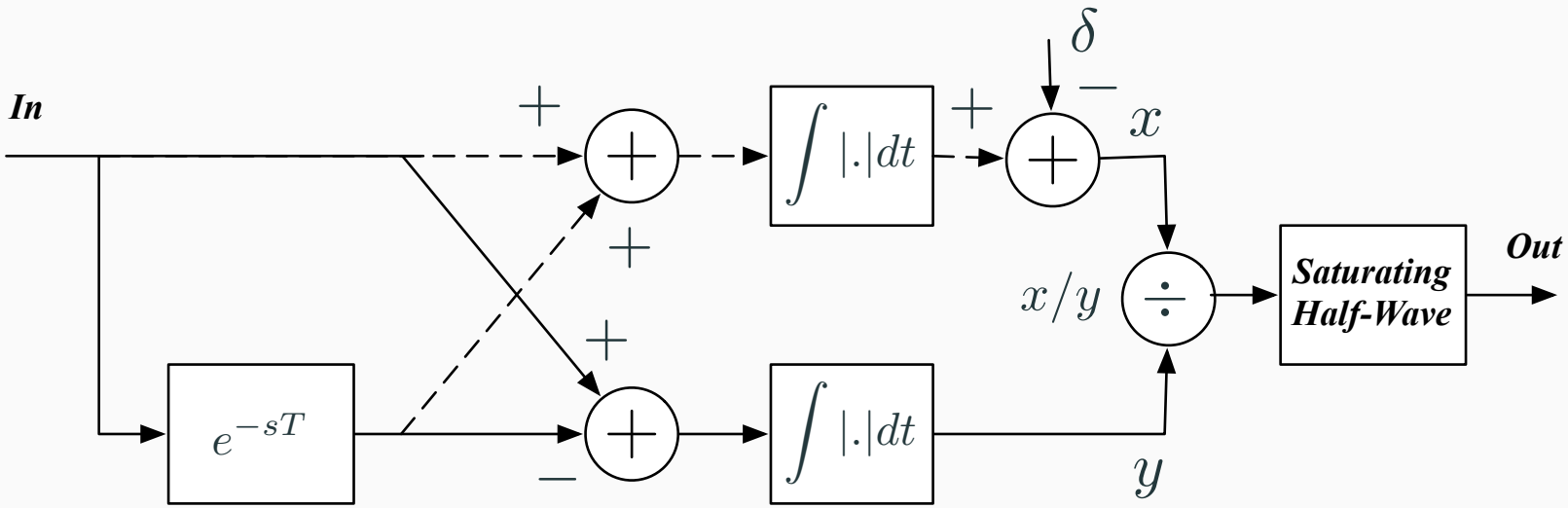
- Based on the Generalized Synchrony Detector (GSD) (Seneff, 1990).

Modified GDS (MGSD)

- Based on the Generalized Synchrony Detector (GSD) (Seneff, 1990).

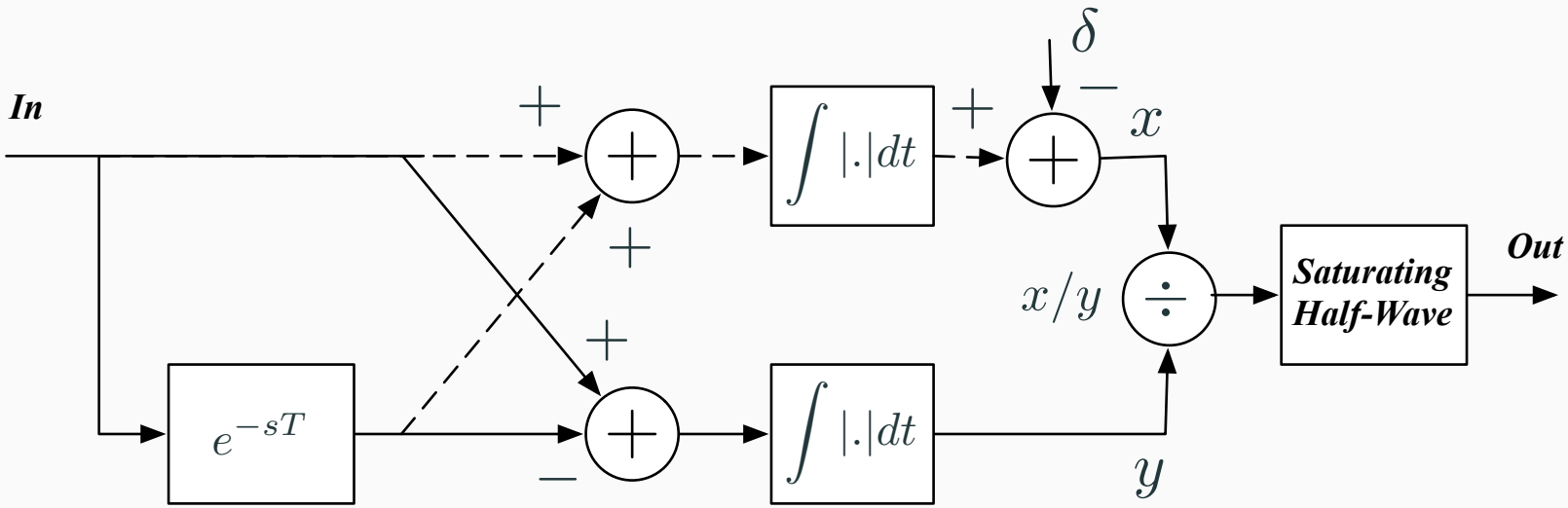
Modified GDS (MGSD)

- Based on the Generalized Synchrony Detector (GSD) (Seneff, 1990).



Modified GDS (MGSD)

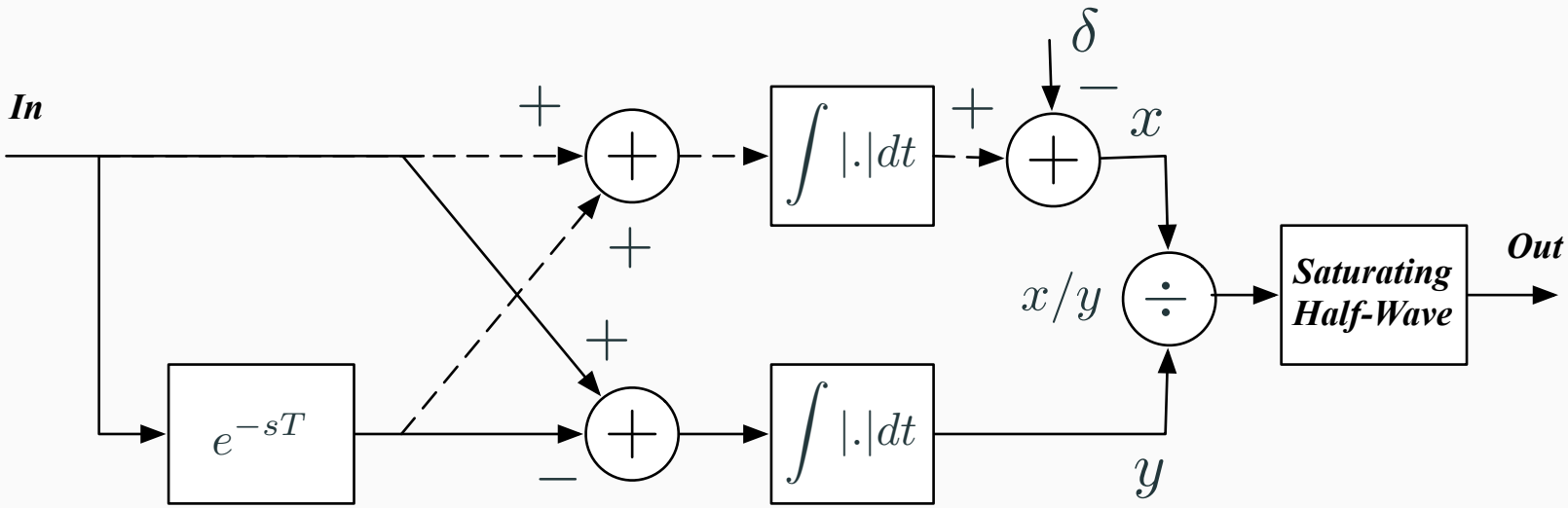
- Based on the Generalized Synchrony Detector (GSD) (Seneff, 1990).



- Time domain.

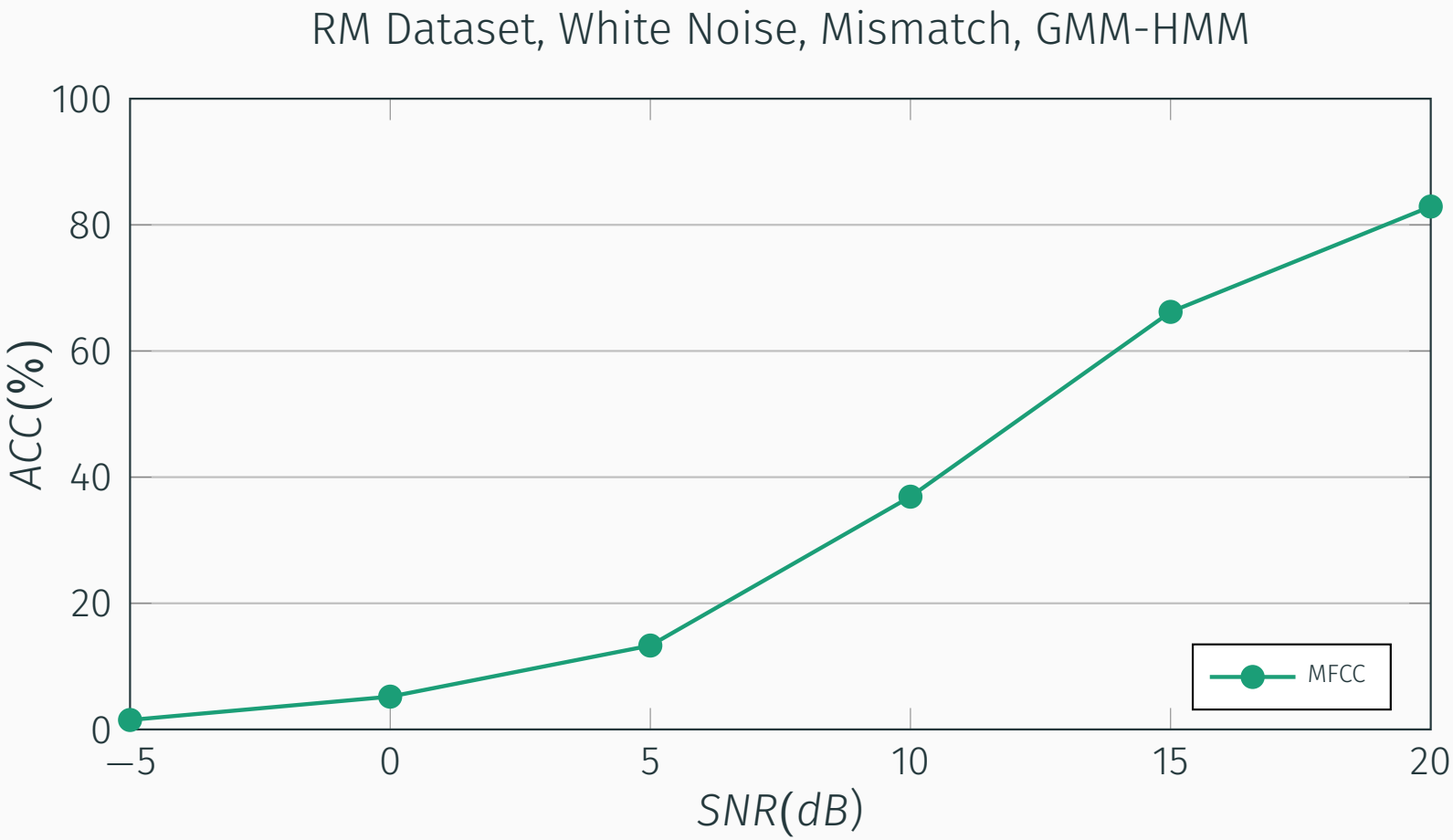
Modified GDS (MGSD)

- Based on the Generalized Synchrony Detector (GSD) (Seneff, 1990).

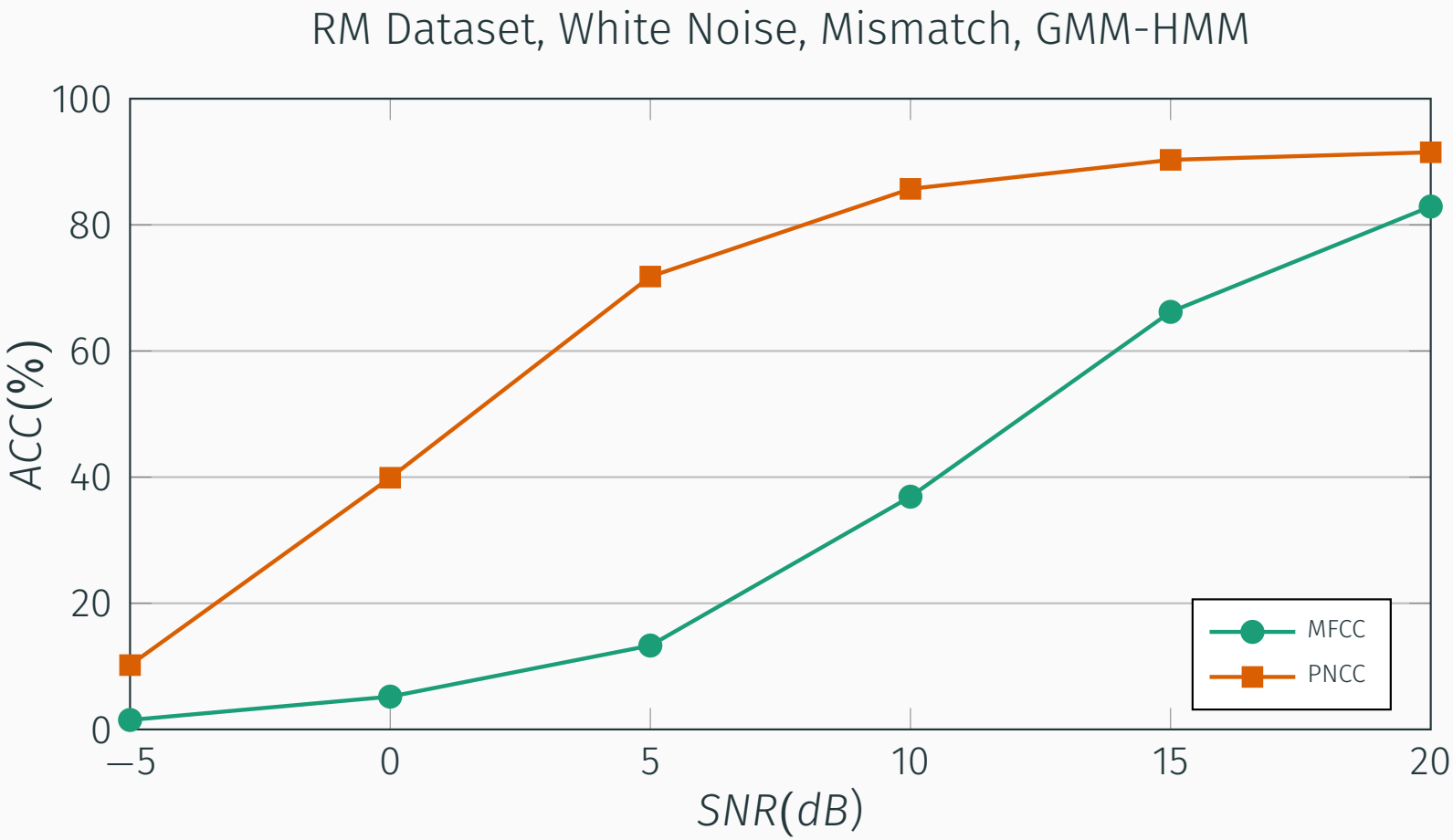


- Time domain.
- Low computational complexity.

