Comparing Speech Recognition Services for HCI Applications in Behavioral Health

Piotr Chlebek
Ellipsis Health, Inc.
San Francisco, CA, USA
piotr@ellipsishealth.com

Elizabeth Shriberg
Ellipsis Health, Inc.
San Francisco, CA, USA
liz@ellipsishealth.com

Yang Lu
Ellipsis Health, Inc.
San Francisco, CA, USA
yang@ellipsishealth.com

Tomasz Rutowski
Ellipsis Health, Inc.
San Francisco, CA, USA
tomek@ellipsishealth.com

Amir Harati
Ellipsis Health, Inc.
San Francisco, CA, USA
amir@ellipsishealth.com

Ricardo Oliveira
Ellipsis Health, Inc.
San Francisco, CA, USA
ricardo@ellipsishealth.com

ABSTRACT

Behavioral health conditions such as depression and anxiety are a global concern, and there is growing interest in employing speech technology to screen and monitor patients remotely. Language modeling approaches require automatic speech recognition (ASR) and multiple privacy-compliant ASR services are commercially available. We use a corpus of over 60 hours of speech from a behavioral health task, and compare ASR performance for four commercial vendors. We expected similar performance, but found large differences between the top and next-best performer, for both mobile (48% relative WER increase) and laptop (67% relative WER increase) data. Results suggest the importance of benchmarking ASR systems in this domain. Additionally we find that WER is not systematically related to depression itself. Performance is however affected by diverse audio quality from users' personal devices, and possibly from the overall style of speech in this domain.

CCS CONCEPTS

• **Human-centered computing** → **Natural language interfaces**; • **Computing methodologies** → **Speech recognition**; • **Applied computing** → **Life and medical sciences**.

KEYWORDS

Speech recognition; word error rate; behavioral health; digital health; telehealth; natural language processing; depression; anxiety

ACM Reference Format:
Piotr Chlebek, Elizabeth Shriberg, Yang Lu, Tomasz Rutowski, Amir Harati, and Ricardo Oliveira. 2020. Comparing Speech Recognition Services for HCI Applications in Behavioral Health. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers (UbiComp/ISWC '20 Adjunct), September 12–16, 2020, Virtual Event, Mexico.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3410530.3414372

UbiComp/ISWC '20 Adjunct, September 12–16, 2020, Virtual Event, Mexico

ACM ISBN 978-1-4503-8076-8/20/09…$15.00
https://doi.org/10.1145/3410530.3414372

1 INTRODUCTION

Depression and anxiety are globally prevalent conditions that significantly impact society [1]. For example, depression can affect an individual's relationships with family and friends, overall health outcomes, mortality and productivity [2]. There is a critical need for scalable screening and monitoring for these, and other behavioral health conditions [3][4][5]. Recent studies suggest that machine-based analysis of spoken language offers promise as an aid to human providers [6][7]. Spoken language is natural, engaging, and can be recorded from any personal device with a microphone. Speech technology solutions have been studied for depression in particular, with good results for both language models [8][9] and acoustic models [6][7][10][11][12].

Ellipsis Health is developing speech-based solutions for remote, ubiquitous behavioral health screening and monitoring. Our current focus is providing decision support to providers within the clinical workflow, for conditions such as depression and anxiety. Our algorithms require a few minutes of natural speech recorded from a patient's personal device, to estimate the patients' risk level [13]. In prior work we have found that models based on natural language processing (NLP) yield strong results for both depression and anxiety prediction. NLP requires speech recognition, and there are now many options for privacy-compliant ASR.

There is little information available, however, on ASR performance benchmarks for behavioral health tasks; studies have primarily used a single ASR service. In this paper we compare ASR performance for four popular commercially-available systems. Given space constraints we focus on ASR only. In general better ASR correlates with better-performing NLP models. Our first goal is to discern whether commercial systems have similar performance for our domain. A second goal is to find out whether a speaker's health state affects ASR; we test this for depression. Finally, we describe reasons for overall ASR performance on our data, with implications for future ubiquitous behavioral health applications in this domain.

