Classifier-based Mask Estimation for Missing Feature Methods of Robust Speech Recognition

Michael L. Seltzer, Bhiksha Raj & Richard M. Stern

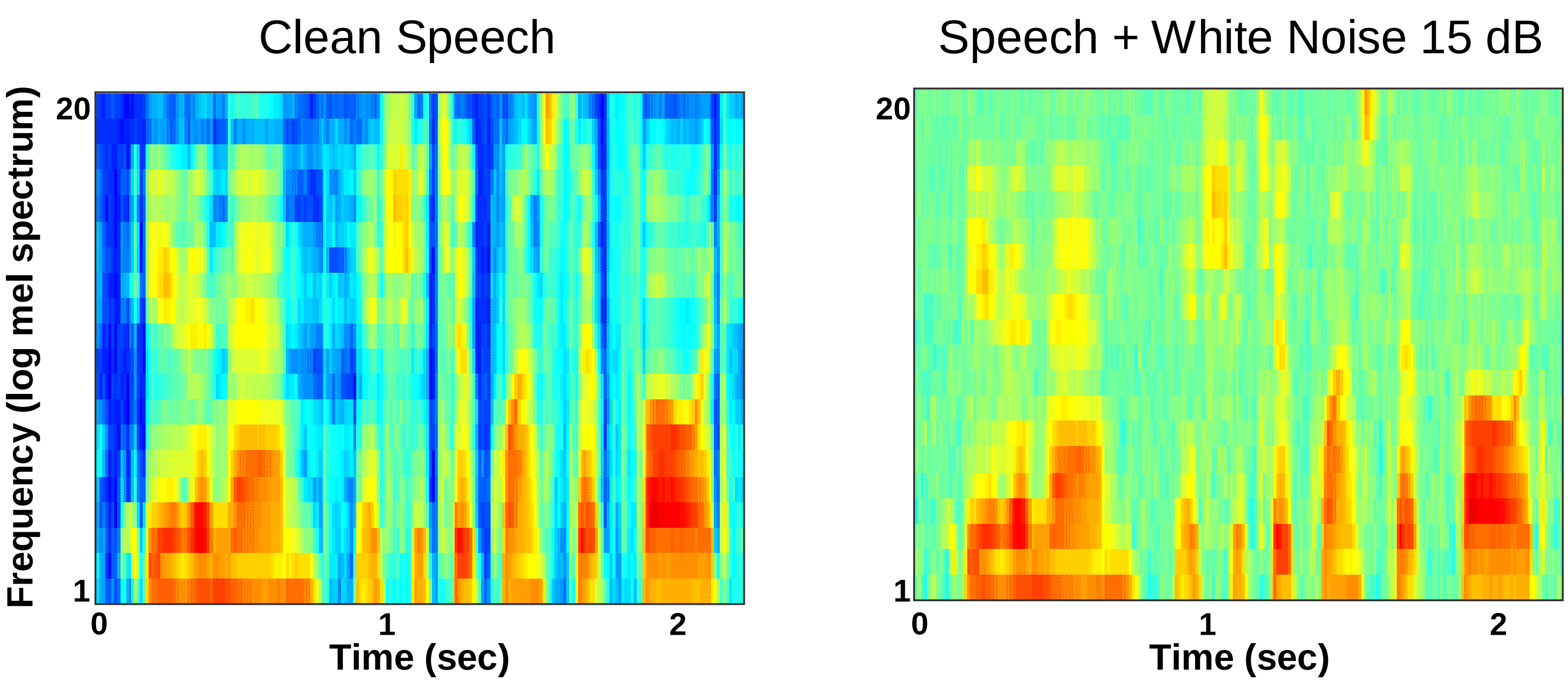
Department of Electrical and Computer Engineering and

