Automated Pathological Speech Assessment with a focus on Large Language Models

Bence Mark HALPERN (Presenter)
Assistant Professor
OSAKA University (SANKEN)

2026. 11. 09 – APSIPA 2026 Winter School - **Trustworthy Speech Foundation Models and LLM-Augmented Modeling for Pathological Speech**

About me

- **2014–2018** **MEng Biomedical Engineering**, Imperial College London
- **2018–2022** **PhD researcher**, Netherlands Cancer Institute (Scharenborg, van Son)
  - **Thesis title**: Making Speech Technology Accessible for Pathological Speakers
- **2019–2023** **Research Consultant**, Oxford Wave Research
  - **Topics**: deepfake detection, speech enhancement
- **2023–2024** **Visiting Assistant Professor**, Nagoya University (Tomoki Toda)
  - **Topics**: controllable pathological speech synthesis
- **2024–2026** **Specially Appointed Assistant Professor**, Nagoya University
  - **Topics**: pathological speech assessment
- **2026–now** **Assistant Professor, Osaka University, SANKEN** (Komatani)



Awards & grants

- **Investigator,** KAKENHI: Development of Reliable and Controllable Speech Media Generation Technology and Expansion of International Research Activities
- **Principal Investigator**, NWO (Dutch Research Council): “I don’t sound like myself”: Creating voice conversion-based speech technology for Healthcare
- **Best Workshop Paper**, 6th ACM International Conference on Multimedia in Asia Workshops

Outline (40 mins + 5 min Q/A)

- **Pathological speech assessment task (5 min)**
  - Background of the task
- **LLM-based reformulations (30 min)**
  - ① Black-box assessment → LLM as a judge
  - ② Anomaly detection-based assessment → LLM-based anomaly detection
  - ③ Automatic speech recognition-based assessment → error-corrected ASR
  - And there could be more to talk about (prompt engineering, autoresearch, …)
- **Open problems (5 min)**
  - What is left to solve?

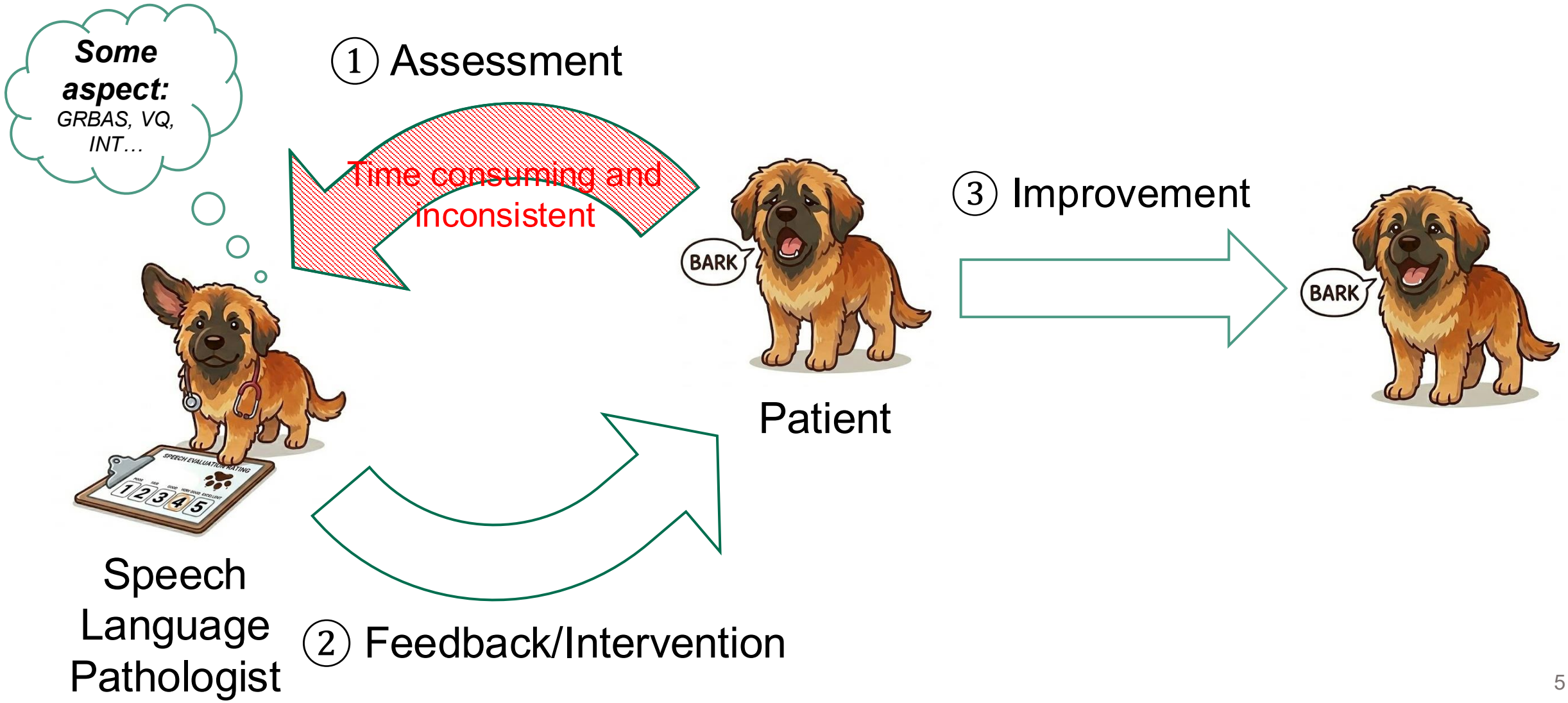
What is pathological speech?

Pathological speech: Speech that deviates from typical due to a disorder.

Typical voice	Voice	Articulation	Language
Sigmatism		strong	
Cleft Lip and Palate	*mild*	strong	
Stuttering	strong	strong	
Alaryngeal voice	strong	*mild*	
Oral cancer	*mild*	strong	
Dysarthria 1x 2x	*mild*	*mild*	*mild*
Aphasia & Dementia	*mild*	*mild*	*mlld*

Table 1. Categorization of different speech disorders depending on what part of the speech production process they affect. Speech samples from COPAS [Noth2018]

An overview of pathological speech assessment in clinic



What aspect of the speech?

- **Focus: intelligibility** which is the communicative ability (more later).
- Reason: Disorder-specific details give diminishing returns. [Halpern2025+]



Language

out of scope today
(LLMs do help there)

[Halpern2025+] Halpern, B.M., Tienkamp, T., Rebernik, T., van Son, R.J.J.H., Wieling, M., Abur, D., Toda, T. (2025) Relationship between objective and subjective perceptual measures. Interspeech 2025.