2 METHOD

We selected and manually transcribed a subset from a large corpus of American English spontaneous speech, collected by Ellipsis Health. Users interacted with an application, and spoke freely in response to specific topics prompts. Sample topics included home

Test set	*Length*	*Unique speakers*	*Utterances*	*Words*	*Unique words*
Mobile	2.5h	171	171	19k	2.4k
Laptop	59h	676	3446	474k	11k

Table 1: Reference test sets

life, relationships, and self-care, among others. The topics are designed to elicit responses containing semantic or other word-based cues to a user's behavioral health state. Word patterns are modeled by deep learning models [13].

For this study we selected 847 unique speakers and used recordings containing on average just over one minute of speech. We created two different reference test sets based on the device used. The *mobile phone & tablet* set contains recordings from Android, iPhone, and iPad. The *laptop & desktop* set includes recordings from computers running Windows, Mac OS, and Linux. For simplicity, we refer to these sets as *mobile* and *laptop*. In all cases, users could record on either their device microphone or an external microphone; this was left to the user's discretion. A summary of our ASR test set is provided in Table 1.

We selected four ASR systems out of the large and growing number of American English speech recognition providers available [20][21][22]. We included two larger and two smaller vendors. For legal reasons, vendor names are not disclosed. The point of this study is to discern whether there are large differences by provider. The actual performance for any particular task will depend on services available at the time, and their empirically determined performance for the application at hand. All of our selected vendors are compliant with the Health Insurance Portability and Accountability Act (HIPAA), support large vocabulary conversational speech recognition (LVCSR), and are popular platforms with well-known commercial clients. Based on information available before the benchmark, we had no reason to assume that the systems would differ significantly from each other in ASR performance. For the purposes of this short paper, we used the ASR services as is — without language model (LM) or acoustic model (AM) adaptation, custom LM/AM, or custom dictionaries.

To compare ASR performance across vendors, we used the standard word error rate metric ($WER = \frac{S+D+I}{N}$) [14] (the lower the better) with its components: numbers of substitutions (S), deletions (D), insertions (I), and words in the reference (N). To calculate the WER required for an apples-to-apples ASR comparison, we used the following approach.

We first employed *text pre-processing*, which allows us to add custom corrections based on regular expressions to prepare the text for more general normalization (GLM) [16]. In this phase, we remove most punctuation, normalize whitespace and standardize numbers, e.g.: 2,000→2000, 31°→31 degrees, 20%→20 percent, $5→5 dollars, 36.7→36 point 7. Numbers are then replaced with spoken words. Second, we used *GLM-like text normalization*, which includes general rules specific to US English, e.g.: British to US English (favourite→favorite), web addresses (.com→dot com), separate acronym components (ADHD→A. D. H. D.), numbers with letters (1st→first, 20s→twenties), and other writing & reading variants (I'll→I shall|I will, 1/2→one half|one over two|a half |half, cannot→can not). We then used *NIST SCTK Sclite [15]*, an open source tool for WER scoring and generation of insertion, deletion, and substitution statistics. Our process is iterative; steps 1–5 are performed several times. After each iteration, statistics and logs are examined, normalization rules are updated, and new WER values are calculated. The measured WER value decreases at each iteration. Once WER stabilizes, and there is no opportunity for improvement through new rules, the process is terminated.

