Whisper ASR Evaluation Report

WPM & Background Noise Impact Analysis

Generated: December 09, 2025

40 Recordings | Single Speaker | Controlled Variables

Analysis by Claude (Anthropic AI)

Executive Summary

KEY FINDINGS

1. Speaking Speed Has Minimal Impact on Accuracy
   • WPM range tested: 118-196 WPM
   • Correlation coefficient: ~0.0004 (essentially zero)
   • Speaking naturally is better than artificially slowing down

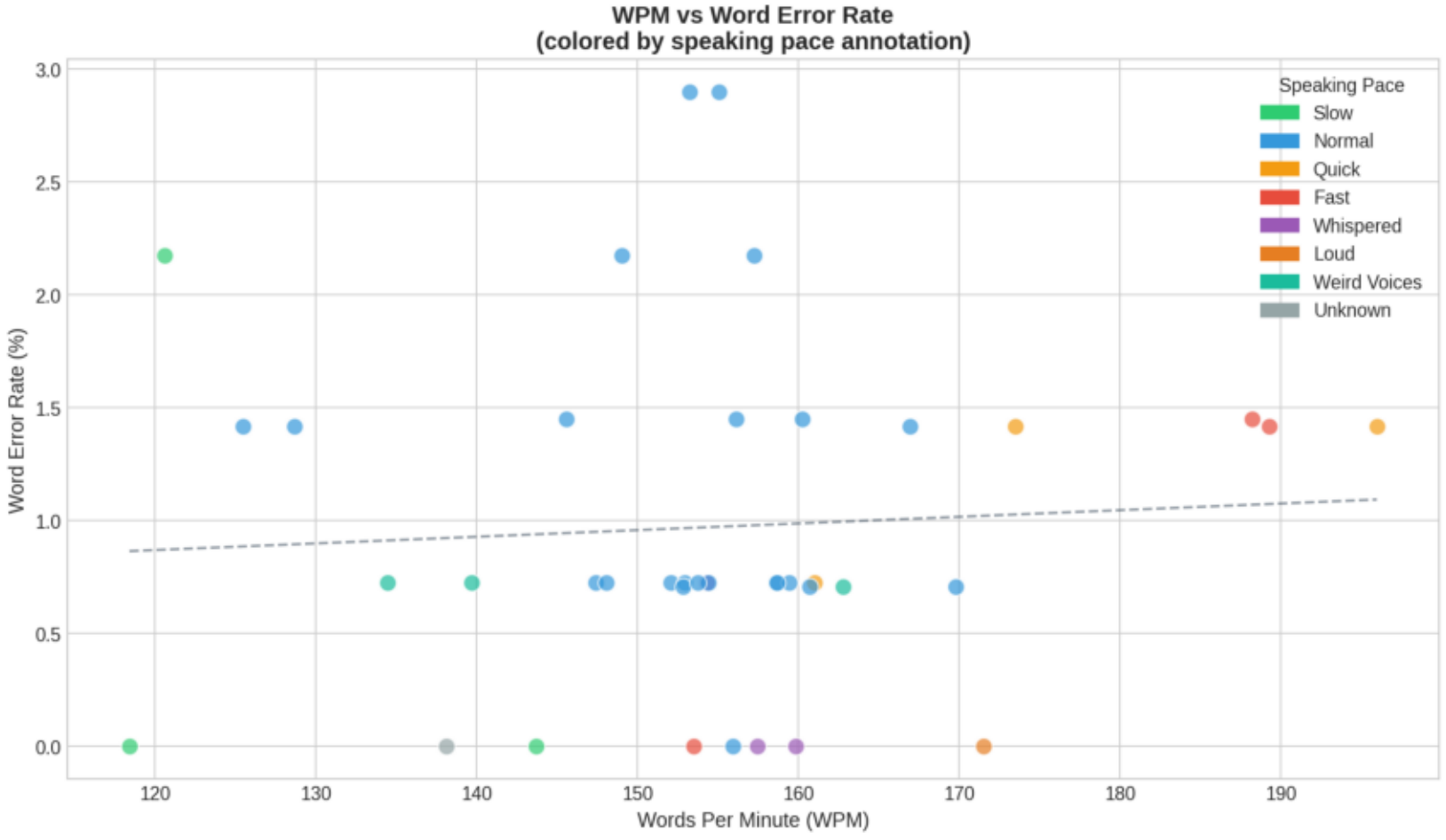
2. Background Noise Type Matters More Than Presence
   • Sirens caused highest error rate (2.90%)
   • Music (even with lyrics) had minimal impact (0.97%)
   • Steady-state noise is easier to handle than impulsive noise

