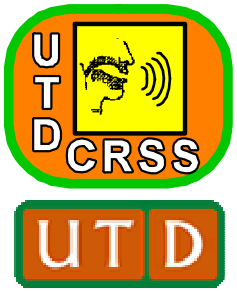
PHONEME SELECTIVE SPEECH ENHANCEMENT USING THE GENERALIZED PARAMETRIC SPECTRAL SUBTRACTION ESTIMATOR

Amit Das and John H.L. Hansen



Center for Robust Speech Systems (CRSS)
Erik Jonsson School of Engineering & Computer Science
Department of Electrical Engineering
University of Texas at Dallas
Richardson, Texas 75083-0688, U.S.A.

ICASSP 2011

May 22-27, 2011 Prague, Czech Republic

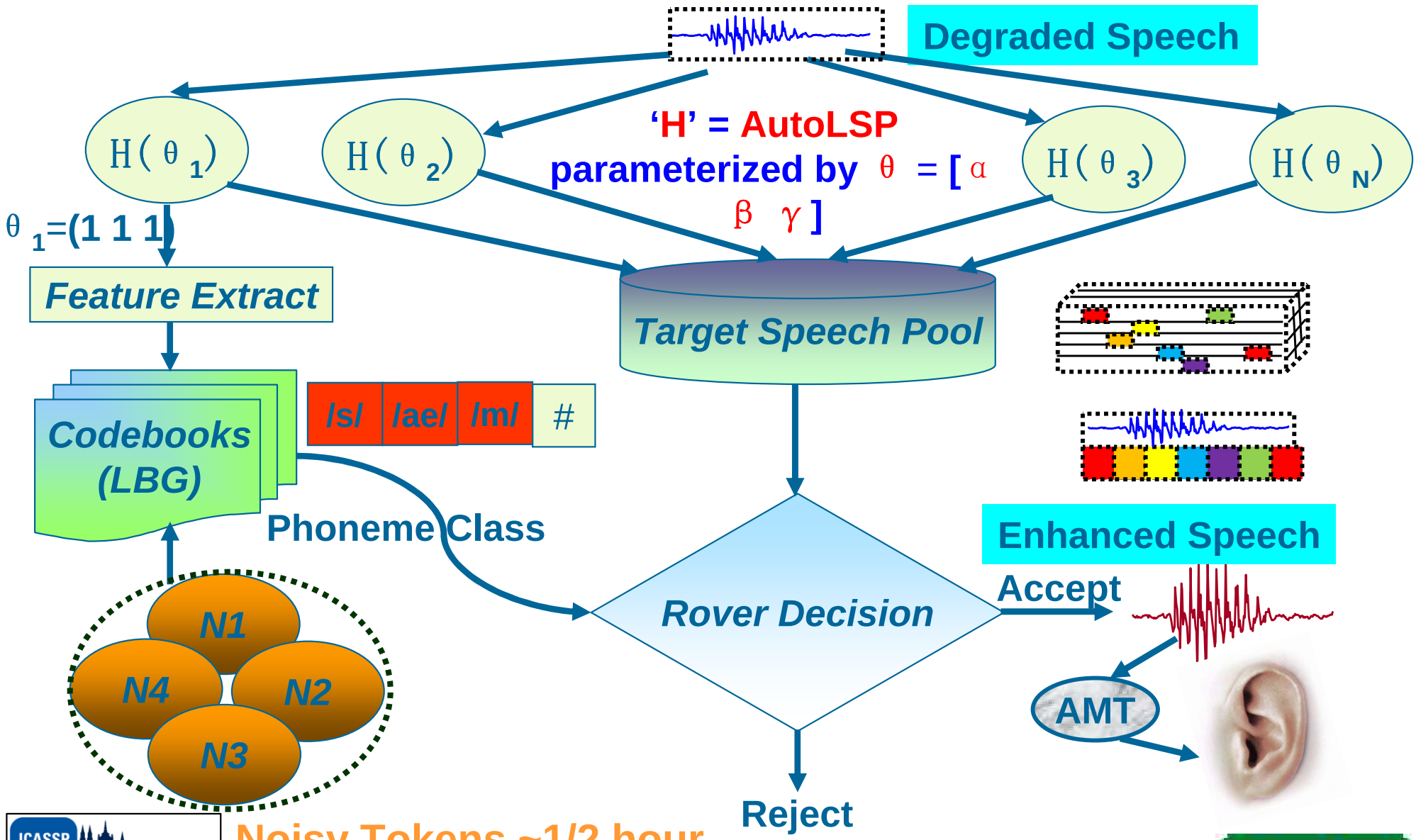
Introduction

- Degree of noise impact is ***not uniform*** across the speech spectrum or phoneme sequence.
- Regions of low SNR in the spectrum more adversely affected than high SNR.
- SNR variation across the spectrum depends on both: ***phoneme class*** and ***noise characteristics***.
- Require ***phoneme class selective enhancement*** algorithm.