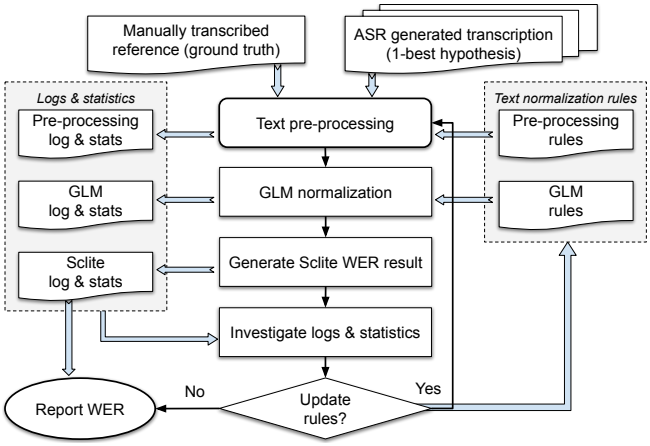

Figure 1: ASR benchmark process

3 RESULTS AND DISCUSSION

Table 2 summarizes WER results for the four ASR systems. In order to compare across systems fairly, we use WER after the normalization just described. To illustrate the importance of using normalization before the comparison, Table 2 also provides raw results for non-normalized text (WER_{Raw}). As can be seen, normalization improved WER results for mobile and laptop tests for all providers. This improvement was not equal, however, with the improvements ranging from 0.5 absolute (1.9% relative) for one vendor, to 3.2 absolute (15.7% relative) for another. By removing the factor of normalization and comparing WER results, our comparisons focus on true system performance differences.

Table 2 shows that overall, WER rates are high for our data—above 10%. Based on reports one would expect WER is below 5% for clean recordings, and near 10% for noisy recordings [19]. We note that the absolute performance reported here does not reflect data pruning, user drop-out, or any type of model adaptation. Our WER results

	Vendor	WER_{Raw}	*WER [%]*	*S [%]*	*D [%]*	*I [%]*
Mobile phone & tablet	A	20.5	20.0	7.1	7.4	**5.5**
	B	15.2	**13.5**	**4.6**	**2.1**	6.8
	C	25.2	23.2	9.4	5.4	8.3
	D	21.9	21.0	9.1	5.0	6.8
Laptop & desktop	A	30.9	30.3	8.0	17.5	**4.8**
	B	20.4	**17.2**	**6.7**	**4.0**	6.5
	C	30.9	28.7	11.5	9.7	7.6
	D	33.1	32.1	12.3	14.7	5.1

Table 2: Word error rate (WER) results

provide realistic estimates of ASR performance that could be expected with this type of application, out of the box. Our recordings cover general domains, such as life and personal concerns. Speakers use everyday speech, without specialized vocabulary. We found low rates of out of vocabulary words (OOVs).

As shown in Table 2, System B has significantly lower WER than the other three systems. The relative WER rate differences between B and the second best system (System A) are striking, corresponding to 48% for mobile/tablet data and 67% for laptop/desktop data, respectively. The different systems also vary in their relative error types, which of course trade off for any particular system. Adjusting error trade offs and retraining of ASR systems is beyond scope for this paper but we note that whereas System B minimizes deletions, System A minimizes insertions. One possibility is that the systems were optimized differently with respect to insertions and deletions. Another, not mutually exclusive possibility is a focus on capturing low-signal speech in B, for example speech that has low amplitude, or speech in noise. Both overall WER and error type distributions are also likely affected by the match of the system language model training data, to the speaking style in our behavioral health data.

The large difference between System B and System A is also seen at the speaker level. As shown in Figure 2, for almost all speakers (each represented by a single point), WER was lower on their speech when using System B. System B also shows less variability in estimates than does System A. Interestingly, when we separate points based on whether or not the speaker is classified as depressed (using binary classification labels from our data), we see no systematic WER differences in the plot nor in mean WER by class. Results for the other two systems (not shown) revealed a similar pattern. Overall, both Figure 2 and Table 2 show a large range of WER results for speakers, with many far above 10%.

Speaking style and audio quality variability are possible reasons for our relatively high WER rates. Furthermore, the latter is also dependent on the former. We have found that in order to detect a user's mental state in their speech, the user needs to share their thoughts and feelings as freely as possible when they speak. This type of sharing is facilitated by speaking in a private, safe and comfortable environment (such as their home), as well as by using a personal device of their choice. As noted earlier, we also left microphone choice (internal or external, and what type) up to the user, so that they would feel comfortable when speaking.