School of Computer Science

Carnegie Mellon University

Pittsburgh, PA 15213 USA

Missing Feature Compensation

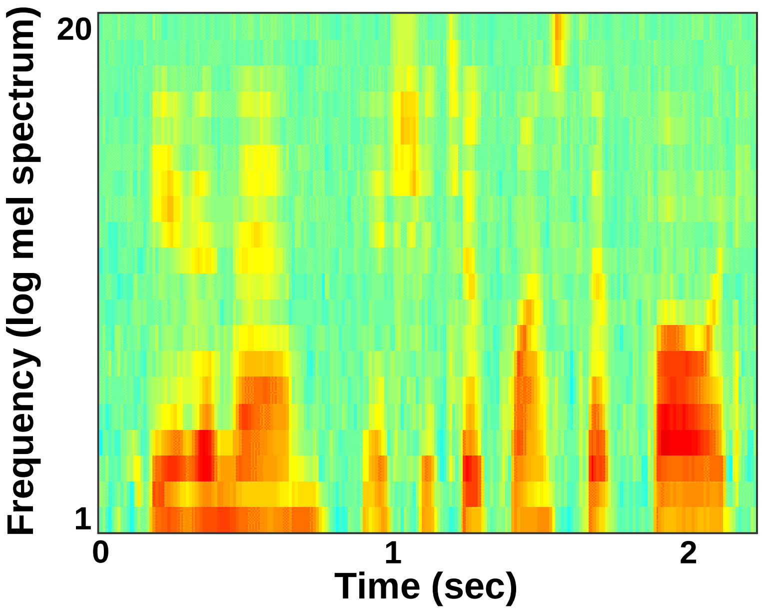


"Even then, if she took one step forward"

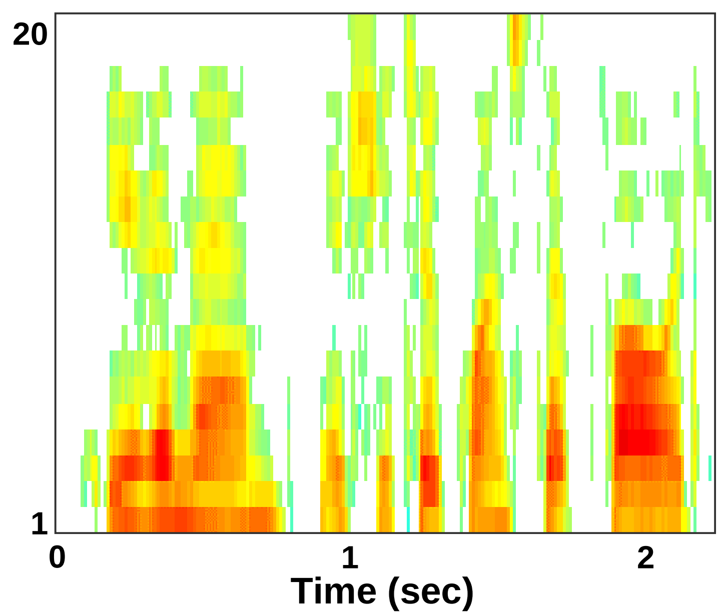
- Noise corrupts some time-frequency locations more than others

Consider noisy regions “missing”

