COMPENSATION FOR NONLINEAR DISTORTION IN NOISE FOR ROBUST SPEECH RECOGNITION

Mark J. Harvilla

Ph.D. Thesis Defense

October 27, 2014

Electrical & Computer ENGINEERING

Carnegie Mellon

Introduction

Topic	Symbol	Fraction of thesis work
Dynamic range compression (DRC) and automatic speech recognition (ASR)	DRC & ASR	11%
Blind amplitude normalization (BAN)	BAN	14%
Blind amplitude reconstruction (BAR)	BAR	28%
Robust estimation of distortion (RED)	RED	28%
Artificially-matched training (AMT)	AMT	9%
The Big Picture	Big Picture	10%

Dynamic Range Compression (DRC)

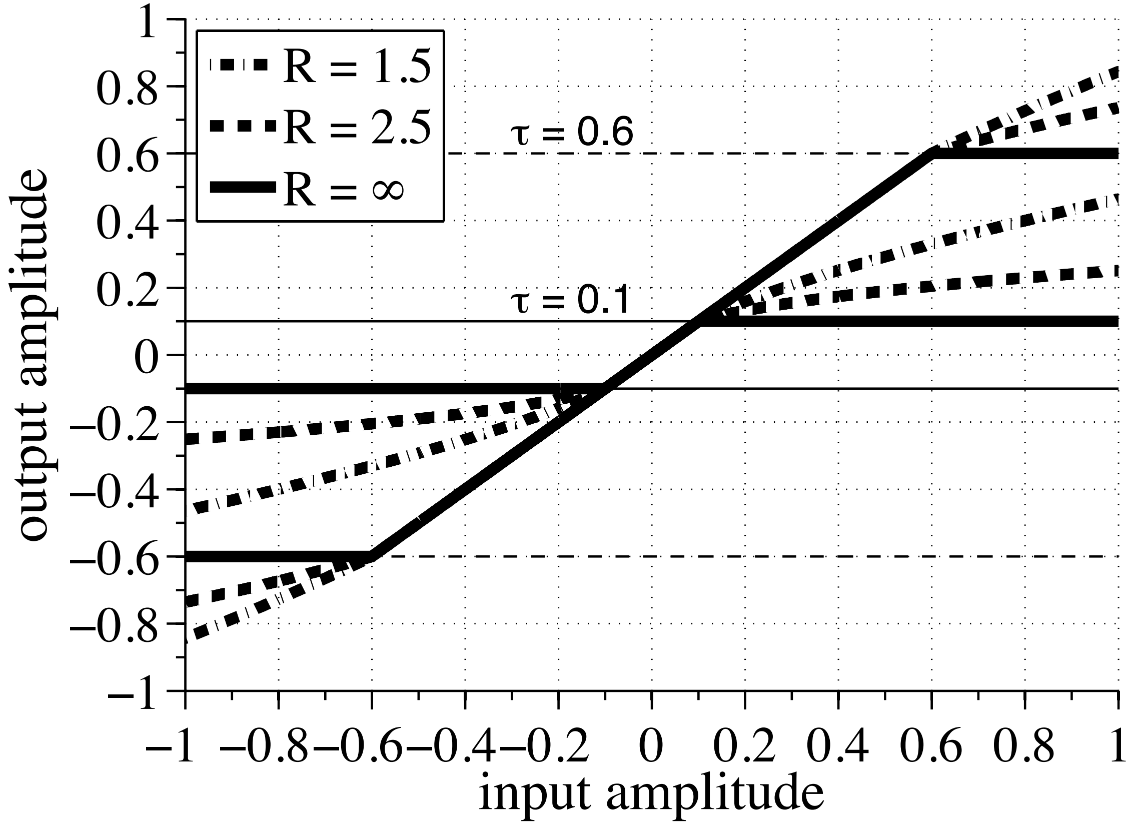
- A form of **nonlinear distortion**
  - Nonlinear systems are common (e.g., AM/FM radio, rectifiers)

- **DRC** is used extensively in audio engineering typically for one of three reasons:
  1. Adhere to **dynamic range limitations** of a signal transmission system, while increasing average signal power
  2. Increase **perceived signal loudness**
  3. Eliminate drastic changes in volume (e.g., **automatic gain control**)

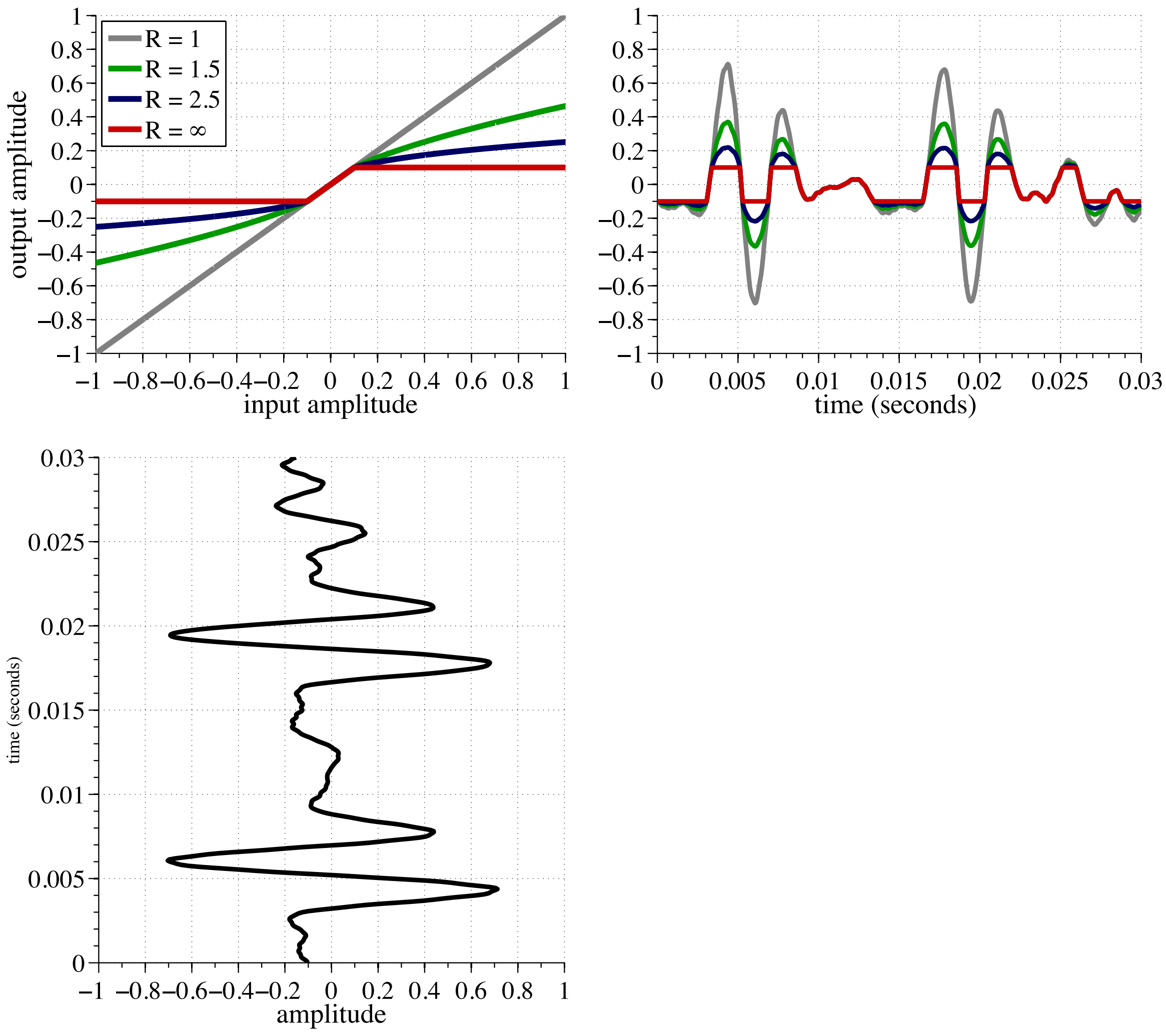
- Because of the ubiquity of DRC, speech systems—like ASR—are likely to encounter compressed speech

Dynamic Range Compression (DRC)

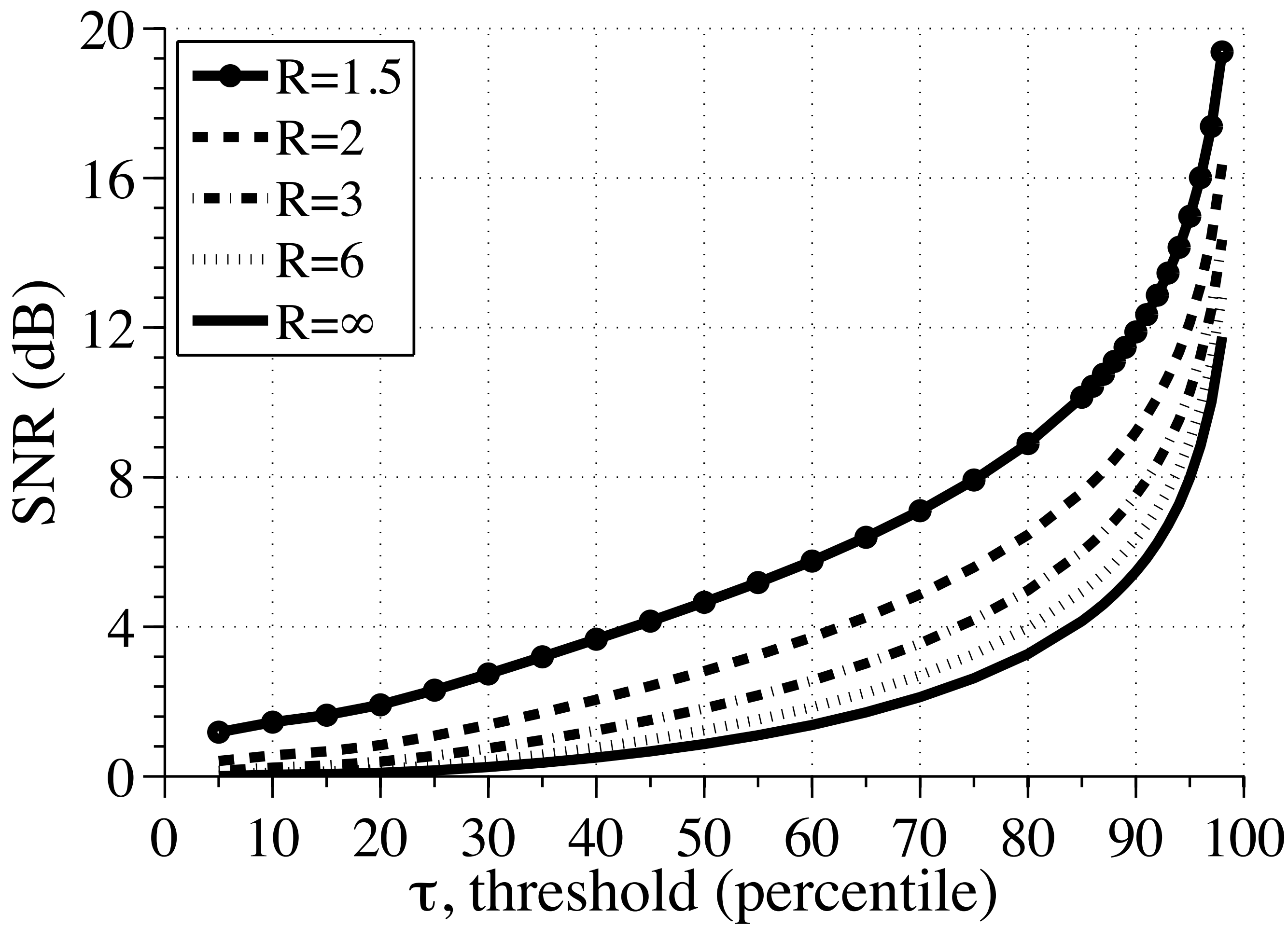
- DRC is characterized by two parameters, **ratio** (*R*) and **threshold** (τ).



Dynamic Range Compression (DRC)



Dynamic Range Compression (DRC)

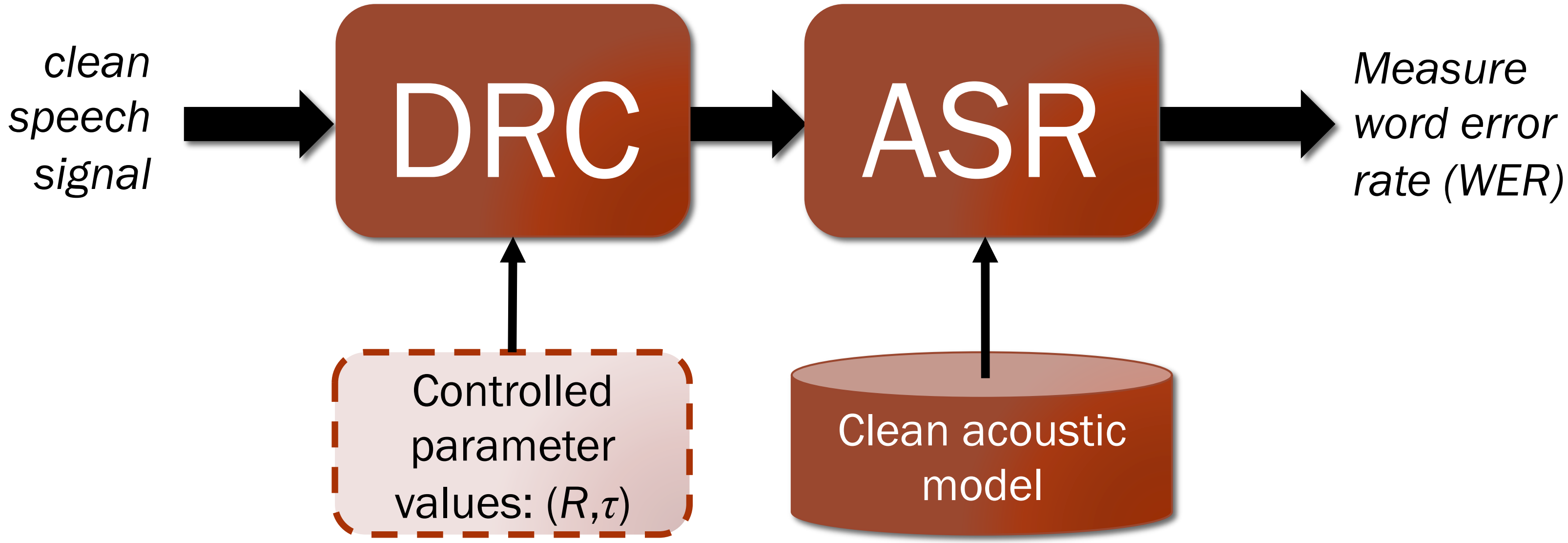


Some examples

Threshold (τ)	Ratio (R)	Audio	Crest Factor	Word Error Rate (WER)	WER after processing
P_{100}	1		17.1 dB	6.4%	6.4%
P_{75}	4		7.7 dB	20.3%	6.4%
P_{75}	∞		4.1 dB	30.8%	13.5%
P_{50}	4		6.7 dB	30.2%	6.4%
P_{50}	∞		2.2 dB	49.5%	23.0%

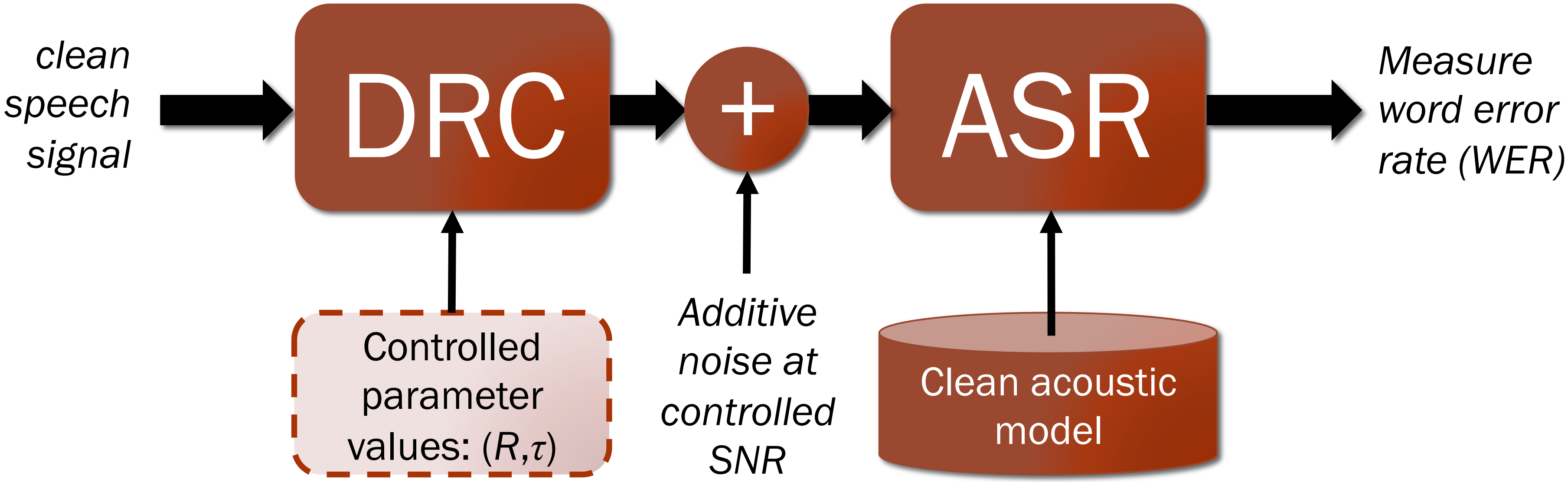
Measuring the effect of DRC on ASR

Experiment 1 (no additive noise):



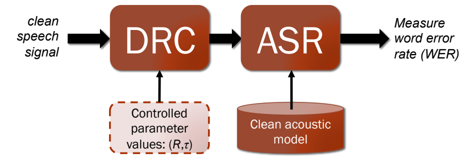
Measuring the effect of DRC on ASR

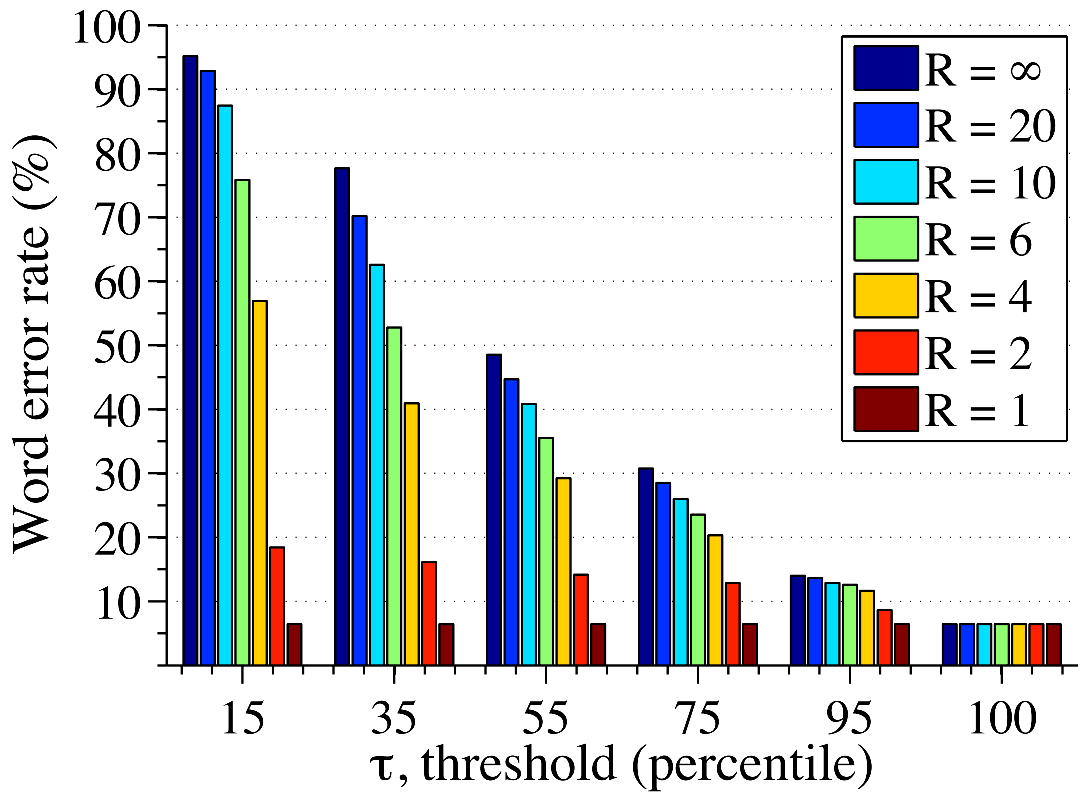
Experiment 2 (additive, channel noise):



Measuring the effect of DRC on ASR

Experiment 1 (no additive noise):

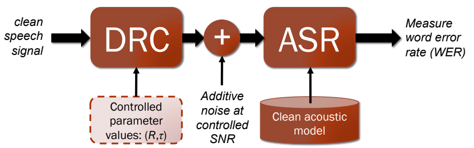


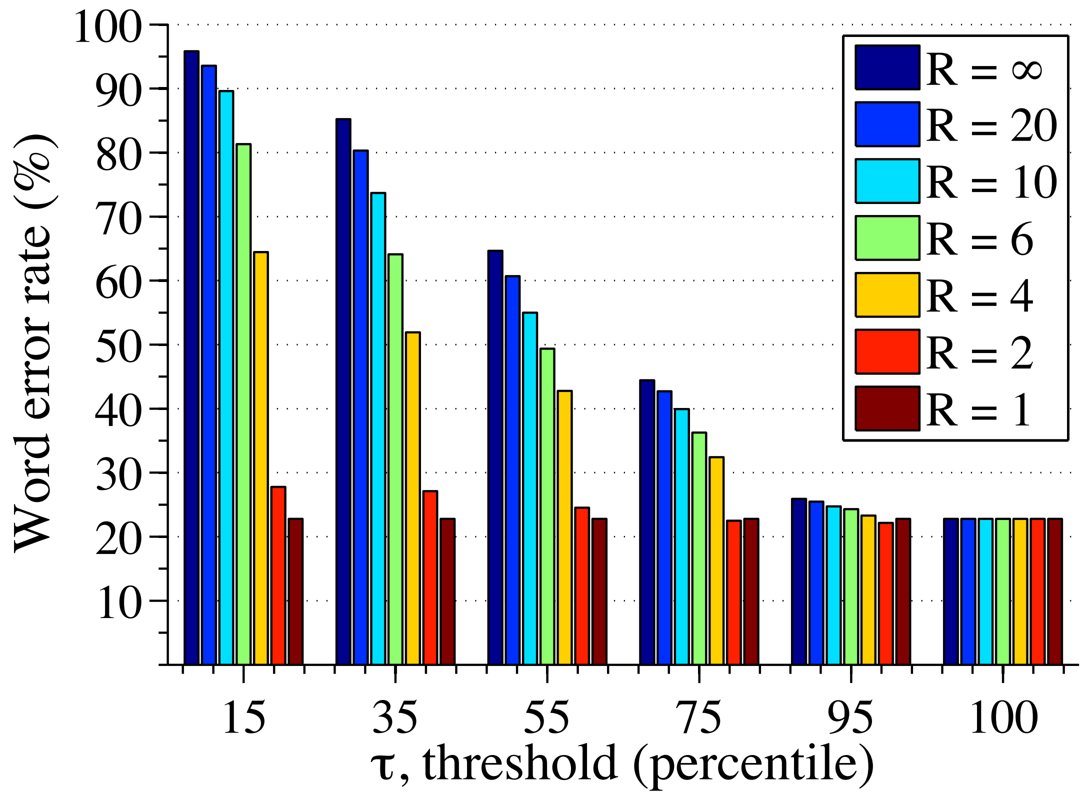


No additive noise

Measuring the effect of DRC on ASR

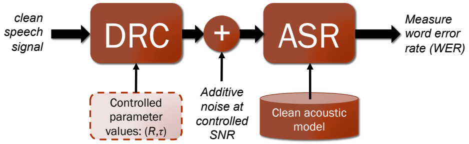
Experiment 2 (additive, channel noise):



