Online Journal of Applied Knowledge Management

A Publication of the International Institute for Applied Knowledge Management

Volume 9, Issue 1, 2021

A cross-language study of speech recognition systems for English, German, and Hebrew

Vered Silber Varod, Open Media and Information Lab, The Open University of Israel, Israel, vereds@openu.ac.il

Ingo Siegert, Mobile Dialog Systems, Otto von Guericke University Magdeburg, Germany, ingo.siegert@ovgu.de

Oliver Jokisch, Institute of Communications Engineering, Leipzig University of Telecommunications, Germany, jokisch@hft-leipzig.de

Yamini Sinha, Mobile Dialog Systems, Otto von Guericke University Magdeburg, Germany, yamini.sinha@st.ovgu.de

Nitza Geri, Department of Management and Economics, The Open University of Israel, Israel, nitzage@openu.ac.il

Abstract

Despite the growing importance of Automatic Speech Recognition (ASR), its application is still challenging, limited, language-dependent, and requires considerable resources. The resources required for ASR are not only technical, they also need to reflect technological trends and cultural diversity. The purpose of this research is to explore ASR performance gaps by a comparative study of American English, German, and Hebrew. Apart from different languages, we also investigate different speaking styles – utterances from spontaneous dialogues and utterances from frontal lectures (TED-like genre). The analysis includes a comparison of the performance of four ASR engines (Google Cloud, Google Search, IBM Watson, and WIT.ai) using four commonly used metrics: Word Error Rate (WER); Character Error Rate (CER); Word Information Lost (WIL); and Match Error Rate (MER). As expected, findings suggest that English ASR systems provide the best results. Contrary to our hypothesis regarding ASR's low performance for under-resourced languages, we found that the Hebrew and German ASR systems have similar performance. Overall, our findings suggest that ASR performance is language-dependent and system-dependent. Furthermore, ASR may be genre-sensitive, as our results showed for German. This research contributes a valuable insight for improving ubiquitous global consumption and management of knowledge and calls for corporate social responsibility of commercial companies, to develop ASR under Fair, Reasonable, and Non-Discriminatory (FRAND) terms.

Keywords: Automatic Speech Recognition (ASR), performance measures, speech-recognition evaluation metrics, ASR engine, cross-language, genre, error rate.

Introduction

The need for Automatic Speech Recognition (ASR) is becoming essential as voice assistants, multimedia mobile messaging services, and particularly video and audio clips are ubiquitous. The use of video clips as a preferred communication medium goes far beyond social media platforms such as YouTube. Facebook and Instagram are focusing more on video, especially with Stories and Instagram TV (IGTV) rapidly growing in popularity. By 2020, online videos were predicted to make up more than 80% of all consumer internet traffic (Cisco, 2020). Already in 2018 approximately 10% of the internet population used voice control (Roberts, 2018). In 2018, two billion minutes of voice and video messages were sent per day (Al-Heeti, 2018). Furthermore, speech recognition has a huge potential for improving business processes and ASR can be harnessed to mine massive data of sales calls, and generate valuable business insights (Gandomi & Haider, 2015). Despite the growing importance of ASR, its application is still challenging, limited to specific domains, language-dependent, and requires considerable resources. Depriving ASR technology from populations that speak low-resourced languages creates a knowledge dam (Silber-Varod et al., 2016), of which scientists are trying to cope with (Chaudhary et al., 2019). In this sense, adaptation of ASR technology raises global and cross-cultural aspects of knowledge management. The resources required for ASR are not only technical and reflect technological trends and cultural diversity. On the technological side, development of multilingual ASR systems enables to share existing speech and text corpora among languages (Abate et al., 2020; Besacier et al., 2014). Nevertheless, even ASR of high-resourced languages is still not satisfactory. In this respect, Tihelka et al. (2020) discussed the grappling with web technologies and remote speech recording. A previous study of Hebrew ASR performance (Silber-Varod & Geri, 2014a) provided an initial proof of concept that ASR engines may deliver satisficing transcription that would enable effective search of video and audio content. Hebrew ASR demonstrated nearly 80% recognition of keywords, therefore, suggesting that under-resourced languages should focus more on keyword recognition, since those ASR systems have not reached yet a satisfactory accuracy level of full transcription. But not only users consume ASR technology, social sciences researchers need this channel transformation for their discourse analysis research, e.g., for the acceleration of manual annotation (Egorow et al., 2017).

The purpose of this research is to explore ASR performance gaps by a comparative study of American English (henceforth, English), which is a high-resourced language, versus German and Hebrew as a mid-resourced and low-resourced language, respectively. Performance evaluation is a major issue in scientific research and technology development (NIST, 2009; Siegert et al., 2020).

Research Questions and Hypotheses

Our research questions concern language-related and supposedly language-neutral factors, which may affect ASR system performance.

RQ1: Is the performance of current ASR systems language-dependent?