Speech + White Noise 15 dB



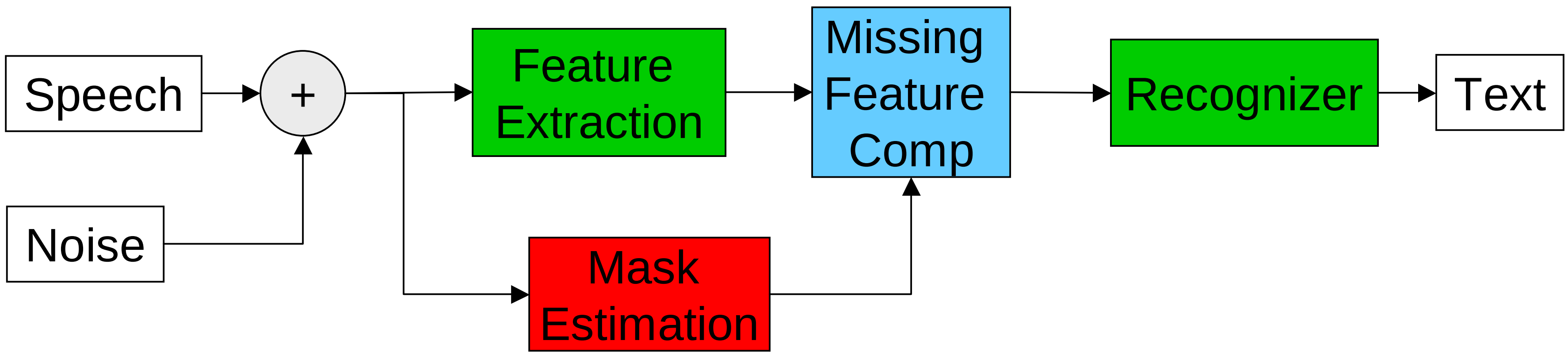
“Corrupt” regions removed



- All regions of local SNR less than 0 dB considered missing.
- Missing Feature Methods perform compensation using remaining reliable regions.
- No stationarity assumptions are made.

Missing Feature Compensation

- For missing feature methods to be successful, we need a *spectrographic mask*, a binary mask that accurately labels the reliable and corrupt features.



How do we estimate masks?

- Conventional mask estimation methods estimate local SNR
  - Methods assume noise is pseudo-stationary
- Is this *really* a noise estimation problem?
  - No!
  - Mask estimation is a *binary decision process*
- Solution: Build a 2-class classifier
  - Use all available information to make a decision
  - No stationarity assumptions about noise

Voiced Speech Feature Extraction

- Most of the energy of voiced speech is centered around the harmonics of the fundamental frequency

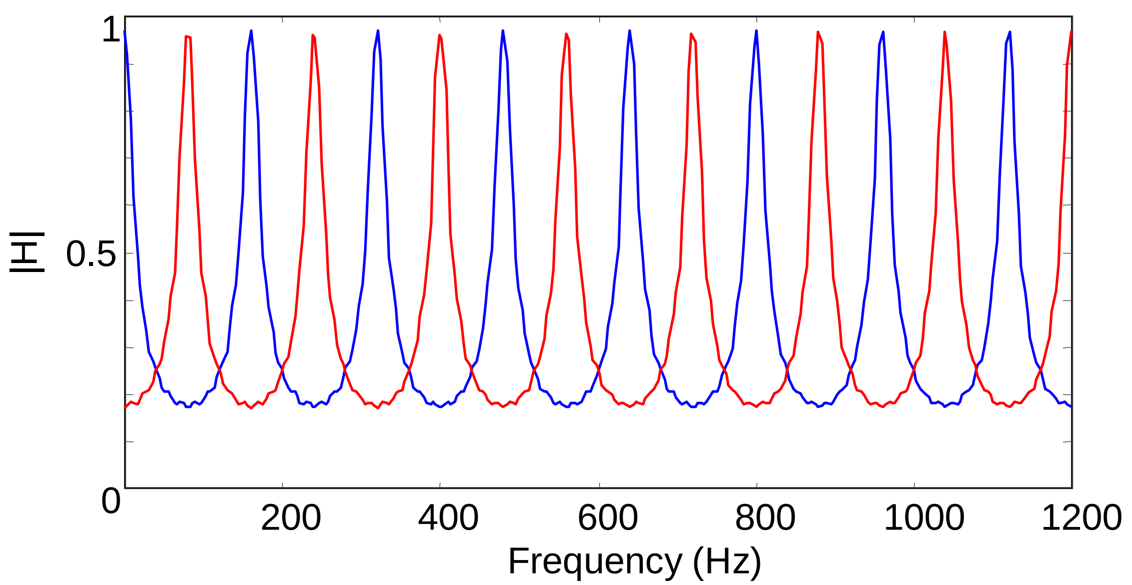
- Noise may or may not contain energy at these frequencies.

- Can we measure how much energy is at the harmonics (speech) and how much is not (noise)?