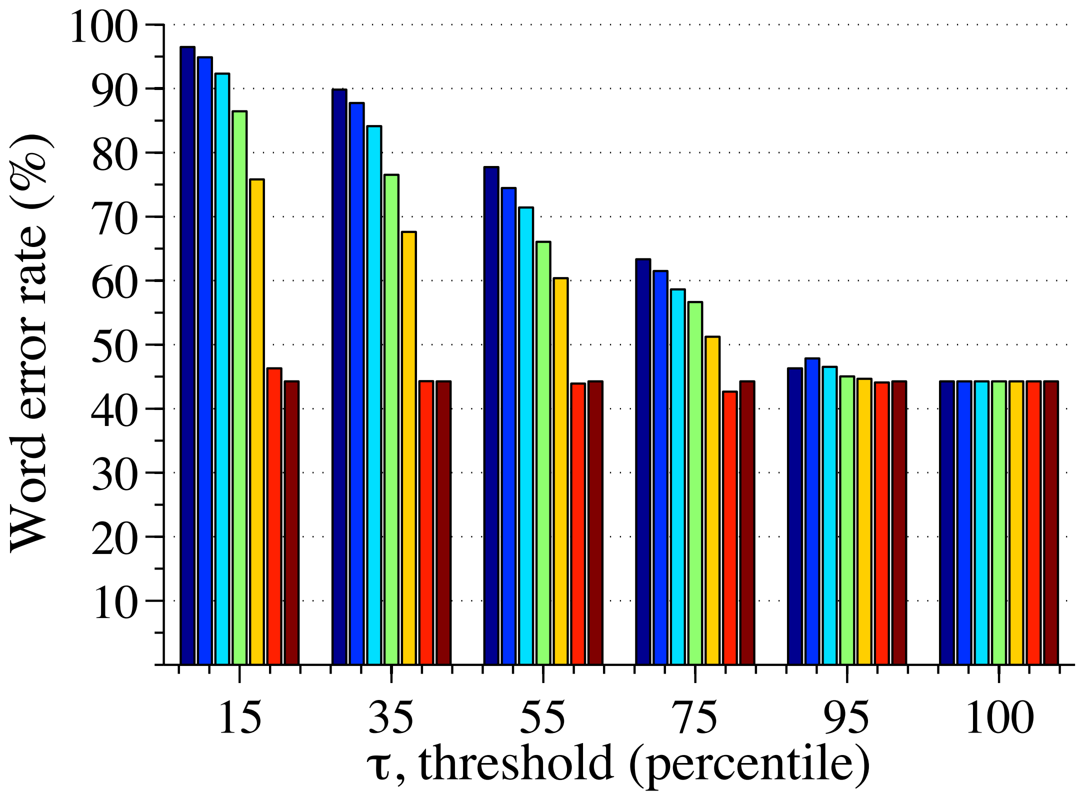


Additive noise at 20-dB SNR w.r.t. compressed signal

Measuring the effect of DRC on ASR

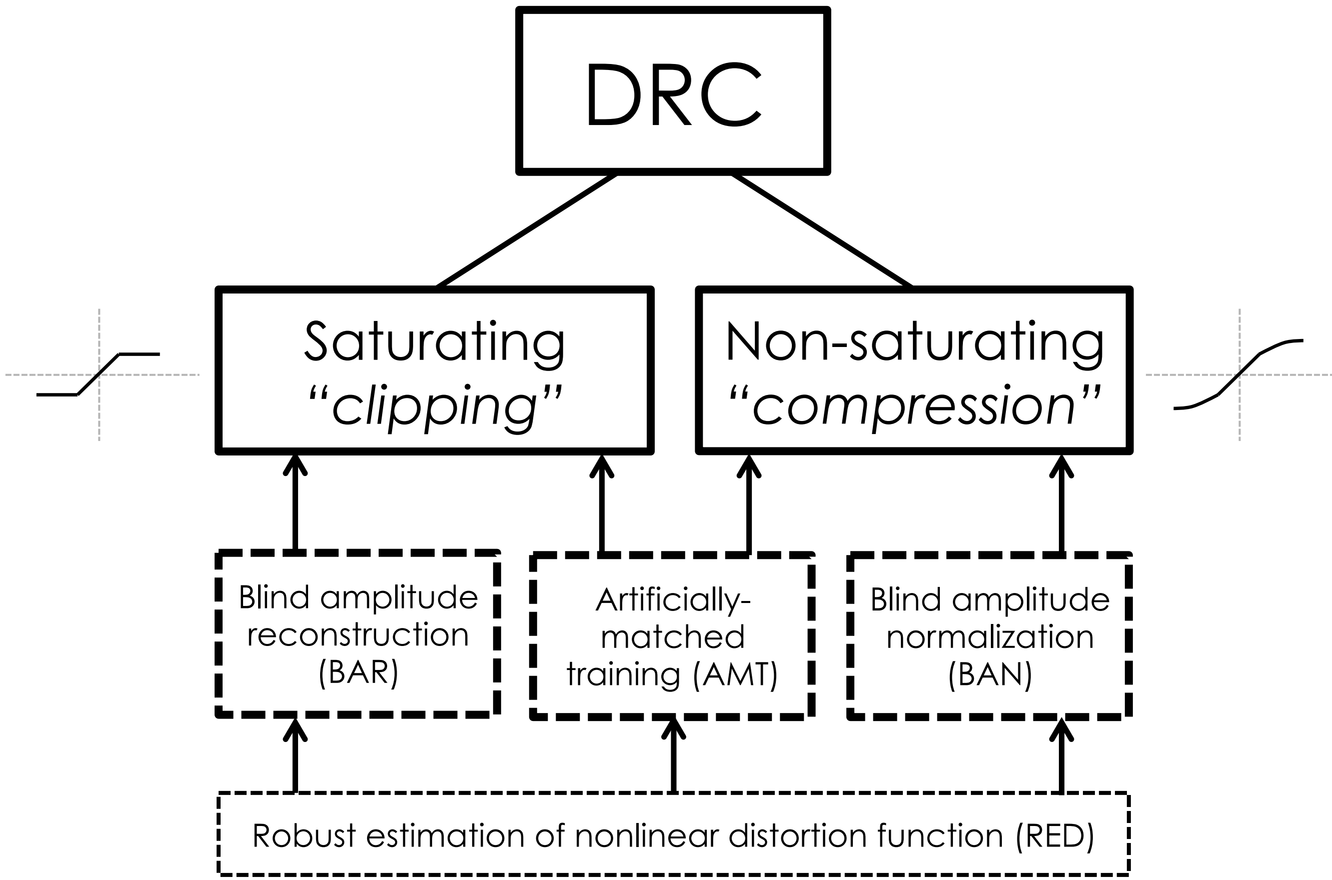


Experiment 2 (additive, channel noise):



Additive noise at 15-dB SNR w.r.t. compressed signal

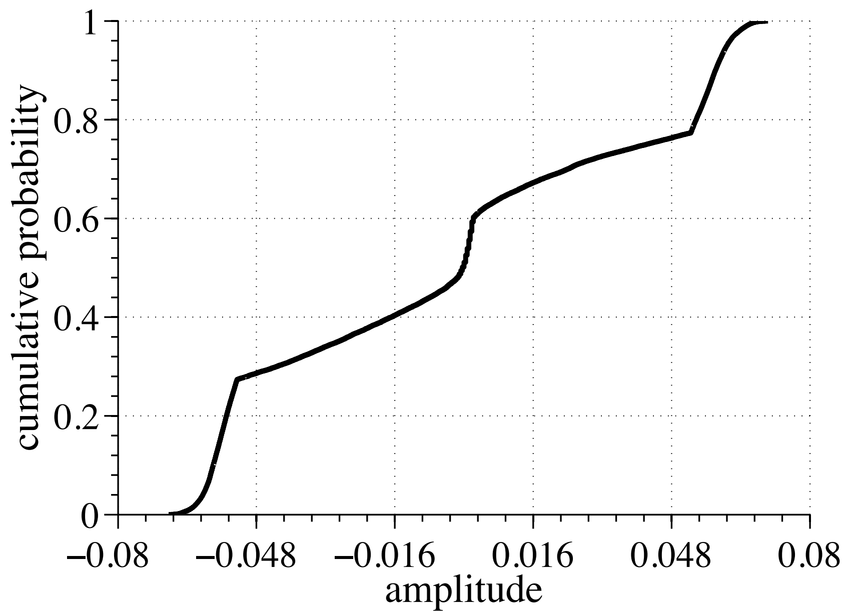
Counteracting the effects of DRC



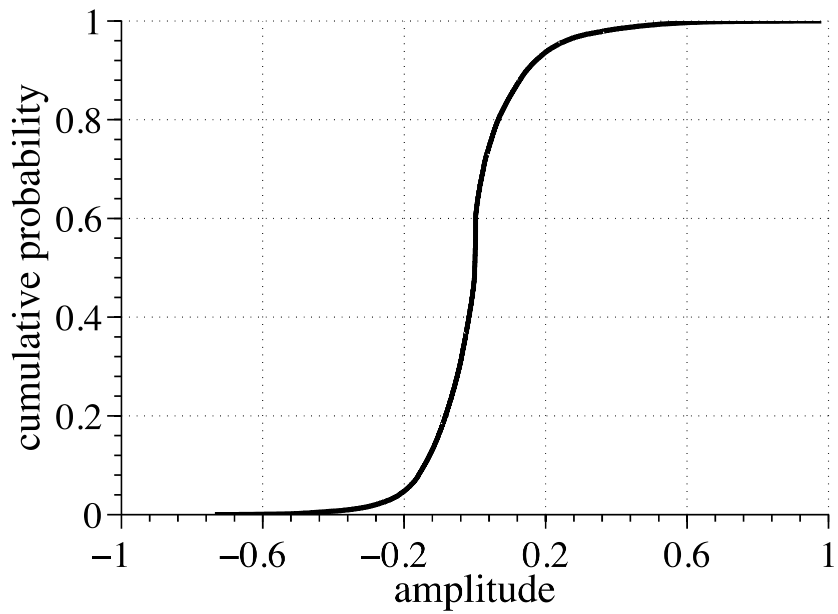
Blind Amplitude Normalization (BAN)

(Balchandran & Mammone; ICASSP 1998)

- **Step 1:** Obtain estimate of the cumulative distribution function (CDF) of the **observed speech**, and of clean, unadulterated **reference speech.**



Observed speech (R = 10, τ = P_{50})

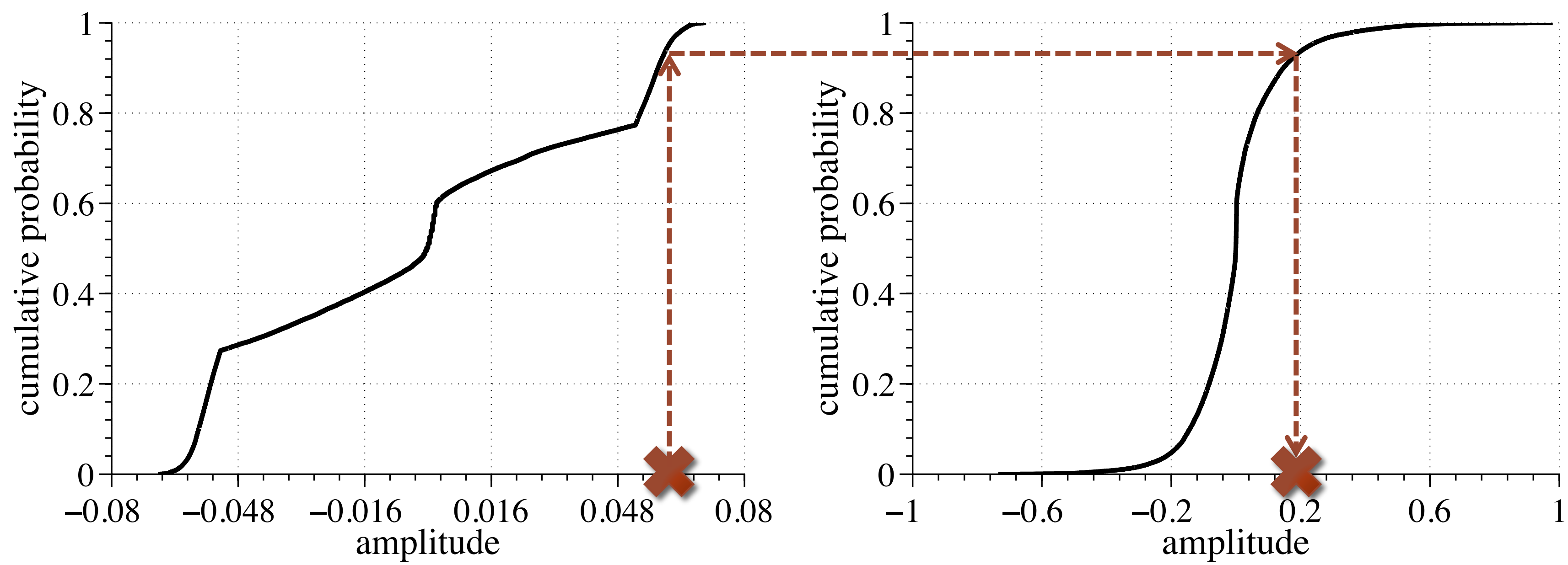


Clean speech

Blind Amplitude Normalization (BAN)

(Balchandran & Mammone; ICASSP 1998)

- **Step 2:** For a given reference signal amplitude, find the amplitude in the observed CDF with the same cumulative probability.

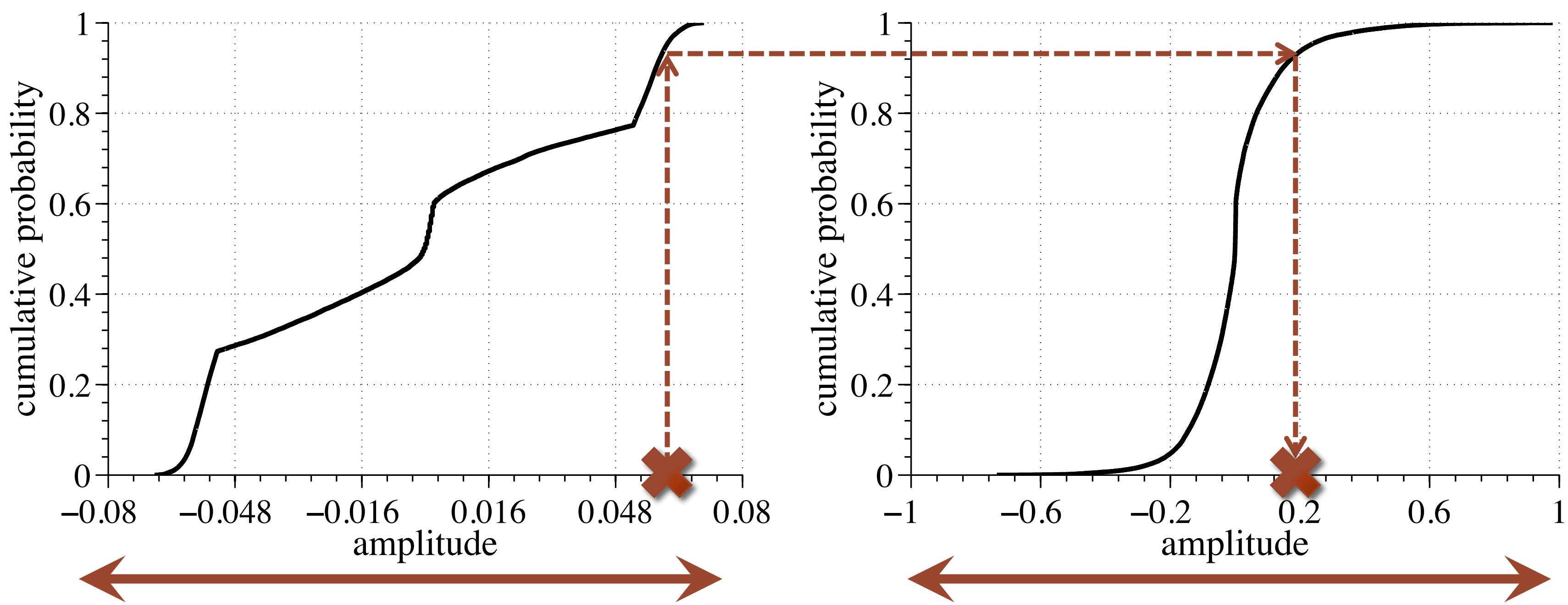


➢ Input amplitude of 0.061 maps to 0.2

Blind Amplitude Normalization (BAN)

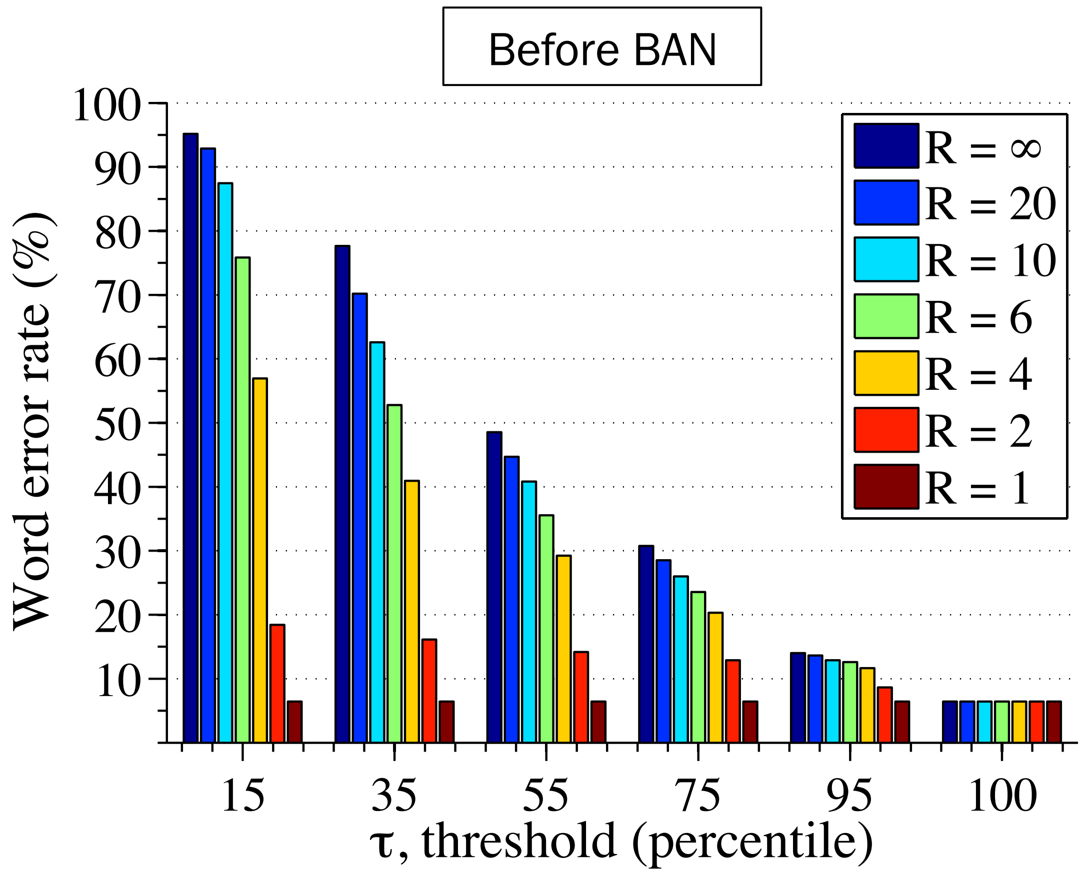
(Balchandran & Mammone; ICASSP 1998)

- **Step 3:** Repeat for each input signal amplitude to obtain a full non-parametric estimate of the nonlinear mapping.



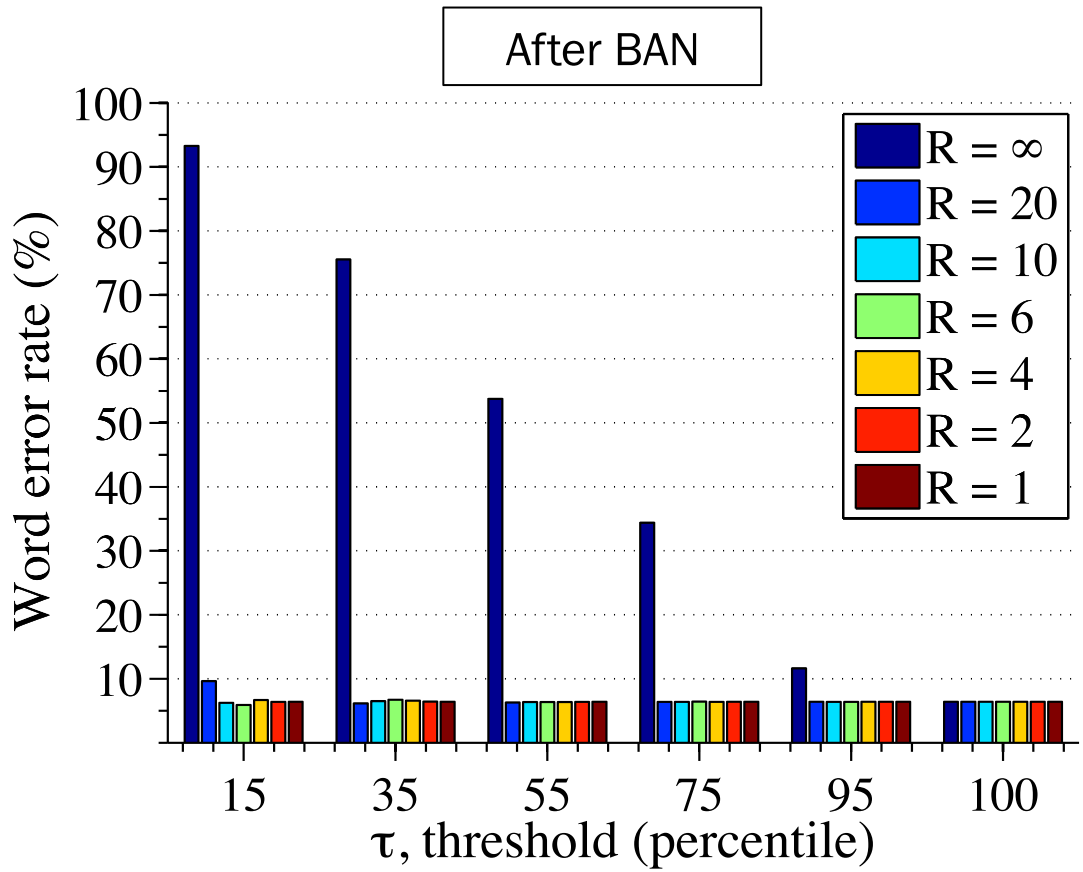
How well does BAN work?

- Experiment 1 (no additive noise):



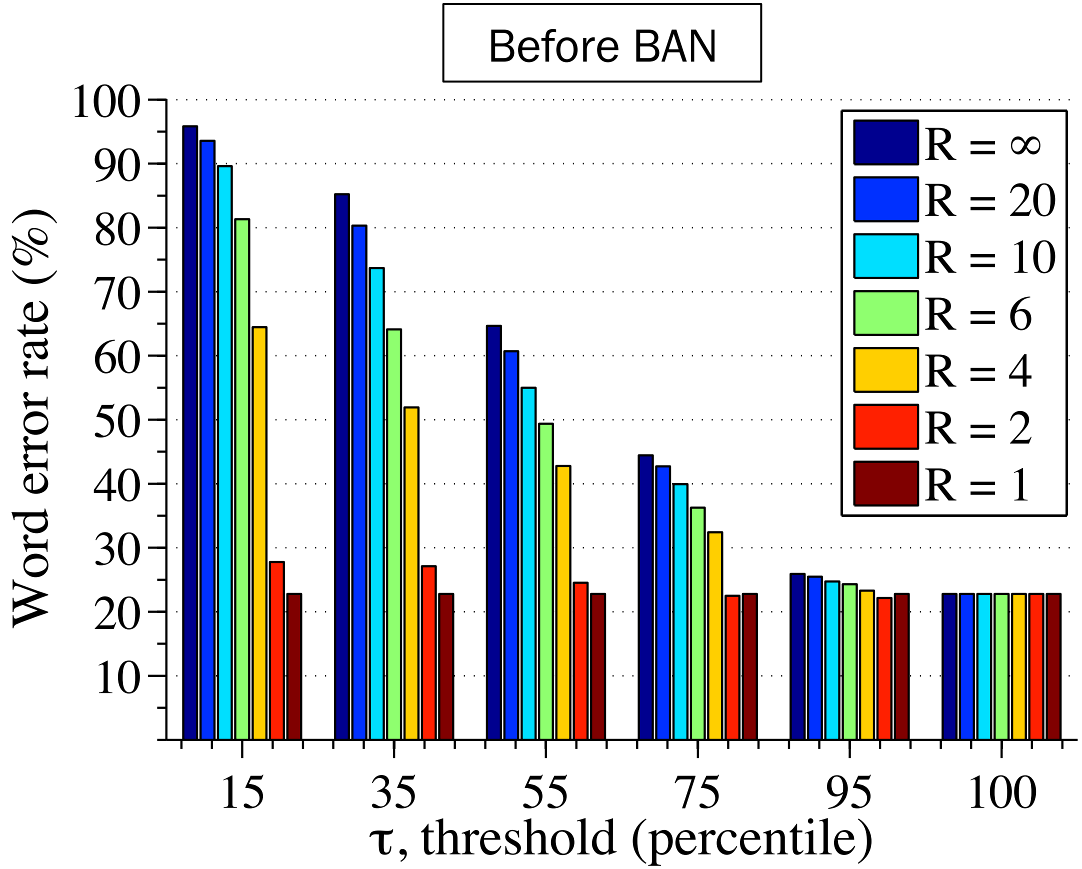
How well does BAN work?