02

Black-box approaches: LLM-as-judge

Black-box formulation of the task

Reference-Free

Score

Model

Pathological Recording



Patient

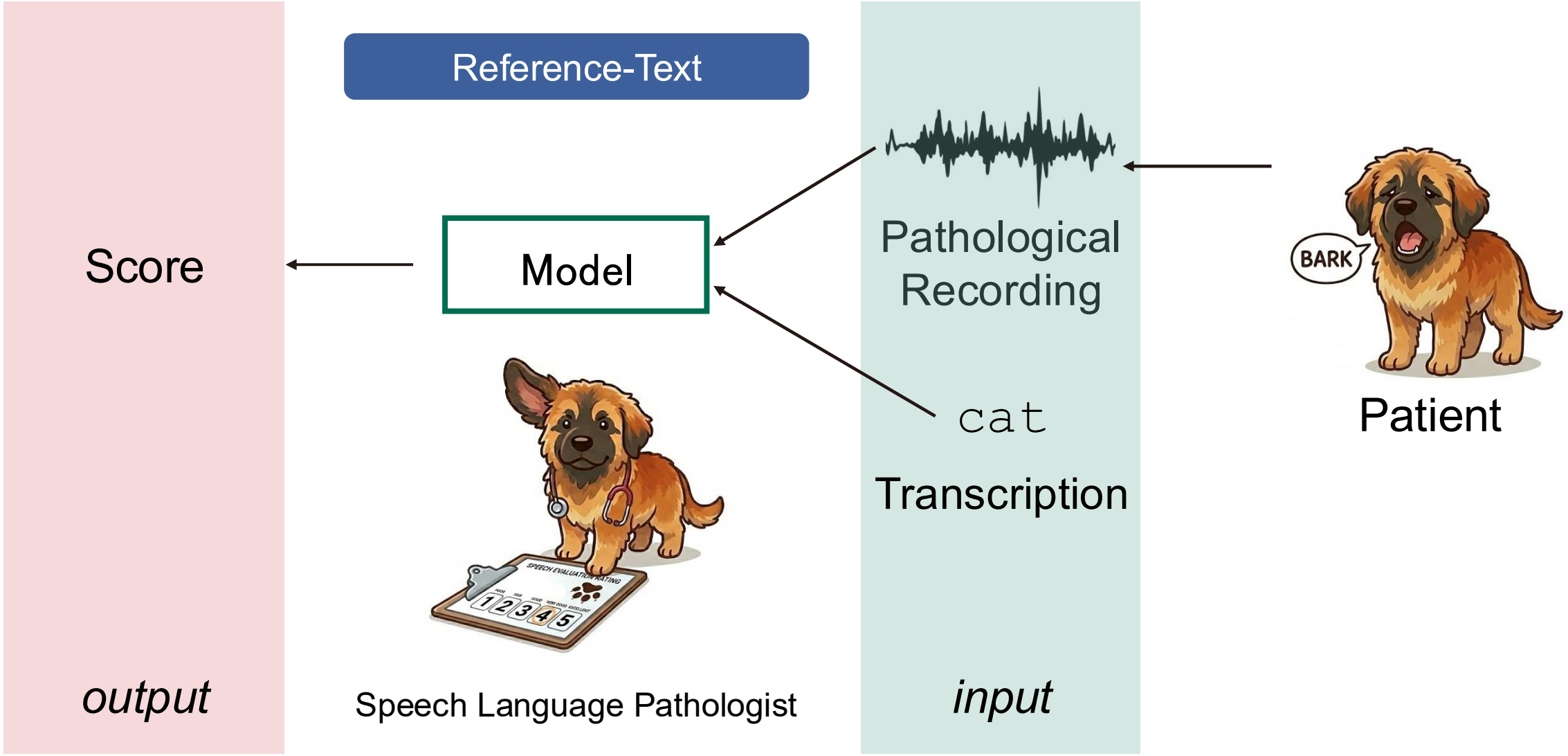


Speech Language Pathologist

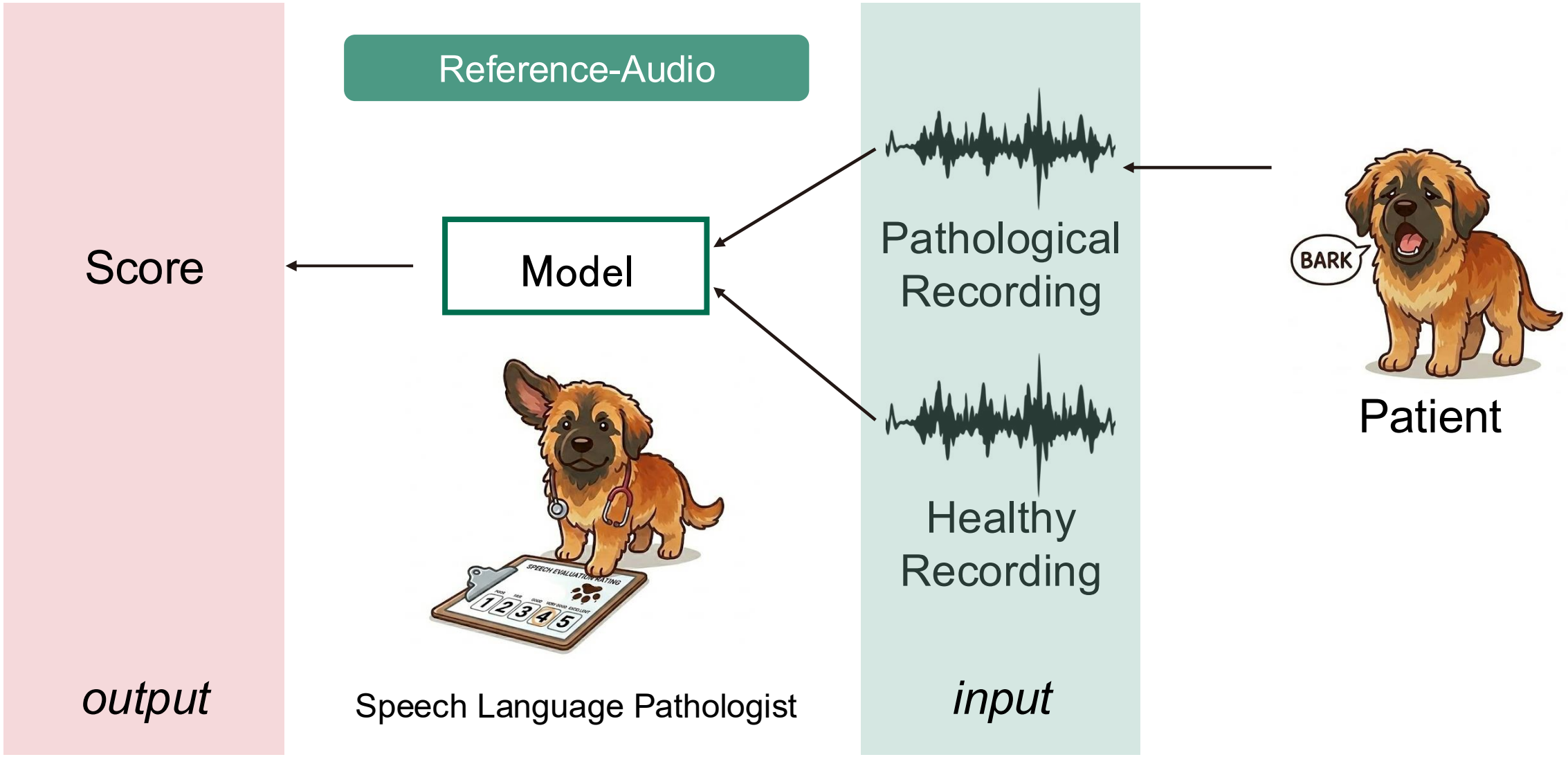
output

input

More complicated black box formulation

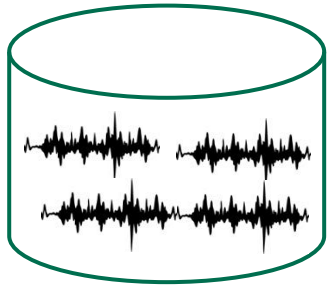


More complicated black box formulation



Standard: Supervised Machine Learning

Training



Healthy + Pathological Corpus

Feature

Classifier / Regressor

1

Ground Truth Score

copy

Performance is not perfect
→ LLMs are worth exploring

Inference



Pathological Speech

Feature

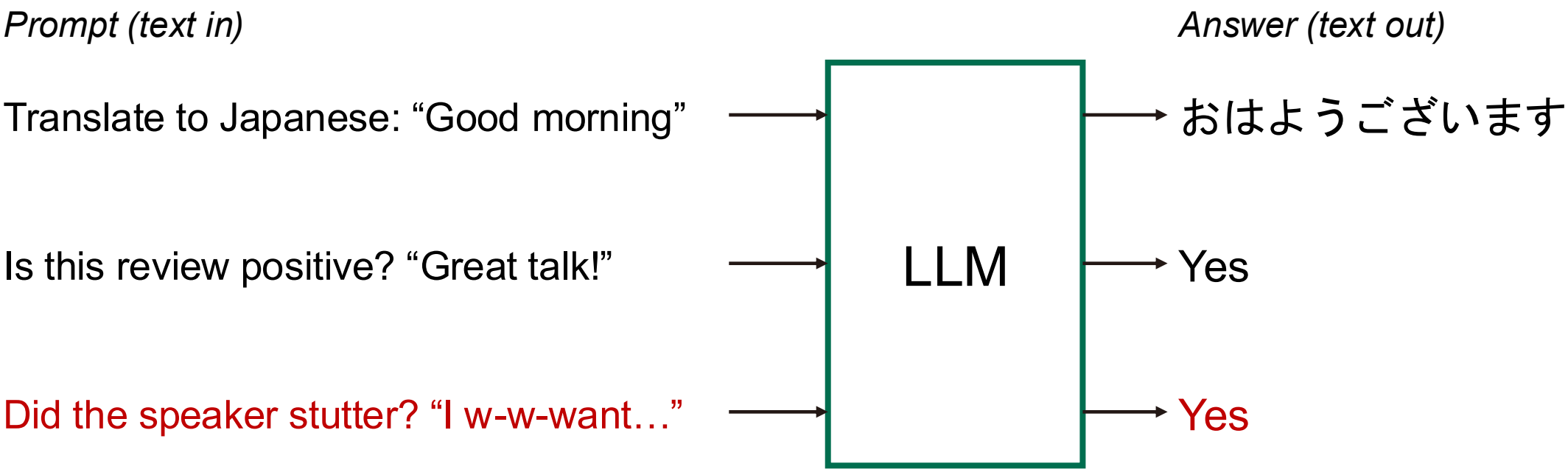
Classifier / Regressor

3

Predicted Score

So where does LLM come into the picture?

- **Universal interface:** any task phrased as text in, text out.
- The prompt defines the task, not the model.



Reference-free reformulation with LLMs (Outline)

Text-based inference

Example prompt (illustrative)
Here is an ASR transcript of a speaker with a speech disorder:
"the burt iss green"
Rate their intelligibility from 0 (unintelligible) to 1 (fully intelligible). Reply with a number.



Pathological Speech

ASR

ASR output: "the burt iss green"

Text-based LLM

Output text

"I'd say 0.6."

Parser

0.6

Reference-free reformulation with LLMs (Outline)

Speech-based inference

Example prompt (illustrative)
You will hear a speaker with a speech disorder:
<audio>
Rate their intelligibility from 0 (unintelligible) to 1 (fully intelligible). Reply with a number.



Pathological Speech

Speech–based LLM

Output text

"I'd say 0.6."

Parser

0.6

So, how do black-box LLMs perform on the task?

- In the speaker-level black-box LLM catches up to PathBench SotA.