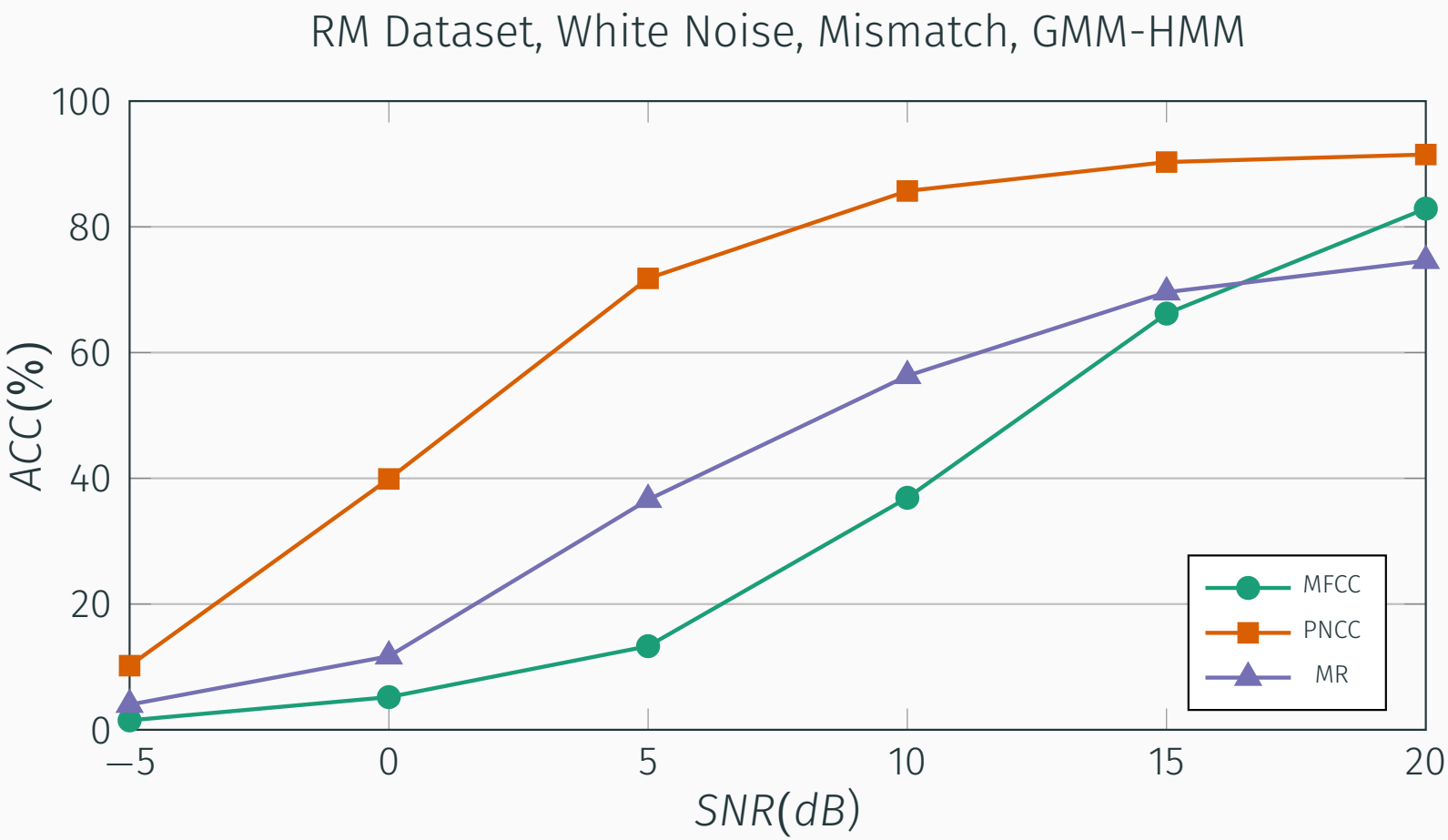
Results: MR, MASLR, MGSD



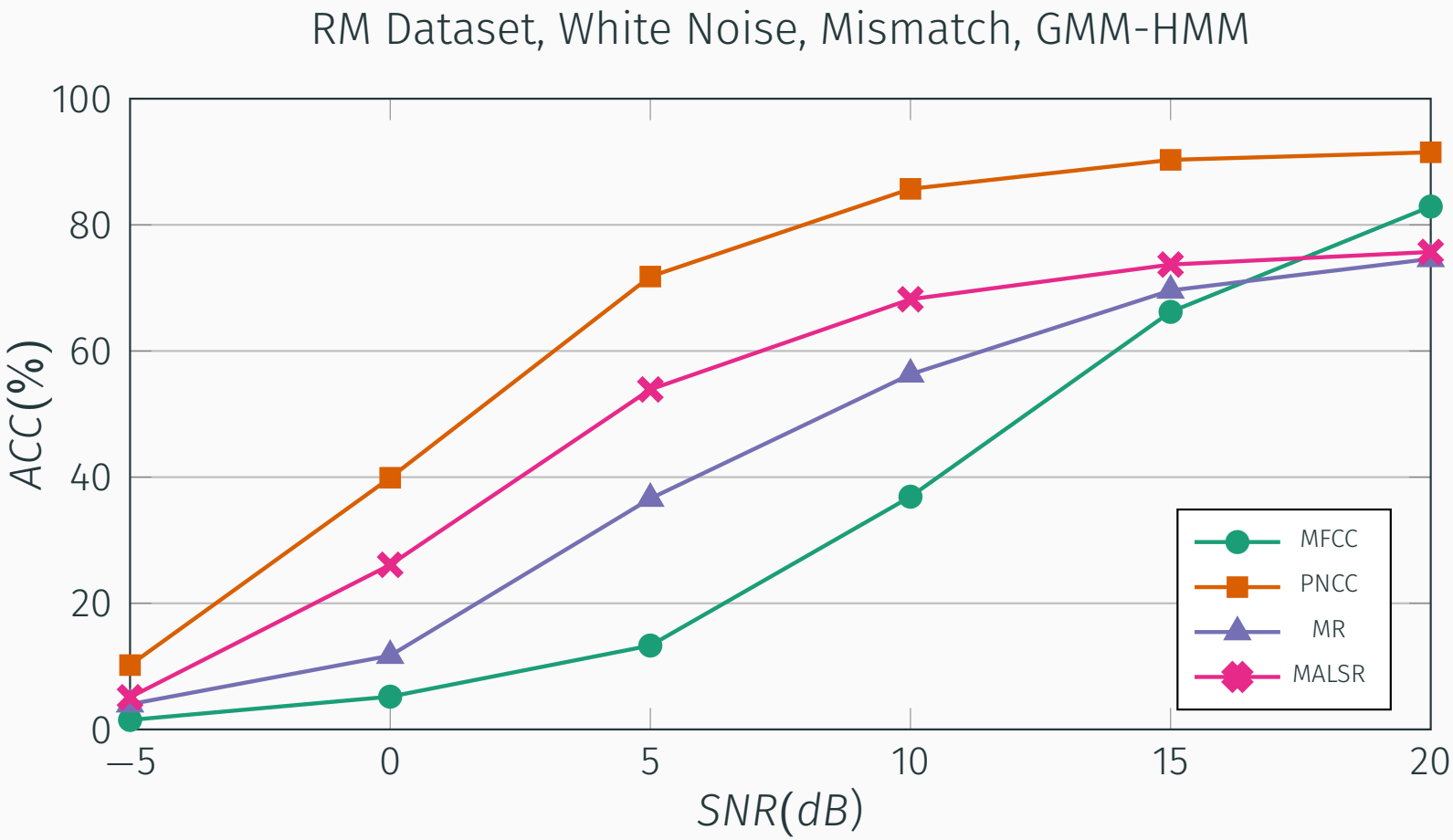
Results: MR, MASLR, MGSD



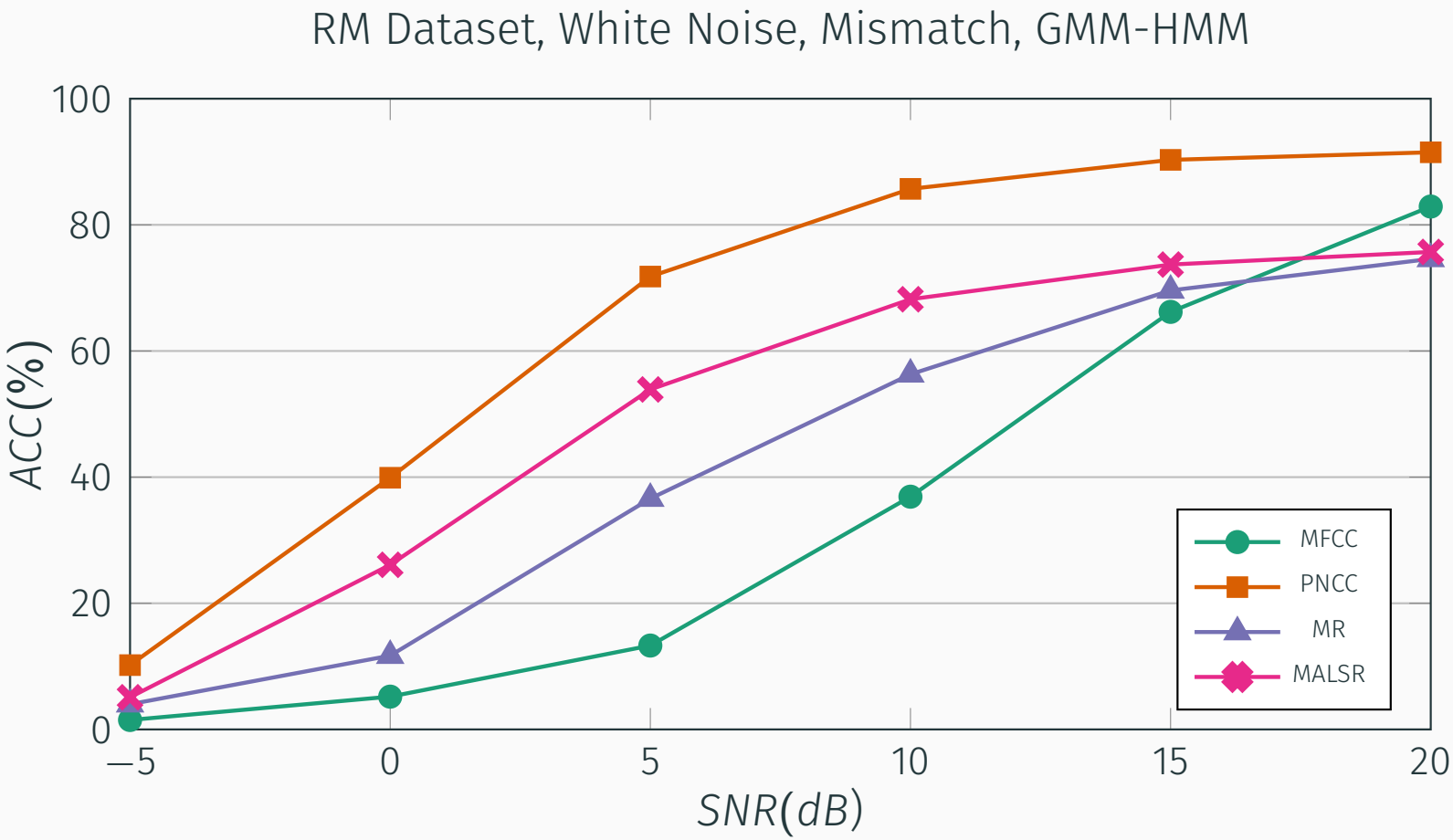
Results: MR, MASLR, MGSD



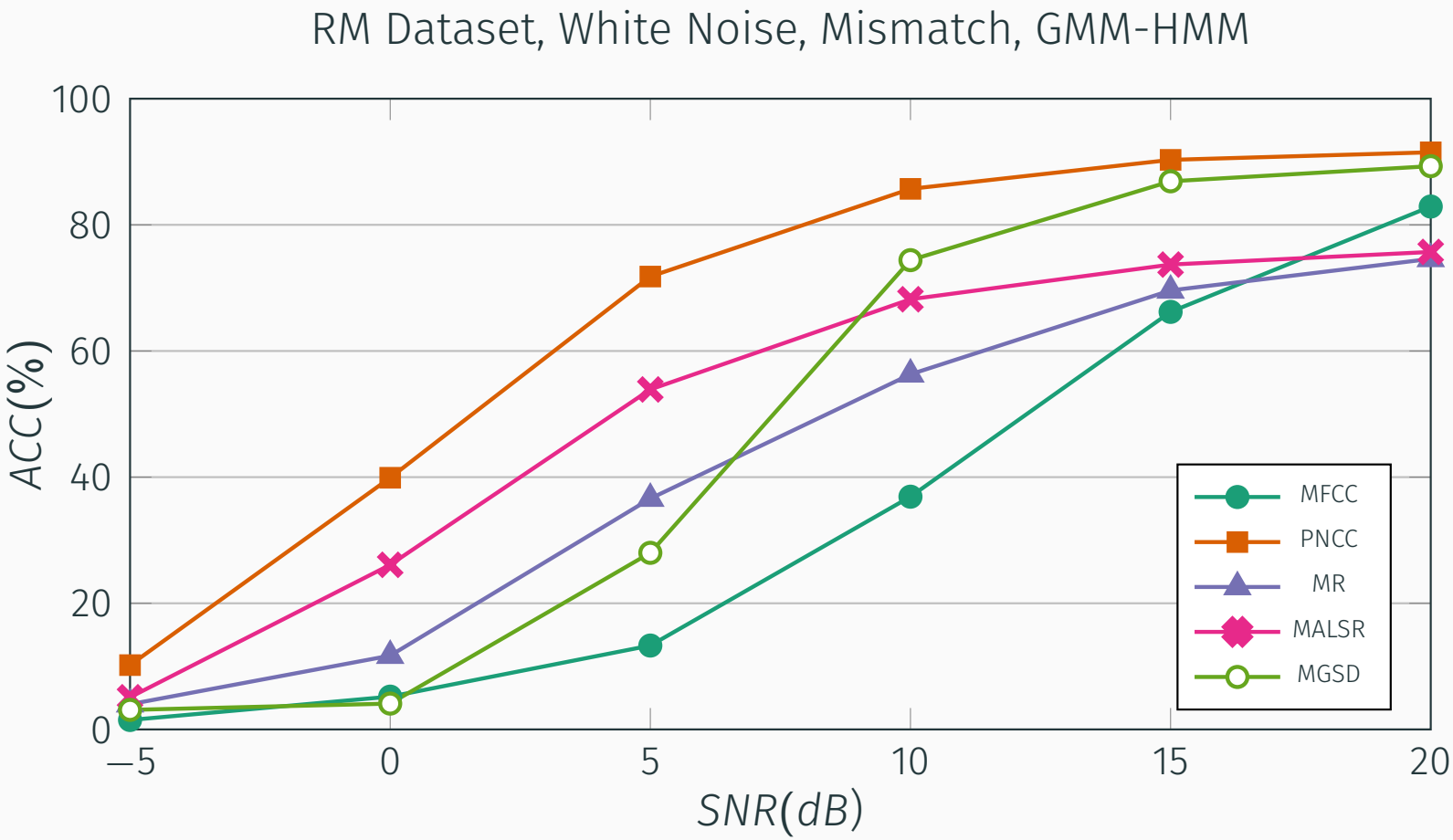
Results: MR, MASLR, MGSD



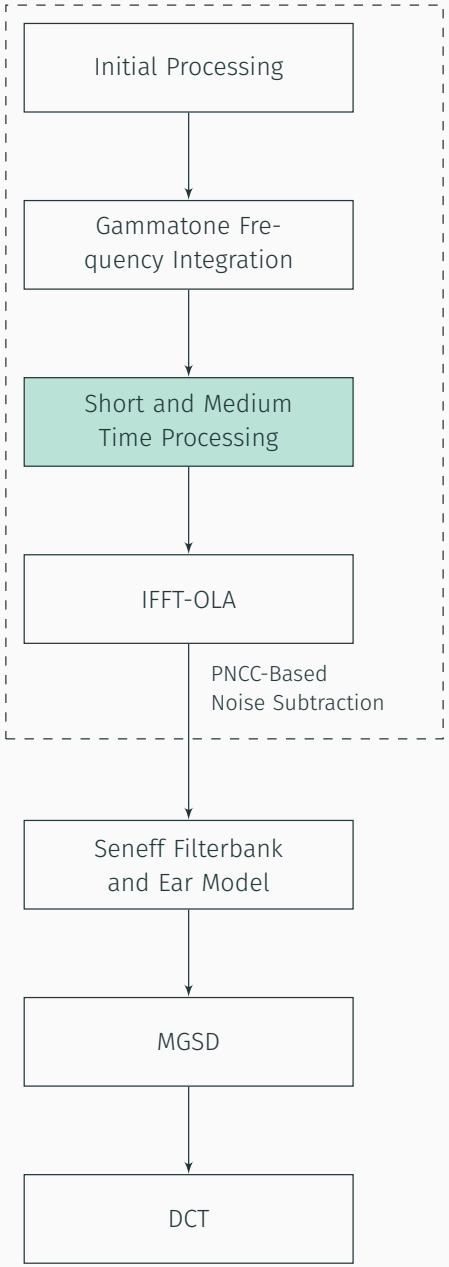
Results: MR, MASLR, MGSD



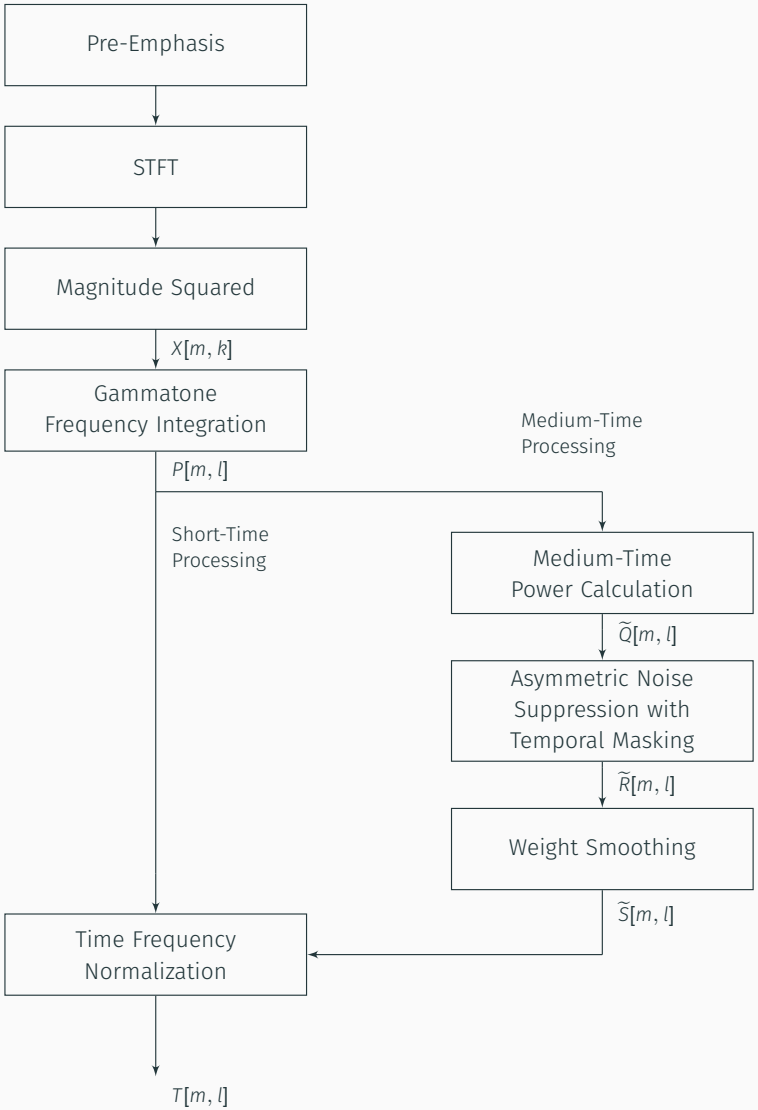
Results: MR, MASLR, MGSD



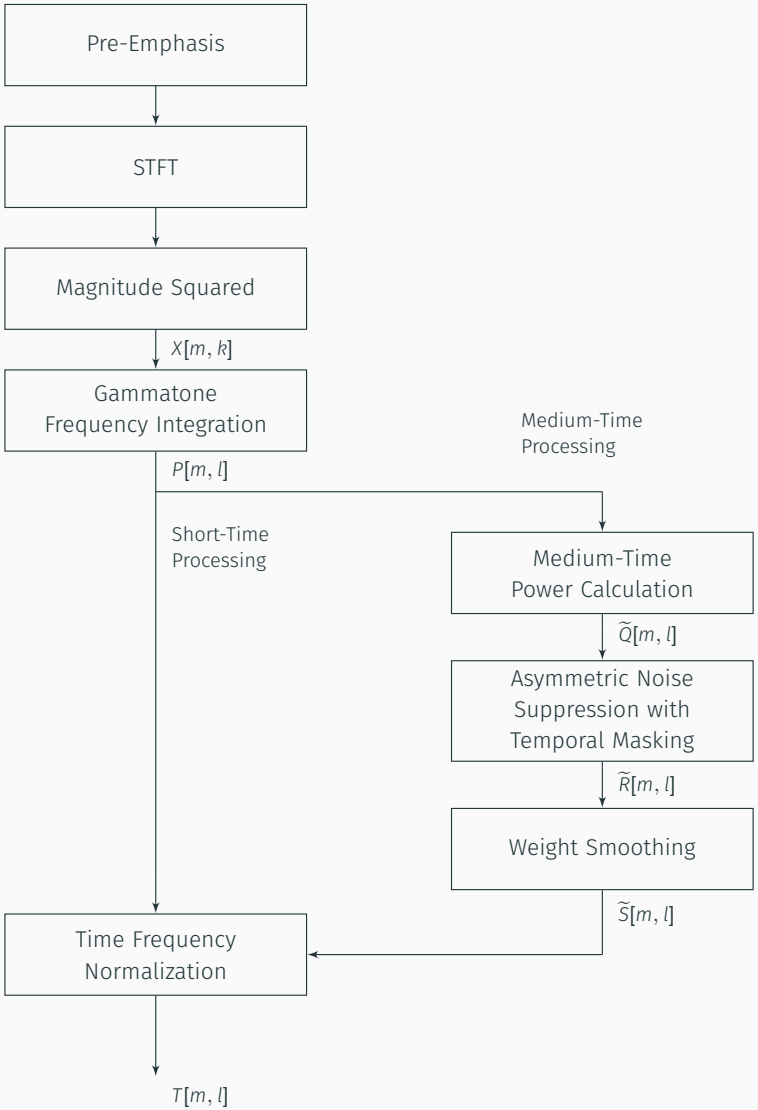
PNCC Integration



PNCC-Based Noise Reduction (PNCC-NR)

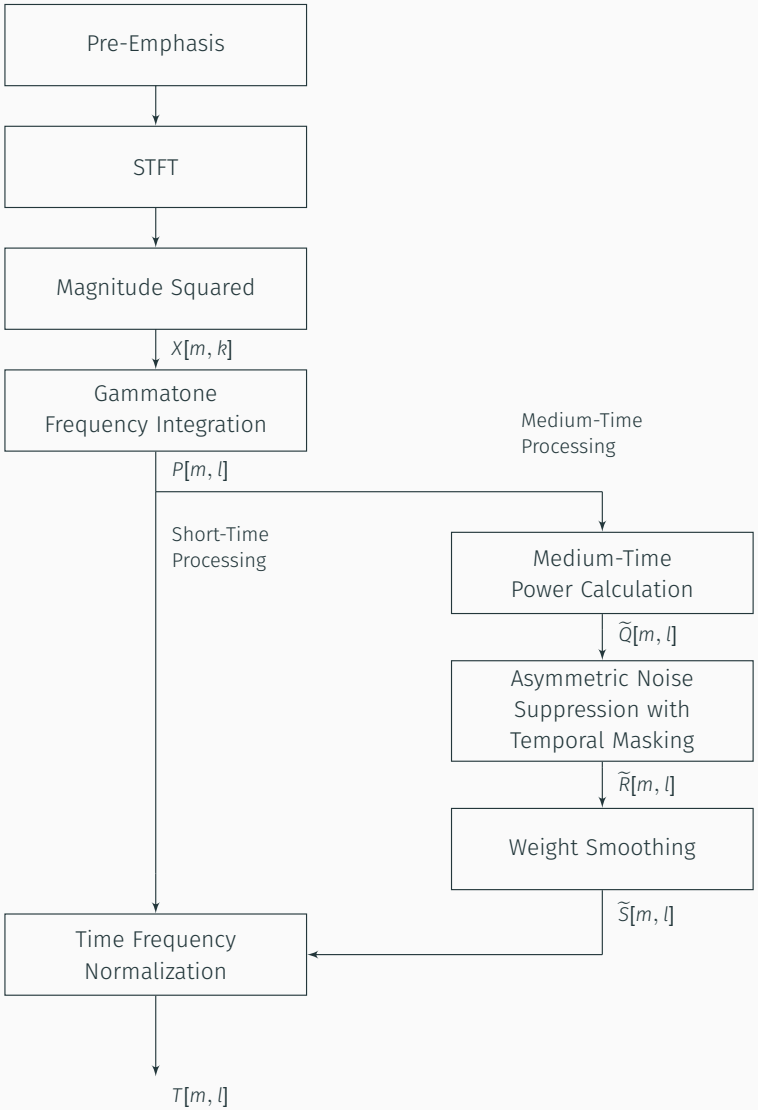


PNCC-Based Noise Reduction (PNCC-NR)



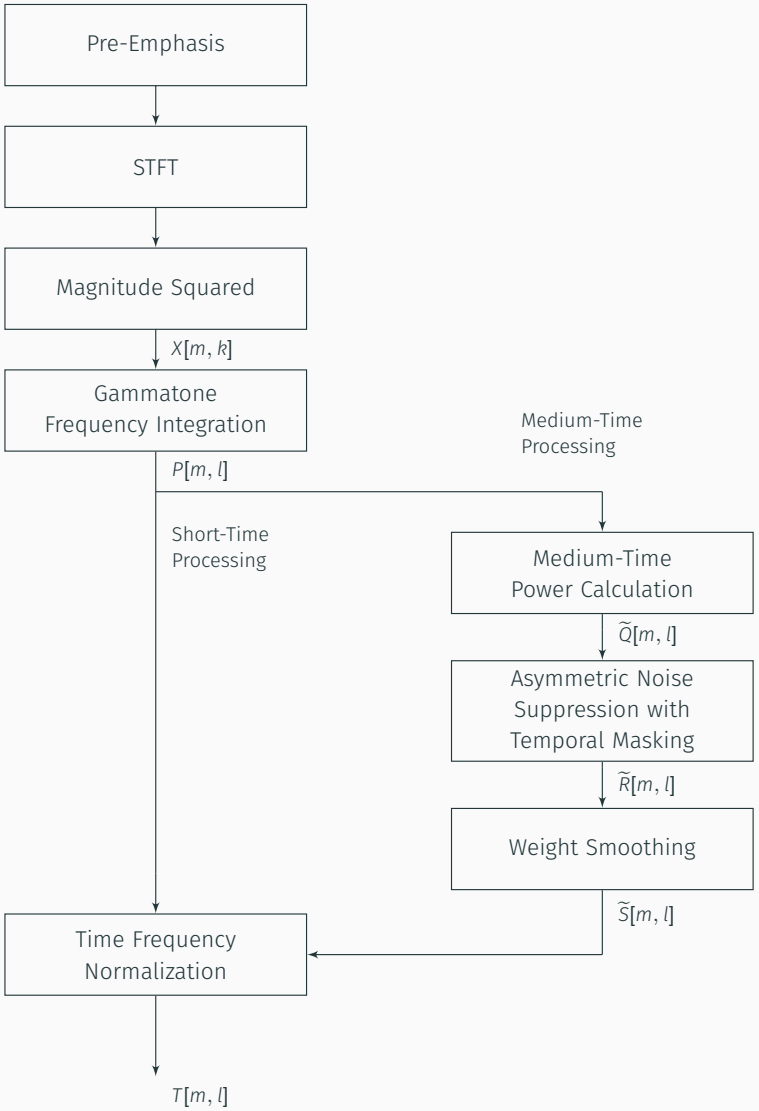
1. $x[n]$ pass through the PNCC pipeline.

