Comparison of Automatic Speech Recognition Systems

Joshua Y. Kim, Chunfeng Liu, Rafael A. Calvo, Kathryn McCabe, Silas C. R. Taylor, Björn W. Schuller, and Kaihang Wu

Abstract High-quality transcription systems are required for conversational analysis systems. We compared two manual transcribers with five automatic transcription systems using video conferences from a medical domain and found that (1) manual transcriptions significantly outperformed the automatic services, and (2) the automatic transcription of YouTube Captions significantly outperformed the other ASR services.

Keyword Speech recognition

1 Introduction

Conversational analysis systems require high-quality transcription systems to extract the verbatim transcripts. The verbatim transcripts could then be used to train deep learning models as a separate modality in addition to audio and video streams [9, 10, 21, 24, 34], or the transcripts can be weaved together with other modalities to form a multimodal narrative that is human-centric [15, 16] and facilitate conversation visualization [14]. Although ASR systems are continually improving, there is little work that compares the performance of the widely available commercial systems.

J. Y. Kim · K. Wu
The University of Sydney, Sydney, Australia

C. Liu
Hello Sunday Morning, Surry Hills, Australia

R. A. Calvo (✉) · B. W. Schuller
Imperial College London, London, United Kingdom
e-mail: r.calvo@imperial.ac.uk

K. McCabe
University of California, Los Angeles, CA, USA

S. C. R. Taylor
University of New South Wales, Kensington, NSW, Australia

S. Stoyanchev et al. (eds.), *Conversational AI for Natural Human-Centric Interaction*,
Lecture Notes in Electrical Engineering 943,
https://doi.org/10.1007/978-981-19-5538-9_8

In this paper, we aim to provide empirical evidence on the performance of five ASR providers—namely, Google Cloud, IBM Watson, Microsoft Azure, Trint, and YouTube.

2 Related Works

ASR systems have seen significant improvements over the past few years [33]. The Switchboard telephone speech dataset is often used to benchmark the performance of the transcription [28]. Microsoft Research reports a WER of 5.1% on the NIST 2000 Switchboard task [33]. IBM Research reports 6.6% WER on the Switchboard subset of the Hub5 2000 evaluation test set [28]. Google Research reports a 6.7% WER on a 12,500-hour voice search dataset and 4.1% on a dictation task [3], both of which are not part of the Switchboard telephone speech dataset. Some works [23] relied on such published statistics which could be misleading.

Applications of the ASR in teleconferences are more challenging as the speaker is speaking at some distance from the microphone—this is known as distant speech recognition. Research on distant speech recognition includes the application of convolutional neural networks (CNN) [17] on the Augmented Multi-party Interaction (AMI) meeting corpus [2], where a word error rate of 40.9% was achieved with a single distant microphone [30]. More recently, Renals and Swietojanski [26] used the AMI corpus to compare ASR approaches using multiple distant microphones and individual headset microphones. The difference in WER is significant—the eight distant microphone setup achieved a WER of 52.0% verses 29.6% (individual microphone). The distant microphone performance was recently surpassed by UNet++ (WER: 42.2%) [35].

Këpuska and Bohouta [13] performed a comparison between CMU Sphinx, Microsoft Speech and Google Cloud and found that the Google Cloud API performs the best with a mean WER of 9%. In that study, the authors used the Texas-Instruments/Massachusetts Institute of Technology (TIMIT) corpus [5]. In this study, we expand the number of online transcription services for comparison and utilize a dataset that is intended to mirror real-world doctor-patient interviews, which has been increasing [7, 8, 25].

3 EQClinic Dataset

3.1 Data Collection

This study used data from the EQClinic platform [20]. Students in an Australian medical school were required to complete the program aimed at improving clinical communication skills during their first and second years of study. Within the

EQClinic platform, the students were required to complete at least one medical consultation with a simulated patient on the online video conferencing platform EQClinic [19]. Participants consist of twelve second-year undergraduate medical students (six female and six male) and two simulated patients (SP, one male and one female). The two SP were professional actors, recruited online and paid AUD35 per hour for participating. The study was approved by the UNSW Human Research Ethics Committee (Project Number HC16048).

