An Objective Intelligibility Metric Evaluation on Spanish Speech

Iván López-Espejo[1], Jesper Jensen[2,3]

[1]Dept. of Signal Theory, Telematics and Communications, University of Granada, Spain
[2]Dept. of Electronic Systems, Aalborg University, Denmark
[3]Oticon A/S, Denmark

Introduction

- **Speech intelligibility** (**SI**) is a critical factor in the development and evaluation of speech enhancement, hearing aids, speech synthesis, etc.
- SI is ideally assessed through **listening tests** → Time-consuming, labor-intensive, and costly!
- *Alternative*: **Objective intelligibility metrics** (**OIMs**), which can be **intrusive** or **non-intrusive**
- *In this work*: We evaluate 5 intrusive and 2 non-intrusive OIMs on a novel **Spanish speech intelligibility dataset**, **SpInt**

Discussion & Conclusion

- **ASR and SI assessment are related tasks** (*speech-to-text mapping*) → MOSA-Net+ and W2V-SIP use ASR models for feature extraction
- Spanish is linguistically closer to English than to Taiwanese Mandarin, which may partly explain the poorer performance of MOSA-Net+ w.r.t. W2V-SIP
- Intrusive OIMs' access to clean reference signals may make them more robust to training–test acoustic mismatches, such as **language mismatch** (*relevant in our setting*)
- OIMs generally capture the impact of SNR on SI, but struggle to reliably rank processing conditions (*in agreement with previous work*)
- *Speech enhancement evaluation*: Human intelligibility assessments cannot be systematically replaced by OIMs!
- **To support the development of more robust and generalizable** deep learning-based **non-intrusive OIMs**, we have made **SpInt publicly available**:

https://zenodo.org/records/20763580

Contact Information

E-mail: **iloes@ugr.es**
Signal Processing, Multimedia Transmission and Speech/Audio Technologies Group, **SigMAT**
Dept. of Signal Theory, Telematics and Communications, University of Granada, Spain

SpInt: A Spanish Speech Intelligibility Dataset

- **SpInt**: A **Spanish speech intelligibility dataset** collected using a matrix sentence test
- SpInt contains 5,148 utterances paired with behavioral intelligibility scores



Listening test stimuli

» Oldenburg Sentence Test (OLSA) speech material

» Bus station additive noise

» *SNRs*: -15, -12.5, -10, -7.5, -5, -2.5, 0, 2.5, 5, 10, and $+\infty$ dB

» Unprocessed (**Unp**), FullSubNet+ (**FSN+**), and SGMSE+ (**SGM+**)

» 1 noise type $\times$ 11 SNRs $\times$ 3 processing conditions $\times$ 6 sentences $=$ **198 stimuli**

Listening test procedure

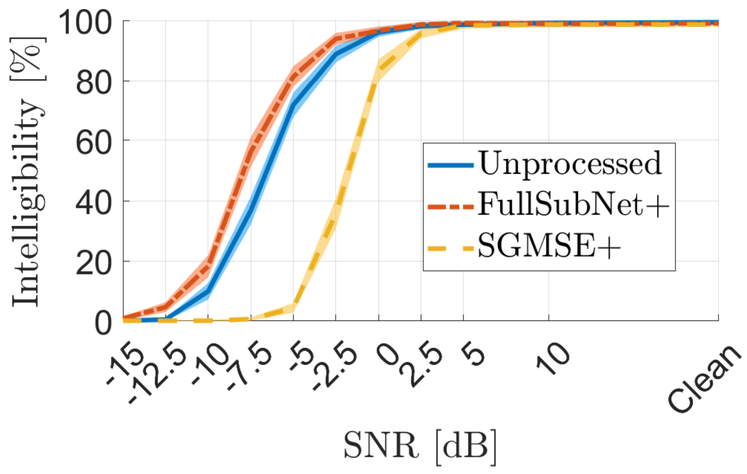
» Quiet room + Beyerdynamic DT 770 Pro headph. + Focusrite Scarlett 2i2 audio int. + laptop

» `COMENZAR` → stimulus playback (*only once*) + word selection → `SIGUIENTE`

» *Training* (18 stim.) → **1st block** (99/198 stim.) → ☕ → **2nd block** (99/198 stim.)

» 26 native Spanish speakers (13F/13M, age: 33.6 $\pm$ 13.5 years) with *self-reported normal hearing*

- SI is computed as the percentage of correctly recognized words
- Results are aggregated across participants
- SI degradation by SGMSE+ (*not reflected in OIMs*): Language mismatch, training SNR range ($[0, 20]$ dB), and diffusion model *hallucination*



Evaluation of Intelligibility Metrics on SpInt

- 5 intrusive (**STOI**, **ESTOI**, **STGI**, **HASPI**, and **SIIB**) and 2 deep learning-based non-intrusive (**MOSA-Net+** and **W2V-SIP**) OIMs
- OIMs should have a strong positive monotonic relationship with SI → Evaluation by **Pearson** (ρ) and **Spearman rank** (r_s) correlation coefficients

Overall performance

	STOI	ESTOI	STGI	HASPI	SIIB	MOSA-Net+	W2V-SIP
ρ	0.94	0.89	**0.96**	0.87	0.53	0.83	0.82
r_s	0.95	0.95	0.96	0.92	**0.97**	0.84	0.93

Performance per processing condition

	Pearson corr. (ρ)			*Spearman rank (r_s)*		
Metric	Unp	FSN+	SGM+	Unp	FSN+	SGM+
STOI	0.87	0.95	**0.99**	**1.00**	**0.97**	0.95
ESTOI	0.82	0.93	**0.99**	**1.00**	**0.97**	**1.00**
STGI	**0.98**	**0.99**	0.97	**1.00**	**0.97**	**1.00**
HASPI	0.83	0.98	0.97	**1.00**	**0.97**	**1.00**
SIIB	0.44	0.50	0.67	**1.00**	**0.97**	**1.00**
MOSA-Net+	0.83	0.94	0.97	0.85	0.96	0.86
W2V-SIP	0.87	0.67	0.92	0.99	0.94	**1.00**

Performance per SNR [dB]