- However, the gap in the harder, utterance-level gap is still considerable.

Group	Metric	Speaker r (50 spk)	Utterance r (793 utt)
Reference Free	ASRIC	**0.86**	0.39
	DArtP	0.79	0.37
	Confidence	0.78	0.35
Reference Text	PER	0.85	**0.43**
	ArtP	0.78	0.40
Speech-LLM (Qwen2.5-Omni-7B)	ref-free	0.84	0.19
	ref-text	**0.86**	0.19
	ref-audio K=1	0.77	0.21

Table 2. Comparison of PathBench and Speech-LLM evaluators on the speaker and utterance-level NeuroVoz dataset.

[HalpernPrelim] Halpern, B. M., Tienkamp, T., Abur, D., & Toda, T. Preliminary results from the PathBench journal paper.

In-context learning: the idea

- **Idea:** adapt an LLM to a new task without retraining it.
- **How:** put labelled examples (exemplars) into the prompt.

Rate how accurately the speaker pronounces the text, from 0 (worst) to 10 (best).

Text: HE LEFT THE FRUIT STAND
Audio: <audio 1>
Pronunciation accuracy: 9

Text: MARK IS GOING TO SEE ELEPHANT
Audio: <audio 2>
Pronunciation accuracy:

← **exemplar**: recording + its human score

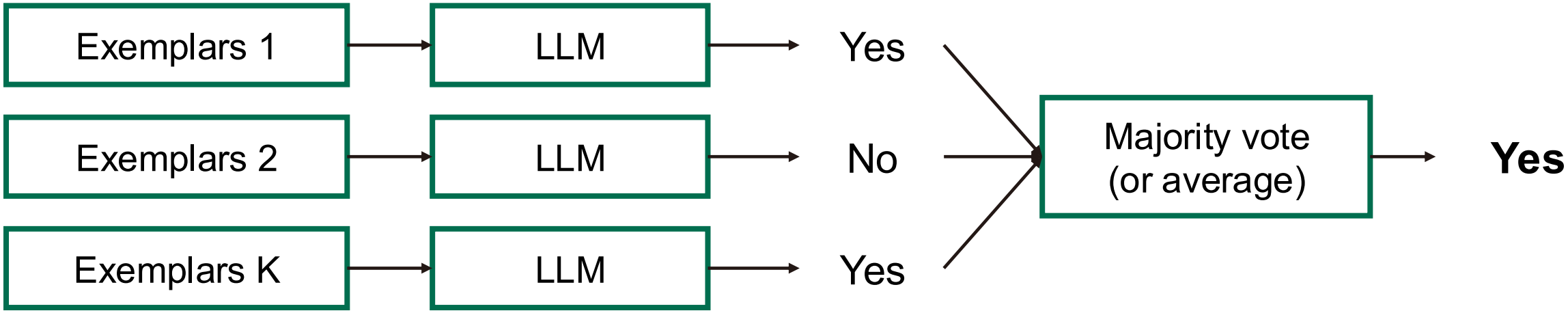
← **test recording**: the LLM fills in the score

Example prompt for pronunciation assessment, shortened from [Wong2025+].

[Wong2025+] Wong, J. H. M., Huzaifah, M., Chen, N. F., & Aw, A. T. (2025). Speech in-context learning of paralinguistic tasks.

Using more examples: prompt ensembling

- **Problem:** the prompt length limits how many exemplars fit.
- **Idea:** run in-context learning K times with different exemplars, combine the answers.



[Wong2025+] Wong, J. H. M., Huzaifah, M., Chen, N. F., & Aw, A. T. (2025). Speech in-context learning of paralinguistic tasks.

Does prompt ensembling help?

- **Yes:** combining 5 runs beats the average single run on both tasks.
- Still a gap between supervised topline they have tested.

