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 signal

Front-End

VAD

logE

Back-End

FD

FN

Recog.
Speech/Non-Speech Detection (I)

- Long Term Spectral Estimation VAD algorithm
- LTSE estimation using a sliding window of 3 frames

\[
LTSE(k) = \max \{ X(k, n + l) \}_{l=-N}^{l=N}
\]

- 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 \( \gamma \)
Speech/Non-Speech Detection (II)

- Threshold $\gamma$ function of the noise energy
  \[
  \gamma = \begin{cases} 
  \gamma_0 & E \leq E_0 \\
  \frac{\gamma_0 - \gamma_1}{E_0 - E_1} (E - E_0) + \gamma_0 & E_0 < E < E_1 \\
  \gamma_1 & E \geq E_1
  \end{cases}
  \]

- VAD parameters
  - $N = 3$, $NFFT = 256$
  - $\gamma_0 = 5dB$, $E_0 = 30dB$ (low noise energy)
  - $\gamma_1 = 1.5dB$, $E_1 = 50dB$ (high noise 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)

Threshold

VAD decision

LTSD

Frame number

Amplitude

Sample number
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 \( y = T[x] \)

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

\[
C_X(x) = \int_{-\infty}^{x} p_X(u) \, du \quad C_Y(y) = \int_{-\infty}^{y} p_Y(u) \, du
\]

\[
C_X(x) = C_Y(y) \quad \Rightarrow \quad x = T^{-1}[y] = C_X^{-1}(C_Y(y))
\]
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
    \[
    \hat{C}(y_{(r)}) = \frac{r-0.5}{2T+1} \quad r = 1, \cdots, 2T + 1
    \]
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

<table>
<thead>
<tr>
<th></th>
<th>Set A</th>
<th>Set B</th>
<th>Set C</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi</td>
<td>16.47%</td>
<td>21.79%</td>
<td>20.70%</td>
<td>19.44%</td>
</tr>
<tr>
<td>Clean</td>
<td>30.46%</td>
<td>30.59%</td>
<td>28.78%</td>
<td>30.18%</td>
</tr>
<tr>
<td>Average</td>
<td>23.46%</td>
<td>26.19%</td>
<td>24.74%</td>
<td>24.81%</td>
</tr>
</tbody>
</table>

### Aurora 3 Relative Improvement

<table>
<thead>
<tr>
<th></th>
<th>Finnish</th>
<th>Spanish</th>
<th>German</th>
<th>Danish</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Well (x40%)</td>
<td>23.62%</td>
<td>6.57%</td>
<td>19.52%</td>
<td></td>
<td>16.57%</td>
</tr>
<tr>
<td>Mid (x35%)</td>
<td>20.12%</td>
<td>-8.98%</td>
<td>15.34%</td>
<td></td>
<td>8.83%</td>
</tr>
<tr>
<td>High (x25%)</td>
<td>52.81%</td>
<td>21.36%</td>
<td>19.19%</td>
<td></td>
<td>31.12%</td>
</tr>
<tr>
<td>Overall</td>
<td>29.69%</td>
<td>4.82%</td>
<td>17.97%</td>
<td></td>
<td>17.50%</td>
</tr>
</tbody>
</table>

- Spectral Subtraction ----- Wiener filtering
- Quantile based VAD ----- LTSD VAD
- Histogram Equalization ----- ECDF
Comparative results over AFE

### Aurora 2 Word Error Rate

<table>
<thead>
<tr>
<th></th>
<th>Set A</th>
<th>Set B</th>
<th>Set C</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi</td>
<td>6.16%</td>
<td>6.50%</td>
<td>7.27%</td>
<td>6.52%</td>
</tr>
<tr>
<td>Clean</td>
<td>12.83%</td>
<td>12.07%</td>
<td>13.63%</td>
<td>12.69%</td>
</tr>
<tr>
<td>Average</td>
<td>9.49%</td>
<td>9.28%</td>
<td>10.45%</td>
<td>9.60%</td>
</tr>
</tbody>
</table>

### Aurora 2 Relative Improvement

<table>
<thead>
<tr>
<th></th>
<th>Set A</th>
<th>Set B</th>
<th>Set C</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi</td>
<td>-8.59%</td>
<td>-4.21%</td>
<td>-3.86%</td>
<td>-5.89%</td>
</tr>
<tr>
<td>Clean</td>
<td>-21.13%</td>
<td>-10.50%</td>
<td>-4.46%</td>
<td>-13.54%</td>
</tr>
<tr>
<td>Average</td>
<td>-14.86%</td>
<td>-7.35%</td>
<td>-4.16%</td>
<td>-9.72%</td>
</tr>
</tbody>
</table>