PNCC-Based Noise Reduction (PNCC-NR)



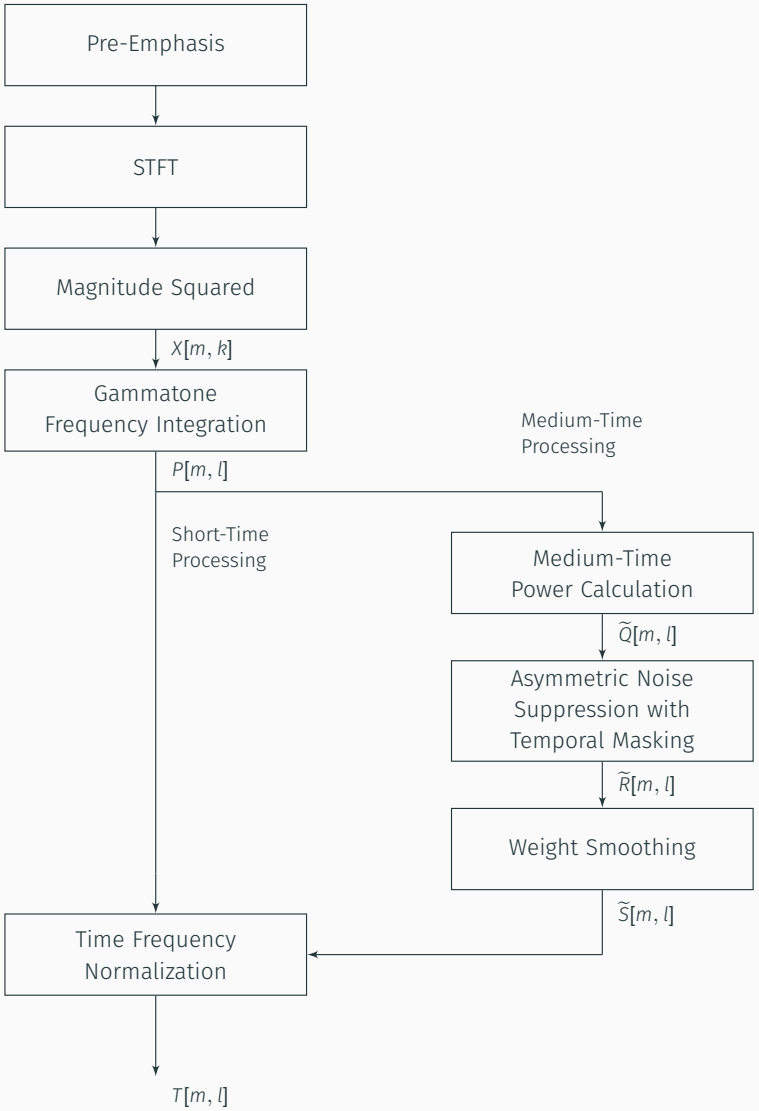
1. $x[n]$ pass through the PNCC pipeline.
2. Weighting coefficient: $w[m,l] = \frac{T[m,l]}{P[m,l]}$

PNCC-Based Noise Reduction (PNCC-NR)



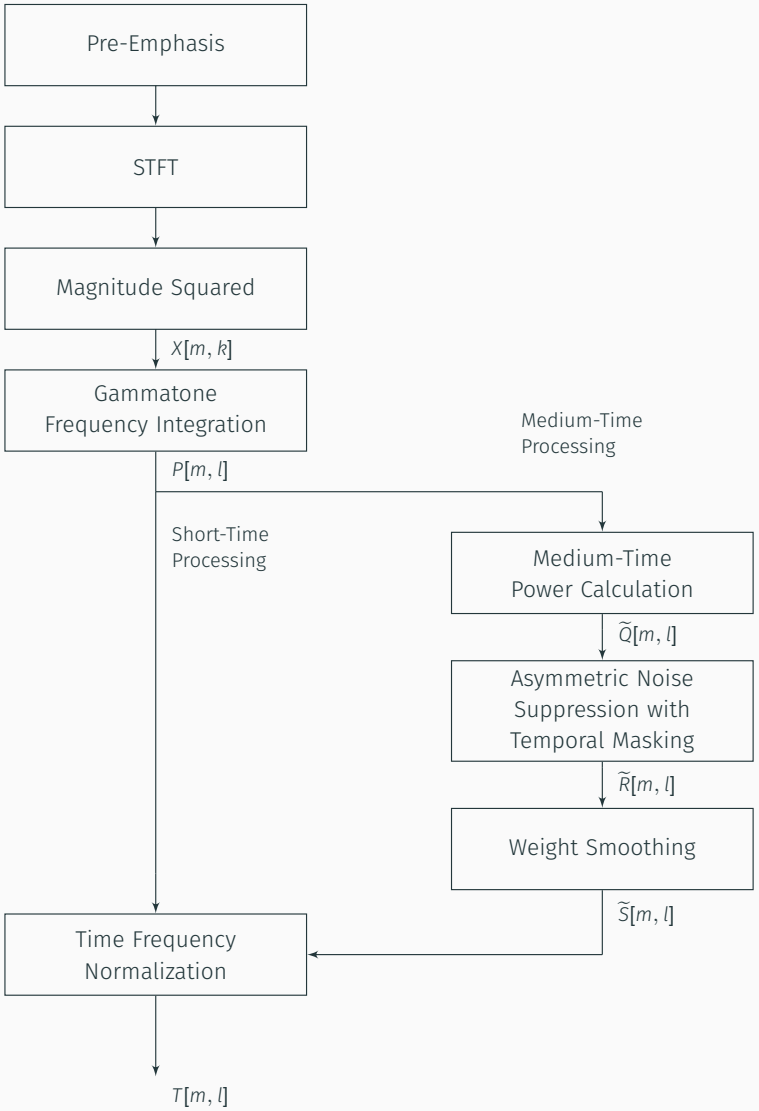
1. $x[n]$ pass through the PNCC pipeline.
2. Weighting coefficient: $w[m,l] = \frac{T[m,l]}{P[m,l]}$
3. Spectral weighting: $\mu[m,k] = \frac{\sum_{l=0}^{L-1} w[m,l]|H_l(k)|}{\sum_{l=0}^{L-1} |H_l(k)|}$

PNCC-Based Noise Reduction (PNCC-NR)



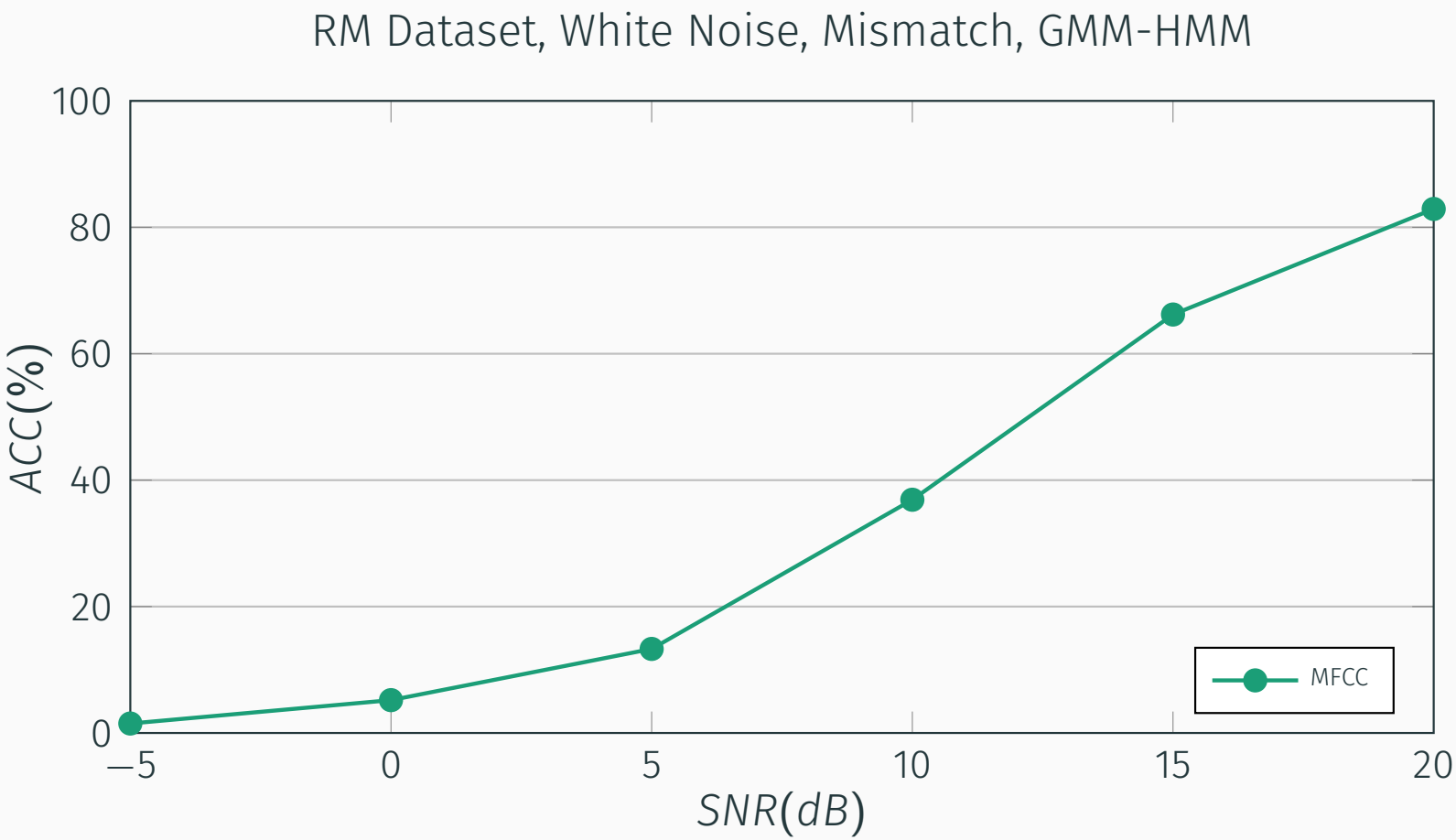
1. $x[n]$ pass through the PNCC pipeline.
2. Weighting coefficient: $w[m,l] = \frac{T[m,l]}{P[m,l]}$
3. Spectral weighting: $\mu[m,k] = \frac{\sum_{l=0}^{L-1} w[m,l]|H_l(k)|}{\sum_{l=0}^{L-1} |H_l(k)|}$
4. Reconstructed spectrum: $\widetilde{X}[m,k] = \mu[m,k]X[m,k]$

PNCC-Based Noise Reduction (PNCC-NR)

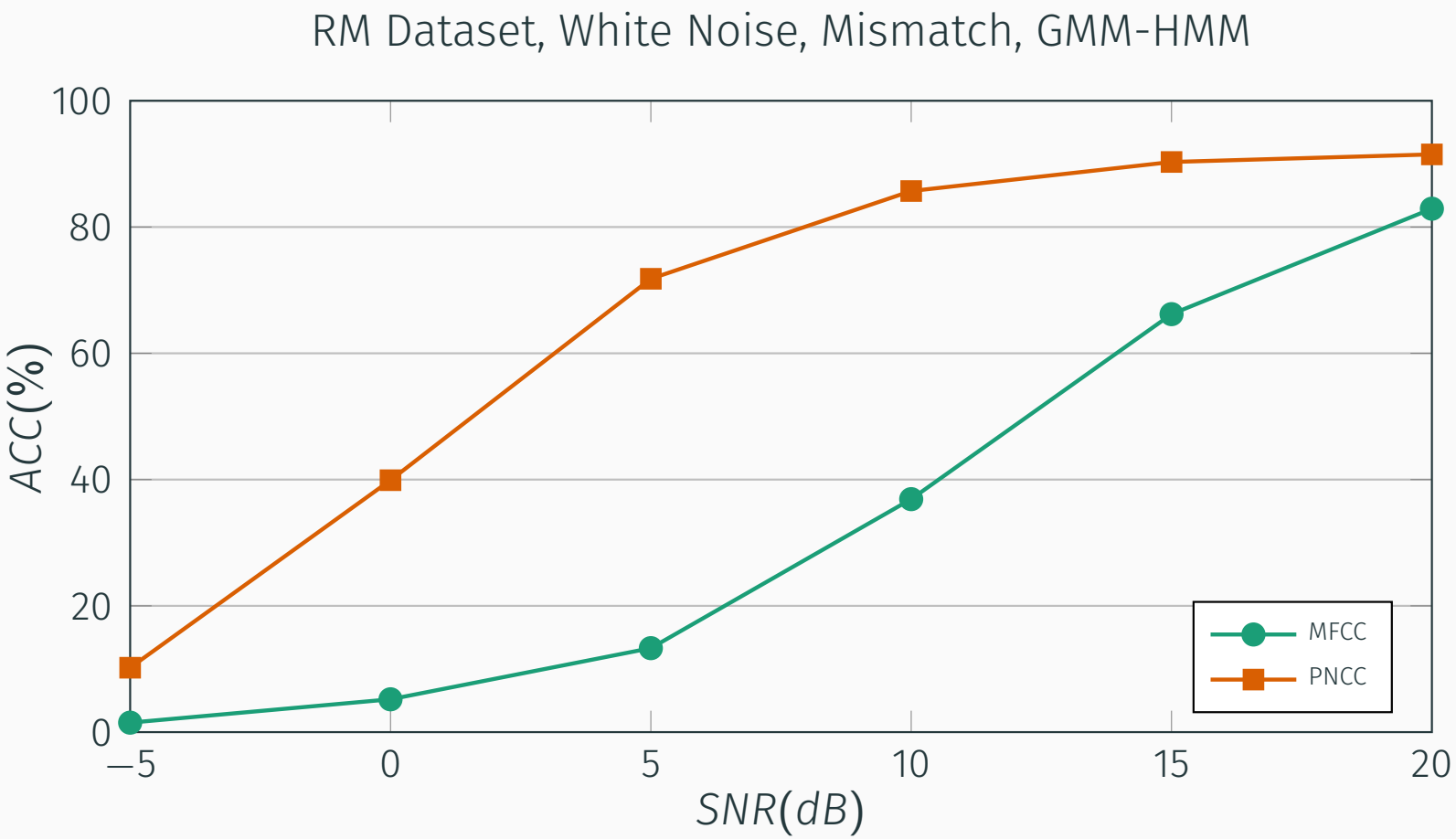


1. $x[n]$ pass through the PNCC pipeline.
2. Weighting coefficient: $w[m,l] = \frac{T[m,l]}{P[m,l]}$
3. Spectral weighting: $\mu[m,k] = \frac{\sum_{l=0}^{L-1} w[m,l]|H_l(k)|}{\sum_{l=0}^{L-1} |H_l(k)|}$
4. Reconstructed spectrum: $\widetilde{X}[m,k] = \mu[m,k]X[m,k]$
5. $\hat{x}[n]$ is re-synthesized from $\widetilde{X}[m,k]$ by applying and IFFT and using OLA.