Introduction

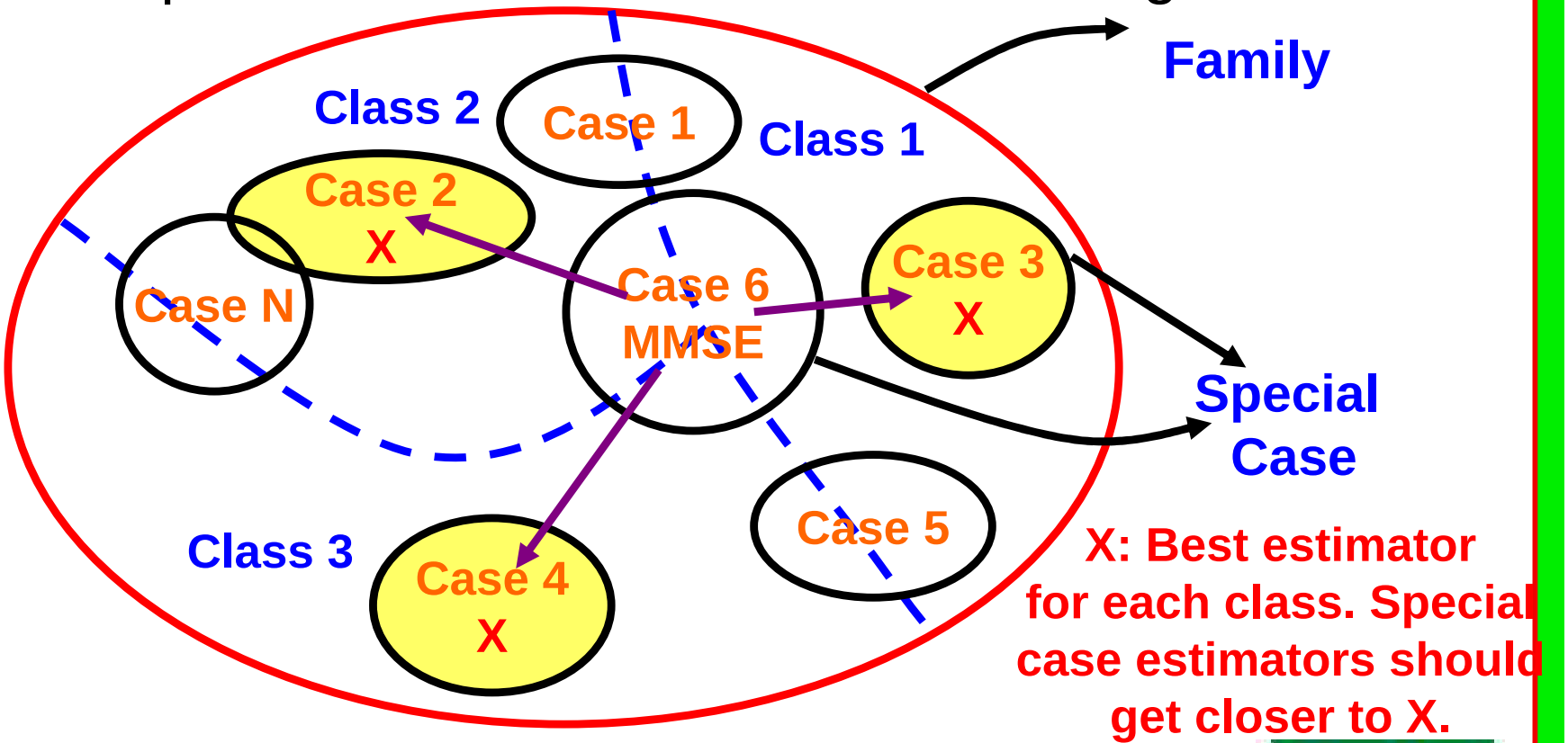
- **Each phoneme class is distorted differently for each noise type**
- **Based on Frequency content, articulatory structure, influence of noise for that phoneme, and stationarity of noise**
- **Speech Class type adaptation for enhancement:**
  - McAulay & Malpass (IEEE Trans. ASSP 1980) – soft-decision noise suppression
  - Hansen & Arslan (IEEE Trans. SAP 1995) – HMM based phone class partitioning for AutoLSP
  - Das & Hansen, (Interspeech-07) - Class Constrained ROVER Based Speech Enhancement

ROVER – System Description



Problem?

- Gain functions of current algorithms (e.g. MMSE) rely on Apriori/Aposteriori SNR. Gains are limited.
- Can we do better? Could generate a family of gain functions & provide more enhancement coverage.



