A New Objective Intelligibility Measure For Time-Frequency Weighted Noisy Speech

Cees Taal\textsuperscript{1}, Richard Hendriks\textsuperscript{1}, Richard Heusdens\textsuperscript{1}, Jesper Jensen\textsuperscript{2}

7-1-2010

\textsuperscript{1} Delft University of Technology, Signal Information & Processing Lab, Delft, the Netherlands.

\textsuperscript{2} Oticon A/S, Smørum, Copenhagen, Denmark.
Introduction

Background

Clean Speech

- Background Noise
- Hearing Aid

Signal Modification, Degradation
- Coding
- Enhancement

Evaluation

- Subjective Listening Test: e.g. % of correctly understood words
- Objective Measure: Objective score
Introduction

Motivation

• In this research, the focus is on time-frequency (TF) weighted noisy speech
  • e.g., single-channel noise reduction, speech separation etc.

• Why?
  • Most conventional objective measures are not reliable for this type of processing
  • Such a reliable measure is desired in the field of noise-reduction

We propose a new objective measure which,
  • ... shows high correlation with intelligibility of noisy and TF-weighted noisy speech
  • ... is simple (very few parameters)
  • ... based on short-time segments (∼400 ms)
Method

- First, TF-decomposition is applied to clean and processed speech
  - 15, 1/3 octave bands, by merging short-time (~25 ms) DFT-bins
  - Bands cover a relevant frequency range for speech intelligibility (~150-4500 Hz)

- Notation:
  - Band index: j, time index: m
  - Clean speech TF-unit: \( X_j(m) \), processed speech TF-unit: \( Y_j(m) \)
Method

Intermediate Intelligibility Measure

- Model depends on intermediate intelligibility measure: $d_j(m)$
  - $d_j(m)$ depends on short segments (~400 ms) of $X_j(n)$ and $Y_j(n)$, per band
  - Where $n \in \{m - N + 1, m - N + 2, \ldots, m\}$ and $N=30$

- Before comparison, $Y_j(n)$ is first modified as follows:
  - Normalization: Compensate for local energy differences
  - Clipping: To make sure speech is inside range relevant for intelligibility
Method

- $Y_j(n)$ is normalized such that its energy equals the energy of $X_j(n)$:

$$
\alpha Y_j(n) = \frac{1}{\sqrt{\sum_n Y_j(n)^2}} \sqrt{\sum_n X_j(n)^2} Y_j(n)
$$

Normalization

- $Y_j(n)$ is normalized such that its energy equals the energy of $X_j(n)$:
Method

- $\alpha Y_j(n)$ is clipped to lower-bound the signal to distortion ratio to -15 dB which gives $Y'_j(n)$

$$SDR(A, B) = 10 \log_{10} \left( \frac{A^2}{(B - A)^2} \right)$$
Method

- $d_j(m)$ equals correlation coefficient between clean and processed speech short-time segments

\[
d_j(m) = \frac{\sum_n (X_j(n) - \mu_x)(Y'_j(n) - \mu_y)}{\sqrt{\sum_n (X_j(n) - \mu_x)^2 \sum_n (Y'_j(n) - \mu_y)^2}}
\]
Method

Eventual outcome

- Eventual outcome is defined as the average over all intermediate intelligibility measures:

\[
d = \frac{1}{JM} \sum_{m,j} d_j(m)
\]
Subjective Data

- Subjective data origins from Kjems *et al.* (2009)
  - Speech is degraded with additive noise
  - Noisy speech is processed with a technique called ‘Ideal Time Frequency Segregation’ (ITFS), Brungart *et al.* (2006)

- In total 167 different conditions are evaluated
  - 3 SNRs
  - 4 noise types
  - Various settings of ITFS-algorithm
Experiment

- Proposed method is compared with three reference objective measures:
  - DAU: Dau auditory model (Dau et. al, 1996)
  - NSEC: (Boldt & Ellis, 2009)
  - CSTI: Normalized covariance based STI (Goldsworthy & Greenberg, 2006)

- All these measures are promising candidates for TF-weighted noisy speech
Results

- Figure of merits:
  - RMSE ($\sigma$)
  - Correlation Coefficient ($\rho$)

<table>
<thead>
<tr>
<th></th>
<th>PROP</th>
<th>CSTI</th>
<th>DAU</th>
<th>NSEC</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma$</td>
<td>10.2%</td>
<td>21.8%</td>
<td>16.4%</td>
<td>17.1%</td>
</tr>
<tr>
<td>$\rho$</td>
<td>0.95</td>
<td>0.73</td>
<td>0.86</td>
<td>0.84</td>
</tr>
</tbody>
</table>
Results

- Reference objective measures underestimate intelligibility of noisy unprocessed speech
- Proposed method good results with both noisy and TF-weighted noisy speech
Conclusions

• A new objective intelligibility measure was presented, based on an intermediate measure for short time-frequency regions (~400 ms)

• The proposed method:
  • ...showed high correlation with TF-weighted noisy speech
  • ...showed better performance then three other reference objective measures
  • ... does not underestimate the intelligibility of the unprocessed noisy speech, which was the case for the three reference objective measures

• Matlab code available: http://www.ceestaal.nl/stoi.zip
Experimental results

![Graph showing experimental results with various values of \( \beta \) and corresponding times.]

- \( \beta = -\infty \)
- \( \beta = -30 \)
- \( \beta = -20 \)
- \( \beta = -15 \)
- \( \beta = -10 \)
Subjective Data
Ideal Time-Frequency Segregation

- Binary time-frequency weighting is applied to noisy speech (Ideal Binary Mask, IBM)
- Mask set to ‘1’ when local SNR within TF-unit exceeds user-defined local criterion (LC):

\[
IBM(f,t) = \begin{cases} 
1, & \text{if } \frac{\text{clean}(f,t)}{\text{noise}(f,t)} > LC \\
0, & \text{otherwise}
\end{cases}
\]
Subjective Data
Ideal Time-Frequency Segregation

- In total 167 different conditions are evaluated:
  - Speech shaped noise, café noise, car interior noise, noise from bottling factory hall
  - 8 different LC-values
  - 3 SNRs: 20% SRT, 50% SRT, -60 dB