3.2 Data Analysis

For each consultation, EQClinic generated one MP4 video recording for each speaker with a resolution of 640x480 pixels and a frame rate of 25fps. Audio recordings were extracted using the FFMpeg software. We selected twelve interview sessions randomly and we ensured that there are three videos for each of the possible gender pairing (male-male, male-female, female-male, and female-female).

The duration of these sessions ranges from 12 to 18 min (mean duration (SD) = 14.8 (2.0)). Each session contained two videos, and each of these video pairs had one speaker (the student or the SP). Each video comprised 668 to 1705 words (mean words (SD) = 1187 (316). In total, 24 videos and a total of 28,480 words were analyzed. Disfluencies like "um" are captured in the transcripts. We sent these 24 videos to seven transcription services—two of which were manual, and the other five were ASR systems. The costs and file formats required for transcription are summarized in Table 1 in the supplementary material. Although the file formats differ, we are interested in also testing services that could not accept videos as inputs.

For the two manual transcription services, one was an independent professional transcriber (CB), and the other was from an online network of hand-picked freelancers available at Rev.com (Rev). For both manual transcription services, video files were provided in the MP4 format for transcription.

Table 1 Summary of required file formats and costs for transcription services. CB denotes the independent professional transcriber. Rev denotes transcribers from Rev.com

Service	File Format	USD per video minute
Manual (CB)	MP4 Video	1.920
Manual (Rev)	MP4 Video	1.500
Automatic (Google Cloud)	Mono-channel FLAC audio	0.048
Automatic (IBM Watson)	Mono-channel FLAC audio	0.020
Automatic (Microsoft Azure)	Mono-channel WAV audio	0.008
Automatic (Trint)	MP4 Video	0.025
Automatic (YouTube)	MP4 Video	0.000

Each of the five ASR services (Google Cloud, IBM Watson, Microsoft Azure, Trint, and YouTube) required a different format of the input file to perform the transcription. For all of the five ASR services, we elected to perform asynchronous transcription service calls because YouTube and Trint do not offer synchronous transcription service calls. Synchronous service calls refer to the ability of the ASR to stream text results, immediately returning text as it is recognized from the audio.

We compared the quality of transcripts gathered from different transcription services. Word Error Rate (WER) is a popular performance measure in automatic speech recognition [4]. We first determined which of the two sets of manual transcriptions would be the reference transcript. We then compared the five sets of automatic transcriptions against this reference transcript to identify the best-performing ASR system. We posit that if multiple transcribers produce similar transcripts as indicated by low WER, they have likely converged on the correct transcription [27]. Therefore, the set of manual transcriptions with the lower WER as compared with each of the five sets of automatic transcription was considered the best choice as the set of reference transcripts.

In our analysis, ten pairwise WER were generated between each of the five hypothesis transcripts and the two manual sets of transcripts (Manual CB and Manual Rev) [1]. For the ten pairwise WER estimates, we determined which of the WER-reference pairs were statistically significantly different. To do that, we needed the 95% WER confidence interval. Since the assumption of independent error rates [6] are not applicable when we fixed the hypothesis transcript to be from one ASR service, we elected to use bootstrapping to generate confidence intervals. The bootstrap technique is used to quantify the uncertainty associated with the WER in our application and involves creating 10,000 bootstrap datasets [29] produced by random sampling with replacement [12]. With the 10,000 bootstrap samples, we computed an average WER. Then, we created the 95% WER confidence interval by eliminating the top and bottom 2.5% values.

After establishing the set of manual transcription was of higher quality, we used this set of manual transcription as our reference transcription to examine the WER of all other transcription services. Next, we investigated whether differences in WER performance between each transcription service were statistically significant. We used one set of reference transcriptions and computed the difference in WER between service X and service Y for each of the 24 transcriptions. Similarly, we then bootstrapped the differences in WER between the two services (service X and Y) and generated the confidence intervals for the differences using 10,000 samples.

