Feature Extraction Combining Spectral Noise Reduction and Cepstral Histogram Equalization For Robust ASR

J.C. Segura, M.C. Benítez, A. de la Torre, A.J. Rubio

Signal Processing and Communications Group

University of Granada (SPAIN)

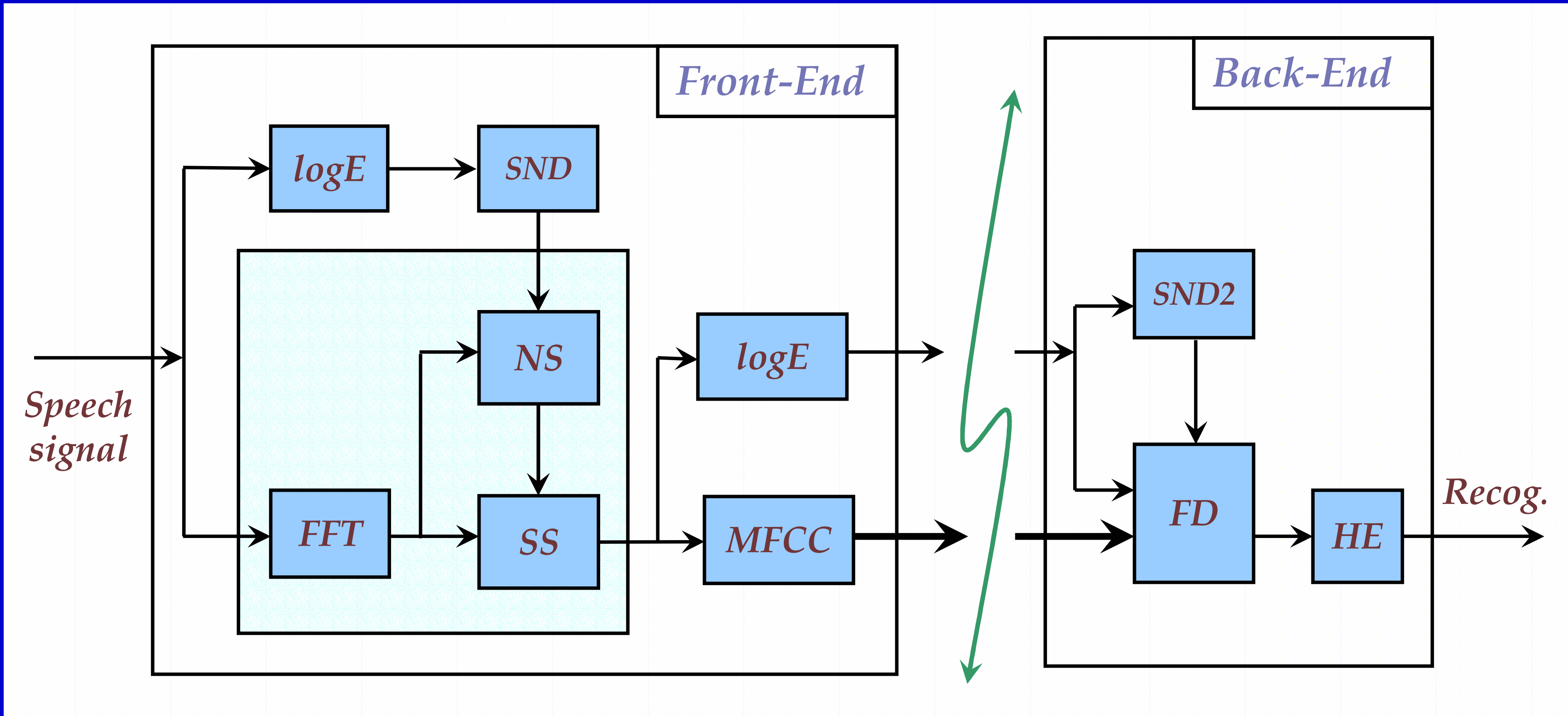
Introduction

- Results for Noisy TI-Digits at ICASSP'02
  - Histogram Equalization (HE) can reduce the mismatch of noisy speech better than CMS and CMVN
  - Its performance is increased when applied over partially compensated speech features
- In this work we explore HE performance in combination with Spectral Subtraction