Cost Function: MSE vs WED

- Modify Error Function to place more weight on spectral valleys (low SNR region) than peaks (high SNR region).

- Hence, new error function is ***Weighted Euclidean distortion (WED)***, can account for perceptual criterion be

$$C_\epsilon(\vec{X^\alpha}, \hat{\vec{X}^\alpha}) = (\vec{X^\alpha} - \hat{\vec{X}^\alpha})^T W (\vec{X^\alpha} - \hat{\vec{X}^\alpha}),$$

$$\text{where, } W = \text{diag}(X_1^\beta, X_2^\beta, ..., X_K^\beta), \text{K} = \text{FFT size.}$$

- Emphasize errors during valleys:

$$\alpha > 0, \beta < 0$$

$$\text{If } X < 1 \rightarrow C_\epsilon = \frac{(X^\alpha - \hat{X}^\alpha)^2}{X^\beta} \Uparrow$$

- Emphasize errors during peaks:

$$\alpha > 0, \beta > 0$$

$$\text{If } X > 1 \rightarrow C_\epsilon = (X^\alpha - \hat{X}^\alpha)^2 X^\beta \Uparrow$$

Background

- Generalized Spectral Subtraction (GSS) (Sim, et al [1]) model:

$$\hat{X}_k^\alpha = a_k Z_k^\alpha - b_k E[D_k^\alpha]$$

$\hat{X}_k =$ magnitude of clean speech estimate

$Z_k =$ magnitude of noisy speech

$D_k =$ magnitude of noise

$a_k, b_k =$ frequency dependent weighting coefficients

$\alpha =$ spectrum exponent

- Optimize a_k, b_k to minimize MSE: $(\hat{X}_k^\alpha - X_k^\alpha)^2$
- GSS-U (Unconstrained) model: $a_k \neq b_k$
- GSS-C (Constrained) model: $a_k = b_k$

[1] B. L. Sim, Y. C. Tong, J. S. Chang, and C. T. Tan, "A parametric formulation of the generalized spectral subtraction method," IEEE Trans. Speech Audio Process., vol. 6, pp. 328–337, July 1998.

Parametric Estimator

- Propose optimizing a_k, b_k to minimize WED:

$$X^{\beta}(\hat{X}_k^{\alpha} - X_k^{\alpha})^2$$

- WED optimized coefficients:

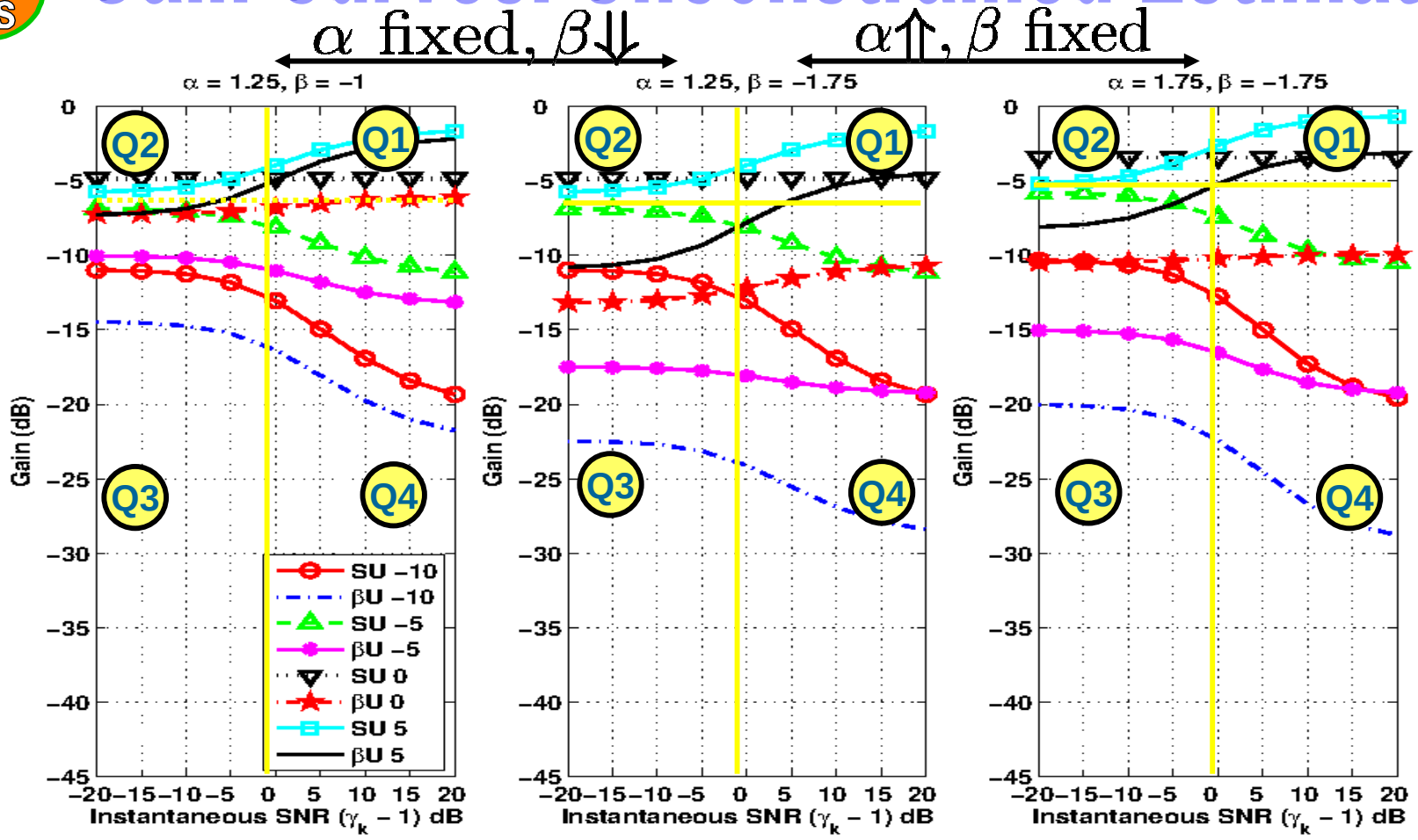
$$a_k = \frac{E[X_k^{\beta+2\alpha}]E[X_k^{\beta}] - E^2[X_k^{\beta+\alpha}]}{E[X_k^{\beta+2\alpha}]E[X_k^{\beta}] - E^2[X_k^{\beta+\alpha}] + E^2[X_k^{\beta}]\left(E[D_k^{2\alpha}] - E^2[D_k^{\alpha}]\right)}$$