4 Results

Figure 1 compares the hypothesis transcripts and each of the two manual transcripts (Manual CB and Rev). We found that the two sources of manual transcription did not differ significantly. For a given set of hypothesis transcripts (generated by selected

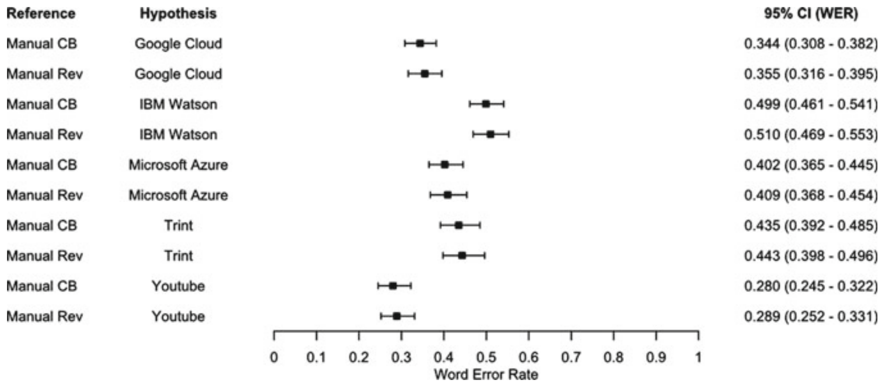

Fig. 1 Forest plot of WER of automatic transcription services, using two sets of reference transcripts from each of the two manual transcription services (Manual CB and Manual Rev)

ASR systems), the confidence interval of Manual CB does not differ from Manual Rev.

We selected Manual CB as the reference transcript and completed a pairwise analysis for the remaining transcription services comparing the quality of all of the transcription services. Figure 2 shows the differences in WER between service pairs. For each of the pairwise differences in WER at a video level, we performed bootstrapping to generate 10,000 samples and compute the 95% confidence intervals.

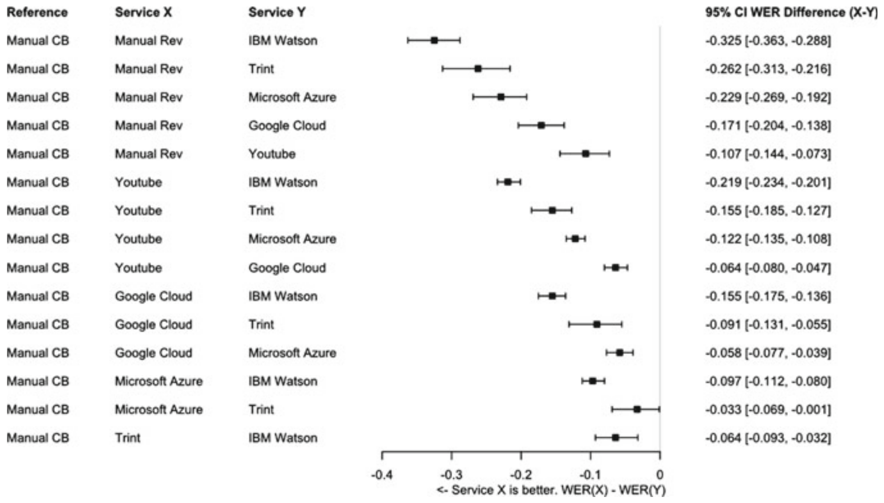

Fig. 2 Forest plot of pair-wise differences in WER of the various transcription services. Only comparisons where Service X is better are illustrated. The plot is ordered by the best performing service in Service X, followed by the mean WER difference between Service X and Service Y

Figure 2 shows that the Manual Rev was the best transcription service, exhibiting significantly better performance relative to the other transcription services. We found that manual transcription was better than all of the automatic transcription services and all pair-wise differences are statistically significant. Amongst the automatic transcription services, we found that YouTube exhibited significantly better performance relative to the other automatic transcription services, and all pair-wise differences are statistically significant.

5 Discussion

Amongst the automatic transcription services, YouTube offers the most accurate automated transcription service, though this is not as accurate as the professional transcription service. We found that the two manual transcriptions demonstrated similar quality with a WER of 17.4%. This is higher than the WER of previous studies based on the standard telephone audio recording dataset where the manually transcribed WER was between 5.1% and 5.9% [32].

Several potential factors may cause the lower accuracy (that is high WER) of human/manual transcription in this study. First, the conversation environment could have influenced the recording quality. The WER in Xiong et al.'s work [32] was tested based on telephone audio recordings, in which the microphone was located near the speaker. However, the medical conversations of this study were conducted over video conferencing on PC or tablets. There was likely to be greater variability in recording quality as some of the speakers were likely seated further away from the microphone. In addition, the medical conversation could be held anywhere; therefore environmental noise and audio feedback in the conversation may have impacted the human transcription. The WER of 17.4% is more similar to benchmarks tackling ASR in far-field, noisy environments [18, 31]. Lastly, we posit that the medical nature of the conversations in our study caused the higher WERs from both the manual transcribers and ASR services. This is in line with the literature. For example, Mani et. al [22] found that Google ASR substituted "taste maker" with "pacemaker", and Henton [11] found that ASR and humans could make mistakes when transcribing drugs (e.g., Feldene vs. Seldane).