- Experiment 1 (no additive noise):



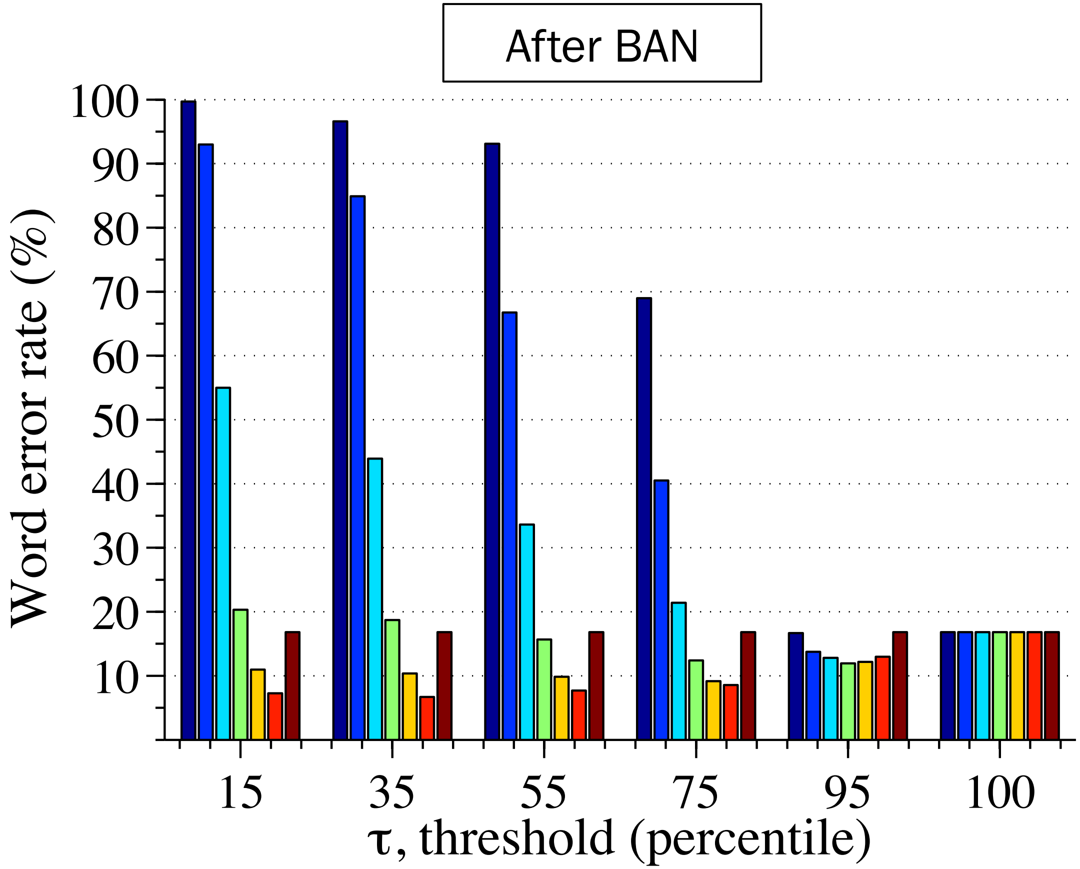
How well does BAN work?

- Experiment 2 (additive, channel noise at 20-dB SNR):



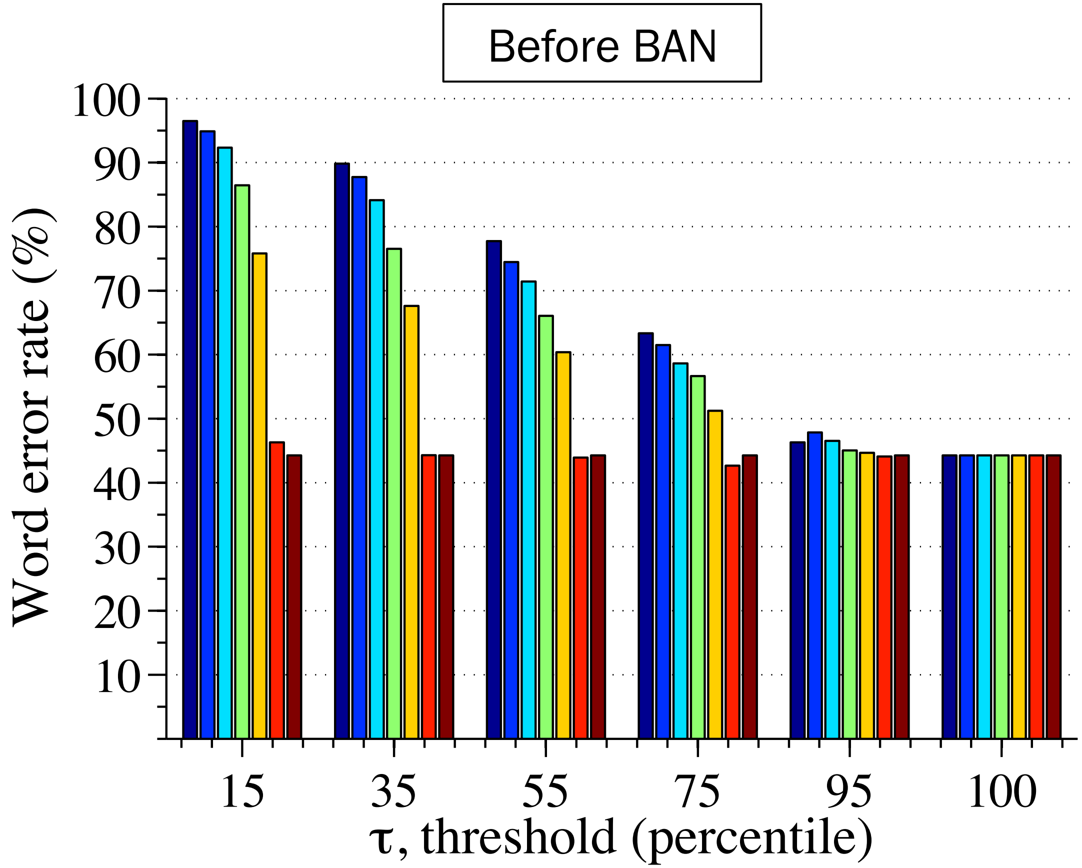
How well does BAN work?

- Experiment 2 (additive, channel noise at 20-dB SNR):



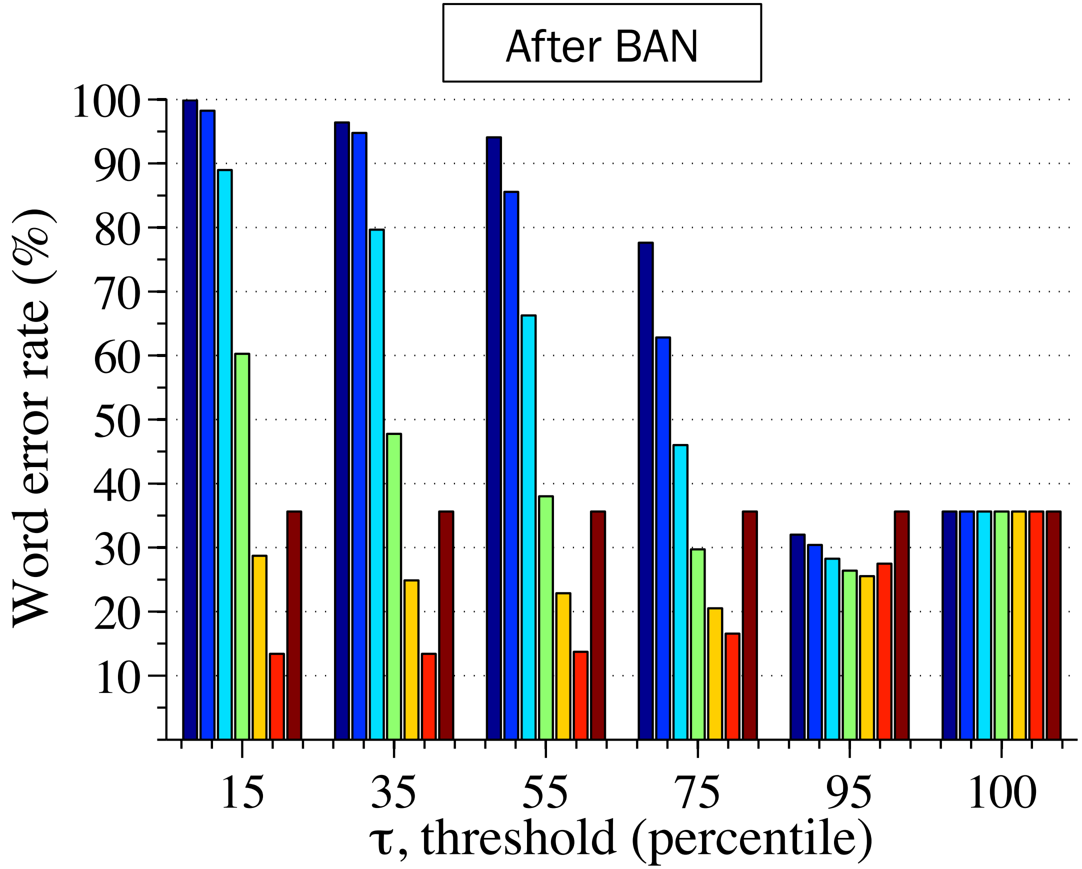
How well does BAN work?

- Experiment 2 (additive, channel noise at 15-dB SNR):



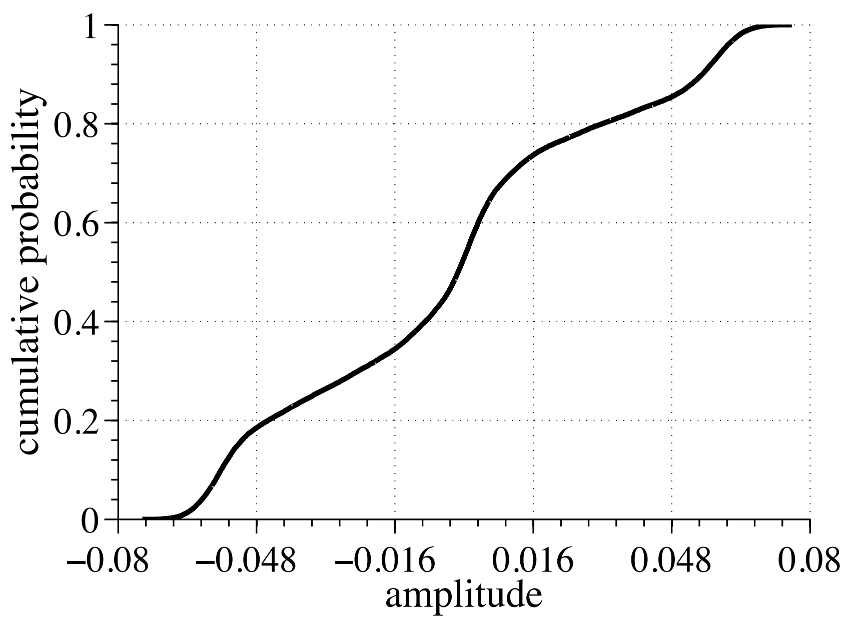
How well does BAN work?

- Experiment 2 (additive, channel noise at 15-dB SNR):

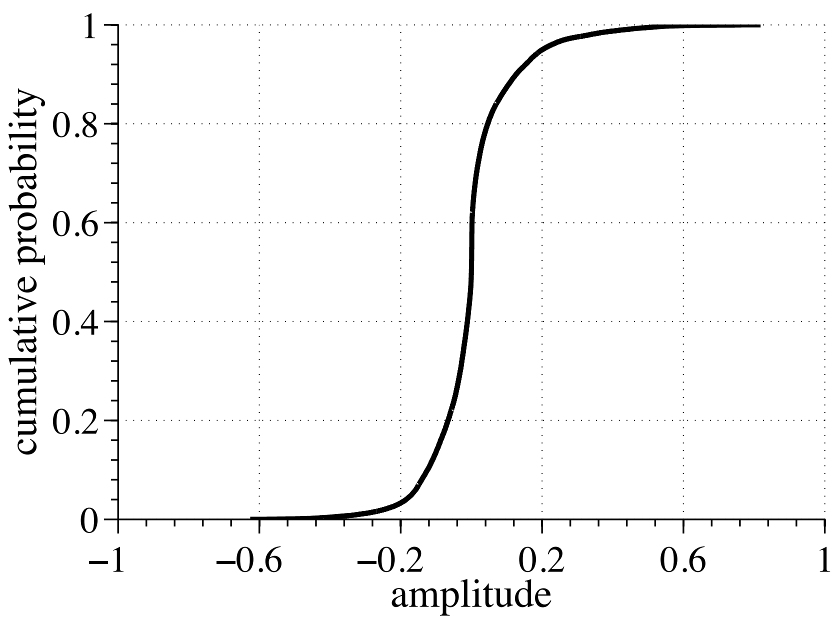


Robust BAN (Harvilla & Stern; unpub.)

- **Idea:** Shift each input sample by the amount the centroid of it and its neighbors is changed when inverting the nonlinearity.



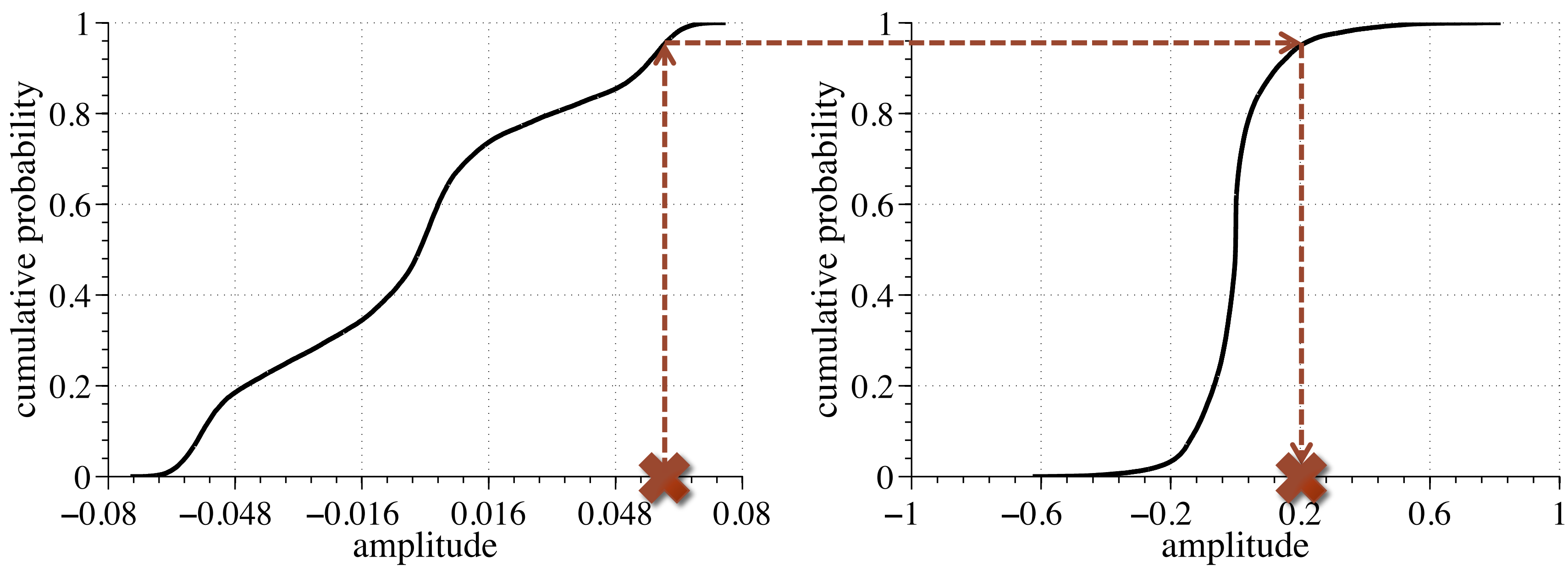
Observed speech after low-pass filter
(R = 10, τ = P_{50}, SNR = 15 dB)



Clean speech after low-pass filter

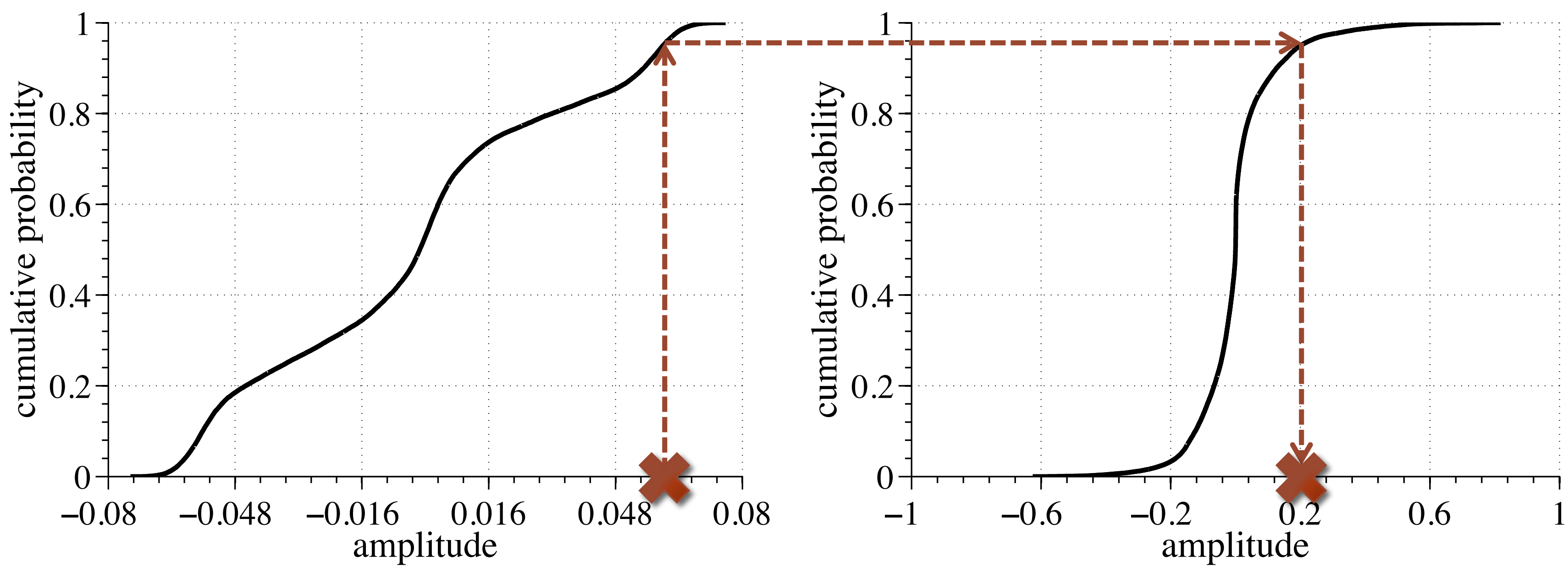
Robust BAN (Harvilla & Stern; unpub.)

- **Step 1:** As before, for a given reference signal amplitude, find the amplitude in the observed CDF with the same cumulative probability.



Robust BAN (Harvilla & Stern; unpub.)

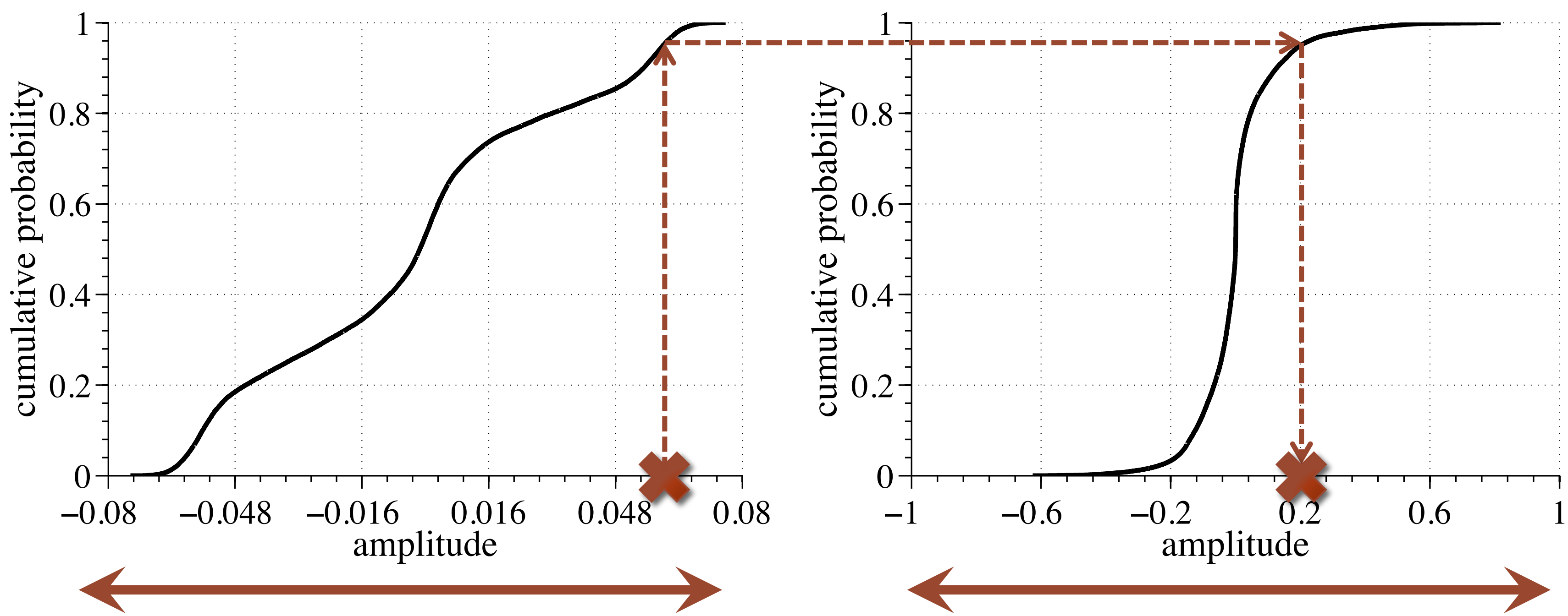
- **Step 2:** The difference between the output and the input is the **offset** to be added to the original, noisy and compressed waveform.



Offset = output – input
= 0.2 – 0.061 = 0.139

Robust BAN (Harvilla & Stern; unpub.)

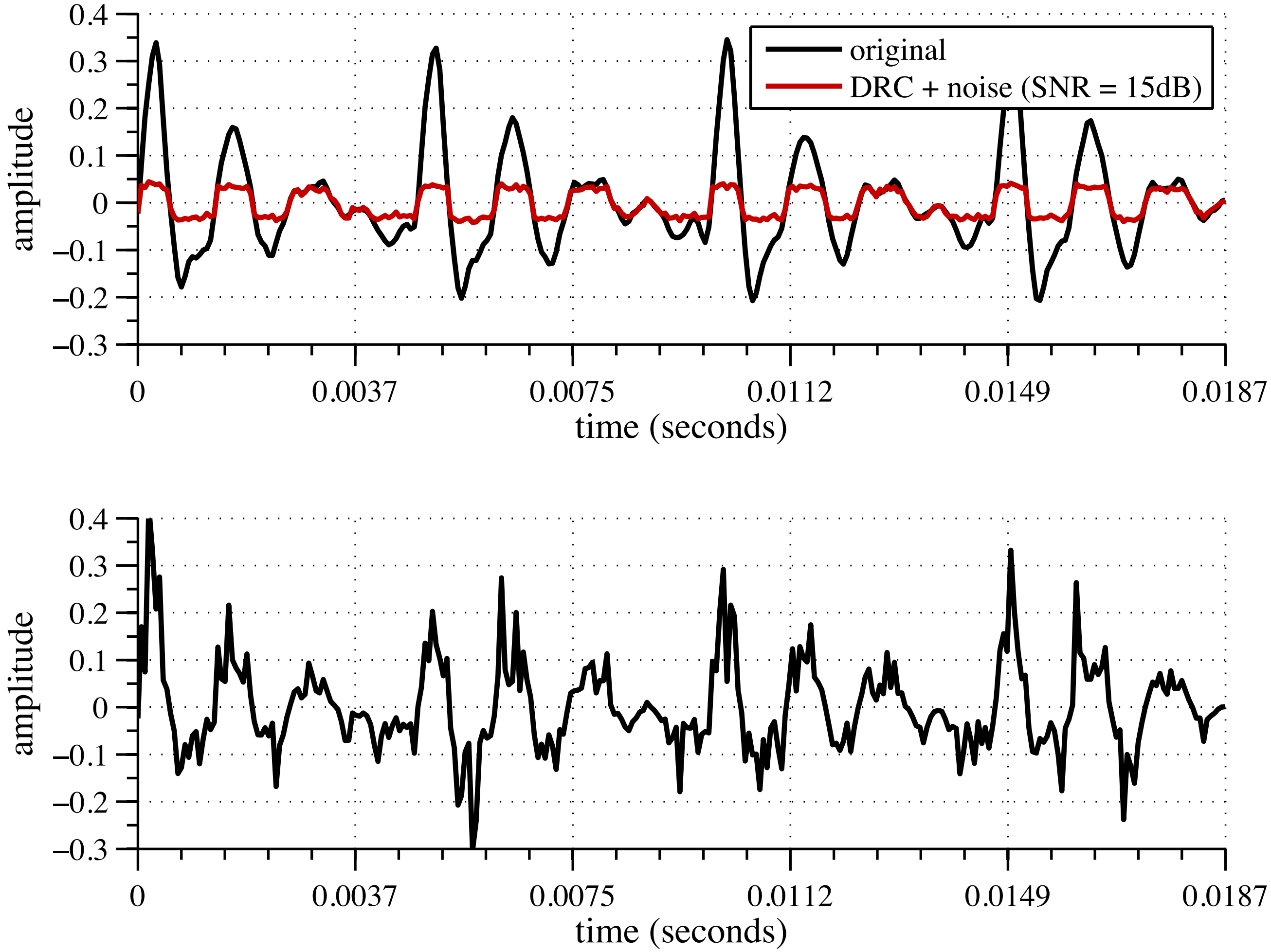
- **Step 3:** Repeat for each input signal amplitude, always using the inverse mapping defined by the smoothed signals.



Robust BAN (Harvilla & Stern; unpub.)