Yes! Use Comb Filters

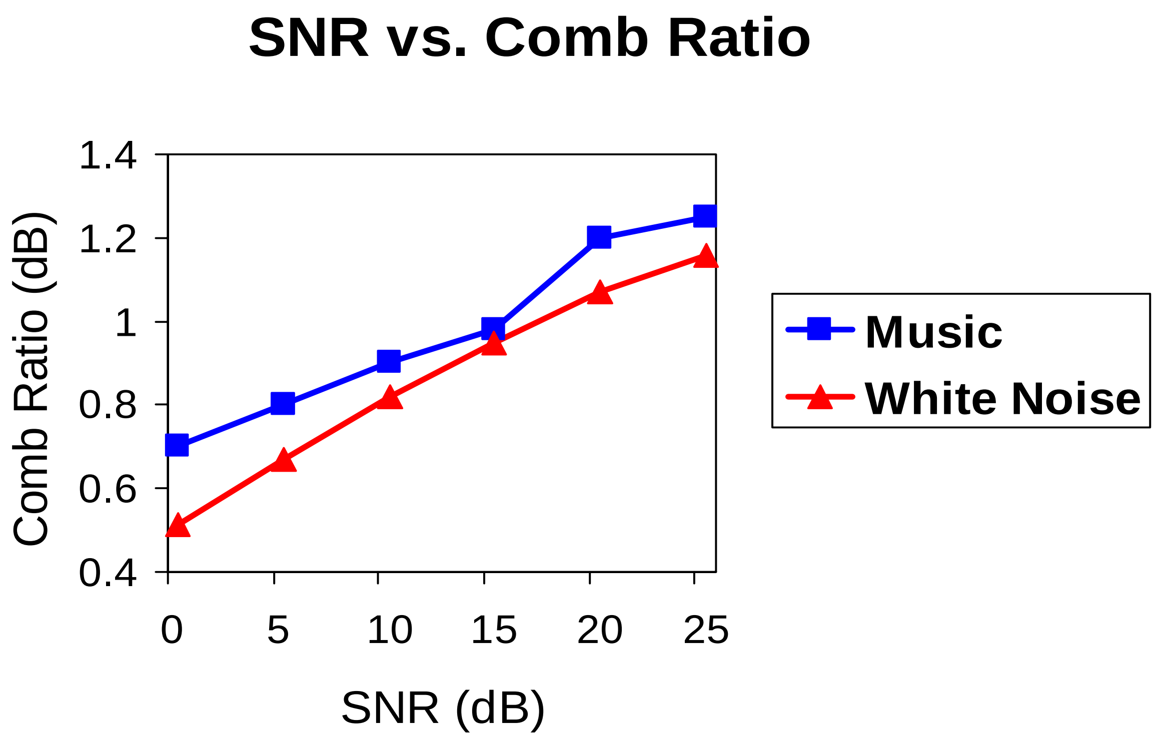
- Capture the energy at and between the harmonics
  - The ratio of the energies of these two filters give us a measure of noise content, the *Comb Ratio*.

$$H_{comb}(z) = \frac{z^{-p}}{1 - gz^{-p}} \qquad H_{combshift}(z) = \frac{-z^{-p}}{1 + gz^{-p}}$$



Comb Ratio as a measure of SNR

- Average Comb Ratio vs. global SNR for the voiced frames of a single utterance
  - Clear relationship between SNR and the Comb Ratio



What about the pitch?