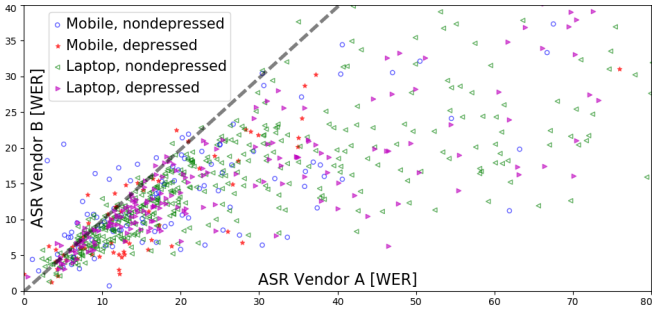

Figure 2: Per-speaker WER for Systems A and B, by depression class and device type. Axes have been truncated

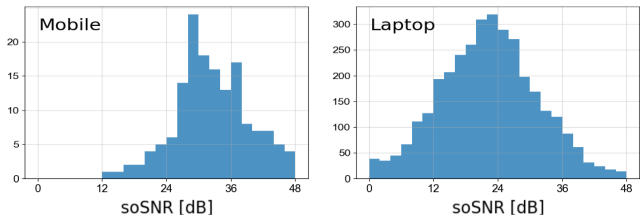

Figure 3: soSNR histograms for reference test sets

To examine audio quality for Systems A and B, we estimated signal to noise ratio (SNR) [17]. Distributions are shown in Figure 3. Due to the nature of our real-world data collection, we incur variability in device hardware and software, including differences in various sound pre-processing algorithms. Given these sources of variability, we use a signal-to-noise (SNR) estimate that applies only to speech regions, or speech-only SNR (soSNR). For our noise level estimation, we took the 10th percentile frame level of the speech frames. We applied A-weighting [18] to reduce the sensitivity of the measurement to frequencies for which the human ear is not sensitive. This estimate is usually much higher than the actual noise level measured in non-speech regions. For example, the typical raw recording in quiet conditions with classic SNR 30dB is equivalent to soSNR 20dB when noise is estimated in the speech areas. Overall, soSRN is a good measure in audio recordings with speech.

As shown in Figure 3, *Mobile phone & tablet* results are better than those of *laptop & desktop*. These results are in line with our SNR estimates, in which the average soSNR value was 34.7 dB for mobile and 22.5 dB for laptop. They can also be seen in Figure 2. We expected better performance on mobile not only due to SNR but also because speech is generally used to a greater degree on mobile than on laptop devices. Another possible explanation for the large differences could be that Vendor B did a better job making the ASR engine resistant to various acoustic phenomena. Such phenomena could include the impact of different sound processing algorithms (audio preprocessing), effects of microphone movement during recording, background noises, or noise-related events.

4 SUMMARY AND FUTURE WORK

Using first-person conversational data from patients with and without behavioral health conditions, we found that one ASR vendor stood out from the other three that we benchmarked. This large WER difference was observed for both mobile and laptop data, was seen for most speakers, and was not solely due to a reduction in deletions. Interestingly, WER was not systematically correlated with depression class itself, for either mobile or laptop data. Overall, our WER is higher than 10%, which may reflect audio quality, speaking style, or both factors.

In future work, we seek to further understand these factors. We also plan to repeat the benchmarking process over adapted (retrained) models for the subset of services that provide this option. We conclude that commercial ASR results regarding speech for this type of behavioral health application can vary widely. It is therefore critical to benchmark available services with data from the application domain.