Task	Metric	Single run	5-run ensemble	Supervised
Stutter detection	F1 ↑	0.70	0.73	≈0.76
Pronunciation assessment	r ↑	0.26	**0.41**	0.75*
Pathological speech assessment	**?**	**?**	**?**	**?**

Table 3. Single run = mean of 5 runs. Supervised = stand-alone model trained on the adaptation data. 31 exemplars in total. [Wong2025+]

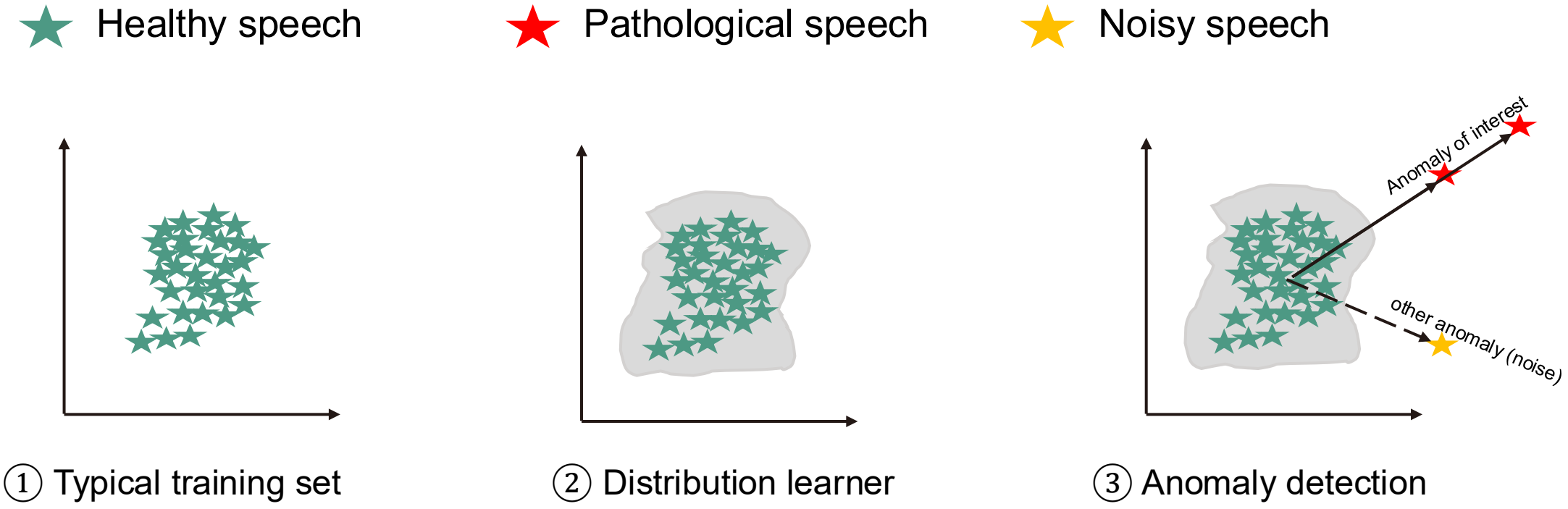
[Wong2025+] Wong, J. H. M., Huzaifah, M., Chen, N. F., & Aw, A. T. (2025). Speech in-context learning of paralinguistic tasks.

03

Anomaly detection as an approach

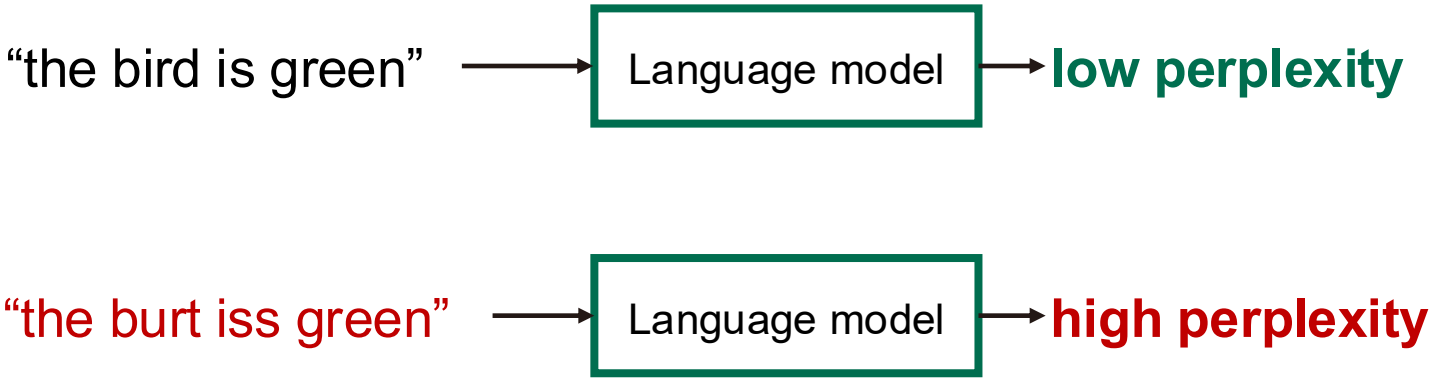
Another approach: Anomaly detection

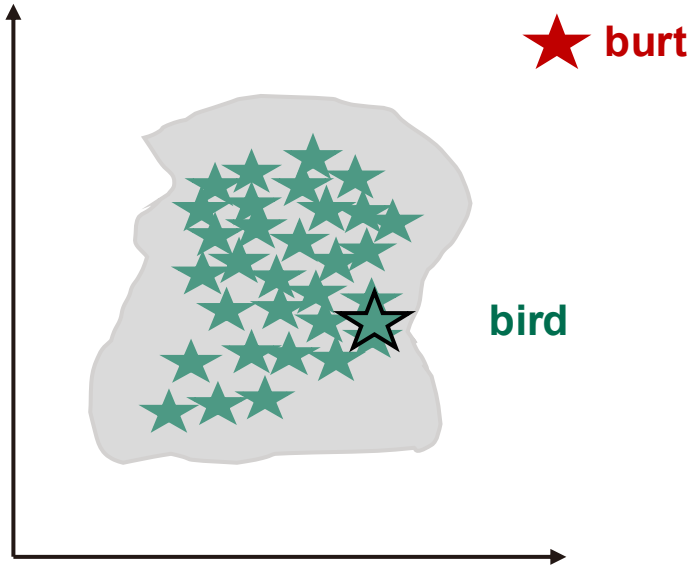
- **Idea:** learn what typical speech looks like, flag what deviates.
- **Reference-free:** needs only typical speech, no transcript.



A language model is a distribution learner

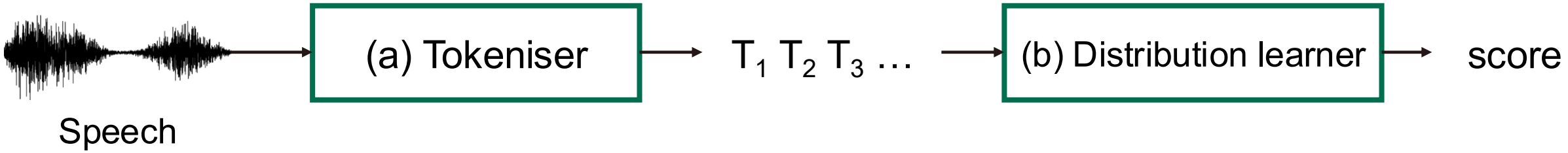
- So far: language models as a **universal text interface**.
- But they are also a **probability distribution** fitted to typical text.
- **Perplexity:** how well new text fits that distribution.





Schematic: tokeniser + distribution learner

- Every method makes these two choices; ours in green.



E.g. words in normal LLM setting
Ours: HuBERT units (k-means)

E.g. gpt-4.1
Ours: LSTM trained on typical speech (LibriLight)
~ no resource to train LLM at that time ☹

[Halpern2024+] Halpern, B. M., & Toda, T. (2024). Reference-free automatic speech severity evaluation using acoustic unit language modelling.

