Computational speech segregation based on an auditory-inspired modulation analysis

Tobias May
Torsten Dau

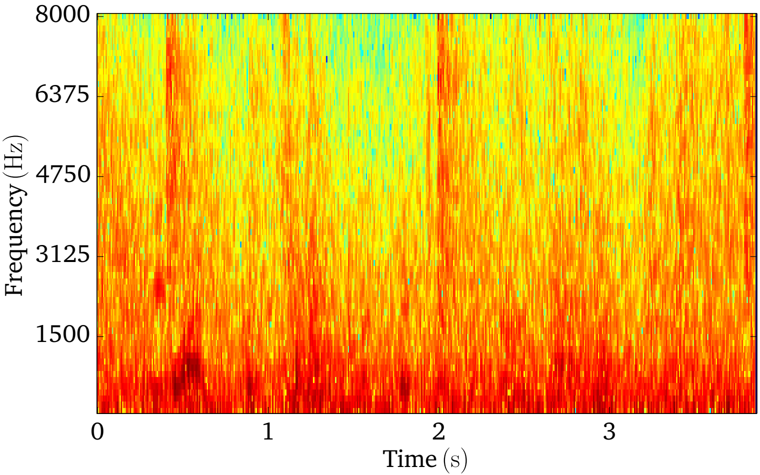
Centre for Applied Hearing Research
Department of Electrical Engineering
Technical University of Denmark

8.1.2015

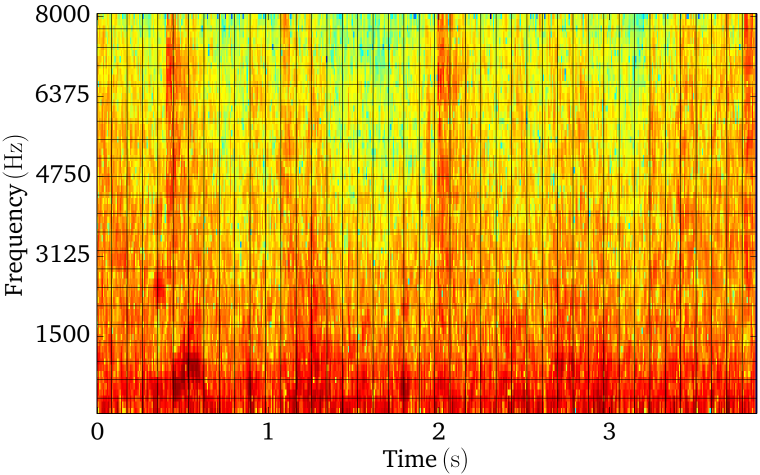
Problem definition

Problem definition

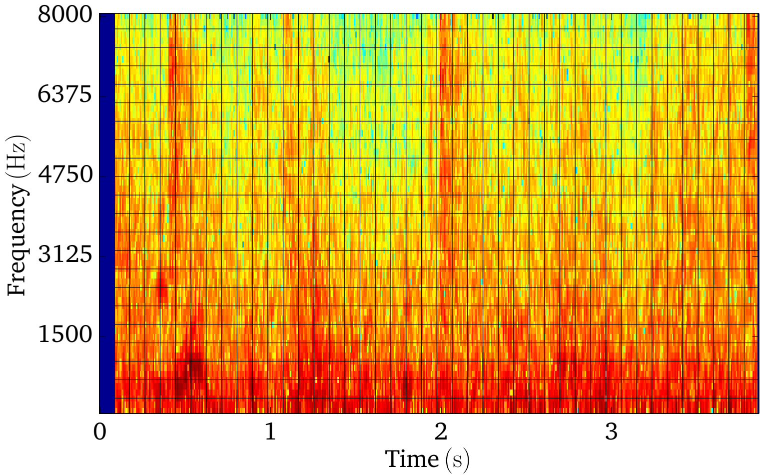
Problem definition



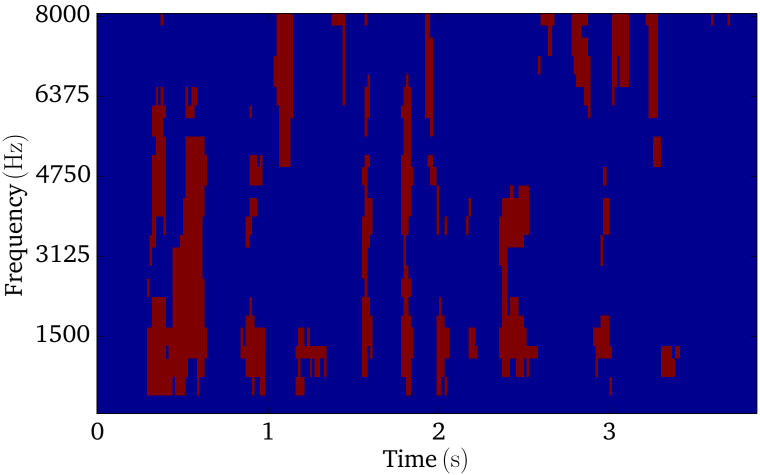
Problem definition



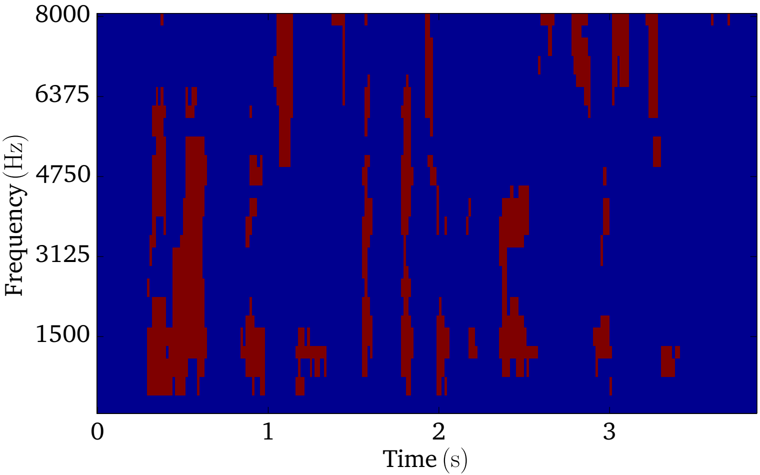
Problem definition



Problem definition



Problem definition



Segregation in the time-frequency (T-F) domain

The concept of the ideal binary mask (IBM):

1. **Segmentation:** Decompose input into individual T-F units

Segregation in the time-frequency (T-F) domain

The concept of the ideal binary mask (IBM):

1. **Segmentation:** Decompose input into individual T-F units
2. **Grouping:** Identify reliable T-F units based on *a priori* SNR

Segregation in the time-frequency (T-F) domain

The concept of the ideal binary mask (IBM):

1. **Segmentation:** Decompose input into individual T-F units
2. **Grouping:** Identify reliable T-F units based on *a priori* SNR

Applications of the IBM:

- Improve speech intelligibility in noise
  (Brungart *et al.*, 2006; Li and Loizou, 2008; Kjems *et al.*, 2009)

Segregation in the time-frequency (T-F) domain

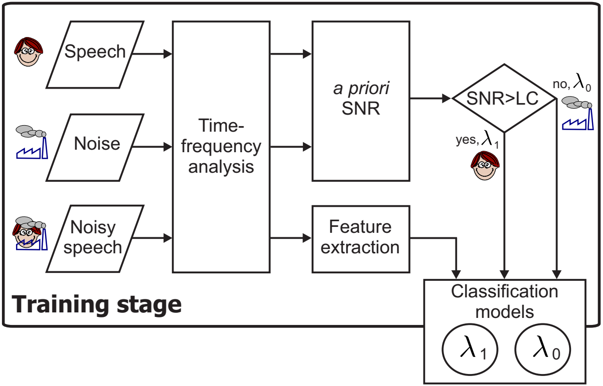
The concept of the ideal binary mask (IBM):

1. **Segmentation:** Decompose input into individual T-F units
2. **Grouping:** Identify reliable T-F units based on *a priori* SNR

