CDF-matching based Nonlinear Feature Transformations for Robust Speech Recognition

José C. Segura

Signal Processing and Communications Group

University of Granada (SPAIN)

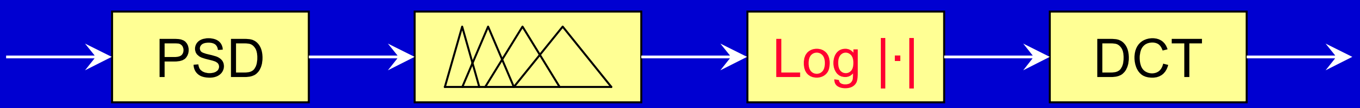
Outline

- Nonlinear effects in speech and speaker recognition
- Mismatch reduction techniques
- CDF-matching based feature transformations
- Cepstral domain nonlinear equalization
- Some experimental results
- Conclusion

Nonlinear effects

- At the signal level
  - Transducer and acquisition hardware

- At the feature level
  - MFCC are generally used as features



Time domain $$y(t) = h(t) * x(t) + n(t)$$

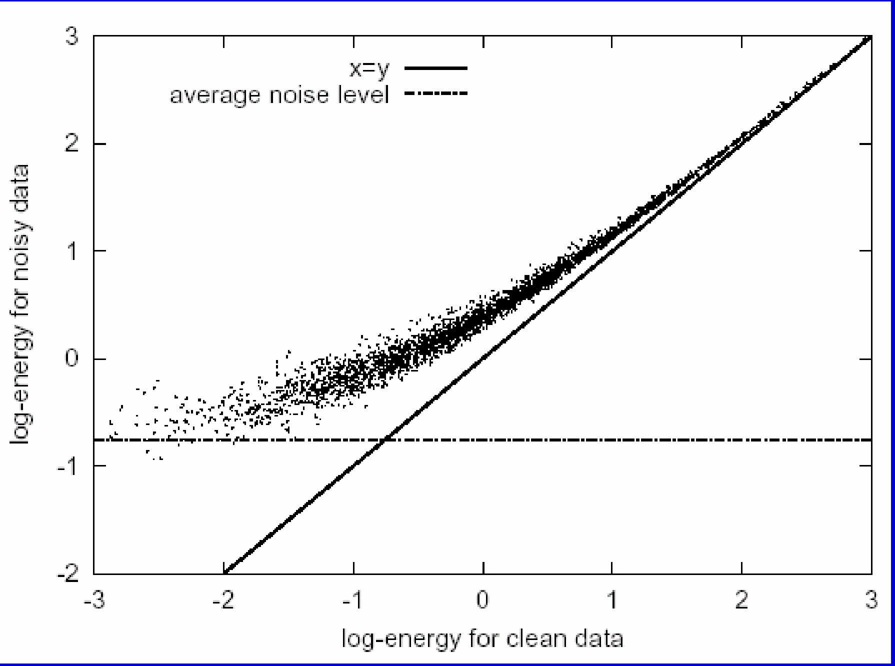
Spectral power domain $$S_y = S_x \cdot |H|^2 + S_n$$

Log-spectral power domain $$y = \log(\exp(x+h) + \exp(n))$$

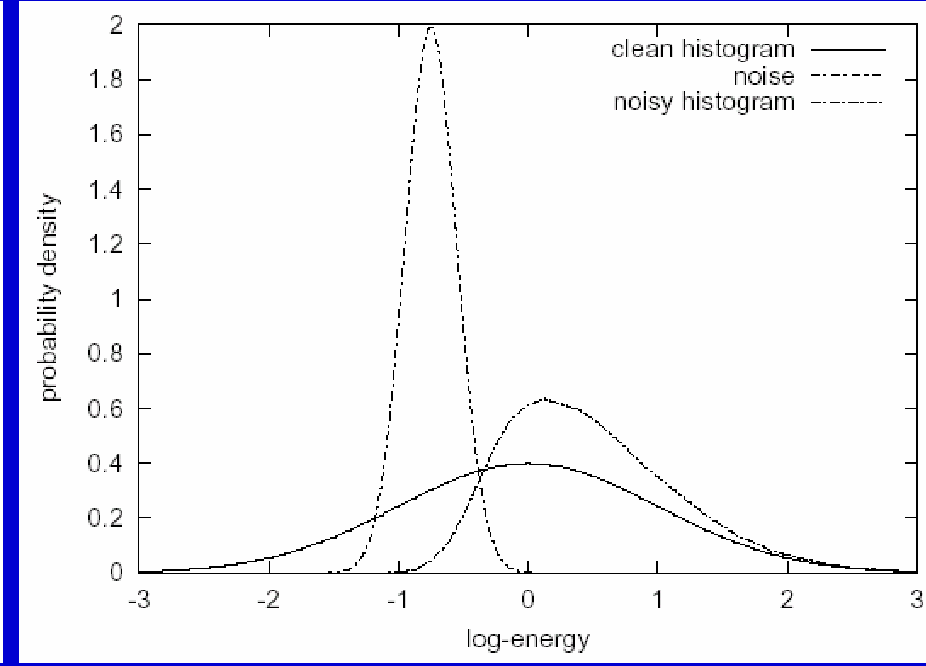
$$x = \log(S_x) \qquad y = \log(S_y) \qquad n = \log(S_n) \qquad h = \log(|H|^2)$$

