DLSU Research Congress 2024
De La Salle University, Manila, Philippines
June 20 to 22, 2024

Evaluating the Performance of a Commercial Speech-to-Text Application for Filipino Language as an Aid in Encoding Healthcare Data

Ronald Pascual*, Adrienne Apuyod[1], Kathleen Bainto[2], Mia Samantha Panit[3], Julienne Llamado[4]
[1]adrienne.apuyodegg@gmail.com
[2]kathleen_gayle_bainto@dlsu.edu.ph
[3]mia_samantha_panit@dlsu.edu.ph
[4]juliennedllamado@gmail.com
*Corresponding Author: ronald.pascual@dlsu.edu.ph

Automatic speech recognition (ASR) is a growing Artificial Intelligence platform being implemented in different sectors of the world. This study focuses on the evaluation of the performance of a state-of-the-art commercial ASR and the characterization of a Filipino speech corpus for healthcare. The dataset employed for this study is a Filipino speech corpus collected from female adult speakers, and contains simulated responses to health assessment questions on physical wellness. The performance evaluation of the commercial ASR was based on word error rate (WER), and involved the analysis of the types of errors found in using the ASR for Filipino language. Test results showed that the ASR has an average WER of 5.43% and that word substitution including medications and medical-related terms is the most prevalent type of error.

Key Words: automatic speech recognition; speech-to-text; healthcare; Filipino speech corpus

1. INTRODUCTION

1.1 Background of the Study

Since the world is evolving day-by-day, people are innovating technologies that somewhat make lives more convenient or make way for new discoveries. One of the advancements that has proven to be very useful in today's world is Artificial Intelligence (AI). Artificial Intelligence is the simulation of human intelligence processes by computer systems and machines, which gathers data to analyze and create algorithms to make patterns. AI requires specialized hardware and software for writing machine learning algorithms. In the recent years, machine learning and deep learning have found broad applications and have exhibited good performance in various domains including automatic speech recognition (ASR) and natural language processing (Kelley, 2024; Laskowski & Tucci, 2024).

With the introduction and availability of ASR technologies, users of information systems can now recite responses or entries rather than typing them into a keyboard. Advanced ASR systems let users input direct questions or answers, like a request for driving instructions or the phone number of a hotel in a certain location. As a result, there are fewer decision points in the menu navigation process.

ASR can benefit both patients and healthcare professionals by making the application interface and interaction more efficient. According to Hodgson et al. (2017), the use of Electronic Health Records (EHRs) is becoming mandatory for clinical documentation, and now widely implemented around the world. Speech recognition can help doctors become more productive, by allowing notes to be translated from speech to text, instead of having them be taken by hand. This allows doctors to spend more time serving patients. Payne et al. (2018) presented the development and design of a

mobile application-based system that is integrated with a commercial ASR software and EHR for physicians who prefer voice interface over typing using a keyboard. More recently, Fox et al. (2021) presented their study on the evaluation of clinical transcription accuracies of various methods for children's narrative that included the use of Google Cloud Speech ASR.

Big data analytics has many benefits but it raises many challenges to be faced. In the study presented by Abouelmehdi et al. (2018), they stated that the most important and common issues in handling big amounts of data is weak security, privacy breaches, and the lack of technical support. Joseph et al. (2020) investigated on the impact and issues of implementing ASR technology in generating clinical documentations by nurses. Common issues and challenges in using ASR technology for clinical documentations include the following: software reliability, ASR accuracy, high initial costing, training requirements, interface modification, transcription error editing and workflow (Joseph et al., 2020; Goss et al., 2019). These issues and challenges need to be addressed before implementing these ASR-based platforms in the healthcare system.

The preceding studies cited in this paper focused on the use of English language as it is a universal language that is supported by commercial ASR systems and the language option that exhibits one of the best performances. A few years ago, Google Cloud Speech launched its ASR that supports Filipino language. However, the performance of Google's ASR for Filipino still remains to be evaluated especially for its application in transcribing or encoding healthcare data. It is also worth noting that a low-resourced custom-built ASR system for Filipino and Bisaya languages for application in a healthcare chatbot was presented in the studies by Pascual et al. (2023a) and Pascual et al. (2023b).

This study focuses on the evaluation of the performance of a state-of-the-art commercial ASR in the Filipino language, as an aid in encoding healthcare data. The lack of healthcare workers in the Philippines and the healthcare workers' increased workload since the pandemic started in 2020 have motivated the authors to conduct a study on the use of ASR technology as an alternative interface that can help in efficiently transcribing speech for health records documentation purposes. This research presents an insight on the performance of Google Cloud Speech-to-Text (ASR) in terms of transcribing audio files containing Filipino speech in the healthcare domain.