The underlying assumption is that current ASR systems are not equally developed across languages. Therefore, current ASRs perform better for some languages compared to others. English, as a global language, is assumed to have the most effective ASR systems, with up to 5% word error rate (WER) (Kurata et al., 2017). ASR of Hebrew, as an under-resourced language, was reported with a 70% WER using Google ASR in 2014 (Silber-Varod & Geri, 2014b), and two years later with a 36.5% WER (Silber-Varod et al., 2016). Though in the following years, Hebrew ASR performance may have improved, we expect that the WER of Hebrew will still be higher than typical WERs of English. For German, Siegert et al. (2020) reported 7% WER. Hence, we hypothesize:

H1: A comparison of English, Hebrew, and German ASR would yield the highest performance for English, the lowest performance for Hebrew, and an in-between performance for German.

RQ2: Do different evaluation metrics yield the same ordinal results (i.e., ranking) of ASR system performance, or are the results language-dependent?

Performance measurement is a well-known general challenge (Brynjolfsson, 1993; Geri & Ronen, 2005). In the context of ASR, Gravier et al. (2004) unfolded the challenges of performance evaluation in the fields of speech and natural language processing: To be objective, evaluation requires considerable resources, which may not be available. Moreover, they argue that: "Comparing performance can only be carried out on a well-defined task, i.e., using standard databases and evaluation metrics" (Gravier et al., 2004, p. 885). Although WER is the most common metric for comparing ASR performance, it has several drawbacks, which are further discussed in the performance measures section below, along with four other metrics used in this research. If using several metrics yields contradicting results, then it is important to determine which metrics are more appropriate. However, if the ordinal results of ASRs performance are similar across metrics (e.g., if an ASR English was rated higher than German by a certain metric, the same order will prevail when using other metrics), then it may suggest that these metrics are language-neutral. Since ASR output is the transcribed text, metrics are affected by the writing system. Therefore, when comparing languages that belong to the same writing system, we would expect to find similar order among metrics. If the compared languages belong to different writing systems, there may be ordinal differences among metrics. German and English belong to the same (Latin) writing system. Hence, we hypothesize:

H2a: There are no ordinal differences of German and English between ASR evaluation metrics.

Hebrew belongs to a different writing system (abjad) than German and English (Latin). Therefore:

H2b: There are ordinal differences of Hebrew and German between ASR evaluation metrics.

H2c: There are ordinal differences of Hebrew and English between ASR evaluation metrics.

Experimental Method

In the current study, we compare the performance of four ASR engines on three languages using four commonly used metrics.

ASR engines

1. *Google Cloud (GC) Speech to Text Application Programming Interface (API)* can process audio data in real-time as well as prerecorded audio. It supports approximately 140 languages including variants. Additionally, GC supports four languages (British English, American English, Russian, and American Spanish) under special conditions (phone call**,** enhanced phone call, and enhanced video**)**. Moreover, GC is able to handle noisy audios without an additional noise cancellation beforehand (Google Cloud).

2. *Google Speech (GS) Recognition API* is specifically designed for testing personal purposes and may be revoked by Google in the future. GS does not have access to the latest Google Cloud ASR engine features but supports the same 140 languages and their variants as GC (Natal et al., 2020).

3. *Wit.ai Speech to Text API* is a formerly open source based chatbot framework with advanced natural language processing features. It is now owned by Facebook. Although mainly intended to develop intelligent chatbots, mobile apps, and IoT devices, WIT also offers the possibility to build conversational applications that respond to voice commands (Liao, 2019).

4. *IBM Watson Speech to Text API* (IBM, 2020) (henceforth, IBM) is a cloud-based service based on deep learning. As GC and GS, IBM offers a real-time transcription as well as the usage with pre-recorded files. But in comparison to GC and GS, the number of supported languages is much smaller (19, including variants). However, IBM's number of supported audio formats is much larger with Ogg, WebM, MP3, or FLAC, to name just a few.

Datasets

Apart from different languages, we also wanted to investigate different speaking styles. We used available datasets from earlier experiments: One dataset comprising speech samples from Map Task settings (Anderson et al., 1991) – spontaneous interactions in semi-structures dialogues. The three Map Task projects are: MATACOP for Hebrew (Azougi et al., 2016; Weise et al., 2020); Montclair Map Task Corpus (Pardo, 2018); German: The L1 sub corpus of the BeMaTaC 3.0 release (Sauer, & Lüdeling, 2016). The other dataset is comprised of lectures in front of an audience (either TEDxTalks in English and German or academic talks in Hebrew). We did not examine ASR under technical challenges, e.g., noisy environments.

Altogether, we extracted utterances from 108 speakers. The number of utterances per speaker ranges from 2 to 100 (total 1,871). The gender distribution varies from 42% female and 58% male in Hebrew, to 45% / 55% in German, and 49% / 51% in English.

