ICML
International Conference
On Machine Learning

MondegreensEval

A Phonetic Benchmark for Measuring Language-model Bias in Automatic Speech Recognition

Wan Ju Kang

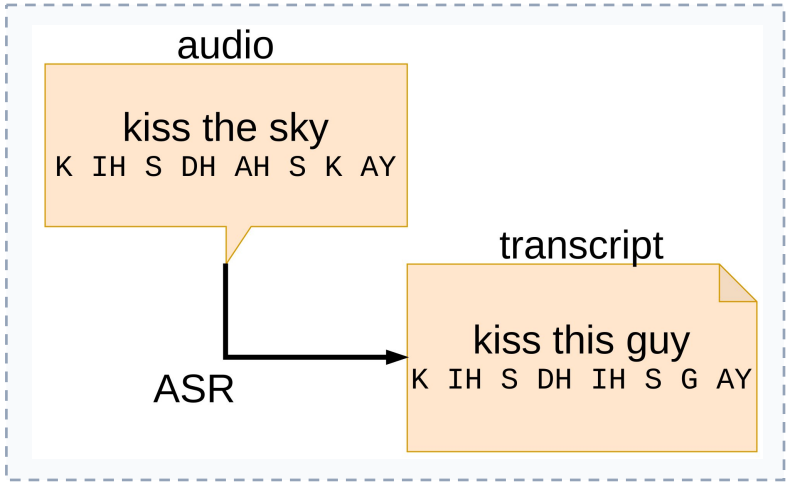
ICML 2026 Workshop on ML for Audio · Seoul, Korea

Mondegreens* are misheard expressions

Played (with noise): **“...kiss the sky”**

Transcribed: **“...kiss this guy”** - fluent, grammatical, wrong

- The two phrases differ in real, detectable phonemes, but the waveform (barely) supports the output
- WER and CER cannot flag this: the substitution is fluent by construction
- *Audio demo at the poster session. Come visit*



*Coined by Wright (1954), from “Laid him on the green” being misheard as “Lady Mondegreen” in *The Bonnie Earl o’ Moray*

Problem and Related Work

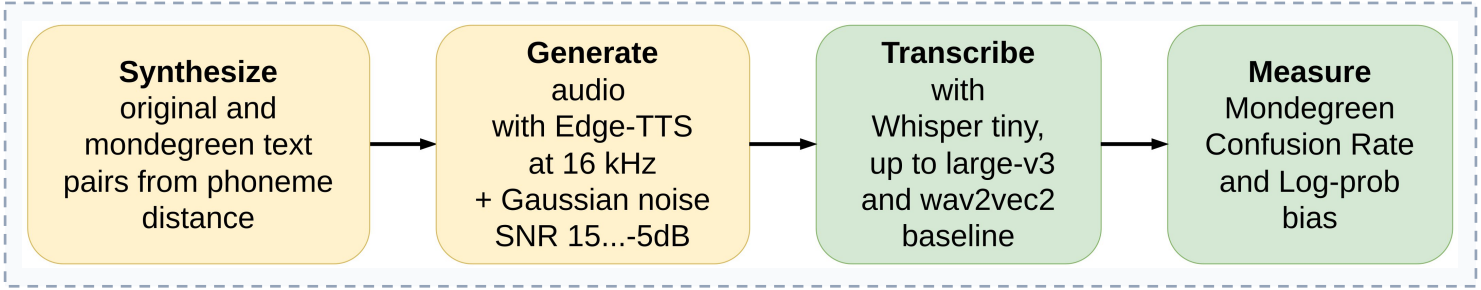
- Decoder-based ASR couples an acoustic encoder with an implicit language model; under degraded audio, the LM prior can override acoustic evidence
- Mondegreens ≠ homophones: real acoustic differences survive (min. phoneme distance **0.038** — no pair is identical), so a confusion is prior override, not ambiguity
- HER asks *is the output grounded?* We ask *which of two grounded readings does the decoder prefer?*
- ILME works assume the implicit LM is a separable component; our ablation finding says otherwise

① Scale
Does LM bias grow with model size?

② Similarity
Does phonetic proximity predict confusion?