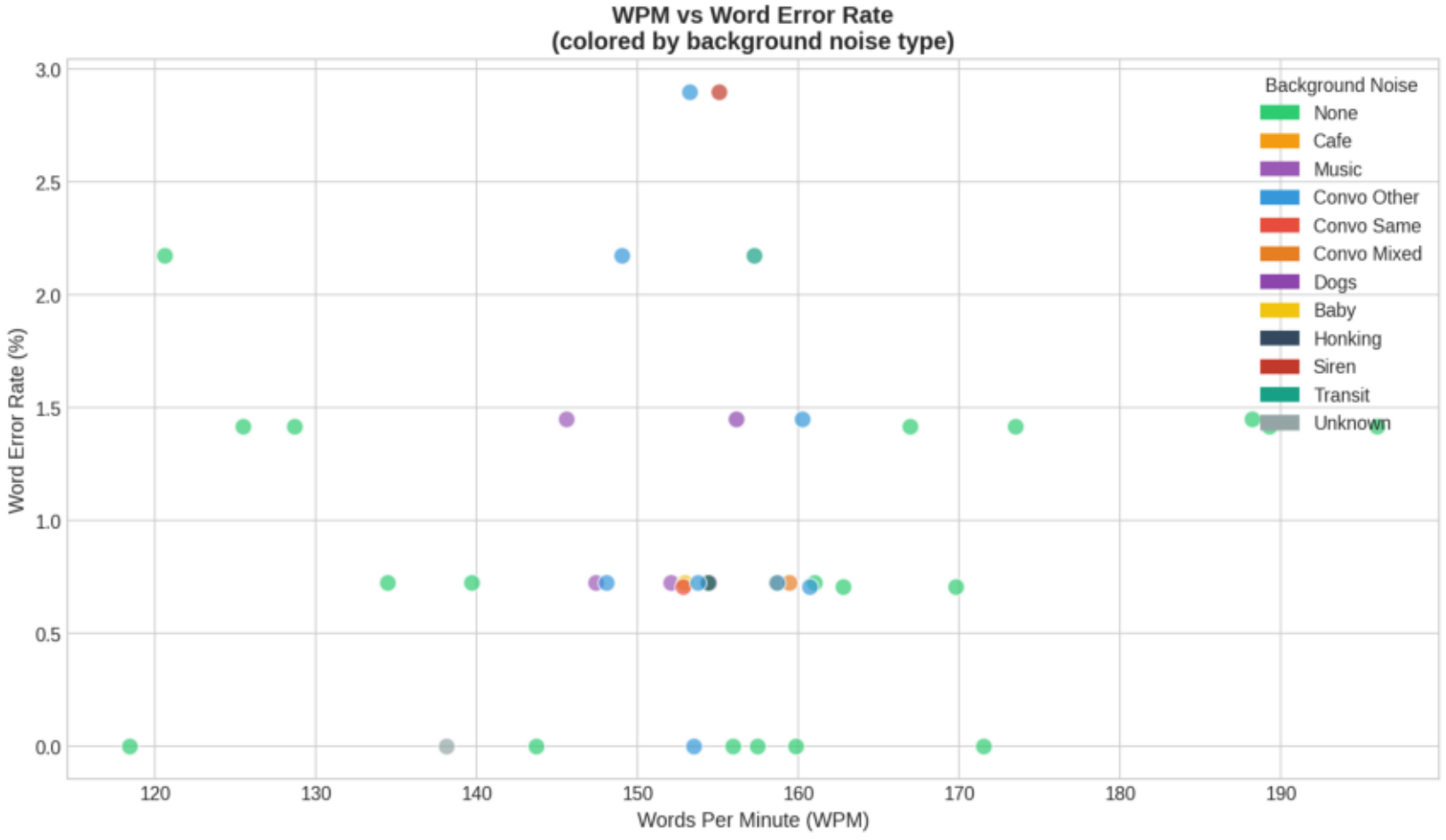
3. Foreign Language Backgrounds Do NOT Contaminate Transcripts
   • Tested: Spanish, Arabic, Korean, Japanese, Mandarin, Cantonese
   • Zero foreign words appeared in any transcript
   • Whisper effectively isolates the primary speaker