Experimental procedure

We first extracted about 200 unique utterances from each language (~100 per genre, cf. Table 1). The utterance extraction was carried out using an automatic random-selection process for English and German, and manually without guidelines for Hebrew. The resulting transcriptions were post-processed by spelling out numbers, abbreviations, apostrophes, etc. to match the reference writing method. For example, a reference utterance "In the year two thousand nine" was automatically transcribed as "2009"; "dreißig grad" was transcribed as "30°". In total, the English utterance length averages 7.629 words or 36.381 characters (range 1-17 words / 5-76 characters); German mean utterance length = 12.455 words or 67.798 characters (4-36 words / 14-176 characters); and the Hebrew utterance length averages 6.202 words or 29.890 characters (variation of 1-16 words and 5-86 characters). Table 1 summarizes the descriptive statistics of the six datasets – two genres per language. We also added lists of the ten most frequent words in each dataset to demonstrate the inherent different orthography systems between English and German compared to Hebrew, which might affect the ASR results: While in English and German the most frequent words are regulatory function words (e.g., the, to, and), in Hebrew we see content words, such as *hamaslul* (the-track). This is evident from the *Tokens per types ratios* that are shown in the table, as well. The ratios in Hebrew are much higher (over 20% in both genres) compared to the ratios in English

(12.29% in TEDxTalks and 6.82% in MapTask) and German (13.41% in TEDxTalks and 6.15% in MapTask), hence the linguistic diversity in Hebrew is higher compared to English and German.

Table 1. Language and genre features of the three databases, including ten most frequent words of each dataset.

Language	**English**		**German**		**Hebrew**	
Genre	**TEDxTalks**	**MapTask**	**TEDxTalks**	**MapTask**	**Lectures**	**MapTask**
Unique utterances	100	99	100	100	95	68
Word tokens	3,164	2,550	2,998	6,516	772	1,212
Word types (unique)	389	174	402	401	258	323
Tokens per types ratio	12.29%	6.82%	13.41%	6.15%	33.42%	26.65%
Words per utterance (range)	1-17	4-17	4-16	5-36	1-14	2-16
Words per utterance (mean)	8.304	6.929	7.571	17.290	5.956	6.379
Characters per utterance (range)	5-76	12-76	14-91	22-176	8-86	5-67
Characters per utterance (mean)	42.192	30.364	47.553	87.840	28.897	30.600
Ten most frequent words						
1	the	the	und	du	*et* (accusative article)	*et* (accusative article)
2	to	that	ich	nach	*ze* (this)	*az* (so)
3	and	is	wir	und	*shel* (of)	*yesh* (there-is)
4	I	and	ist	dann	*ha-* (the-)	*okey* (ok)
5	of	like	die	so	*kol* (every)	*ani* (I)
6	in	its	das	die	*batsura* (in-the-shape)	*hamaslul* (the-track)
7	you	line	in	also	*yesh* (there-is)	*lemata* (down)
8	that	to	es	der	*kmo* (like)	*santimeter* (santimeter)
9	a	of	haben	rechts	*lanu* (to-us)	*shel* (of)
10	right	it	was	gehst	*al* (on)	*anaxnu* (we)

Performance measures

We used four performance measures: Word Error Rate (WER), Character Error Rate (CER), Match Error Rate (MER), and Word Information Lost (WIL). To calculate these measures, we used two python modules (jpuigcerver, 2014; Vaessen, 2020).

The WER is the minimum edit distance between changes to the hypothesis text, expressed as the number of substitutions (S), deletions (D), and insertions (I) of words, and the number of words in the reference text (N), i.e., the Levenshtein distance (Levenshtein, 1966) for words:

$$WER = \frac{S+D+I}{N} \qquad (1)$$

WER is commonly used when comparing different ASR systems, but it does not provide details of the source of errors (Morris et al., 2004). Also note, as N is the number of words in the reference, and the hypothesis can fail in all words with additional insertion, the WER can be larger than value 1. The zero value (minimum of WER) means an exact match of reference and hypothesis.

The Character Error Rate (CER) is a similar measure as WER using characters instead of words:

$$CER = \frac{S_C+D_C+I_C}{N_C} \qquad (2)$$

There is no direct relation between CER and WER, as even for a very good, i.e., low CER one can observe a high WER, e.g., if words are correct despite the suffix (Hernandez et al., 2018). The simple comparison is not ‘the whole story’ when weighting the number of characters in each language. The probability to recognize the correct character by chance are higher in a language with a smaller number of characters per word. Therefore, we normalized the character-based metrics, according to the written systems: Hebrew has 22 letter, English 26, and German 30 letters, which leads to the normalization formula:

*Normalized CER: (CER results / number of letters in language x) * number of letters in English.*

The Match Error Rate (MER) of Morris (2002) is an absolute measure of the ASR performance, and indicates the probability of a given hypothesis being incorrect, by taking into account the correct matched words (C):

$$MER = \frac{S+D+I}{C+S+D+I} \qquad (3)$$

By the incorporation of substitutions, deletions and insertions also in the denominator, MER is limited to the range of [0,1].