Outline

- System description
- Front-End Spectral Noise Reduction
  - Speech/Non-Speech Detection
  - Spectral Subtraction
- Back-End Processing
  - Frame-Dropping
  - Feature Normalization
- Experimental set-up
- Results and discussion

System Description



Spectral Subtraction

- Standard implementation on the magnitude spectrum

$$\left|\hat{X}_t(w)\right| = \max\left\{ \left(\left|Y_t(w)\right| - \alpha \left|\hat{N}_t(w)\right| \right), \beta \left|Y_t(w)\right| \right\}$$

$$\left|\hat{N}_t(w)\right| = \begin{cases} \lambda \left|\hat{N}_{t-1}(w)\right| + (1-\lambda)\left|Y_t(w)\right| & \textit{Non-Speech} \\ \left|\hat{N}_{t-1}(w)\right| & \textit{Speech} \end{cases}$$

Over-subtraction	$\alpha = 1.1$	$\hat{N}(w)$:	*Noise estimate*
Maximum attenuation	$\beta = 0.3$	$Y(w)$:	*Noisy speech*
Forgetting factor	$\lambda = 0.95$	$\hat{X}(w)$:	*Clean speech estimate*

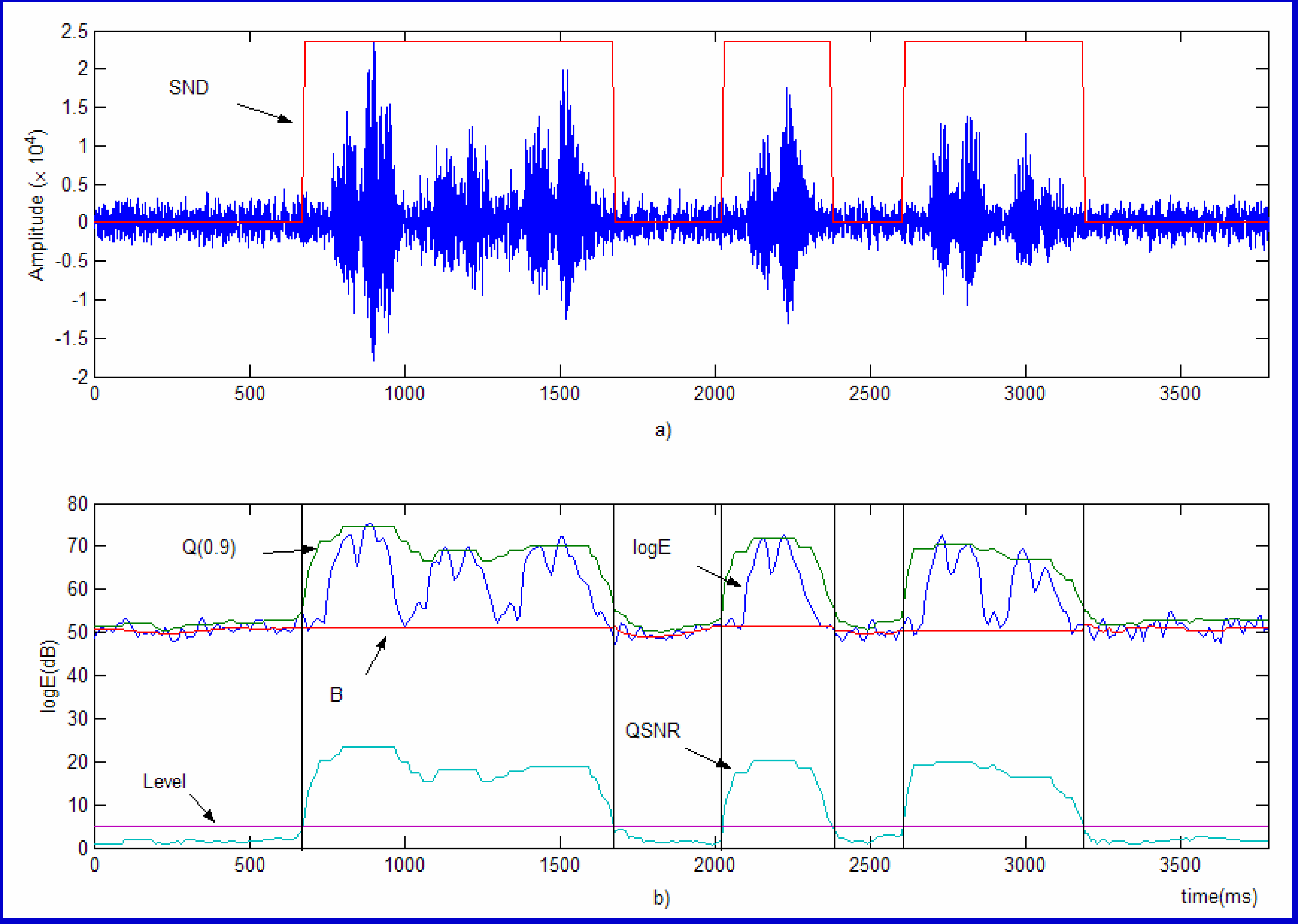
Speech/Non-Speech Detection (I)

