A Differential Study of the Performance of Whisper ASR and Human on Transcription Task of whispered Mandarin Speech

Haonan Chen
University of Zurich
haonan.chen@uzh.ch

Abstract

The emergence of automatic speech recognition (ASR) is an important branch in the field of human-computer interaction, which provides material for further text processing and offers extensive support for speech and text technologies. Its effective operation relies heavily on the clarity, consistency and predictability of the speech signal. Therefore, dealing with varying speech quality is the key and difficult point for ASR systems. This paper will focus on performance comparison of the current leading Whisper ASR system and Human Speech Recognition (HSR) in transcribing whispered speech, with Mandarin selected as the target language. Considering the fact that Mandarin is a tonal language and the tones play a crucial role in conveying meaning, the challenges posed by this feature to the ASR system will be more significant and worth studying. This study attempts to investigate the strengths and weaknesses of Whisper and manual transcription by comparing the results of 16 sentences (comprising 8 daily utterances and 8 literary recitations) to provide insights to identify areas for potential improvement. The experimental results shows that the Whisper ASR system performed uniformly across different text categories, but it was evidently outperformed by human, particularly with daily utterances. This suggest that Whisper still needs to improve its ability to generate transcription with context and to adapt the model to the acoustic characteristics of whispered Mandarin speech.

1 Introduction and Background

Automatic Speech Recognition (ASR) systems, pivotal in contemporary technology, convert spoken language into text. Their applications span various domains, from voice-activated assistants to transcription services. One notable ASR system is WHISPER (Vaswani et al., 2017), which is trained on diverse and extensive multilingual datasets, so is capable of recognizing and processing various speech patterns, accents, and dialects and is known for its adaptability and accuracy. These systems, traditionally optimized for normal speech, face unique challenges when dealing with whispered speech, a variant significantly different in acoustic properties.

Mandarin as a tonal language relies heavily on pitch variations to convey meaning. Tones in Mandarin are crucial for distinguishing between words that would otherwise be homophones. The absence of vocal fold vibration leads to a considerable reduction in speech energy and speaking rate, a significant factor in speech recognition technologies and human perception.

Many previous studies have focused on the special acoustic features of whispered Mandarin speech compared to normal speech. In addition to the obvious absence of f0, studies by Chang and Yao (2007) emphasize the shrinking but still significant differences in duration and intensity among the four lexical tones in Mandarin. Complementing these findings, research by Lv and Zhao (2010) has demonstrated that whispered Mandarin vowels exhibit a notable upward shift in formant frequency, expanded and nearly constant formant bandwidths, much lower gain and flatter spectrum. With regard to consonants in whispered Mandarin, key findings from Xu et al. (2022) show that consonants tend to have longer durations and varied intensities, with nasals and semivowels being notably lower in intensity and others showed a notably increase. Fricatives and affricatives perceived largely unchanged spectrum with a narrower distinction in duration among them.

In addition to this, Gao and Esling (2003) use Laryngoscopic endoscopy to demonstrate the physiological aspects of whispered mandarin. It shows that the height of the larynx primarily dictates pitch shifting, serving as a substitute for the vocal folds vibration, rising tonal contours correlates with rising larynx, and the sphincteric opening correlates with the production of noise in whisper, which controls tone intensity, thus playing a role similar to the loudness in phonated tones.

Given all the unique properties of whispered speech compared to phonated speech, along with the current sparse corpus of whisper Mandarin available as training material, it is reasonable to assume that WHISPER's performance will be impaired, as ASR systems largely depend on the acoustic properties of speech, which undergo significant changes in whispering. Assessment of normal speech recognition is familiar topic, in this paper I will design experiments to compare the performance of the leading ASR system Whisper with human speech recognition (HSR).

Hypothesis. Since humans rely more on contextual cues and previous language experience than the ASR system, this may give them an advantage in understanding whispered speech with unique acoustic properties. We therefore hypothesize that there is a significant difference in performance between WHISPER and HSR, and that human transcription outperforms WHISPER. Similarly, since daily phrases are used more and are simpler than literary texts, it can be hypothesized that transcription of daily texts performs significantly better than literary texts.