Log-FBE nonlinear distortion effects

Nonlinear transformation



Transformed PDF



Mismatch reduction

- Linear approaches
  - Spectral subtraction (SS), Wiener filtering (WF)
  - Cepstral Mean Subtraction (CMS)
  - Cepstral Mean and Variance Normalization (CMVN)
  - Time filtering of log-FBE's (RASTA, LDA)
- Nonlinear approaches
  - Linear approximations (CDCN, VTS, SPLICE,…)
  - Neural networks (RBF, MLP)

Feature normalization

- Tries to reduce the mismatch normalizing the feature space
- Linear approaches
  - Cepstral Mean Subtraction
  - Cepstral Mean and Variance Normalization
  - Time filtering of log-FBE's
- Nonlinear extension
  - Compensate not only the location and scale (first and second moment) but also the shape of the PDF's (higher order moments)
  - Our approach is based on CDF-matching

CDF-matching (I)

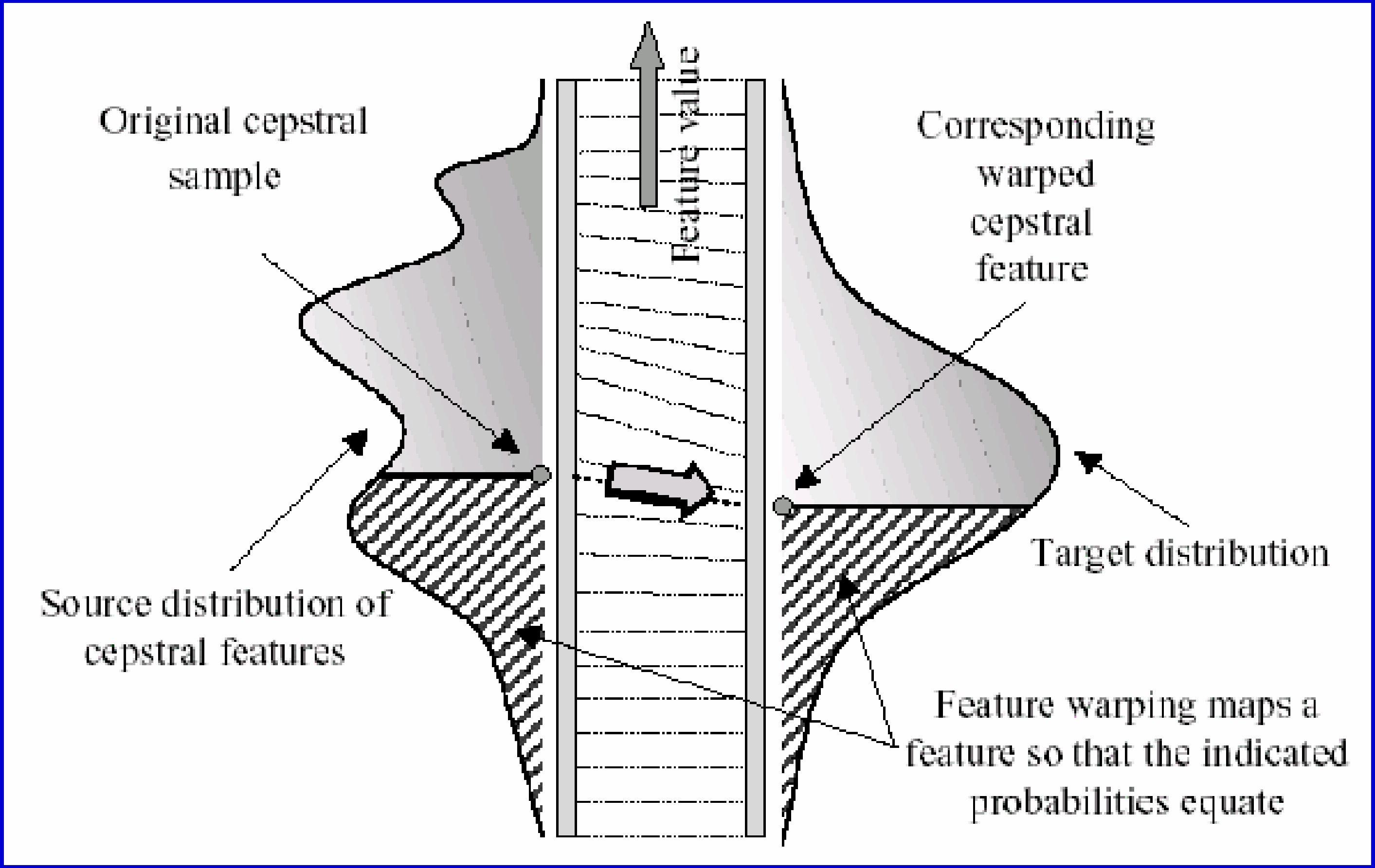
- Given a zero-memory one-to-one general transformation $y=T[x]$