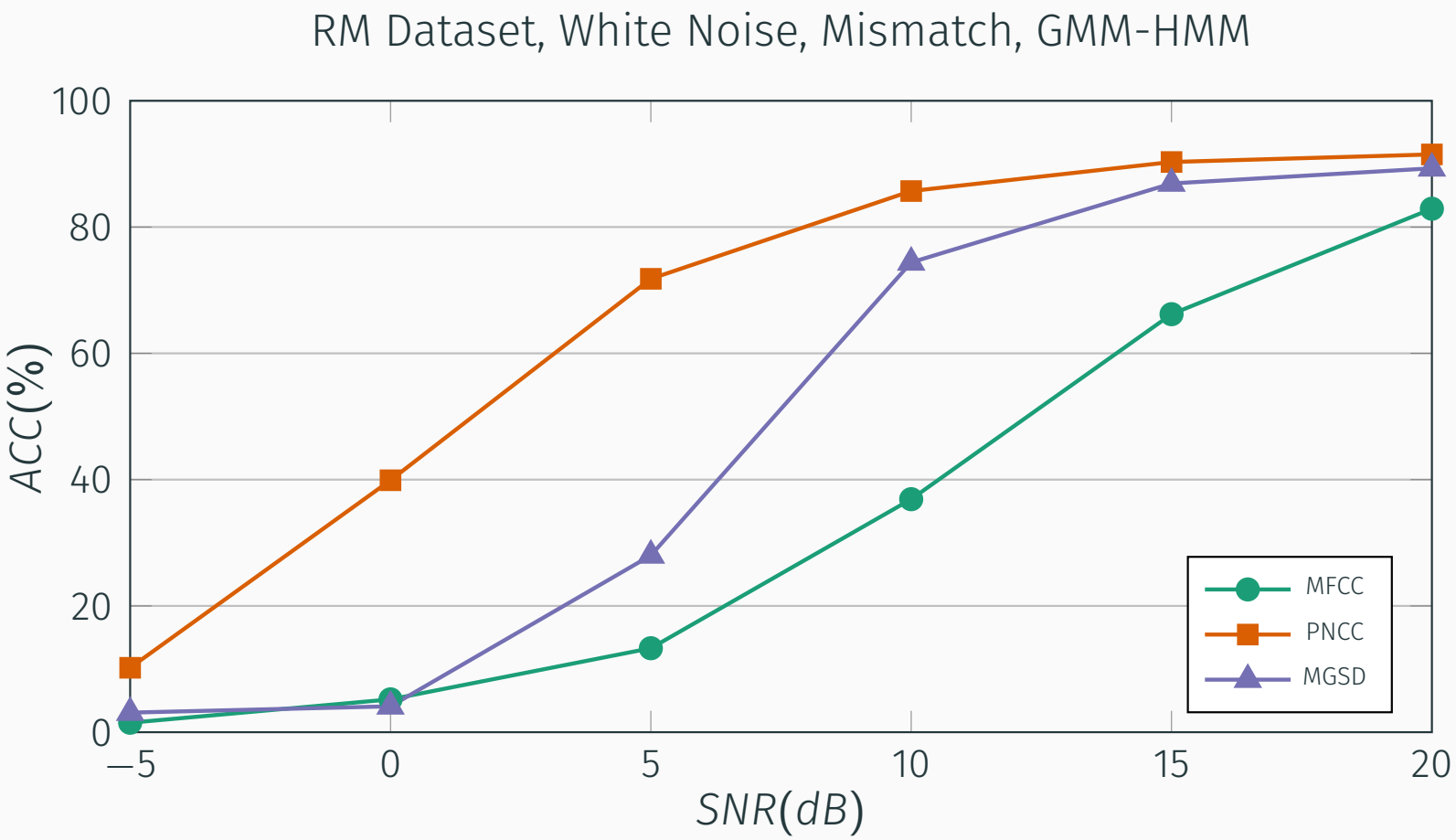
PNCC-NR Results



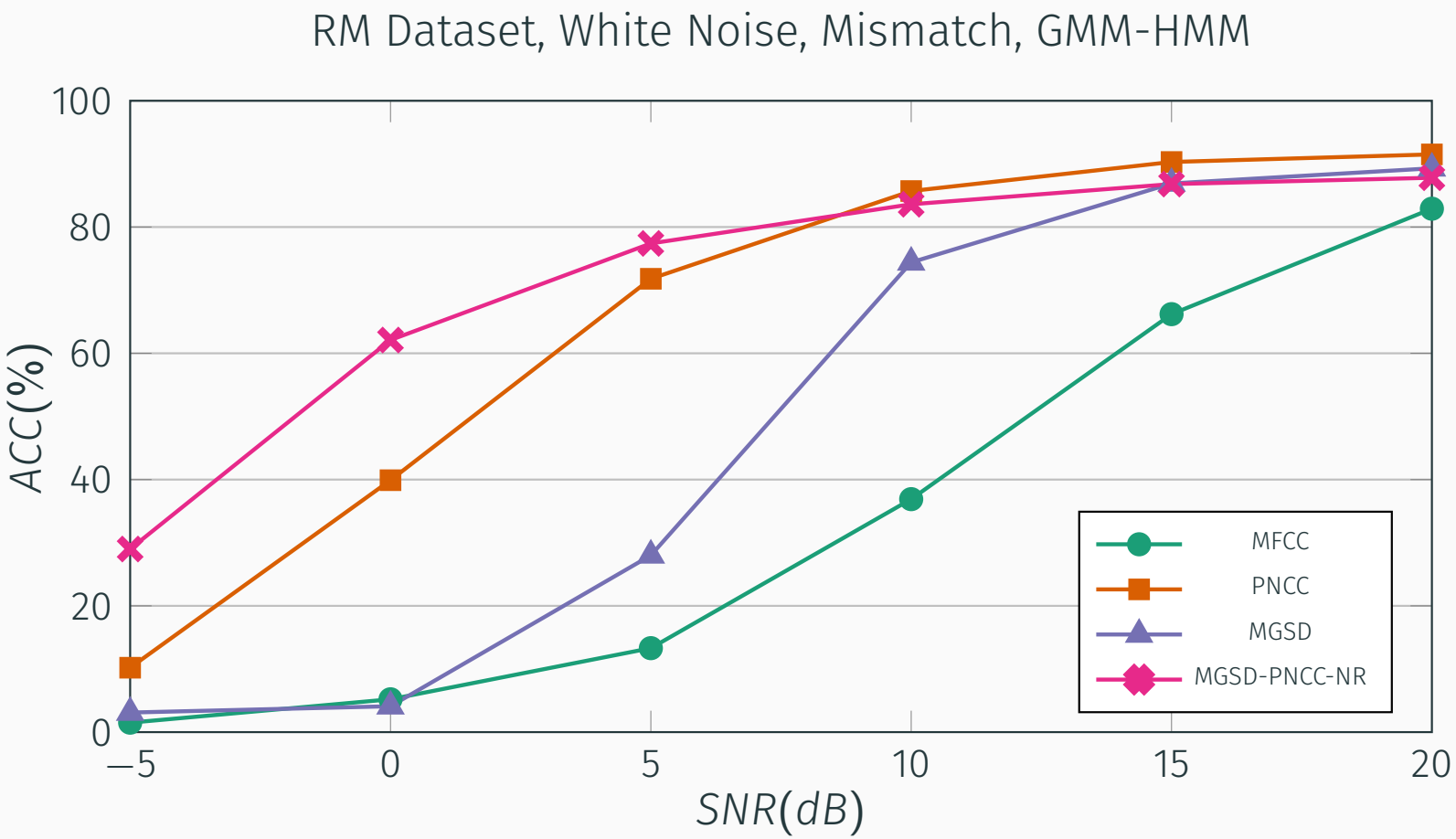
PNCC-NR Results



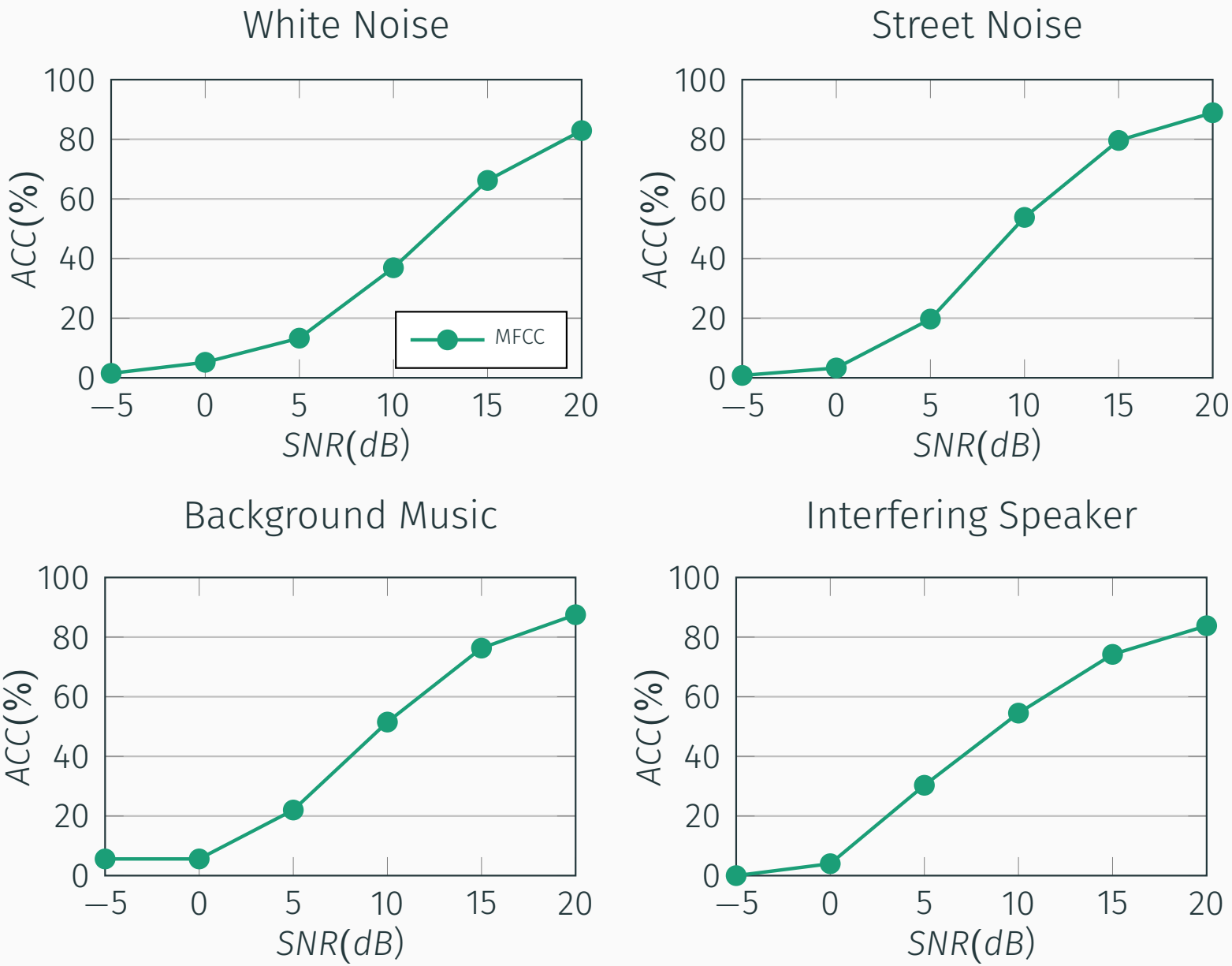
PNCC-NR Results



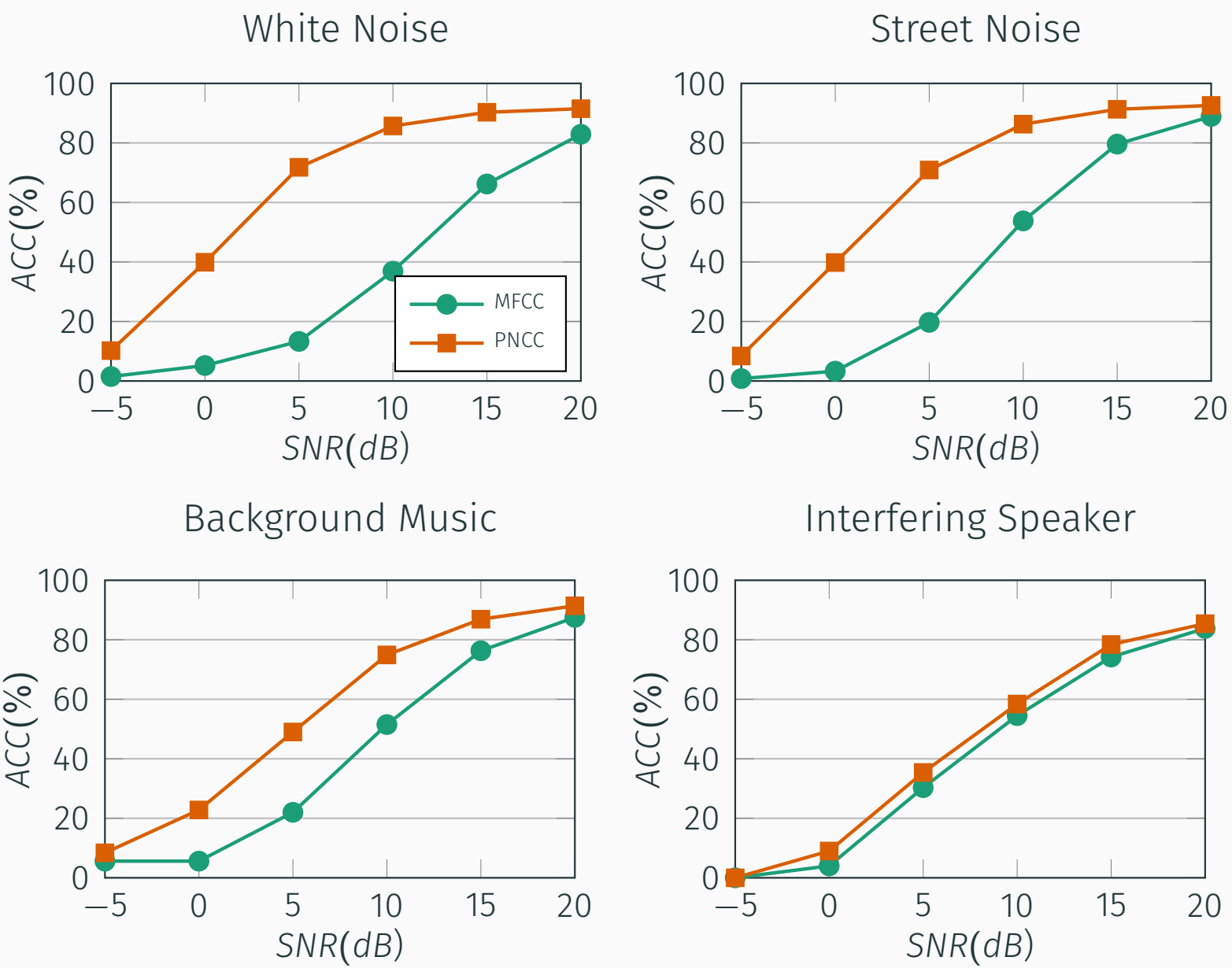
PNCC-NR Results



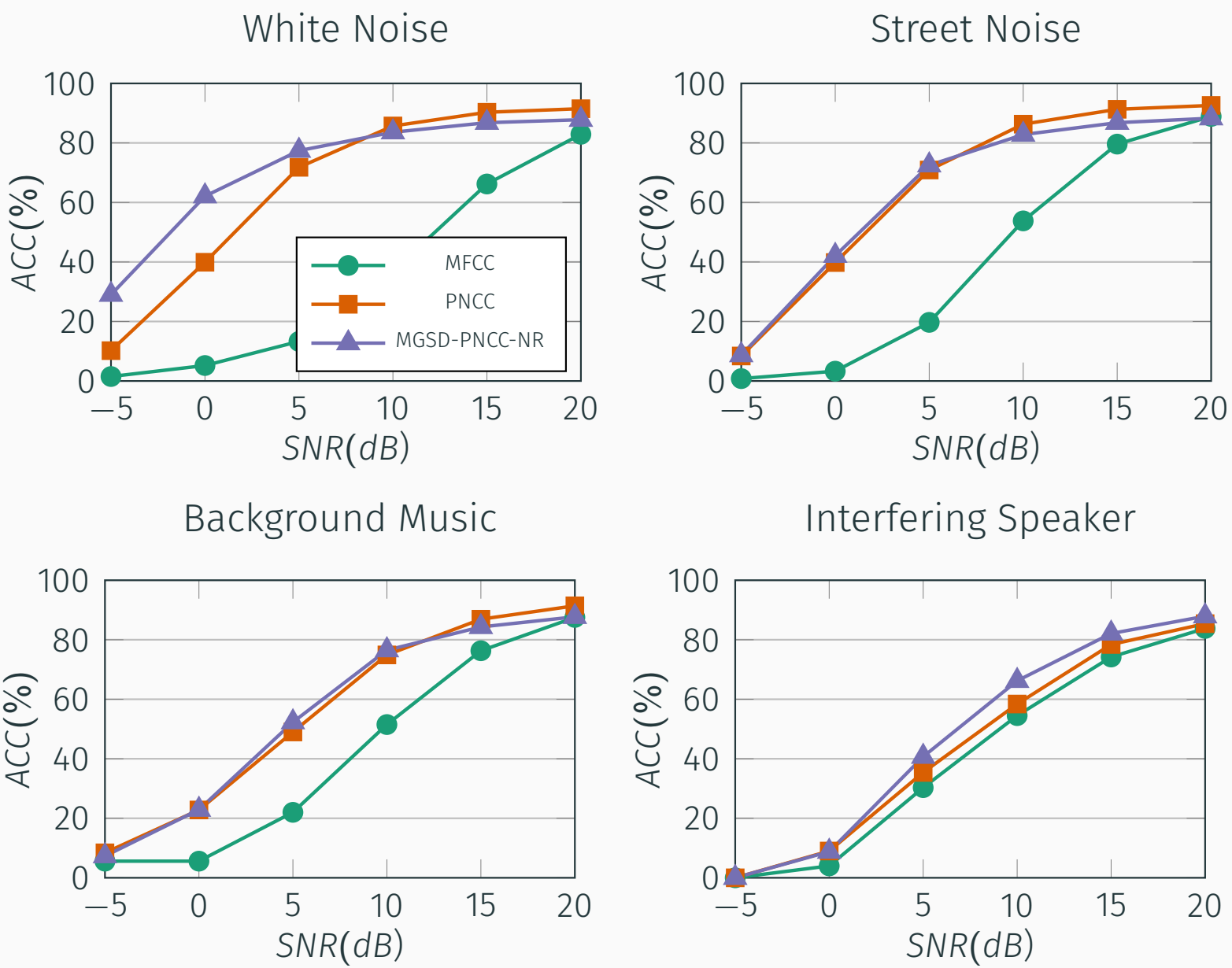
Results RM, Mismatch, GMM-HMM



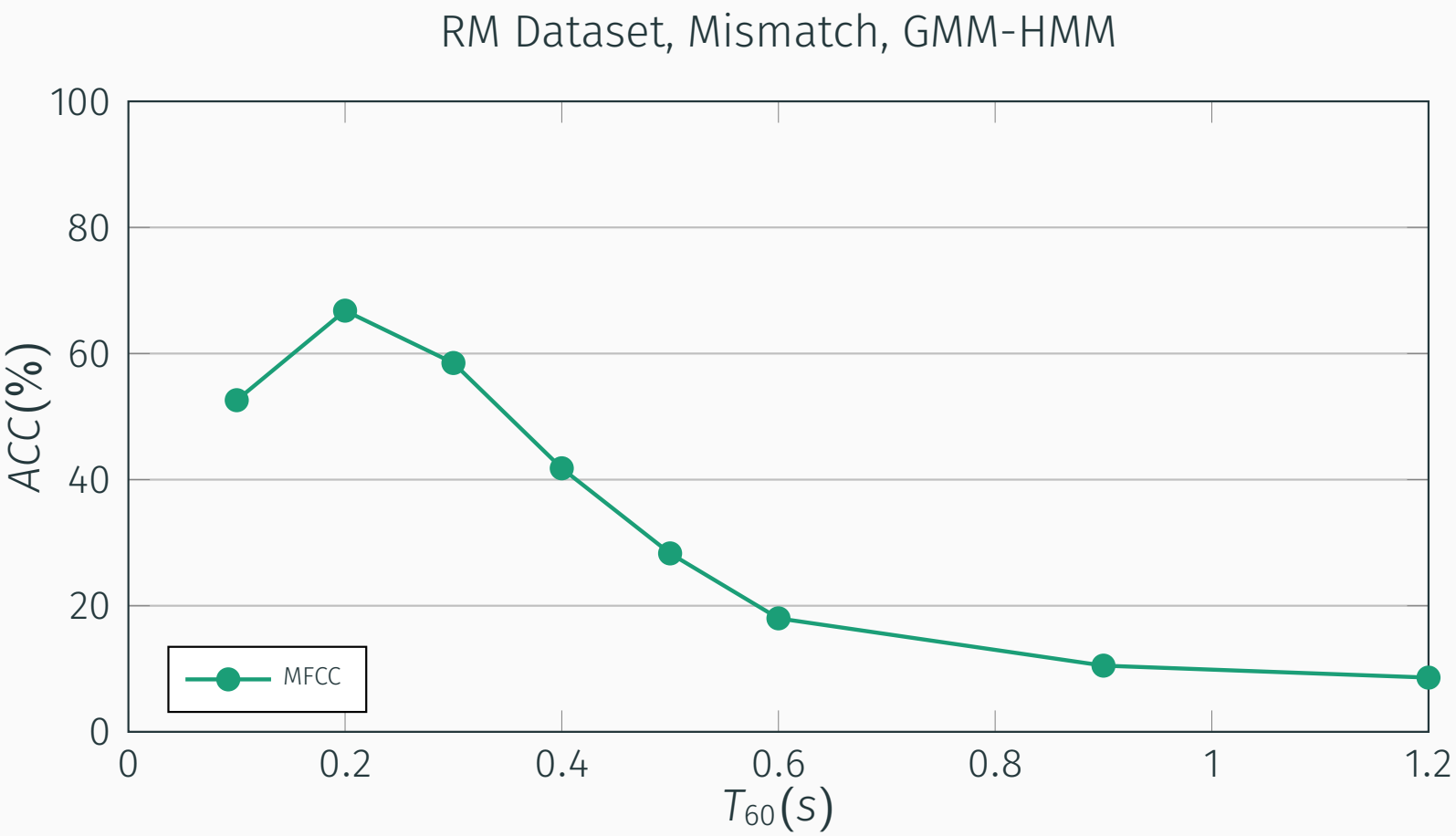
Results RM, Mismatch, GMM-HMM



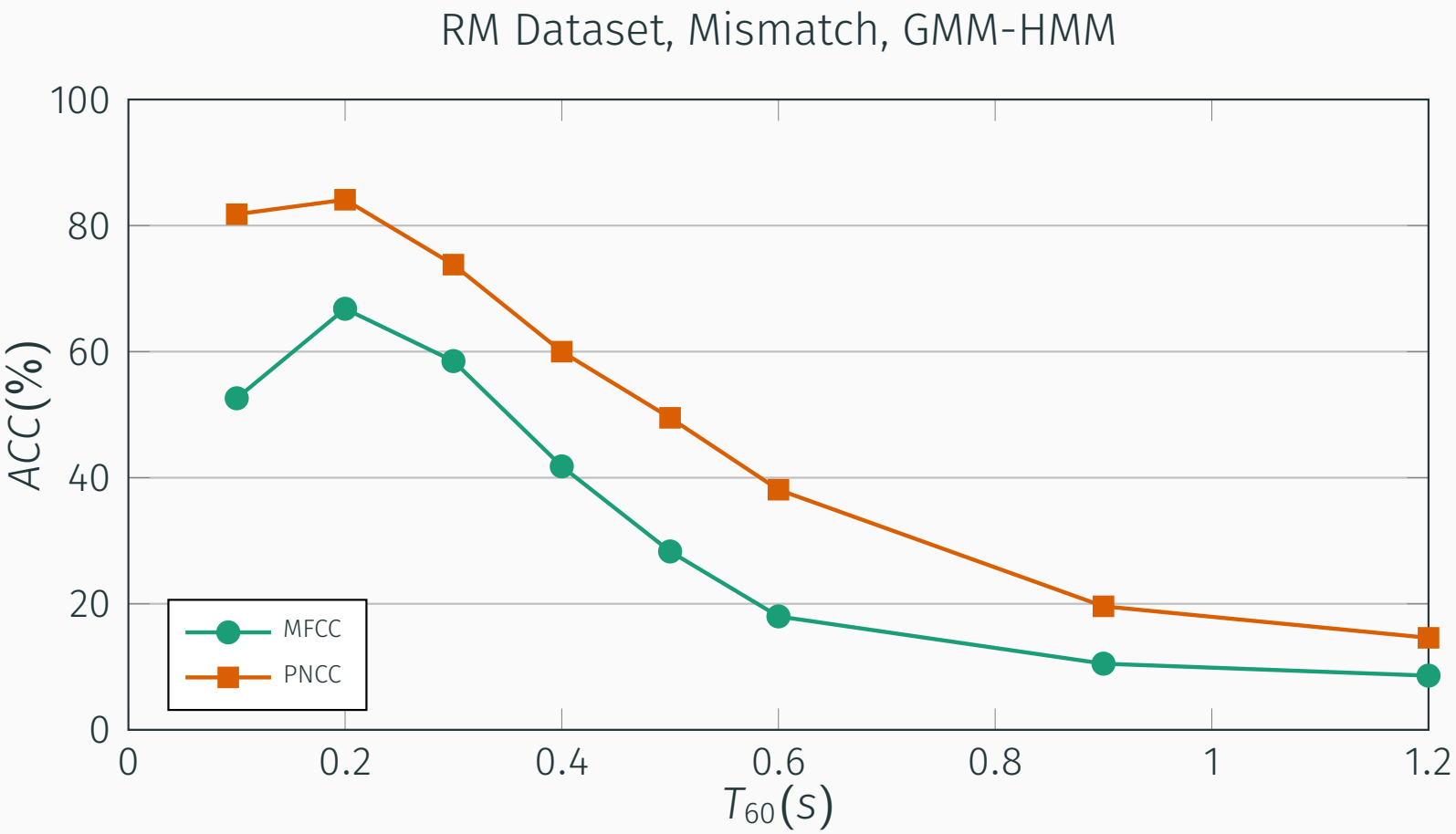
Results RM, Mismatch, GMM-HMM



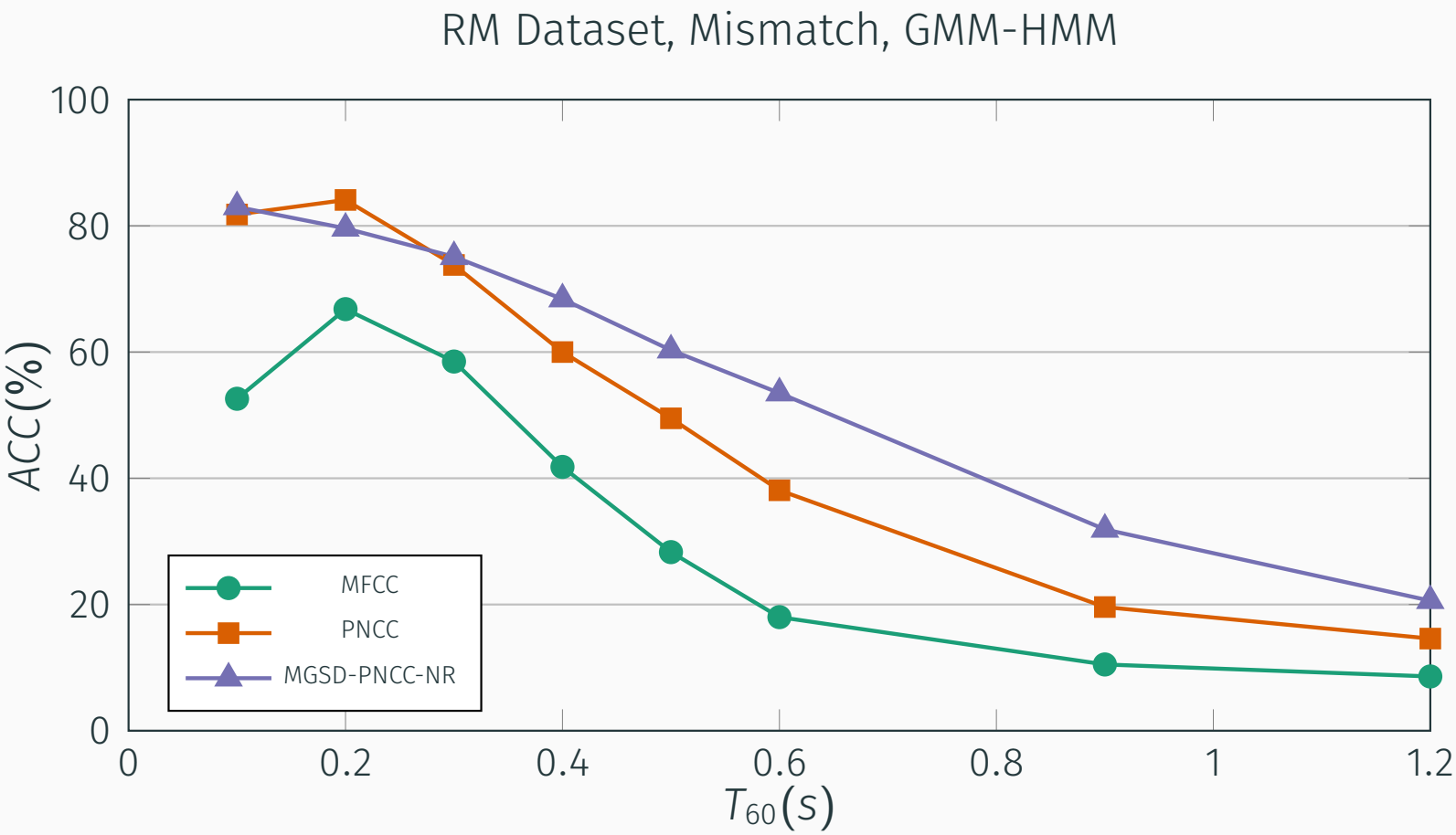
Results: Reverberant condition



Results: Reverberant condition



Results: Reverberant condition



CNNs and Bio-Inspired Features Combination

Motivation

- Deep learning changes the paradigm in speech recognition.

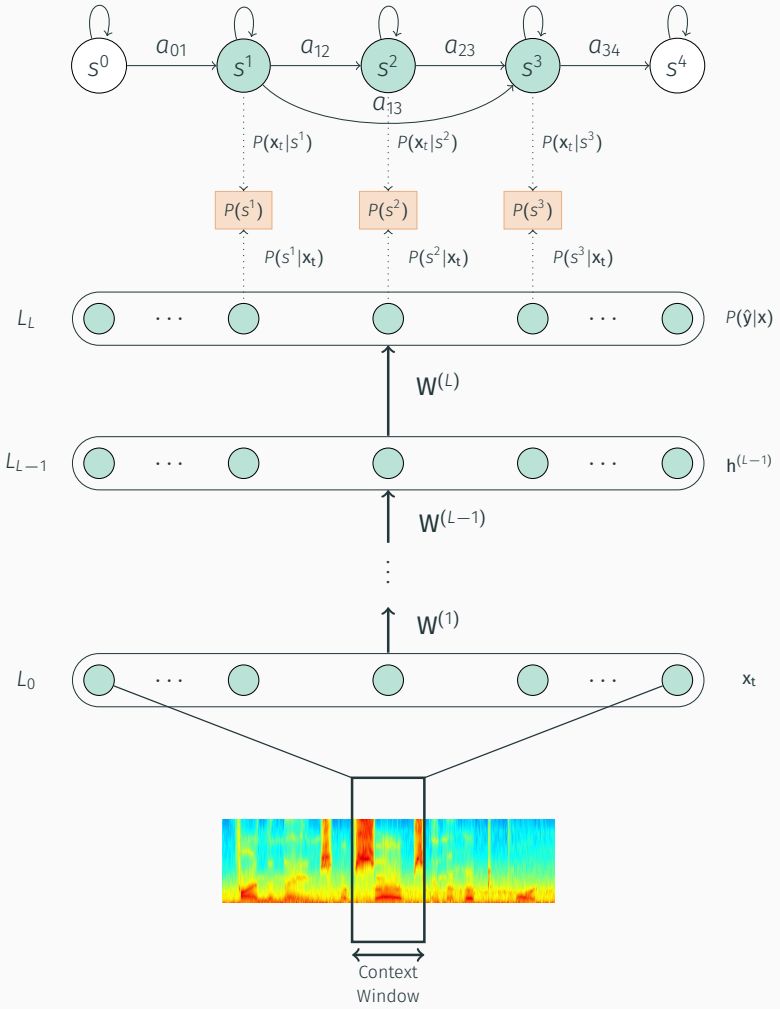
Motivation

- Deep learning changes the paradigm in speech recognition.
- ASR was the first industrial application.

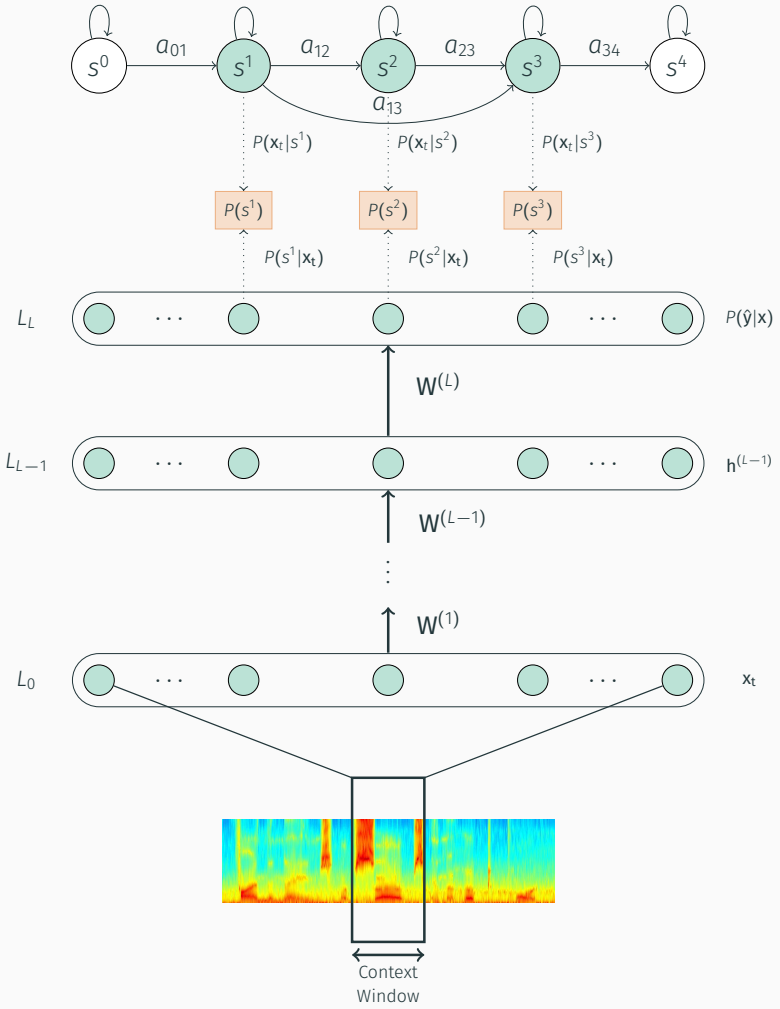
Motivation

- Deep learning changes the paradigm in speech recognition.
- ASR was the first industrial application.
- Not many works have addressed the robustness of these systems under noisy conditions.

Deep Learning in Speech Recognition: CD-DNN-HMM Hybrid model

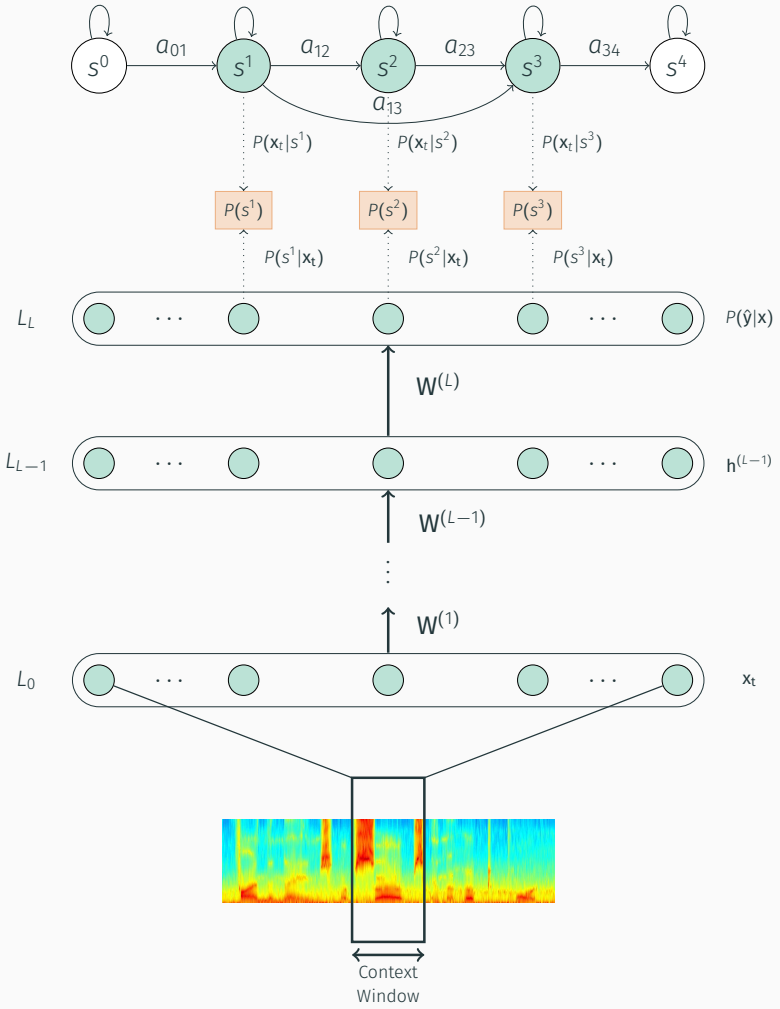


Deep Learning in Speech Recognition: CD-DNN-HMM Hybrid model



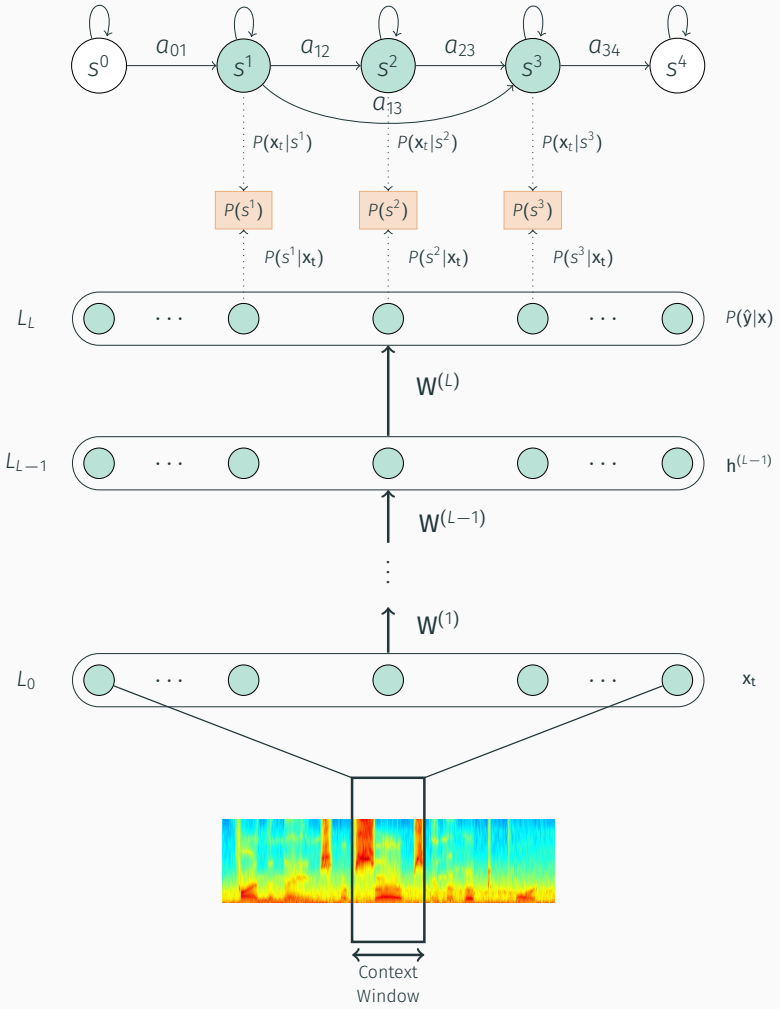
- Bourlard and Morgan (1994)

Deep Learning in Speech Recognition: CD-DNN-HMM Hybrid model



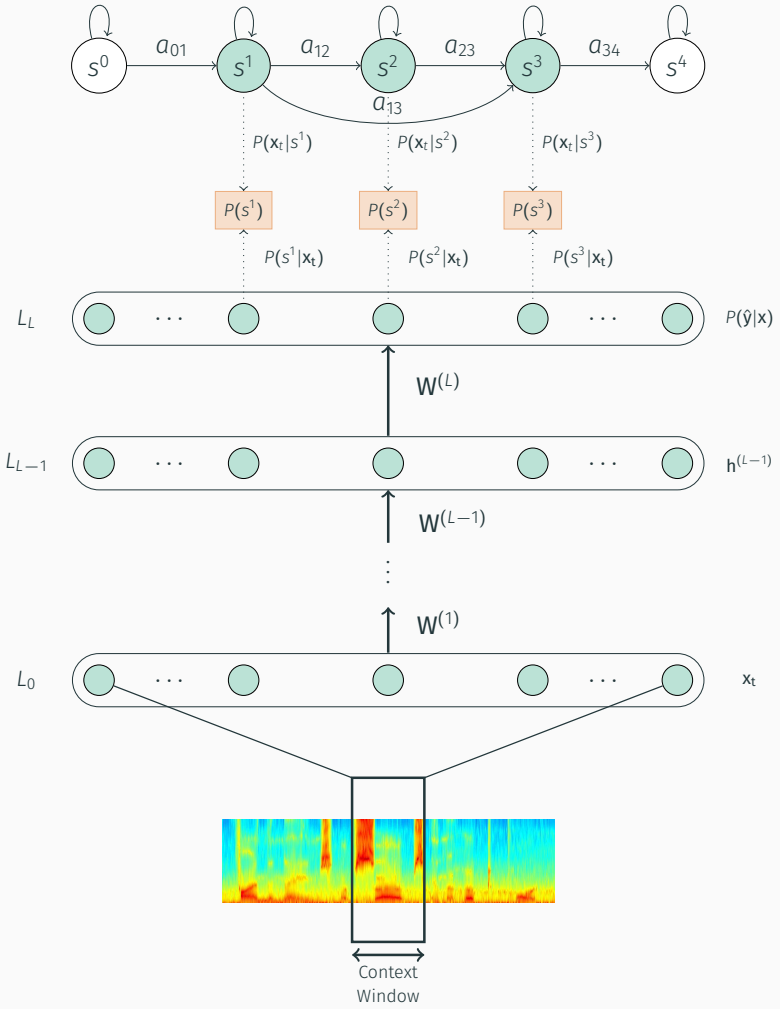
- Bourlard and Morgan (1994)
- DNN with larger number of hidden layers.

Deep Learning in Speech Recognition: CD-DNN-HMM Hybrid model



- Bourlard and Morgan (1994)
- DNN with larger number of hidden layers.
- Models senones (tied states) directly.

Deep Learning in Speech Recognition: CD-DNN-HMM Hybrid model



- Bourlard and Morgan (1994)
- DNN with larger number of hidden layers.
- Models senones (tied states) directly.