$$b_k = a_k - (1 - a_k)\frac{E[X_k^{\beta+\alpha}]}{E[X_k^{\beta}]E[D_k^{\alpha}]}$$

- Substitute these coefficients in the GSS model to form parametric estimators. Offer better flexibility.

- MSE optimized coefficients = Special case of WED optimized coefficients when $\beta = 0$.

Gain Curves: Unconstrained Estimator

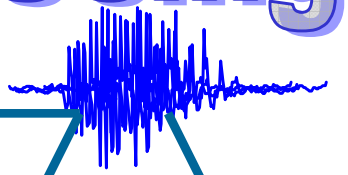


- Q1: speech region,
- Q2: unlikely region,
- Q3: noise-only region,
- Q4: musical noise region,

$$\xi_k > -5\text{dB}, \gamma_k > 0\text{dB}$$
$$\xi_k > -5\text{dB}, \gamma_k < 0\text{dB}$$
$$\xi_k < -5\text{dB}, \gamma_k < 0\text{dB}$$
$$\xi_k < -5\text{dB}, \gamma_k > 0\text{dB}$$

- SU:Sim et al. (Baseline) GSS-Unconstrained
- βU: Proposed GSS-beta Unconstrained

ROVER Using MIXMAX



Degraded Speech

$G(\theta_3)$

$G(\theta_2)$

$G(\theta_1)$

VAD: Noise-Only?

Yes

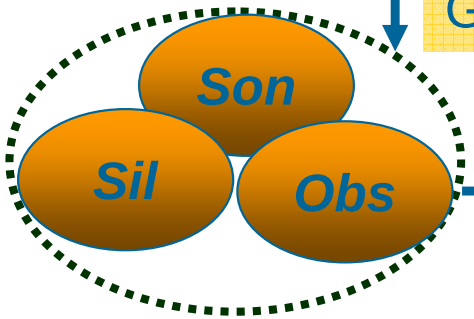
No

Tuned estimators
$G(\theta_1)$: Good for Sonorants
$G(\theta_2)$: Good for Obstruents
$G(\theta_3)$: Good for Silence
Note: $\theta_i = (\alpha_i, \beta_i)$

MFCC

(use VAD) Update Sil GMM if noise-only

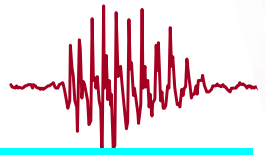
Son/Obs: Proceed for MIXMAX evaluation



Reference GMMs

MIXMAX: max(Son,Obs,Noise)

MFCC to Mag. Spec



Enhanced Speech

ROVER Using MIXMAX

- Three Broad Phoneme Class (BPC) types:
  - Sonorants (vowels, nasals, semivowels).
  - Obstruents (fricatives, affricates, stops).
  - Silence.
- Parametric beta estimators (GSS-BU, GSS-BC) may be tuned to adapt to each BPC.
  - Sonorant estimator → Best enhances sonorants **only**.
  - Obstruent estimator → Best enhances obstruents **only**.
- Outputs from each estimator converted to MFCC. Decide weights of each estimator at each frame. Soft combine weights and generate composite utterance. Mechanism is similar to ROVER (Das & Hansen, Interspeech-07).
- Noisy speech can be modeled by the MIXMAX (mixture maximum) model and can be used to classify sonorants/obstruents.