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

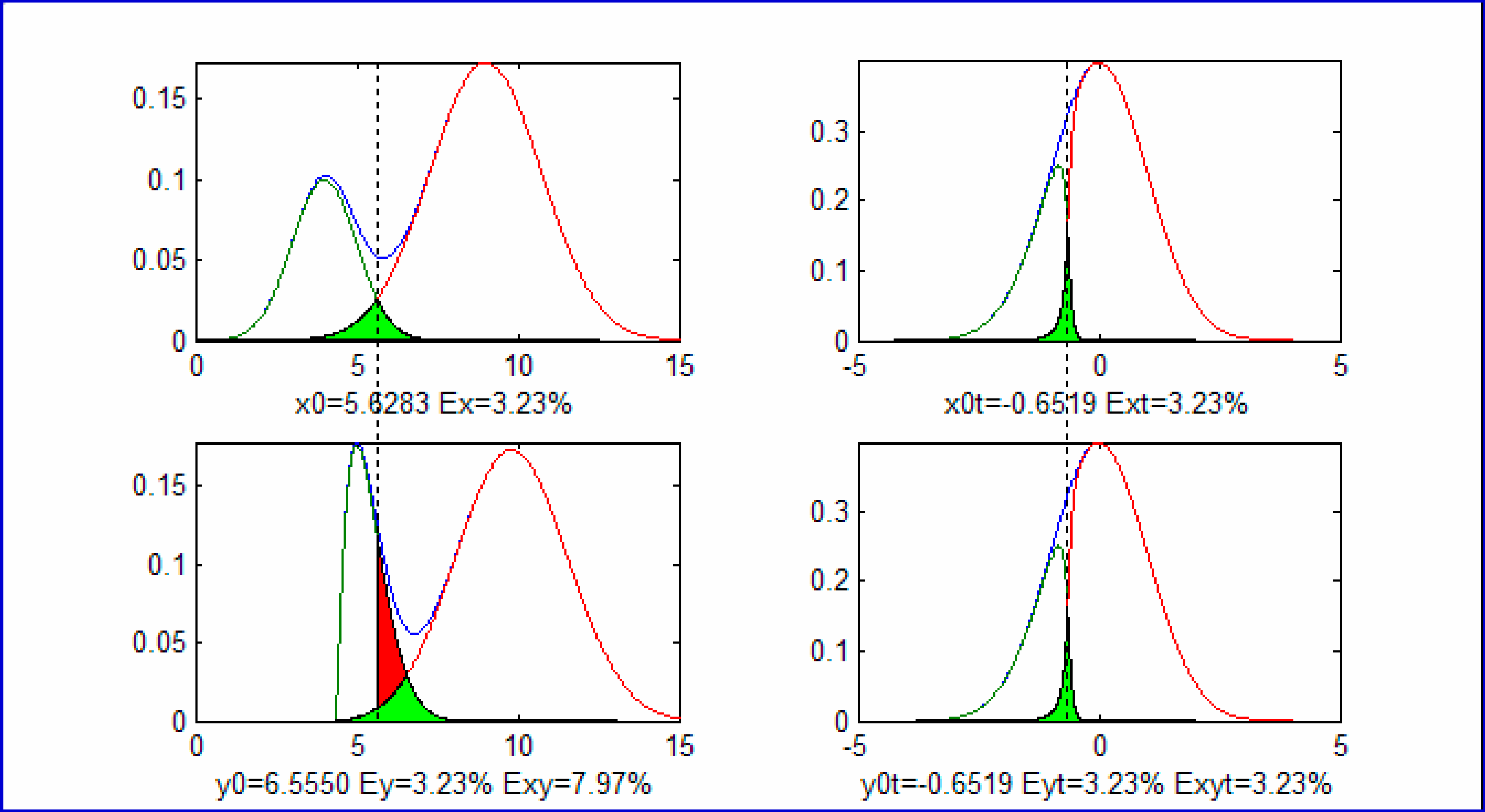
$$C_X(x) = \int_{-\infty}^{x} p_X(u)\,du \qquad\qquad C_Y(y) = \int_{-\infty}^{y} p_Y(u)\,du$$