- Long context windows (10 to 20 frames)

DNN for Robust Speech Recognition

- Alternatives:

DNN for Robust Speech Recognition

- Alternatives:
  - Data augmentation with multi-condition data.

DNN for Robust Speech Recognition

- Alternatives:
  - Data augmentation with multi-condition data.
  - Incorporating a noise model.

DNN for Robust Speech Recognition

- Alternatives:
  - Data augmentation with multi-condition data.
  - Incorporating a noise model.
  - Systems combination.

DNN for Robust Speech Recognition

- Alternatives:
  - Data augmentation with multi-condition data.
  - Incorporating a noise model.
  - Systems combination.
  - Robust Architectures.

DNN for Robust Speech Recognition

- Alternatives:
  - Data augmentation with multi-condition data.
  - Incorporating a noise model.
  - Systems combination.
  - Robust Architectures.
  - Robust features.

DNN for Robust Speech Recognition

- Alternatives:
  - Data augmentation with multi-condition data.
  - Incorporating a noise model.
  - Systems combination.
  - Robust Architectures.
  - Robust features.
- Our approach:

DNN for Robust Speech Recognition

- Alternatives:
  - Data augmentation with multi-condition data.
  - Incorporating a noise model.
  - Systems combination.
  - Robust Architectures.
  - Robust features.
- Our approach:
  - Convolutional Neural Networks.

DNN for Robust Speech Recognition

- Alternatives:
  - Data augmentation with multi-condition data.
  - Incorporating a noise model.
  - Systems combination.
  - Robust Architectures.
  - Robust features.
- Our approach:
  - Convolutional Neural Networks.
  - Previously presented features.

CNNs in ASR

- Convolution replaces the general matrix multiplication.

Redrawn Dumoulin and Visin (2016).

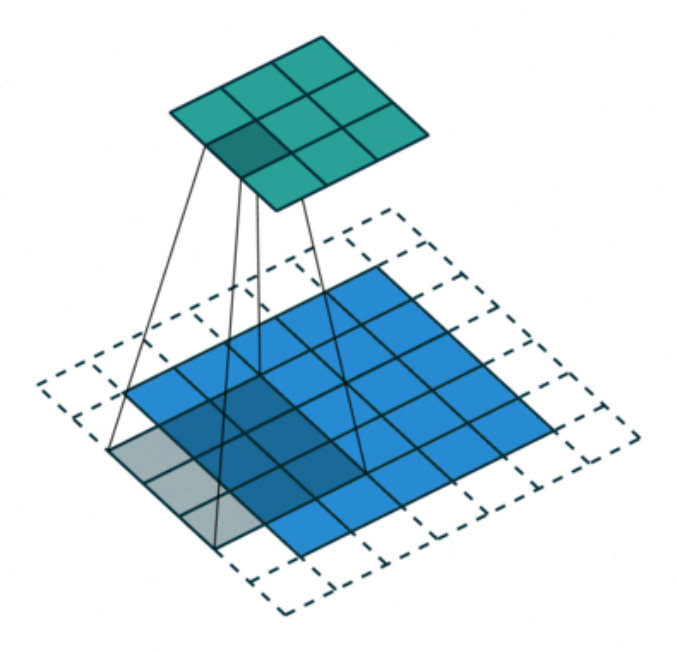
CNNs in ASR

- Convolution replaces the general matrix multiplication.