- Based on log-Energy quantile difference
- Quantiles are estimated over a sliding window of 21 frames (at a frame rate of 100Hz)
  - $Q_{0.5}$ (median) is used to track the noise level B
  - $Q_{0.9}$ is used to track the speech level
- $Q_{SNR} = Q_{0.9} - B$ is thresholded to detect speech
- Noise level B is updated with $Q_{0.5}$ whenever non-speech is detected

Speech/Non-Speech Detection (II)

- Characteristics of the SND algorithm
  - Easy and fast implementation
  - Fast tracking of noise level
  - Q_{SNR} is smooth enough to prevent false speech detections
  - Implicit symmetric hang-over

Speech/Non-Speech Detection (III)



Frame-Dropping

- The objective is to remove long speech pauses
- Based on same SND algorithm
  - It works over the noise reduced speech
- One frame is removed only if in the middle of a non-speech segment of predefined length
  - This prevents over-dropping
  - 11 frames are used in this work

Feature Normalization (I)

- CDF-matching for non-linear distortion compensation
  - Given a zero-memory one-to-one general transformation $y=T[x]$

$$x \rightarrow p_X(x) \qquad\qquad y = T[x] \rightarrow p_Y(T[x]) = p_Y(y)$$

$$C_X(x) = \int_{-\infty}^{x} p_X(u)\,du \qquad\qquad C_Y(y) = \int_{-\infty}^{y} p_Y(u)\,du$$

$$C_X(x) = C_Y(y) \qquad \Rightarrow \qquad x = T^{-1}[y] = C_X^{-1}(C_Y(y))$$

Feature Normalization (II)

- Two ways of using CDF-matching for mismatch reduction

- CDF-matching for feature compensation
  - $C_X(x)$ is estimated during training
  - During test, $C_Y(y)$ estimate is used to compensate for the mismatch

$$\hat{x} = \hat{T}^{-1}[y] = C_X^{-1}(\hat{C}_Y(y))$$

- CDF-matching for feature normalization
  - A predefined $C_X(x)$ is selected (usually Gaussian)
  - For both training and test, features are transformed to match the reference distribution using an estimate of $C_Y(y)$
  - Can be viewed as an extension of CMVN