2. METHODOLOGY

2.1 Filipino Speech Corpus

The authors utilized a dataset that is a subset of a Filipino speech corpus collected for a multilingual healthcare chatbot research project by Azcarraga et al. (De La Salle University, 2021) and was also reported in Pascual et al. (2023a). The speech corpus subset contains speech collected from 8 female adult speakers. All the volunteer speakers consented to the speech recording activity, and all their personal identities were anonymized in the speech corpus.

We used a total of 8,725 segmented audio files containing a total of 15,353 words. All speech recordings are sampled at 16 KHz and stored in uncompressed format (*.wav) using PCM-signed 16-bit mono channel encoding. Each segmented audio file contains a simulated response to one of the 381 general physical wellness assessment questions asked by a nurse and can be categorized as shown in Table 1 (Pascual et al., 2023a). All the audio files have been transcribed to word- and phoneme-levels, with the word-level transcription serving as the gold standard or reference transcript.

Table 1. General Physical Wellness Assessment Question Categories and Information (Pascual et al., 2023a).

Category	Information
Chief Complaint	Chief Complaint Fever, Cough, Colds, Pain, Nausea
Review of Past Medical History	Past Surgeries, Hospitalizations, Familial History
Medications	Generic and brand names of medications
Instrumental Activities of Daily Living Scale	Feeding, Grooming, Toileting, Dressing, General Hygiene, Transferring, Playtime, Sleep
Review of Systems	General Appearance, Allergies, Skin, Neuro, Ears, Eyes, Nose, Mouth and Throat, Heart and Lungs, Gastrointestinal, Genitourinary, Multiple Sclerosis, Endocrine

The organization of the audio files in the speech corpus can be verified through the file name format: *rec_XXX_YYY_ZZZ_AAA.wav*, where XXX = speaker number, YYY = question number, ZZZ = language code, and AAA = response segment number.

2.2 Speech-to-Text and Error Rates

As mentioned earlier in the introduction, Google Cloud Speech-to-Text (STT) introduced a few years back one of the best available AI-powered automatic speech recognition (ASR) software application that supports Filipino language. In this study, the performance of Google Cloud's ASR for Filipino is evaluated for application in transcribing or encoding healthcare data using the Filipino speech corpus described in the previous section. We used the demo version of Google Cloud STT with default models and settings for command/search mode. Figure 1 shows a screenshot of the Google Cloud STT demo page.

Fig 1. Google Cloud Speech-to-Text

We used output text data from Google Cloud STT and compare each of these to the reference transcripts in order to analyze the ASR errors made while decoding words contained in the Filipino speech corpus for healthcare applications. The types of errors produced by Google Cloud ASR may include insertion, deletion, and substitution. Substitutions occur when an expected word in the reference transcript gets replaced by a different word. Insertions occur when a word is added, even if the word was never uttered in the speech according to the reference transcript. Deletions occur when a word is left out of the reference transcript.

We used the word error rate (WER), a standard metric for measuring the performance of an ASR system. WER is defined as the sum of the number of word substitutions, insertions, and deletions, divided by the total number of words in the test dataset. Equation 1 shows the standard WER formula in percentage. We may also calculate the specific error rates according to type such as substitution error rate, insertion error rate, and deletion error rate using the formula provided in Equation 2. Google Cloud STT's performance in terms of the aforementioned error rates are presented in the next section.

$$WER = \frac{\sum(S + I + D)}{Total\ number\ of\ words} \times 100\% \qquad \text{(Eq. 1)}$$

where: WER = word error rate;
S = number of substitutions;
I = number of insertions;
D = number of deletions.

$$Error\ Rate_{(k)} = \frac{\sum Errors(k)}{Total\ number\ of\ words} \times 100\% \qquad \text{(Eq. 2)}$$

where: Error Rate $_{(k)}$ = type-k error rate;
k = error type {S | I | D}.

3. RESULTS AND DISCUSSION

In this section, we present and discuss the resulting performance of Google Cloud Speech-to-Text (STT) in terms of word error rates (WER) and the specific error rates for each error type.

We calculated the individual WER for each of the 8 speakers using the formula given in Equation 1. Table 2 shows the calculated WERs per speaker. The best performance obtained is 2.6% WER using speaker #8 data, while the worst performance gave a WER of 9.6% using speaker #4 data. It is worth noting that is a relatively small variance in WER, and that no WER for any of the 8 speakers exceeded 10%. The overall average WER for all the 8 speakers is 5.43%. This result agrees to the best WER of 3.96% that Pascual et al. (2023b) reported in their work regarding the development of a TDNN-HMM Filipino ASR system for a healthcare chatbot using the same speech corpus. Generally, in the field of speech recognition, a WER of 5% or less may be considered as a good performance for an ASR system as compared to a human's ability to decode spoken words from an audio signal.

