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

Cees H. Taal, Richard C. Hendriks, Richard Heusdens and Jesper Jensen
Delft, University of Technology, Mediamatics Dept. Mekelweg 4, CD Delft, The Netherlands

Contact: C.H.Taal@tudelft.nl

Background

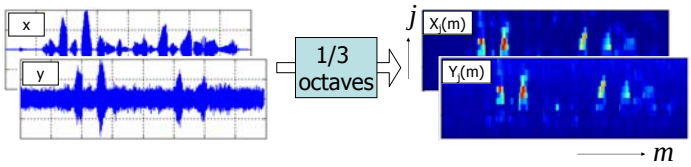
An objective intelligibility measure predicts the effect on speech-intelligibility, due to some speech degradation (e.g. additive noise).

Most conventional objective measures are not reliable for time-frequency (TF) weighted noisy speech (e.g., noise reduction, speech-separation). A reliable intelligibility measure is desired in the field of noise-reduction for evaluation and development purposes.

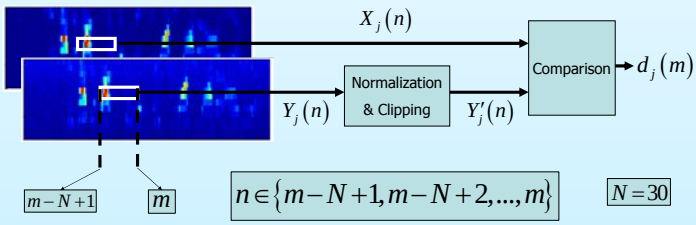
A new objective measure is presented based on short-time segments (~400 ms), which shows high correlation with TF-weighted noisy speech.

Proposed method

Clean (x) and processed (y) speech are first decomposed into 15, DFT-based (~25 ms) 1/3 octave bands (~150-4500 Hz.).



Proposed method depends on intermediate intelligibility measure for short-time (~400 ms) segments, per band:



$$n \in \{m-N+1, m-N+2, ..., m\} \qquad N = 30$$

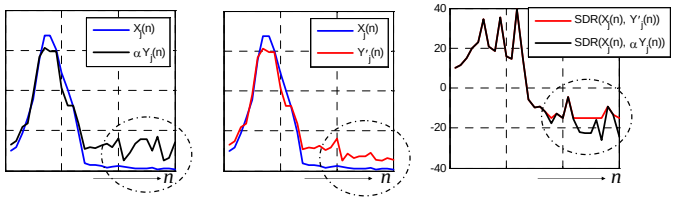
First, normalization is applied to compensate for any local level differences:

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

followed by clipping to lower-bound the SDR to -15 dB per TF-unit:

$$Y'_j(n) = \max\left(\alpha Y_j(n), X_j(n) + 10^{15/20} X_j(n)\right)$$

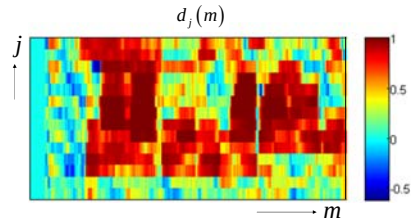
$$SDR(A,B) = 10\log_{10}\left(\frac{A^2}{(B-A)^2}\right)$$



Both signals are compared by means of a correlation coefficient:

$$d_j(m) = \frac{\sum_n \left(X_j(n) - \mu_X\right)\left(Y'_j(n) - \mu_{Y'}\right)}{\sqrt{\sum_n \left(X_j(n) - \mu_X\right)^2 \sum_n \left(Y'_j(n) - \mu_{Y'}\right)^2}}$$

Eventual outcome is obtained by the average over all TF-units:



$$d = \frac{1}{JM} \sum_{j,m} d_j(m)$$

Evaluation

Subjective data origins from Kjems *et al.* (2009):

- Speech is degraded with 4 different noise types, 3 SNRs
- Noisy speech is TF-weighted with a technique called 'Ideal Time Frequency Segregation' (ITFS), Brungart *et al.* (2006)
- In total, 167 different conditions (including unprocessed noisy speech)

Proposed method compared with three reference objective measures

- Dau-auditory model
- Normalized-covariance based STI
- NSEC

Results

	PROP	CSTI	DAU	NSEC
RMSE	10.2%	21.8%	16.4%	17.1%
ρ	0.95	0.73	0.86	0.84