Multilingual Speech Intelligibility Testing in Crowdsourced and Laboratory Settings

Nerio Moran, Miguel Plaza, Ginette Prato, Shirley Pestana, Daniel Arismendi, Jose Kordahi, Cyprian Wronka, Laura Lechler, Kamil Wojcicki.
Cisco Systems, Inc.

Problem Description

- Generative AI bears high potential but also risk.
- Laboratory-based measures are costly and do not scale well.
- No crowd-sourced test implementations for speech intelligibility.

Poster Presentation:
Thu, 18 April, 08:20 - 10:20
Poster Zone 5A

ICASSP 2024 KOREA

Diagnostic Rhyme Test

- Word pairs differ in one phoneme covering various distinctive features.
- Performance can be analysed overall, by feature, by phoneme.
- Repeatable: no memory effects.
- Independent of semantic context, but limited to single words.
- Scalable: duration and difficulty suitable for crowdsourcing.
- Noise can be added to make the task harder and amplify differences.

Distinctive Feature	Sample-1			
	VWL	CONS	Present	Absent
Voicing	ɑ	b/p (i)	BOND	POND
	ʊ	z/s (i)	ZOO	SUE
	ɛ	v/f (f)	REV	REF
	ɪ	v/f (f)	SHEAVE	SHEAF

Audio Data

- Audio data crowdsourced, annotated, quality control, meta data.
- 5 languages available: English, German, Spanish, French, Chinese (consonant and tonal) (Arabic and Japanese in the pipeline).
- 6 recordings/speakers per word.
- Variety of voices and accents.

Crowdsourced DRT Implementation

Easy test preparation:

1. Process audio base data, upload to host, add URLs to meta file.
2. Configure test and run script.
3. Upload QSF file to Qualtrics.
4. Distribute link (internal or via crowdsourcing).

Figure 3. View of a typical test item from the main task.

Participant pre-filtering:

e.g., first language, normal hearing, no dyslexia, high approval rate, etc.

Survey Flow:

Figure 2. Survey Flow

Option: disqualify participant based on attention checks.

Incentive: participation reward and bonus for best performance.

Results

Results download: as CSV from Qualtrics (manual and API).
Post-filtering: questionnaire responses, hearing test, attention checks.
Score calculation ($\text{percent}_{\text{correct}}$):

$$P(c) = \frac{[R - W]}{R + W} * 100$$

where R is the number of correct and W the number of incorrect responses [1]. Scores are calculated per test item.

References

[1] ITU-T Rec. P.807, "Subjective test methodology for assessing speech intelligibility," 2016.
[2] L. Lechler and K. Wojcicki, "Crowdsourced multilingual speech intelligibility testing," in *Proc. ICASSP*, Seoul, South Korea, 2024, pp. 1441–45.

Results

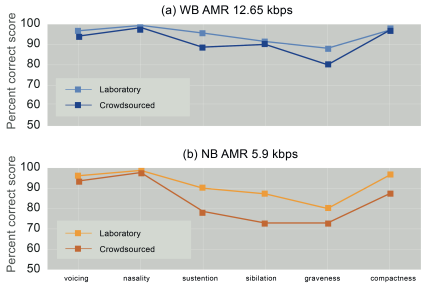

Figure 5. Comparison of two codecs in laboratory [1] and crowdsourced [2] conditions.

Our experiments showed expected patterns of intelligibility degradation from applying narrow-band (NB) codecs:
e.g., no degradation where distinguishing information is below 4kHz (i.e., tone, voicing, nasality).

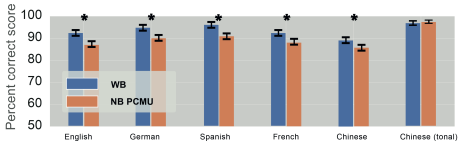

Figure 4. Overall intelligibility scores for several languages. Significance level $p<0.05$.

- The above is true across languages and features.
- The crowdsourced results are in good agreement with lab tests.
- The approach showed good re-test consistency in repeated experiments.

Conclusions

We present in our open-source release:
- Multilingual audio data.
- Tools to create surveys on Qualtrics.
- Tools to analyse the results.
- Documentation, tutorials, experimental results.

Our paper demonstrates good accuracy and consistency of the approach. **We invite the research community to collaborate with us!**

Online Resources

GitHub

Paper

Supplementary Materials