What is the impact of varying the number of training samples on the accuracy of Qwen-Audio versus Whisper-Larg

Assignee Research

June 10, 2026

Abstract

Recently, instruction-following audio-language models have received broad attention for audio interaction with humans. However, the absence of pre-trained audio models capable of handling diverse audio types and tasks has hindered progress in this field. Consequently, most existing works have only been able to support a limited range of interaction capabilities. In this paper, we develop the Qwen-Audio model and address this limitation by scaling up audio-language pre-training to cover over 30 tasks and various audio types, such as human speech, natural sounds, music, and songs, to facilitate

1 Introduction

This paper examines: Qwen-Audio: Advancing Universal Audio Understanding via Unified Large-Scale Audio-Language Models. Research question: What is the impact of varying the number of training samples on the accuracy of Qwen-Audio versus Whisper-Large-V3 and OpenPangu-7B-MLA in low-resource language speech understanding tasks as measured by WER on MMSU?.

2 Methodology

Systematic literature search across multiple databases yielded 5 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.7/10.

3 Results

5 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 3.7/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
Qwen-Audio achieves a WER of 1.8 on Librispeech dev-clean, 4.0 on dev-other, 2.0 on test-clean, and 4.2 on test-other.	×	0.03
Qwen-Audio achieves a WER of 1.2 on Aishell1 dev and 1.3 on test.	×	0.04
Qwen-Audio achieves a WER of 3.3 on Aishell2 Mic, 3.1 on iOS, and 3.3 on Android.	×	0.04
Qwen-Audio achieves a BLEU score of 25.1 on CoVoST2 en-de, 33.9 on de-en, 41.5 on en-zh, and 15.7 on zh-en.	×	0.02
Qwen-Audio achieves a BLEU score of 39.7 on CoVoST2 es-en, 38.5 on fr-en, and 36.0 on it-en.	×	0.03
Qwen-Audio achieves a CIDEr score of 0.441, SPICE of 0.136, and SPIDEr of 0.288 on Clotho.	×	0.03
Qwen-Audio achieves an AAS of 60.3 ms on Industrial Data for SRWT.	×	0.04
The audio encoder in Qwen-Audio is initialized based on the Whisper-large-v2 model.	×	0.05
Qwen-Audio employs a single audio encoder to process various types of audio.	×	0.07
The training objective of Qwen-Audio is to maximize the next text token probability conditioning on audio representation	×	0.07

References

- http://arxiv.org/abs/2207.08179v1
- http://arxiv.org/abs/2311.07919v2
- http://arxiv.org/abs/2109.14357v1