IEEE Signal Processing Letters

An Instrumental Intelligibility Metric Based on Information Theory

Steven Van Kuyk[Vict], W. Bastiaan Kleijn[Vict,TUD] and Richard C. Hendriks[TUD]

Problem

- When designing a speech-based communication system it is important to understand how the system will affect intelligibility (i.e., the proportion of correctly identified words).
- Formal listening tests provide valid data, but are time-consuming and expensive.
- Intelligibility metrics that predict the intelligibility of speech signals have been proposed, but their usefulness is limited to specific types of distortion (e.g., noise, reverb, enhancement).
- Needed: an intelligibility metric that generalizes to many types of distortions.

Contributions

- We propose a monaural intrusive intelligibility metric called SIIB (speech intelligibility in bits).
- SIIB estimates the amount of information shared between a talker and a listener in bits per second.
- Unlike existing information theoretic intelligibility metrics, SIIB accounts for talker variability and time-frequency dependencies.

Communication Model

- A talker randomly selects a message, $\{\mathbf{M}_t\}$, e.g., a phoneme, word, or neural state, where t is the time index.
- The talker encodes the message into a speech signal, $\{\mathbf{X}_t\}$, according to a conditional probability distribution: $p(\{X_t\}|\{M_t\})$. In this way, talker variability is incorporated into the communication model.
- The speech signal is transmitted to a listener through a communication channel. Let $\{\mathbf{Y}_t\}$ denote the received signal.
- We call $\{\mathbf{M}_t\} \to \{\mathbf{X}_t\}$ the speech production channel, and call $\{\mathbf{X}_t\} \to \{\mathbf{Y}_t\}$ the environmental channel.
- We represent $\{\mathbf{X}_t\}$ and $\{\mathbf{Y}_t\}$ as sequences of log-spectra on an ERB frequency scale.
- SIIB is based on the hypothesis that intelligibility is a function of the mutual information rate of $\{\mathbf{M}_t\}$ and $\{\mathbf{Y}_t\}$.

The Information Rate

- Let $\mathbf{M}^K = [(\mathbf{M}_1)^T, (\mathbf{M}_2)^T, \cdots, (\mathbf{M}_K)^T]^T$ be a vector obtained by stacking K consecutive message vectors and similarly for $\mathbf{X}^K$ and $\mathbf{Y}^K$.
- The mutual information rate is defined by

$$I(\{\mathbf{M}_t\};\{\mathbf{Y}_t\}) = \lim_{K\to\infty} \frac{1}{K} I(\mathbf{M}^K;\mathbf{Y}^K),$$

  where $I(\mathbf{M}^K;\mathbf{Y}^K)$ denotes mutual information.
- An upper bound for the rate can be obtained by applying the data processing inequality twice:

$$I(\{\mathbf{M}_t\};\{\mathbf{Y}_t\}) \leq \min\left(I(\{\mathbf{M}_t\};\{\mathbf{X}_t\}), I(\{\mathbf{X}_t\};\{\mathbf{Y}_t\})\right)$$

- Define $\tilde{\mathbf{X}}^K = f(\mathbf{X}^K)$, where f is an invertible transform that removes statistical dependencies between the elements of $\mathbf{X}^K$ and similarly for $\tilde{\mathbf{Y}}^K$. To this end, we use the Karhunen-Loève Transform (KLT).
- The information rate of the environmental channel can then be written as a summation:

$$I(\{\mathbf{X}_t\};\{\mathbf{Y}_t\}) = \lim_{K\to\infty} \frac{1}{K} I(\tilde{\mathbf{X}}^K;\tilde{\mathbf{Y}}^K)$$

$$= \lim_{K\to\infty} \frac{1}{K} \sum_{j=1}^{KJ} I(\tilde{\mathbf{X}}_j^K;\tilde{\mathbf{Y}}_j^K).$$

- Approximating $\{\mathbf{M}_t\}$ and $\{\mathbf{X}_t\}$ as Gaussian, the information rate of the speech production channel is

$$I(\{\mathbf{M}_t\};\{\mathbf{X}_t\}) = \lim_{K\to\infty} -\frac{1}{K} \sum_{j=1}^{KJ} \frac{1}{2}\log_2(1-r_j^2),$$

  where the *production correlation coefficient*, $r_j = 0.75$, describes the efficiency of encoding a message according to $p(\{X_t\}|\{M_t\})$.
- **SIIB typically ranges from 0 b/s (zero intelligibility) to 150 b/s (high intelligibility).**

Evaluation

- An ideal intelligibility metric would have a monotonic increasing relationship with intelligibility scores.
- We quantify the strength of the relationship using Kendall's tau, τ, and Pearson's correlation, ρ.