MIXMAX Model [2]

- GMMs for sonorants (X), obstruents (Y):

  MFCCs, N mixtures, K components:

$$f(\mathbf{x}) = \sum_i c_x(i) \prod_k \mathcal{N}(\mu_x(i,k), \sigma_x(i,k))$$

$$t(\mathbf{y}) = \sum_j c_y(j) \prod_k \mathcal{N}(\mu_y(j,k), \sigma_y(j,k))$$

- GMM for silence (D): MFCC,1 mixture, K components:

$$g(\mathbf{d}) = \prod_k \mathcal{N}(\mu_d(k), \sigma_d(k))$$

- MIXMAX model for noisy speech (Z):

$$\mathbf{Z} \approx \max(\mathbf{X}, \mathbf{Y}, \mathbf{D})$$

[2] A. Nadas, D. Nahamoo, M.A. Picheny, ``Speech recognition using noise-adaptive prototype," IEEE Trans. Speech & Audio Proc., 37(10):1495-1505, Oct. 1989.

Experimental Set-up

- Results are based on 32 tokens from TIMIT test corpus
- Metrics: Segmental-SNR, Itakura-Saito (IS) Distortion
- GMMs trained from 300 tokens, # mixtures = 16 (sonorants, obstruents), 1 (silence); 39-dim MFCC based GMMs.
- Noise Types: Flat communications channel noise (FLN, mostly stationary), large crowd noise (LCR, mostly non-stationary).
- Acronyms used in figures (next slides):

Baseline	Parametric	ROVER
MMSE (Ephraim-Malah) [3]	**WC (WED Chi) [5]**	**RWC (ROVER WED Chi) [5]**
JMAP (Wolfe-Godsil Joint-MAP) [4]	**JC (JMAP Chi) [5]**	**RJC (ROVER JMAP Chi) [5]**
GU (Sim et al., GSS Unconstrained)	**GBU (GSS-Beta unconstrained)**	**RGBU (ROVER GBU)**
GC (Sim et al., GSS Constrained)	**GBC (GSS-Beta Constrained)**	**RGBC (ROVER GBC)**

[3] Y. Ephraim, D. Malah, "Speech enhancement using a minimum mean-square error short-time amplitude estimator," IEEE Trans. ASSP, 32(6):443-445, Dec 1984.
[4] P. J. Wolfe, S. J. Godsill, "Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement," *EURASIP J. App. Sig. Process.*, vol. 10.
[5] A. Das. J.H.L. Hansen, "Broad phoneme class based speech enhancement using the Mixture Maximum Model," ICASSP-10, 4762-4765, March 2010.

SegSNR Increase , FLN (0dB SNR)

Baseline(GU) vs Parametric (GBU) vs ROVER (RGBU)

	Enhancement$\langle\alpha,\beta\rangle$	Rise in SegSNR (dB)			
		Son	Obs	Sil	Ovl
Parametric	GU$\langle 1.00, 0.00\rangle$	4.39	3.36	0.20	3.85
	GBU$_S\langle 1.50, 1.00\rangle$	4.85	5.25	0.88	4.79
	GBU$_O\langle 1.00, -1.75\rangle$	3.94	7.77	4.23	5.41
	GBU$_N\langle 2.00, -1.75\rangle$	-2.70	7.12	7.45	1.45
	RGBU	5.18	7.89	4.97	6.58