The Word Information Lost (WIL) measures the proportion of word information preserved in recognition and is the approximation of the Relative Information Lost measure proposed by Papoulis (1991). It is based on the Mutual Information (MI), which provides a measure of the statistical dependence between the input words X and output words Y in the unordered set of I/O word pairs obtained by I/O alignment. For more details see Morris et al. (2004):

$$WIL = 1\frac{I(X,Y)}{H(Y)} \qquad (4)$$

To calculate the different measures CER and WER we used the command-line tool *xer* (jpuigcerver, 2014). For MER and WIL, we applied the python library jiwer 2.2.0 (Vaessen, 2020). In total, we have 9,355 evaluation results for 1,871 ASR processed samples.

Results

ASR systems

Regarding the question of which engine gives the best results, we found that IBM is generally the engine with the best results for English. GC shows the best performance when measuring CER and little behind for the other metrics. For German, GC, GS, and WIT give the best results, depending on the metrics, while IBM has the highest error rates. For Hebrew, IBM and WIT are not available; GC and GS give the same results with a small advantage towards GC. For German, the one-way Analysis of Variance (ANOVA) showed significant difference among the engines in all the four metrics and a post hoc test showed that IBM is worse significantly (p < 0.01) compared to the three other engines, while the other engines are similar for all metrics. To conclude, when taking the engine with the best performances for each language (IBM for English, GS for German and GC for Hebrew), we found that as expected, English results are the best. Surprisingly and contrary to our assumption, the German average results are the lowest. Figure 1 presents the results of the four metrics.

The ANOVA results of the CER metric were found insignificant ($p < 0.05$). A post hoc analysis showed no significant difference between language pairs. The ANOVA results of the WER metric were found significantly different ($p < 0.01$). A post hoc analysis showed a significant difference between English and Hebrew ($p < 0.01$) and between English and German ($p < 0.01$), but not between German and Hebrew.

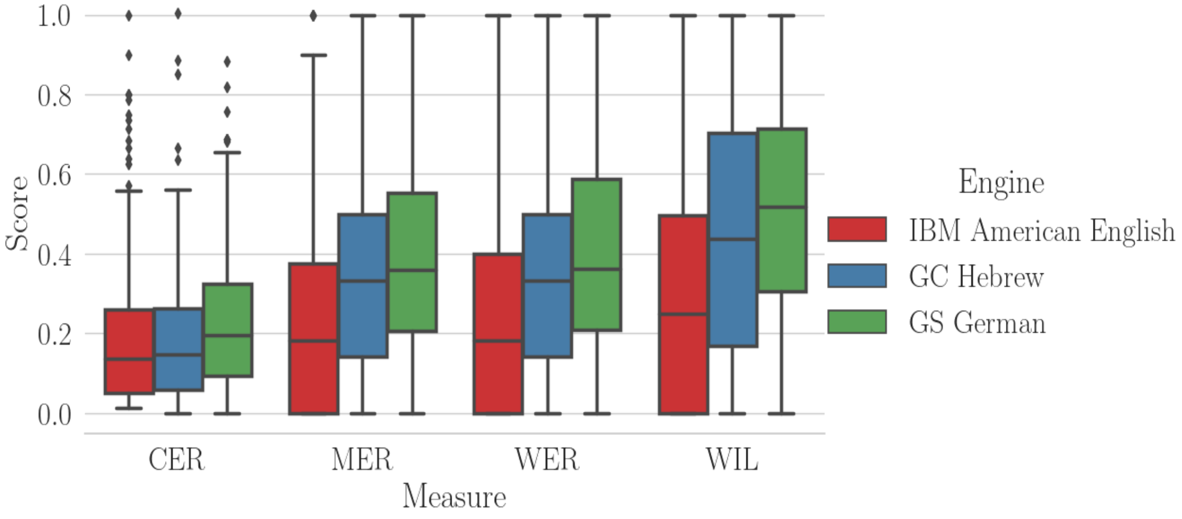


Figure 1. Error rates of the four metrics according to the engine with best performance per language.

Comparison of metrics

We calculated the correlation coefficient of the results of the four metrics, in each language. Almost all correlations are high and positive ($r > 0.5$; in German and Hebrew). This is an indication that the evaluations have the same trend.

Regarding the question which metric gives the best results, when comparing the results of one system, the GC engine, CER, again, shows best results for all languages (Figure 2). MER gives better results than WER.