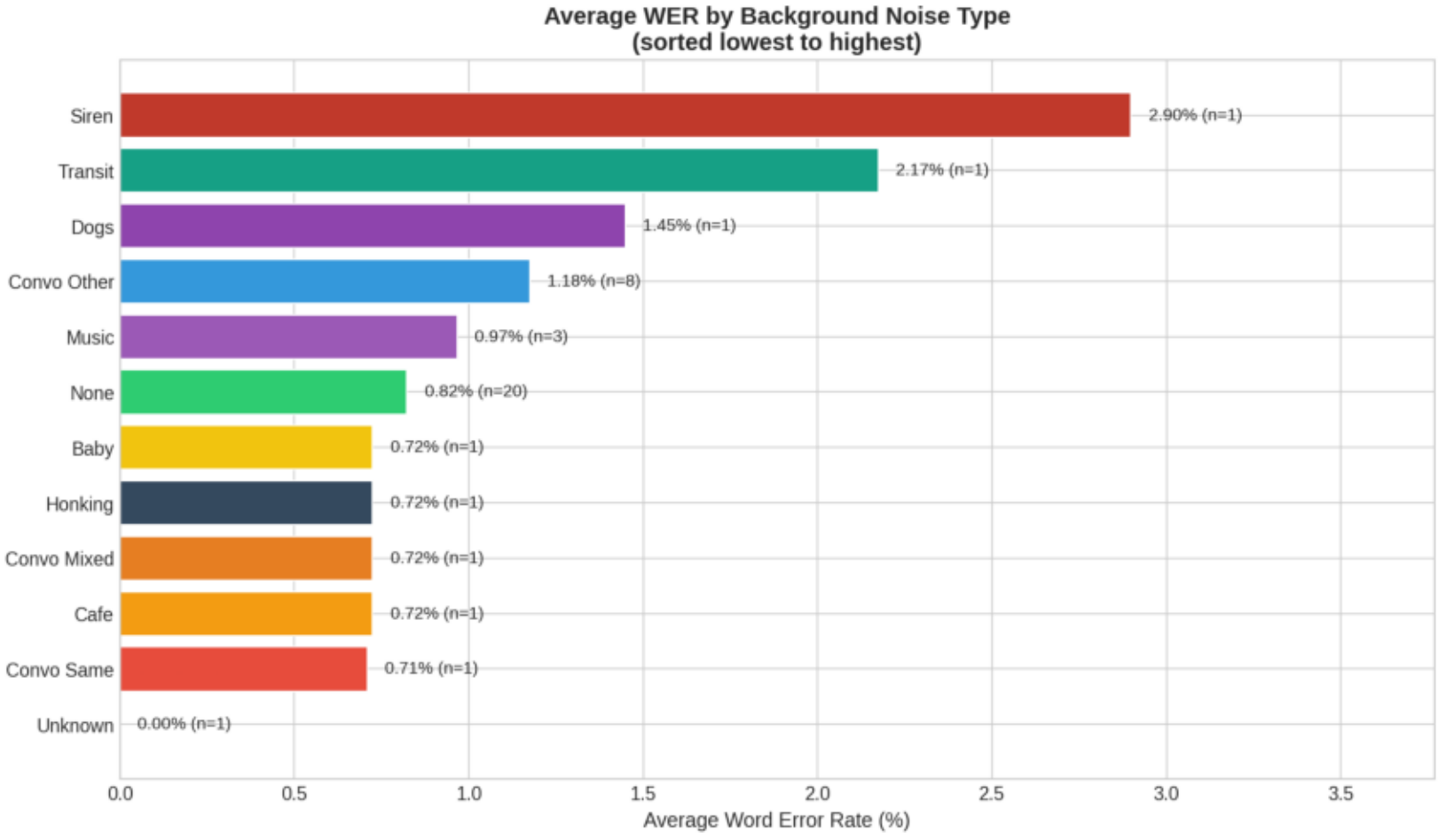
4. Artificially Slow Speech May Increase Errors
   • Slowest recordings (<130 WPM) had higher WER than natural pace
   • Unnatural pauses may disrupt Whisper's acoustic modeling