Green: Good

Yellow: Reasonable

Red: Poor

Goal: More greens and/or fewer reds per row

GU : Sim et al. Baseline GSS-U estimator

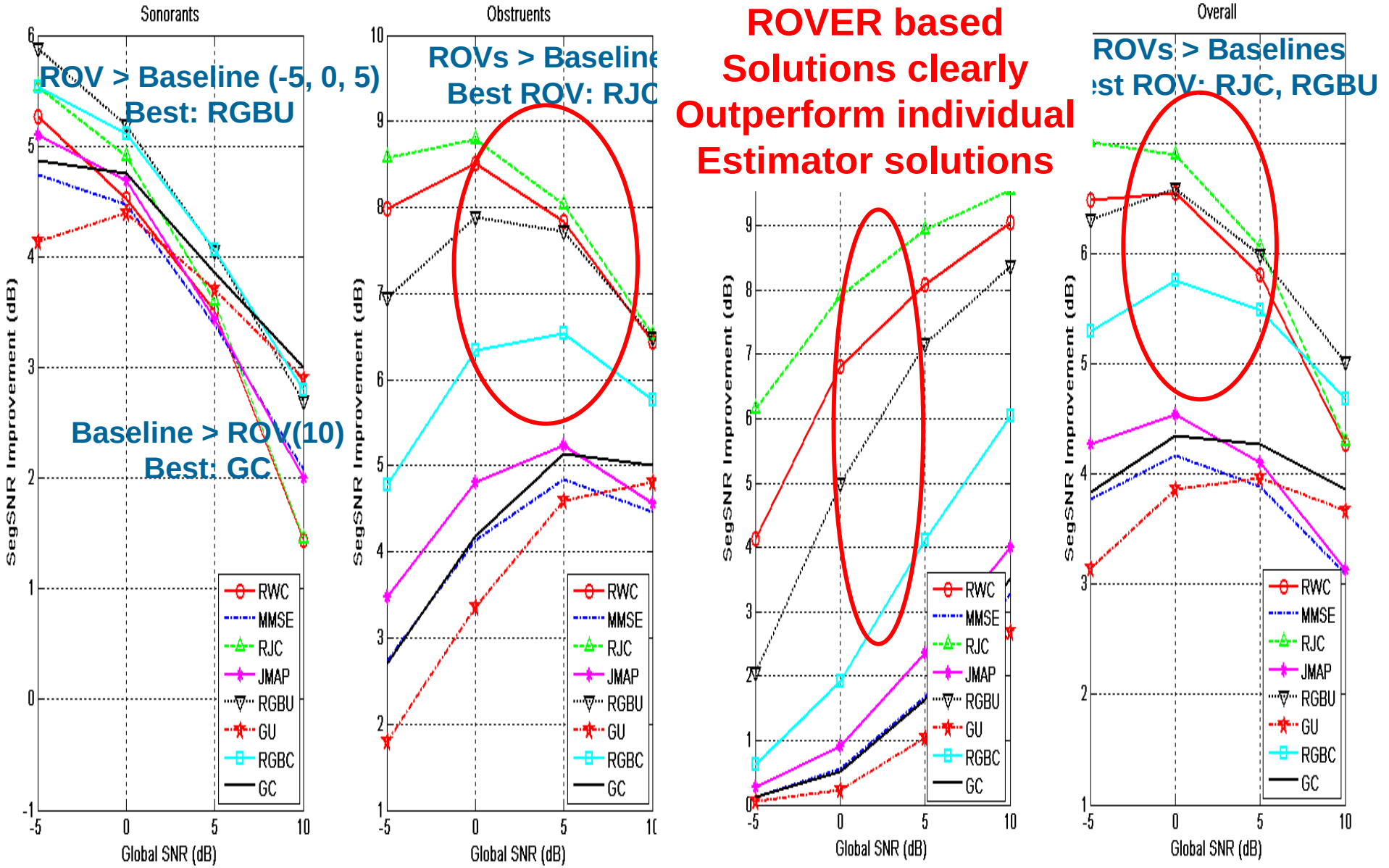
GBU_S : GSS-BU estimator with best configurable parameters for sonorants

GBU_O : GSS-BU estimator with best configurable parameters for obstruents

GBU_N : GSS-BU estimator with best configurable parameters for silence

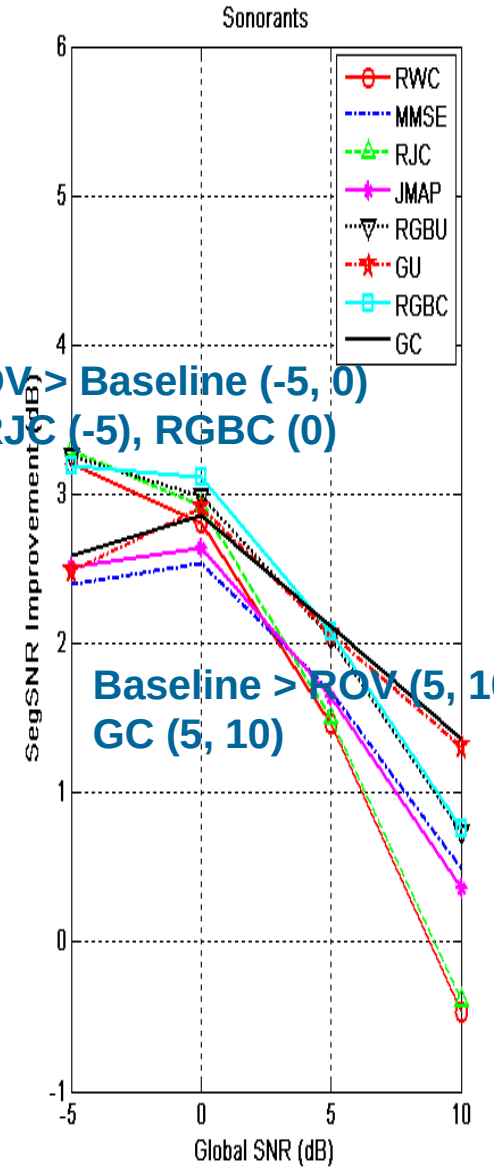
$RGBU$: ROVER GSS-BU estimator

SegSNR Increase: FLN Noise (-5 to 10 dB)