Redrawn Dumoulin and Visin (2016).

CNNs in ASR

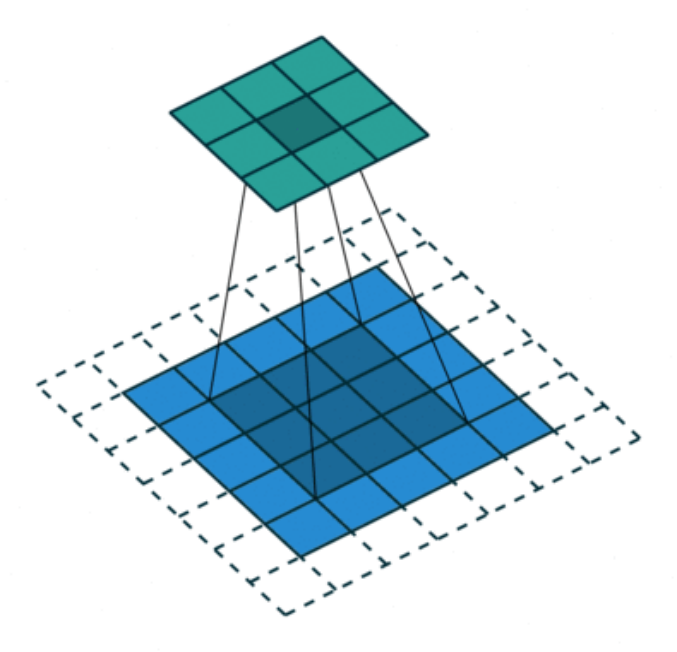
- Convolution replaces the general matrix multiplication.



Redrawn Dumoulin and Visin (2016).

CNNs in ASR

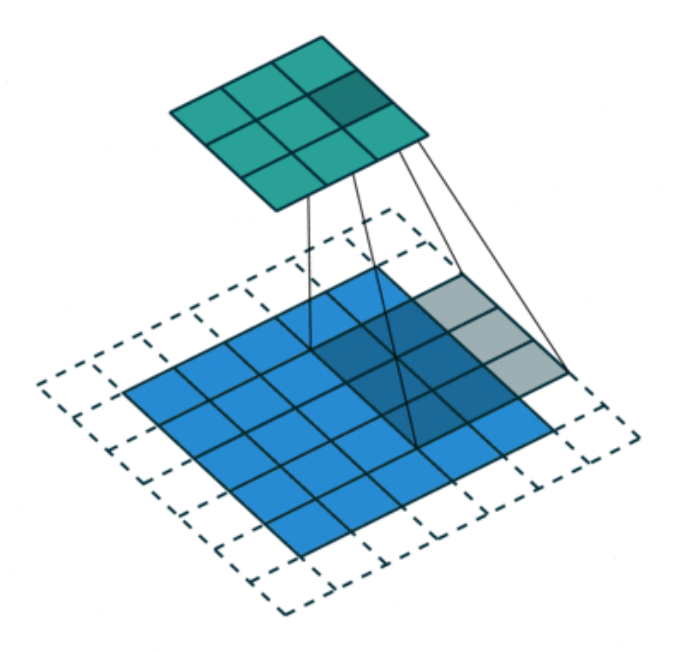
- Convolution replaces the general matrix multiplication.



Redrawn Dumoulin and Visin (2016).

CNNs in ASR

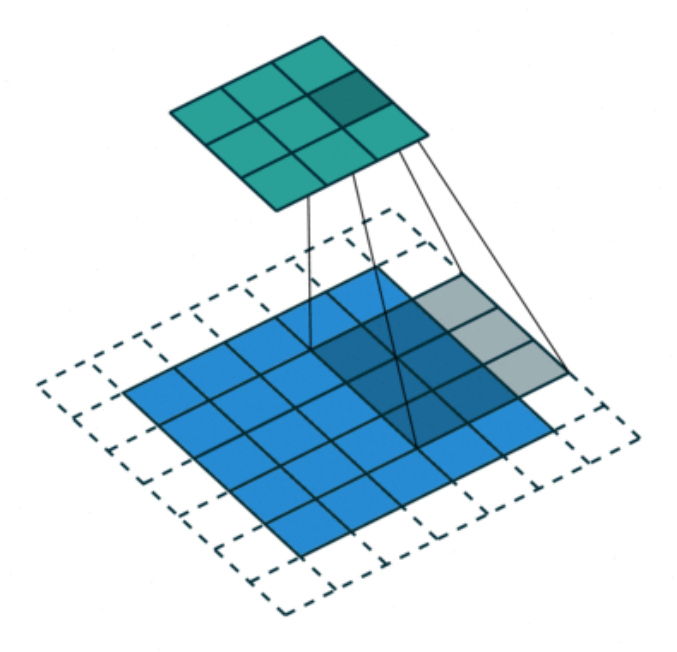
- Convolution replaces the general matrix multiplication.



Redrawn Dumoulin and Visin (2016).

CNNs in ASR

- Convolution replaces the general matrix multiplication.

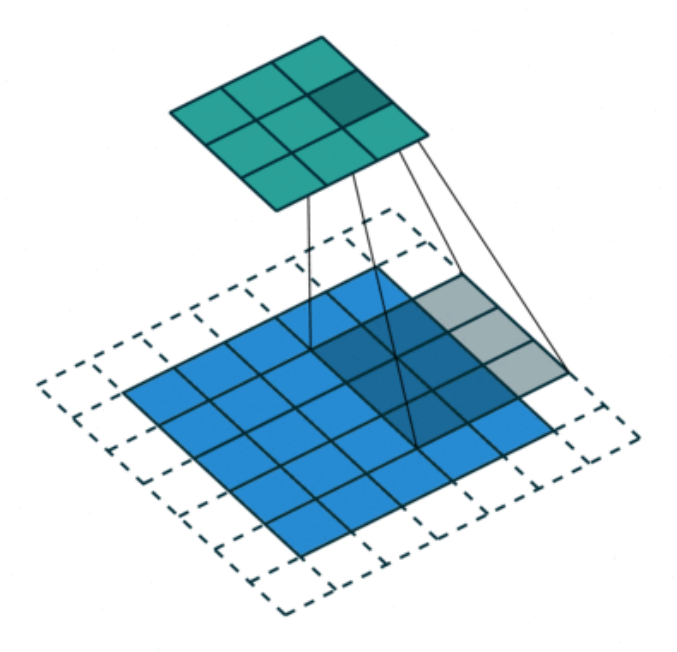


- Robust architecture due to:

Redrawn Dumoulin and Visin (2016).

CNNs in ASR

- Convolution replaces the general matrix multiplication.

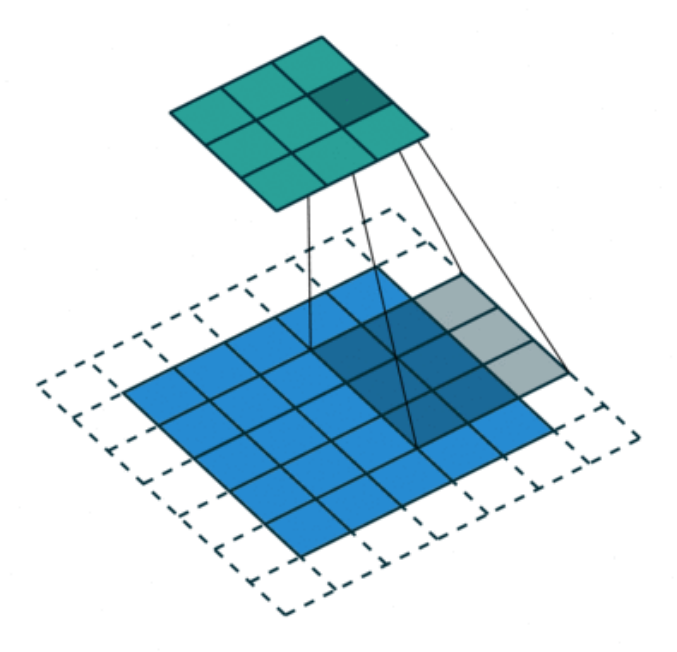


- Robust architecture due to:
  - Less parameters.

Redrawn Dumoulin and Visin (2016).

CNNs in ASR

- Convolution replaces the general matrix multiplication.

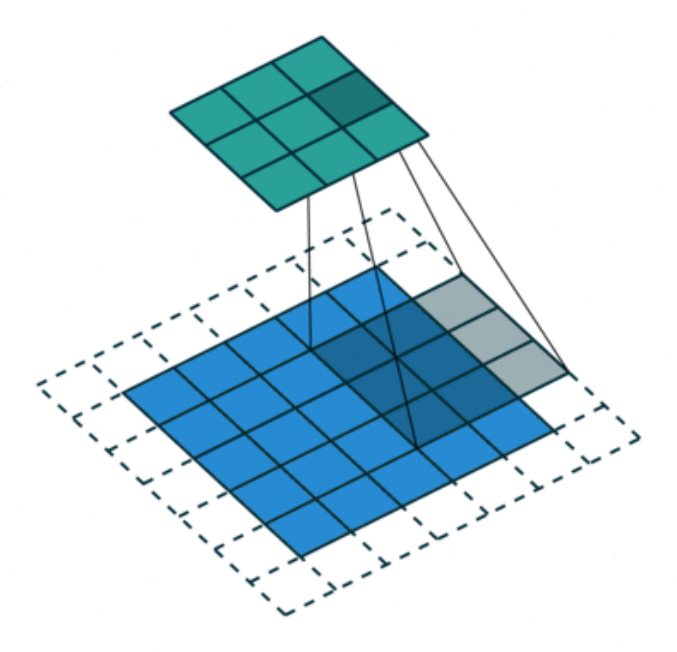


- Robust architecture due to:
  - Less parameters.
  - Translation invariance characteristics.

Redrawn Dumoulin and Visin (2016).

CNNs in ASR

- Convolution replaces the general matrix multiplication.



- Robust architecture due to:
  - Less parameters.
  - Translation invariance characteristics.
- We propose a new architecture based on RestNets and its combination with our robust features.

Redrawn Dumoulin and Visin (2016).

ResNets

- Computer vision tasks (He et al., 2016).

ResNets

- Computer vision tasks (He et al., 2016).
- Approximate the residual function:

$$F(x) = H(x) - x$$

ResNets

- Computer vision tasks (He et al., 2016).
- Approximate the residual function:

$$F(x) = H(x) - x$$

- Easier to push the residual to zero.

ResNets

- Computer vision tasks (He et al., 2016).
- Approximate the residual function:

$$F(x) = H(x) - x$$

- Easier to push the residual to zero.
- Shortcuts connection:

$$H(x) = F(x) + x$$

ResNets

- Computer vision tasks (He et al., 2016).
- Approximate the residual function:

$$F(x) = H(x) - x$$

- Easier to push the residual to zero.
- Shortcuts connection:

$$H(x) = F(x) + x$$

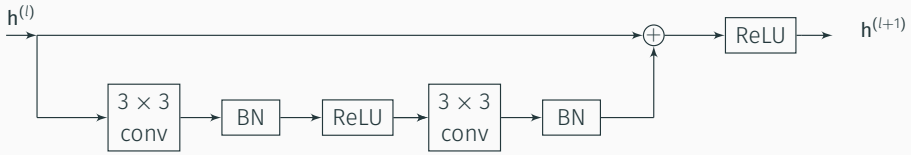
ResNets

- Computer vision tasks (He et al., 2016).
- Approximate the residual function:

$$F(x) = H(x) - x$$

- Easier to push the residual to zero.
- Shortcuts connection:

$$H(x) = F(x) + x$$



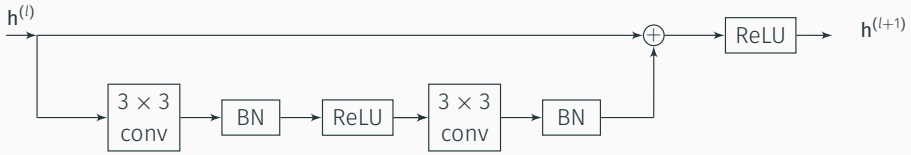
ResNets

- Computer vision tasks (He et al., 2016).
- Approximate the residual function:

$$F(x) = H(x) - x$$

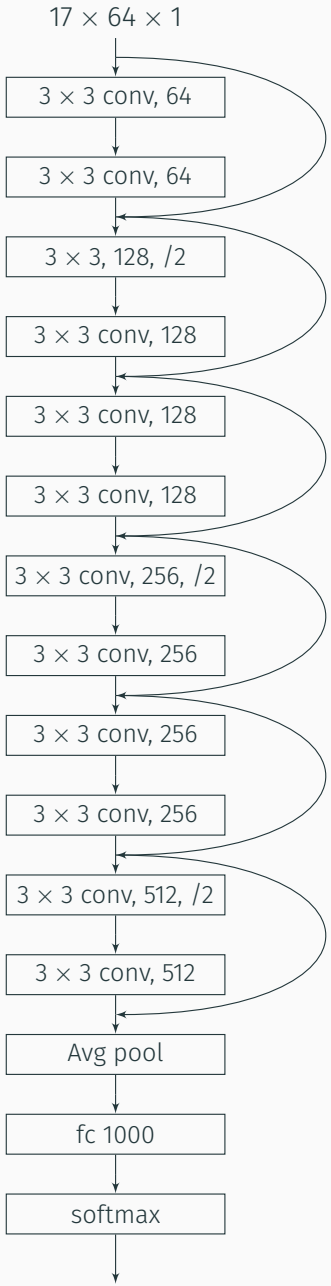
- Easier to push the residual to zero.
- Shortcuts connection:

$$H(x) = F(x) + x$$



- Easy to optimize and performance gains when networks are deep.

Proposed ResNet architecture



Features Modification

1. Increases output dimension.

Features Modification

1. Increases output dimension.
    - Removing the DCT: non-null cross-correlations between the components of the feature vectors.

Features Modification

1. Increases output dimension.
    - Removing the DCT: non-null cross-correlations between the components of the feature vectors.
    - Increasing the number of filters.

Features Modification

1. Increases output dimension.
   - Removing the DCT: non-null cross-correlations between the components of the feature vectors.
   - Increasing the number of filters.
2. Logarithmic non-linearity function performs better in conjunction with deep-learning back-ends.

Tested Features

1. **MelFB:** log filter-bank representation based on MFCC.

Tested Features

1. **MelFB:** log filter-bank representation based on MFCC.
2. **PNFB:** filter-bank version of PNCC where the power-law non-linearity is replaced with the logarithmic non-linearity.

Tested Features

1. **MelFB:** log filter-bank representation based on MFCC.
2. **PNFB:** filter-bank version of PNCC where the power-law non-linearity is replaced with the logarithmic non-linearity.
3. **MF-PNFB:** PNFB with the masking modeling.

Tested Features

1. **MelFB:** log filter-bank representation based on MFCC.
2. **PNFB:** filter-bank version of PNCC where the power-law non-linearity is replaced with the logarithmic non-linearity.
3. **MF-PNFB:** PNFB with the masking modeling.
4. **MGSD-MF-PNCC-NR:** synchrony modeling where the synchrony spectrum is filtered using the MF technique.

Experimental Setup

- Aurora 4 clean and multi-condition training sets.

Experimental Setup

- Aurora 4 clean and multi-condition training sets.
- Triphone GMM-HMM.

Experimental Setup

- Aurora 4 clean and multi-condition training sets.
- Triphone GMM-HMM.
- Network architectures:

Experimental Setup

- Aurora 4 clean and multi-condition training sets.
- Triphone GMM-HMM.
- Network architectures:
  - Fully connected DNN: 5 layers, 2048 hidden units.

Experimental Setup

- Aurora 4 clean and multi-condition training sets.
- Triphone GMM-HMM.
- Network architectures:
  - Fully connected DNN: 5 layers, 2048 hidden units.
  - State of the art CNN (Sainath et al., 2015).

Experimental Setup

- Aurora 4 clean and multi-condition training sets.
- Triphone GMM-HMM.
- Network architectures:
  - Fully connected DNN: 5 layers, 2048 hidden units.
  - State of the art CNN (Sainath et al., 2015).
  - Very deep CNN, vd6 model proposes by Qian et al. (2016).

Experimental Setup

- Aurora 4 clean and multi-condition training sets.
- Triphone GMM-HMM.
- Network architectures:
  - Fully connected DNN: 5 layers, 2048 hidden units.
  - State of the art CNN (Sainath et al., 2015).
  - Very deep CNN, vd6 model proposes by Qian et al. (2016).
  - ResNet.

Experimental Setup

- Aurora 4 clean and multi-condition training sets.
- Triphone GMM-HMM.
- Network architectures:
  - Fully connected DNN: 5 layers, 2048 hidden units.
  - State of the art CNN (Sainath et al., 2015).
  - Very deep CNN, vd6 model proposes by Qian et al. (2016).
  - ResNet.
- Kaldi + Tensorflow.

Experimental Setup

- Aurora 4 clean and multi-condition training sets.
- Triphone GMM-HMM.
- Network architectures:
  - Fully connected DNN: 5 layers, 2048 hidden units.
  - State of the art CNN (Sainath et al., 2015).
  - Very deep CNN, vd6 model proposes by Qian et al. (2016).
  - ResNet.
- Kaldi + Tensorflow.
- Batch Normalization.

Experimental Setup

- Aurora 4 clean and multi-condition training sets.
- Triphone GMM-HMM.
- Network architectures:
  - Fully connected DNN: 5 layers, 2048 hidden units.
  - State of the art CNN (Sainath et al., 2015).
  - Very deep CNN, vd6 model proposes by Qian et al. (2016).
  - ResNet.
- Kaldi + Tensorflow.
- Batch Normalization.
- Adam optimizer (Kingma and Ba, 2014).

Experimental Setup

- Aurora 4 clean and multi-condition training sets.
- Triphone GMM-HMM.
- Network architectures:
  - Fully connected DNN: 5 layers, 2048 hidden units.
  - State of the art CNN (Sainath et al., 2015).
  - Very deep CNN, vd6 model proposes by Qian et al. (2016).
  - ResNet.
- Kaldi + Tensorflow.
- Batch Normalization.
- Adam optimizer (Kingma and Ba, 2014).
- Cross-entropy as a loss function.

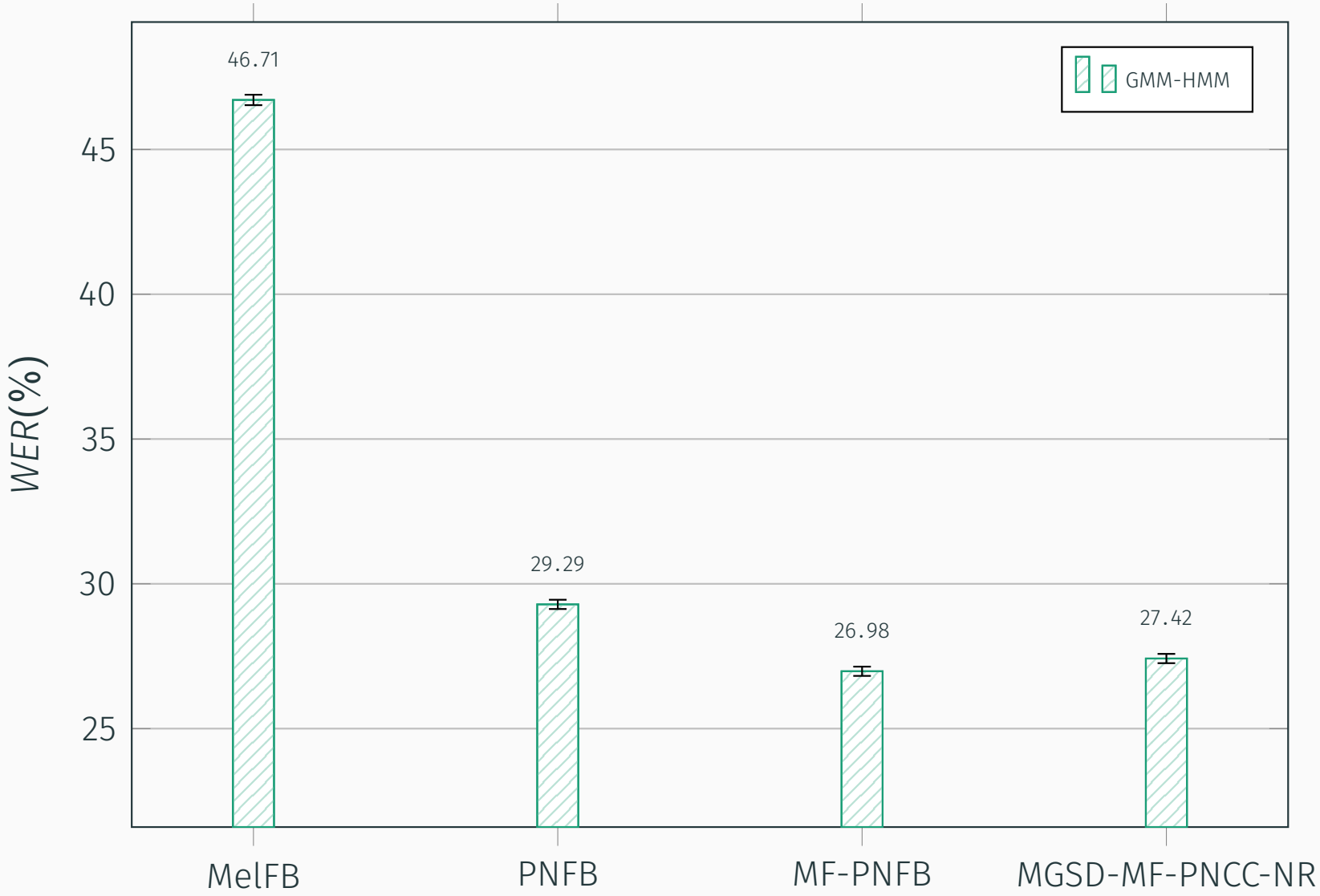
Experimental Setup

- Aurora 4 clean and multi-condition training sets.
- Triphone GMM-HMM.
- Network architectures:
  - Fully connected DNN: 5 layers, 2048 hidden units.
  - State of the art CNN (Sainath et al., 2015).
  - Very deep CNN, vd6 model proposes by Qian et al. (2016).
  - ResNet.
- Kaldi + Tensorflow.
- Batch Normalization.
- Adam optimizer (Kingma and Ba, 2014).
- Cross-entropy as a loss function.
- Xavier initialization.