By investigating the acoustic properties of whispered Mandarin and reviewing related studies, the complexities involved in recognizing and transcribing whispered speech can be better understood. Combined with the performance comparisons and results analyzed in this study, these may contribute to the improvement of speech recognition in specific domains. Moreover, improving the recognition rate of whispered speech is also beneficial to help patients with vocal cord vibration deficiency to communicate with others and to solve the problem of speech transcription in some special scenarios, thus it is also an application-worthy problem.

2 Method

2.1 Data Collection

Stimuli preparation. Considering the low intelligibility of isolated monosyllables in whispered Mandarin (Holbrook and Lu, 1969), this study utilizes complete sentences as experimental material to enhance intelligibility. In the pre-experiments, the transcription of slow, clear whispered recordings was very effective, especially for daily utterances, and both Whisper ASR and human were able to achieve character error rates (CER) approaching zero. So, the difficulty of the transcription task was increased by increasing the speed of speech to amplify the difference.

Since gender has also been shown to affect ASR recognition (Boito et al., 2022), 16 recordings from one female speaker aged 24 were selected as experimental material, which also ensures a consistent whispering level and speaking rate across all recordings.

The 16 sentences contain characters ranging from 19-34, with an average length of 24.6 characters. The first eight sentences are daily sentences and the last eight sentences are prose extracts. The first eight sentences are set with certain whispered homophones that are difficult to distinguish and require a combination of contexts to transcribe correctly.

All stimuli were recorded as mono recordings with resolution of 44100Hz on Praat using computer microphone.

Participants Selection. Transcribers includes 6 native Mandarin speakers, 4 female and 2 male, 22-41 years.

Data collection. With respect to ASR systems, the *whisper-large-v3* released in November 2023 were chosen to represent the leading performance of ASR. All recordings were transcribed through the model on Colab. Similarly, all recordings were given to 6 participants and their results were recorded and processed.

Pre-processing. All HSR results are saved in the *Transcription.txt*, and the preprocessed results after removing non-characters including punctuations and blank spaces are saved in *Transcription.csv*. Afterward data analysis was based on the *Transcription.csv* file.

2.2 Evaluation Matrix and Data Analysis

The Character Error Rate (CER) will be the metric for assessing transcription performance. CER is defined as the percentage ratio of the minimum number of insertions (i), substitutions (s) and deletions (d) of characters that are required to obtain the gold answer and the total number of characters (n):

$$CER = \frac{i+s+d}{n} \tag{1}$$

CER of the transcription results from 6 participants were averaged as the performance of the human transcription. CER calculations were done using python package *jiwer*.

The Wilcoxon Signed-Rank Test was conducted as a measure of whether there was a significant difference between the ASR and HSR performance. Furthermore, transcription performance of 8 daily utterances and 8 literary recitations were also compared to assessed the effect of text category. The Mann-Whitney U Test was conducted to see whether there was a significant difference between the 2 text categories regarding ASR and HSR performance. Data analysis was done in python.

3 Results

3.1 Basic Statistics and Normality Test

All of the CER data was stored in a *CER_Results.csv* file, and we will be focusing on the analysis and comparison of the WHISPER and HSR columns of data in the following. Here HSR performance is represented by the mean performance of all participants.

	Count	Mean	Std	p
WHISPER	16	0.2591	0.1537	0.9864
HSR	16	0.1487	0.1208	0.0326

Table 1: Basic information of WHISPER and HSR performance

According to the basic information of CER in Table 1, ASR model WHISPER has higher Mean and higher Std of CER than HSR, which indicates a lower accuracy of transcription of whispered Mandarin and more erratic performance for multiple sentences. This tentatively suggests that humans may be more adept at transcription and comprehension of whispered Mandarin speech.