Results with acoustic language modelling

Feature	NKI-SpeechRT	NKI-OC-VC
Reference Free		
Shimmer	0.15	-0.13
σF0	-0.17	0.32
Jitter	0.13	0.45
WADA SNR	-0.29	-0.64
Voicing%	0.03	-0.18
HNR	-0.30	0.14
CPP	-0.16	-0.27
SpeechLMScore	**0.38**	**0.69**
Reference-Based		
Phoneme Error Rate	-0.82	-0.92

- **Reference-free:** needs only typical speech, no transcript.
- **Dataset-dependent:** r=0.69 for oral cancer, r=0.38 after chemoradiotherapy.
- **Still well below reference-based approaches.**
- Probably different tokenizer and distribution learner choices would have driven up performance.

Table 4. Pearson's r of listener scores with automatic scores. [Halpern2024+]

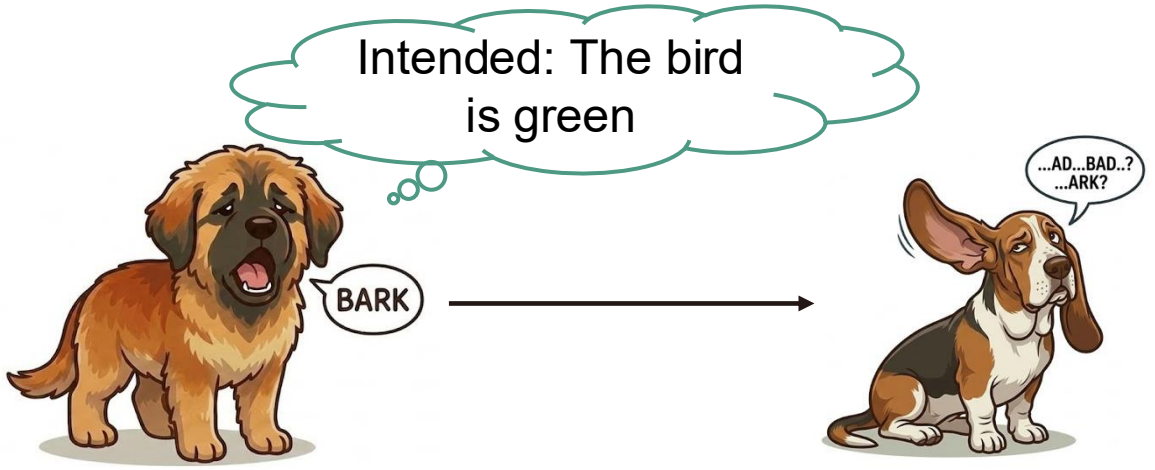
[Halpern2024+] Halpern, B. M., & Toda, T. (2024). Reference-free automatic speech severity evaluation using acoustic unit language modelling.

04

Automatic speech recognition-based assessment

What is intelligibility?

- One way to understand intelligibility is to define it by the number of words the listener gets right.



Speaker → Human listener → the **burt** is green

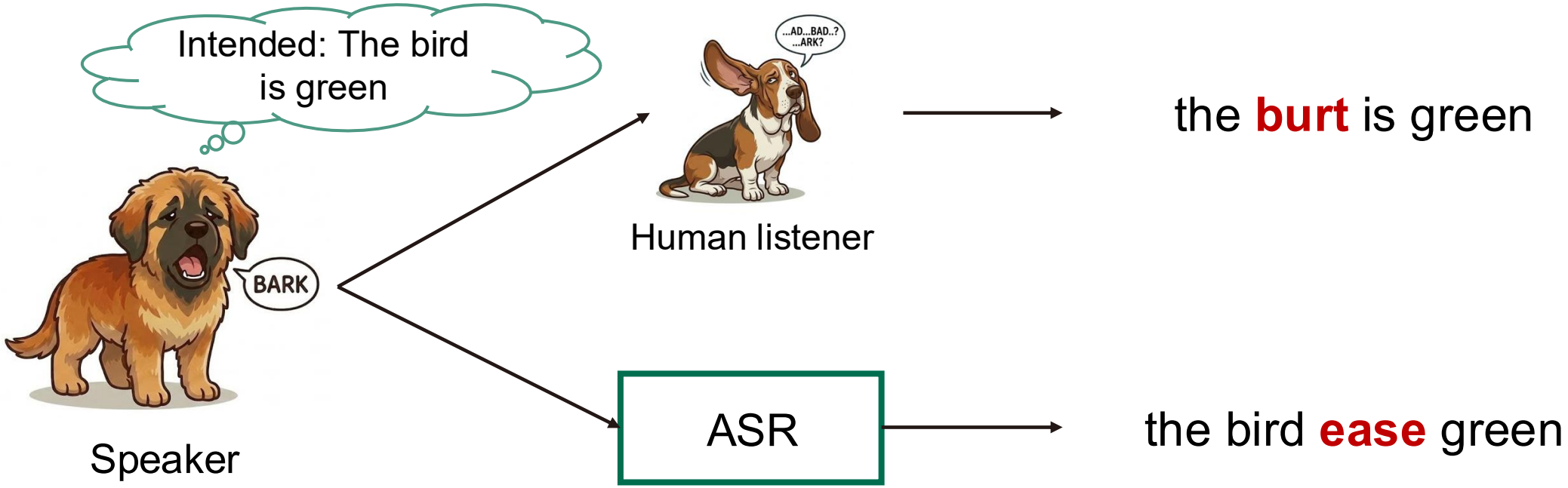
written down

3 / 4 words correct → 75 %

- This measure will highly correlate with people's perception of intelligibility they would answer in a questionnaire. [Abur2019+]

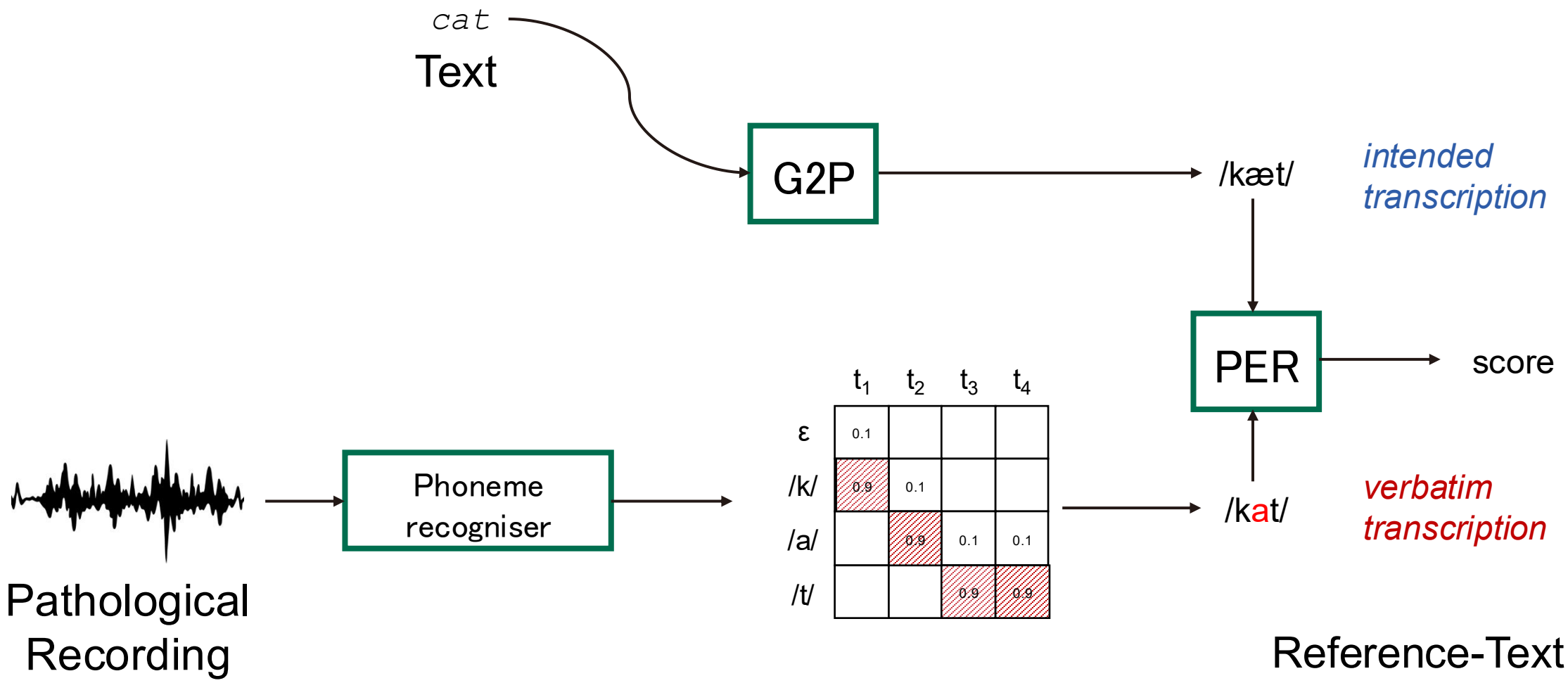
Can a machine be the listener?