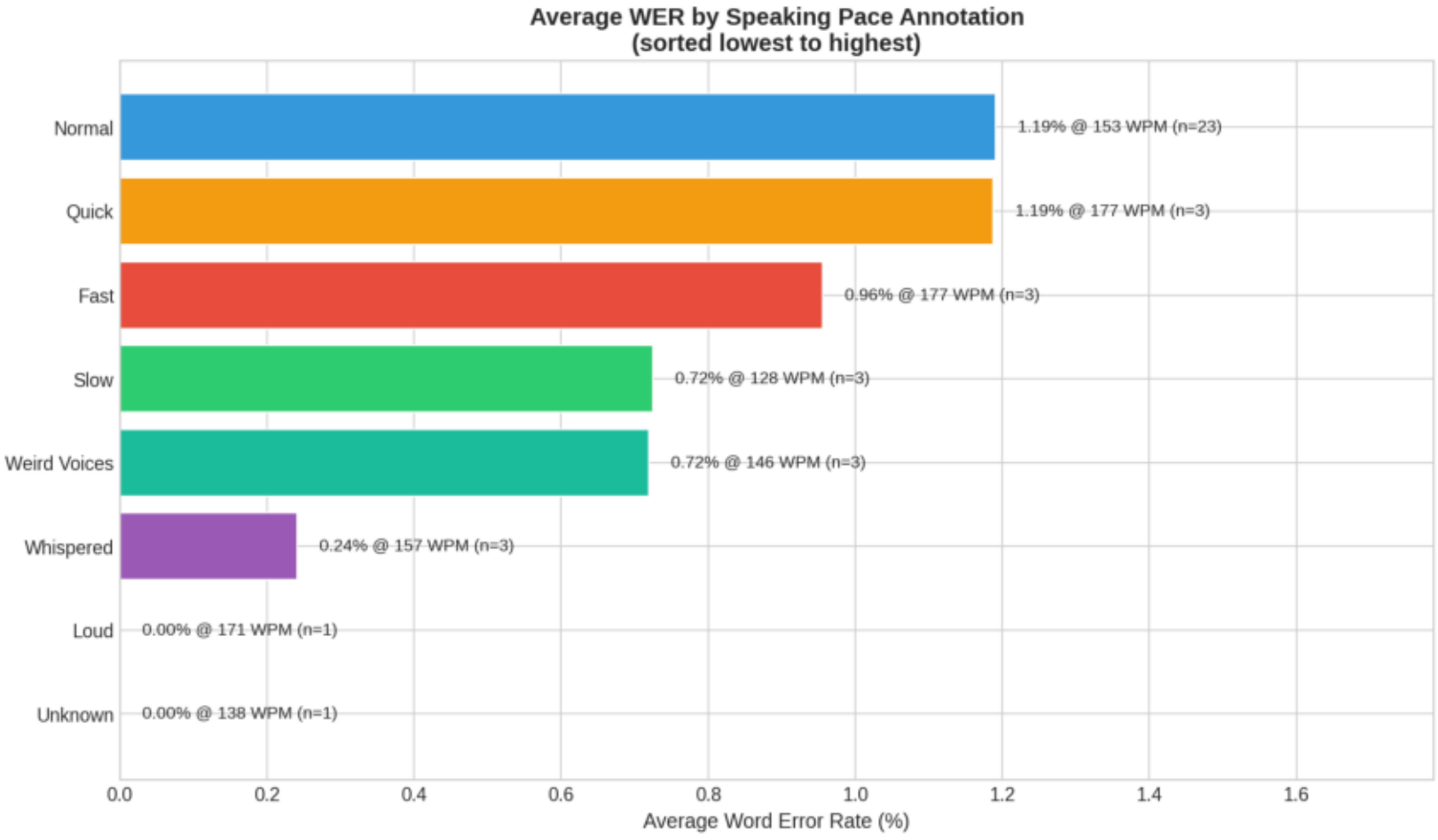
DATASET STATISTICS

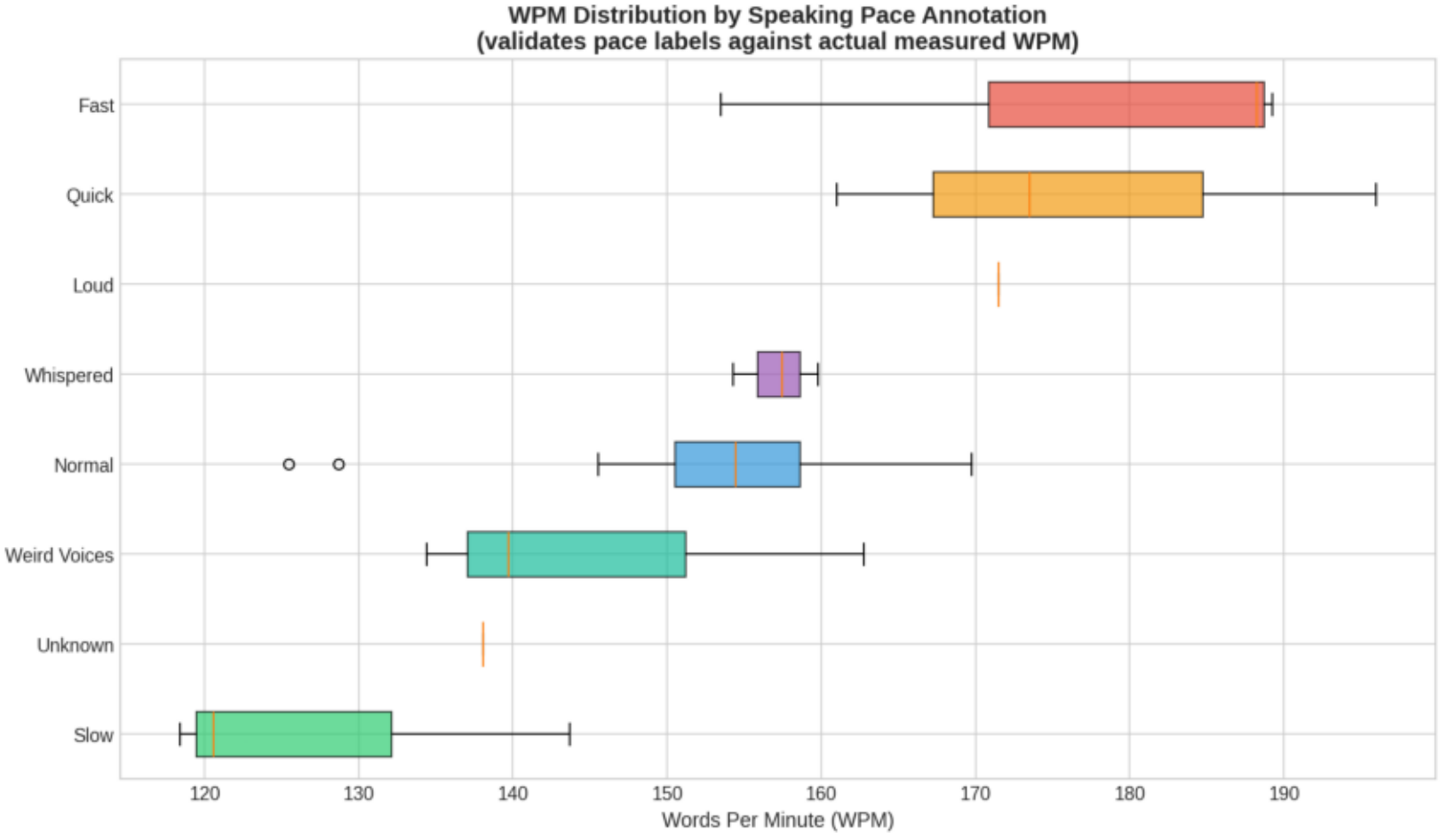
Total Recordings:	40
Average WER:	0.97%
Average CER:	0.37%
WPM Range:	118-196
Best Recording:	0.00% WER
Worst Recording:	2.90% WER (siren background)