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

CDF-matching (II)



Two Gaussian class example



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

CDF-matching (III)

- 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

CDF-matching based approaches (I)

- 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

CDF-matching based approaches (II)

- 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
  - J.C. Segura, M.C. Benítez, A. de la Torre, S. Dupont, A.J. Rubio, ***VTS residual noise compensation*** [ICASSP'02]
    - Domain: Cepstrum
    - Task: Speech Recognition / AURORA

CDF-matching based approaches (III)

- Some recent works
  - S. Molau, F. Hilger, D. Kayser, H. Ney. ***Enhanced Histogram Equalization in the acoustic feature space*** [ICSLP'02]
    - Domain: log-FBE
    - Task: Speech Recognition in noise
  - F. Hilger, S. Molau, H. Ney. ***Quantile based histogram equalization for online applications*** [ICSLP'02]
    - Domain: Filter-bank Energy
    - Task: Speech Recognition / AURORA
  - J.C. Segura, A. de la Torre, M.C. Benítez,… ***Feature extraction combining spectral noise reduction and cepstral histogram equalization*** [ICSLP'02]
    - Domain: Cepstrum
    - Task: Speech Recognition / AURORA

Implementation details

- Domain selection
  - Log-FBE
  - Cepstrum (has the advantage that features are almost uncorrelated)
- CDF estimation
  - Using Cumulative Histograms
  - Using the Empirical Cumulative Distribution Function
  - Using sampling quantiles (a reduced number 4-10)
- Reference density
  - Learned from clean data
  - Fixed (usually Gaussian)
- Adaptation data
  - From several sentences to short windows (2-3s)

Efficient implementation with ECDF

$\{x_1, \cdots, x_t, \cdots x_T\}$	Time sequence of features
$\{x_{(1)}, \cdots, x_{(r)}, \cdots x_{(T)}\}$	Sorted sequence
$ECDF(x_{(r)}) = \dfrac{(r-0.5)}{T}$	CDF estimation
$Q(u)$	Reference quantile function

$$T(x_t) = Q\left(\frac{(r-0.5)}{T}\right) \quad \forall \quad x_t = x_{(r)}$$

- For T fixed we only need

$$q_r = Q\left(\frac{(r-0.5)}{T}\right) \quad \forall \quad r = 1, \cdots, T$$

Variable silence lengths (I)

- CDF-matching main assumption
  - The global statistics of speech is independent of the phonetic content

- Problem
  - When using a single sentence to estimate the transformation, this is not true

  - The silence fraction has a special influence
    - If higher than the mean, equalization tends to transform silence into speech increasing the insertion rate
    - If shorter than the mean, equalization tends to transform speech into silence increasing deletions