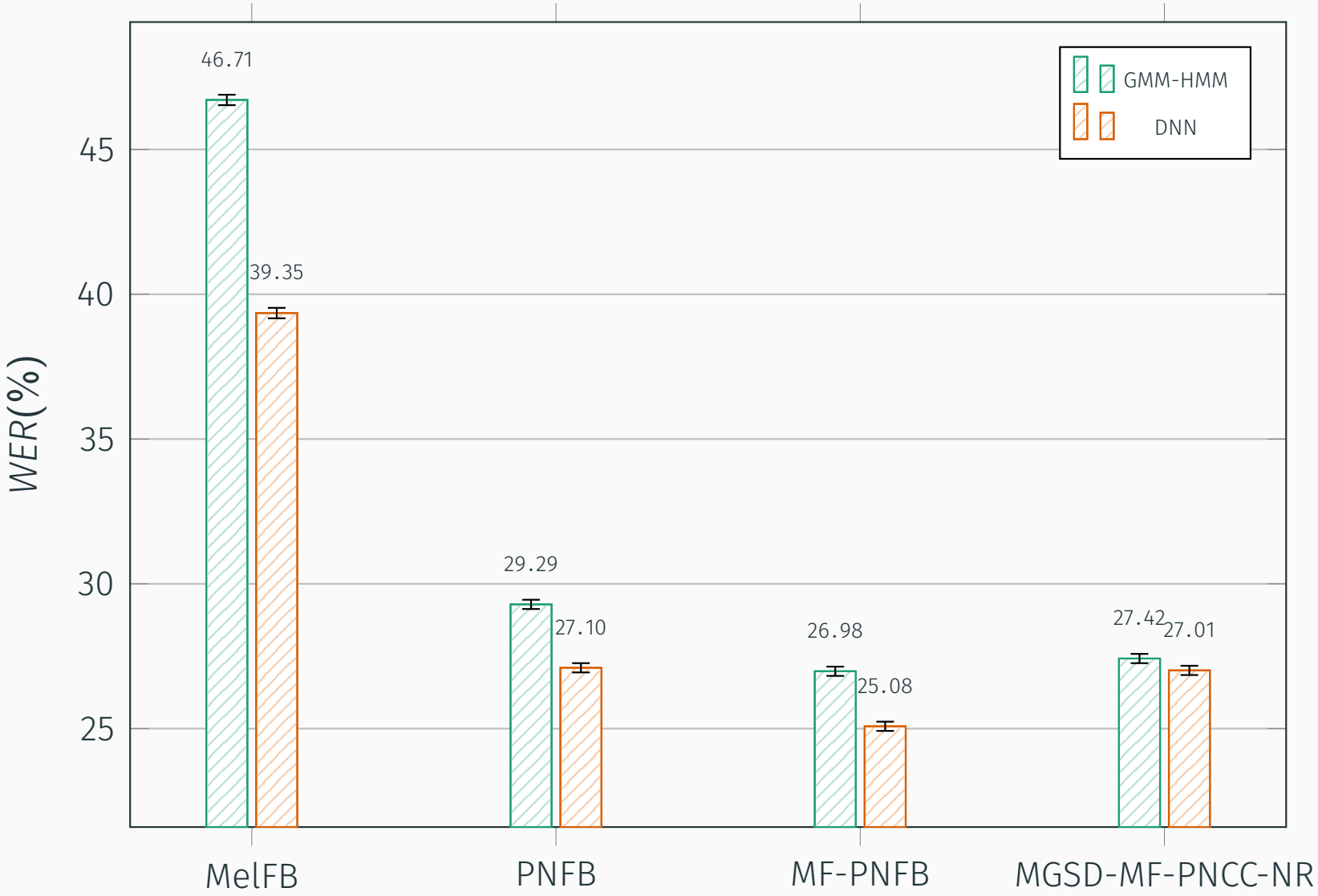
Experimental Setup

- Aurora 4 clean and multi-condition training sets.
- Triphone GMM-HMM.
- Network architectures:
  - Fully connected DNN: 5 layers, 2048 hidden units.
  - State of the art CNN (Sainath et al., 2015).
  - Very deep CNN, vd6 model proposes by Qian et al. (2016).
  - ResNet.
- Kaldi + Tensorflow.
- Batch Normalization.
- Adam optimizer (Kingma and Ba, 2014).
- Cross-entropy as a loss function.
- Xavier initialization.
- Early stopping with 3 retries of patience.