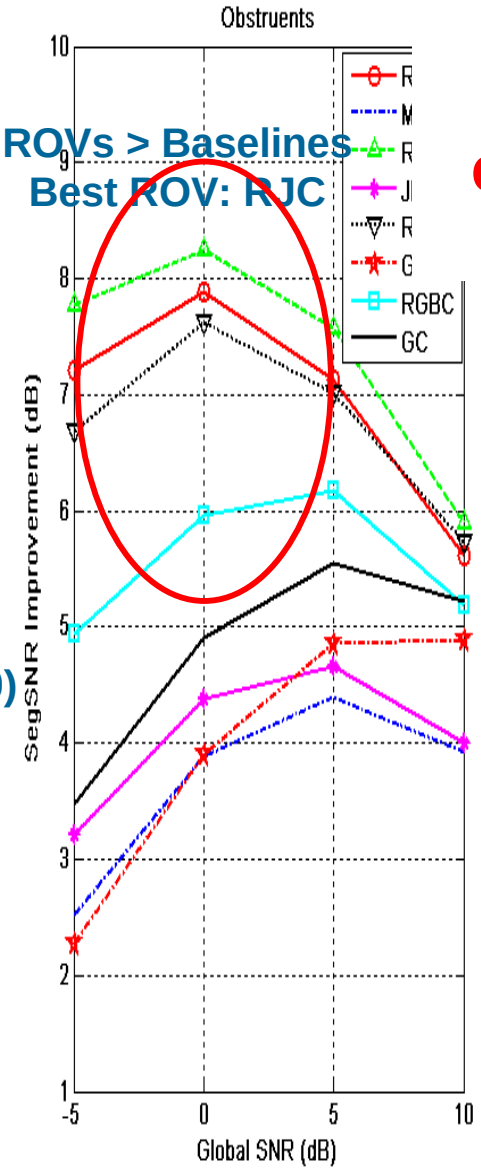
SegSNR Increase = SegSNR Enhanced Speech – SegSNR Noisy Speech

SegSNR Increase: LCR Noise (-5 to 10 dB)



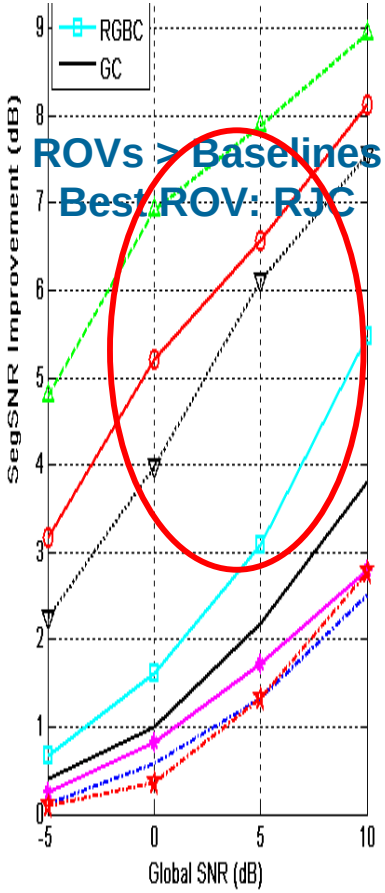
ROV> Baseline (-5, 0)
RJC (-5), RGBC (0)

Baseline > ROV (5, 10)
GC (5, 10)

