Improved feature extraction based on spectral noise reduction and nonlinear feature normalization

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

Signal Processing and Communications Group

University of Granada (SPAIN)

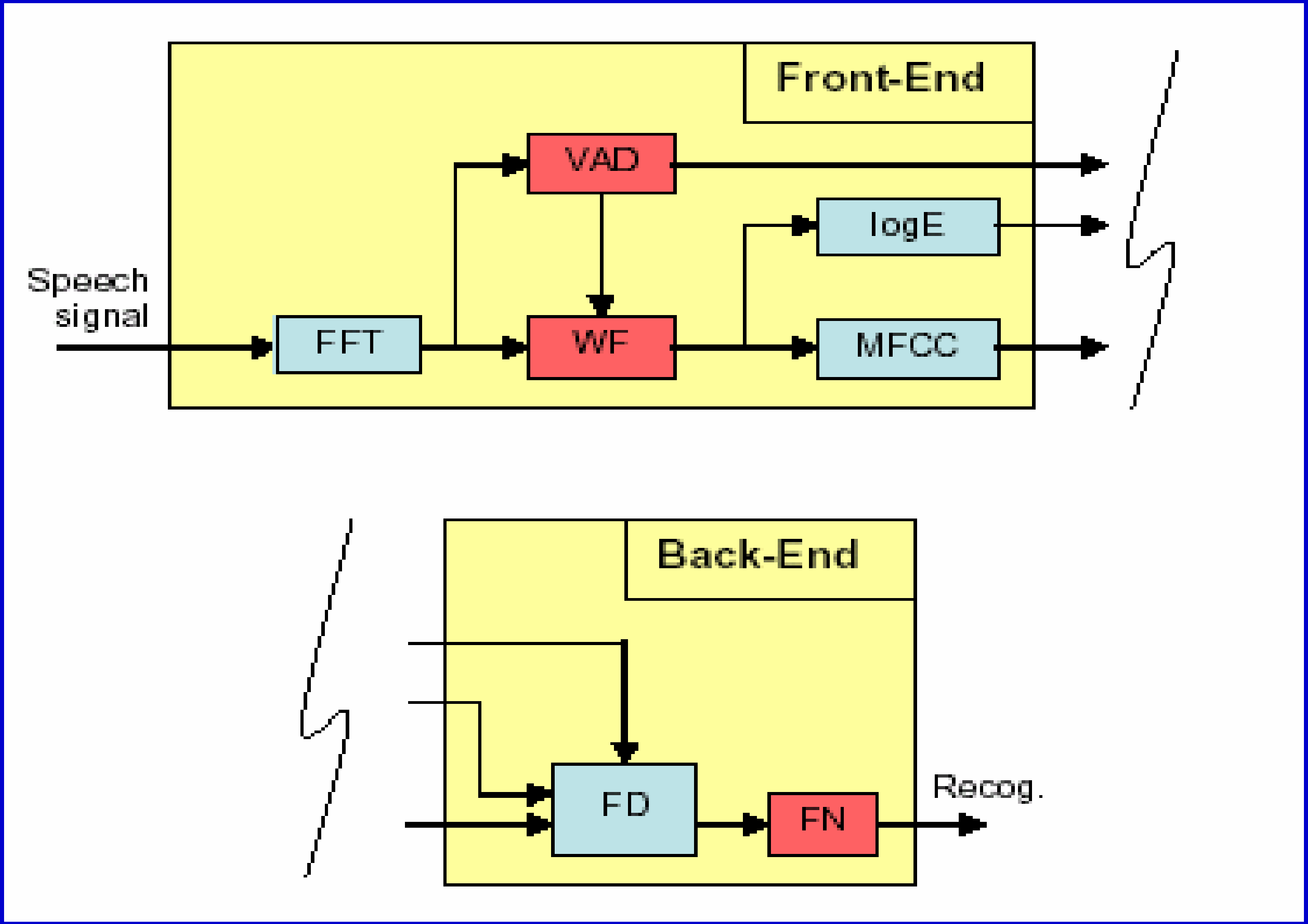
Introduction

- 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
- Results for AURORA 2 and 3 at ICSLP'2002
  - Feature extraction combining spectral noise reduction and cepstral histogram equalization for robust ASR
- In this work we explore CDF-matching performance in combination with Wiener filtering