- Replace the human listener with an ASR → automatic intelligibility.

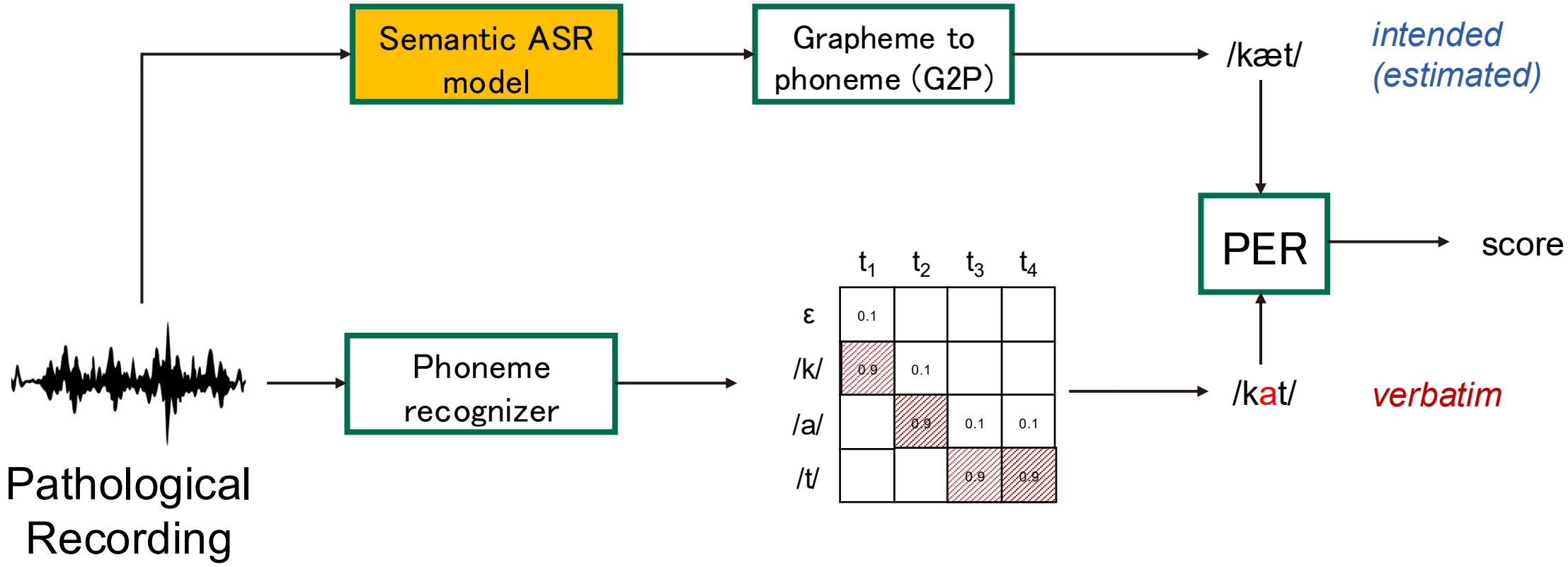


The ASR **does not neccessarily make the same mistakes** but the number of mistakes it makes tends to depend on the intelligibility of the speaker.

Phoneme Error Rate (PER)



ASR Inconsistency Score



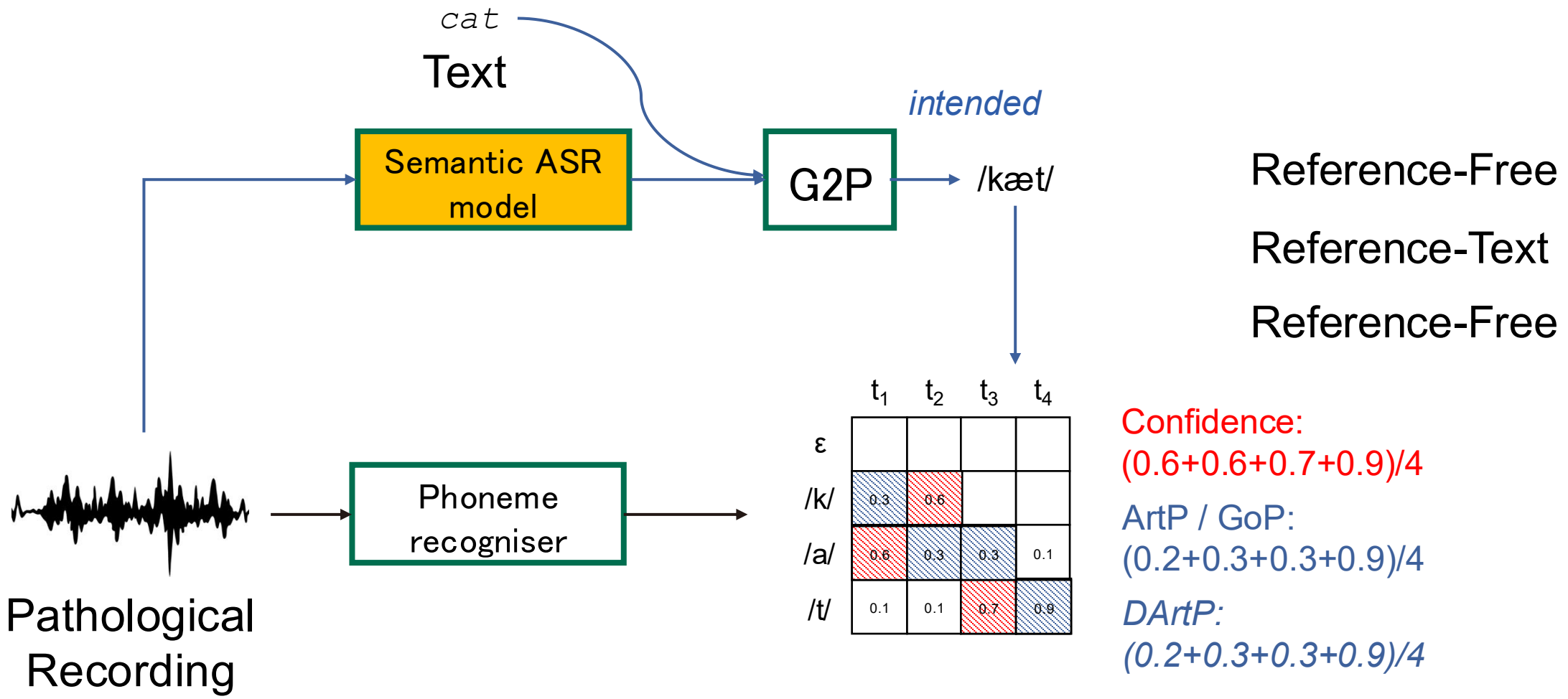
[Halpern2026+] Halpern, B. M., Tienkamp, T., & Abur, D. (2026). Towards explainable reference-free speech intelligibility evaluation of people with pathological speech.