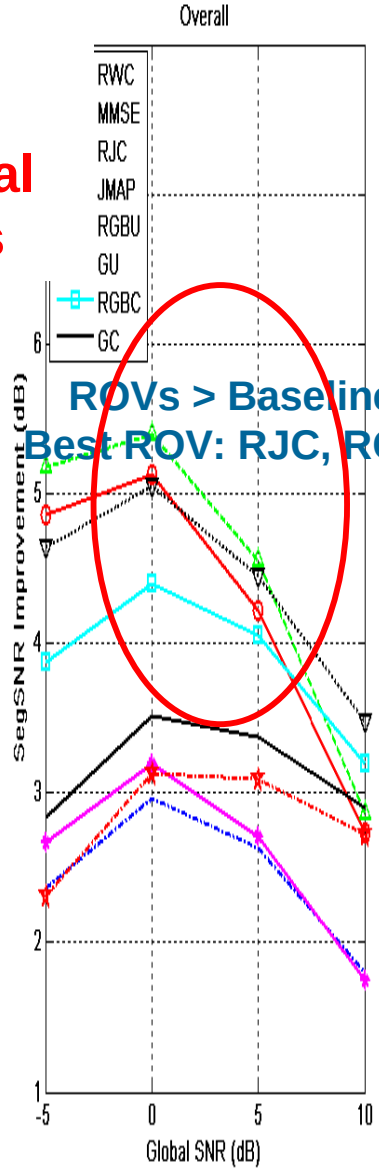


ROVs > Baselines
Best ROV: RJC

ROVER based Solutions clearly Outperform individual Estimator solutions



ROVs > Baselines
Best ROV: RJC



ROVs > Baselines
Best ROV: RJC, RGBU

SegSNR Increase = SegSNR Enhanced Speech – SegSNR Noisy Speech

Top Two Algorithms

FLN Noise

BPC	SegSNR				IS Distortion			
	-5 dB	0 dB	5 dB	10 dB	-5 dB	0 dB	5 dB	10 dB
Son ♣	RGBU	RGBU	RGBU	GC	RGBU	RGBU	RGBU	RGBU
Son ♠	RGBC	RGBC	RGBC	GU	RJC	RGBC	RGBC	RGBC
Obs ♣	RJC	RJC	RJC	RJC	RGBU	RGBU	RGBU	RGBU
Obs ♠	RWC	RWC	RWC	RWC	RWC	RWC	RWC	RWC
Sil ♣	RJC	RJC	RJC	RJC	RGBU	RGBU	RGBU	RGBU
Sil ♠	RWC	RWC	RWC	RWC	RJC	RJC	RJC	RWC
Ovl ♣	RJC	RJC	RJC	RGBU	RGBU	RGBU	RGBU	RGBU
Ovl ♠	RWC	RWC	RGBU	RGBC	RWC	RWC	RWC	RGBC

1st
2nd

LCR Noise

BPC	SegSNR				IS Distortion			
	-5 dB	0 dB	5 dB	10 dB	-5 dB	0 dB	5 dB	10 dB
Son ♣	RJC	RGBC	GC	GC	RGBC	RGBC	RGBC	RGBC
Son ♠	RGBC	RGBU	GU	GU	RGBU	RGBU	RGBU	RGBU
Obs ♣	RJC	RJC	RJC	RJC	RGBU	RGBU	RGBU	RGBU
Obs ♠	RWC	RWC	RWC	RGBU	RWC	RWC	RWC	RGBC
Sil ♣	RJC	RJC	RJC	RJC	RGBU	RGBU	RGBU	RGBU
Sil ♠	RWC	RWC	RWC	RWC	RJC	RWC	RWC	RJC
Ovl ♣	RJC	RJC	RJC	RGBU	RGBU	RGBU	RGBU	RGBU
Ovl ♠	RWC	RWC	RGBU	RGBC	RWC	RGBC	RGBC	RGBC

Conclusions

- Parametric GSS-βU, GSS-βC estimators.
- Parametric estimators can be pre-tuned per phoneme class but may not perform well across all classes.
- ROVER based paradigm to pick phoneme class segments from parametric pre-tuned estimators and form a single composite utterance.
- ROVER based estimators outperform baseline and parametric estimators across most combinations of FLN/LCR noise types and global SNRs.

References

[1] J. Deller, J.H.L. Hansen, J. Proakis, *Discrete Time Processing of Speech Signals*, Prentice-Hall Publishers, NY, 2000.

[2] J.H.L. Hansen, L. Arslan, "Markov model based phoneme class partitioning for improved constrained iterative speech enhancement", IEEE Trans. Speech & Aud. Proc., 3(1):98-104, Jan 1995.

[3] B. L. Sim, Y. C. Tong, J. S. Chang and C. T. Tan, "A parametric formulation of the generalized spectral subtraction method," IEEE Trans. Speech & Audio Proc., 6(4):328-337, July 1998.

[4] A. Das, J.H.L. Hansen, "Generalized parametric spectral subtraction using weighted Euclidean distortion," Interspeech-2008, pp. 399-402, Sept. 2008.

[5] A. Nádas, D. Nahamoo, M.A. Picheny, "Speech recognition using noise-adaptive prototype," IEEE Trans. Speech & Aud. Proc., 37(10):1495-1505, Oct. 1989.

[6] D. Burshtein, S. Gannot, "Speech enhancement using a mixture-maximum model," IEEE Trans. Speech & Aud. Proc., 10(6):341-351, Sept. 2002.

[7] A. Das, J.H.L. Hansen, "Broad phoneme class based speech enhancement using mixture maximum model," Proc. IEEE ICASSP, 4762-4765, Mar 2010.