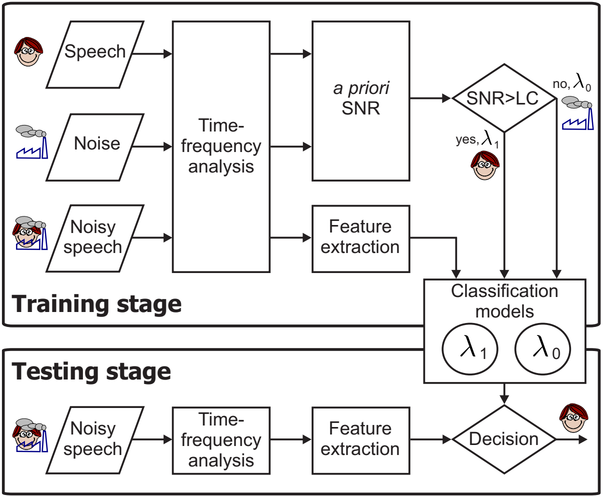
Applications of the IBM:

- Improve speech intelligibility in noise
  (Brungart *et al.*, 2006; Li and Loizou, 2008; Kjems *et al.*, 2009)
- Automatic speech recognition and speaker identification
  (Cooke *et al.*, 2001; May *et al.*, 2012)

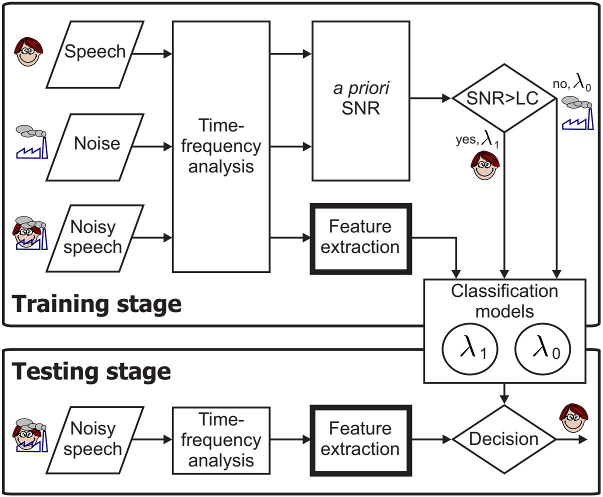
How to estimate the IBM in realistic scenarios?



How to estimate the IBM in realistic scenarios?

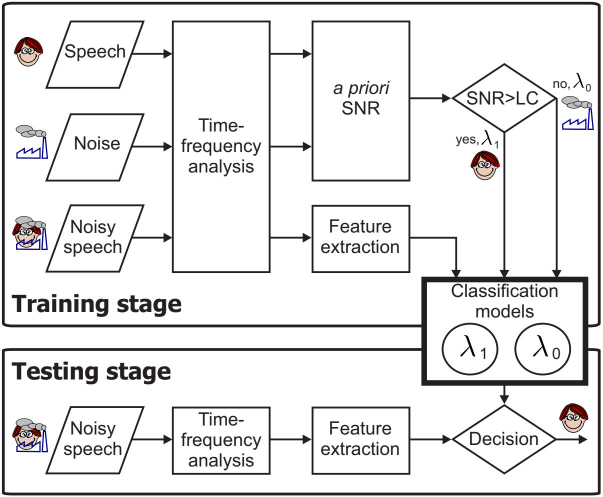


How to estimate the IBM in realistic scenarios?



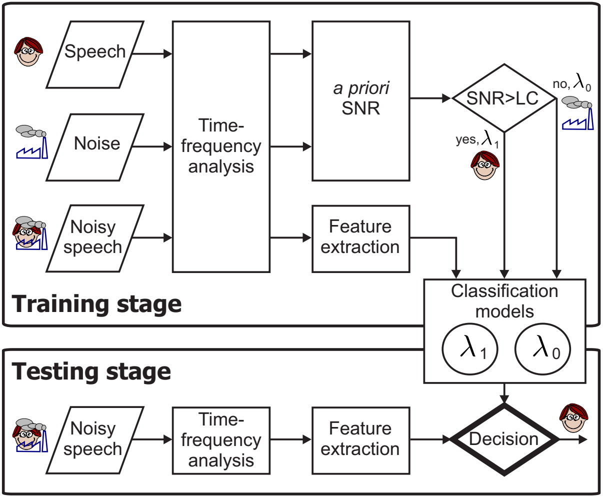
? Influence of feature representation

How to estimate the IBM in realistic scenarios?



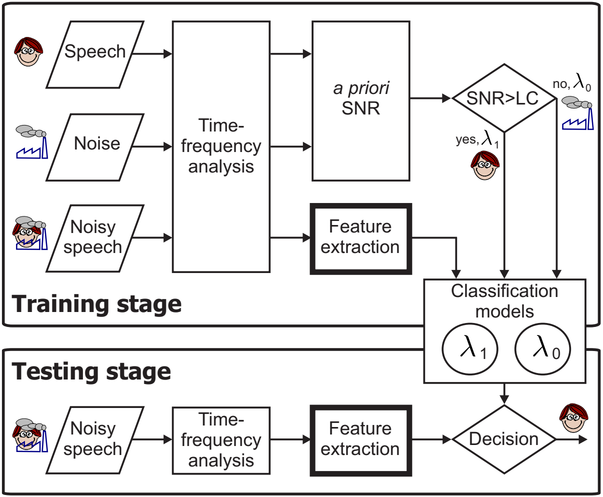
- ? Influence of feature representation
- ? Generalization to *unseen* acoustic conditions

How to estimate the IBM in realistic scenarios?



- ? Influence of feature representation
- ? Generalization to *unseen* acoustic conditions
- ? Contribution of spectro-temporal context

How to estimate the IBM in realistic scenarios?



- ? Influence of feature representation
- ? Generalization to *unseen* acoustic conditions
- ? Contribution of spectro-temporal context

Auditory-inspired features for speech segregation

- Recent studies exploit between 45-90 feature dimensions (Kim *et al.*, 2009; Han and Wang, 2012; Wang and Wang, 2013; Healy *et al.*, 2013)

Auditory-inspired features for speech segregation

- Recent studies exploit between 45-90 feature dimensions
  (Kim *et al.*, 2009; Han and Wang, 2012; Wang and Wang, 2013; Healy *et al.*, 2013)
  - Contribution of individual features is difficult to assess

Auditory-inspired features for speech segregation

- Recent studies exploit between 45-90 feature dimensions (Kim *et al.*, 2009; Han and Wang, 2012; Wang and Wang, 2013; Healy *et al.*, 2013)
  - Contribution of individual features is difficult to assess
- All studies used linear amplitude modulation spectrogram (AMS) features (Kollmeier and Koch, 1994; Tchorz and Kollmeier, 2003)

Auditory-inspired features for speech segregation

- Recent studies exploit between 45-90 feature dimensions
  (Kim *et al.*, 2009; Han and Wang, 2012; Wang and Wang, 2013; Healy *et al.*, 2013)
  - Contribution of individual features is difficult to assess
- All studies used linear amplitude modulation spectrogram (AMS) features (Kollmeier and Koch, 1994; Tchorz and Kollmeier, 2003)
  - Not consistent with psychoacoustic data on modulation detection
    (Bacon and Grantham, 1989; Dau *et al.*, 1997; Ewert and Dau, 2000)

Auditory-inspired features for speech segregation

- Recent studies exploit between 45-90 feature dimensions (Kim *et al.*, 2009; Han and Wang, 2012; Wang and Wang, 2013; Healy *et al.*, 2013)
  - ⊖ Contribution of individual features is difficult to assess
- All studies used linear amplitude modulation spectrogram (AMS) features (Kollmeier and Koch, 1994; Tchorz and Kollmeier, 2003)
  - ⊖ Not consistent with psychoacoustic data on modulation detection (Bacon and Grantham, 1989; Dau *et al.*, 1997; Ewert and Dau, 2000)

Approach:

1. Analyze role of modulation features for speech segregation

Auditory-inspired features for speech segregation

- Recent studies exploit between 45-90 feature dimensions (Kim *et al.*, 2009; Han and Wang, 2012; Wang and Wang, 2013; Healy *et al.*, 2013)
  - ⊖ Contribution of individual features is difficult to assess
- All studies used linear amplitude modulation spectrogram (AMS) features (Kollmeier and Koch, 1994; Tchorz and Kollmeier, 2003)
  - ⊖ Not consistent with psychoacoustic data on modulation detection (Bacon and Grantham, 1989; Dau *et al.*, 1997; Ewert and Dau, 2000)

Approach:

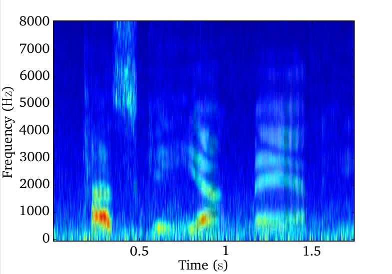
1. Analyze role of modulation features for speech segregation
2. Compare linearly- and logarithmically-scaled modulation filters

Amplitude modulation spectrogram (AMS)

1. Compute spectrogram based on $4\,\mathrm{ms}$ segments

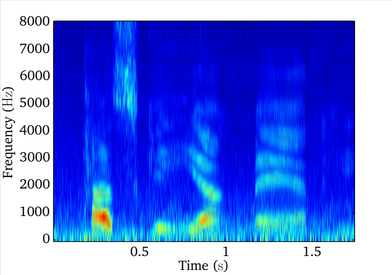
linear AMS features

- 2D spectrogram



logarithmic AMS features

- 2D spectrogram

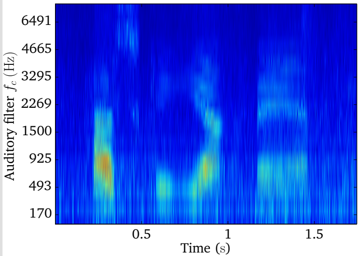


Amplitude modulation spectrogram (AMS)

1. Compute spectrogram based on $4\,\mathrm{ms}$ segments
2. Analyze 25 auditory filters between 80 and $8000\,\mathrm{Hz}$

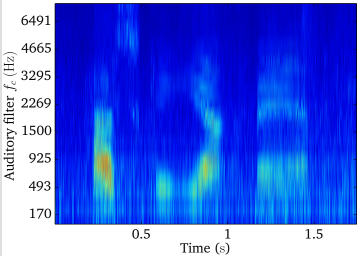
linear AMS features

- 2D auditory spectrogram



logarithmic AMS features

- 2D auditory spectrogram

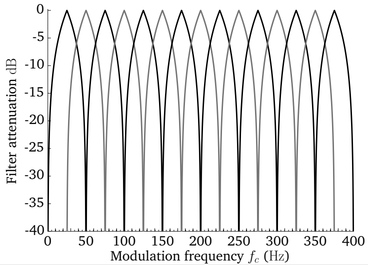


Amplitude modulation spectrogram (AMS)

1. Compute spectrogram based on $4\,\mathrm{ms}$ segments
2. Analyze 25 auditory filters between 80 and $8000\,\mathrm{Hz}$
3. Apply modulation filterbank

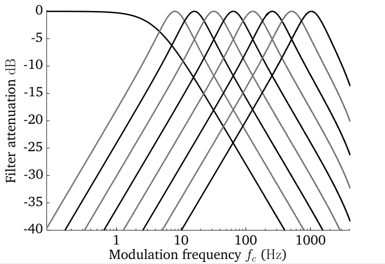
linear AMS features

- 15 filters, linear scale



logarithmic AMS features

- 9 filters, logarithmic scale

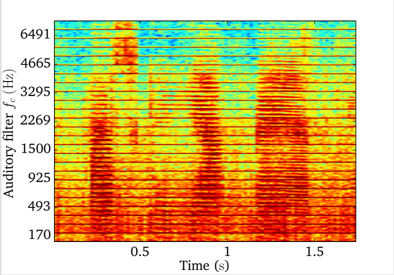


Amplitude modulation spectrogram (AMS)

1. Compute spectrogram based on $4\,\mathrm{ms}$ segments
2. Analyze 25 auditory filters between 80 and $8000\,\mathrm{Hz}$
3. Apply modulation filterbank

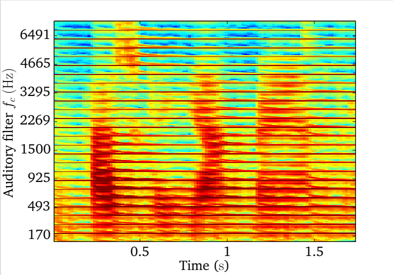
linear AMS features

- 3D modulation spectrogram



logarithmic AMS features

- 3D modulation spectrogram



Segregation experiment I

Speech segregation system:

- GMM classifier trained with linear or logarithmic AMS features

Segregation experiment I

Speech segregation system:

- GMM classifier trained with linear or logarithmic AMS features

Training:

- 100 HINT sentences
- mixed at $-5, 0, 5\,\mathrm{dB}$ SNR
- ICRA1, ICRA7, PSAM $8\mathrm{Hz}$, traffic, music, destroyer, factory

Segregation experiment I

Speech segregation system:

- GMM classifier trained with linear or logarithmic AMS features

Training:

- 100 HINT sentences
- mixed at $-5, 0, 5\,\mathrm{dB}$ SNR
- ICRA1, ICRA7, PSAM $8\mathrm{Hz}$, traffic, music, destroyer, factory

Testing:

- 60 HINT sentences
- mixed at $-5\,\mathrm{dB}$ SNR
- ICRA1, ICRA7, PSAM $8\mathrm{Hz}$, traffic, music, destroyer, factory (*unknown* realizations)

Segregation experiment I

Speech segregation system:

- GMM classifier trained with linear or logarithmic AMS features

Training:

- 100 HINT sentences
- mixed at $-5, 0, 5\,\mathrm{dB}$ SNR
- ICRA1, ICRA7, PSAM $8\mathrm{Hz}$, traffic, music, destroyer, factory

Testing:

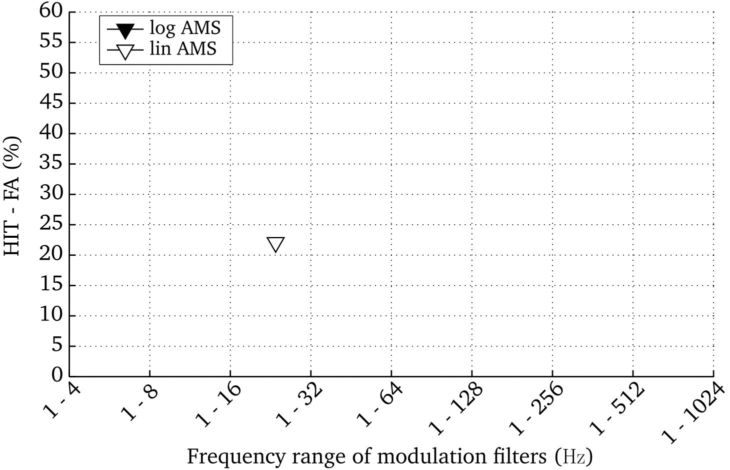
- 60 HINT sentences
- mixed at $-5\,\mathrm{dB}$ SNR
- ICRA1, ICRA7, PSAM $8\mathrm{Hz}$, traffic, music, destroyer, factory (*unknown* realizations)

Evaluation:

✓ Measure HIT-FA, which correlates with speech intelligibility

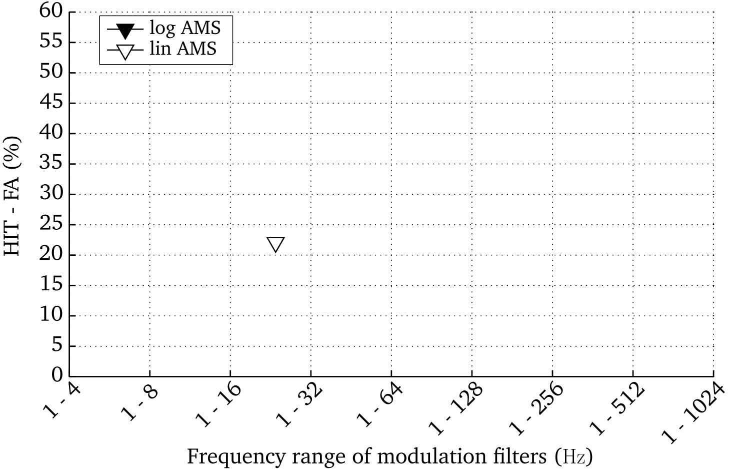
Contribution of individual modulation filters