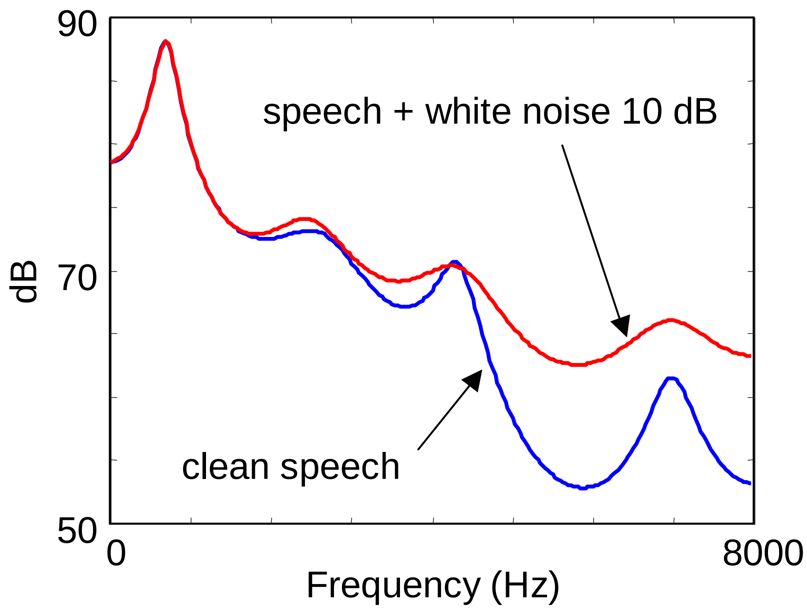
- Comb filtering assumes we know the fundamental frequency of the speech signal. (We don't.)

- There are several pitch tracking algorithms that we can use to estimate the pitch.

More Voiced Speech Features

- Voiced speech has a distinctive spectral contour
  - Noise will change this contour.

/EH/ in “then”



Features to capture spectral contour

- Sub-band Energy to Frame Energy Ratio
- Flatness: variance of the energy in a local spectrographic region

Voiced Speech Feature Summary

- Voiced Feature Set:
  - Comb Ratio
  - Sub-band Energy to Frame Energy Ratio
  - Flatness
  - Ratio of secondary and primary autocorrelation peaks
  - Ratio of sub-band energy to estimate of noise floor energy

- Using *ratios* rather than absolute values for features enables the classifier to be *invariant to overall signal level*

What about the unvoiced speech?

- For unvoiced speech we only use the features that characterize spectral shape:
  - Sub-band Energy to Frame Energy Ratio
  - Flatness
  - Sub-band Energy to Sub-band Noise Floor Ratio

Classification Strategy

- Multivariate Gaussian classifier
- Separate classifier for voiced and unvoiced regions
- Separate classifier per sub-band
- Trained with oracle masks that label training data as reliable or unreliable

How well do we do?

- Speech corrupted by noise
  - 3 noise environments: white noise, factory noise, music
    - Assumption: Known operating environment
  - Training Set:
    - 2880 utterances from Resource Management corrupted with noise at various SNRs.
  - Test Set:
    - 1600 utterances from Resource Management corrupted with noise at a single SNR
  - Oracle masks for Evaluation:
    - If local SNR is < -5dB, consider mask location to be corrupt

Mask Estimation Performance

- Performance compared to "oracle masks" via confusion matrix.