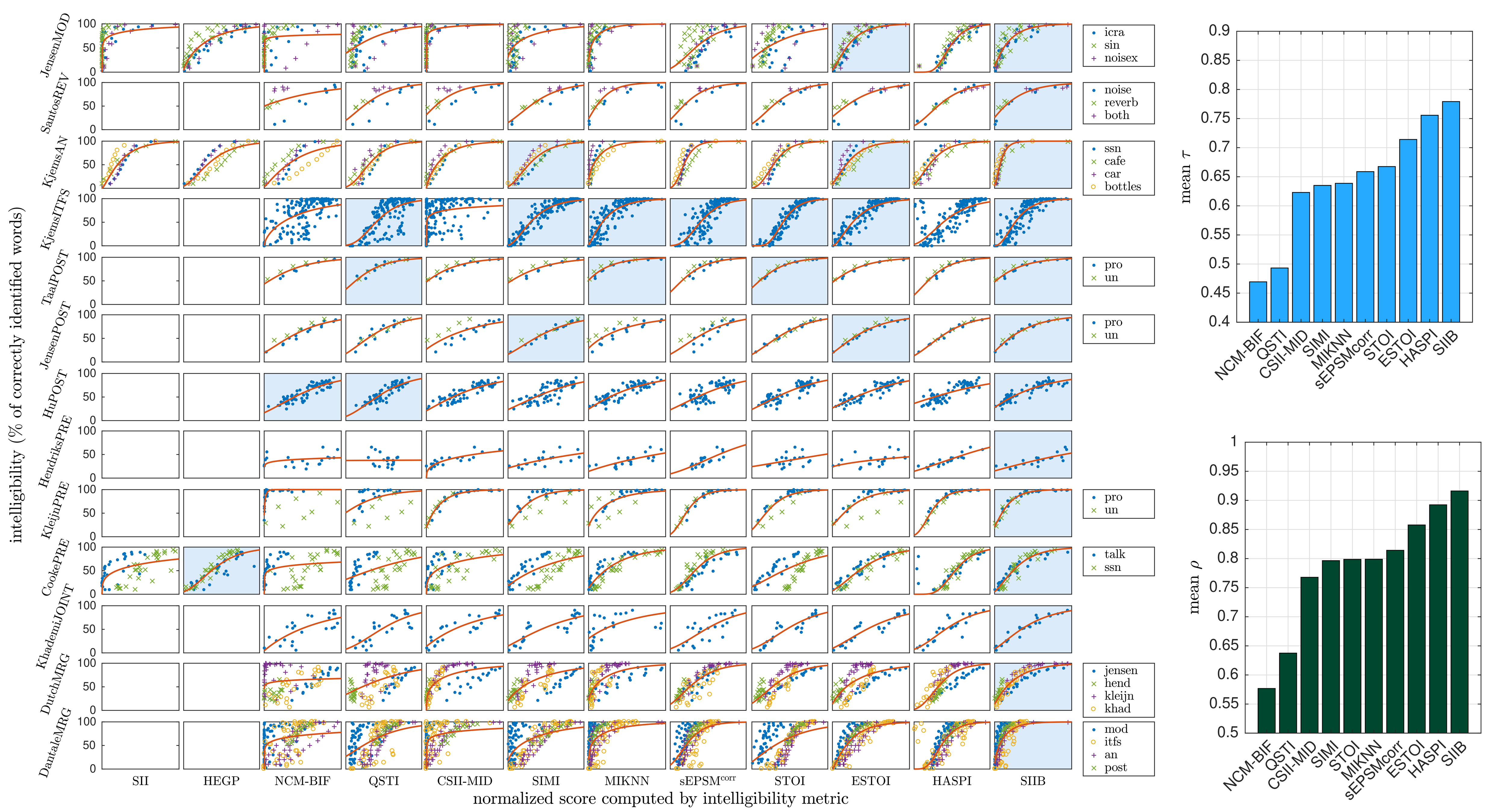
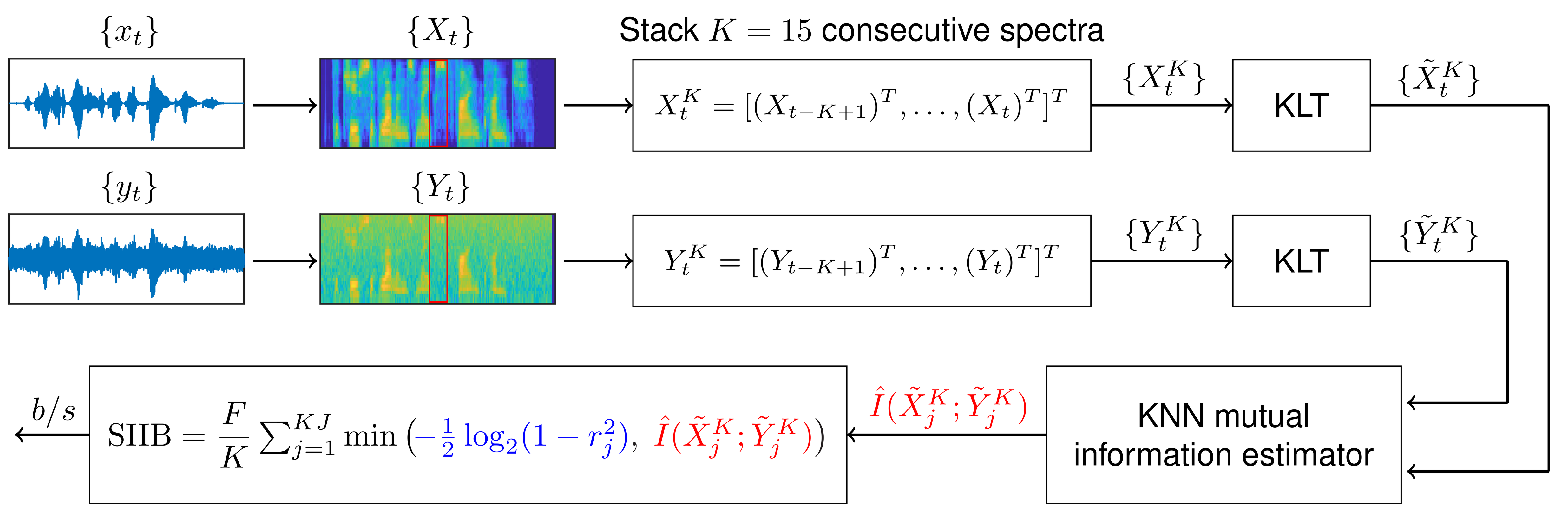