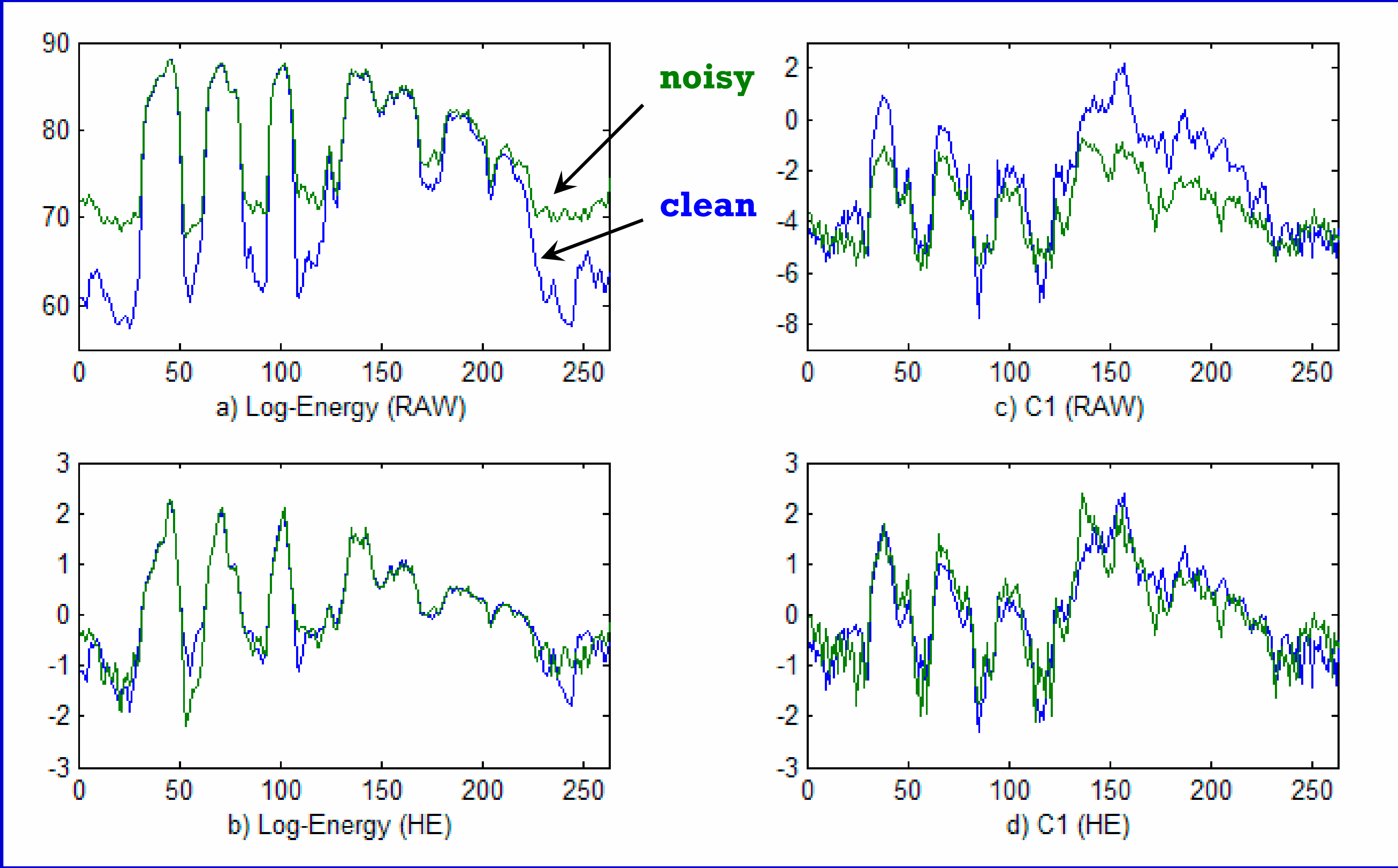
Variable silence lengths (II)

- Possible solutions
  - Adapt the reference histogram
    - This needs an estimation of the silence fraction
      - Using a VAD
      - Perform two pass recognition
  - Use frame-dropping
    - Using a VAD to discard non-speech frames
    - This approach also improves the performance of almost any speech recognition system by limiting the insertion rate

Cepstral domain Nonlinear EQ

- In our current approach
  - Equalization is performed in the cepstral domain
  - For each sentence
    - Each cepstral coefficient is processed independently
    - The reference distribution is a standard Gaussian
  - Frame-Dropping is used to deal with variable silence lengths
    - Equalization is performed after frame-dropping

A real example



Results (I)