- Noisy speech at $-5\,\mathrm{dB}$ SNR



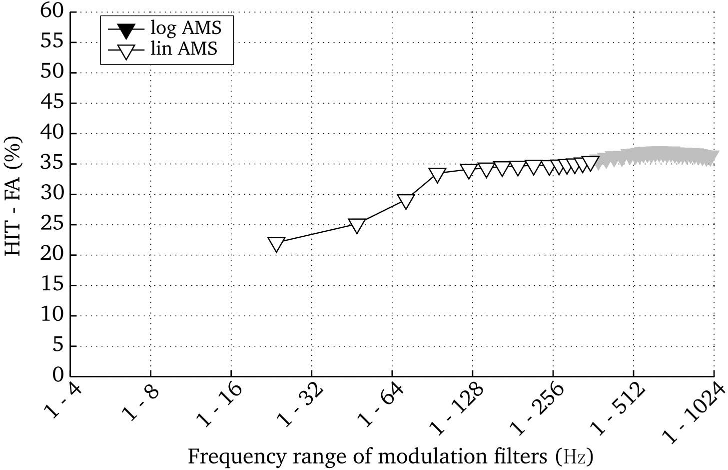
Contribution of individual modulation filters

- Noisy speech at $-5\,\mathrm{dB}$ SNR



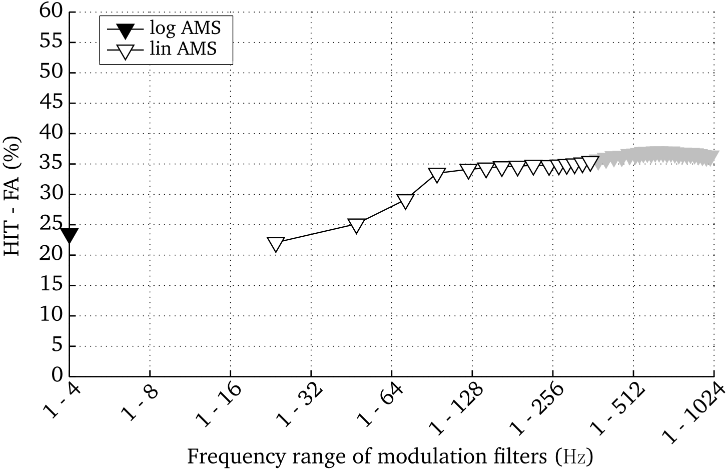
Contribution of individual modulation filters

- Noisy speech at $-5\,\mathrm{dB}$ SNR



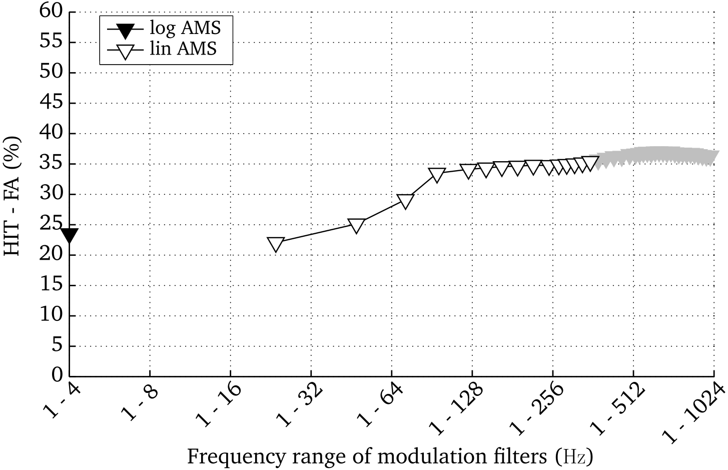
Contribution of individual modulation filters

- Noisy speech at $-5\,\mathrm{dB}$ SNR

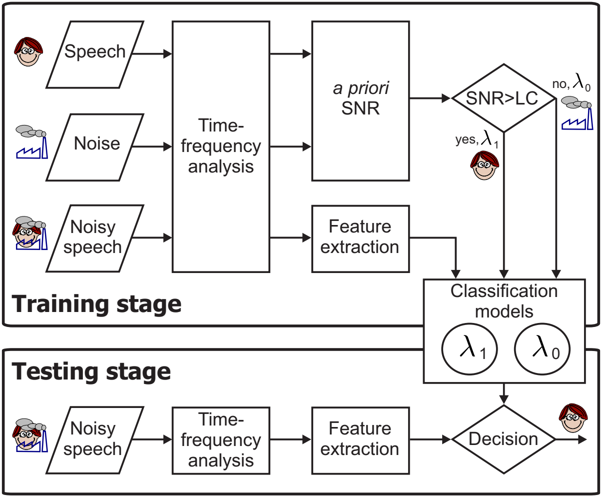


Contribution of individual modulation filters

- Noisy speech at $-5\,\mathrm{dB}$ SNR

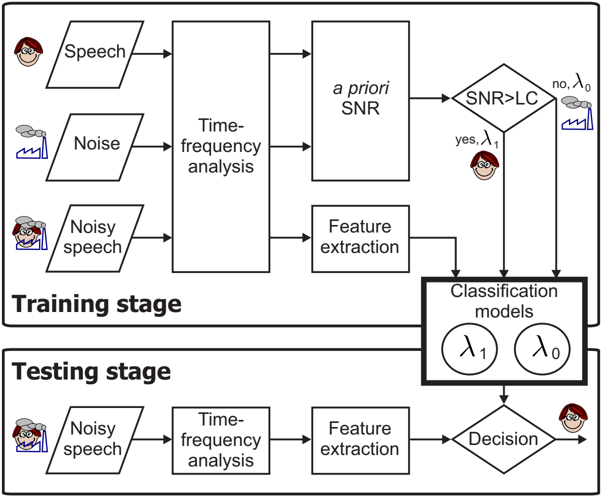


How to estimate the IBM in realistic scenarios?



- ✓ Influence of feature representation
- ? Generalization to *unseen* acoustic conditions
- ? Contribution of spectro-temporal context

How to estimate the IBM in realistic scenarios?



- ✓ Influence of feature representation
- ? Generalization to *unseen* acoustic conditions
- ? Contribution of spectro-temporal context

Segregation experiment I

Segregation system:

- GMM classifier trained with linear or logarithmic AMS features

Training:

- 100 HINT sentences
- mixed at $-5, 0, 5\,\mathrm{dB}$ SNR
- ICRA1, ICRA7, PSAM $8\,\mathrm{Hz}$, traffic, music, destroyer, factory

Testing:

- 60 HINT sentences
- mixed at $-5\,\mathrm{dB}$ SNR
- ICRA1, ICRA7, PSAM $8\,\mathrm{Hz}$, traffic, music, destroyer, factory (*unknown* realizations)

Evaluation:

✓ Measure HIT-FA, which correlates with speech intelligibility

Segregation experiment I

Segregation system:

- GMM classifier trained with linear or logarithmic AMS features

Training:

- 100 HINT sentences
- mixed at $-5, 0, 5\,\mathrm{dB}$ SNR
- ICRA1, ICRA7, PSAM $8\,\mathrm{Hz}$, traffic, music, destroyer, factory

Testing:

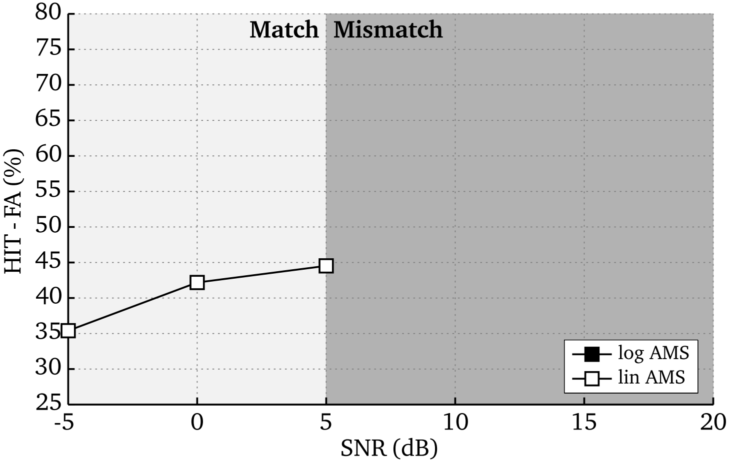
- 60 HINT sentences
- mixed at $-5 : 5 : 20\,\mathrm{dB}$ SNR
- ICRA1, ICRA7, PSAM $8\,\mathrm{Hz}$, traffic, music, destroyer, factory (*unknown* realizations)

Evaluation:

✓ Measure HIT-FA, which correlates with speech intelligibility

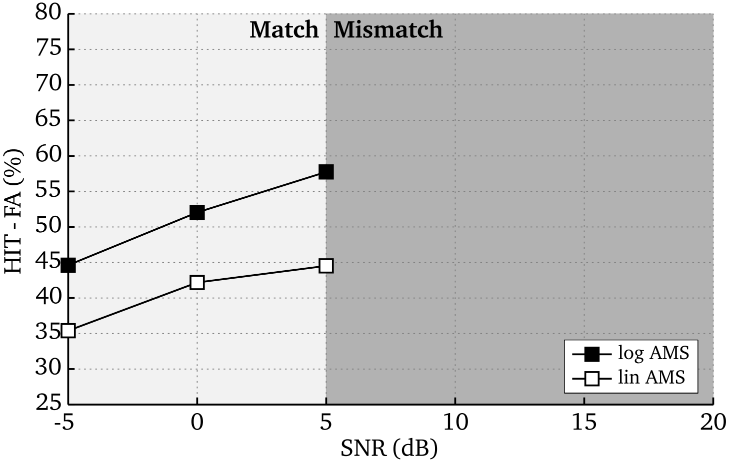
Speech segregation performance

- Performance averaged across all 7 background noises



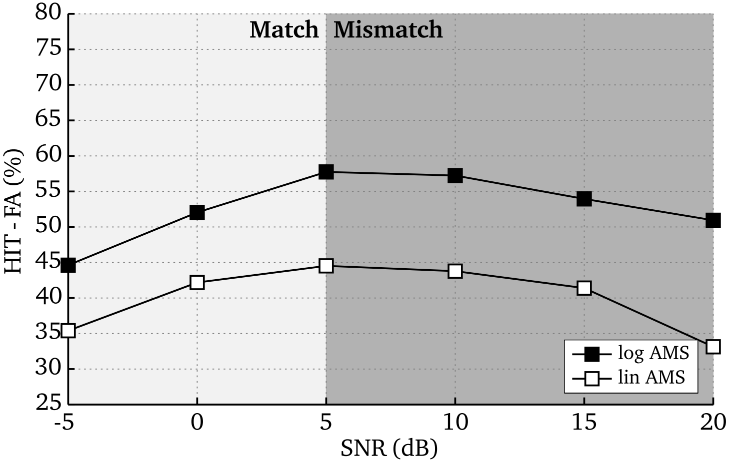
Speech segregation performance

- Performance averaged across all 7 background noises



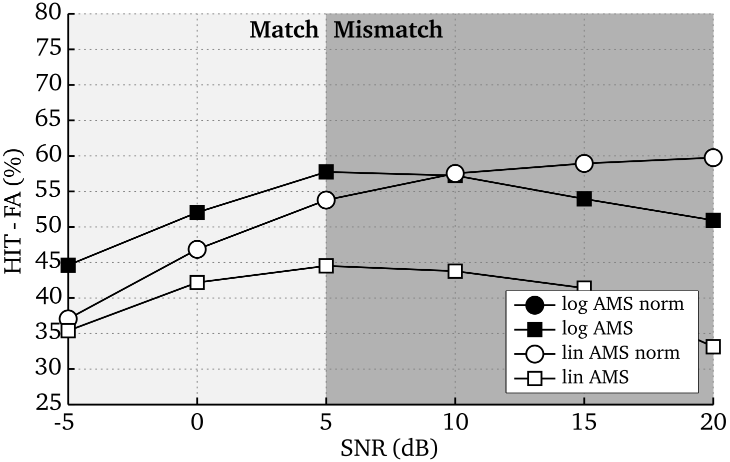
Speech segregation performance

- Performance averaged across all 7 background noises



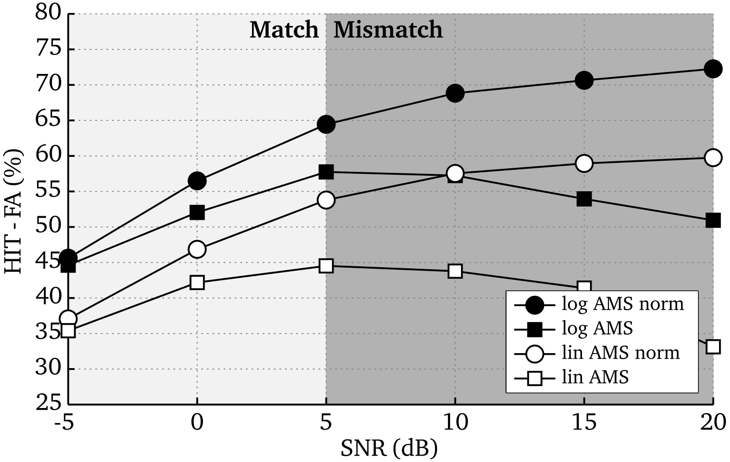
Speech segregation performance

- Performance averaged across all 7 background noises

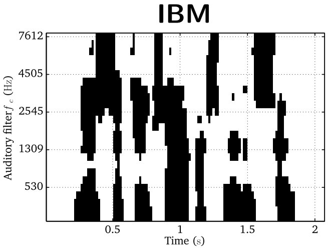


Speech segregation performance

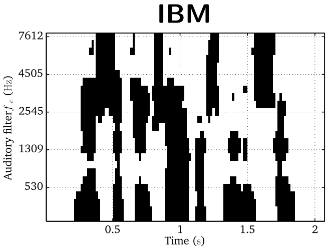
- Performance averaged across all 7 background noises

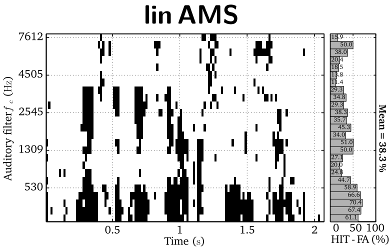


IBM estimation: Noisy speech at 0 dB SNR

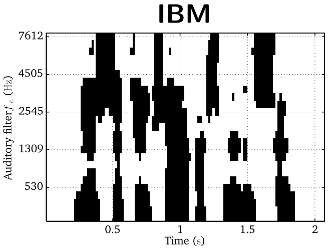


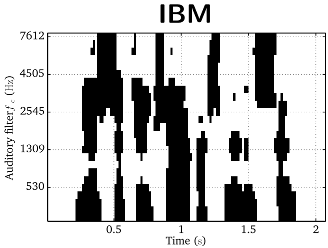
IBM estimation: Noisy speech at 0 dB SNR

