الكلية متعددة التخصصات بآسفي
ⵜⴰⵖⵉⵡⴰⵏⵜ ⵜⴰⴳⵜⵉⵣⵍⵉⵜ ⵏⵓⵙⴼⵉ
Faculté Polydisciplinaire de Safi
MARRAKECH
جامعة القاضي عياض
UNIVERSITÉ CADI AYYAD

Six papers - five themes - one field under recalibration

Automatic Speech Recognition

Supervised by

Dr. soufiane Hourri

Presented by

Elaichi Imane
Landaoui Maroua
Chakrane Fouziya
MOhammed Yassine Nouri

Année universitaire : 2026/ 2027

OVERVIEW

Five research themes

How the six papers fit together: from measuring ASR, to making it robust, small, multi-speaker and adaptable

01	HALLUCINATIONS **HALAS + Whisper-CD**	Detect them on real speech (dataset) · reduce them at decoding time
02	EVALUATION & BENCHMARKING **Open ASR Leaderboard**	One reproducible yardstick: 86 models × 11 datasets, WER + RTFx
03	MODEL COMPRESSION **Whisper PTQ**	80+ INT / FP quantization configs for on-device Whisper
04	OVERLAPPED SPEECH **Shuffles**	Single-pass, speaker-attributed transcription and alignment
05	DOMAIN ADAPTATION **Multi-View Text Denoising**	Adapt LLM-based ASR with text only, without forgetting

HALAS: human-annotated ASR hallucinations

Barański, Jasiński, Bartolewska, Witkowski, Kowalczyk · AGH University of Krakow

1 · BACKGROUND / CONTEXT

ASR models sometimes invent words that were never spoken.

2 · PROBLEM / RESEARCH GAP

Previous methods were mainly tested on artificial audio, not real speech.

3 · OBJECTIVE

Create a public dataset to study and detect hallucinations in real speech.

4 · METHODOLOGY

- **3,611 audio clips and 7 ASR models.**
- **10 human annotators labeled hallucinations.**
- **Two annotators per clip, with a third resolving disagreements.**

5 · RESULTS & CONCLUSION

56.1%
F1 of best detector (Whisper v3)

0.81
ROC-AUC of CER as a proxy

- All 7 models produced hallucinations in the selected audio clips.
- Different models often invented similar short phrases, such as “thank you”.
- Detecting hallucinations remains difficult, even with advanced methods.

Whisper-CD: multi-negative contrastive decoding

Ahn, Chae, Park, Shim · Sungkyunkwan University

1 · BACKGROUND / CONTEXT

Whisper can repeat words, invent text during silence, or omit spoken content in long recordings.

2 · PROBLEM / RESEARCH GAP

Existing methods often target one error type or require retraining. Contrastive decoding had not been applied to ASR hallucinations.

3 · OBJECTIVE

Reduce transcription errors without retraining Whisper.

4 · METHODOLOGY

Compare original audio with three modified versions: noise, silence, and shifted audio. Test on five English datasets.

5 · RESULTS & CONCLUSION

−24.3 pp
Maximum WER reduction on CORAAL

+48%
tokens/s vs beam search

- Repetition loops and hallucinations are reduced.
- Combining all three modified audio inputs works better overall than using just one.
- The method is faster than beam search in the reported comparison, but adds some computation compared with standard decoding.
-

THEME 2 · EVALUATION & BENCHMARKING

Open ASR Leaderboard

Srivastav, Zheng, Bezzam et al. · Hugging Face, NVIDIA, Univ. of Cambridge, Mistral AI, OpenAI

1 · BACKGROUND / CONTEXT

Many ASR models and tools exist, but comparing their accuracy and speed is difficult.

2 · PROBLEM / RESEARCH GAP

Existing benchmarks often focus on English and short audio. WER alone is not enough to compare model performance.

3 · OBJECTIVE

Create an open platform to compare ASR models in accuracy and speed across languages and audio lengths.

4 · METHODOLOGY

- **86 models from 26 organizations.**
- **11 datasets.**
- **Three evaluation tracks: English short-form, multilingual, and English long-form.**
- **Two metrics: WER (accuracy) and RTFx (speed).**

5 · RESULTS & CONCLUSION

86
models compared on 11 datasets

3,386
RTFx of Parakeet TDT 0.6B v2

- Conformer encoders + LLM decoders give the lowest WER; CTC / TDT decoders are far faster
- Closed models lead on long-form (9.05% vs 12.2% best open)
- Wider language coverage coincides with higher English WER

THEME 3 · MODEL COMPRESSION

Systematic PTQ study of Whisper

Choi, Lee, Kim, Lee · LG Electronics, Sogang University

1 · BACKGROUND / CONTEXT

- Whisper converts speech into text.
- Its large size limits its use on mobile devices.