Apart from that, HSR data do not follow a normal distribution ($p < 0.05$), which may affect the validity of parametric tests because many of them (e.g. t-tests) assume normal distribution of the data. There are two approaches to address this case: (1) applying data transformations to make the data more consistent with a normal distribution, and (2) using non-parametric methods such as the Mann-Whitney U test or the Wilcoxon rank sum test, which do not rely on the assumption of a normal distribution. In response to method 1, not all data can necessarily be transformed to achieve a normal distribution, and this transformation may introduce new issues or inaccuracies that change the nature of the data relationship. After applying some data transformations it was found that no satisfactory normal distribution could be achieved, so non-parametric tests was adopted.

In this context, the Wilcoxon Signed-Rank Test was first employed to compare the performance difference between CERs derived from pairs of sentences using WHISPER and HSR. The Mann-Whitney U Test was then used to compare the transcription performance of WHISPER and HSR between daily utterances and literary text, respectively.

3.2 Comparison of WHISPER and Human Transcription Performance

Figure 1 shows the character error rate (CER) of WHISPER and human transcribers (HSR and Participants 1 to 6) for the transcription of 16 sentences. A lower CER indicates fewer errors in transcription and better performance.

WHISPER had a wide distribution of CERs, with a minimum value of 0, a maximum value close to 0.6, and a median around 0.25. The boxplot for WHISPER showed large variability, implying that its performance fluctuates widely from sentence to sentence. The CERs for HSR were generally lower than those for WHISPER, suggesting that the performance of the human transcriber was generally superior.

For individuals, participants 1, 3 and 6 achieved zero error in most cases, which may indicate a high level of mastery of the transcription task. For the CER close to 1 that appeared in participant 1 and participant 5, the transcriber did not understand the entire sentence, so the answer was mostly blank, with only a few characters. The difference in CER between participants indicates that performance on the transcription task may differ between individuals due to experience, familiarity, or other factors. This was verified in my callback with participant 3, who had memorized the literary excerpts due to her previous translation test preparation.

Generally speaking, human transcribers performed better and with less variability than WHISPER, probably because humans are better at using contextual information and linguistic knowledge to recognize whispers.

The boxplot with paired sentences in Figure 2 shows that HSR obtains significantly lower CERs than WHISPER for 12 of the 16 sentences in the group. Of the 4 sentences where WHISPER performs better, only the 11th sentence has a CER that

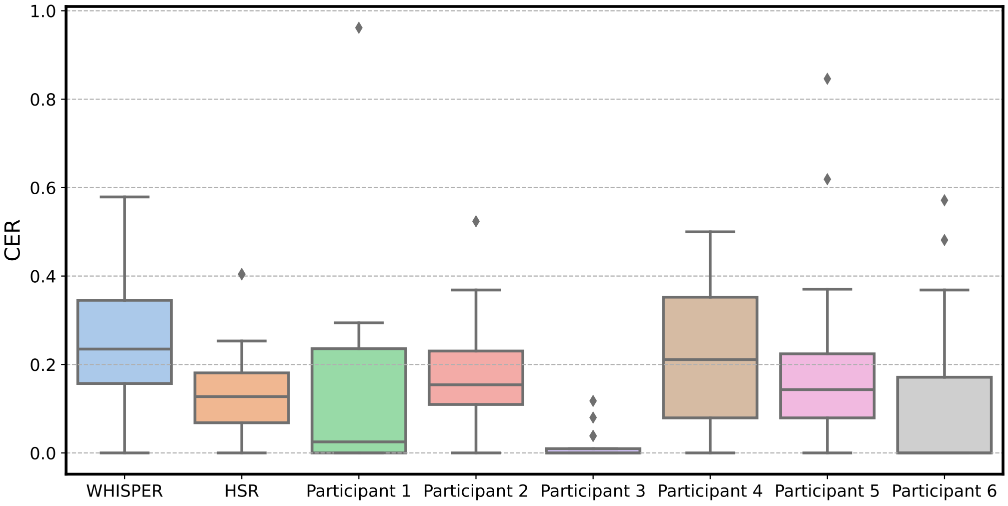

Figure 1: The CER of WHISPER and human transcribers (HSR and six participants) for the transcription of 16 sentences