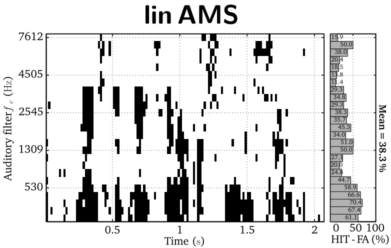




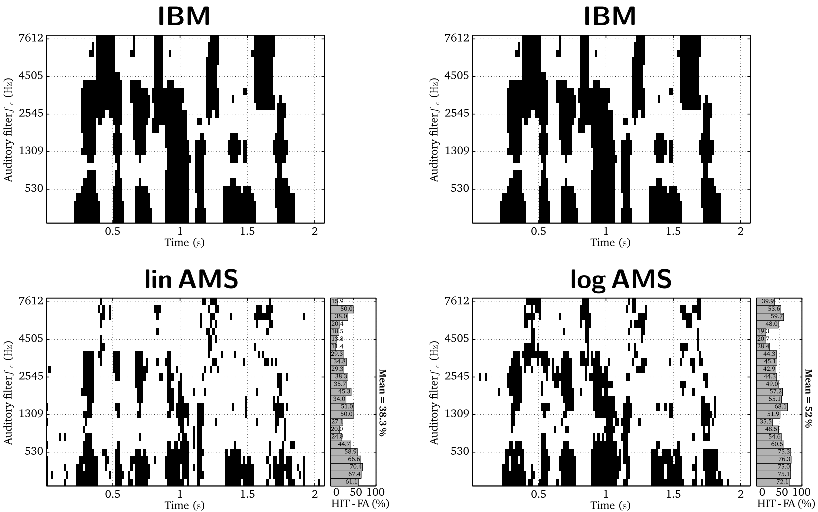
IBM estimation: Noisy speech at 0 dB SNR



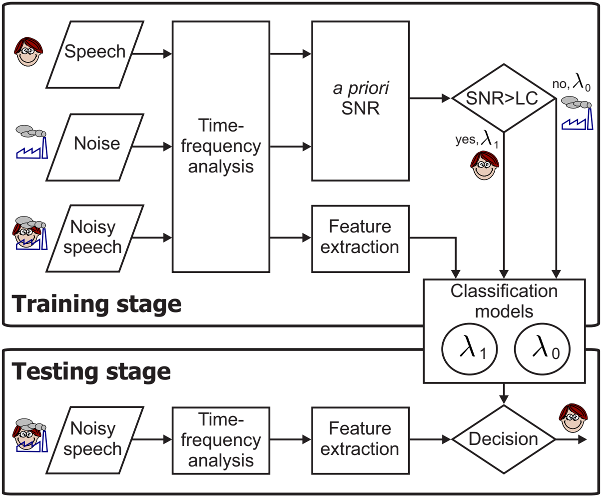




IBM estimation: Noisy speech at 0 dB SNR

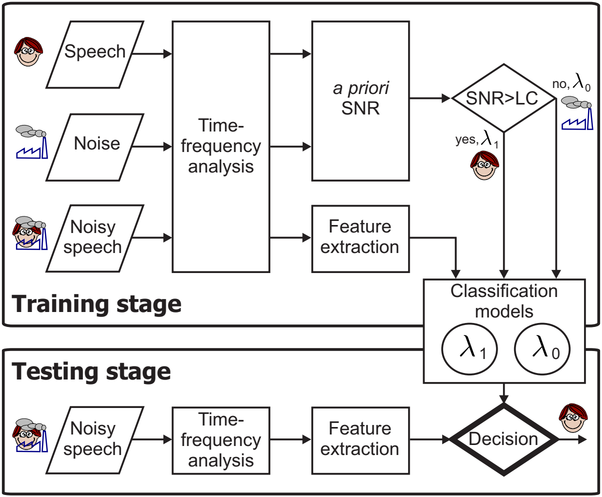


How to estimate the IBM in realistic scenarios?



- ✓ Influence of feature representation
- ✓ Generalization to *unseen* acoustic conditions
- ? Contribution of spectro-temporal context

How to estimate the IBM in realistic scenarios?



- ✓ Influence of feature representation
- ✓ Generalization to *unseen* acoustic conditions
- ? Contribution of spectro-temporal context

Exploiting contextual information

- The IBM is usually estimated in each T-F unit independently
  (Kim *et al.*, 2009; Han and Wang, 2012)

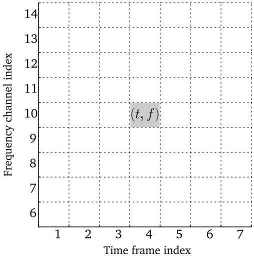
Exploiting contextual information

- The IBM is usually estimated in each T-F unit independently (Kim *et al.*, 2009; Han and Wang, 2012)

  - Speech occupies neighboring T-F units, so-called *glimpses* (Cooke, 2005, 2006)

Exploiting contextual information

- The IBM is usually estimated in each T-F unit independently (Kim *et al.*, 2009; Han and Wang, 2012)
  - Speech occupies neighboring T-F units, so-called *glimpses* (Cooke, 2005, 2006)



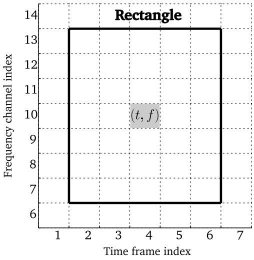
Approach:

- Use *posterior* of GMM-based segregation system as new feature

Exploiting contextual information

- The IBM is usually estimated in each T-F unit independently (Kim *et al.*, 2009; Han and Wang, 2012)

  - Speech occupies neighboring T-F units, so-called *glimpses* (Cooke, 2005, 2006)

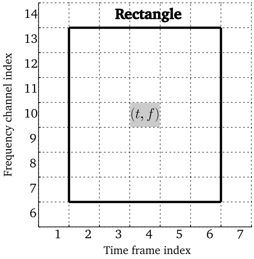


Approach:

- Use *posterior* of GMM-based segregation system as new feature
- Analyze the effect of across-time and frequency integration

Exploiting contextual information

- The IBM is usually estimated in each T-F unit independently (Kim *et al.*, 2009; Han and Wang, 2012)
  - Speech occupies neighboring T-F units, so-called *glimpses* (Cooke, 2005, 2006)

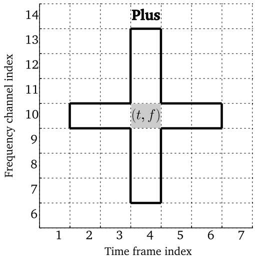


Approach:

- Use *posterior* of GMM-based segregation system as new feature
- Analyze the effect of across-time and frequency integration
- Investigate influence of different window shapes

Exploiting contextual information

- The IBM is usually estimated in each T-F unit independently (Kim *et al.*, 2009; Han and Wang, 2012)
  - Speech occupies neighboring T-F units, so-called *glimpses* (Cooke, 2005, 2006)



Approach:

- Use *posterior* of GMM-based segregation system as new feature
- Analyze the effect of across-time and frequency integration
- Investigate influence of different window shapes

Segregation experiment II

Segregation system:

- GMM classifier trained with linear or logarithmic AMS features
- Investigate role of spectro-temporal integration window

Segregation experiment II

Segregation system:

- GMM classifier trained with linear or logarithmic AMS features
- Investigate role of spectro-temporal integration window

Training:

- 100 HINT sentences
- mixed at $-5, 0, 5\,\mathrm{dB}$ SNR
- ICRA1, ICRA7, PSAM $8\mathrm{Hz}$, traffic, music, destroyer, factory

Segregation experiment II

Segregation system:

- GMM classifier trained with linear or logarithmic AMS features
- Investigate role of spectro-temporal integration window

Training:

- 100 HINT sentences
- mixed at $-5, 0, 5\,\mathrm{dB}$ SNR
- ICRA1, ICRA7, PSAM $8\mathrm{Hz}$, traffic, music, destroyer, factory

Testing:

- 60 HINT sentences
- mixed at $-5 : 5 : 20\,\mathrm{dB}$ SNR
- ICRA1, ICRA7, PSAM $8\mathrm{Hz}$, traffic, music, destroyer, factory (*unknown* realizations)

Segregation experiment II

Segregation system:

- GMM classifier trained with linear or logarithmic AMS features
- Investigate role of spectro-temporal integration window

Training:

- 100 HINT sentences
- mixed at $-5, 0, 5\,\mathrm{dB}$ SNR
- ICRA1, ICRA7, PSAM $8\mathrm{Hz}$, traffic, music, destroyer, factory