③ Ablation
Can disabling token conditioning remove the prior?

Benchmark: 190 pairs, five noise levels, six models



190 (original, mondegreen) pairs: song lyrics 114 · conversational 41 · liturgical 35

- Four phoneme-distance tiers: near-homophone (n=36) · ambiguous (n=64) · weakly similar (n=16) · dissimilar (n=74)
- Audio: Edge-TTS (en-US-JennyNeural, 16 kHz), both forms per pair, Gaussian noise at SNR 15…−5 dB
- Models: Whisper tiny → large-v3, plus wav2vec2 CTC as an LM-free baseline

Ablation runs: *condition_on_prev_tokens=False*

Two metrics: a hard count and a soft signal

MCR-mono

Mondegreen audio played;

Confusion is defined as ‘transcript lands closer to the canonical text than to mondegreen text’

Edit-distance filter excludes transcription failures, thresholded at 0.5

Counts substitutions that actually happened

Logprob bias

Teacher-forced decoding:

log P(original | audio) – log P(mondegreen | audio)

Positive = model prefers the phrase it didn’t hear

Captures the pull before it manifests

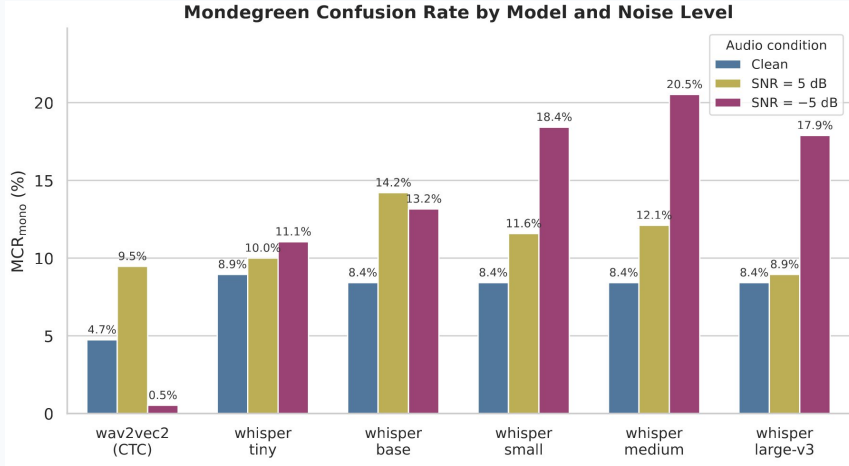
We also ran the symmetric control **MCR-orig**, confusion in the opposite direction

Result ① Noise drives confusion; model scale does not

Clean: all Whisper sizes at **8.4-8.9%**; wav2vec2 baseline 4.7%

SNR -5 dB: medium **20.5%** · small 18.4% · large-v3 17.9% · tiny 11.1%; not monotonic in parameter count

– wav2vec2 collapse at -5 dB is decoder failure, not LM correction



Directional control (MCR-orig)

Clean audio, opposite direction: **0.0–1.6%** vs. 8.4–8.9%

≤ MCR-mono in **all 18** model×condition cells

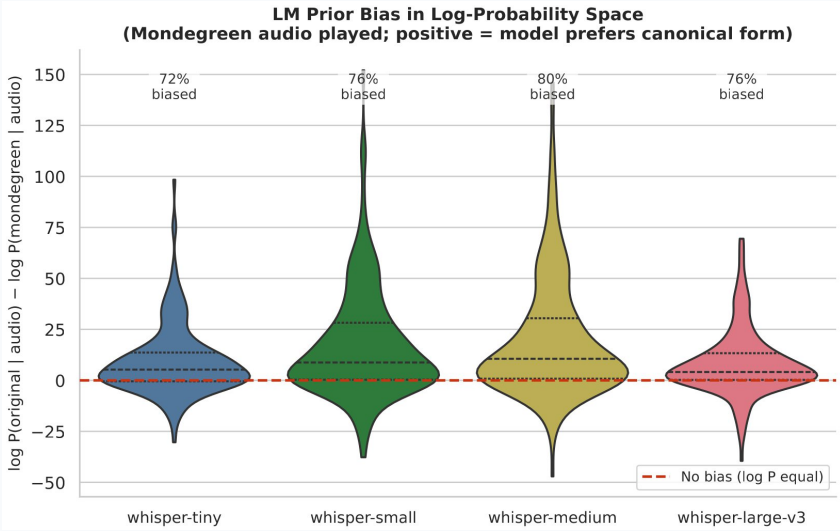
Gap widens with noise: **7.9 → 10.7 pts**, meaning that confusion flows one way, toward the prior