Voiced

AWGN

	"1"	*"0"*
1	*87%*	*13%*
0	*16%*	*84%*

Factory

	"1"	*"0"*
1	*79%*	*21%*
0	*21%*	*79%*

Music

	"1"	*"0"*
1	*72%*	*28%*
0	*33%*	*67%*

Unvoiced

AWGN

	"1"	*"0"*
1	*76%*	*24%*
0	*13%*	*87%*

Factory

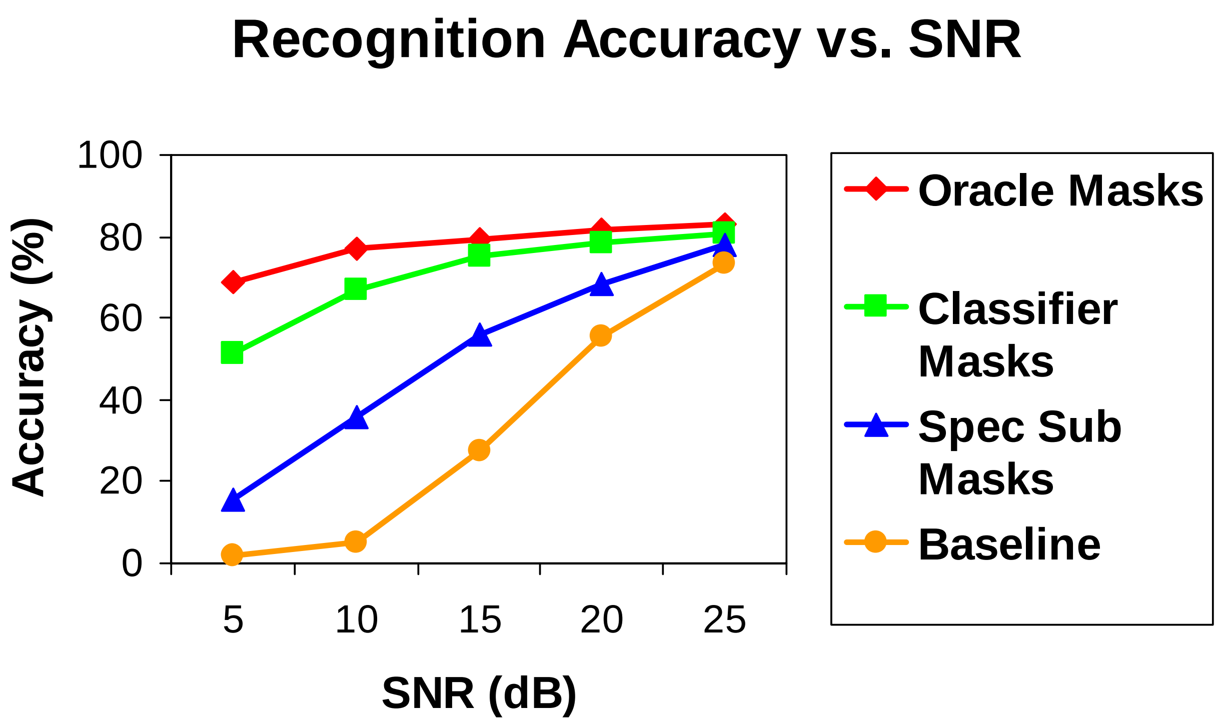
	"1"	*"0"*
1	*71%*	*29%*
0	*22%*	*78%*

Music

	"1"	*"0"*
1	*64%*	*36%*
0	*28%*	*72%*

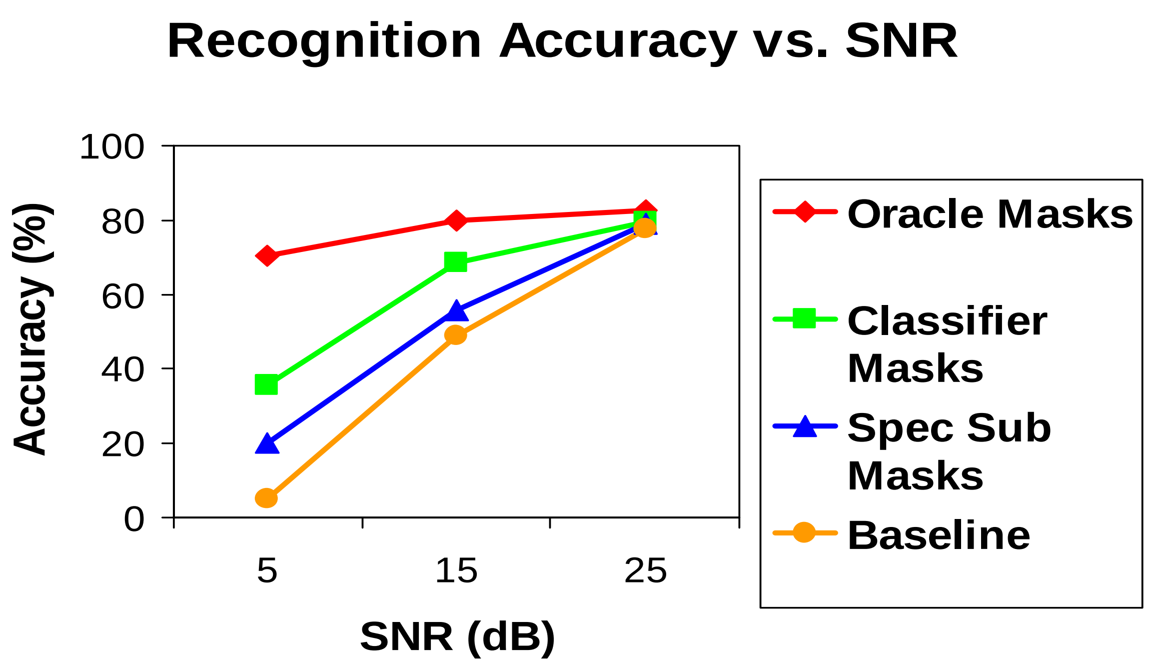
Speech Recognition with Estimated Masks

- Speech + White Noise



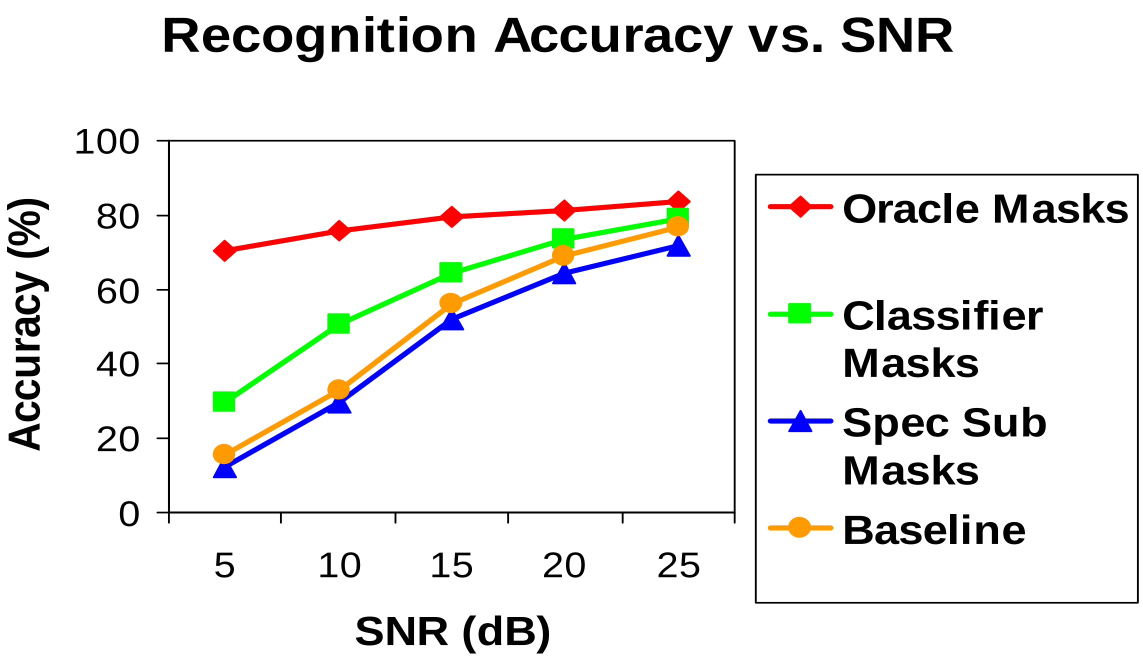
Speech Recognition with Estimated Masks

- Speech + Factory Noise



Speech Recognition with Estimated Masks

- Speech + Music



Conclusions

- Missing Feature Methods have great potential for compensation for *stationary and non-stationary noises*, if the spectrographic masks are known.
- We have developed a classification scheme for mask estimation that is free of the stationarity assumptions made by previous methods.
- We obtained substantial improvements in recognition accuracy with classifier-based masks over conventional mask estimation methods in all three noise conditions.