A semi-spontaneous Dutch speech dataset for speech enhancement and speech recognition

Dimme de Groot, Yuanyuan Zhang, Jorge Martinez, Odette Scharenborg
Delft University of Technology

Automatic speech recognition (ASR) systems are widely used in diverse scenarios, including emergency centres, health care services, and education. To ensure reliable performance in practice, it is crucial that ASR systems are able to recognize speech from diverse speakers and under diverse acoustic conditions, including noisy environments with background talkers. However, most speech datasets for developing or evaluating noise-robust ASR and speech enhancement (SE) algorithms consist of synthetic noisy speech, created by mixing clean recordings with real or artificially generated noise. Despite their obvious usefulness, such synthetic noisy speech does not fully capture the complex acoustic characteristics of real noisy speech.

Most publicly available Dutch speech datasets, such as the Corpus Gesproken Nederlands (CGN), Jasmin-CGN, and the Dutch portion of Common Voice 12, were recorded in quiet environments, occasionally with a few noisy samples. Consequently, the performance of state-of-the-art (SotA) ASR systems on Dutch is typically only evaluated on clean speech. For example, OpenAI's Whisper models were evaluated on Dutch speech from Jasmin-CGN and Common Voice 12. Meta's Wav2vec 2.0 model was evaluated on Jasmin-CGN, while two Dutch ASR models trained from scratch were evaluated on Jasmin-CGN.

To answer the open question of how well SotA ASR systems perform on real noisy Dutch speech, we designed and collected DRES. DRES is a freely available 1.5-hour Dutch realistic elicited (semi-spontaneous) speech dataset from 80 speakers recorded in noisy, public indoor environments with background talkers. The speech was recorded with a four-channel linear microphone array and was elicited by requesting the speakers to complete two tasks:

1. Draw a picture card and tell a short story fitting the drawn card,
2. Draw a topic card and shortly talk about that topic.

Using DRES, we evaluated the ASR performance of eight SotA off-the-shelf ASR models developed by Google, Microsoft, Meta, OpenAI, NVIDIA, and the authors. We found that five out of the eight ASR models achieve word error rates (WERs) lower than 22% on DRES, despite the challenging conditions. Additionally, we also evaluated the speech quality and ASR performance after applying six well-known single-channel SE algorithms. In contrast to recent work, we did not find a positive effect of modern single-channel SE on ASR performance, emphasizing the importance of evaluating in realistic conditions.

This work has been accepted for presentation at IEEE SLT 2026.

References

Zhang, Y., De Valck, T., & Scharenborg, O. (2026). Speech recognition performance disparities between Dutch diverse speaker groups. Phonetica, (0). https://doi-org.tudelft.idm.oclc.org/10.1515/phon-2025-0061

Fuckner, M., Horsman, S., Wiggers, P., & Janssen, I. (2023, October). Uncovering bias in ASR systems: Evaluating wav2vec2 and whisper for Dutch speakers. In *2023 International conference on speech technology and human-computer dialogue (SpeD)* (pp. 146-151). IEEE. 10.1109/SpeD59241.2023.10314895