Outline

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

System Description



Speech/Non-Speech Detection (I)

- Long Term Spectral Estimation VAD algorithm

- LTSE estimation using a sliding window of 3 frames

$$LTSE(k) = \max\left\{X(k, n+l)\right\}_{l=-N}^{l=+N}$$

- Decision rule

$$LTSD = 10\log_{10}\left(\frac{1}{NFFT}\sum_{k=0}^{NFFT-1}\frac{LTSE^2(k)}{Ne^2(k)}\right)$$

- LTSD is compared with an adaptive threshold γ

Speech/Non-Speech Detection (II)

- Threshold γ function of the noise energy

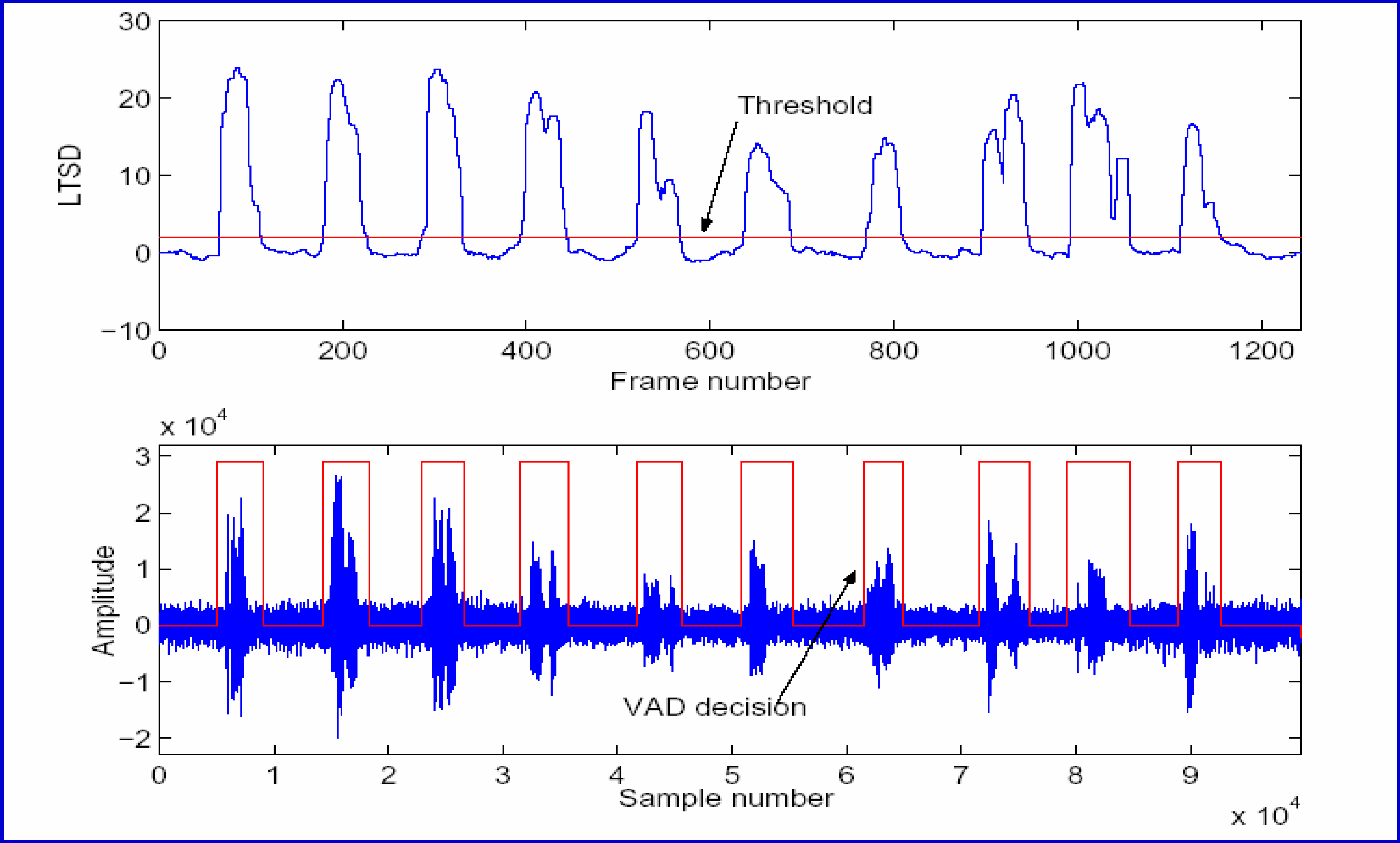
$$\gamma = \begin{cases} \gamma_0 & E \le E_0 \\ \dfrac{\gamma_0 - \gamma_1}{E_0 - E_1}(E - E_0) + \gamma_0 & E_0 < E < E_1 \\ \gamma_1 & E \ge E_1 \end{cases}$$