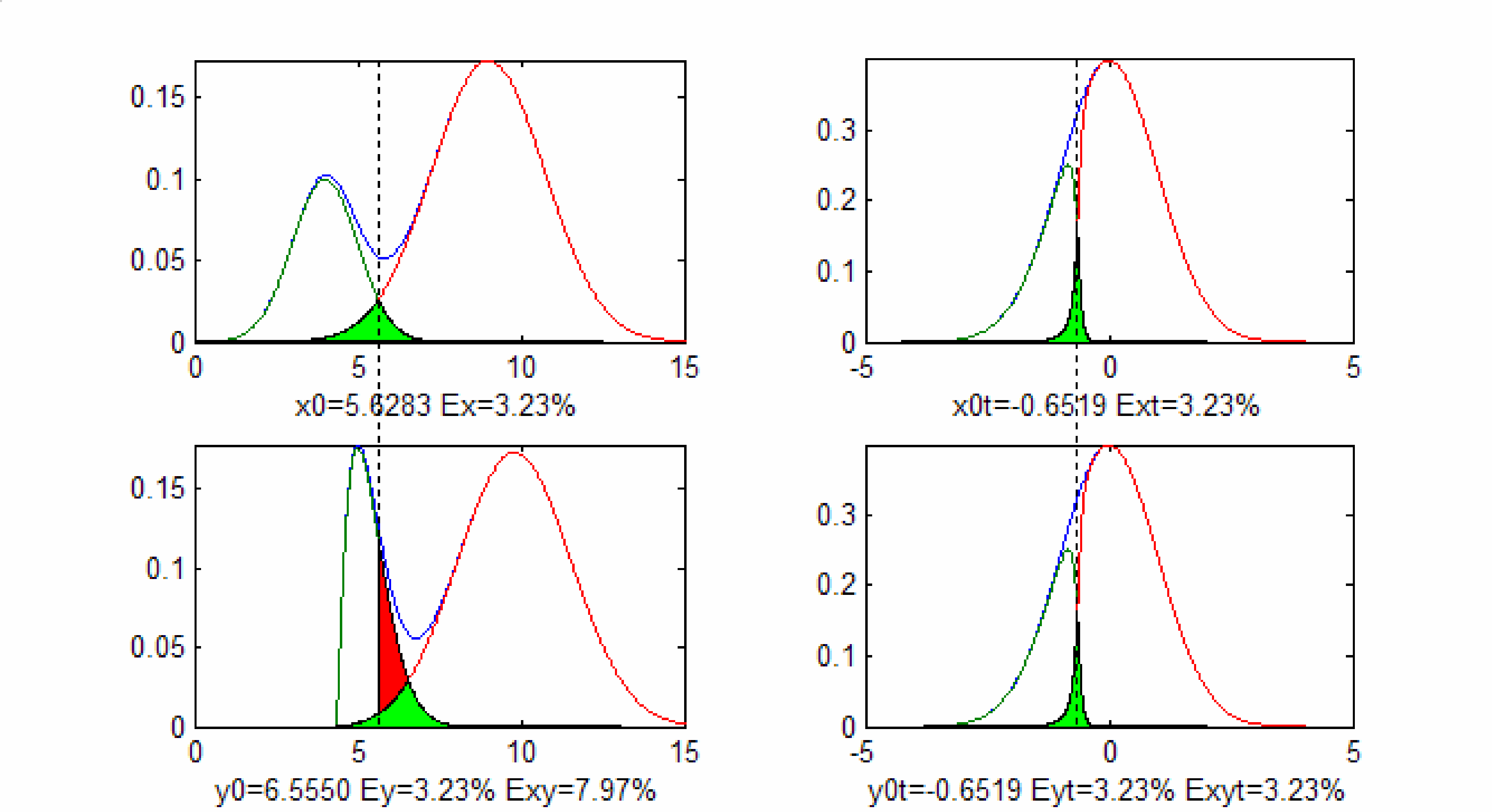
Feature Normalization (III)