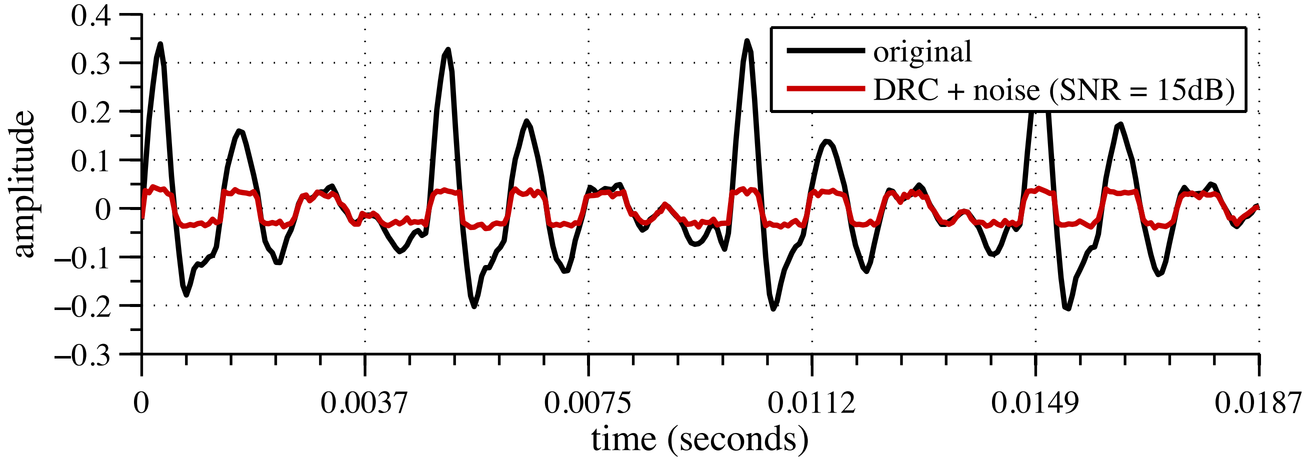
- **Step 1:** For each sample, find the centroid of the value and its surrounding 4 samples.
- **Step 2:** Pass the centroid value through the inverse nonlinearity estimate.
- **Step 3**: Find the difference ("offset") between the output of the inverse nonlinearity and the centroid.
- **Step 4:** Add the offset to the original noisy and compressed sample value from Step 1.
- **Step 5:** Repeat for each sample in the input signal.

Robust BAN (Harvilla & Stern; unpub.)

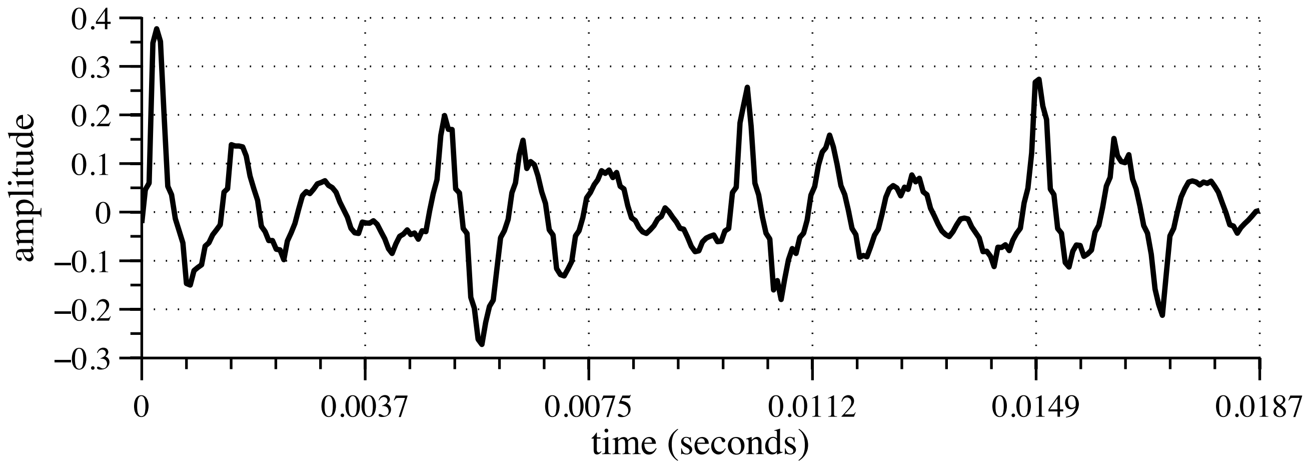


Repaired using BAN:

Robust BAN (Harvilla & Stern; unpub.)

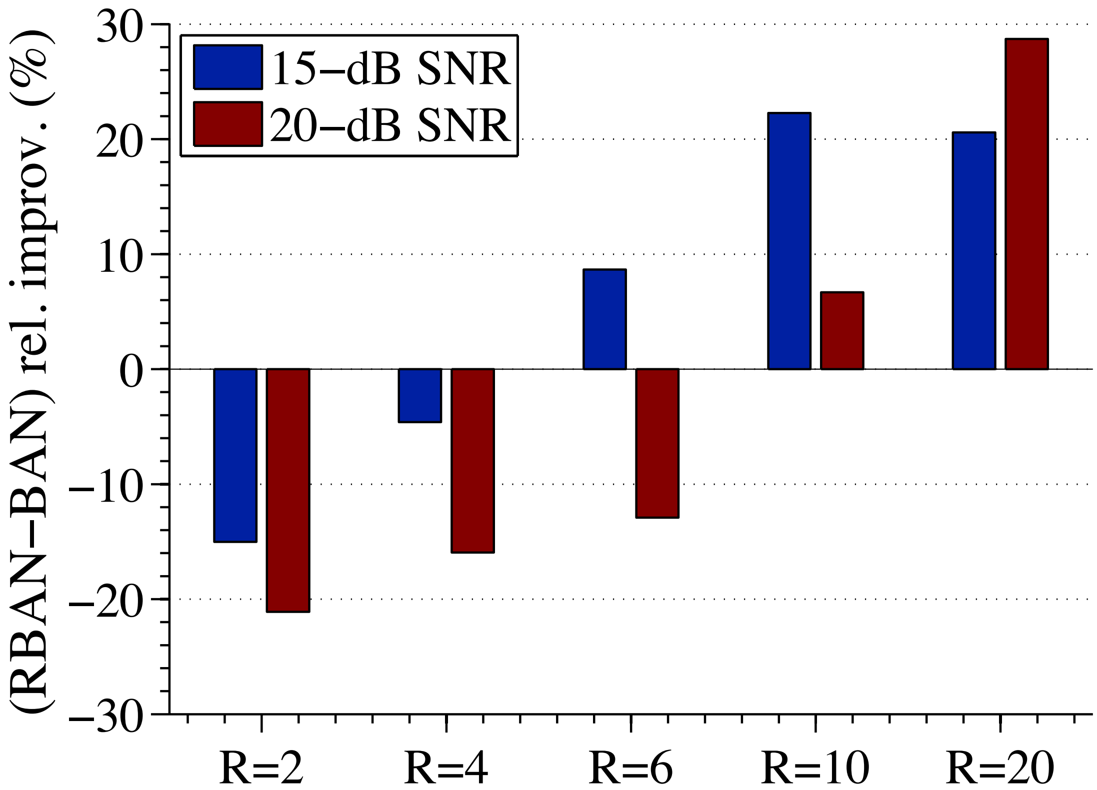


Repaired using Robust BAN:



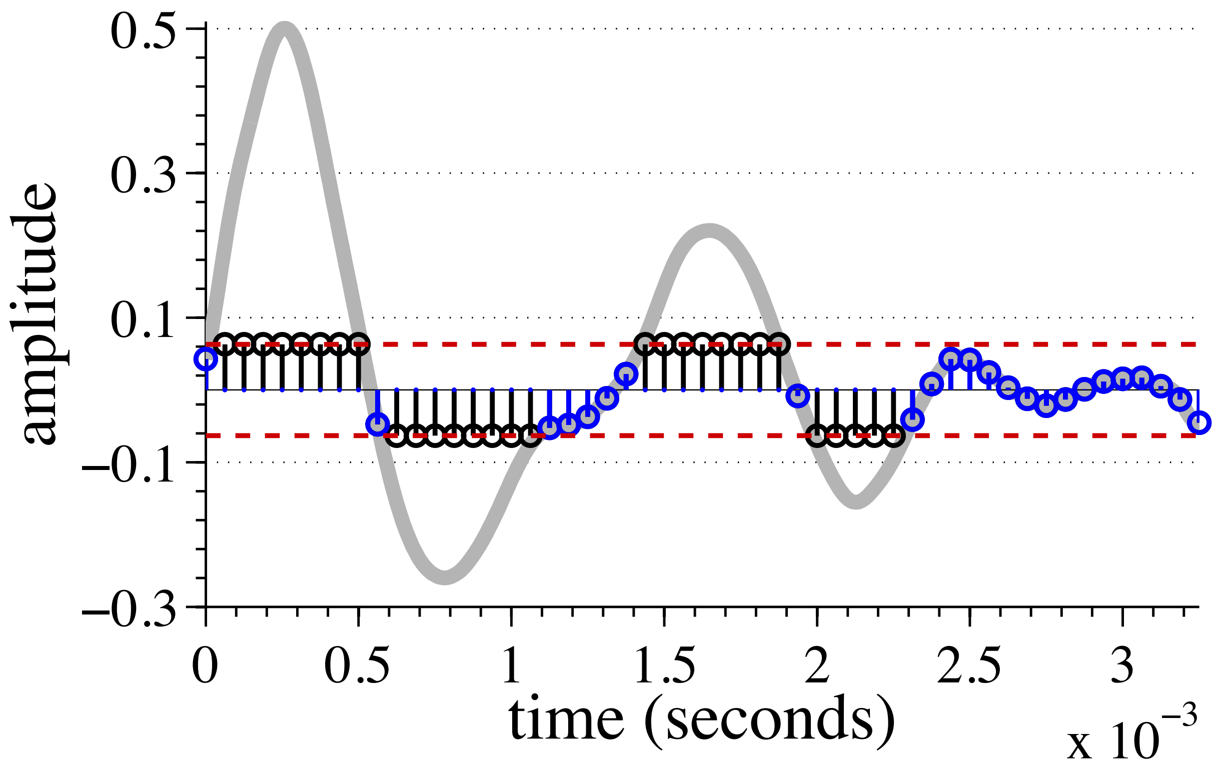
Results summary

- RBAN is more useful as *R* becomes large and SNR decreases:



Blind Amplitude Reconstruction (BAR)

- When R = ∞, BAN techniques are ineffective.
- All samples greater than |τ| are completely lost (“clipping”).



Consistent Iterative Hard Thresholding

(Kitic *et al.*; ICASSP 2013)

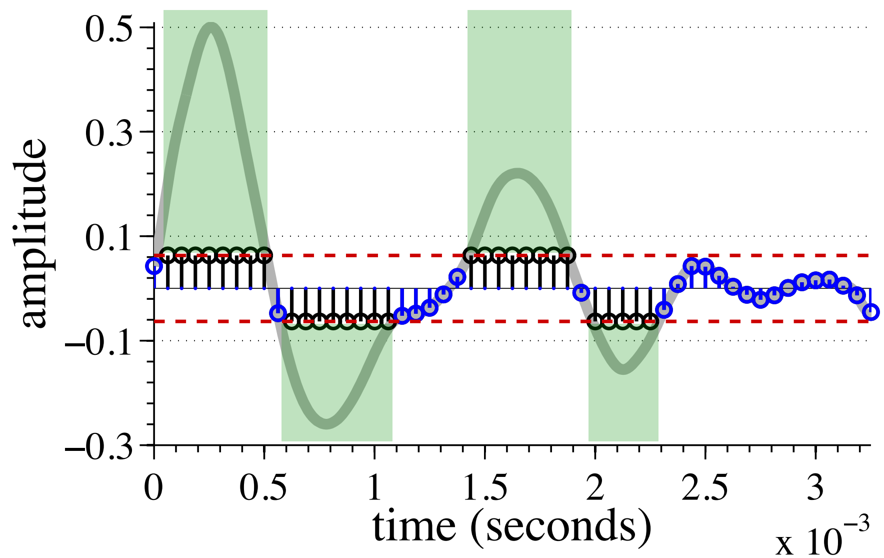
- **Kitic-IHT** works by learning a sparse representation of the incoming clipped speech in term of Gabor basis vectors.
- Learning is done using a modified version of the **Iterative Hard Thresholding** (IHT) algorithm.
- The learned sparse representation is then used to reconstruct the signal on a frame-by-frame basis.

$$\underbrace{\hat{\boldsymbol{x}}}_{\text{Repaired signal frame}} \approx \underbrace{\boldsymbol{\Psi}}_{\text{Gabor basis vectors}} \overbrace{\boldsymbol{\alpha}}^{\text{Sparse representation, learned from clipped observation}}$$

Kitic-IHT will be used as a baseline to compare novel declipping algorithm performance.

Constrained BAR (Harvilla & Stern; Interspeech 2014)

- Declip the signal by interpolating missing samples such that the **energy in the second derivative is minimized** (i.e., for smoothness).
- Ensure the interpolation **matches the sign of the clipped signal and is greater than |τ|** in the absolute sense.



Constrained BAR (Harvilla & Stern; Interspeech 2014)

- Explaining masking matrices

$$\boldsymbol{x} = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$$

$$\boldsymbol{S}_r = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Isolates reliable samples

$$\boldsymbol{S}_c = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Isolates clipped samples

$$\boldsymbol{S}_r\boldsymbol{x} = \boldsymbol{x}_r = \begin{bmatrix} x_0 \\ x_2 \\ x_4 \end{bmatrix} \qquad \boldsymbol{S}_c\boldsymbol{x} = \boldsymbol{x}_c = \begin{bmatrix} x_1 \\ x_3 \end{bmatrix} \qquad \boldsymbol{x} = \boldsymbol{S}_r^T\boldsymbol{x}_r + \boldsymbol{S}_c^T\boldsymbol{x}_c$$

Constrained BAR (Harvilla & Stern; Interspeech 2014)

CBAR objective function:

$$\underset{\boldsymbol{x}_c}{\text{minimize}} \quad \left|\left|\boldsymbol{D}_2(\boldsymbol{S}_r^T\boldsymbol{x}_r + \boldsymbol{S}_c^T\boldsymbol{x}_c)\right|\right|_2^2$$

$$\text{subject to} \quad \boldsymbol{x}_c \circ \operatorname{sgn} \boldsymbol{S}_c\boldsymbol{x} \geq +\tau\mathbf{1}$$

Constrained BAR (Harvilla & Stern; Interspeech 2014)

- Because Constrained BAR (CBAR) imposes a hard constraint when minimizing the objective function, **it is very slow.**
- **A line search** algorithm is used to solve the constrained optimization separately for every frame.
- In the worst case, **it is 400 times slower than real time.**
- This motivates the development of a declipping algorithm that **does not require a hard constraint.**

Regularized BAR (Harvilla & Stern; ICASSP 2015)

- Replace CBAR's hard constraint with regularization terms:

CBAR objective function:

$$\underset{\boldsymbol{x}_c}{\text{minimize}} \quad \left|\left|\boldsymbol{D}_2(\boldsymbol{S}_r^T\boldsymbol{x}_r + \boldsymbol{S}_c^T\boldsymbol{x}_c)\right|\right|_2^2$$

$$\text{subject to} \quad \boldsymbol{x}_c \circ \operatorname{sgn} \boldsymbol{S}_c\boldsymbol{x} \geq +\tau\boldsymbol{1}$$

Regularized BAR (Harvilla & Stern; ICASSP 2015)

- Replace CBAR's hard constraint with regularization terms:

$$\underset{\boldsymbol{x}_c}{\text{minimize}} \quad \left|\left|\boldsymbol{D}_2(\boldsymbol{S}_r^T\boldsymbol{x}_r + \boldsymbol{S}_c^T\boldsymbol{x}_c)\right|\right|_2^2$$

Regularized BAR (Harvilla & Stern; ICASSP 2015)

- Replace CBAR's hard constraint with regularization terms:

$$\underset{\boldsymbol{x}_c}{\text{minimize}} \quad \left|\left|\boldsymbol{D}_2(\boldsymbol{S}_r^T\boldsymbol{x}_r + \boldsymbol{S}_c^T\boldsymbol{x}_c)\right|\right|_2^2 + \lambda\left|\left|\boldsymbol{t}_0 - \boldsymbol{S}_c^+\boldsymbol{x}_c\right|\right|_2^2 + \lambda\left|\left|\boldsymbol{t}_1 - \boldsymbol{S}_c^-\boldsymbol{x}_c\right|\right|_2^2$$

Regularized BAR (Harvilla & Stern; ICASSP 2015)

- Replace CBAR's hard constraint with regularization terms:

RBAR objective function:

$\boldsymbol{x}_c$ can be solved for in closed form!

$$\underset{\boldsymbol{x}_c}{\text{minimize}} \quad \left|\left|\boldsymbol{D}_2(\boldsymbol{S}_r^T\boldsymbol{x}_r + \boldsymbol{S}_c^T\boldsymbol{x}_c)\right|\right|_2^2 + \lambda\left|\left|\boldsymbol{t}_0 - \boldsymbol{S}_c^+\boldsymbol{x}_c\right|\right|_2^2 + \lambda\left|\left|\boldsymbol{t}_1 - \boldsymbol{S}_c^-\boldsymbol{x}_c\right|\right|_2^2$$

Regularized BAR (Harvilla & Stern; ICASSP 2015)

- Replace CBAR's hard constraint with regularization terms:

Frame-specific solution:

$\boldsymbol{x}_c$ can be solved for in closed form!

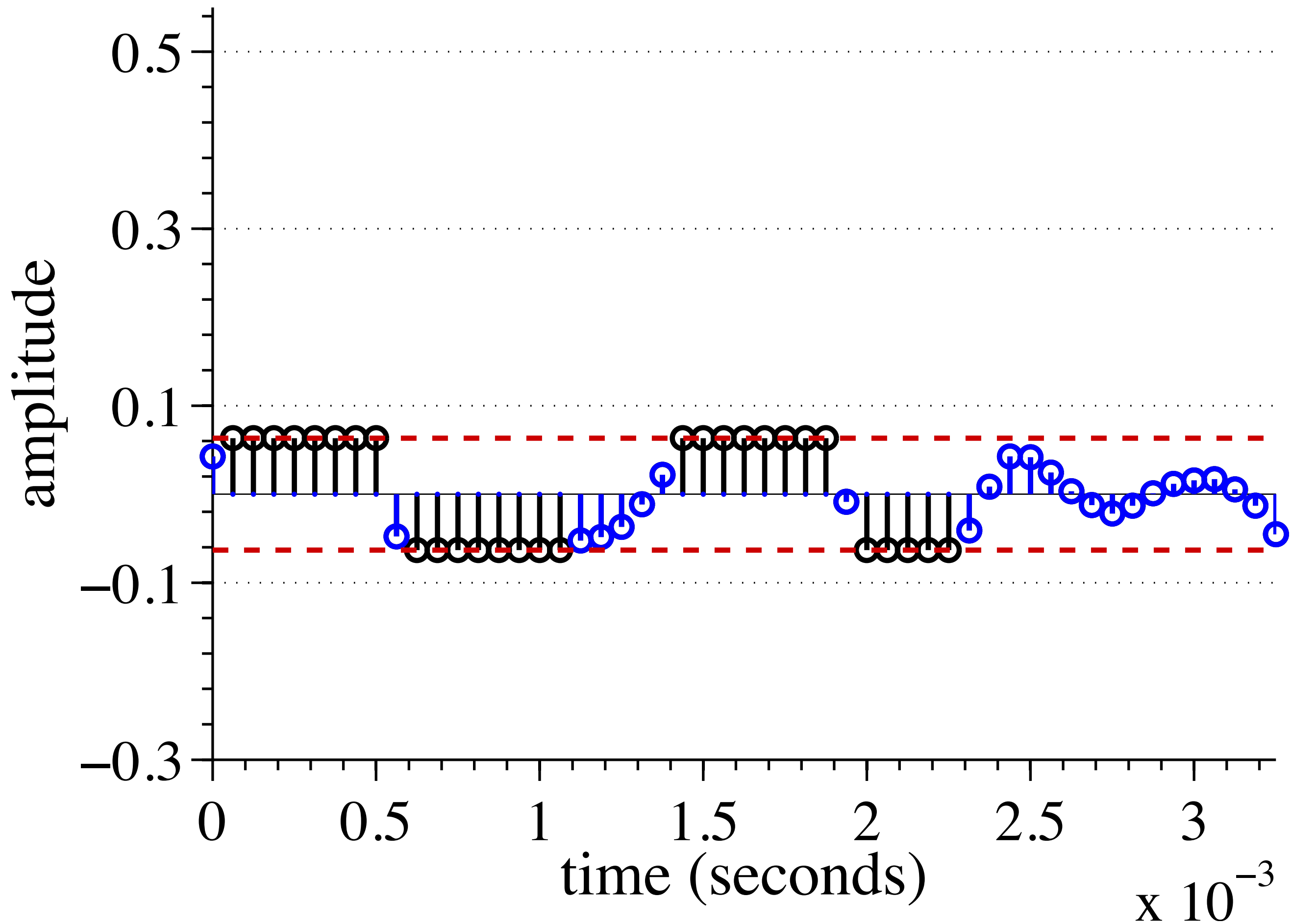
$$\hat{\boldsymbol{x}}_c = -(\boldsymbol{S}_c\boldsymbol{D}_2^T\boldsymbol{D}_2\boldsymbol{S}_c^T + \lambda(\boldsymbol{S}_c^+)^T\boldsymbol{S}_c^+ + \lambda(\boldsymbol{S}_c^-)^T\boldsymbol{S}_c^-)^{-1} \times(\boldsymbol{S}_c\boldsymbol{D}_2^T\boldsymbol{D}_2\boldsymbol{S}_r^T\boldsymbol{x}_r - \lambda(\boldsymbol{S}_c^+)^T\boldsymbol{t}_0 - \lambda(\boldsymbol{S}_c^-)^T\boldsymbol{t}_1)$$

$$\hat{\boldsymbol{x}} = \boldsymbol{S}_r^T\boldsymbol{x}_r + \boldsymbol{S}_c^T\hat{\boldsymbol{x}}_c$$

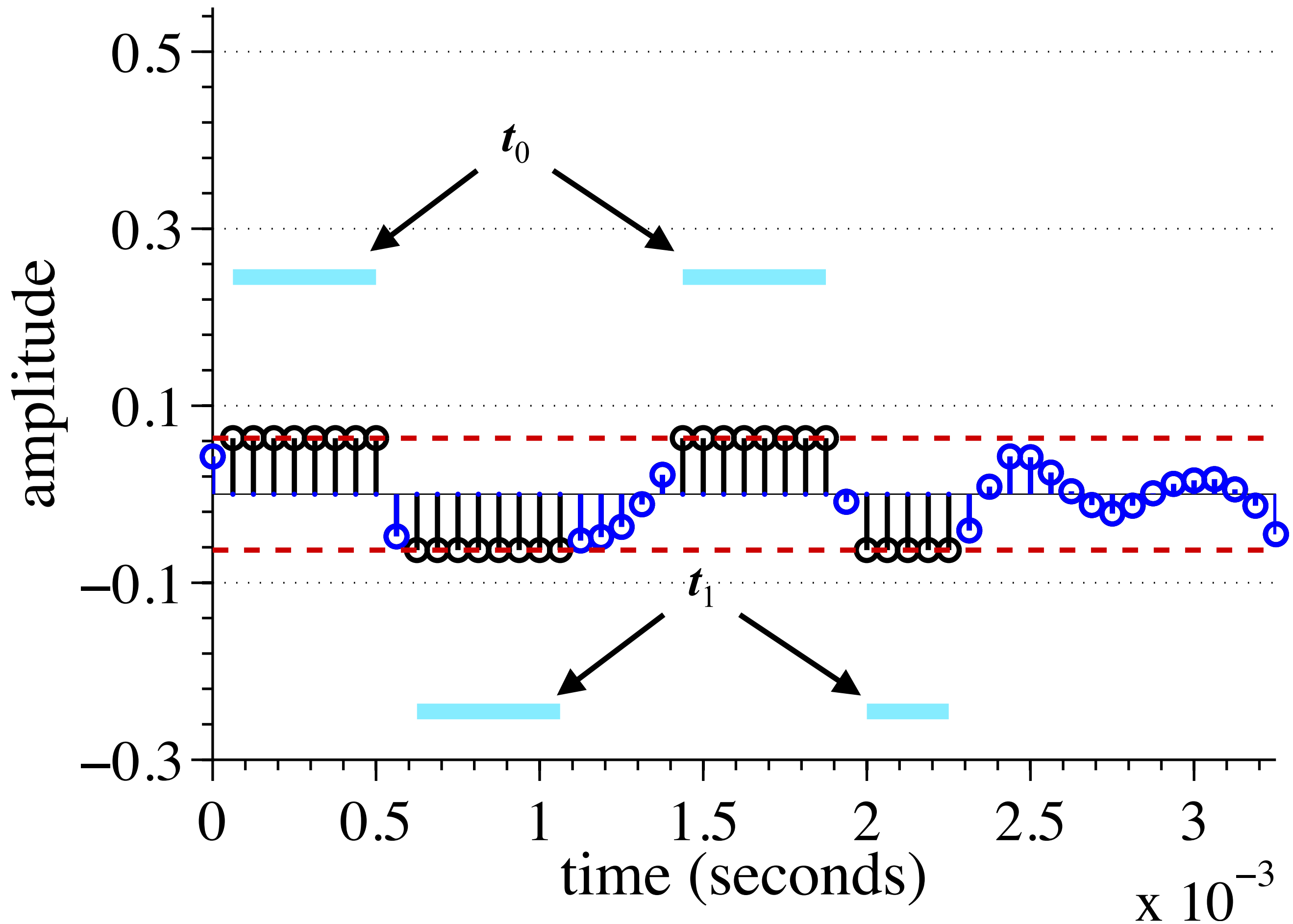
Regularized BAR (Harvilla & Stern; ICASSP 2015)

- The $\boldsymbol{t}_0$ and $\boldsymbol{t}_1$ terms are target vectors.
- They "float" above the clipped segments at the target amplitude.
- They are defined as a function of the fraction of clipped samples in a frame.