### Aurora 3 Word Error Rate

<table>
<thead>
<tr>
<th></th>
<th>Finnish</th>
<th>Spanish</th>
<th>German</th>
<th>Danish</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Well (x40%)</td>
<td>4.14%</td>
<td>3.13%</td>
<td>4.37%</td>
<td>6.01%</td>
<td>4.41%</td>
</tr>
<tr>
<td>Mid (x35%)</td>
<td>10.60%</td>
<td>6.43%</td>
<td>10.10%</td>
<td>14.31%</td>
<td>10.36%</td>
</tr>
<tr>
<td>High (x25%)</td>
<td>12.69%</td>
<td>10.20%</td>
<td>8.93%</td>
<td>21.07%</td>
<td>13.22%</td>
</tr>
<tr>
<td>Overall</td>
<td>8.54%</td>
<td>6.05%</td>
<td>7.52%</td>
<td>12.68%</td>
<td>8.70%</td>
</tr>
</tbody>
</table>

### Aurora 3 Relative Improvement

<table>
<thead>
<tr>
<th></th>
<th>Finnish</th>
<th>Spanish</th>
<th>German</th>
<th>Danish</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Well (x40%)</td>
<td>-5.88%</td>
<td>6.85%</td>
<td>10.63%</td>
<td>9.35%</td>
<td>5.24%</td>
</tr>
<tr>
<td>Mid (x35%)</td>
<td>44.44%</td>
<td>-5.76%</td>
<td>-10.26%</td>
<td>22.69%</td>
<td>12.78%</td>
</tr>
<tr>
<td>High (x25%)</td>
<td>5.23%</td>
<td>-20.71%</td>
<td>-2.06%</td>
<td>-3.23%</td>
<td>-5.19%</td>
</tr>
<tr>
<td>Overall</td>
<td>14.51%</td>
<td>-4.45%</td>
<td>0.15%</td>
<td>10.87%</td>
<td>5.27%</td>
</tr>
</tbody>
</table>
## Segmental Implementation

### Aurora 2 Word Error Rate

<table>
<thead>
<tr>
<th></th>
<th>Set A</th>
<th>Set B</th>
<th>Set C</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi</td>
<td>6.33%</td>
<td>6.55%</td>
<td>7.51%</td>
<td>6.65%</td>
</tr>
<tr>
<td>Clean</td>
<td>13.16%</td>
<td>12.04%</td>
<td>13.64%</td>
<td>12.81%</td>
</tr>
<tr>
<td>Average</td>
<td>9.74%</td>
<td>9.29%</td>
<td>10.57%</td>
<td>9.73%</td>
</tr>
</tbody>
</table>

### Aurora 2 Relative Improvement

<table>
<thead>
<tr>
<th></th>
<th>Set A</th>
<th>Set B</th>
<th>Set C</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi</td>
<td>-12.68%</td>
<td>-7.21%</td>
<td>-8.87%</td>
<td>-9.73%</td>
</tr>
<tr>
<td>Clean</td>
<td>-28.78%</td>
<td>-14.51%</td>
<td>-9.10%</td>
<td>-19.14%</td>
</tr>
<tr>
<td>Average</td>
<td>-20.73%</td>
<td>-10.86%</td>
<td>-8.99%</td>
<td>-14.43%</td>
</tr>
</tbody>
</table>

### Aurora 3 Word Error Rate

<table>
<thead>
<tr>
<th></th>
<th>Finnish</th>
<th>Spanish</th>
<th>German</th>
<th>Danish</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Well (x40%)</td>
<td>4.14%</td>
<td>3.31%</td>
<td>5.09%</td>
<td>6.68%</td>
<td>4.80%</td>
</tr>
<tr>
<td>Mid (x35%)</td>
<td>10.60%</td>
<td>6.61%</td>
<td>11.27%</td>
<td>16.86%</td>
<td>11.34%</td>
</tr>
<tr>
<td>High (x25%)</td>
<td>13.25%</td>
<td>8.99%</td>
<td>10.78%</td>
<td>20.44%</td>
<td>13.37%</td>
</tr>
<tr>
<td>Overall</td>
<td>8.68%</td>
<td>5.89%</td>
<td>8.68%</td>
<td>13.68%</td>
<td>9.23%</td>
</tr>
</tbody>
</table>

### Aurora 3 Relative Improvement

<table>
<thead>
<tr>
<th></th>
<th>Finnish</th>
<th>Spanish</th>
<th>German</th>
<th>Danish</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Well (x40%)</td>
<td>-5.88%</td>
<td>1.49%</td>
<td>-4.09%</td>
<td>-0.75%</td>
<td>-2.31%</td>
</tr>
<tr>
<td>Mid (x35%)</td>
<td>44.44%</td>
<td>-8.72%</td>
<td>-23.03%</td>
<td>8.91%</td>
<td>5.40%</td>
</tr>
<tr>
<td>High (x25%)</td>
<td>1.05%</td>
<td>-6.39%</td>
<td>-23.20%</td>
<td>-0.15%</td>
<td>-7.17%</td>
</tr>
<tr>
<td>Overall</td>
<td>13.46%</td>
<td>-4.05%</td>
<td>-15.50%</td>
<td>2.78%</td>
<td>-0.83%</td>
</tr>
</tbody>
</table>
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