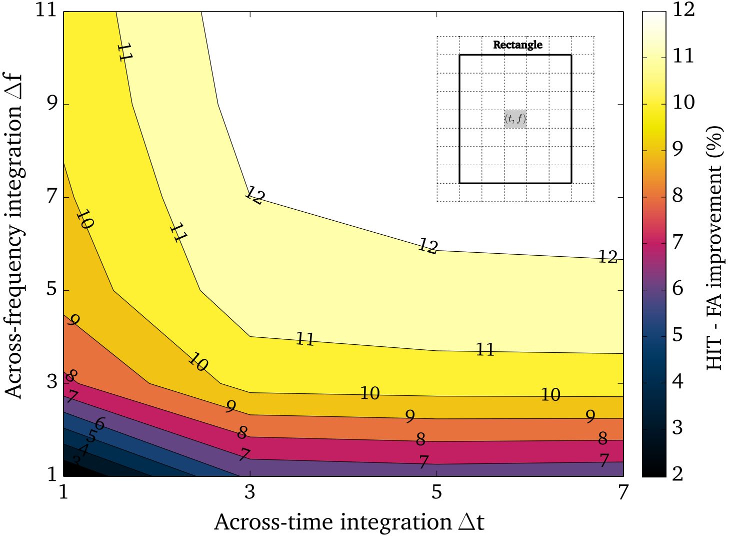
Testing:

- 60 HINT sentences
- mixed at $-5 : 5 : 20\,\mathrm{dB}$ SNR
- ICRA1, ICRA7, PSAM $8\mathrm{Hz}$, traffic, music, destroyer, factory (*unknown* realizations)

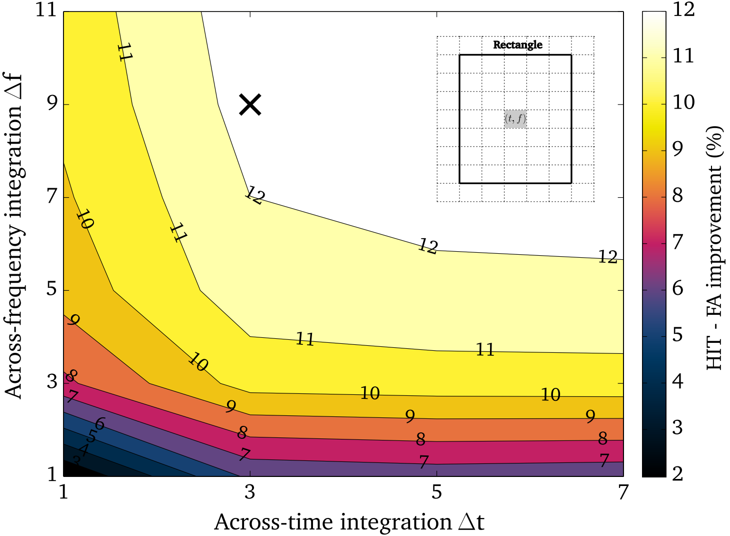
Evaluation:

✓ Measure HIT-FA, which correlates with speech intelligibility

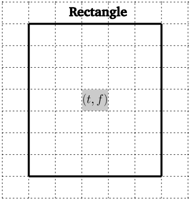
Effect of spectro-temporal window size



Effect of spectro-temporal window size



Effect of spectro-temporal window shape



Effect of spectro-temporal window shape

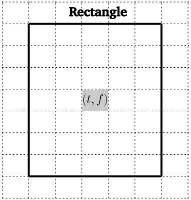
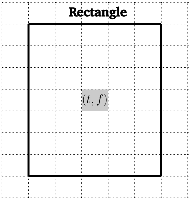


Table: HIT - FA % for different window shapes using $\Delta\mathrm{t} = 3$ and $\Delta\mathrm{f} = 9$.

Window shape	# T-F units	lin AMS	log AMS
Rectangle	24.6	63.0	67.5

Effect of spectro-temporal window shape



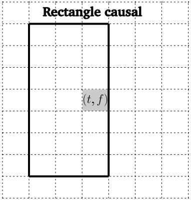


Table: HIT - FA $\%$ for different window shapes using $\Delta \mathrm{t} = 3$ and $\Delta \mathrm{f} = 9$.

Window shape	# T-F units	lin AMS	log AMS
Rectangle	24.6	63.0	67.5
Rectangle causal	16.4	60.0	67.2

Effect of spectro-temporal window shape

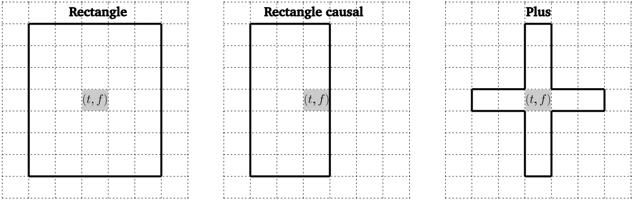


Table: HIT - FA % for different window shapes using $\Delta t = 3$ and $\Delta f = 9$.

Window shape	# T-F units	lin AMS	log AMS
Rectangle	24.6	63.0	67.5
Rectangle causal	16.4	60.0	67.2
Plus	10.2	60.8	66.8

Effect of spectro-temporal window shape

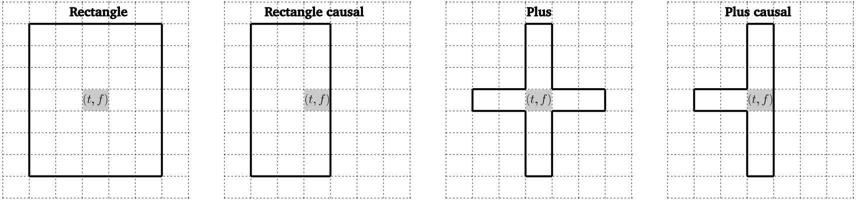
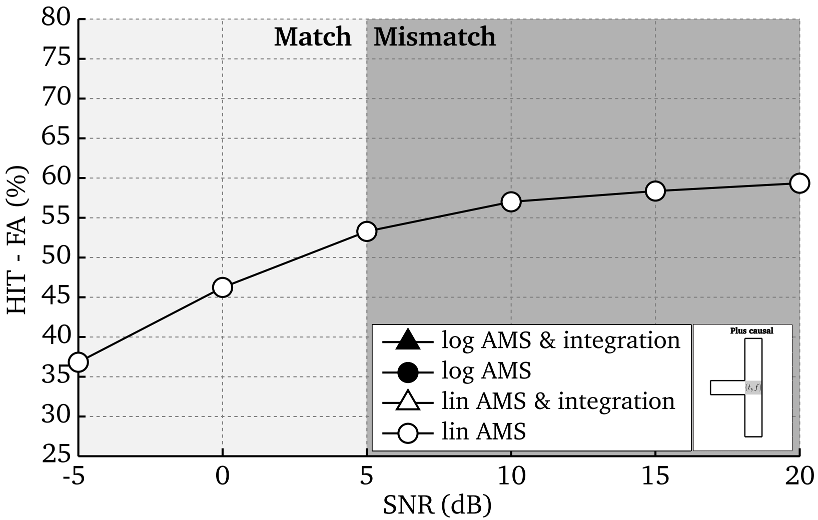


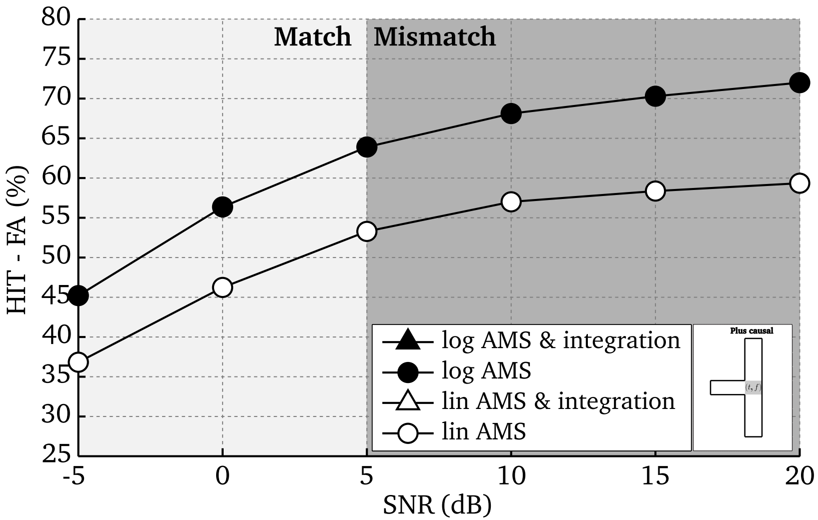
Table: HIT - FA % for different window shapes using $\Delta t = 3$ and $\Delta f = 9$.