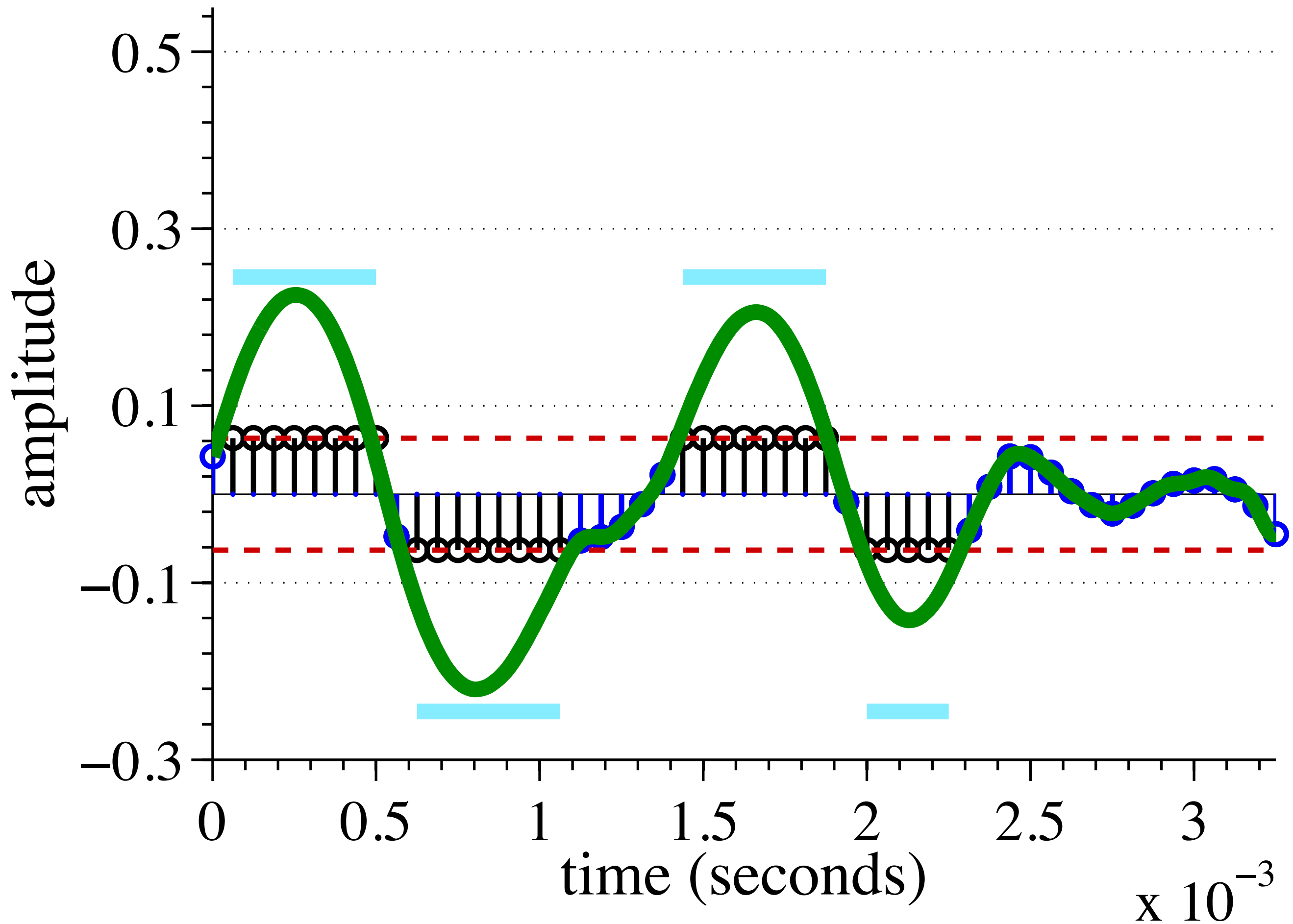
Regularized BAR (Harvilla & Stern; ICASSP 2015)



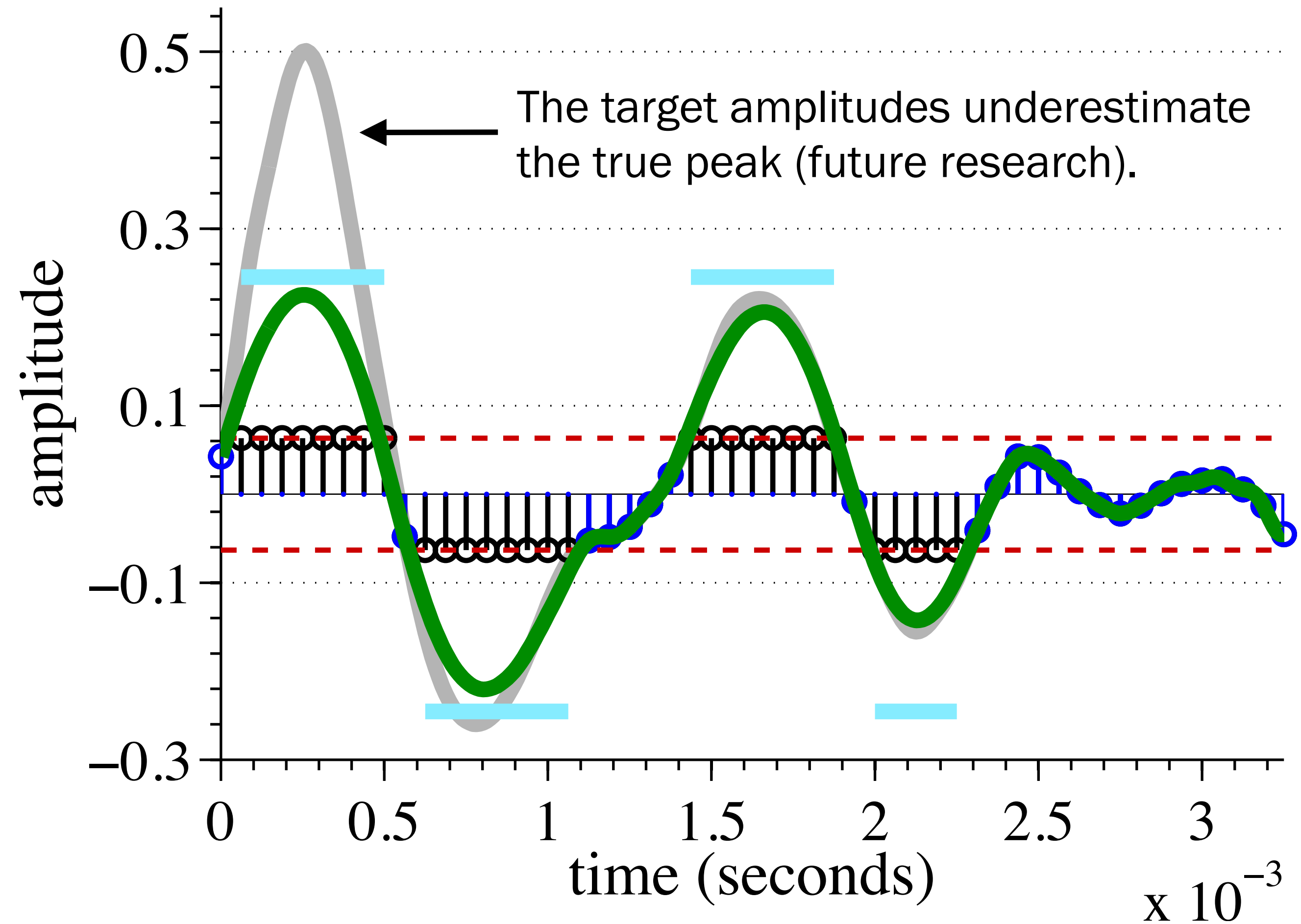
Regularized BAR (Harvilla & Stern; ICASSP 2015)



Regularized BAR (Harvilla & Stern; ICASSP 2015)

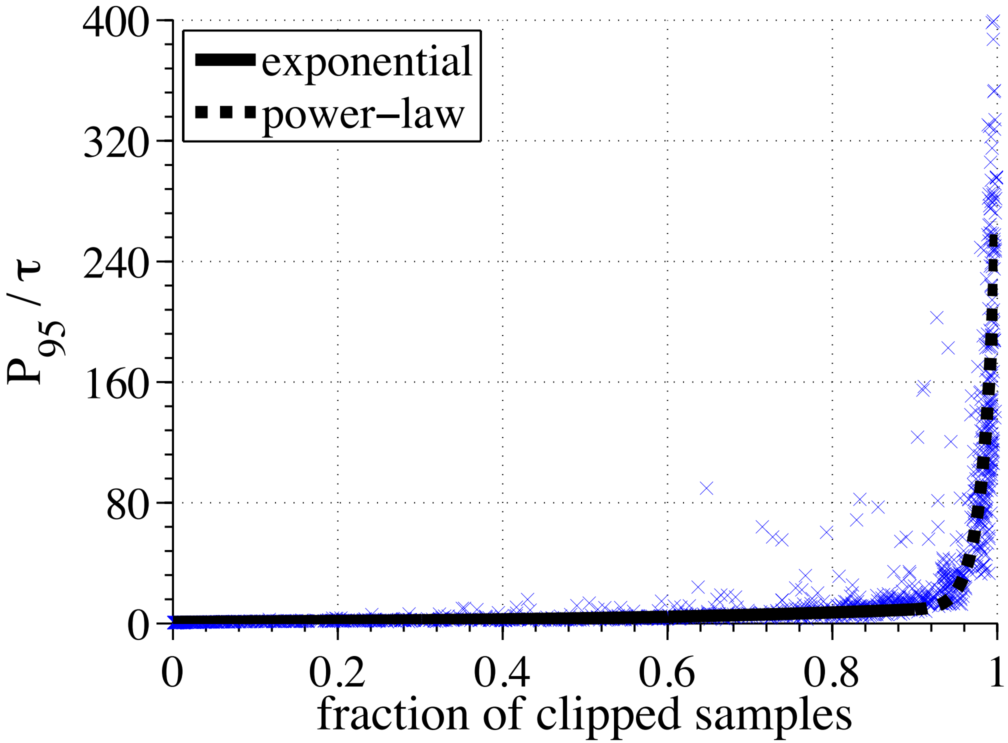


Regularized BAR (Harvilla & Stern; ICASSP 2015)



Regularized BAR (Harvilla & Stern; ICASSP 2015)

- Amplitude prediction



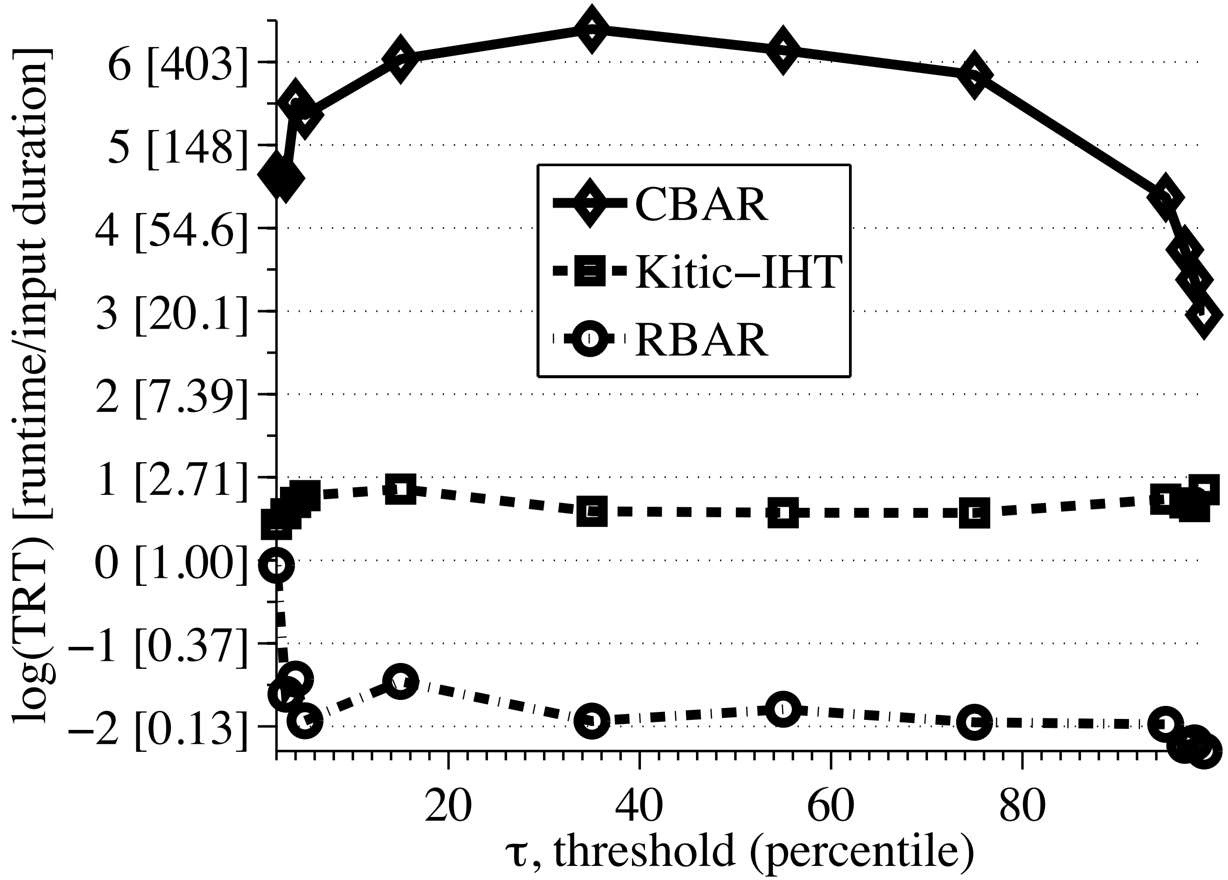
ρ: fraction of clipped samples in frame

$$\phi(\rho) \approx \begin{cases} e^{2.5\rho} & \text{for } \rho \leq 0.9 \\ 272\rho^{60} + 9.0 & \text{for } \rho > 0.9 \end{cases}$$

$$\boldsymbol{t}_0 = \phi(\rho)\tau\mathbf{1}$$

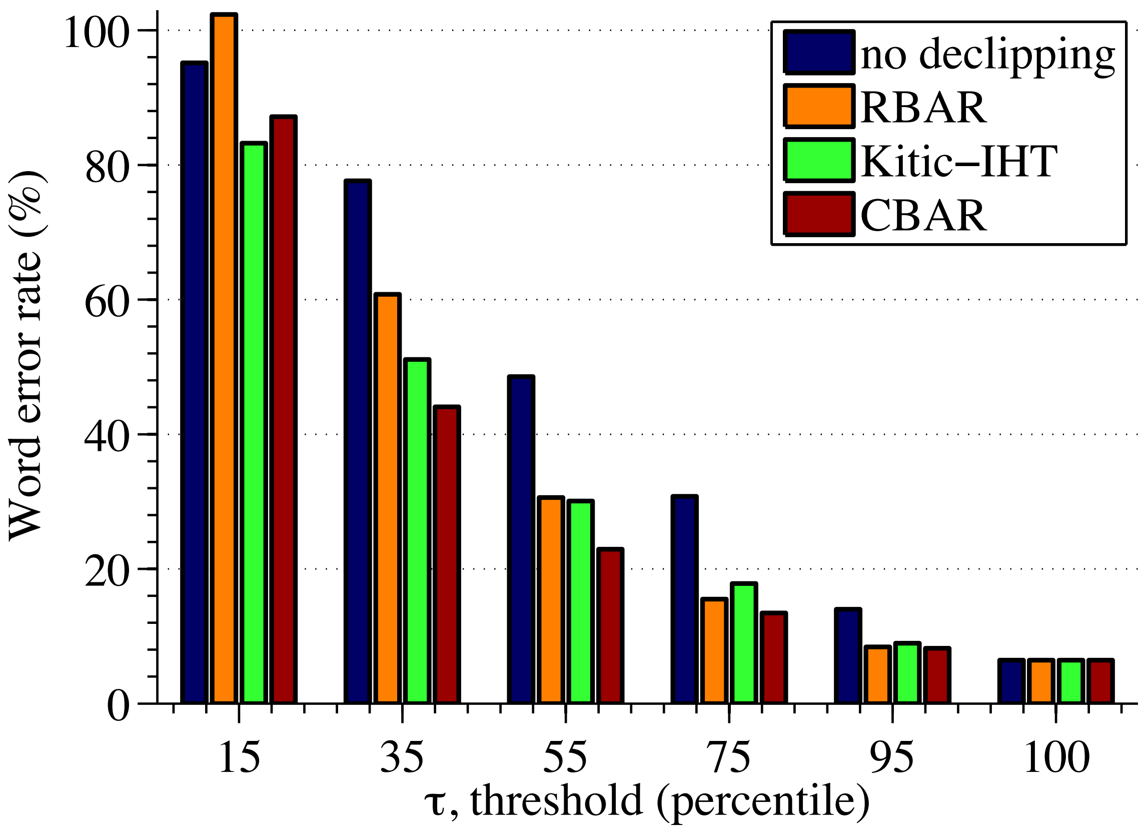
$$\boldsymbol{t}_1 = -\phi(\rho)\tau\mathbf{1}$$

Processing speed



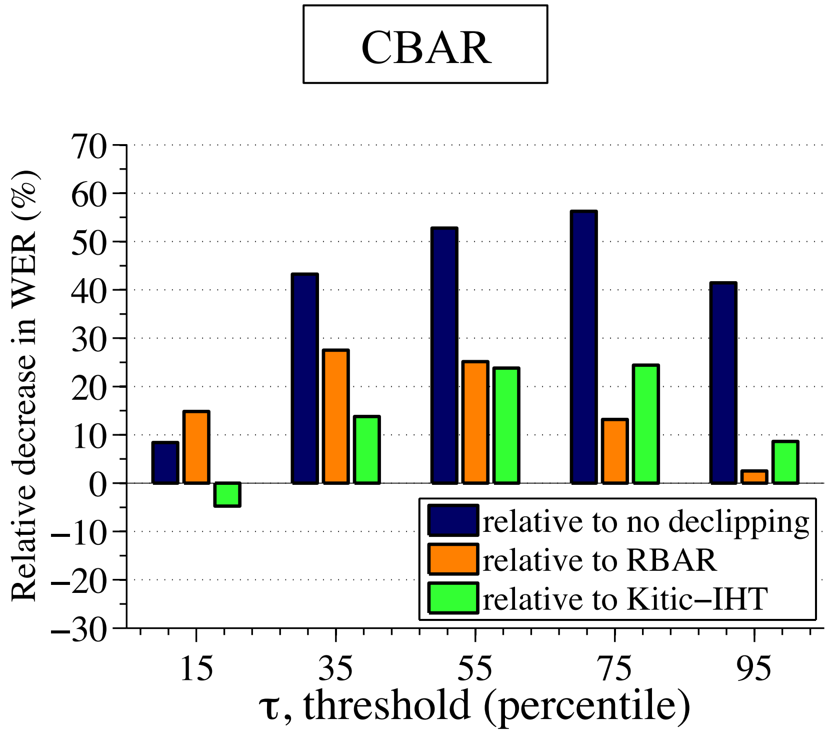
Declipping performance

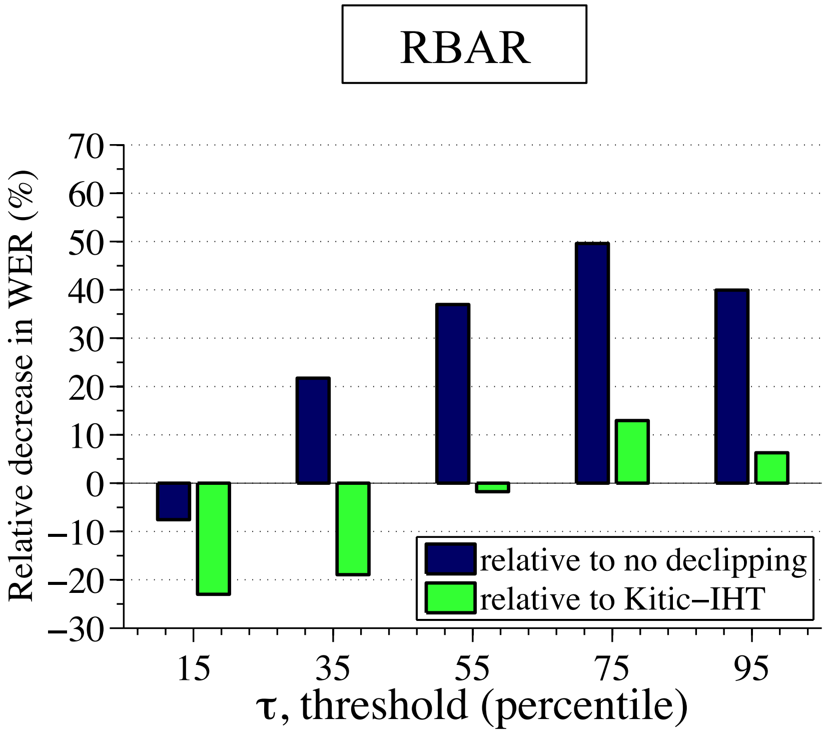
- Experiment 1 (no additive noise):



Declipping performance

- Experiment 1 (no additive noise), relative improvements:

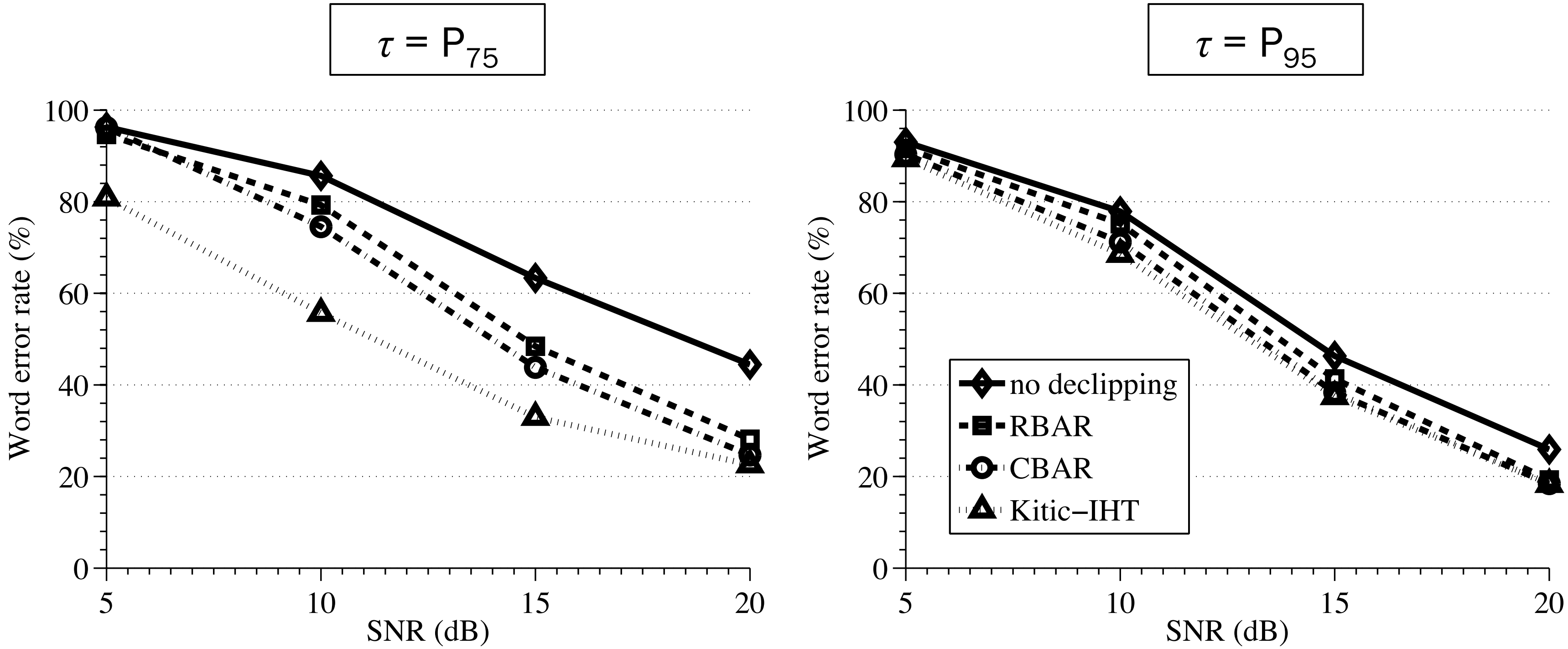




Declipping performance

- Experiment 2 (additive noise):

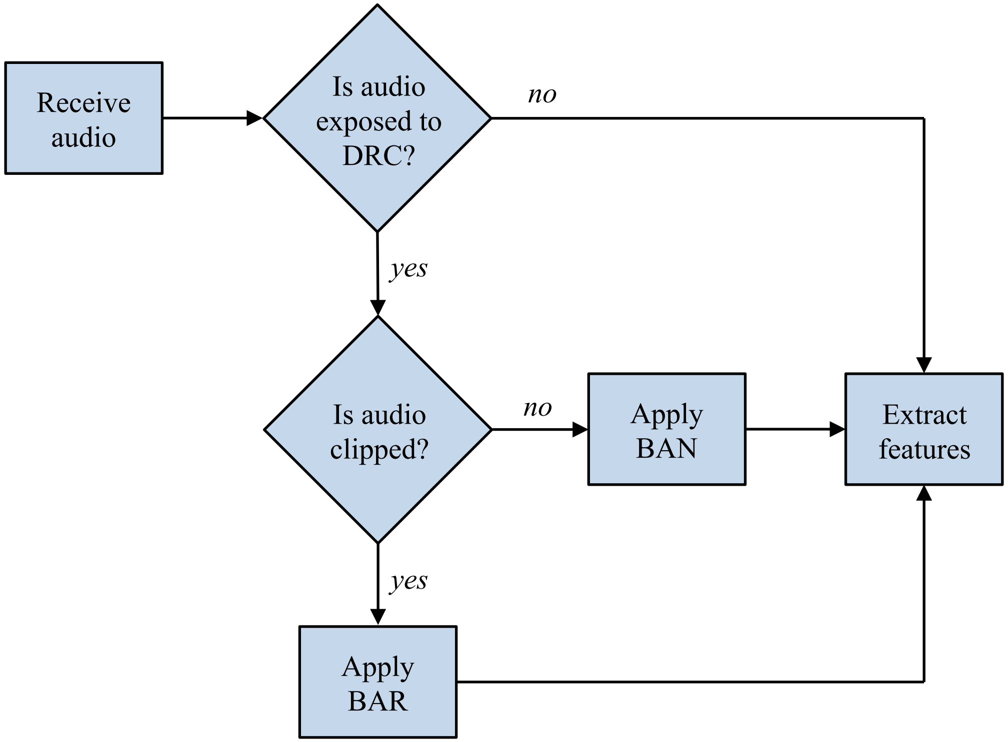
The location of all clipped samples is assumed known.