is significantly greater for WHISPER than HSR. The transcription of this sentence involved two participants giving nearly blank answers with only a few Chinese characters, which greatly increased the mean CER of HSR. WHISPER achieved full accuracy on the 5th sentence while the CER for HSR was 0.075, which was because 3 participants transcribed a homophone incorrectly. For sentences 13 and 14, WHISPER's performance is only 0.03 and 0.02 lower than the CER for the human transcription respectively, so is not significantly better. Overall, WHISPER will always give a complete answer whether it is correct or not, while blank character in human transcripts will affect the overall average performance.

Statistically, according to Wilcoxon Signed-Rank Test result p-value is 0.02496, which indicates that overall HSR is more accurate than WHISPER in transcription task of whispered Mandarin speech.

3.3 Comparison of Transcription Performance of Different Text Category

The sentences for the experimental design consisted of eight daily utterance and eight literary excerpts, and this section compares the transcription results of WHISPER and HSR of the two genres respectively, from Figure 3 and using the Mann-Whitney U test.

For WHISPER, the median CERs for the first and last eight sentences were close to each other,

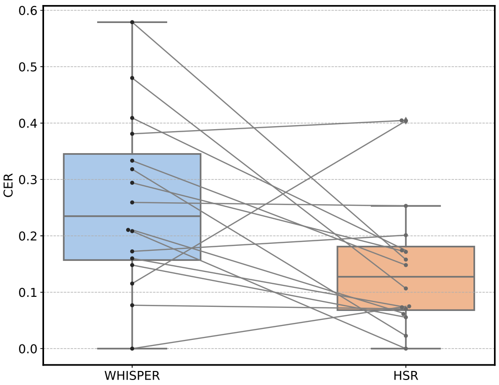

Figure 2: The CER comparison with paired sentences in WHISPER and HSR

and the size and shape of the boxplots were similar, which means that there was no significant difference in the performance of the ASR system on daily phrases and literary excerpts. For HSR, the CERs for the first eight sentences were significantly lower than those for the last eight sentences, and the box plots showed less variability, suggesting that human transcribers perform significantly better on daily utterances than on literary excerpts.

The p-value of the Mann-Whitney U test was used to determine whether there was a statistically significant difference between the medians of two independent samples. For WHISPER, the p-value is 0.7209, much higher than 0.05, which indicates

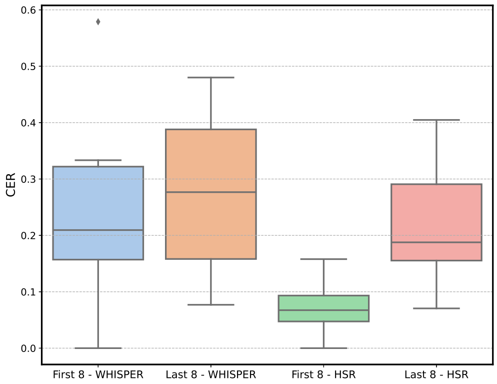

Figure 3: The CER of the first eight sentences (daily utterances) and the last eight sentences (literary recitations) in WHISPER and HSR, respectively

that there is insufficient evidence for a difference in ASR performance between daily utterances and literary excerpts.

For HSR, the p-value is 0.0047, much less than 0.05, so there exists a significant difference in transcription performance between human transcribers on daily utterances and literary excerpts. This is consistent with the intuitive conclusions we draw from the Figure.

4 Conclusion

The aim of this study was to explore the ability of the Whisper ASR system and human transcribers (HSR) in transcribing whispered Mandarin speech, focusing on a comparative evaluation across 16 sentences comprising both daily utterances and literary excerpts.

The Character Error Rate (CER) served as the primary metric for assessing transcription accuracy. Whisper ASR exhibited higher mean CER values, indicating a lower overall accuracy in the transcription of whispered Mandarin speech compared to human transcribers.