- Previous works: Feature compensation
  - R. Balchandran, R. Mammone. ***Non-parametric estimation and correction of non-linear distortion in speech systems*** [ICASSP'98]
    - Domain: Speech samples
    - Task: Speaker ID / Sigmoid and cubic distortions
  - S. Dharanipragada, M. Padmanabhan. ***A nonlinear unsupervised adaptation technique for speech recognition*** [ICSLP'00]
    - Domain: Cepstrum
    - Task: Speech Recognition / Handset / Speaker-phone mismatch
  - F. Hilger, H. Ney. ***Quantile based histogram equalization for noise robust speech recognition*** [EUROSPEECH'01]
    - Domain: Filter-bank Energy
    - Task: Speech Recognition / AURORA task

Feature Normalization (IV)

- Previous works: Feature normalization
  - J. Pelecanos, S. Sridharan. ***Feature warping for robust speaker verification*** [Speaker Odyssey'01]
    - Domain: Cepstrum
    - Task: NIST 1999 Speaker Recognition Evaluation database
  - B. Xiang, U.V. Chaudhari,… ***Short-time gaussianization for robust speaker verification*** [ICASSP'02]
    - Domain: Cepstrum / Short-time
    - Task: Speaker Verification
  - J.C. Segura, A. de la Torre, M.C. Benítez,… ***Non-linear transformations of the feature space for robust speech recognition*** [ICASSP'02]
    - Domain: Cepstrum
    - Task: Speech Recognition / AURORA

Feature Normalization (V)