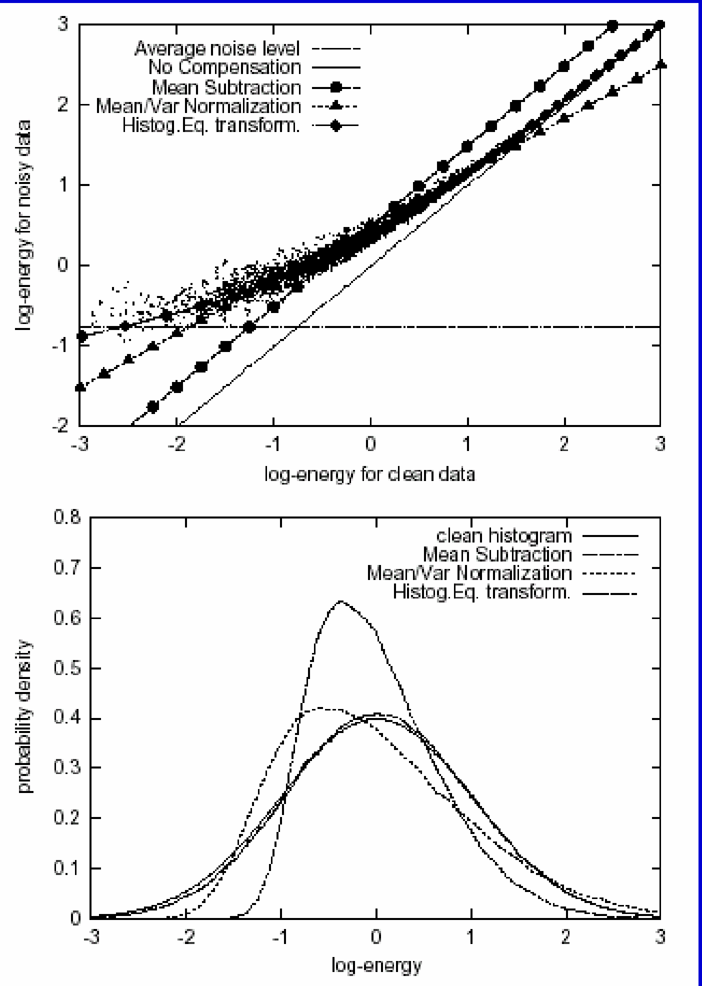
- Experimental set-up: ETSI AURORA tasks
  - Noisy TI-digits (artificially added noise)
    - Experiments: Multi-Condition and Clean-Condition training
  - SpeechDat Car databases (2 microphones in 3 noise conditions)
    - Experiments: Well-Match, Medium-Mismatch, High-Mismatch

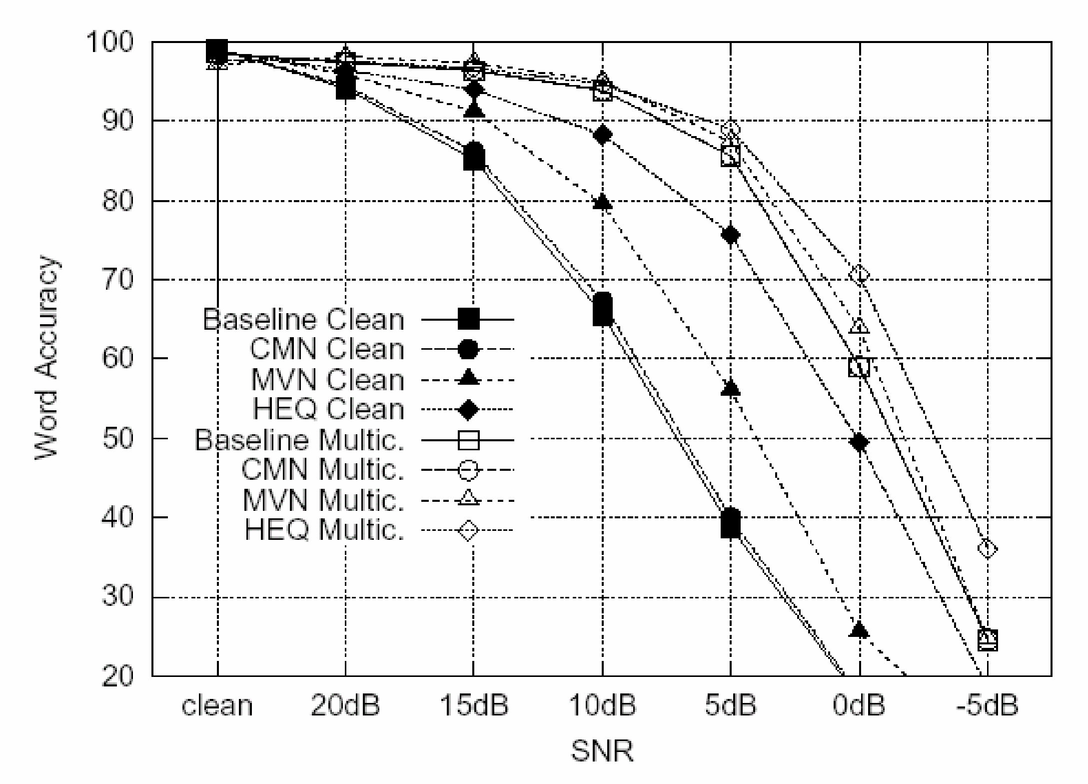
- Acoustic features
  - Standard front-end: 12 MFCC + logE
  - Delta and acceleration coefficients are appended at the recognizer with regression lengths of 7 and 11 frames respectively