The Whisper ASR system's performance did not show a significant difference when transcribing daily utterances versus literary excerpts. However, the human transcribers demonstrated superior performance, particularly with daily utterances, which can be attributed to their ability to leverage contextual cues and linguistic knowledge more effectively than the Whisper ASR systems.

In summary, human performed better than the Whisper ASR system. This highlights the ability of the human transcriber to deal with the nuances of whispered Mandarin, especially with confusing homophones, as well as the ability to understand literary texts, which remains a challenge for the ASR system.

5 Discussion

The results of the study show that while ASR systems like Whisper have made great progress in general speech recognition, there is still a considerable gap in their ability to process whispered speech.

Further advances in ASR technology will be necessary to approach the flexibility and comprehension skills demonstrated by human transcription. This should focus on improving the ability to interpret contextual information and adjust transcription results accordingly, which is crucial for the accurate transcription of whispered Mandarin speech with absent acoustic cues.

In addition, attention should also be focused on the problem of "homophone" confusion in whispered Mandarin. Future research directions in this area could focus on exploring advanced algorithms that are able to distinguish subtle pitch changes in the absence of typical articulations. For example, a research by Xueqin et al. (2016) has proposed spectrum sparse-based approaches for feature extraction that have shown promise in improving whispered speech recognition.

Research on adapting the structure of ASR models to recognize different speech types, such as children's speech investigated by Jain et al. (2023) also reminds us of the importance of model fine-tuning.

And most importantly, the effectiveness of an ASR system depends on its training dataset, so the inclusion of a variety of speech samples (including whispers) is also crucial to the development of a robust and versatile ASR system. In this regard, apart from a Mandarin whisper database proposed by Lee et al. (2014), there has not yet been rich research on the subject. However, some recent articles, e.g. by Lin et al. (2023), have proposed some methods for building a corpus of whispered speech using a normal corpus with specific processing method, which might compensate for the shortage of data in this direction.

The goal of addressing these challenges is to close the gap between human and machine speech recognition capabilities, paving the way for ASR systems to accurately transcribe Mandarin whispered speech, enhance communication for people with vocal cord impairments, and provide reliable transcription in specialized scenarios.

References

Marcely Zanon Boito, Laurent Besacier, Natalia Tomashenko, and Yannick Estève. 2022. A study of gender impact in self-supervised models for speech-to-text systems. *arXiv preprint arXiv:2204.01397*.

Charles Chang and Yao Yao. 2007. Tone production in whispered mandarin. *UC Berkeley Phonology Lab Annual Report*, pages 326–329.

Man Gao and John H Esling. 2003. Articulatory features of tones in whispered chinese. In *Proceedings of the 15th International Congress of Phonetic Sciences*, volume 3, pages 2629–2632.

Anthony Holbrook and Hsiao-Tung Lu. 1969. A study of intelligibility in whispered chinese.

Rishabh Jain, Andrei Barcovschi, Mariam Yiwere, Peter Corcoran, and Horia Cucu. 2023. Adaptation of whisper models to child speech recognition. *arXiv preprint arXiv:2307.13008*.

Pei Xuan Lee, Darren Wee, Hilary Si Yin Toh, Boon Pang Lim, Nancy F Chen, and Bin Ma. 2014. A whispered mandarin corpus for speech technology applications. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Zhaofeng Lin, Tanvina Patel, and Odette Scharenborg. 2023. Improving whispered speech recognition performance using pseudo-whispered based data augmentation. *arXiv preprint arXiv:2311.05179*.

Gang Lv and Heming Zhao. 2010. Acoustic analyses of whispered mandarin. In *2010 3rd International Congress on Image and Signal Processing*, volume 7, pages 3486–3489. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Min Xu, Jing Shao, Hongwei Ding, and Lan Wang. 2022. Acoustic-perceptual correlates of whispered mandarin consonants. In *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 195–199. IEEE.

Chen Xueqin, Zhao Heming, and Fan Xiaohe. 2016. Performance analysis of mandarin whispered speech recognition based on normal speech training model. In *2016 Sixth International Conference on Information Science and Technology (ICIST)*, pages 548–551. IEEE.