The bias is everywhere, even when MCR looks modest

72-80% of pairs carry positive bias, at every size

Mean bias: +9.3 (tiny) · +17.6 (small) · **+20.3 (medium)** · +8.4 (large-v3)

- Large-v3's attenuation vs. medium suggests encoder capacity or pretraining changes help, but the effect is not monotone across scale
- A systematic architectural property, not an occasional failure the hard metric happens to catch



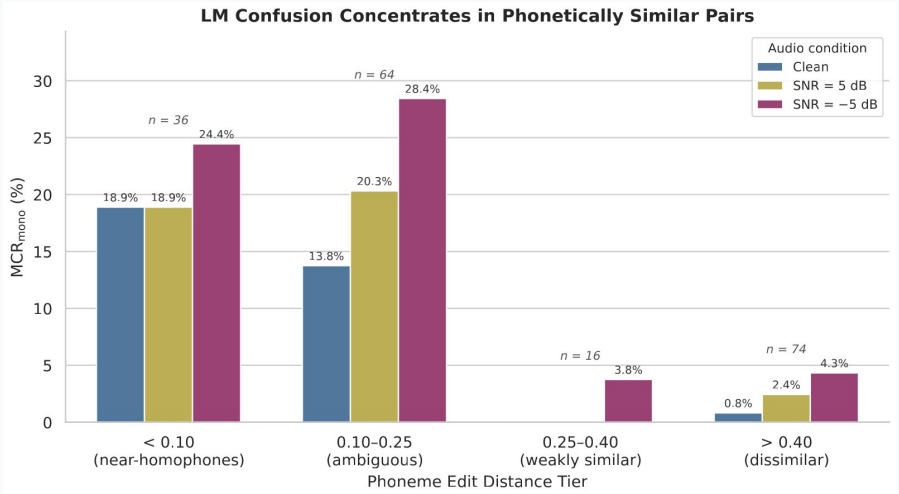
Result ② Phonetic proximity predicts confusion

r = −0.36 between phoneme distance and confusion rate ($p = 4.7\times10^{-7}$)

Near-homophone tier: 19-21% clean, **24-28%** degraded

Dissimilar tier: **under 4%** in every condition

– The gradient is the design working: confusion tracks the controlled confound, not random failure



Result ③ Negative finding from ablation

condition_on_prev_tokens=False ineffective: medium at **8.4% clean / 12.1% at 5 dB under both settings**; large-v3 likewise

– What the flag actually removes is cross-utterance context only; the intrinsic prior stays fully operational

Implication: the prior lives in the **attention weights**, not the autoregressive token stream

– A problem for ILME-style subtraction: there may be no separable component to subtract

– Consistent with encoder-decoder asymmetry, and with Whisper's decoder behaving as a weak standalone LM

MCR-mono (%)	condTrue	condFalse
whisper-medium clean	8.4	8.4
whisper-medium 5 dB	12.1	12.1

Limitations

- Single TTS voice, single synthesis per phrase; voice idiosyncrasy not yet excluded; multi-voice and recorded-speech reruns are next
- Gaussian noise only; babble, reverb, codec artifacts may shift the profiles
- The ablation is partial by nature; a full control needs a CTC-only Whisper encoder, which doesn’t exist

Takeaways

ICML
International Conference
On Machine Learning

- Up to **80%** of pairs biased toward the phrase the ASR model did not hear
- Up to **28%** confusion at near-homophone tier
- **Phonetic proximity** correlates with MCR
- Asymmetric confusion **directionality**:
  - Mondegreen audio -> (hallucinatory) correct transcript confusion much higher than Original audio -> hallucinatory mondegreen transcript
- MondegreensEval complements WER / CER / HER: a diagnostic for the acoustic-fidelity vs. linguistic-fluency tradeoff.