Kitic-IHT is more robust to additive noise (future research).

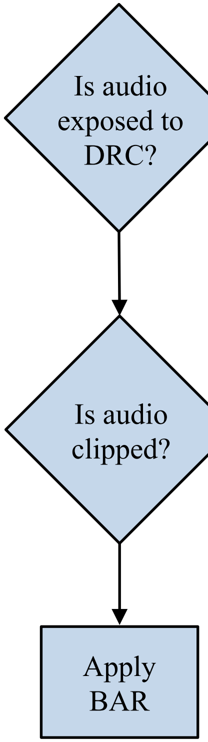
Robust Estimation of Distortion (RED)

- Given a received speech signal, how does one determine if declipping (BAR) or decompression (BAN) need to be performed?



Robust Estimation of Distortion (RED)

- Given a received speech signal, how does one determine if declipping (BAR) or decompression (BAN) need to be performed?



Search for peaks in the probability distribution of the waveform amplitudes ✔

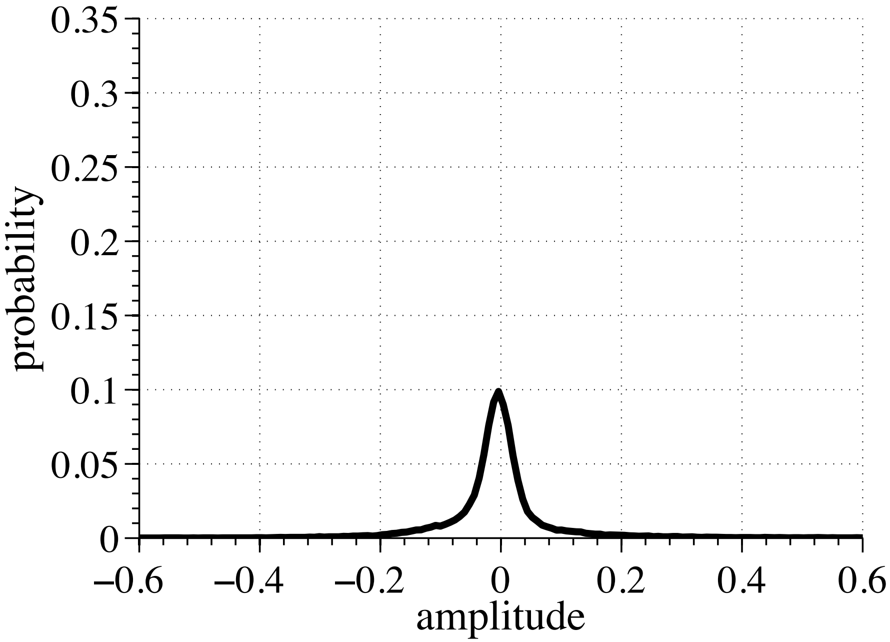
Accurately estimate the value of *R* (recall: if *R* is “very” large, speech is effectively clipped) ✗

Requires estimation of which samples are clipped and must assume the possibility of noise (e.g., as in Experiment 2) ✔

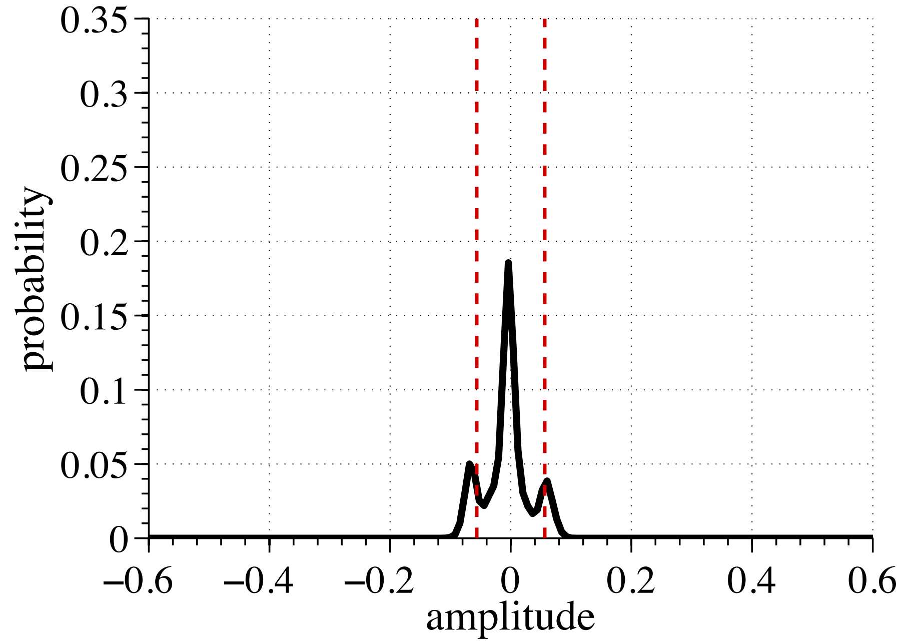
Clipped speech detection & τ estimation

(Harvilla & Stern; ICASSP 2015)

- Exposure to DRC significantly modifies the waveform amplitude distribution of the speech



Uncompressed speech with noise at 15-dB SNR

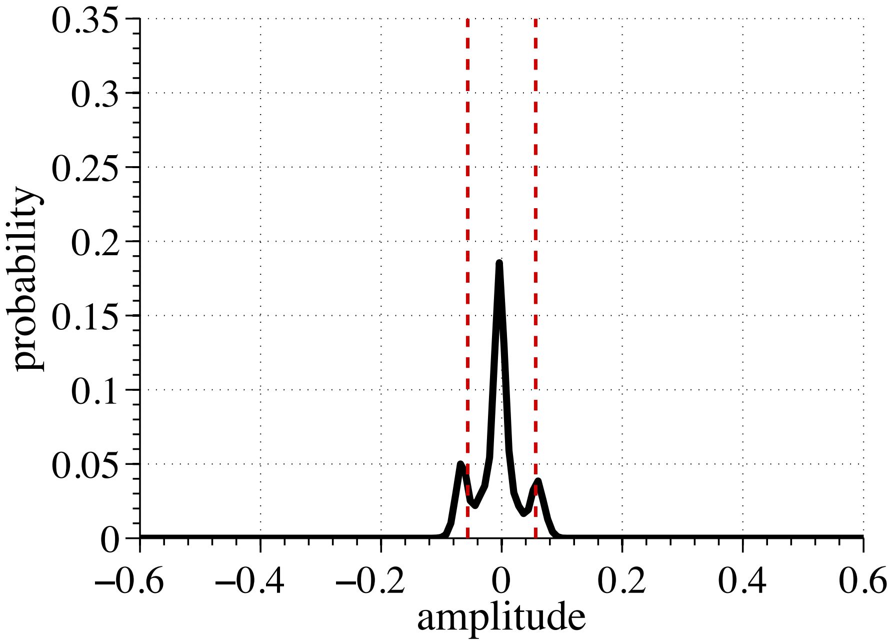


DRC'ed speech (R=6, τ=0.06) + noise at 15 dB

Clipped speech detection & τ estimation

(Harvilla & Stern; ICASSP 2015)

- Exposure to DRC significantly modifies the waveform amplitude distribution of the speech



DRC'ed speech (R=6, τ=0.06)
+ noise at 15 dB

<u>Clipping detection and τ estimation algorithm:</u>

1. Detect peaks in the distribution
2. Compute:

$$\hat{\tau} = \frac{1}{K-1}\sum_{i=0}^{K-1}|k_i|$$

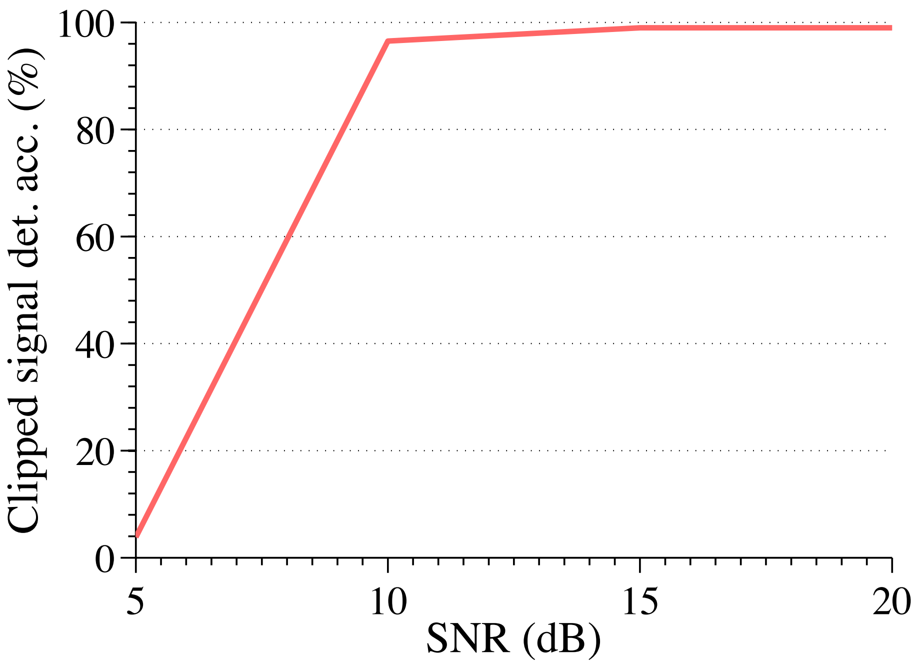
3. Output indicates clipping occurrence and amplitude value of τ (0.5*(|-τ| + 0 + |τ|))

(if output is ∞, no clipping)

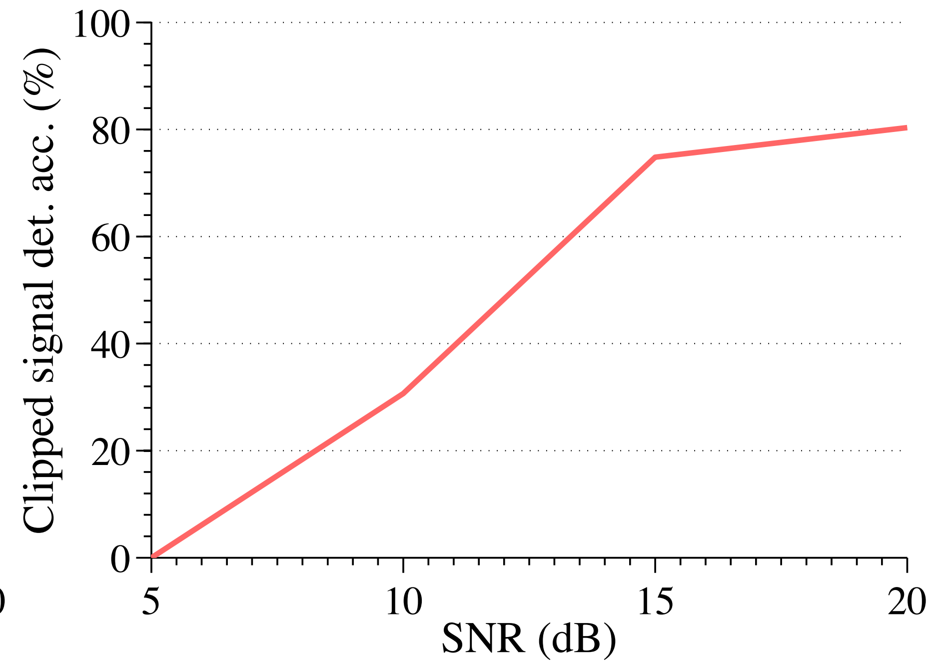
Clipped speech detection & τ estimation

(Harvilla & Stern; ICASSP 2015)

Clipped signal detection accuracies



$\tau = P_{95}$



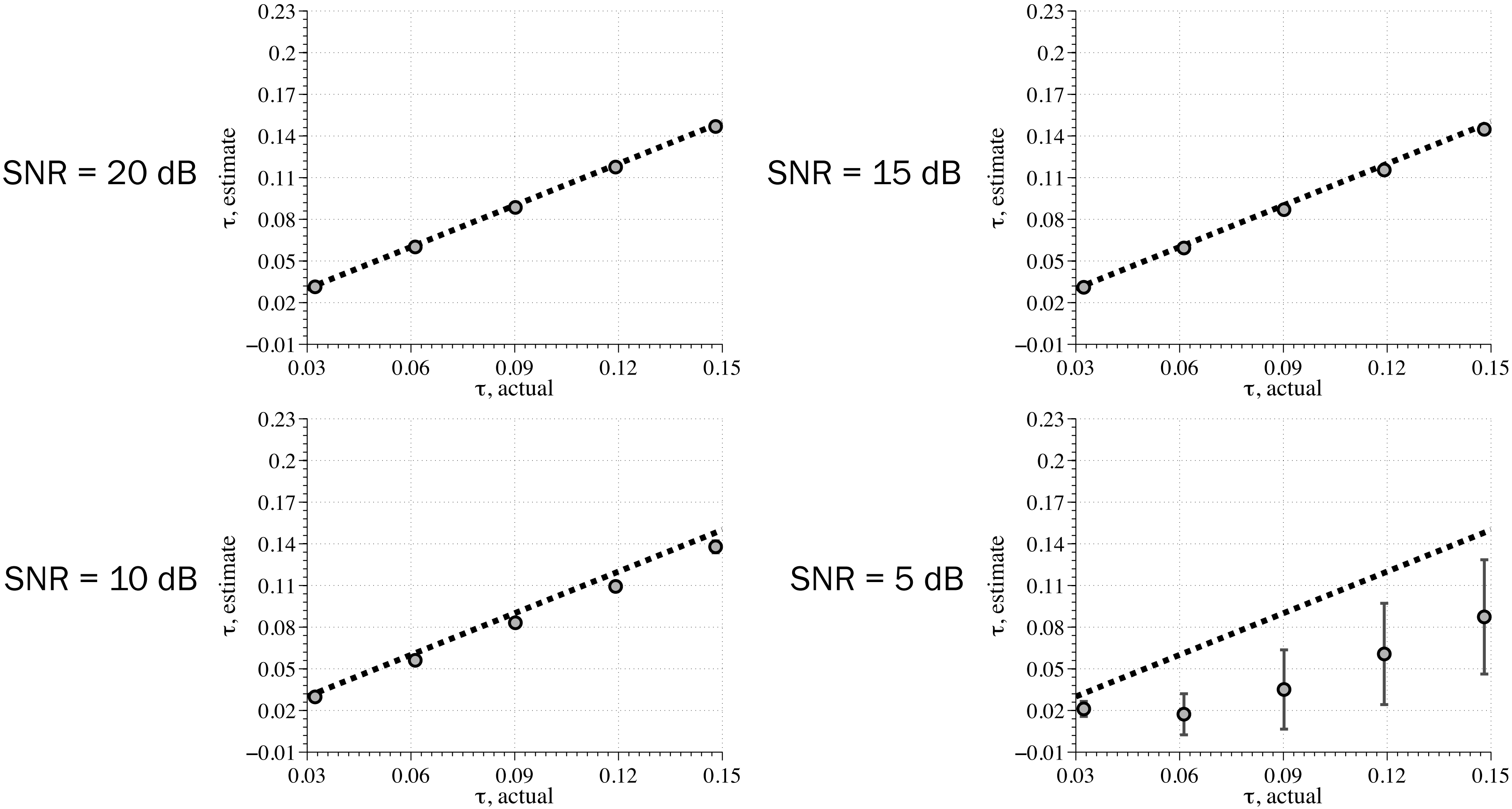
$\tau = P_{75}$

Because the amplitude distribution merges into one lobe (thus, one peak) with decreasing SNR and τ, detection accuracy correspondingly decreases.

Clipped speech detection & τ estimation

(Harvilla & Stern; ICASSP 2015)

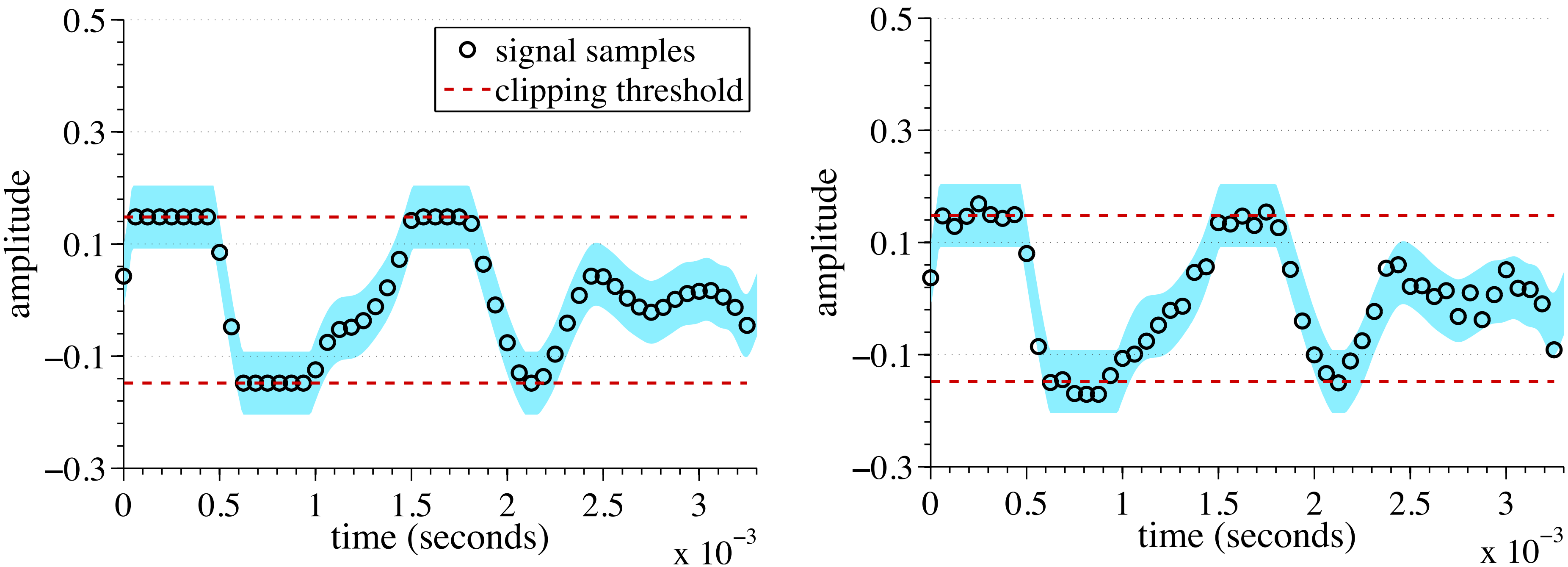
τ-estimation accuracies for $R = \infty$



Clipped sample estimation

(Harvilla & Stern; ICASSP 2015)

- Given the amplitude value of τ, how do we determine the location of clipped samples?



Clipped speech, no noise

Clipped speech + noise at 10-dB SNR

Clipped sample estimation

(Harvilla & Stern; ICASSP 2015)

- Given the amplitude value of τ, how do we determine the location of clipped samples?
- Solution:

Given,

amplitude value of τ

percentile value of τ

variance of the additive noise (σ_w^2)

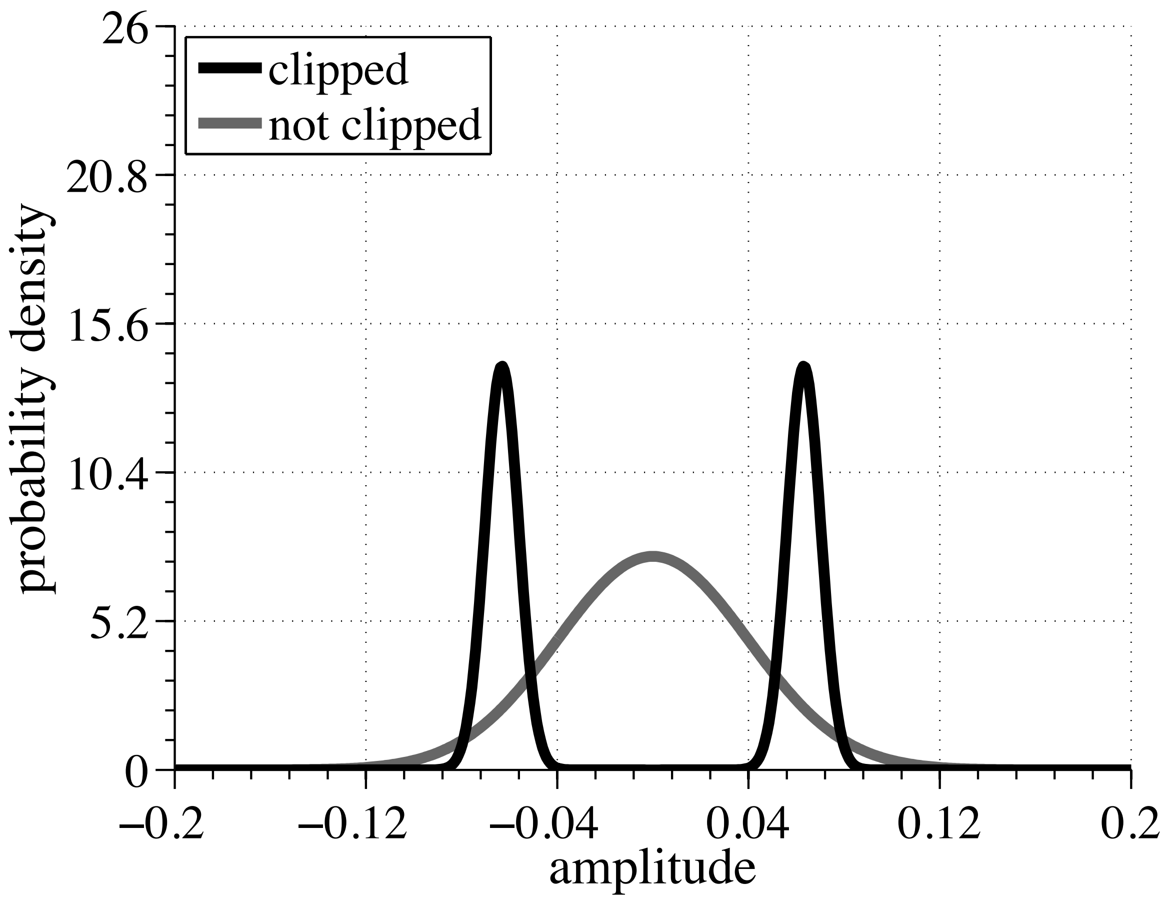
variance of the observed signal (σ_y^2)

- Model the clean speech and noise with separate Gaussians
- For each sample, classify as clipped if

$$Pr(\text{ clipped} \mid \text{observed sample}, \tau, \sigma_w^2, \sigma_y^2) > \Pr(\text{ not clipped} \mid \text{observed sample}, \tau, \sigma_w^2, \sigma_y^2)$$

Clipped sample estimation

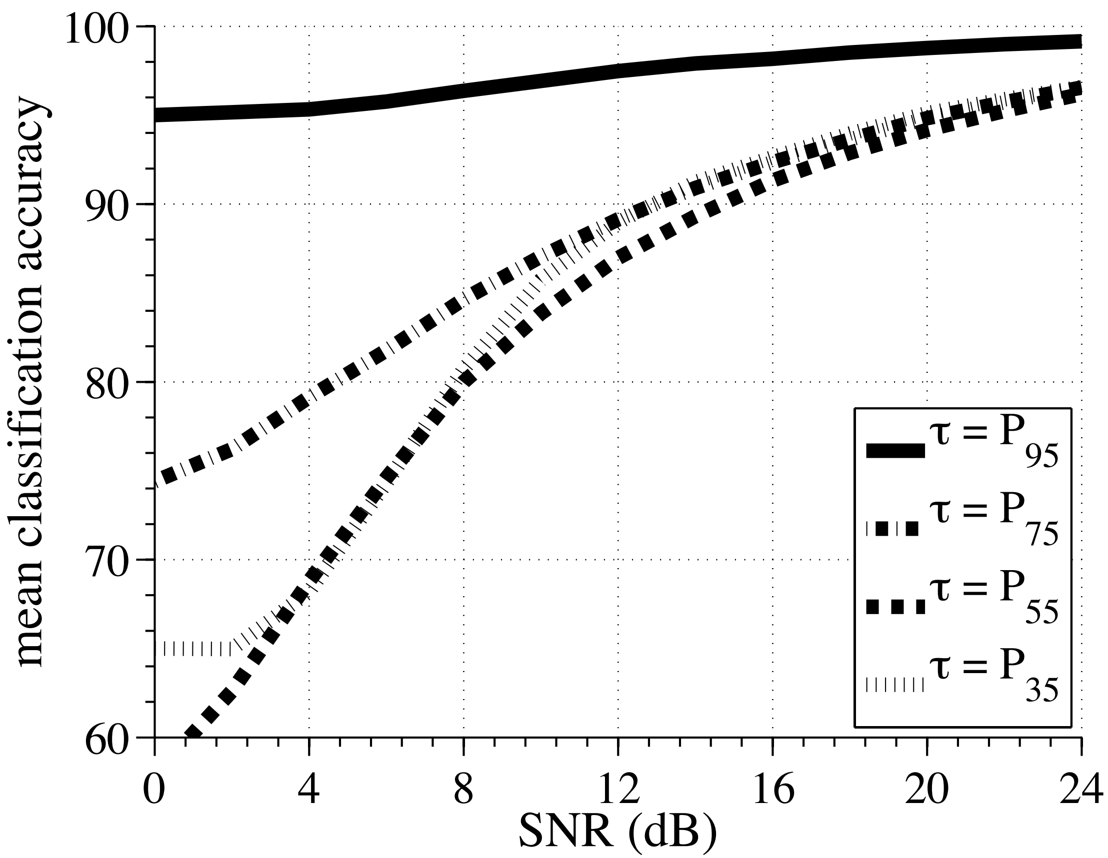
(Harvilla & Stern; ICASSP 2015)