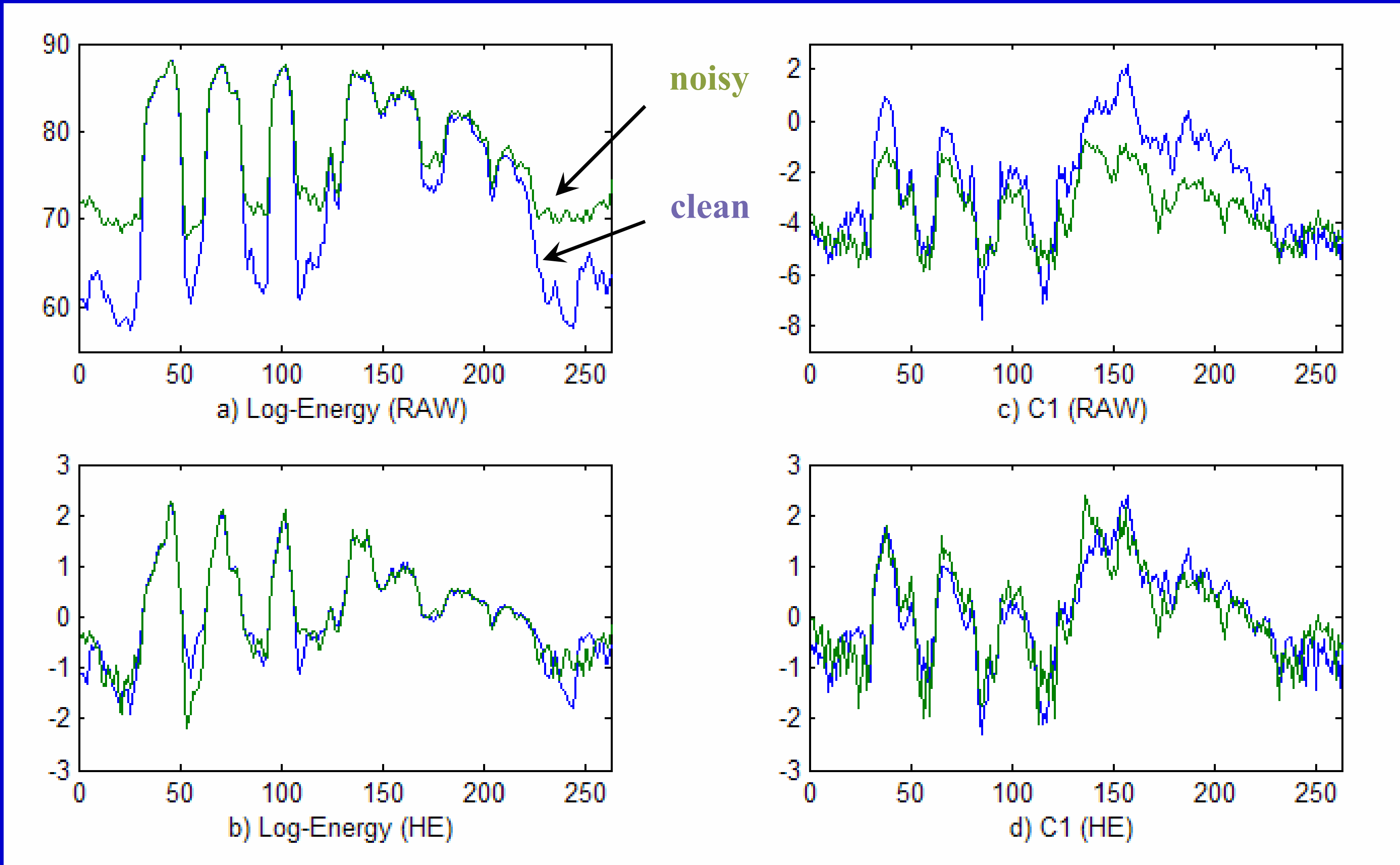


$$y = \log(\exp(x+h)+\exp(n)) \qquad h = 0.8 \qquad n = 3.5$$

Feature Normalization (VI)

- Implementation details
  - CDF-matching is applied in the cepstrum domain in a feature transformation scheme
  - Each cepstral coefficient is transformed independently to match a Gaussian reference distribution
  - Algorithm
    - $C_Y(y)$ is estimated for each feature of each utterance using cumulative histograms
    - The bins centers are transformed and a piecewise linear transformation is constructed
    - The transformation is applied to the input features to get the transformed ones

Feature Normalization (VII)



Experimental set-up

- Database end-pointing
  - Noisy TI-digits and SpeechDat Car databases have been automatically end-pointed
  - SND algorithm is used on clean speech (channel 0) utterances
  - 200ms of silence have been added at the end-points
- Acoustic features
  - Standard front-end: 12 MFCC + logE
  - Delta and acceleration coefficients are appended at the recognizer with regression lengths of 7 and 11 frames respectively
- Acoustic modeling
  - One 16 emitting states left-to-right continuous HMM per digit
  - 3 Gaussian mixture per state

Aurora 2 results

TI-Digits Multi-condition Training					
	A	B	C	Average	Rel.Imp.
Baseline	88.07	87.22	84.56	87.03	----
SS	90.94	88.69	86.29	89.11	9.43%
SS+HE	90.72	89.74	90.03	90.19	15.42%
SS+FD+HE	90.89	89.80	90.11	90.30	17.99%

TI-Digits Clean-condition Training					
	A	B	C	Average	Rel.Imp.
Baseline	58.74	53,40	66.00	58.06	----
SS	73.71	69.35	75.63	72.35	37.71%
SS+HE	82.08	82.61	81.73	82.22	55.59%
SS+FD+HE	82.51	82.78	81.87	82.49	56.45%