soarhigh.github.io

Backup A - MCR-orig: the full directional control

Model	Clean	5 dB	–5 dB
wav2vec2 (CTC)	0.5	5.3	0.5
whisper-tiny	1.6	4.2	4.2
whisper-base	0.5	3.2	6.3
whisper-small	0.5	1.6	6.3
whisper-medium	0.5	1.1	4.7
whisper-large-v3	0.0	2.1	4.7

MCR-orig ≤ MCR-mono in **all 18 cells** (n = 190 per cell)

Whisper-average gap widens with noise: **7.9 → 9.2 → 10.7 pts** (clean → 5 dB → -5 dB)

- wav2vec2 at -5 dB: both directions saturate at 0.5% - acoustic failure, not correction
- If MCR-mono were an artifact, this saturation pattern would appear everywhere; it does not

Backup B - Terminology

mondegreen_text

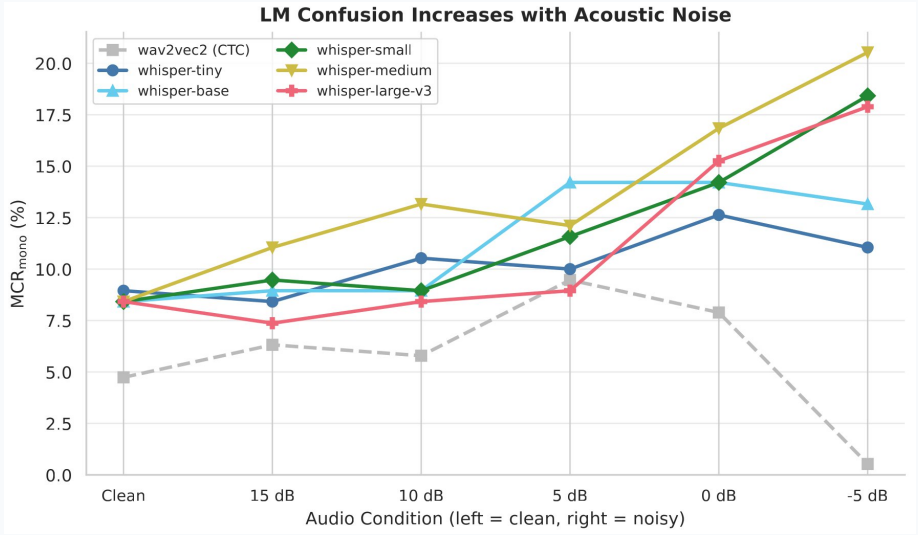
The phonetically plausible but lower-frequency alternative. In every “mondegreen audio” trial, this is what was **synthesized and played**; the acoustic ground truth the model is scored against.

original_text

The higher-frequency, canonical phrasing (the actual lyric, idiom, or liturgical line). The target the LM prior pulls toward.

“Mondegreen” coined by Sylvia Wright (1954), mishearing “laid him on the green” as “Lady Mondegreen” in the ballad The Bonny Earl o’ Moray.

Backup C - Full SNR sweep & decoding configuration



	tiny	base	small	medium	large-v3
Architecture					
Parameters	39M	74M	244M	764M	1,543M
Encoder layers	4	6	12	24	32
Decoder layers	4	6	12	24	32
d_{model}	384	512	768	1024	1280
Attention heads	6	8	12	16	20
Mel filterbanks	80	80	80	80	128
Weights dtype	fp32	fp32	fp32	fp32	fp16

Inference (all models)
Input sampling rate: 16,000 Hz Hop length: 160 samples Audio chunk: 30 s
Language: `en` Task: `transcribe` Timestamps: disabled
Beam search: disabled (greedy) Batch size: 1
`condition_on_prev_tokens`: `True` (standard) / `False` (ablation, medium & large-v3 only)

- Greedy decoding, batch 1, identical constraints across models
- Prefix tokens: <|startoftranscript|> <|en|> <|transcribe|> <|notimestamps|>