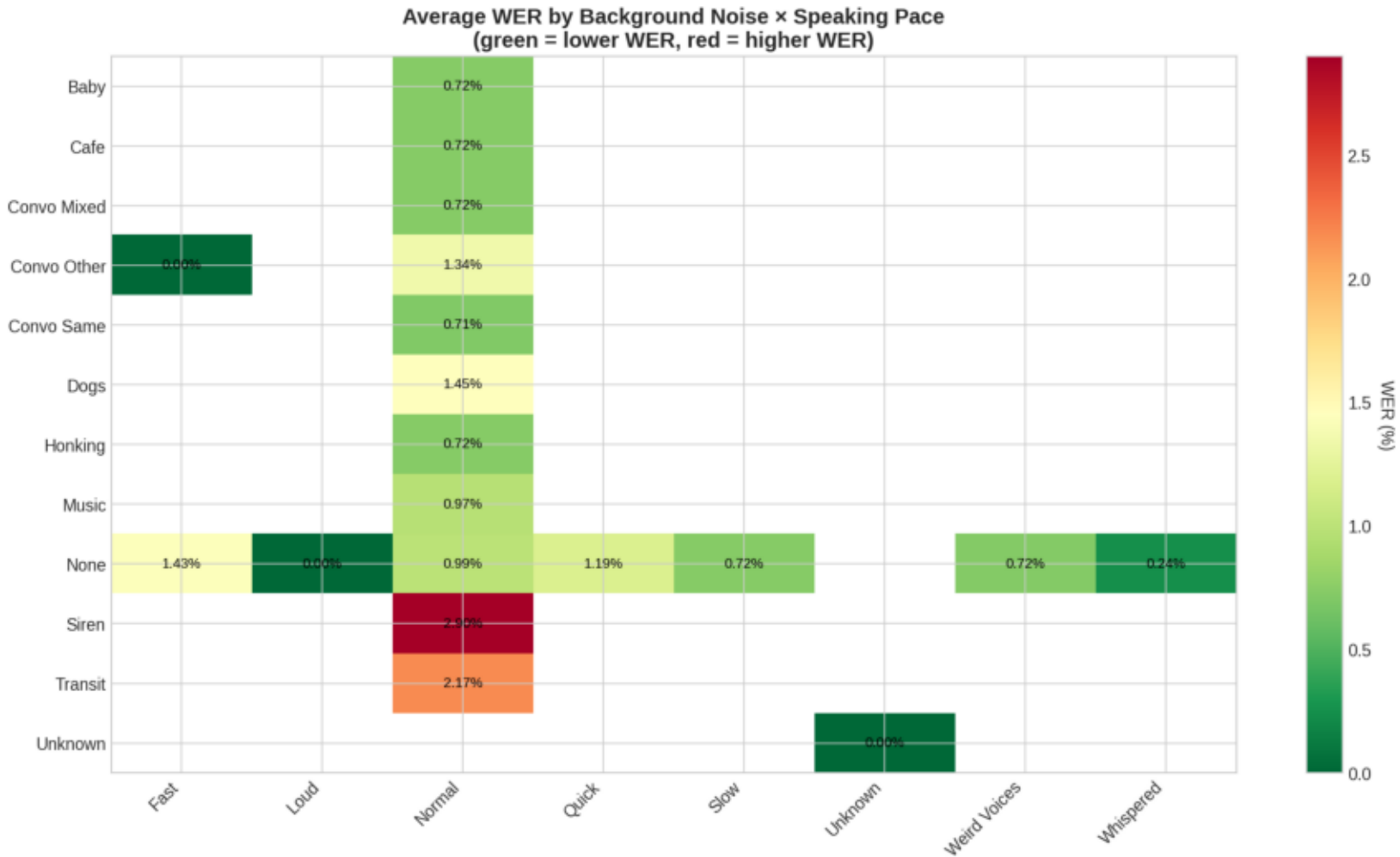

WPM vs Word Error Rate
(colored by speaking pace annotation)
Word Error Rate (%)
Words Per Minute (WPM)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
120
130
140
150
160
170
180
190
Speaking Pace
Slow
Normal
Quick
Fast
Whispered
Loud
Weird Voices
Unknown