Speech clipped at τ = 0.07 and added to noise at 15-dB SNR

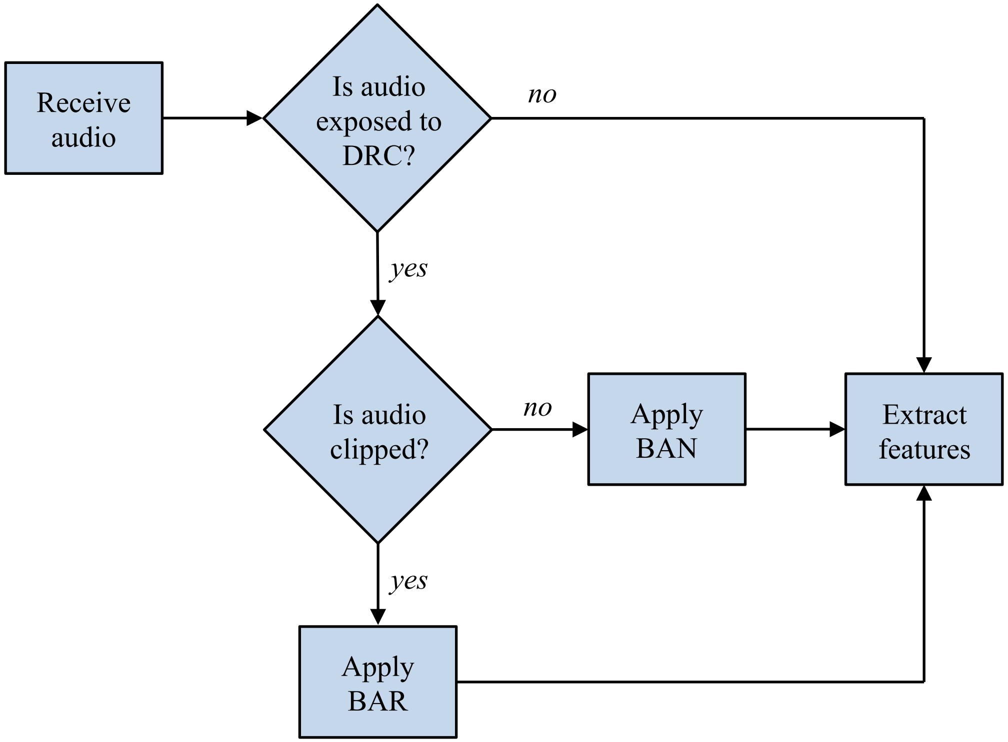
Clipped sample estimation

(Harvilla & Stern; ICASSP 2015)



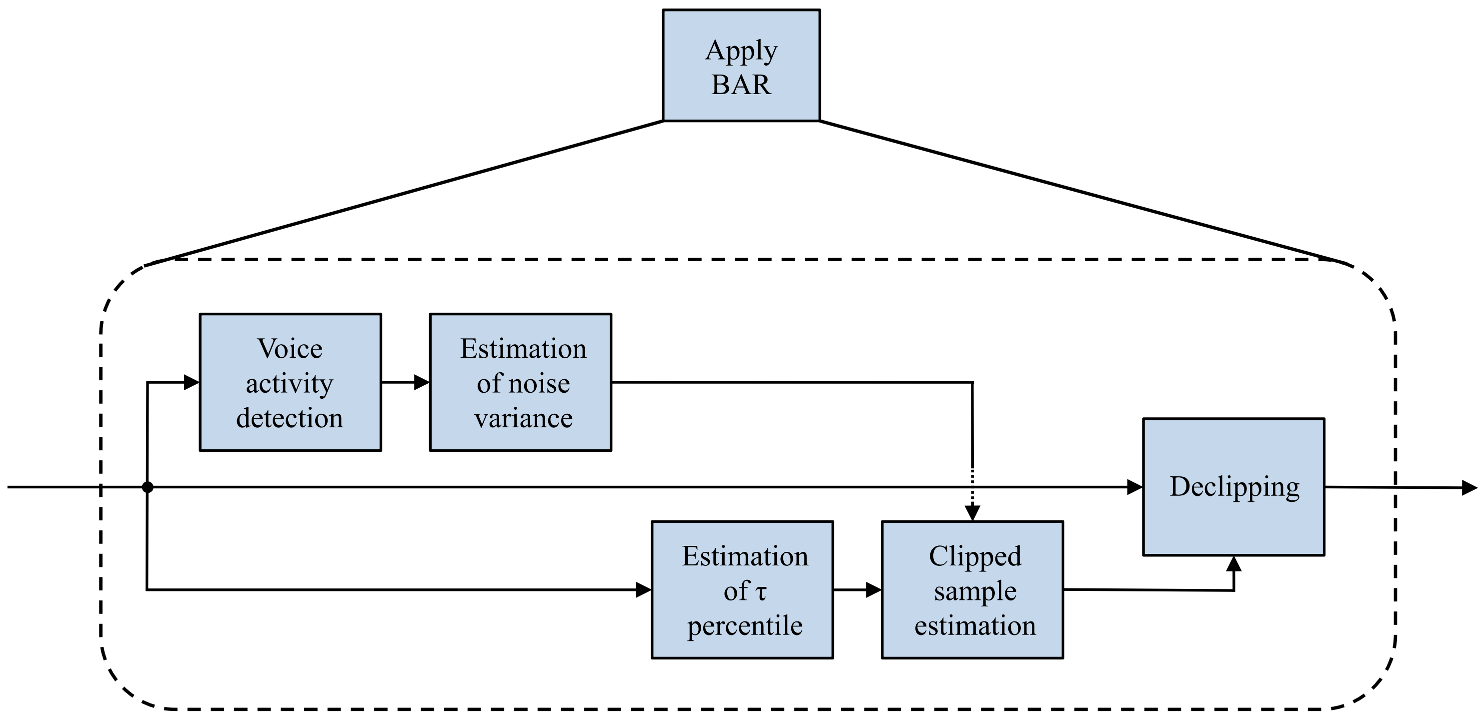
Robust Estimation of Distortion (RED)

- Given a received speech signal, how does one determine if declipping (BAR) or decompression (BAN) need to be performed?



Clipped sample estimation

(Harvilla & Stern; ICASSP 2015)

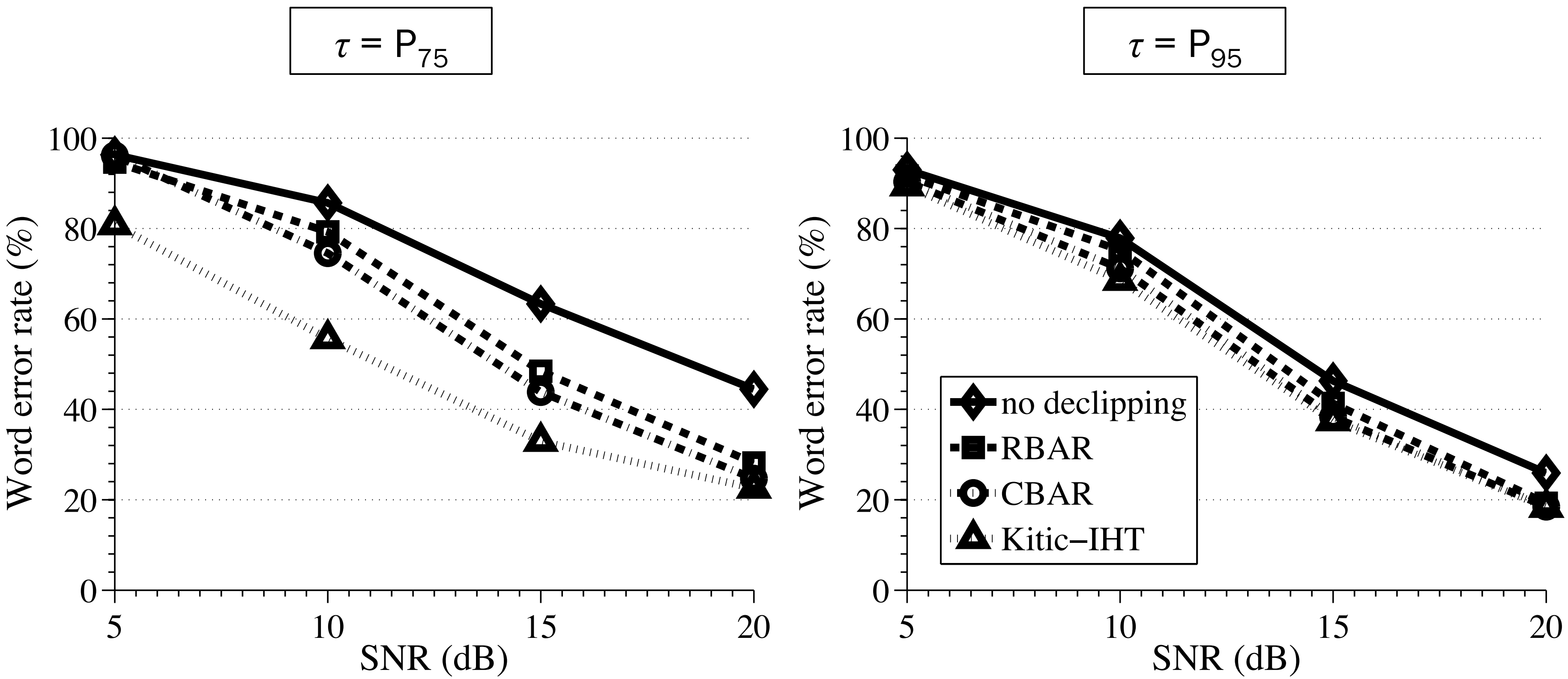


Clipped sample estimation

(Harvilla & Stern; ICASSP 2015)

- Experiment 2 (additive noise):

The location of all clipped samples is assumed known.

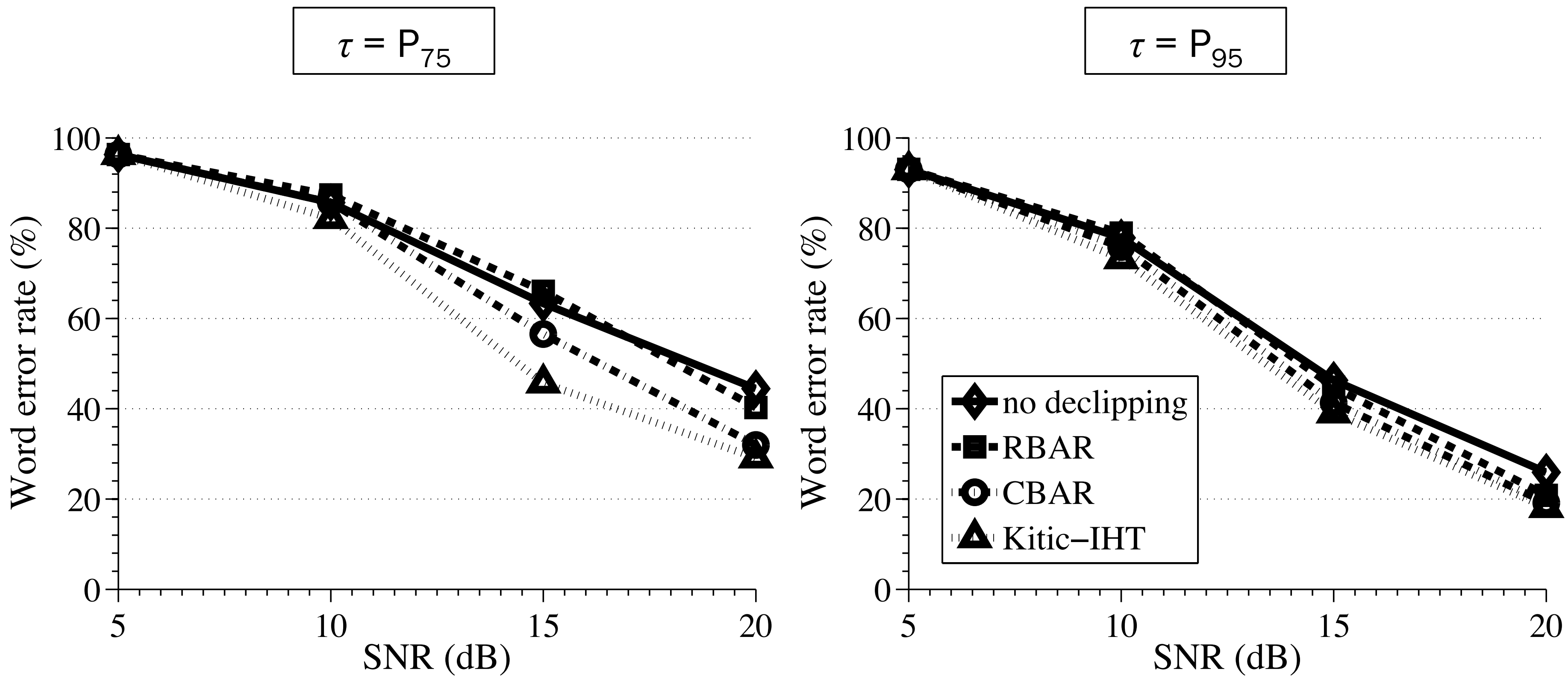


Clipped sample estimation

(Harvilla & Stern; ICASSP 2015)

- Experiment 2 (additive noise):

Clipping occurrence and location is detected using RED techniques

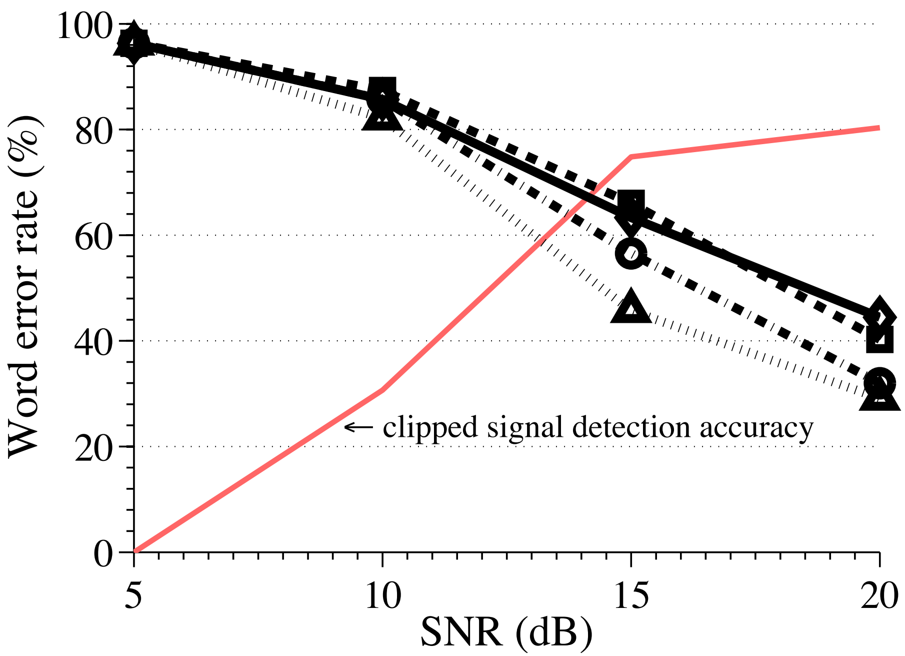


Clipped sample estimation
(Harvilla & Stern; ICASSP 2015)

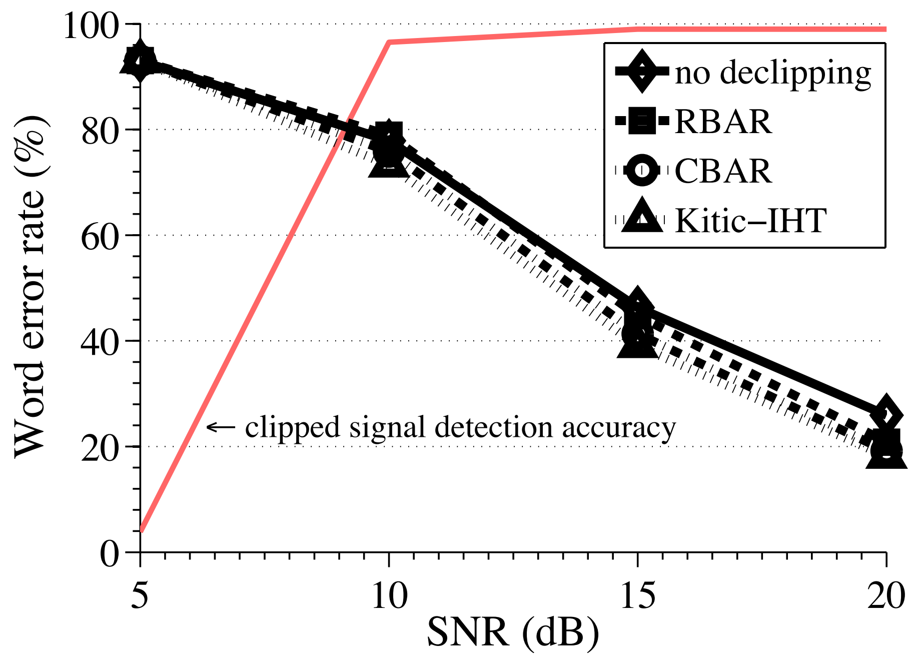
- Experiment 2 (additive noise):

Clipping occurrence and location is detected using RED techniques

τ = P_{75}

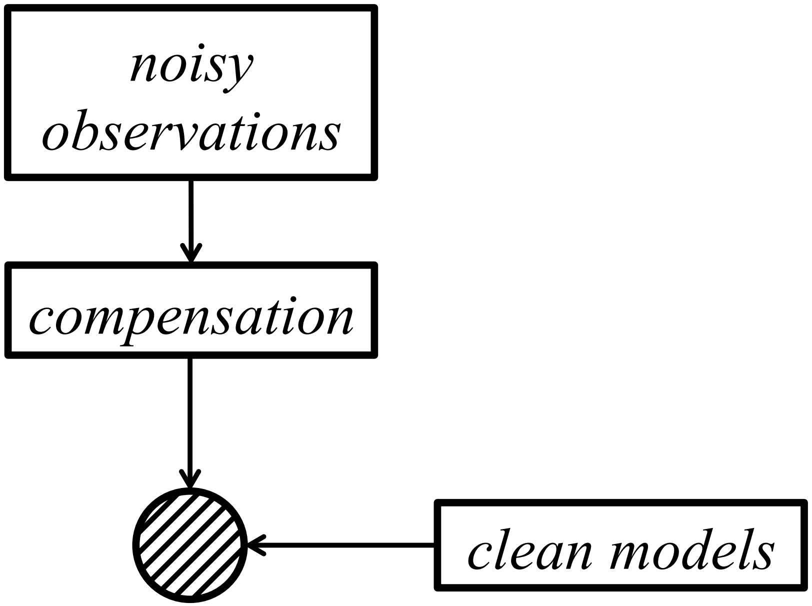


τ = P_{95}



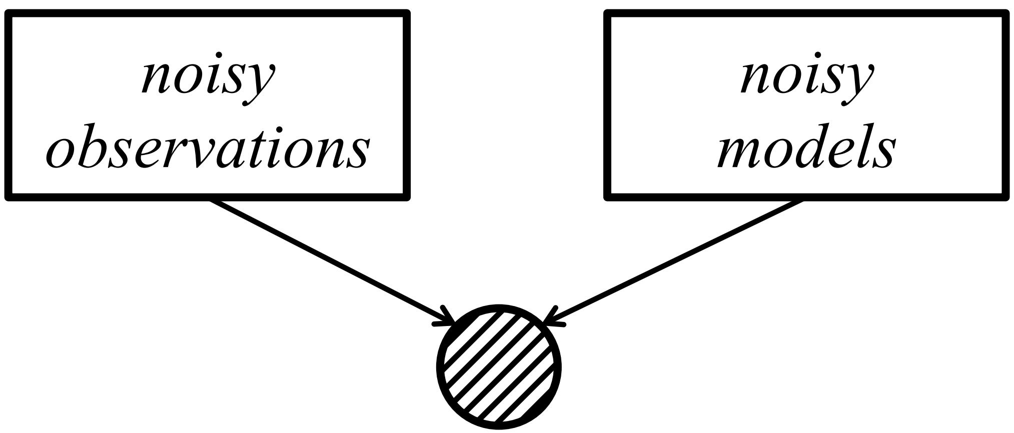
Artificially-Matched Training (AMT)

- So far, the developed techniques have sought to repair clipped, compressed and noisy speech to "look like" clean speech:



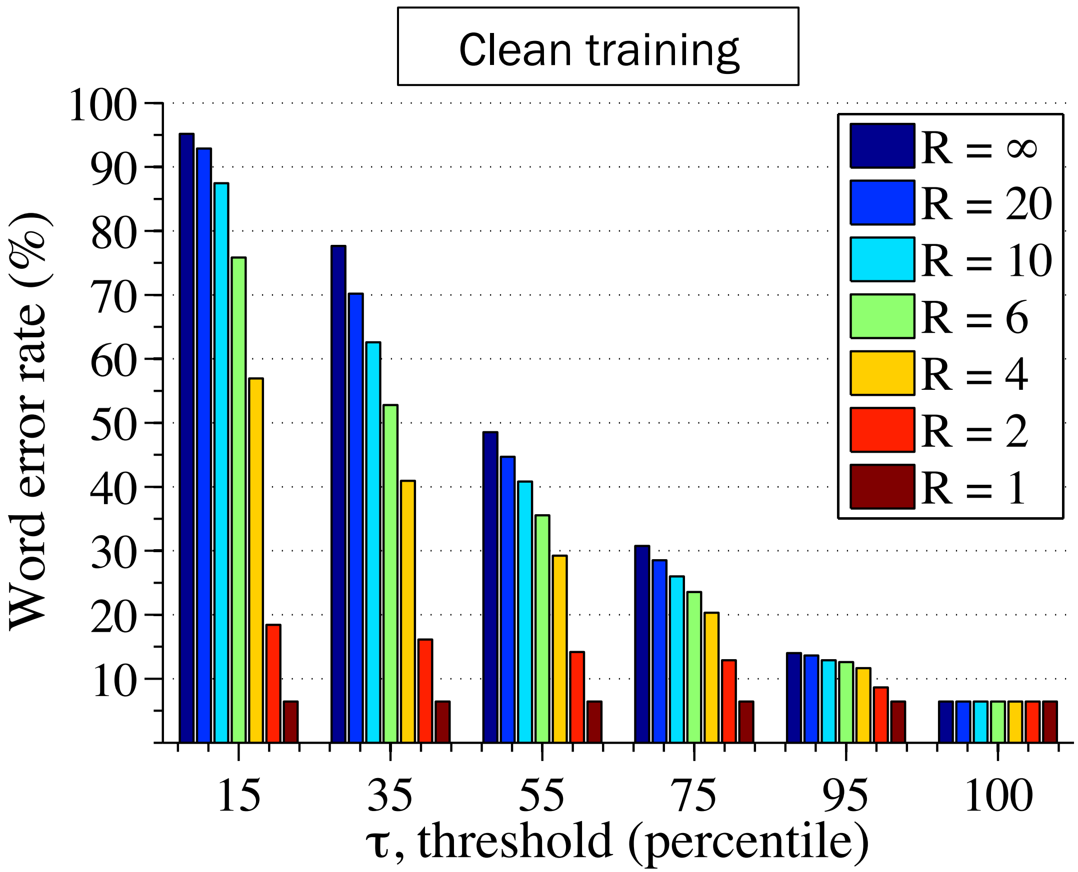
Artificially-Matched Training (AMT)

- Ultimately, it's only important for the Acoustic Model and testing data conditions to match. They both need not be "clean."