- Acoustic modeling
  - One 16 emitting states left-to-right continuous HMM per digit
  - 3 Gaussian mixture per state

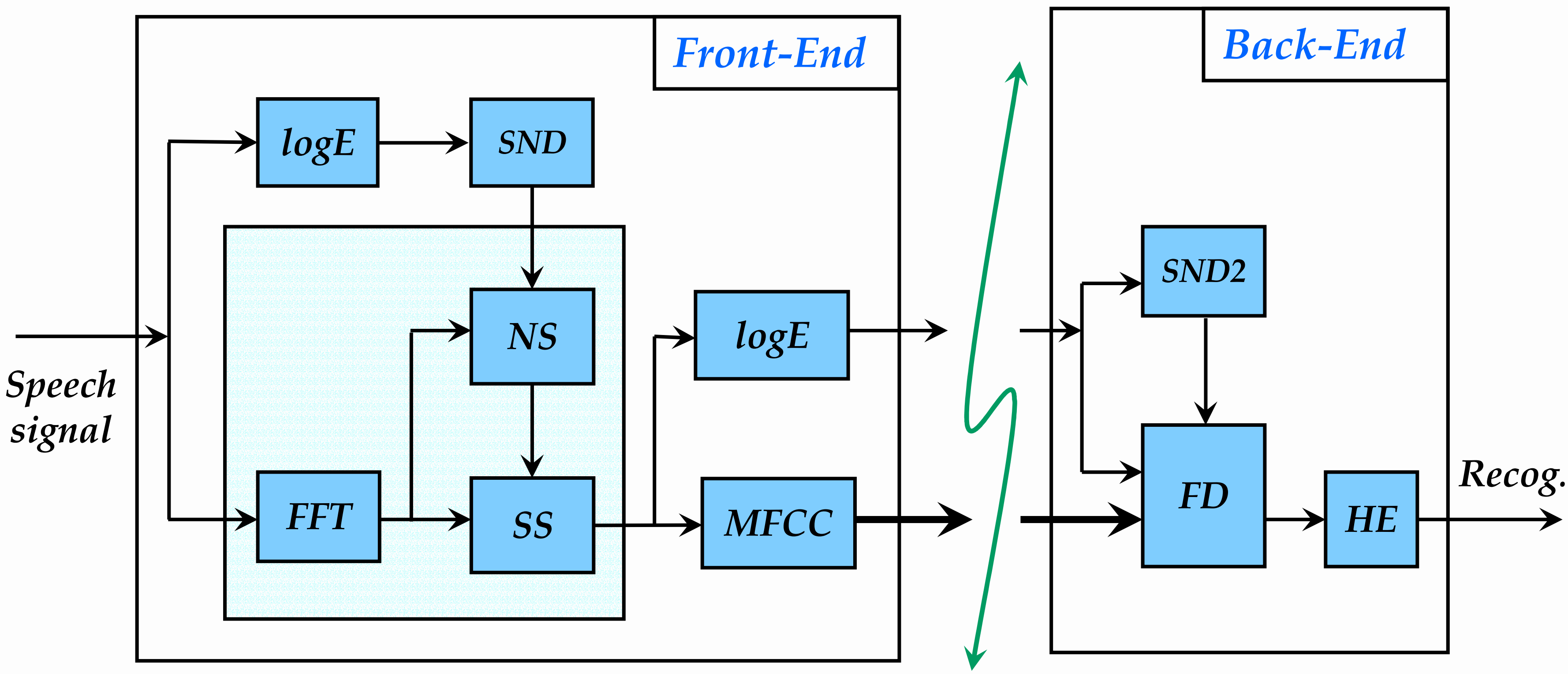
Results (II)

Cepstral equalization (Gaussian reference) compared with CMS and CMVN for noisy TI-digits