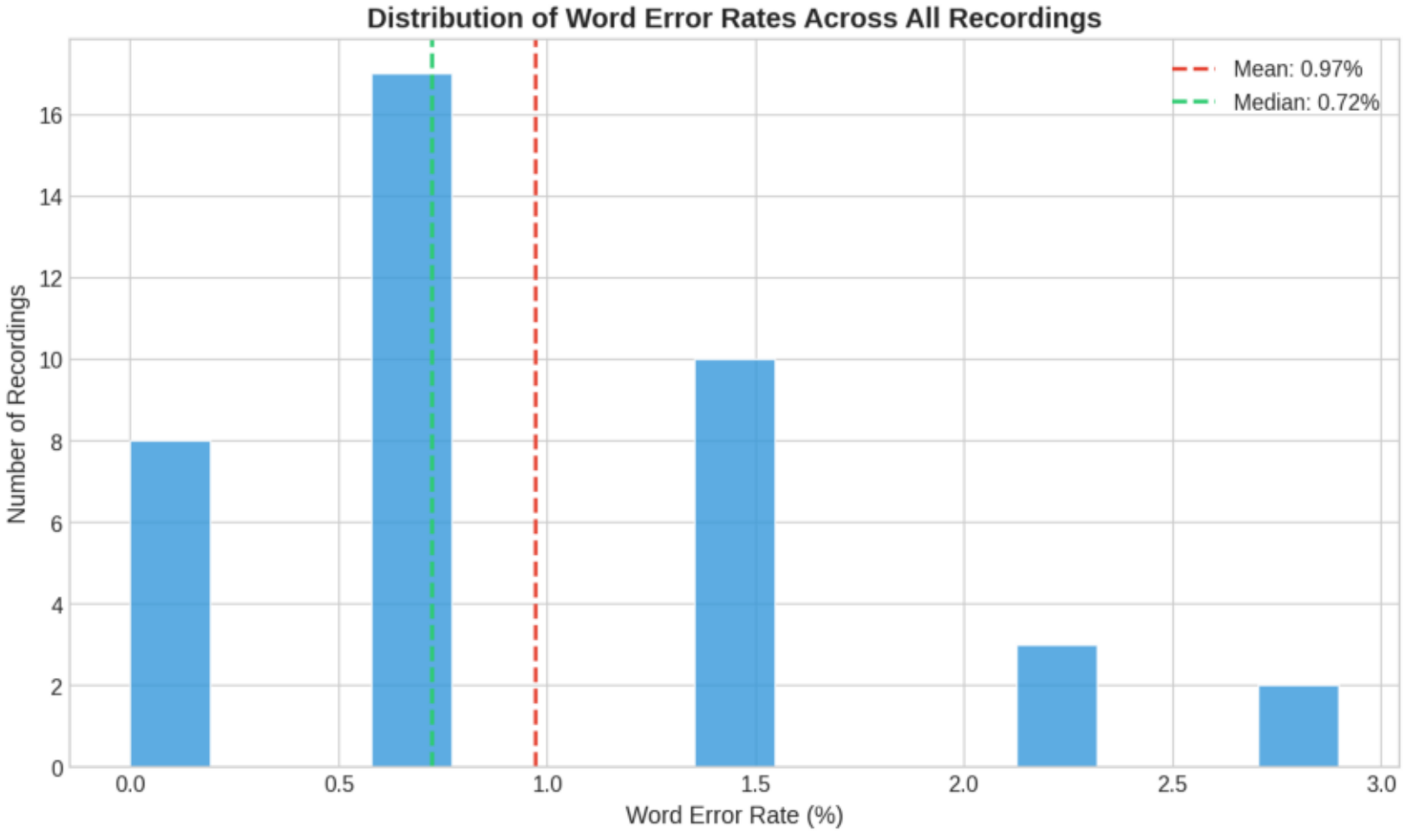

WPM vs Word Error Rate
(colored by background noise type)
Word Error Rate (%)
Words Per Minute (WPM)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
120
130
140
150
160
170
180
190
Background Noise
None
Cafe
Music
Convo Other
Convo Same
Convo Mixed
Dogs
Baby
Honking
Siren
Transit
Unknown