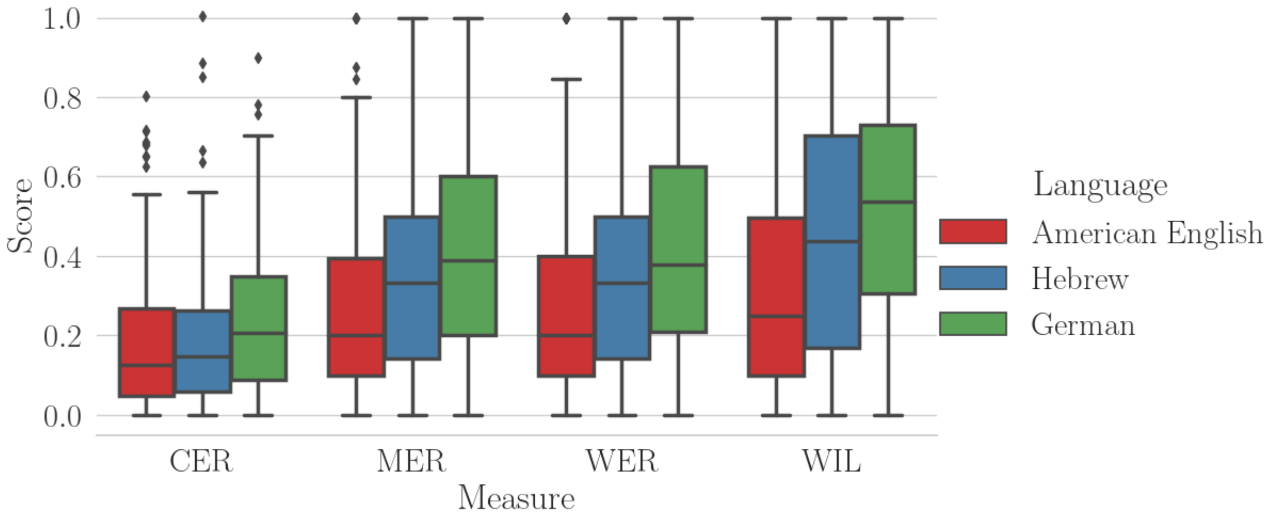


Figure 2. A comparison of the three languages according to the average error rate of the four metrics using Google Cloud (GC) API

Language differences

We ran ANOVA to compare the three languages in GC engine. The differences through all the metric results were found significant ($p \leq 0.05$). Table 2 summarizes the results of these comparisons and the Tukey HSD ("Honestly Significant Difference") post-hoc tests. These tests indicate which groups were significantly different from each other (R Foundation, 2002). It can be shown that English and German are significantly different through all the metrics and English and Hebrew are significantly different in three out of the four metrics (CER excluded); German and Hebrew are not significantly different in all four metrics.

Table 2. ANOVA between languages in GC engine and Tukey HSD Post-hoc Tests. An asterisk (*) represents statistical significance.

Metric	English	German	Hebrew	P value	Tukey HSD Post-hoc Test
Number of observations	194	200	163		
CER Mean	0.191	0.255	0.217	0.0543*	Eng. vs Ger.: p=0.044* Eng. vs Heb.: p=0.633 Ger. vs Heb.: p= 0.353
CER Standard deviation	0.205	0.234	0.359		
MER Mean	0.256	0.398	0.343	<0.0001***	Eng. vs Ger.: p<0.001*** Eng. vs Heb.: p=0.01*** Ger. vs Heb.: p=0.101
MER Standard deviation	0.236	0.252	0.277		
WER Mean	0.265	0.444	0.373	<0.0001***	Eng. vs Ger.: p<0.001*** Eng. vs Heb.: p=0.003** Ger. vs Heb.: p=0.079
WER Standard deviation	0.247	0.303	0.382		
WIL Mean	0.325	0.510	0.441	<0.0001***	Eng. vs Ger.: p<0.001*** Eng. vs Heb.: p=0.001*** Ger. vs Heb.: p<0.067
WIL Standard deviation	0.282	0.283	0.317		

Annotation	p-value	Significance level
***	(0.0001, 0.001]	0.001
**	(0.001, 0.01]	0.01
*	(0.01, 0.05]	0.05

Comparison of genres

Regarding genre, the differences in Hebrew (GC engine) and English (IBM engine) are not significant for any of the metrics. In German (GS engine), WIL is significantly different in the two genres ($p < 0.01$ and $p = 0.01$, respectively) while the other three metrics are not. Figure 3 shows the results of GS ASR system. It seems that for German, at least for the analyzed samples, there is an effect of the genre while in English and Hebrew genre does not affect the results. In general, the comparisons between the genres show a more robust engine for English and Hebrew, and less for German, which is genre-sensitive. Otherwise, lectures have a "simple" and structured language model, while the MapTask is relatively a more spoken language with a complex language model. Lectures are structured while the MapTask syntax has a complex spontaneous syntax (at least for German).