23.57%
35.51%
37.22%

Aurora 3 results

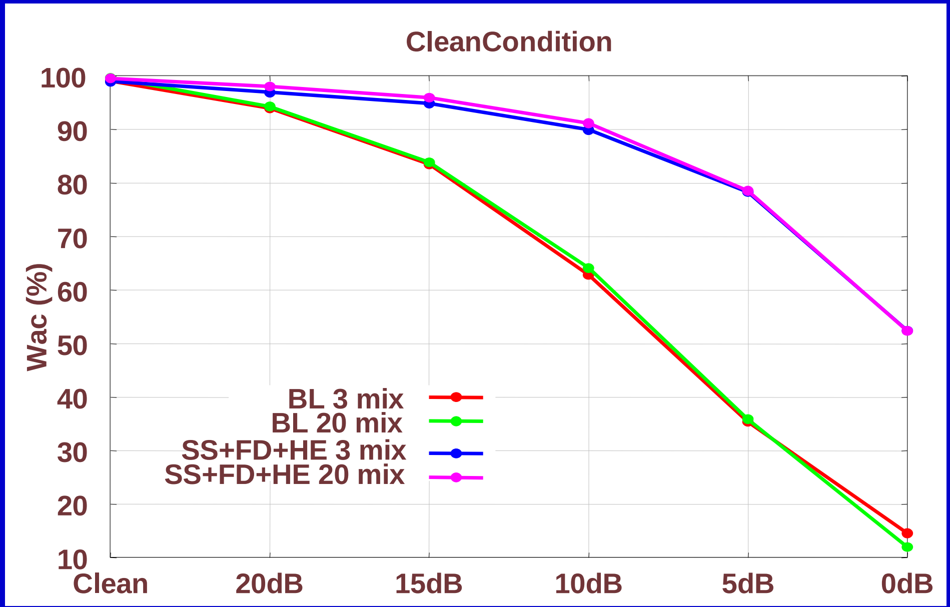
Finnish					
	WM	MM	HM	Average	Rel.Imp.
Baseline	92.74	80.51	40.53	75.41	-----
SS	95.09	78.80	69.19	82.91	21.92%
SS+HE	94.58	86.53	74.20	86.67	35.10%
SS+FD+HE	94.58	86.73	73.11	86.46	35.00%

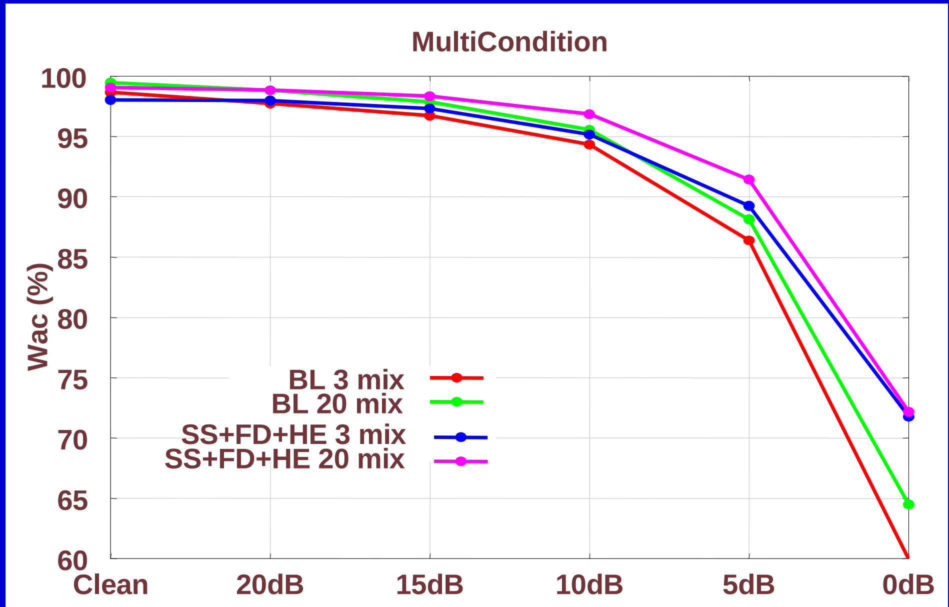
Spanish					
	WM	MM	HM	Average	Rel.Imp.
Baseline	92.94	83.31	51.55	79.22	-----
SS	95.58	89.76	71.94	87.63	39.00%
SS+HE	96.15	93.15	86.77	93.00	57.00%
SS+FD+HE	96.65	94.10	87.03	93.35	61.95%