2 · PROBLEM / RESEARCH GAP

How can Whisper be compressed without significantly reducing accuracy?

3 · OBJECTIVE

- Compare integer and floating-point quantization formats.
- Find the best balance between model size and accuracy.

4 · METHODOLOGY

- Test Whisper tiny.en and base.en.
- Evaluate more than 80 quantization configurations.
- Use LibriSpeech and measure WER.

5 · RESULTS & CONCLUSION

6.4×
smaller with NVFP4 W4A16

+0.07 pt
WER vs FP32 (base.en)

- Activation precision strongly affects accuracy.
- MXFP4 and INT3 give poor results.
- base.en performs better than tiny.en above ~36 MB at equal memory budget.

Modeling overlapped speech with shuffles

Wiesner, Cornell, Polok, Ondel Yang, Burget, Khudanpur · JHU, LISN-CNRS, CMU, Brno Univ. of Technology

1 · BACKGROUND / CONTEXT

- Multiple speakers can talk at the same time.
- It is difficult to recognize the words and identify who said them.

2 · PROBLEM / RESEARCH GAP

How can we transcribe overlapping speech and identify each speaker?

3 · OBJECTIVE

Develop a method to transcribe and align speech from multiple speakers.

4 · METHODOLOGY

- Use the WavLM model.
- Test the method on simulated speech overlaps using LibriSpeech.
- Use shuffles to handle different word orders across speakers.

5 · RESULTS & CONCLUSION

18.5%

tcpWER on Synth (κ = 4 s) vs 36.6 for t-SOT

69 ms

boundary error, 2-speaker alignment

- Results comparable to SD-CTC.
- Enables single-pass alignment of overlapping speech.
- Performance decreases with three speakers, and memory usage is high.

THEME 5 · DOMAIN ADAPTATION

Text-only adaptation of LLM-based ASR

Burdisso, Villatoro-Tello et al. · Idiap Research Institute, EPFL, Univ. of Zurich, Uniphore, Brno Univ. of Technology

1 · BACKGROUND / CONTEXT

LLM-based ASR uses a speech encoder and a language model. Adapting it requires large audio-text datasets.

2 · PROBLEM / RESEARCH GAP

Text-only fine-tuning can make the model forget how to recognize speech.

3 · OBJECTIVE

Adapt ASR to new domains using text only, without losing speech recognition ability.

4 · METHODOLOGY

Train the model to correct noisy text using a mix of audio-text pairs and corrupted text from source and target domains.

5 · RESULTS & CONCLUSION

25.4%
best relative WER gain (cross-domain)

+29–63%
WER if the audio view is removed

- Gains: 18.6–18.9% in-domain, 4.2–7.0% out-of-domain, 20.6–25.4% cross-domain
- Beats the Fang et al. and Ma et al. text-only baselines
- Still behind audio-based adaptation (Banking 6.53 vs 4.55% WER)

OUR CHOICE · WHY HALAS

Four reasons we chose HALAS

Relevant, well-founded and still open: a good paper to study in depth

01

REAL-WORLD PROBLEM

Hallucinations hurt trust

ASR models can write words nobody said, and different models often invent the same short phrases, such as “thank you”.

02

CLEAR RESEARCH GAP

Tested on real speech

Earlier methods were mainly tested on artificial audio; HALAS offers a public, human-labeled dataset of real speech.

03

RIGOROUS METHOD

Clear and verifiable

3,611 audio clips, 7 ASR models and 10 annotators: two labels per clip, with a third resolving disagreements.

04

OPEN CHALLENGE

Room to go further

The best detector reaches only 56.1% F1, and the paper pairs naturally with Whisper-CD, which mitigates hallucinations.

Numbers as reported in the source papers · Interspeech 2026

CONCLUSION

Six papers, one choice

ASR is being recalibrated: reliability, evaluation, efficiency, robustness and adaptability

1 · THE BIG PICTURE

Six papers cover five themes: hallucinations, evaluation, compression, overlapped speech and domain adaptation.

2 · OUR CHOICE

HALAS offers a public, human-annotated dataset of real speech; detection is still hard (56.1% F1).

3 · NEXT STEP

Study HALAS in depth: the dataset, the annotation protocol and the detection results.

IN NUMBERS

6 papers reviewed

5 research themes

1 paper chosen: HALAS

Thank you

Questions?

Numbers as reported in the source papers · Interspeech 2026