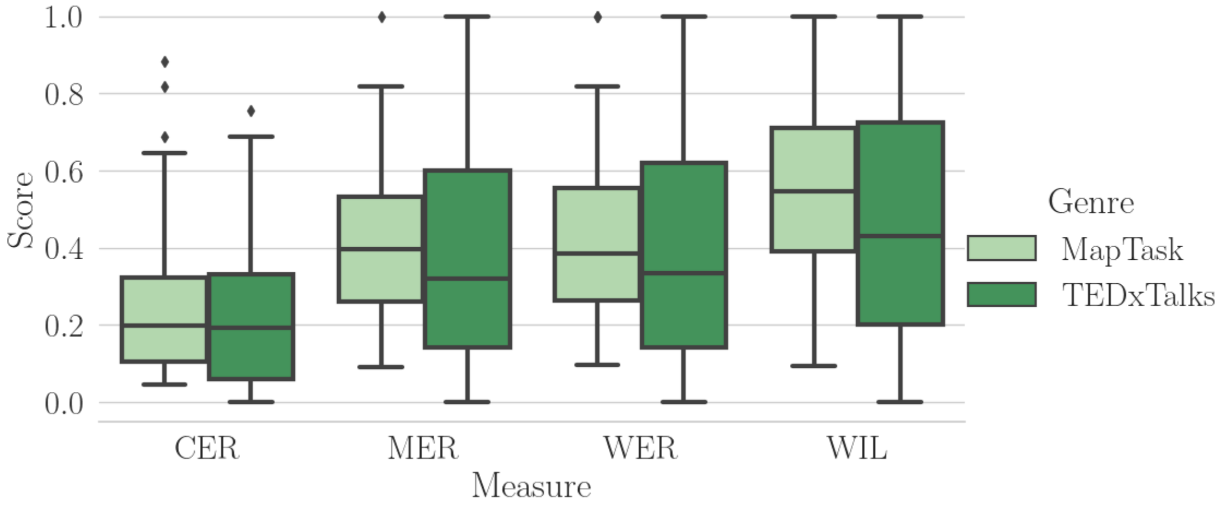


Figure 3. Genre differences in German GS across four metrics.

Discussion

Our research aims to shed light on the relative performance of state-of-the-art ASR engines on two different genres for three different languages. Regarding the language-dependence of current ASR systems performance (RQ1), as we hypothesized in H1, English ASRs provides the best results. As a language that mediates world trade – not surprisingly, commercial companies have invested in English ASR. Contrary to our hypothesis H1, we found that the Hebrew and German ASR systems have similar performance. Regarding the examined metrics, the CER metric shows the best results, suggesting that GC and IBM engines perform best for English ASR. Thus, to summarize, our findings suggest that ASR performance is language-dependent and system-dependent. Furthermore, Hebrew might not be categorized as an under-resourced language anymore. The observed overall error rates suggest that current ASR engines are not optimal for the chosen data set.

Regarding the second research question RQ2, we found no ordinal differences of German and English between ASR evaluation metrics. Thus, H2a was confirmed, but H2b and H2c were not. We found no ordinal differences of Hebrew and German between ASR evaluation metrics; nor ordinal differences of Hebrew and English between ASR evaluation metrics. In fact, an error diagnosis shows that German and Hebrew results are not different, except for the genre differences found in German, where we found the lowest performance for German's MapTask when measuring WIL metric.

Regarding the metrics, normalized CER is more sensitive to the writing systems, the orthography, where German has a disadvantage due to longer stretches of characters. WER on the other hand, depends more on the vocabulary, in which English has an advantage being a well-trained and structured language. For Hebrew, the advantage of CER metric upon WER might be explained due to its orthographic system (i.e., clitics are attached to the word stem and orthography is vowelless to a certain extent). Thus, regarding RQ2, our two hypotheses can be applied.

Last, our results show that current general-purpose ASRs still lack performance for satisfying mass transcription. Nonetheless, this research contributes a valuable insight for improving ubiquitous global consumption and management of knowledge. First, we call for a corporate social responsibility of commercial companies, which should take care of their ASR-related development under Fair, Reasonable, and Non-Discriminatory (FRAND) terms. If a company ignores a particular language, it violates human equality. Moreover, in this study, we worked under a limited "black box" type of access to the ASR systems. For research purposes, we call for companies to enable access to the parameters and configuration of the ASR systems.

The cross-cultural aspect of the ASR technology does not end in the development process; it goes further to the users' experience. A future study might aim to emulate and uncover the fundamental user practices, for example, in real work with various search engines. Such a study might shed light on the practice of breaking down queries into "searchable" keywords or discursive units as a widespread sociocultural pattern of our time. Exploring how the "googling" paradigm delimits the horizon of the material to be searched can be an interesting sociocultural exploration.

Conclusions

In this cross-language study, we showed how performance of ASR engines is still varied across languages and genres, and across ASR systems and metrics. Currently, the best engines for American-English are IBM, which gives better results for spontaneous speech (i.e., MapTask) and GC which gives better results for TedxTalks. The best engine for German is GS, which works better with TedxTalks spoken genre. For Hebrew, only two engines were available and GC performed better for both genres. Our research calls for developing ASR engines under FRAND terms in order to improve ubiquitous global consumption and management of knowledge.