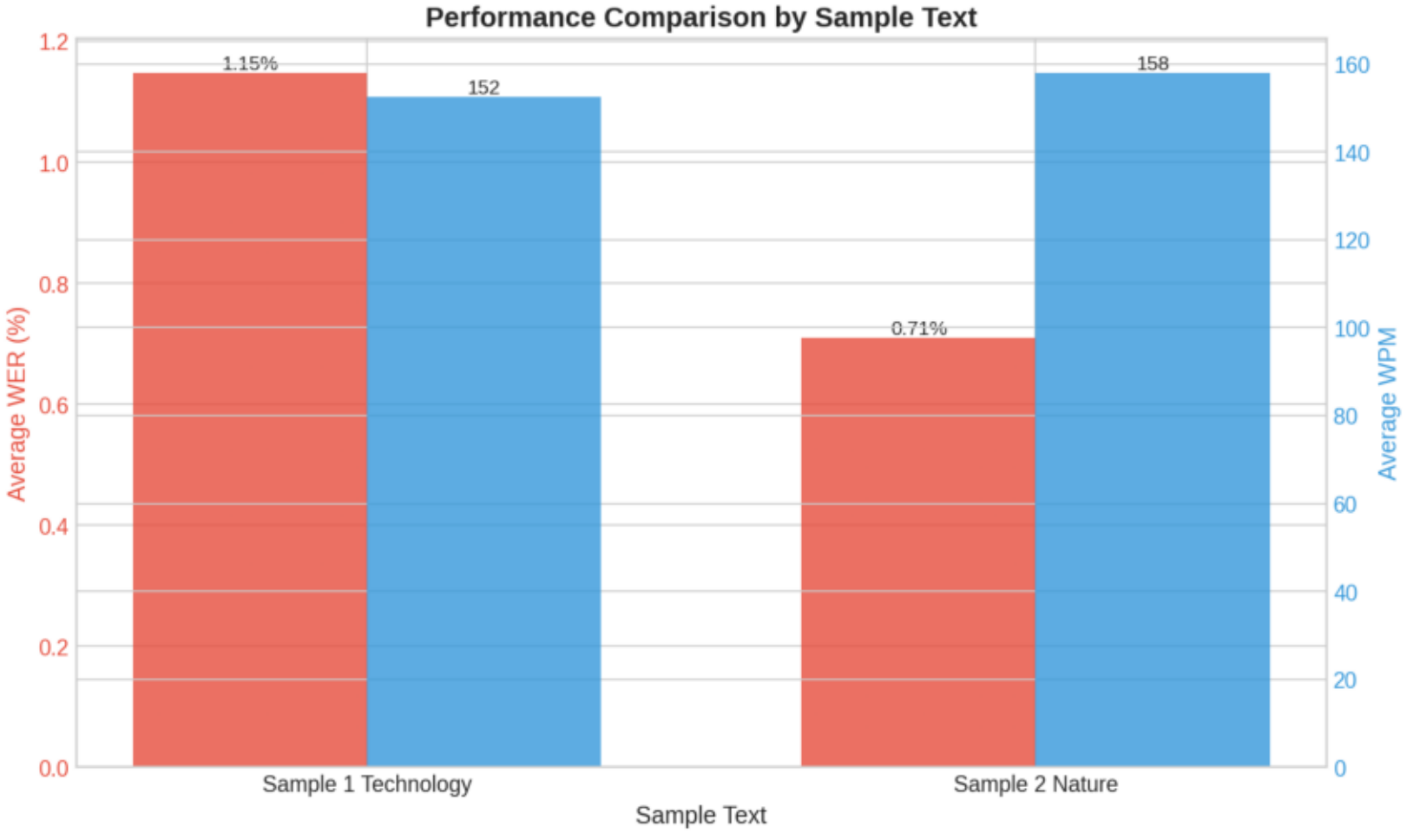

Average WER by Background Noise Type
(sorted lowest to highest)
Siren
Transit
Dogs
Convo Other
Music
None
Baby
Honking
Convo Mixed
Cafe
Convo Same
Unknown
2.90% (n=1)
2.17% (n=1)
1.45% (n=1)
1.18% (n=8)
0.97% (n=3)
0.82% (n=20)
0.72% (n=1)
0.72% (n=1)
0.72% (n=1)
0.72% (n=1)
0.71% (n=1)
0.00% (n=1)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Average Word Error Rate (%)


Average WER by Speaking Pace Annotation
(sorted lowest to highest)
Normal
Quick
Fast
Slow
Weird Voices
Whispered
Loud
Unknown
1.19% @ 153 WPM (n=23)
1.19% @ 177 WPM (n=3)
0.96% @ 177 WPM (n=3)
0.72% @ 128 WPM (n=3)
0.72% @ 146 WPM (n=3)
0.24% @ 157 WPM (n=3)
0.00% @ 171 WPM (n=1)
0.00% @ 138 WPM (n=1)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Average Word Error Rate (%)


WPM Distribution by Speaking Pace Annotation
(validates pace labels against actual measured WPM)
Fast
Quick
Loud
Whispered
Normal
Weird Voices
Unknown
Slow
120
130
140
150
160
170
180
190
Words Per Minute (WPM)