Table 2. WERs for the 8 Speakers in the Filipino Speech Corpus

Speaker No.	Word Error Rate (WER %)
1	7.5%
2	2.8%
4	9.6%
6	5.4%
8	2.6%
9	3.9%
11	4.8%
12	4.6%
Overall	5.43%

Figure 2 shows the distribution of the audio files by speaker in the test set, while Table 3 gives the number of words spoken by speaker. Note that we used a total of 8,725 segmented audio files for all the 8 speakers in the dataset while there is a total of 15,353 words in the same data set. This would mean that on the average, there are around 2 words per response or audio file. However single-word responses such as a yes/no response, and even 3- or 4-word responses are also common in the dataset.

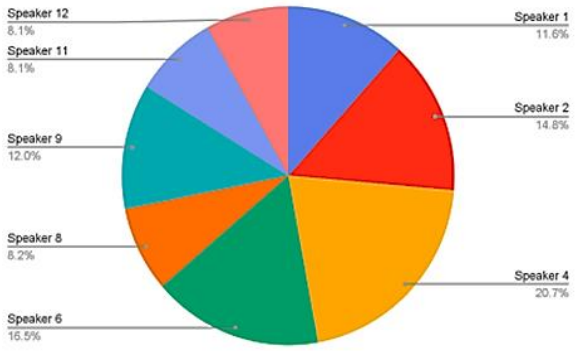

Fig 2. Distribution of the number of audio files by speaker in the test set.

Table 3. Number of Words Spoken by Speaker

Speaker No.	Number of Words Spoken
1	2020
2	2667
4	2765
6	1656
8	1422
9	2081
11	1380
12	1362

To expound on the nature of errors that the Google Cloud ASR obtained in decoding of words contained in the Filipino speech corpus, we present some examples of outputs with errors of various types. Table 4 shows some selected reference-hypothesis text pairs together with the edit operation error codes. The reference text is the ground truth transcription while the hypothesis text is the output of the Google Cloud ASR decoder. The last line in each row contains the edit operations and tells whether each word in the reference text is decoded correctly (C), or is a substitution (S), an insertion (I), or a deletion (D) error. We may note from the examples that some medication names and even medical-related terms can sometimes be erroneously decoded by the ASR. These kinds of errors may be critical for healthcare data encoding application especially when the meaning of the hypothesized text output is highly different from the meaning of the words or phrase that were spoken by a patient.

For further analysis of the frequency of errors, Table 5 presents the actual word error counts for each type of error, namely substitution, deletion, and insertion, for each of the respective speakers in the dataset. The most prevalent type of error is the substitution error, which accounts to 75.6% of all the word errors. The deletion errors and the insertion errors account respectively to 13.2% and 11.3% of the total word errors.

Table 4. Examples of Google Cloud ASR Outputs with Errors.
("Ref" = reference text; "Hyp" = ASR's hypothesized or output text; "Op" = edit operations with error codes; C = correctly decoded; S = substitution error; I = insertion error; D = deletion error.)

Category	Transcripts and Error Codes
General Wellness	(File: rec_001_026_FIL_004.wav) Ref : kain ng prutas Hyp : kaining none prutas Op : S S C
Fever	(File: rec_001_050_FIL_006.wav) Ref : lampas 38 degrees Hyp : lampas 38 *** Op : C C D
General Wellness	(File: rec_004_013_FIL_014.wav) Ref : nakalimutan ko Hyp : nakalimutan ko na Op : C C I
Cough/Colds	(File: rec_001_028_FIL_002.wav) Ref : loviscol Hyp : ruby school Op : S S
Cough/Colds	(File: rec_002_044_FIL_001.wav) Ref : cetirizine Hyp : city rizal Op : S S
Ear	(File: rec_002_134_FIL_001.wav) Ref : pag malakas na tunog o sigaw Hyp : pag malakas na tunog ** *** Op : C C C C D D
Allergy / Skin	(File: rec_001_018_FIL_002.wav) Ref : pangpahid sa balat Hyp : ang sakit sa balat Op : S I C C
Cough/Colds	(File: rec_004_016_FIL_002.wav) Ref : mag-inhaler Hyp : magin healer Op : S S

Table 5. Errors Counts Per Speaker and Error Type

Speaker No.	Substitution	Deletions	Insertions
1	98	31	23
2	66	6	5
4	189	33	44
6	61	18	11
8	29	4	4
9	64	13	6
11	63	3	1
12	61	2	0