REFERENCES

[1] R. C. Kessler, M. Petukhova, N. A. Sampson, A. M. Zaslavsky, and H. Wittchen, 2012. *Twelve-month and lifetime prevalence and lifetime morbid risk of anxiety and mood disorders in the United States*. International Journal of Methods in Psychiatric. Research 21, 3, 169–184.

[2] McLaughlin, K.A., 2011. *The Public Health Impact of Major Depression: A Call for Interdisciplinary Prevention Efforts*. Prevention Science volume 12, 361–371.

[3] S. El-Den, T.F. Chen, Y.L. Gan, E. Wong and C.L. O'Reilly, 2018. *The psychometric properties of depression screening tools in primary healthcare settings: A systematic review*. Journal of Affect Disorders; 225:50, 3–22.

[4] A. Halfin, 2007. *Depression: The benefits of early and appropriate treatment*. American J. of Managed Care, S92–S97.

[5] D. Ben-Zeev, E. A. Scherer, R. Wang, H. Xie and A. and T. Campbell, 2015. *Next-Generation Psychiatric Assessment: Using Smartphone Sensors to Monitor Behavior and Mental Health*. Psychiatr. Rehabil. J., vol. 38, no. 3, pp. 218–226.

[6] M. Ghai, S. Lal, S. Duggal and S. Manik, 2017. *Emotion recognition on speech signals using machine learning*. Proceedings of International Conference on Big Data Analytics and Computational Intelligence (ICBDAC), pp. 34–39.

[7] A. S. Cohen and B. Elvevag, 2014. *Automated computerized analysis of speech in psychiatric disorders*. Current Opinion in Psychiatry, 27(3):203–209.

[8] A.H. Orabi, P. Buddhitha, M.H. Orabi and D. Inkpen, 2018. *Deep Learning for Depression Detection of Twitter Users*. In proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic (pp. 88–97).

[9] A. Yates, A. Cohan and N. Goharian, 2017. *Depression and self-harm risk assessment in online forums*. https://aclweb.org/anthology/D17-1322

[10] R. P. Lippmann, 1997. *Speech recognition by machines and humans*. Speech Communication, vol. 22, no. 1, pp. 1–16.

[11] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, 2016. *Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network*. In Proc. of ICASSP, Shanghai, China, March, IEEE pages 5200–5204.

[12] A. Nilsonne, J. Sundberg, S. Ternstrom and A. Askenfelt, 1988. *Measuring the rate of change of voice fundamental frequency in fluent speech during mental depression*. The Journal of the Acoustical Society of America 1988; 83: 716–728.

[13] T. Rutowski, A. Harati, Y. Lu and E. Shriberg, 2019. *Optimizing Speech-Input Length for Speaker-Independent Depression Classification*. Proc. Interspeech 2019.

[14] Wikipedia. 2020. *Word error rate*. Last modified February 7, 2020. https://en.wikipedia.org/wiki/Word_error_rate

[15] GitHub. 2018. *NIST Scoring Toolkit*. Last modified November 12, 2018. https://github.com/usnistgov/SCTK

[16] GitHub. 2008. *GLM Rules*. Last modified April 29, 2008. https://github.com/usnistgov/SCTK/blob/master/doc/GLMRules.txt

[17] H. G. Hirsch and C. Ehricher. 1995. *Noise estimation techniques for robust speech recognition* in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., pp. 153–156.

[18] Wikipedia. 2020. *A-weighting*. Last modified June 30, 2020. https://en.wikipedia.org/wiki/A-weighting

[19] GitHub. 2020. *WER are we?* Last modified April 3, 2020. https://github.com/syhw/wer_are_we

[20] Capterra.ae. 2020. *Speech Recognition Software*. https://capterra.ae/directory/30098/speech-recognition/software

[21] J. Simpson. 2019. *5 Best Speech-to-Text APIs*. https://nordicapis.com/5-best-speech-to-text-apis/

[22] Index.co. 2020. *Companies - Speech Recognition - Index*. https://index.co/market/speech-recognition/companies