Window shape	# T-F units	lin AMS	log AMS
Rectangle	24.6	63.0	67.5
Rectangle causal	16.4	60.0	67.2
Plus	10.2	60.8	66.8
Plus causal	9.2	59.3	66.8

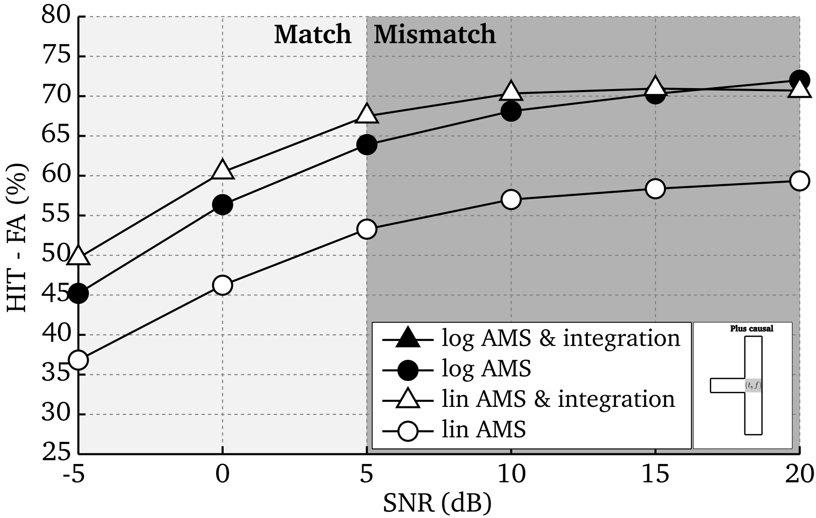
Effect of spectro-temporal integration



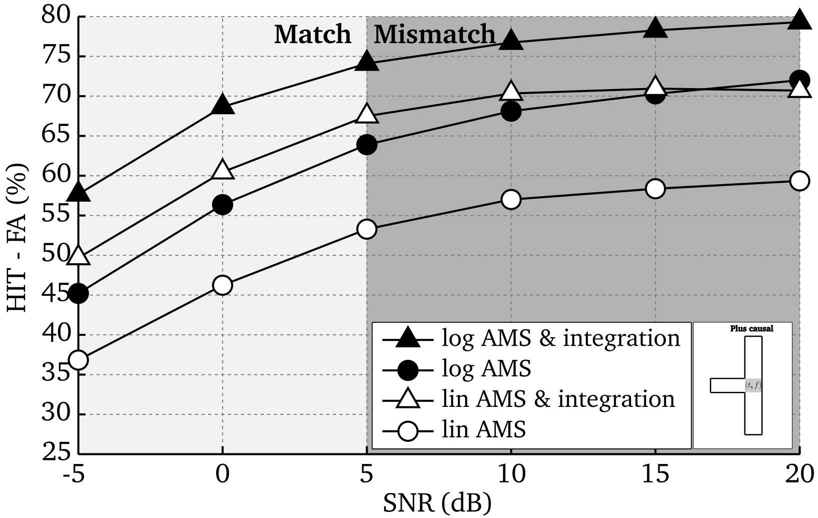
Effect of spectro-temporal integration



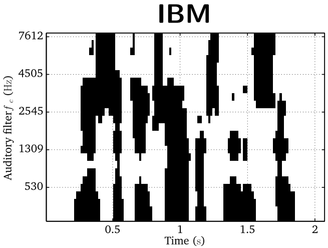
Effect of spectro-temporal integration



Effect of spectro-temporal integration

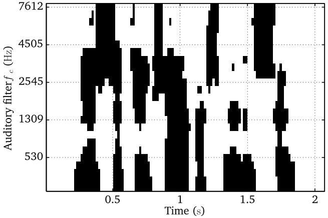


IBM estimation: Noisy speech at 0 dB SNR

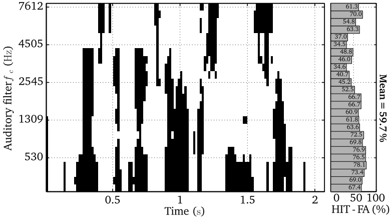


IBM estimation: Noisy speech at 0 dB SNR

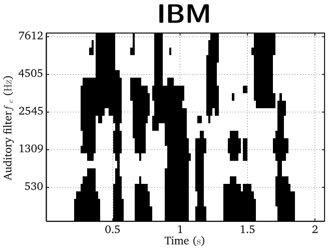
IBM

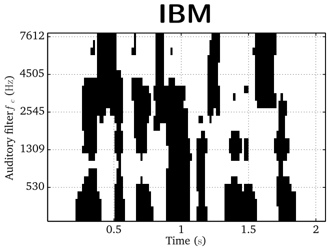


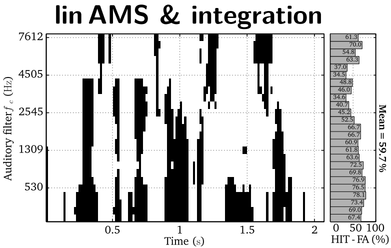
lin AMS & integration



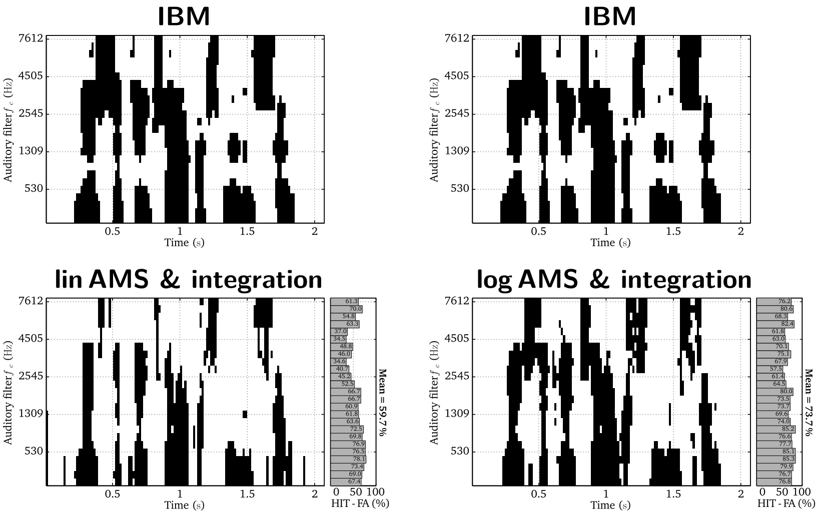
IBM estimation: Noisy speech at 0 dB SNR







IBM estimation: Noisy speech at 0 dB SNR



Conclusions

✓ Approach cocktail-party problem by **combining** knowledge about **auditory processing** with **supervised learning strategies**

Conclusions

- ✓ Approach cocktail-party problem by **combining** knowledge about **auditory processing** with **supervised learning strategies**
- ✓ **Auditory-inspired modulation features** provide higher segregation performance than higher-dimensional variants

Conclusions

- ✓ Approach cocktail-party problem by **combining** knowledge about **auditory processing** with **supervised learning strategies**
- ✓ **Auditory-inspired modulation features** provide higher segregation performance than higher-dimensional variants
- ✓ **Feature normalization** allows **generalization** to unseen SNRs

Conclusions

- ✓ Approach cocktail-party problem by **combining** knowledge about **auditory processing** with **supervised learning strategies**
- ✓ **Auditory-inspired modulation features** provide higher segregation performance than higher-dimensional variants
- ✓ **Feature normalization** allows **generalization** to unseen SNRs
- ✓ **Spectro-temporal integration** substantially improves segregation performance