Figure from Van Kuyk et al., 'An evaluation of intrusive instrumental intelligibility metrics', https://arxiv.org/abs/1708.06027

Intelligibility Data Sets

JensenMOD	10 types of modulated noise.	HuPOST	4 noise types; 8 SCNR algorithms.
SantosREV	2 noise types; reverb.	HendriksPRE	4 pre-processing enhancement algorithms; noise; reverb.
KjemsAN	4 noise types.		
KjemsITFS	4 noise types; ideal binary mask.	KleijnPRE	3 pre-processing enhancement; 2 noise types.
TaalPOST	Noise; 2 single-channel noise reduction (SCNR).	CookePRE	9 pre-processing enhancement; 2 noise types.
JensenPOST	Noise; 3 SCNR algorithms.	KhademiJOINT	Pre-processing enhancement; SCNR; noise.

Proposed Algorithm (SIIB)



Conclusions

- SIIB and HASPI have the highest performance overall and are the only intelligibility metrics that attempt to reduce statistical dependencies between input features.
- The KLT does not remove all of the statistical dependencies. Accounting for the remaining dependencies may give an information rate closer to the lexical information rate of ≈ 50 b/s.
- Intelligibility metrics perform worse on 'unseen' data ($\rho = 0.75$) than on 'seen' data ($\rho = 0.91$).
- A MATLAB implementation is available at: ***https://stevenvankuyk.com/matlab_code***