References

Abate, S. T., Tachbelie, M. Y., & Schultz, T. (2020). Multilingual acoustic and language modeling for ethio-semitic languages. *Proc. Interspeech 2020*, 1047-1051. https://10.21437/Interspeech.2020-2856

Al-Heeti, A. (2018). WhatsApp: 65B messages sent each day, and more than 2B minutes of calls. *CNET*. https://www.cnet.com/news/whatsapp-65-billion-messages-sent-each-day-and-more-than-2-billion-minutes-of-calls/

Anderson, H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S., & Weinert, R. (1991). The HCRC map task corpus. *Language and Speech*, *34*(4), 351-366. https://doi.org/10.1177%2F002383099103400404

Azogui, J., Lerner, A., & Silber-Varod, V. (n.d.). *The Open University of Israel Map Task Corpus (MaTaCOp)*. http://www.openu.ac.il/en/academicstudies/matacop/

Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, *56*, 85-100. https://doi.org/10.1016/j.specom.2013.07.008

Brynjolfsson, E. (1993). The productivity paradox of information technology. *Communications of the ACM*, *36*(12), 66-77. https://doi.org/10.1145/163298.163309

Chaudhary, A., Dalmia, S., Hu, J., Li, X., Matthews, A., Muis, A. O., Otani, N., Rijhwani, S., Sheikh, Z., Vyas, N., Wang, X., Xie, J., Xu, R., Zhou, C., Jansen, P. J., Yang, Y., Levin, L., Metze, F., Mitamura, T., Mortensen, D. R., … McKeown, K. (2019). The ARIEL-CMU systems for LoReHLT18. *arXiv preprint*, arXiv:1902.08899.

Cisco (2020). Cisco annual Internet report (2018–2023) white paper. Retrieved from: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html

Egorow, O., Lotz, A., Siegert, I., Böck, R., Krüger, J., & Wendemuth, A. (2017). Accel erating manual annotation of filled pauses by automatic pre-selection. *Proceedings of the 2017 International Conference on Companion Technology,* pp. 1–6. https://doi.org/10.1109/COMPANION.2017.8287079

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137-144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Geri, N., & Ronen, B. (2005). Relevance lost: The rise and fall of activity-based costing. *Human Systems Management, 24*(2), 133-144.

Google Cloud API (n.d.). Retrieved from: https://cloud.google.com/speech-to-text/docs

Gravier, G., Bonastre, J. F., Geoffrois, E., Galliano, S., McTait, K., & Choukri, K. (2004). *The ESTER evaluation campaign for the rich transcription of French broadcast news. LREC.* http://www.lrec-conf.org/proceedings/lrec2004/pdf/672.pdf

Hernandez F., Nguyen V., Ghannay S., Tomashenko N., & Estève Y. (2018) TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In Karpov A., Jokisch O., Potapova R. (Eds) *Speech and Computer (SPECOM 2018), Lecture Notes in Computer Science*, *11096*. Springer, Cham. https://doi.org/10.1007/978-3-319-99579-3_21

IBM (2020). *Getting started with speech to text.* IBM Cloud Docs, Speech to Text. Retrieved from: https://cloud.ibm.com/docs/speech-to-text

Jpuigcerver, J. (2014). *Xer*. Retrieved from: https://github.com/jpuigcerver/xer

Ramabhadran, B., Saon, G., & Sethy, A. (2017). Language modeling with highway LSTM. *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 244-251. https://doi.org/10.1109/ASRU.2017.8268942

Levenshtein, V. I. (1965). Binary codes capable of correcting deletions, insertions, and reversals. *Dokl. Akad. Nauk SSSR*, *163*(4), 845-848. (in Russian)

Liao, J. (2019). Hello speech, we love you too. *medium.com*. Retrieved from: https://medium.com/wit-ai/hello-speech-we-love-you-too-5851d0b34f9f

Morris, A. C. (2002). An information theoretic measure of sequence recognition performance. *IDIAP-com* 02-03, 2002. ftp://ftp.idiap.ch/pub/reports/2002/com02-03.pdf

Morris, A. C., Maier, V., & Green, P. (2004). From WER and RIL to MER and WIL: Improved evaluation measures for connected speech recognition. *Proceedings of the Eighth International Conference on Spoken Language Processing.* https://www.isca-speech.org/archive/archive_papers/interspeech_2004/i04_2765.pdf

Natal, A., Shires, G., & Jägenstedt, P. (2020). Web speech API draft community group report. https://wicg.github.io/speech-api/

National Institute of Standards and Technology (NIST) (2009), The history of automatic speech recognition evaluations at NIST. Retrieved from: http://itl.nist.gov/iad/mig/publications/ASRhistory/index.html

Papoulis, A. (1991). *Probability, random variables, and stochastic processes*. McGraw-Hill.