German					
	WM	MM	HM	Average	Rel.Imp.
Baseline	91.20	81.04	73.17	83.14	-----
SS	93.41	86.60	84.32	88.75	30.70%
SS+HE	94.79	88.58	89.32	91.25	45.29%
SS+FD+HE	94.57	88.07	88.95	90.89	43.00%

30.54%
45.79%
46.65%

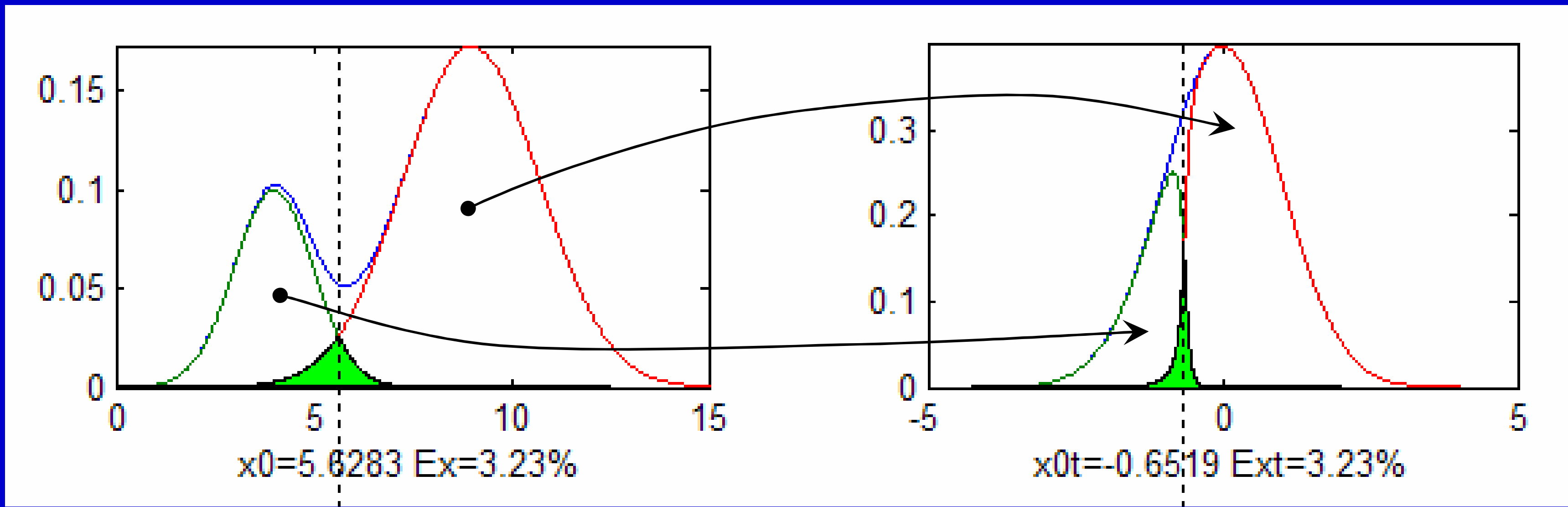
20 mixtures Aurora 2 results





Features	Clean Condition		Multi Condition	
	Absolute	Relative	Absolute	Relative
BL 3mix	58.06	--.--	87.03	--.--
BL 20mix	58.04	4.51%	88.98	26.39%
SS+FD+HE 3mix	82.49	56.45%	90.30	17.99%
SS+FD+HE 20mix	83.22	62.67%	91.53	41.38%

Gaussian class distortion



- Gaussian class densities are transformed into non-Gaussian ones

Conclusions

- A simple and effective SND algorithm based on logarithmic energy quantile difference is presented
- HE is evaluated in combination with classical spectral subtraction with mean relative improvements of 37.22% and 46.65% for AURORA 2 and 3 tasks
- Performance for the 20 mixtures system suggest the need of a higher number of Gaussians after HE

Signal Processing and Communications Group

University of Granada (SPAIN)

This slides are available at

http://sirio.ugr.es/segura/pdfdocs/icslp02_sl.pdf