4. CONCLUSION

In this paper, the authors have presented an evaluation of the performance of a state-of-the-art commercial speech-to-text (STT) application software in encoding healthcare data from spoken words in Filipino language. The Filipino speech corpus used for this study was presented to contain a total of 8,725 audio files or 15,353 words uttered by 8 different speakers. The transcripts of the audio files contain simulated responses to health assessment questions on physical wellness. The performance evaluation of the aforementioned commercial STT application software was implemented using word error rates (WER) and the analysis of the types of errors found in the output text given by the ASR decoder. Test results showed that the commercial ASR has an average WER of 5.43%, which can be considered as a good performance comparable to that of a human's ability to decode spoken words from an audio signal. Care must be taken however for cases where the errors involve wrong transcriptions of medication names and medical-related terms. Moreover, we found that the most prevalent type of word error is the substitution error, which accounts to 75.6% of all the word errors. The deletion errors and the insertion errors account respectively to 13.2% and 11.3% of the total word errors.

The good performance of the available commercial STT application software for Filipino shows a significant promise in changing how healthcare data is encoded. It can be a helpful tool for patients, academics, and healthcare professionals since it is fairly accurately in converting spoken words into text. Numerous advantages may be seen with the use of ASR in healthcare settings, including improved clinical documentation, higher output, improved accessibility for people with disabilities, and the potential to accelerate medical research.

In the future, it would be interesting to conduct a similar study to the one presented here as more Philippine languages become supported by these commercial ASR systems. As of this writing, Pascual et al. (2013a; 2013b) are still currently doing related work on the development of ASR systems for

Bisaya-Cebuano language for application in the healthcare domain.

5. ACKNOWLEDGMENTS

The authors would like to thank Dr. Judith Azcarraga and her team for providing access to the Filipino Speech Corpus used in this study. The authors also thank their friends, family, and acquaintances for providing utmost support and motivation throughout the entire study.

6. REFERENCES

Abouelmehdi, K., Beni-Hessane, A., & Khaloufi, H. (2018). Big healthcare data: preserving security and privacy. Journal of Big Data, 5(1).

De La Salle University (2021). Two DLSU projects get DOST support to boost country's AI, data science. 2401: The official newsletter of De La Salle University, 52(11), 1-2.

Fox, C. B., Israelsen-Augenstein, M., Jones, S., & Gillam, S. L. (2021). An Evaluation of Expedited Transcription Methods for School-Age Children's Narrative Language: Automatic Speech Recognition and Real-Time Transcription. Journal of Speech, Language and Hearing Research (Online), 64(9), 3533-3548.

Goss, F. R., Blackley, S. V., Ortega, C. A., Kowalski, L. T., Landman, A. B., Lin, C. T., Meteer, M., Bakes, S., Gradwohl, S. C., Bates, D. W., & Zhou, L. (2019, October). A clinician survey of using speech recognition for clinical documentation in the electronic health record. International Journal of Medical Informatics, 130, 103938.

Hodgson, T., Magrabi, F., & Coiera, E. (2017). Efficiency and safety of speech recognition for documentation in the electronic health record. Journal of the American Medical Informatics Association, 24(6), 1127–1133.

Joseph, J., Moore, Z. E. H., Patton, D., O'Connor, T., & Nugent, L. E. (2020). The impact of implementing speech recognition technology on the accuracy and efficiency (time to complete) clinical documentation by nurses: A systematic review. Journal of Clinical Nursing, 29(13-14), 2125–2137.

Laskowski, N., & Tucci, L. (2024). What is artificial intelligence (AI)? Everything you need to know. [available online] https://www.techtarget.com.

Kelley, K. (2024). What is Artificial Intelligence and Why It Matters in 2024? [available online] https://www.simplilearn.com.

Pascual, R., Azcarraga, J., Cheng, C., Ing, J. A., Wu, J., & Lim, M. L. (2023a). Filipino and Bisaya Speech Corpus and Baseline Acoustic Models for Healthcare Chatbot ASR. Proc. of 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME). Tenerife, Canary Islands, Spain, July 2023, pp. 1-5, doi: 10.1109/ICECCME57830.2023.10253232.

Pascual, R., Azcarraga, J., & Ing, J. A. (2023b). TDNN-HMM ASR Systems on Under-Resourced Local Languages Towards Application in a Healthcare Chatbot. Proc. of 4th International Conference on Advances in Computational Science and Engineering (ICACSE 2023), Manila, Philippines, December 2023 (In Press).

Payne, T., Alonzo, W.D., Markiel, J.A., Lybarger, K., & White, A. (2018). Using voice to create hospital progress notes: Description of a mobile application and supporting system integrated with a commercial electronic health record. Journal of Biomedical Informatics, 77(1), 91-96.