Pardo, J. S. (2018). Montclair map task corpus. https://digitalcommons.montclair.edu/data/1/

R Foundation (2002). Analysis of variance from summary data. *The Interactive Statistical Pages*. https://statpages.info/anova1sm.html

Roberts, M. (2018). OK Google, Siri, Alexa, Cortana; Can you tell me some stats on voice search? *The Edit Blog*. [Online; 8-Jan-2018]. https://www.slideshare.net/mikejeffs/ok-google-siri-alexa-cortana-can-you-tell-me-some-stats-on-voice-search

Sauer, S., & Lüdeling, A. (2016). Flexible multi-layer spoken dialogue corpora. *Special Issue: Compilation, Transcription, Markup and Annotation of Spoken Corpora*, *International Journal of Corpus Linguistics*, *21*(3), 419-438. https://doi.org/10.1075/ijcl.21.3.06sau

Siegert, I., Sinha, Y., Jokisch, O., & Wendemuth, A. (2020) Recognition performance of selected speech recognition APIs – A longitudinal study. In: Karpov A., Potapova R. (eds) *Speech and Computer*. SPECOM 2020. Lecture Notes in Computer Science: Vol. 12335. Springer, Cham. https://doi.org/10.1007/978-3-030-60276-5_50

Silber-Varod, V., Winer, A., & Geri, N. (2016). Opening the knowledge dam: Speech recognition for video search. *Journal of Computer Information Systems, 57*(2), 106-111. https://doi.org/10.1080/08874417.2016.1183423

Silber-Varod, V., & Geri, N. (2014a). Can automatic speech recognition be satisficing for audio/video search? Keyword-focused analysis of Hebrew automatic and manual transcription. *Online Journal of Applied Knowledge Management*, *2*(1), 104-121.

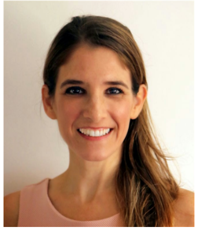
Silber-Varod, V., & Geri, N. (2014b). Error diagnosis and classification of errors in two Hebrew state-of-the-art speech recognition systems. *Proceedings of 2014 Speech Processing Conference Afeka*, Tel-Aviv, Israel July 7-8, 2014.

Tihelka, D., Jůzová, M., & Vít, J. (2020). Grappling with web technologies: The problems of remote speech recording. *Proceedings of the International Conference on Speech and Computer* (pp. 592-602). Springer, Cham. https://doi.org/10.1007/978-3-030-60276-5_57

Vaessen, N. (2020). JiWER: Similarity measures for automatic speech recognition evaluation. Retrieved from: https://pypi.org/project/jiwer/

Weise, A., Silber-Varod, V., Lerner, A., Hirschberg, J., Levitan, R. (2020). Entrainment in spoken Hebrew dialogues. In: J. Pardo, E. Pellegrino, V. Dellwo, & B. Möbius (Eds.), *Special Issue on Vocal Accommodation in Speech Communication*, *Journal of Phonetics, 83*. https://doi.org/10.1016/j.wocn.2020.101005

Authors Biographies



Vered Silber-Varod, Ph.D. Director of the Open Media and Information Lab (OMILab), The Open University of Israel. Former Research Fellow at the Research Center for Innovation in Learning Technologies, The Open University of Israel. Research interests and publications focus on various aspects of speech sciences, with expertise in speech prosody, acoustic phonetics, speech communication and text analytics. Honored to be part of ISCA'S WomenNSpeech list. Currently the treasurer-Secretary, The Haiim B. Rosen Israeli Linguistic Society.

Ingo Siegert, Dr.-Ing. is Juniorprofessor for Mobile Dialog Systems at the Otto von Guericke University Magdeburg. Research interests and publications focus on signal-based analyses and interdisciplinary investigations of human-computer interaction in terms of addressee detection and the utilization of further interaction patterns, such as filled pauses or discourse particles. He has published 90+ peer reviewed papers on several conferences and various journals and is co-organizer of several workshops and conferences.

Oliver Jokisch, Dr.-Ing. is teaching as a full professor for signal and system theory at the Leipzig University of Telecommunications (HfTL), Germany. He studied information technology at TU Dresden in Germany as well as at the Loughborough University in United Kingdom. Oliver graduated as a diploma engineer and holds a PhD degree in information technology from TU Dresden. His research is dedicated to different AI areas as well as to audio, speech and video communication. Oliver has co-founded several IT companies.

Yamini Sinha, B.Sc. is a master's student in electrical engineering and information technology at Otto von Guericke University Magdeburg. Her master's thesis is focused on evaluating and improving performance of cloud-based speech recognition systems for German conversational speech. She holds a bachelor's degree in electronics and instrumentation engineering from Sikkim Manipal Institute of Technology, India. Her research interests include speech processing and recognition, natural language processing, natural language understanding, and AI related topics.

Nitza Geri, Ph.D. Nitza Geri is an Associate Professor at the Open University of Israel, Department of Management and Economics. She is a board member of the Research Center for Innovation in Learning Technologies, and served as Head of the center (2012-2018). Nitza holds a B.A. in Accounting and Economics and a Ph.D. in Technology and Information Systems Management from Tel-Aviv University. Her research interests focus on the value of information and knowledge: strategic information systems, information economics, attention economy, knowledge management, value creation, Theory of Constraints, and effectiveness of e-learning. Personal site: http://www.openu.ac.il/en/personalsites/NitzaGeri.aspx.