Results (III): combined with SS



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

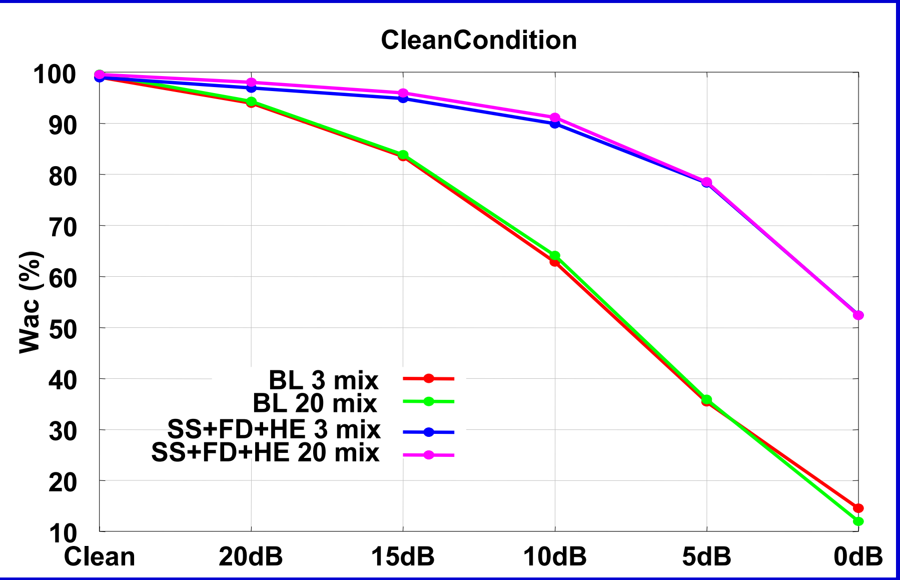
Finnish					
	WM	MM	HM	Average	Rel.Imp.
Baseline	92.74	80.51	40.53	75.41	-----
SS	95.09	78.80	69.19	82.91	21.92%
SS+HE	94.58	86.53	74.20	86.67	35.10%
SS+FD+HE	94.58	86.73	73.11	86.46	35.00%

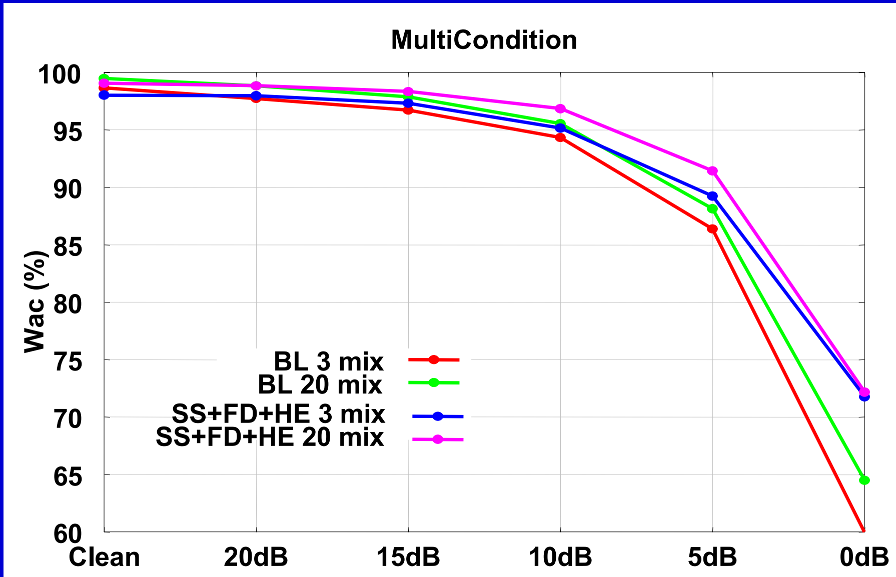
Spanish					
	WM	MM	HM	Average	Rel.Imp.
Baseline	92.94	83.31	51.55	79.22	-----
SS	95.58	89.76	71.94	87.63	39.00%
SS+HE	96.15	93.15	86.77	93.00	57.00%
SS+FD+HE	96.65	94.10	87.03	93.35	61.95%

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%

Conclusion

- Nonlinear cepstral equalization based on CDF-matching is superior to CMS and CMVN

- It can be used as a standalone technique or in combination with noise reduction ones.

- Some open questions
  - Handling variable speech/silence ratios
  - Segmental implementation
  - Selection of the reference distribution
  - Parametric estimation of the CDF
  - Modelling equalized features

Signal Processing and Communications Group

University of Granada (SPAIN)