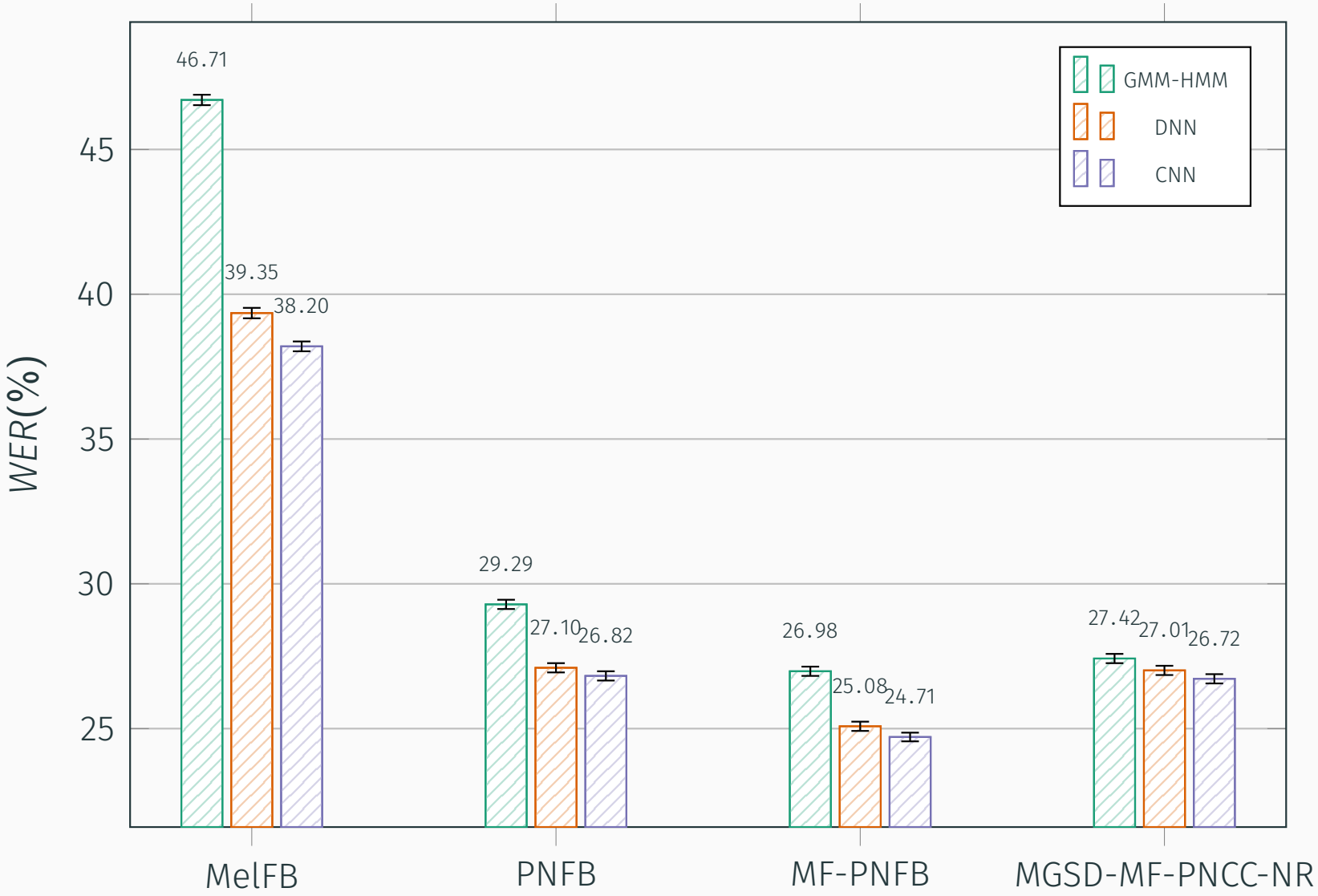
Results Aurora 4: Mismatch Case



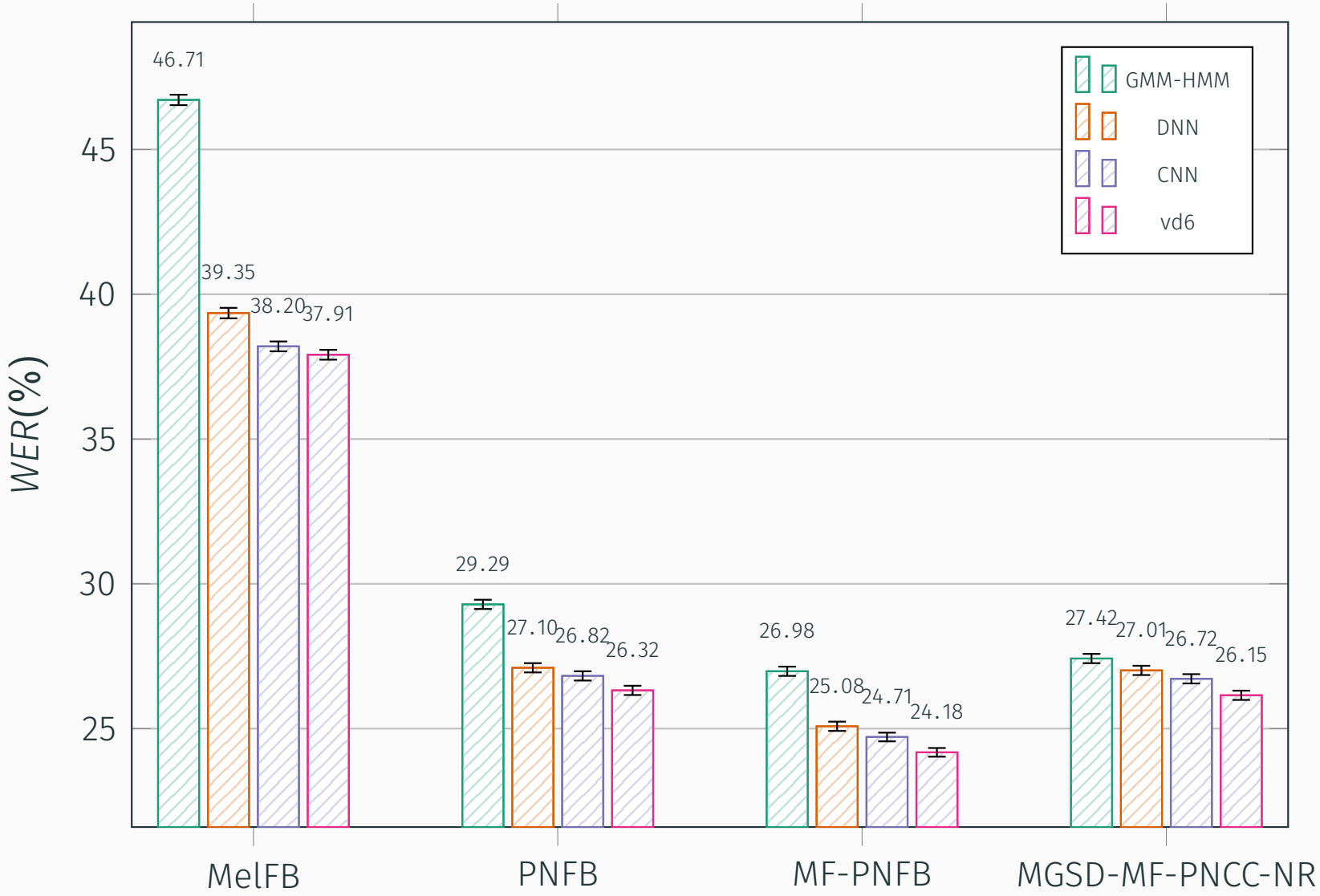
Results Aurora 4: Mismatch Case



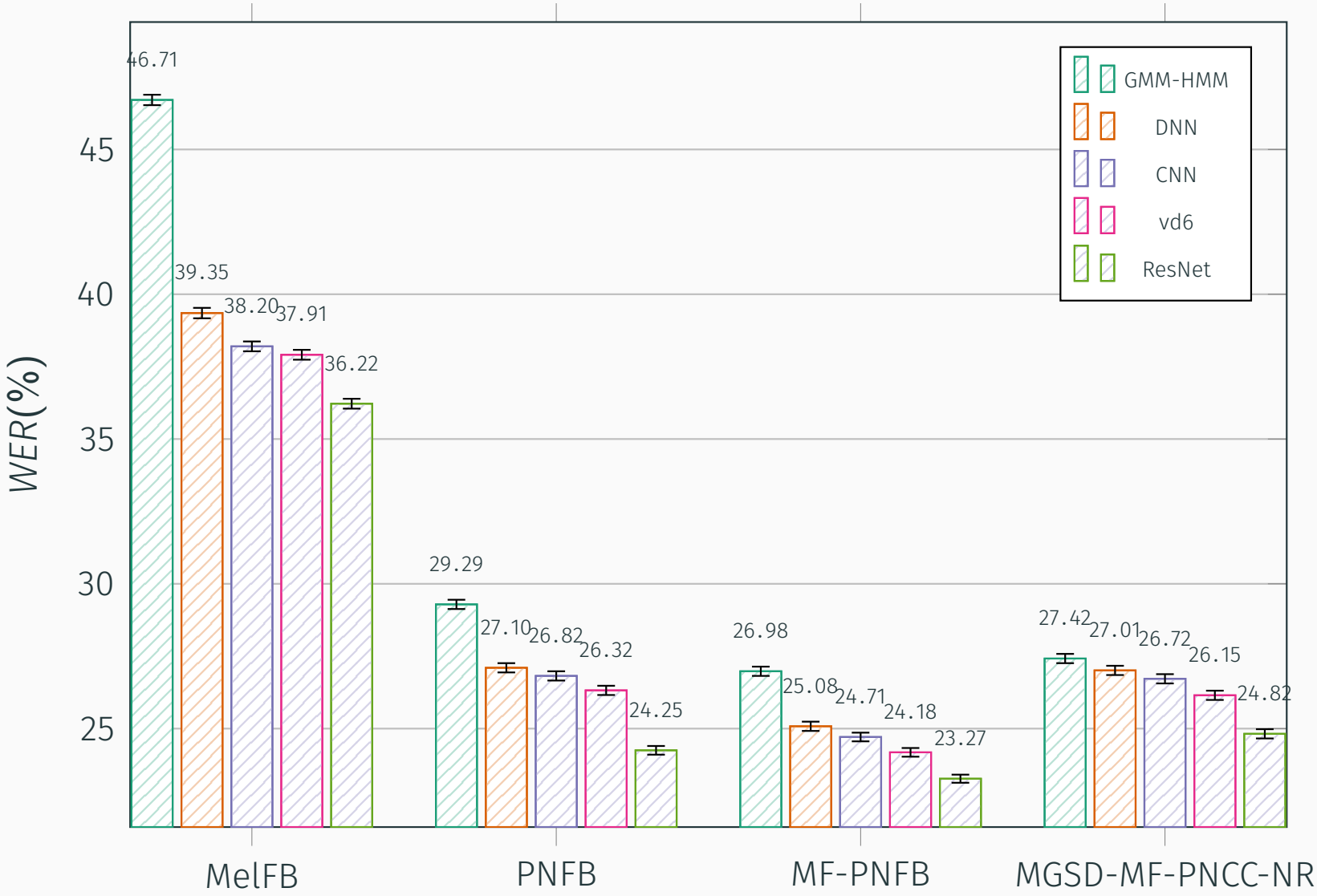
Results Aurora 4: Mismatch Case



Results Aurora 4: Mismatch Case

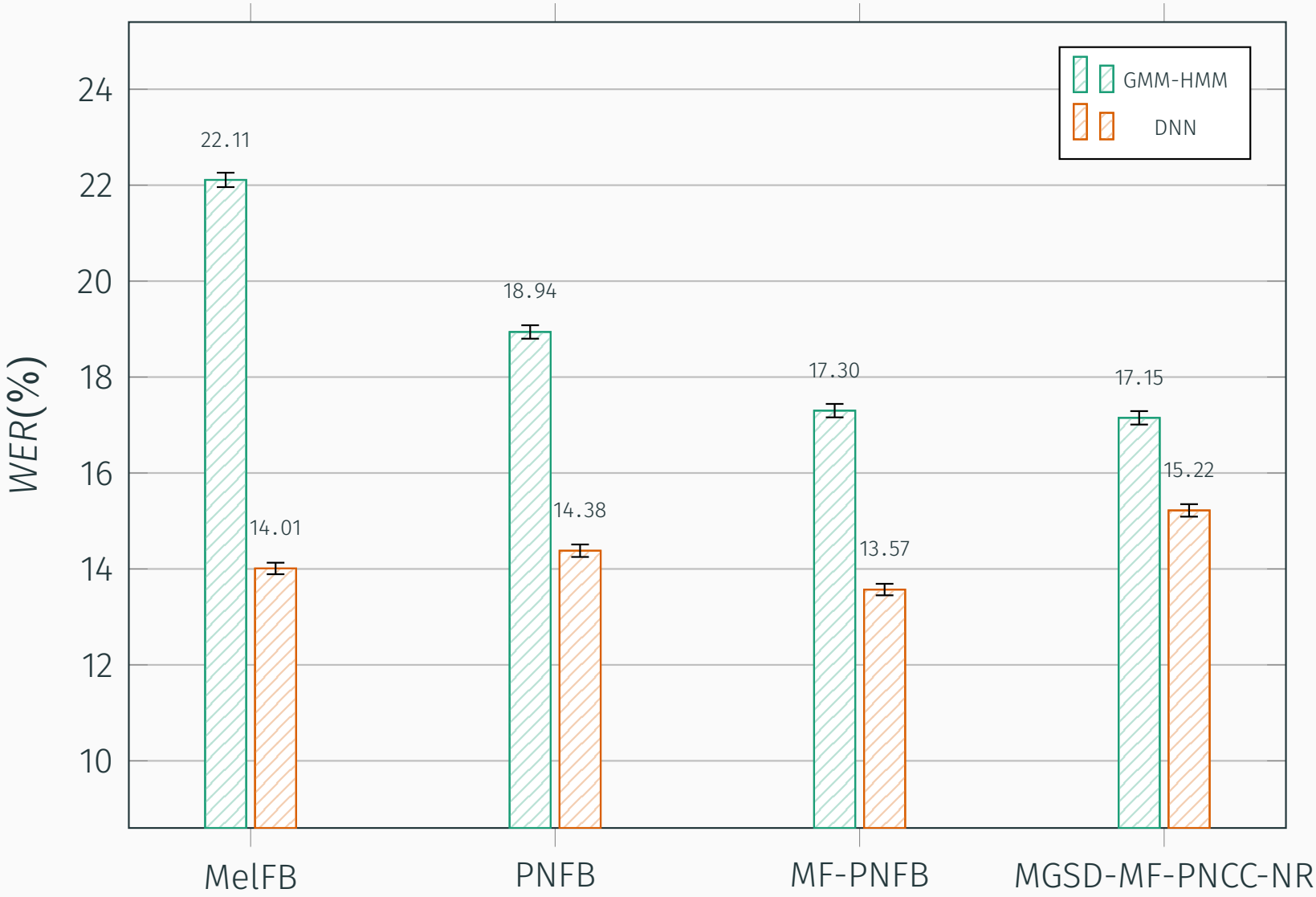


Results Aurora 4: Mismatch Case

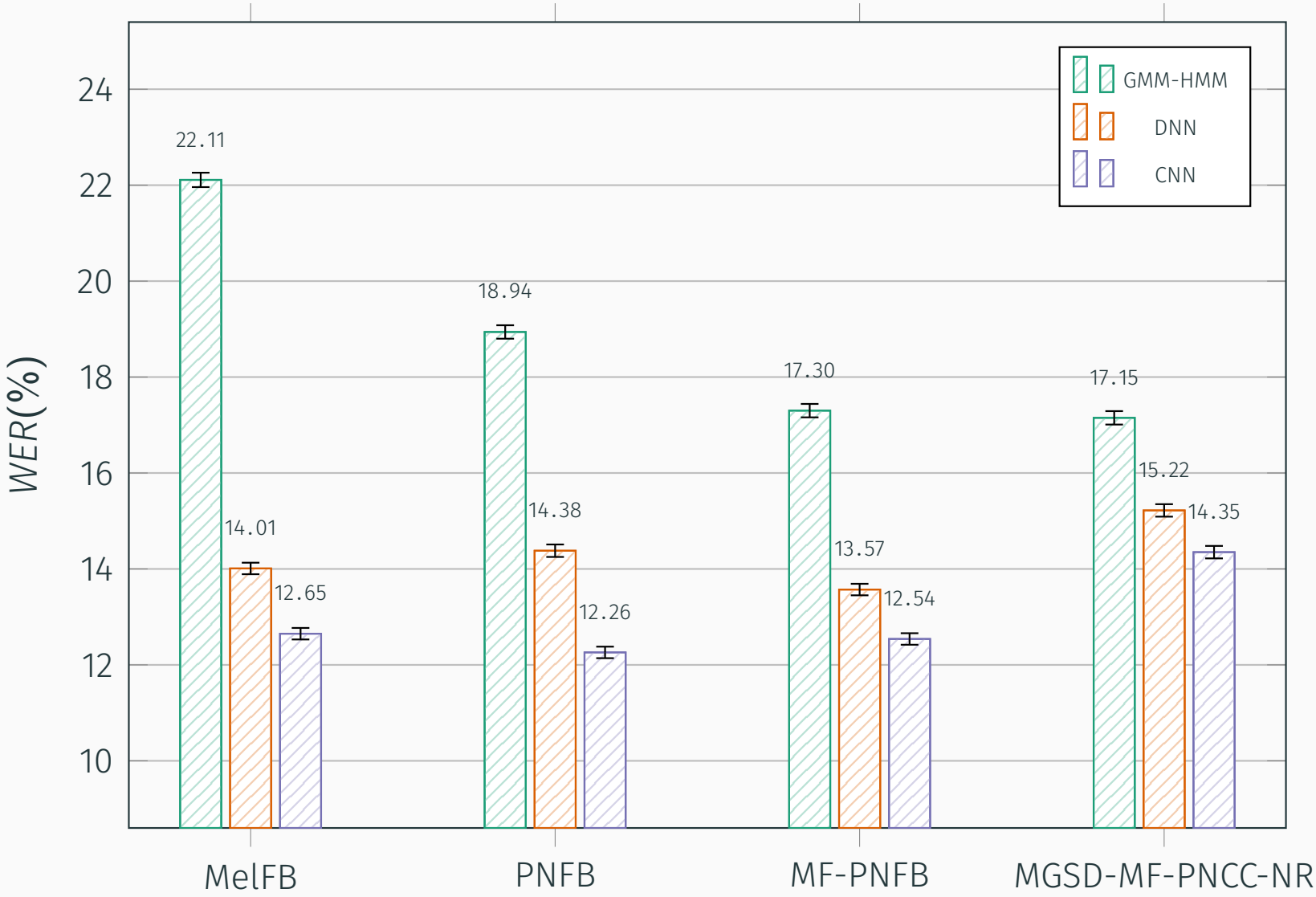


Results Aurora 4: Match Case

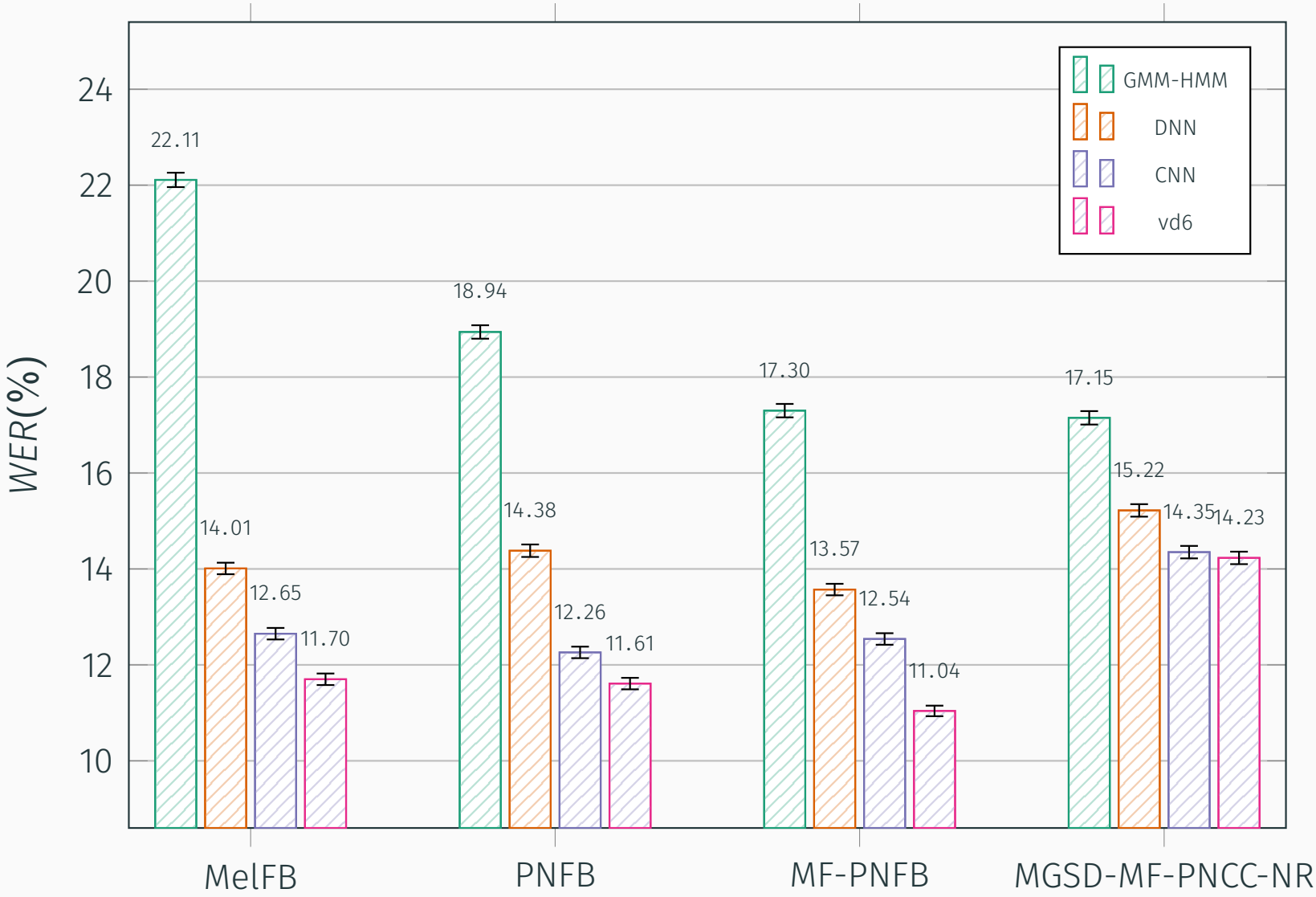
Results Aurora 4: Match Case



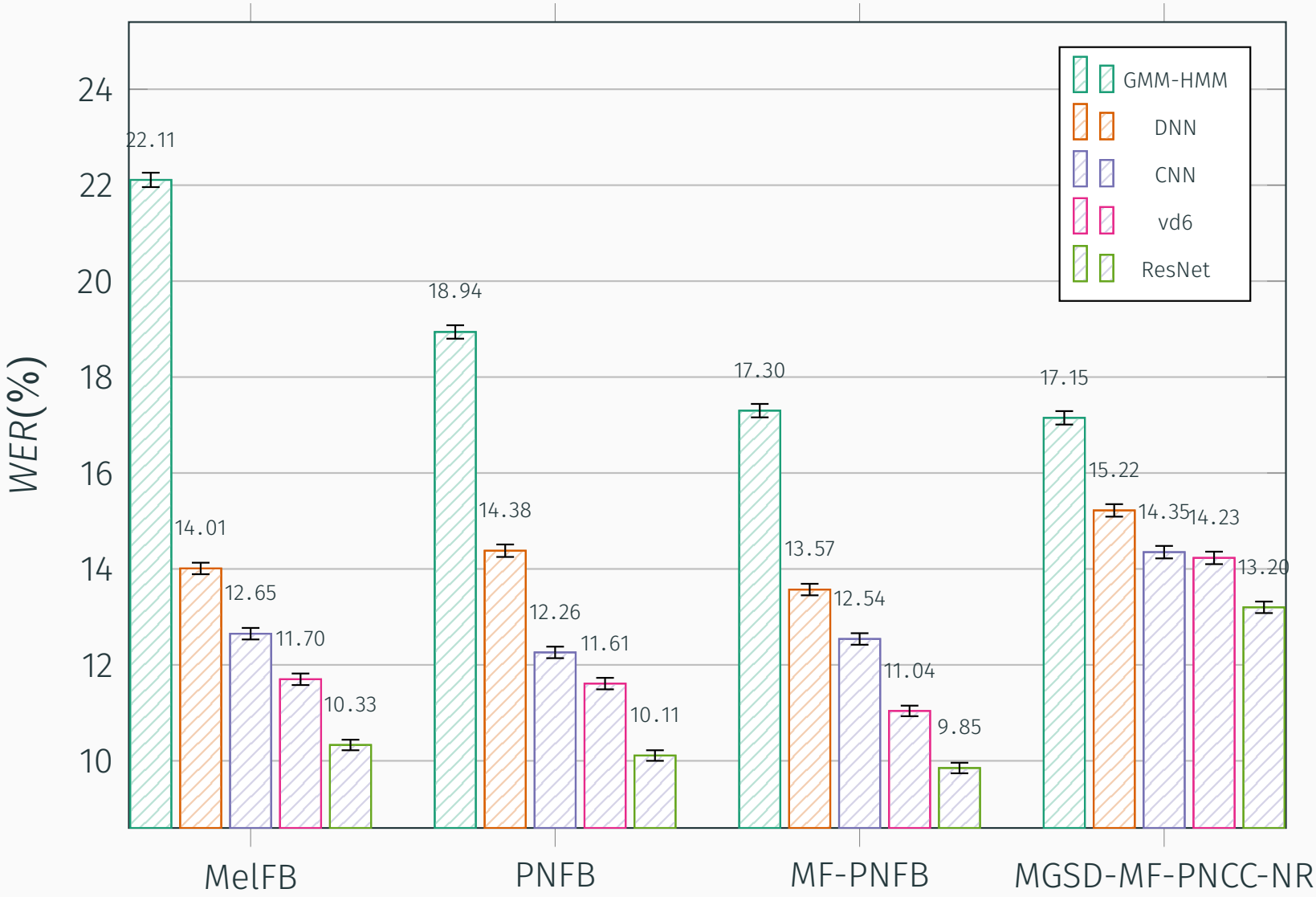
Results Aurora 4: Match Case



Results Aurora 4: Match Case



Results Aurora 4: Match Case



Conclusions

Conclusions: Auditory motivated features: Masking

- Masking modeling based on morphological filtering.

Conclusions: Auditory motivated features: Masking

- Masking modeling based on morphological filtering.
- Integrated into the PNCC.

Conclusions: Auditory motivated features: Masking

- Masking modeling based on morphological filtering.
- Integrated into the PNCC.
- Improves the recognition rates in hybrid and GMM-HMM systems

Conclusions: Auditory motivated features: Masking

- Masking modeling based on morphological filtering.
- Integrated into the PNCC.
- Improves the recognition rates in hybrid and GMM-HMM systems
- Significantly better results than PNCCs in Aurora 2, 4 and WSJ0.

Conclusions: Auditory motivated features: Masking

- Masking modeling based on morphological filtering.
- Integrated into the PNCC.
- Improves the recognition rates in hybrid and GMM-HMM systems
- Significantly better results than PNCCs in Aurora 2, 4 and WSJ0.
- Future work: Initialize first CNNs layers.

Conclusions: Auditory motivated features: Masking

- Masking modeling based on morphological filtering.
- Integrated into the PNCC.
- Improves the recognition rates in hybrid and GMM-HMM systems
- Significantly better results than PNCCs in Aurora 2, 4 and WSJ0.
- Future work: Initialize first CNNs layers.

Conclusions: Auditory motivated features: Masking

- Masking modeling based on morphological filtering.
- Integrated into the PNCC.
- Improves the recognition rates in hybrid and GMM-HMM systems
- Significantly better results than PNCCs in Aurora 2, 4 and WSJ0.
- Future work: Initialize first CNNs layers.

TASLP Journal

F. de-la-Calle-Silos, F.J. Valverde-Albacete, A. Gallardo-Antolín, C. Pelaéz-Moreno. *'Morphologically-filtered power-normalized cochleograms as robust, biologically inspired features for ASR,"* in *IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 11, pp. 2070-2080, Nov. 2015.*

Conclusions: Auditory motivated features: Masking

- Masking modeling based on morphological filtering.
- Integrated into the PNCC.
- Improves the recognition rates in hybrid and GMM-HMM systems
- Significantly better results than PNCCs in Aurora 2, 4 and WSJ0.
- Future work: Initialize first CNNs layers.

TASLP Journal

F. de-la-Calle-Silos, F.J. Valverde-Albacete, A. Gallardo-Antolín, C. Pelaéz-Moreno. *'Morphologically-filtered power-normalized cochleograms as robust, biologically inspired features for ASR,"* in *IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 11, pp. 2070-2080, Nov. 2015.*

Interspeech Conference

F. de-la-Calle-Silos, F.J. Valverde-Albacete, A. Gallardo-Antolín, C. Pelaéz-Moreno. *"ASR Feature Extraction with Morphologically-Filtered Power-Normalized Cochleograms"*, in *Proceedings of Interspeech (Annual Conference of the International Speech Communication Association), Sep. 2014.*

Conclusions: Auditory motivated features: Syncrony

- Model of the synchrony effect.

Conclusions: Auditory motivated features: Syncrony

- Model of the synchrony effect.
- MALSR and MGSD.

Conclusions: Auditory motivated features: Syncrony

- Model of the synchrony effect.
- MALSR and MGSD.
- Noise removal technique based on the PNCC.

Conclusions: Auditory motivated features: Syncrony

- Model of the synchrony effect.
- MALSR and MGSD.
- Noise removal technique based on the PNCC.
- Significantly better results than PNCCs in white noise and reverberant conditions.

Conclusions: Auditory motivated features: Syncrony

- Model of the synchrony effect.
- MALSR and MGSD.
- Noise removal technique based on the PNCC.
- Significantly better results than PNCCs in white noise and reverberant conditions.
- Future work: Other references to synchronize and better PNCC integration.

Conclusions: Auditory motivated features: Syncrony

- Model of the synchrony effect.
- MALSR and MGSD.
- Noise removal technique based on the PNCC.
- Significantly better results than PNCCs in white noise and reverberant conditions.
- Future work: Other references to synchronize and better PNCC integration.

Conclusions: Auditory motivated features: Syncrony

- Model of the synchrony effect.
- MALSR and MGSD.
- Noise removal technique based on the PNCC.
- Significantly better results than PNCCs in white noise and reverberant conditions.
- Future work: Other references to synchronize and better PNCC integration.

SPL Journal

F. de-la-Calle-Silos and Richard M. Stern. *'Synchrony-Based Feature Extraction for Robust Automatic Speech Recognition,"* in *IEEE Signal Processing Letters, vol. 24, no. 8, pp. 1158, Aug. 2017*

Conclusions: Robust Deep Learning

- Novel architecture based on ResNets.

Conclusions: Robust Deep Learning

- Novel architecture based on ResNets.
- Significantly better results than state of the art networks in Aurora 4 dataset.

Conclusions: Robust Deep Learning

- Novel architecture based on ResNets.
- Significantly better results than state of the art networks in Aurora 4 dataset.
- Combination with our features: PNFB with masking modeling achieves a significant improvement with respect to conventional features.

Conclusions: Robust Deep Learning

- Novel architecture based on ResNets.
- Significantly better results than state of the art networks in Aurora 4 dataset.
- Combination with our features: PNFB with masking modeling achieves a significant improvement with respect to conventional features.
- Future work: combination with RNNs

Conclusions: Robust Deep Learning

- Novel architecture based on ResNets.
- Significantly better results than state of the art networks in Aurora 4 dataset.
- Combination with our features: PNFB with masking modeling achieves a significant improvement with respect to conventional features.
- Future work: combination with RNNs

Conclusions: Robust Deep Learning

- Novel architecture based on ResNets.
- Significantly better results than state of the art networks in Aurora 4 dataset.
- Combination with our features: PNFB with masking modeling achieves a significant improvement with respect to conventional features.
- Future work: combination with RNNs

IberSpeech Conference

F. de-la-Calle-Silos, A. Gallardo-Antolín, C. Pelaéz-Moreno. *"Deep Maxout Networks applied to Noise-Robust Speech Recognition"*, *Advances in Speech and Language Technologies for Iberian Languages. Communications in Computer and Information Science, Springer, 2014.*

Conclusions: Robust Deep Learning

- Novel architecture based on ResNets.
- Significantly better results than state of the art networks in Aurora 4 dataset.
- Combination with our features: PNFB with masking modeling achieves a significant improvement with respect to conventional features.
- Future work: combination with RNNs

IberSpeech Conference

F. de-la-Calle-Silos, A. Gallardo-Antolín, C. Pelaéz-Moreno. *"Deep Maxout Networks applied to Noise-Robust Speech Recognition"*, *Advances in Speech and Language Technologies for Iberian Languages. Communications in Computer and Information Science, Springer, 2014.*

Journal

F. de-la-Calle-Silos, A. Gallardo-Antolín, C. Pelaéz-Moreno. *"Deep Residual Networks with Auditory Inspired Features for Robust Speech Recognition"* in *Expert Systems with Applications.*

Conclusions

- Integration into SAVIER research demonstrator at AIRBUS.

Conclusions

- Integration into SAVIER research demonstrator at AIRBUS.

Conclusions

- Integration into SAVIER research demonstrator at AIRBUS.

References

M. Berouti, R. Schwartz, and J. Makhoul. Enhancement of speech corrupted by acoustic noise. In *Acoustics, Speech and Signal Processing (ICASSP), 1979 IEEE International Conference on*, volume 4, pages 208–211, 1979.

H.A. Bourlard and N. Morgan. *Connectionist Speech Recognition: A Hybrid Approach*. Kluwer international series in engineering and computer science: VLSI, computer architecture, and digital signal processing. Springer US, 1994.

References ii

S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust., Speech and Signal Processing*, 28(4): 357–366, 1980.

Vincent Dumoulin and Francesco Visin. A guide to convolution arithmetic for deep learning. *CoRR*, abs/1603.07285, 2016.

Hugo Fastl and Eberhard Zwicker. *Psycho-acoustics: Facts and Models*. Springer, 3 edition, 2007.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.

H. Hermansky and N. Morgan. Rasta processing of speech. *Speech and Audio Processing, IEEE Transactions on*, 2(4):578–589, Oct 1994.

References iii

H. Hermansky, B. Hanson, and H. Wakita. Perceptually based linear predictive analysis of speech. In *Acoustics, Speech and Signal Processing (ICASSP), 1985 IEEE International Conference on*, volume 10, pages 509–512, Apr 1985.

W Jesteadt, S P Bacon, and JR Lehman. Forward masking as a function of frequency, masker level, and signal delay. *The Journal of the Acoustical Society of America*, 71(4):950 – 962, 1982.

C. Kim and R. M. Stern. Power-normalized cepstral coefficients (PNCC) for robust speech recognition. *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, 24(7):1315–1329, July 2016.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

References iv

P. J. Moreno, B. Raj, and R. M. Stern. A vector taylor series approach for environment-independent speech recognition. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 2, pages 733–736 vol. 2, May 1996.

Y. Qian, M. Bi, T. Tan, and K. Yu. Very deep convolutional neural networks for noise robust speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(12):2263–2276, Dec 2016. ISSN 2329-9290.

J E Rose, J E Hind, D J Anderson, and J F Brugge. Some effects of stimulus intensity on response of auditory nerve fibers in the squirrel monkey. *Journal of Neurophysiology*, 34(4):685–699, 1971. ISSN 0022-3077.

References v

Tara N. Sainath, Brian Kingsbury, George Saon, Hagen Soltau, Abdel rahman Mohamed, George Dahl, and Bhuvana Ramabhadran. Deep convolutional neural networks for large-scale speech tasks. *Neural Networks*, 64:39 – 48, 2015. Special Issue on "Deep Learning of Representations".

Stephanie Seneff. A joint synchrony/mean-rate model of auditory speech processing. In Alex Waibel and Kai-Fu Lee, editors, *Readings in Speech Recognition*, pages 101–111. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990. ISBN 1-55860-124-4.

E. D. Young and M. B. Sachs. Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers. *J. Acoustic. Soc. Amer.*, 66:1381–1403, 1979.

Xuedong Zhang, Michael G. Heinz, Ian C. Bruce, and Laurel H. Carney. A phenomenological model for the responses of auditory-nerve fibers: I. Nonlinear tuning with compression and suppression. *The Journal of the Acoustical Society of America*, 109(2):648, 2001.