Although human transcription was not perfect, we found that human accuracy was higher than the tested ASR systems. Of the tested ASR systems, the YouTube Captions service achieved the highest accuracy. These results provided us with a preliminary understanding of the transcription qualities of human and ASR systems on video conferencing data. Our results are in line with Këpuska and Bohouta [13] who found that Google Cloud Speech-To-Text outperformed Microsoft Speech Services.

6 Conclusion

We have provided the first comparison of the performance of automated transcription services in the domain of dyadic medical teleconsultation. We found that manual transcription significantly outperformed the automatic services, and the automatic transcription of YouTube Captions significantly outperformed the other ASR services. There are three limitations to this work. Firstly, the evidence from this paper is limited to a highly professional scenario (medical consultation). Whilst we posit that the finding may be generalizable to non-professional settings, it is left for future work in this area. Secondly, we only transcribed a small number of videos due to financial constraints. Lastly, the systems are continuously improving and this study is only a snapshot of the current state. Future research could compare the results of snapshots at different time periods.

Acknowledgements The authors thank Hicham Moad S for his help rendered in scripting for the Microsoft Azure API, and Marriane Makahiya for typesetting. RAC is partially funded by the Australian Research Council Future Fellowship FT140100824.

References

1. Belambert: Asr-evaluation. https://github.com/belambert/asr-evaluation
2. Carletta J (2007) Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus. Lang Resour Eval 41(2):181–190
3. Chiu CC, Sainath TN, Wu Y, Prabhavalkar R, Nguyen P, Chen Z, Kannan A, Weiss RJ, Rao K, Gonina E, et al (2018) State-of-the-art speech recognition with sequence-to-sequence models. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 4774–4778
4. Gaikwad SK, Gawali BW, Yannawar P (2010) A review on speech recognition technique. Int J Comput Appl 10(3):16–24
5. Garofolo JS, Lamel LF, Fisher WM, Fiscus JG, Pallett DS (1993) Darpa timit acoustic-phonetic continuous speech corpus cd-rom. nist speech disc 1-1.1. NASA STI/Recon technical report n 93, 27403
6. Gillick L, Cox SJ (1989) Some statistical issues in the comparison of speech recognition algorithms. In: International conference on acoustics, speech, and signal processing. IEEE, pp 532–535
7. Gopal RK, Solanki P, Bokhour B, Skorohod N, Hernandez-Lujan D, Gordon H (2021) Provider, staff, and patient perspectives on medical visits using clinical video telehealth: a foundation for educational initiatives to improve medical care in telehealth. J Nurse Practit
8. Gordon HS, Solanki P, Bokhour BG, Gopal RK (2020) “i’m not feeling like i’m part of the conversation” patients’ perspectives on communicating in clinical video telehealth visits. J Gen Intern Med 35(6):1751–1758
9. Hazarika D, Poria S, Mihalcea R, Cambria E, Zimmermann R (2018) Icon: interactive conversational memory network for multimodal emotion detection. In: Proceedings of the 2018 conference on empirical methods in natural language processing, pp 2594–2604
10. Hazarika D, Poria S, Zadeh A, Cambria E, Morency LP, Zimmermann R (2018) Conversational memory network for emotion recognition in dyadic dialogue videos. In: Proceedings of the conference. Association for computational linguistics. North American Chapter. Meeting, vol 2018, p 2122. NIH Public Access