- VAD parameters

$$N = 3 \qquad NFFT = 256$$
$$\gamma_0 = 5dB \qquad E_0 = 30dB \qquad (low \quad noise \quad energy)$$
$$\gamma_1 = 1.5dB \qquad E_1 = 50dB \qquad (high \quad noise \quad energy)$$

- Adaptive VAD to time varying noise environments
- Details of the algorithm
  - A New Adaptive Long-Term Spectral Estimation Voice Activity Detector (EUROSPEECH'03)

Speech/Non-Speech Detection (III)

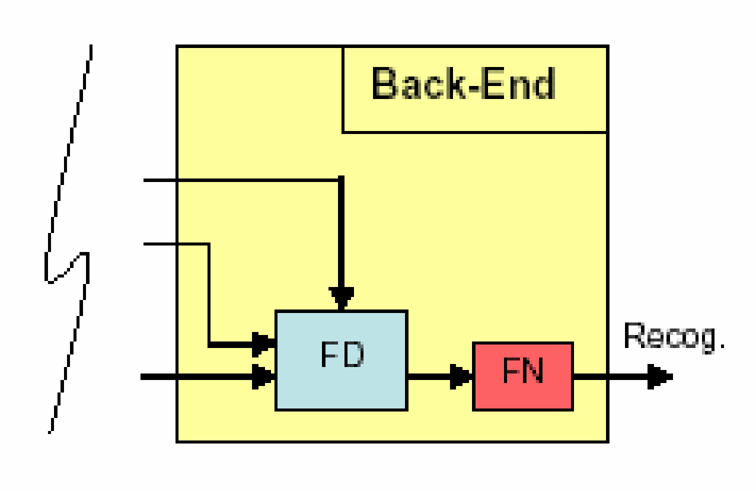


Spectral noise reduction

- Noise reduction implementation as in the first stage of the ETSI ES 202 050 without mel-scale warping.
- Temporal and frequency smoothing of the magnitude spectrum of the noisy frames is applied.
- Maximum attenuation is fixed at 22dB.
- FIR filter with 17 taps is obtained.
- Noise of spectrum estimation as (with $\lambda = 0.99$)

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

Back-end processing



- Frame Dropping
  - Remove all the frames labeled as non-speech
- Feature Normalization
  - ECDF-matching

Feature Normalization (I)

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

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

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

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

Feature Normalization (II)

- 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
- 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

Feature Normalization (III)

- Ecdf Algorithm:
  - Temporal buffer for a given distorted features

$$Y_t = \{ y_{-T}, \cdots, y_t, \cdots, y_T \}$$

  - Order statistics of data

$$y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(r)} \cdots \leq y_{(2T+1)}$$