Disadvantage of evaluating on phonemes

- Resolution is limited by the number of tokens (phonemes) as there is a finite number of tokens.
- Phoneme error rate can be larger than 1 meaning that it might be challenging to calibrate it in a practical scenario.

Predicted	/kæt/	/kat/	/kat/	/kat/	/kaat/
Reference	/kæt/	/kæt/	/kæt/	/kæt/	/kæt/
Intelligibility	0	0.33	0.66	1	1.33

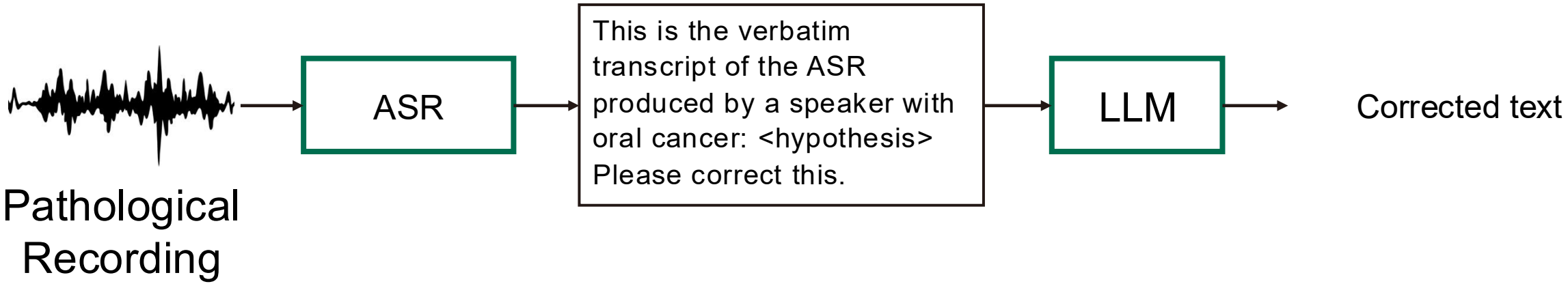
Joint Probability Estimation Methods



[Talkar2025+] Talkar, T., Kawabata, K., Higgins, C., & Tobyne, S. (2025). Development and validation of a Wav2Vec 2.0-based cross-language methodology for measurement of articulatory precision.

What is a semantic ASR model?

- **Goal:** Predict the intended transcription instead of the verbatim.
- **How we implement**: error-correction of ASR model



ASR: he grasped the little **passing** ← *verbatim transcription*

LLM: he grasped the little parcel ← *intended transcription (estimate)*

Error-correcting ASR with LLMs: n-best

- **n-best:** the LLM sees the top-n ASR hypotheses, not just one.

zero-shot uncon

Input: Perform error correction on the top3 outputs generated by an Automatic Speech Recognition (ASR) system. The ASR hypotheses, listed in order of their ASR posterior score, are as follows:
<hypothesis1> he grasped the little passing </hypothesis1>
<hypothesis2> he grasped the little parcel </hypothesis2>
<hypothesis3> he grasped the little puzzle </hypothesis3>

Please provide the corrected top1 ASR transcription of the given utterance only, do not add any explanations or other words.

ChatGPT: he grasped the little parcel

zero-shot constr

Input: Perform language model rescoring based on the top3 outputs generated by an Automatic Speech Recognition (ASR) system. The ASR hypotheses, listed in order of their ASR posterior score, are as follows:
<option1> he grasped the little passing </option1>
<option2> he grasped the little parcel </option2>
<option3> he grasped the little puzzle </option3>

Please output the selected top1 ASR transcription as <option?> The selected top1 ASR transcription </option?>.

ChatGPT: The selected top1 ASR transcription is <option2> he grasped the little parcel </option2>.

Unconstrained:
write a new transcript

Constrained:
pick one of the n (= rescoring)

[Ma2025+] Ma, R., Qian, M., Gales, M., & Knill, K. (2025). ASR error correction using large language models.

Constrained correction as anomaly detection

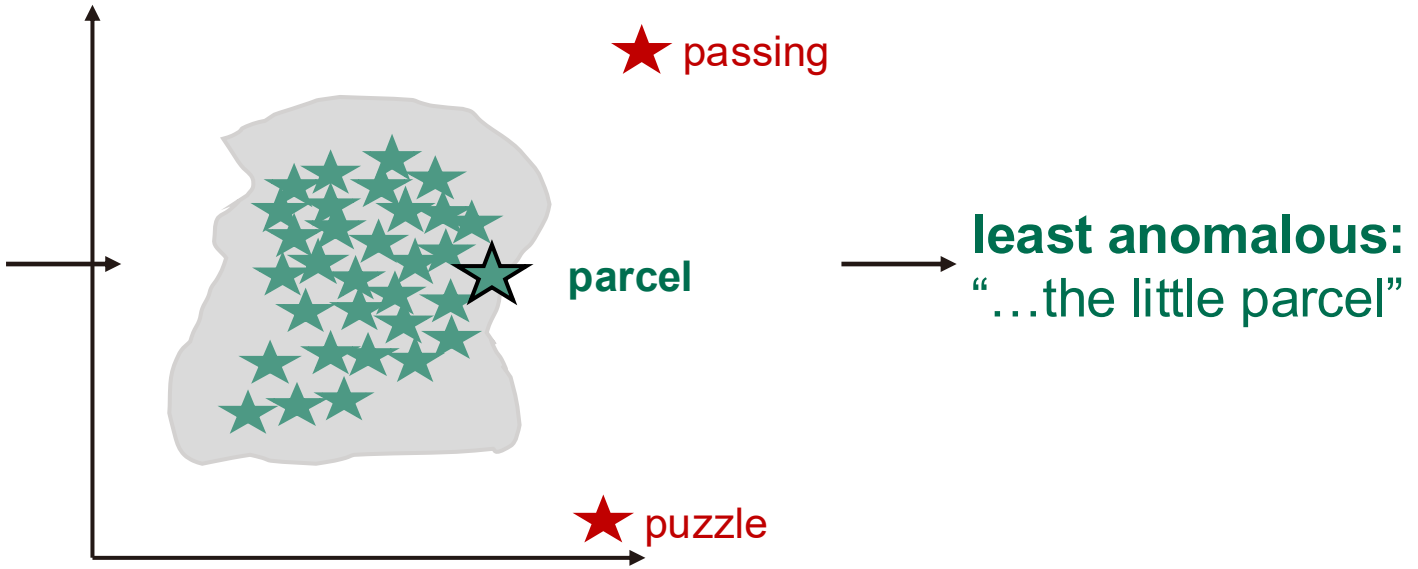
- **More principled:** the LM scores each n-best hypothesis.
- Keep the **least anomalous** hypothesis.

n-best hypotheses

★ ❶ he grasped the little passing

★ ❷ he grasped the little parcel

★ ❸ he grasped the little puzzle



How well error-correction performs?