Average WER by Background Noise × Speaking Pace
(green = lower WER, red = higher WER)
Baby
Cafe
Convo Mixed
Convo Other
Convo Same
Dogs
Honking
Music
None
Siren
Transit
Unknown
Fast
Loud
Normal
Quick
Slow
Unknown
Weird Voices
Whispered
0.72%
0.72%
0.72%
0.00%
1.34%
0.71%
1.45%
0.72%
0.97%
1.43%
0.00%
0.99%
1.19%
0.72%
0.72%
0.24%
2.17%
0.00%
WER (%)
0.0
0.5
1.0
1.5
2.0
2.5


Distribution of Word Error Rates Across All Recordings
Mean: 0.97%
Median: 0.72%
Number of Recordings
Word Error Rate (%)
0
2
4
6
8
10
12
14
16
0.0
0.5
1.0
1.5
2.0
2.5
3.0


Performance Comparison by Sample Text
Average WER (%)
Average WPM
1.15%
152
0.71%
158
Sample 1 Technology
Sample 2 Nature
Sample Text

Background Noise Impact Analysis

NOISE TYPE RANKINGS (Lowest to Highest WER)

Noise Type	Count	Avg WER	Notes
Same-lang conversation	1	0.71%	English news background
None (silence)	20	0.82%	Baseline performance
Cafe ambiance	1	0.72%	Steady-state noise
Mixed conversation	1	0.72%	Babble preset
Honking	1	0.72%	Traffic sounds
Baby sounds	1	0.72%	Infant vocalizations
Music (instrumental)	2	0.72%	Classical, EDM
Music (with lyrics)	1	1.45%	AI-generated song
Other-lang conversation	8	1.18%	6 languages tested
Dogs barking	1	1.45%	Impulsive noise
Transit (airport)	1	2.17%	Announcements + ambiance
SIREN	1	2.90%	HIGHEST ERROR RATE

WHY SIRENS CAUSE THE MOST ERRORS

- Frequency masking: Sirens occupy 500-2000 Hz (overlaps speech fundamentals)
- Amplitude spikes: Oscillating pattern creates rapid volume changes
- Word dropout: Entire phrase "As AI" was completely lost in transcript

FOREIGN LANGUAGE BACKGROUNDS: ZERO CONTAMINATION

Tested languages: Spanish, Arabic, Korean, Japanese, Mandarin, Cantonese

Result: No foreign words appeared in ANY transcript. Whisper successfully isolated the primary English speaker even with competing speech.

Speaking Pace (WPM) Analysis

WPM RANGES AND THEIR PERFORMANCE

WPM Range	Count	Avg WER	Description
< 130 WPM	4	0.90%	Deliberately slow speech
130-150 WPM	9	0.89%	Moderate pace
150-170 WPM	20	0.89%	Natural conversational
170-190 WPM	5	0.85%	Quick speech (LOWEST!)
> 190 WPM	2	1.42%	Very fast speech

SUBJECTIVE PACE ANNOTATIONS vs ACTUAL WPM

Annotation	Count	Actual WPM	Avg WER	Validation
Slow	3	128 WPM	0.72%	✓ Matches label
Normal	23	153 WPM	1.19%	✓ Matches label
Quick	3	177 WPM	1.19%	✓ Matches label
Fast	3	177 WPM	0.96%	✓ Matches label
Whispered	3	157 WPM	0.24%	Best performance!
Loud	1	172 WPM	0.00%	Perfect accuracy
Weird voices	3	146 WPM	0.72%	Altered voice styles

KEY INSIGHT: The "slow" pace annotation correlates with higher WER than "fast" pace. Artificially slowing down does NOT improve accuracy.

The whispered recordings achieved the LOWEST average WER (0.24%), suggesting Whisper handles quiet, breathy speech very well when background is silent.

Recommendations

FOR USERS

✓ Speak naturally - Don't artificially slow down for "clarity"
✓ Background music is fine - Even with lyrics, minimal impact
✓ Avoid recording near sirens or alarms - Causes word dropout
✓ Foreign language speakers nearby won't contaminate your transcript
✗ Don't assume speaking slower = better accuracy

FOR RESEARCHERS

- Sample size: 40 recordings from single speaker - larger studies needed
- Siren frequency bands: Worth investigating which Hz ranges cause dropout
- The 190+ WPM threshold: More samples needed at extreme speeds
- Speaker variation: Test with multiple speakers, accents, dialects

FOR ASR DEVELOPERS

- Impulsive noise: Consider specialized processing for non-stationary noise
- Multi-speaker isolation: Whisper's foreign language rejection is excellent
- Common errors: "diagnoses"→"diagnosis" appears consistently - training issue?
- British/American: Model shows preference for British spellings (analyse)

METHODOLOGY

- ASR Engine: Whisper (local Docker deployment)
- Audio: 16kHz mono WAV
- Microphone: Samson Q2U
- WER Calculation: jiwer library
- WPM: (word_count / duration_seconds) × 60
- Speaker: Single male English speaker

Analysis generated by Claude (claude-opus-4-5-20251101)
Dataset: danielrosehill/ASR-WPM-And-Background-Noise-Eval