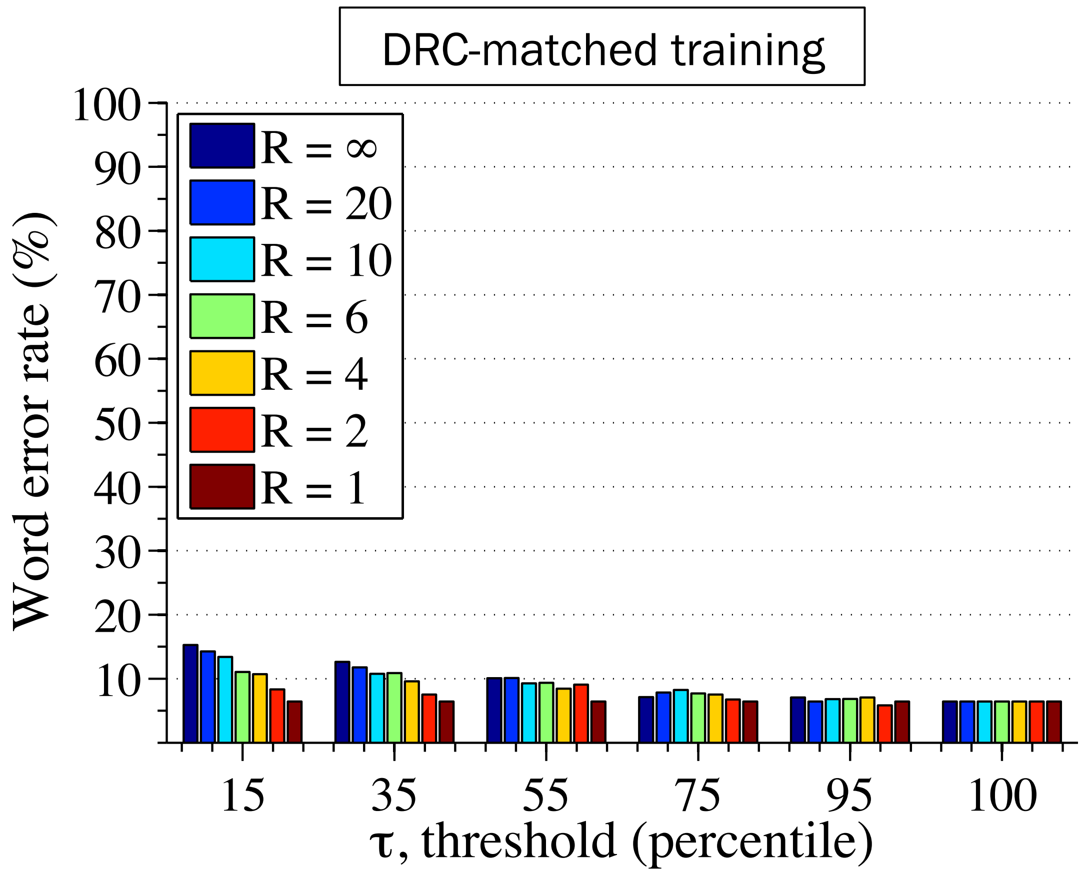
Artificially-Matched Training (AMT)

- Experiment 1 (no additive noise):



Artificially-Matched Training (AMT)

- Experiment 1 (no additive noise):



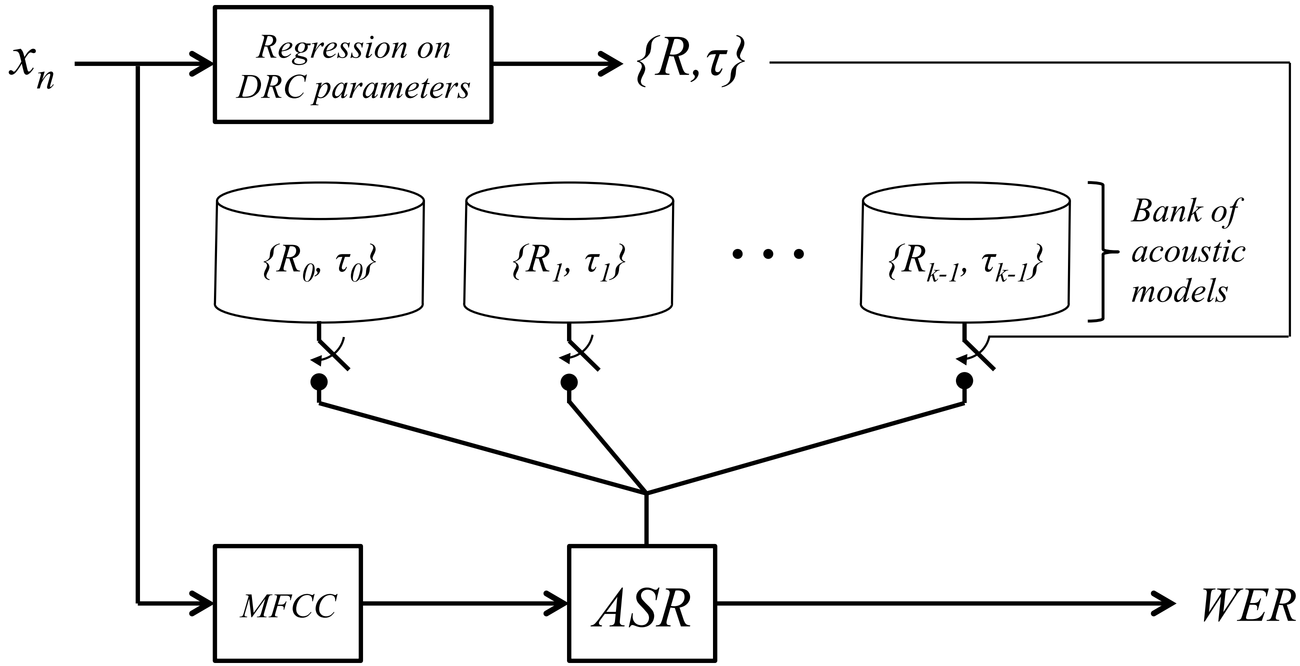
Artificially-Matched Training (AMT)

- One approach to achieving this in practice:

Current implementation uses the following parameter sets:

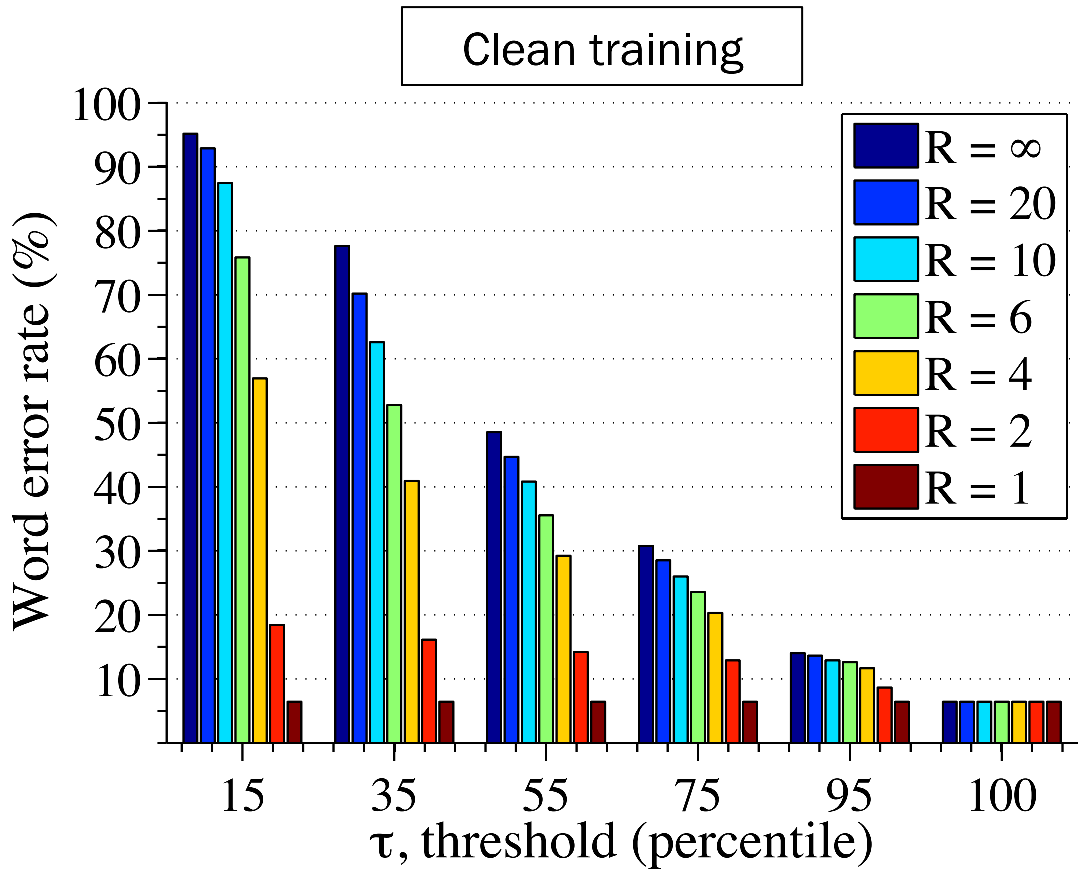
R = {∞}
τ = {P_{15}, P_{35}, P_{55}, P_{75}, P_{95}}

Artificially-Matched Training with Acoustic Model Selection (AMT-AMS)



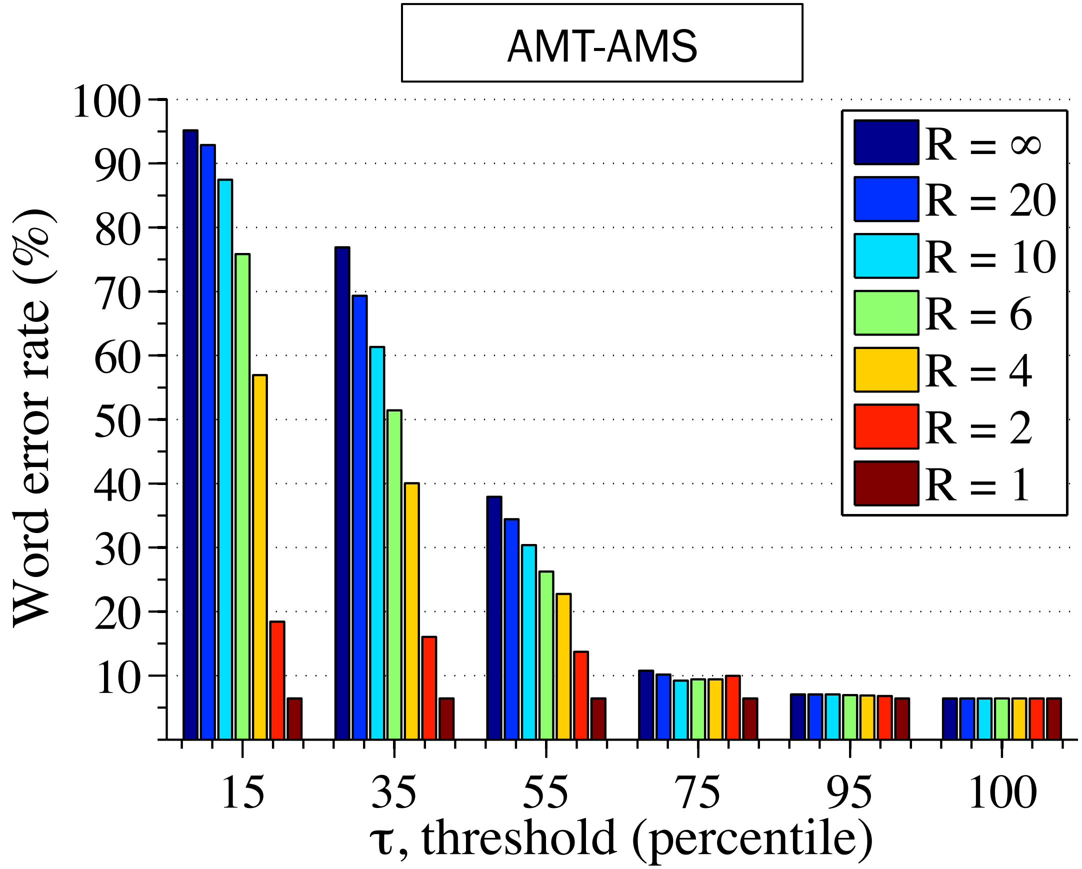
Artificially-Matched Training (AMT)

- Experiment 1 (no additive noise):



Artificially-Matched Training (AMT)

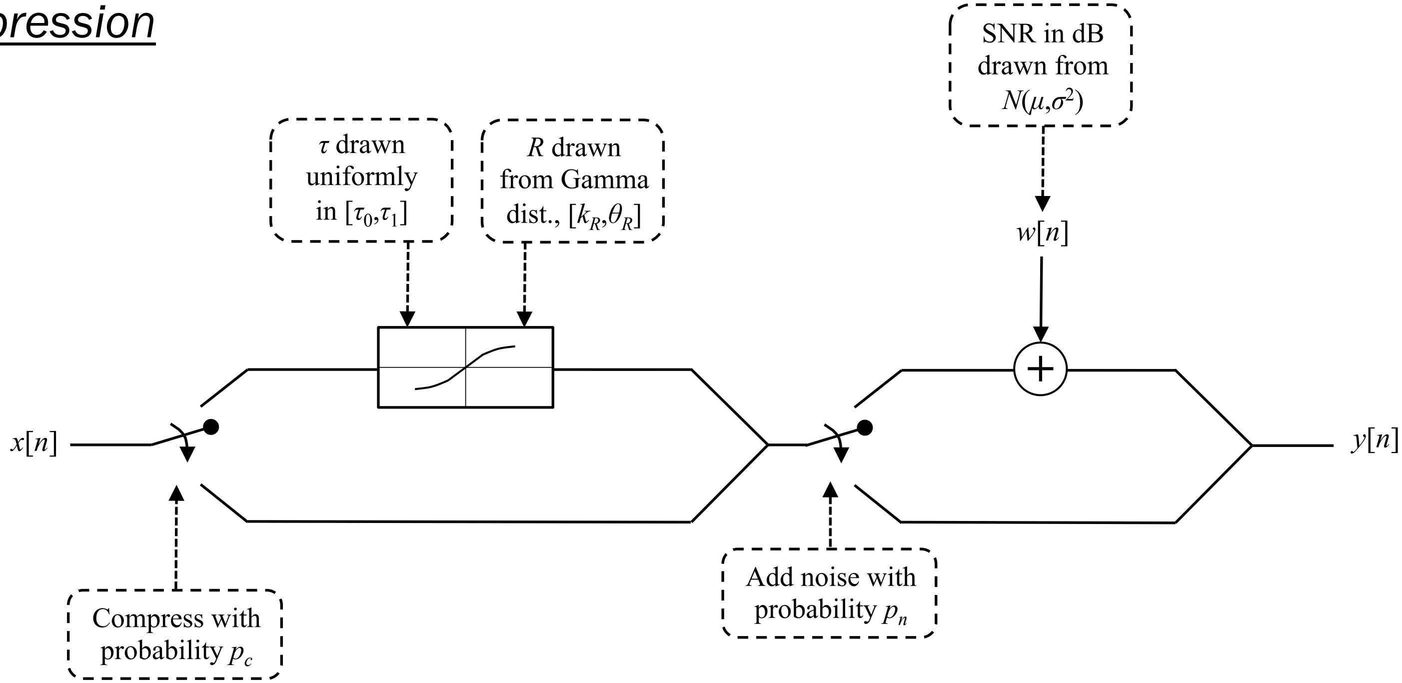
- Experiment 1 (no additive noise):



The Big Picture

- With no knowledge of the noise conditions and characteristics of the incoming speech, how well does the combination of algorithms from the thesis work in practice?

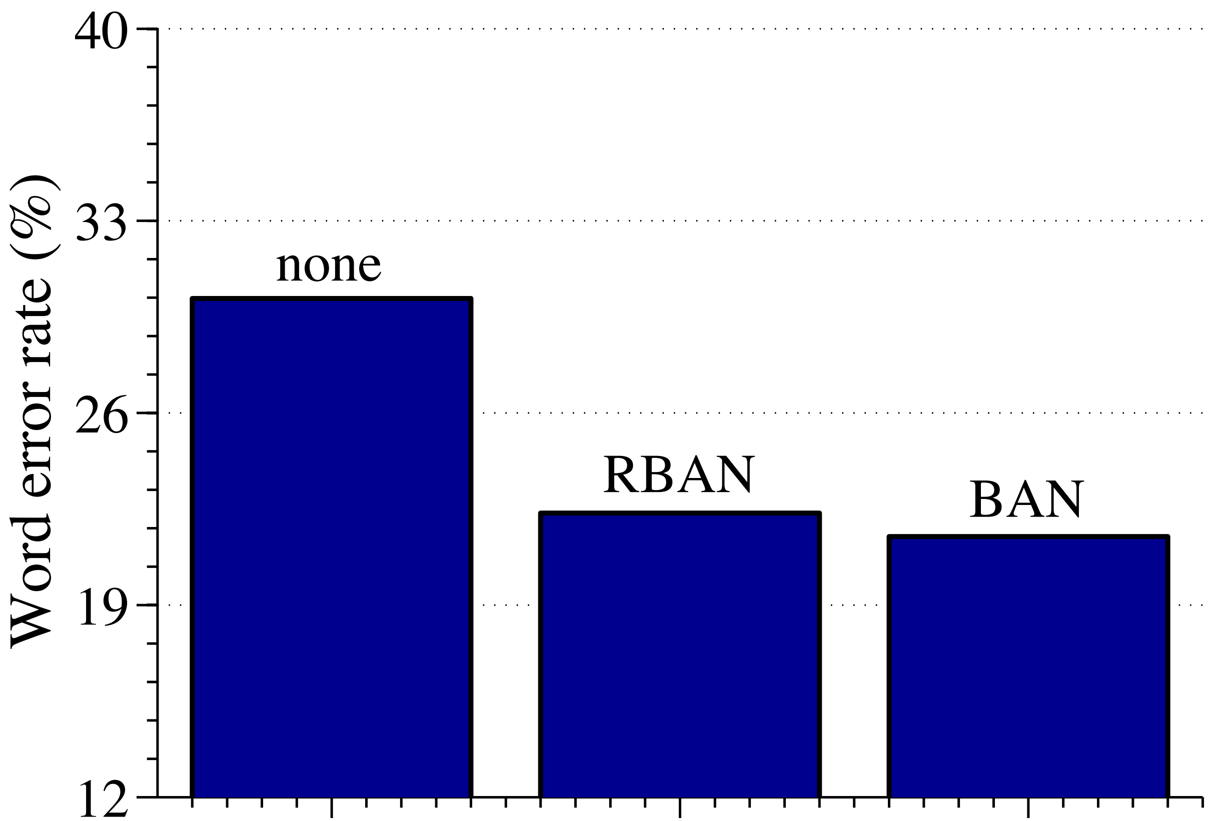
<u>Compression</u>



$p_c = 0.9$
$t_0 = 60$
$t_1 = 98$
$p_n = 0.75$
$\mu = 20$
$\sigma^2 = 25$
$k = 3$
$\theta = 2$

The Big Picture

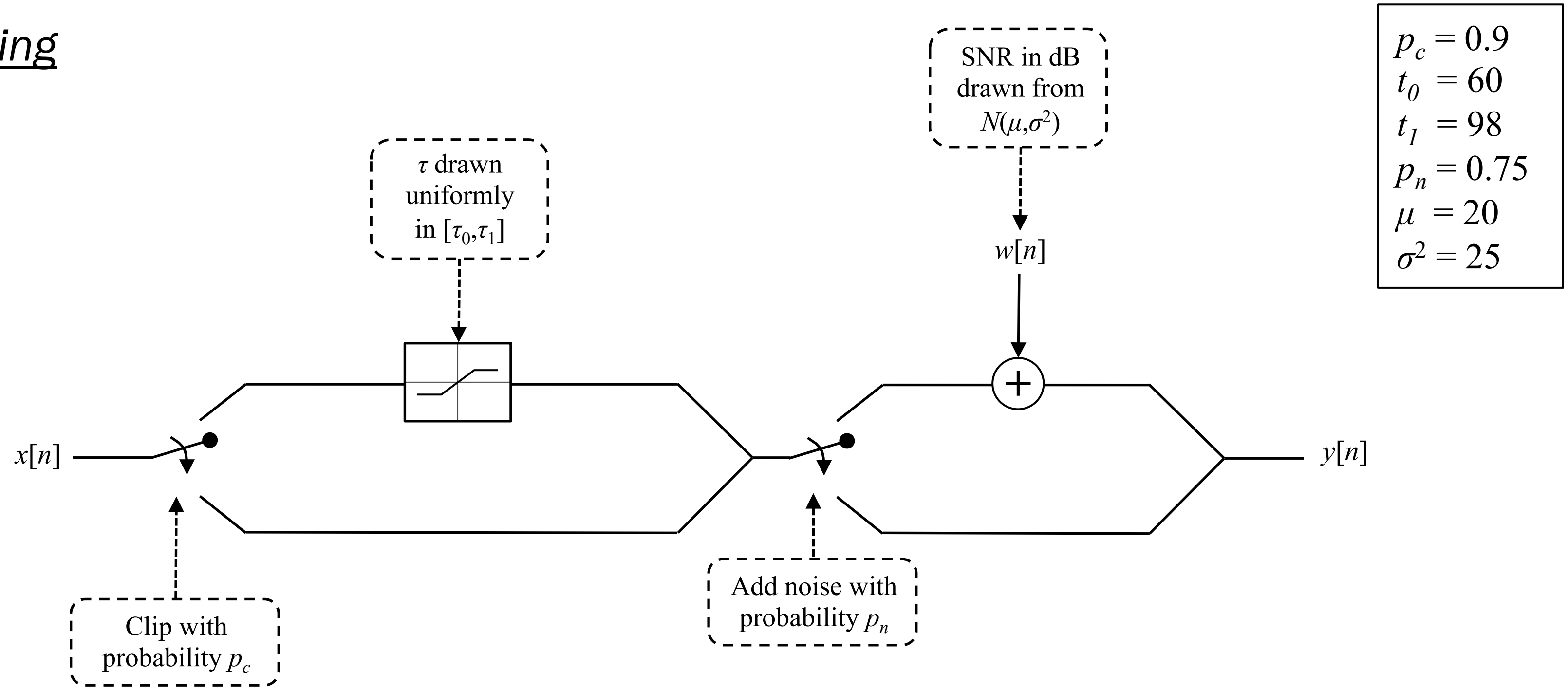
Compression



The Big Picture

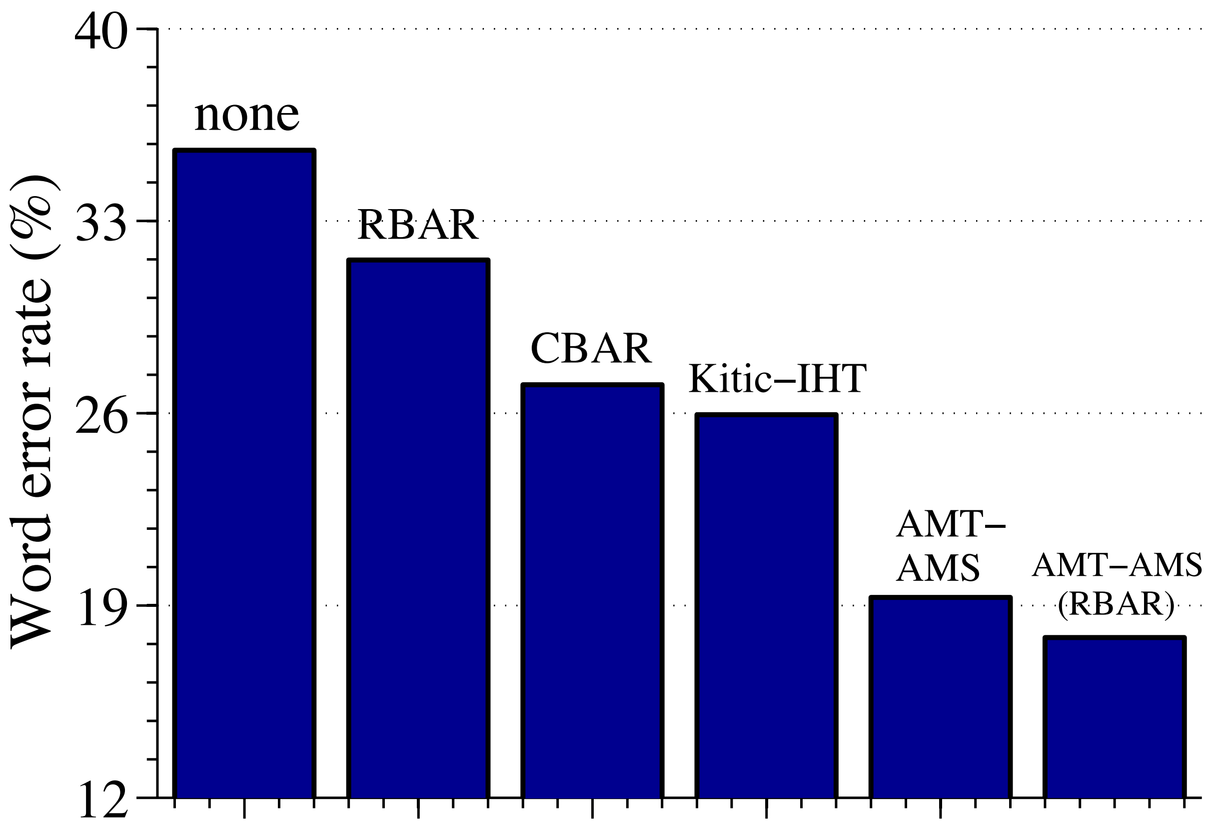
- With no knowledge of the noise conditions and characteristics of the incoming speech, how well does the combination of algorithms from the thesis work in practice?

Clipping



The Big Picture

Clipping



Summary & Conclusions

- A previously-unexplored problem in speech recognition, DRC, was introduced.
- Novel solutions to the two primary aspects of the problem, clipping and compression, were developed.
- Techniques for detecting the occurrence of DRC were considered.
- A comprehensive solution to DRC for speech recognition was proposed.
- DRC, especially in noise, is a very hard problem, but this thesis lays the groundwork for very promising future research.

Summary & Conclusions

- Areas of future research include:
  - Improving target amplitude estimates for RBAR [BAR]
  - Improving the robustness of BAR methods to additive noise [BAR]
  - Improving the robustness of clipped/compressed signal detection to low-valued SNR and τ [RED, Big Picture]
  - Development of an *R*-estimation algorithm [RED, Big Picture]
  - Further investigation of the performance of AMT-AMS with an increasing granularity of acoustic model references [AMT]

Thank you!

- Questions?