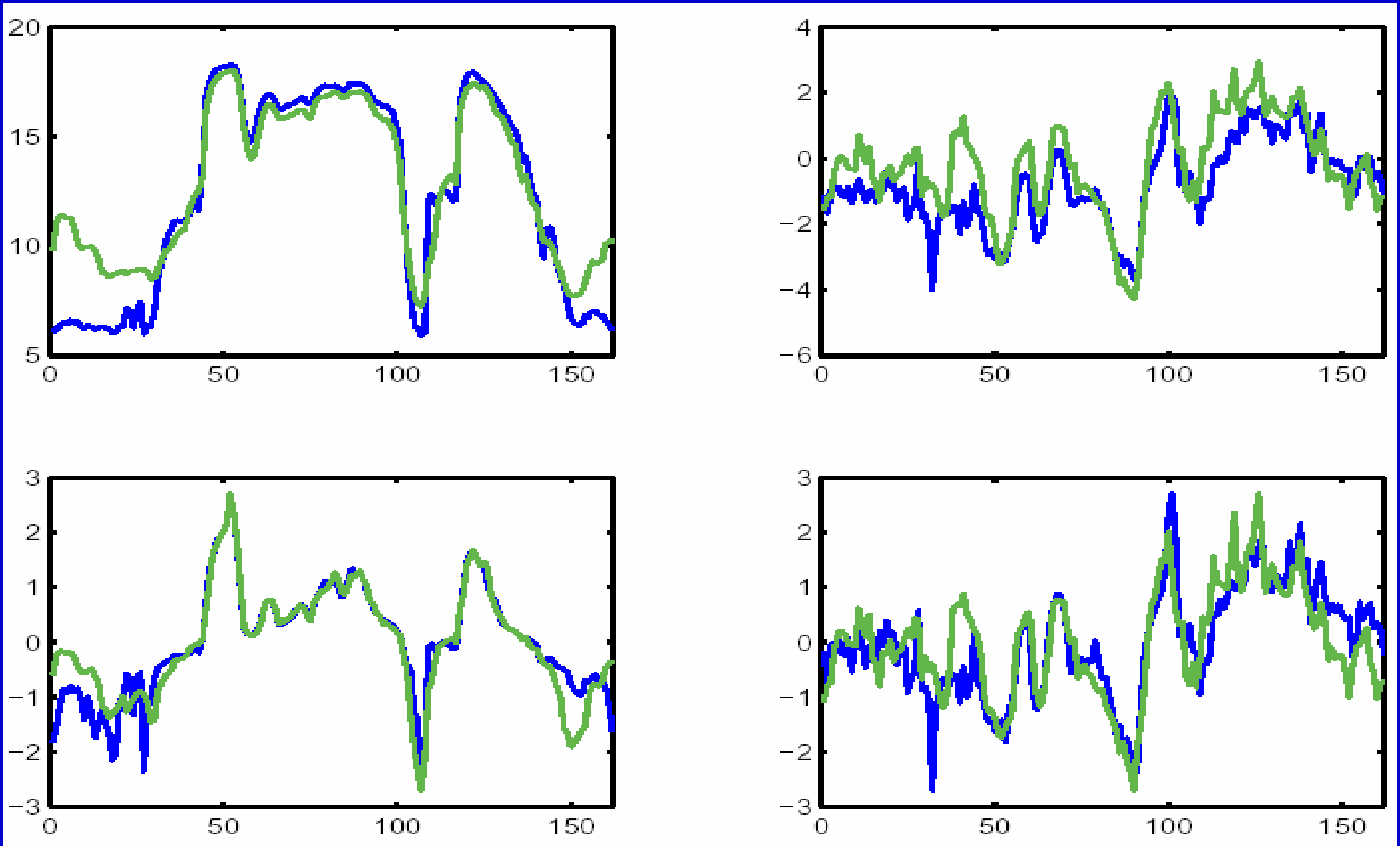
  - Asymptotically unbiased point estimation of the CDF

$$\hat{x}_t = C_x^{-1}\left(\hat{C}(y_t)\right) = C_x^{-1}\left(\frac{r(y_t) - 0.5}{2T+1}\right)$$

  - Estimation of the transformed value of the distorted feature

$$\hat{C}(y_{(r)}) = \frac{r - 0.5}{2T+1} \qquad r = 1, \cdots, 2T+1$$

Feature Normalization (IV)



Experimental set-up (I)

- 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 for AURORA 3
  - 20 Gaussian mixture per state for AURORA 2

Experimental set-up (II)

- Batch implementation
  - Using all the features of a given input utterance to perform the normalization
- Segmental implementation
  - Non-stationary noise
  - Using a short temporal window around the frame to be normalized
  - 121 frames of temporal window

Experimental Results (I)

- Results with *Batch* implementation
  - Comparative results over the previous system (ICSLP'02)

Aurora 2 Relative Improvement				
	Set A	Set B	Set C	Overall
Multi	16,47%	21,79%	20,70%	19,44%
Clean	30,46%	30,59%	28,78%	30,18%
Average	23,46%	26,19%	24,74%	24,81%