11. Henton C (2005) Bitter pills to swallow. asr and tts have drug problems. Int J Speech Technol **8**(3), 247–257
12. James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning, vol 112. Springer
13. Këpuska V, Bohouta G (2017) Comparing speech recognition systems (microsoft api, google api and cmu sphinx). Int J Eng Res Appl 7(03):20–24
14. Kim JY, Calvo RA, Yacef K, Enfield N (2019) A review on dyadic conversation visualizations-purposes, data, lens of analysis. arXiv:1905.00653
15. Kim JY, Kim GY, Yacef K (2019) Detecting depression in dyadic conversations with multi-modal narratives and visualizations. In: Australasian joint conference on artificial intelligence. Springer, pp 303–314
16. Kim JY, Yacef K, Kim G, Liu C, Calvo R, Taylor S (2021) Monah: multi-modal narratives for humans to analyze conversations. In: Proceedings of the 16th conference of the European chapter of the association for computational linguistics: main volume, pp 466–479
17. LeCun Y, Bengio Y et al (1995) Convolutional networks for images, speech, and time series. Handbook of Brain Theory and Neural Netw 3361(10):1995
18. Li J, Zhao R, Chen Z, Liu C, Xiao X, Ye G, Gong Y (2018) Developing far-field speaker system via teacher-student learning. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 5699–5703
19. Liu C, Lim RL, McCabe KL, Taylor S, Calvo RA (2016) A web-based telehealth training platform incorporating automated nonverbal behavior feedback for teaching communication skills to medical students: a randomized crossover study. J Med Internet Res 18(9):e246
20. Liu C, Scott KM, Lim RL, Taylor S, Calvo RA (2016) Eqclinic: a platform for learning communication skills in clinical consultations. Med Educ Online 21(1):31801
21. Majumder N, Poria S, Hazarika D, Mihalcea R, Gelbukh A, Cambria E (2019) Dialoguernn: An attentive rnn for emotion detection in conversations. In: Proceedings of the AAAI conference on artificial intelligence, vol 33, pp 6818–6825
22. Mani A, Palaskar S, Konam S (2020) Towards understanding asr error correction for medical conversations. In: Proceedings of the first workshop on natural language processing for medical conversations, pp 7–11
23. Miao K, Biermann O, Miao Z, Leung S, Wang J, Gai k (2020) integrated parallel system for audio conferencing voice transcription and speaker identification. In: 2020 international conference on high performance big data and intelligent systems (HPBD&IS). IEEE, pp 1–8
24. Mittal T, Bhattacharya U, Chandra R, Bera A, Manocha D (2020) M3er: multiplicative multi-modal emotion recognition using facial, textual, and speech cues. In: AAAI, pp 1359–1367
25. Nielsen C, Agerskov H, Bistrup C, Clemensen J (2020) Evaluation of a telehealth solution developed to improve follow-up after kidney transplantation. J Clin Nurs 29(7–8):1053–1063
26. Renals S, Swietojanski P (2017) Distant speech recognition experiments using the AMI corpus. New Era for robust speech recognition, pp 355–368
27. Roy BC, Roy DK, Vosoughi S (2010) Automatic estimation of transcription accuracy and difficulty
28. Saon G, Kuo HKJ, Rennie S, Picheny M (2015) The IBM 2015 english conversational telephone speech recognition system. arXiv:1505.05899
29. Siohan O, Ramabhadran B, Kingsbury B (2005) Constructing ensembles of asr systems using randomized decision trees. In: Proceedings.(ICASSP'05). IEEE international conference on acoustics, speech, and signal processing, 2005. vol 1. IEEE, pp I–197
30. Swietojanski P, Ghoshal A, Renals S (2014) Convolutional neural networks for distant speech recognition. IEEE Signal Process Lett 21(9):1120–1124
31. Tang Z, Meng HY, Manocha D (2020) Low-frequency compensated synthetic impulse responses for improved far-field speech recognition. In: ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 6974–6978
32. Xiong W, Droppo J, Huang X, Seide F, Seltzer M, Stolcke A, Yu D, Zweig G (2016) Achieving human parity in conversational speech recognition. arXiv:1610.05256

33. Xiong W, Wu L, Alleva F, Droppo J, Huang X, Stolcke A (2018) The microsoft 2017 conversational speech recognition system. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 5934–5938
34. Zadeh A, Liang PP, Mazumder N, Poria S, Cambria E, Morency LP (2018) Memory fusion network for multi-view sequential learning. In: Proceedings of the AAAI conference on artificial intelligence, vol 32
35. Zhao T, Zhao Y, Wang S, Han M (2021) Unet++-based multi-channel speech dereverberation and distant speech recognition. In: 2021 12th international symposium on Chinese spoken language processing (ISCSLP). IEEE, pp 1–5