Pearson's r	NKI-OC-VC	NKI-SpeechRT	NeuroVoz	TORGO
5-gram LM (beam search)	-0.86	-0.76	**-0.89**	-0.84
gpt-3.5-turbo (zero-shot unc)	-0.90	-0.84	-0.71	-0.94
gpt-4.1 (zero-shot unconstrained)	-0.89	-0.84	-0.76†	-0.94
WER (reference-based)	**-0.90**	**-0.85**	-0.82	**-0.95**

Table 5. Speaker-level Pearson's r with perceptual intelligibility; GPT rows are means over runs (s.d. ≤ 0.013). Bold: best per column; underlined: best reference-free (ranked on unrounded values); †: significant GPT difference ($p < 0.05$). [Halpern2026+]

- Overall, fairly high correlations on all datasets (not PathBench protocols!)
- N-gram based methods sometimes still outperform the LLM, however it was not entirely the same setup.
- In one case (NeuroVoz), error corrected beats reference-based evaluation.

So, did we manage to solve pathological ASR?

Metric	NKI-OC-VC	NKI-SpeechRT	NeuroVoz	TORGO
W_{greedy} WER	50.04%	41.62%	27.35%	57.81%
W_{LLM} WER	44.73%	33.79%	20.13%	54.70%
r (W_{LLM} WER, intell.)	-0.90	-0.80	-0.86	-0.94

Table 6. WER across all speakers by dataset; last row: per-speaker correlation of W_{LLM} WER with intelligibility. [Halpern2026+]

- Not all, LLM-corrected transcriptions are still far from correct.
- However, the extent of correctibility depends on severity → even **without correct reference** a useful intelligibility signal is possible.

Summary: Where did LLMs help?

- LLM's provide better error corrections for pathological speech than n-grams, at least more useful, for the downstream speech intelligibility task.
  - Explainable reference-free evaluation is closer now.
- LLM-as-Judge scoring does not outperform but match speaker-level SoTA.
  - There might be scenarios where it is easier to deploy LLM.

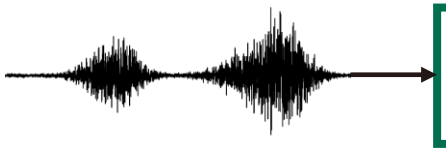
03

Open problems

Challenges of using LLMs for speech assessment

- Data contamination (in terms of error correcting ASR) [Tseng2025+]
  - Plagiarism avoidance only partly mitigates that problem, i.e., LLMs are reluctant to reproduce exact text [Ippolito2023+]
- Data contamination (in terms of testing samples)
  - Dataset could be leaked to the internet (less likely if not public). [Sainz2023+]
- Data governance
  - Frontier models means data sharing with companies, local models mean weaker performance

Map of problems to be solved



Noisy conversation
(SLP + patient)

① Diarisation

patient's speech only



② Speech Enhancement

Noisy Pathological Recording



Pathological Recording

cat

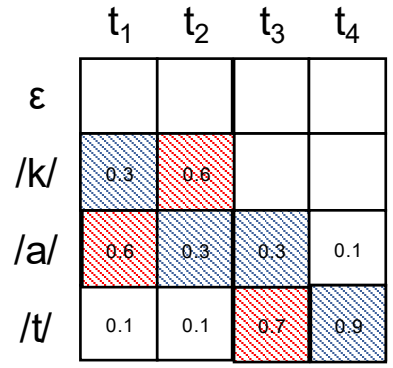
Text

④ Pathological ASR

⑤Grapheme to phoneme

/kæt/

③ Universal Phoneme Recognizers



⑥ Forced alignment

⑦ Representation (IPA? Wav2vec?)

Please contribute to speech assessment research!

- PathBench is an open-source project.
- As a starting researcher, it takes a long time to understand all settings, so we encourage you to use PathBench and leave it to those settings.
- We need to be serious about common reference points if we want to meaningfully advance science.
- Full journal paper and new protocols for PathBench are coming soon (utterance-level first).



Conclusion

- Large language models proved to be useful for error-correction based speech intelligibility assessment but they remain inferior to SotA on utterance-level (*until the next model update*).
- Real view on their performance is heavily distorted due to data compliance reasons (can't use frontier models), and possibility of data leakage.
- Several challenges remain until we can assess the speech effortlessly using speech technology.
- Slides are available upon request.

References

[Abur2019+] Abur, D., Enos, N. M., & Stepp, C. E. (2019). Visual analog scale ratings and orthographic transcription measures of sentence intelligibility in Parkinson's disease with variable listener exposure. American Journal of Speech-Language Pathology, 28(3), 1222-1232.

[Folkertsma2026+] Folkertsma, H., Tienkamp, T., de Visscher, S., Witjes, M., van Son, R., Guo, J., & Halpern, B. M. (2026). Improving automatic speech recognition for speakers treated for oral cancer using data augmentation and LLM error correction. arXiv preprint arXiv:2605.15854.

[Halpern2024+] Halpern, B. M., & Toda, T. (2024). Reference-free automatic speech severity evaluation using acoustic unit language modelling. In Proceedings of the 6th ACM International Conference on Multimedia in Asia Workshops (pp. 1-5).

[Halpern2026+] Halpern, B. M., Tienkamp, T., & Abur, D. (2026). Towards explainable reference-free speech intelligibility evaluation of people with pathological speech. arXiv preprint arXiv:2602.12723.

[PathBench] Halpern, B. M., Tienkamp, T., Abur, D., & Toda, T. (2026). PathBench: Speech intelligibility benchmark for automatic pathological speech assessment. arXiv preprint arXiv:2603.08097.

[Ippolito2023+] Ippolito, D., Tramer, F., Nasr, M., Zhang, C., Jagielski, M., Lee, K., Choquette Choo, C., & Carlini, N. (2023). Preventing generation of verbatim memorization in language models gives a false sense of privacy. In Proceedings of the 16th International Natural Language Generation Conference (pp. 28–53). ACL.

[Janbakhshi2020+] Janbakhshi, P., Kodrasi, I., & Bourlard, H. (2020). Synthetic speech references for automatic pathological speech intelligibility assessment. In ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 6099–6103). IEEE.

[Ma2025+] Ma, R., Qian, M., Gales, M., & Knill, K. (2025). ASR error correction using large language models. IEEE Transactions on Audio, Speech and Language Processing, 33, 1389-1401.

[Ren2019+] Ren, J., Liu, P. J., Fertig, E., Snoek, J., Poplin, R., DePristo, M. A., Dillon, J. V., & Lakshminarayanan, B. (2019). Likelihood ratios for out-of-distribution detection. Advances in Neural Information Processing Systems 32 (NeurIPS 2019).

[Sainz2023+] Sainz, O., Campos, J., García-Ferrero, I., Etxaniz, J., de Lacalle, O. L., & Agirre, E. (2023). NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In Findings of the Association for Computational Linguistics: EMNLP 2023 (pp. 10776–10787).

[Talkar2025+] Talkar, T., Kawabata, K., Higgins, C., & Tobyne, S. (2025). Development and validation of a Wav2Vec 2.0-based cross-language methodology for measurement of articulatory precision. In Proc. Interspeech 2025 (pp. 3748–3752).

[Tseng2025+] Tseng, Y., Parcollet, T., van Dalen, R., Zhang, S., & Bhattacharya, S. (2025). Evaluation of LLMs in speech is often flawed: Test set contamination in large language models for speech recognition. arXiv preprint arXiv:2505.22251.

[Wong2025+] Wong, J. H. M., Huzaifah, M., Chen, N. F., & Aw, A. T. (2025). Speech in-context learning of paralinguistic tasks. In 2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (pp. 1-7). IEEE.