Aurora 3 Relative Improvement					
	Finnish	Spanish	German	Danish	Average
Well (x40%)	23,62%	6,57%	19,52%		16,57%
Mid (x35%)	20,12%	-8,98%	15,34%		8,83%
High (x25%)	52,81%	21,36%	19,19%		31,12%
Overall	29,69%	4,82%	17,97%		17,50%

- Spectral Subtraction ----- Wiener filtering
- Quantile based VAD ----- LTSD VAD
- Histogram Equalization ----- ECDF

★ Comparative results over AFE

Aurora 2 Word Error Rate				
	Set A	Set B	Set C	Overall
Multi	6,16%	6,50%	7,27%	6,52%
Clean	12,83%	12,07%	13,63%	12,69%
Average	9,49%	9,28%	10,45%	9,60%

Aurora 2 Relative Improvement				
	Set A	Set B	Set C	Overall
Multi	-8,59%	-4,21%	-3,86%	-5,89%
Clean	-21,13%	-10,50%	-4,46%	-13,54%
Average	-14,86%	-7,35%	-4,16%	-9,72%

Aurora 3 Word Error Rate					
	Finnish	Spanish	German	Danish	Average
Well (x40%)	4,14%	3,13%	4,37%	6,01%	4,41%
Mid (x35%)	10,60%	6,43%	10,10%	14,31%	10,36%
High (x25%)	12,69%	10,20%	8,93%	21,07%	13,22%
Overall	8,54%	6,05%	7,52%	12,68%	8,70%

Aurora 3 Relative Improvement					
	Finnish	Spanish	German	Danish	Average
Well (x40%)	-5,88%	6,85%	10,63%	9,35%	5,24%
Mid (x35%)	44,44%	-5,76%	-10,26%	22,69%	12,78%
High (x25%)	5,23%	-20,71%	-2,06%	-3,23%	-5,19%
Overall	14,51%	-4,45%	0,15%	10,87%	5,27%

❖ Segmental Implementation

Aurora 2 Word Error Rate				
	Set A	Set B	Set C	Overall
Multi	6,33%	6,55%	7,51%	**6,65%**
Clean	13,16%	12,04%	13,64%	**12,81%**
Average	**9,74%**	**9,29%**	**10,57%**	**9,73%**

Aurora 2 Relative Improvement				
	Set A	Set B	Set C	Overall
Multi	-12,68%	-7,21%	-8,87%	**-9,73%**
Clean	-28,78%	-14,51%	-9,10%	**-19,14%**
Average	**-20,73%**	**-10,86%**	**-8,99%**	**-14,43%**

Aurora 3 Word Error Rate					
	Finnish	Spanish	German	Danish	Average
Well (x40%)	4,14%	3,31%	5,09%	6,68%	**4,80%**
Mid (x35%)	10,60%	6,61%	11,27%	16,86%	**11,34%**
High (x25%)	13,25%	8,99%	10,78%	20,44%	**13,37%**
Overall	**8,68%**	**5,89%**	**8,68%**	**13,68%**	**9,23%**

Aurora 3 Relative Improvement					
	Finnish	Spanish	German	Danish	Average
Well (x40%)	-5,88%	1,49%	-4,09%	-0,75%	**-2,31%**
Mid (x35%)	44,44%	-8,72%	-23,03%	8,91%	**5,40%**
High (x25%)	1,05%	-6,39%	-23,20%	-0,15%	**-7,17%**
Overall	**13,46%**	**-4,05%**	**-15,50%**	**2,78%**	**-0,83%**

Conclusions

- Feature extraction algorithm based on the combination of spectral noise reduction and nonlinear features normalization
- New VAD based on Long Term spectral envelope
  - Improve the noise estimation
  - Frame dropping
  - Better discrimination speech/noise
- More computational efficiency of the feature normalization algorithm
- Segmental version of the feature normalization algorithm
  - Performance is only slightly worse
- Results presented for AURORA 2 and AURORA 3

Signal Processing and Communications Group

University of Granada (SPAIN)

These slides are available at

http://sirio.ugr.es/segura/pdfdocs/eurospeech'03_sl.pdf