Automatic Speech Recognition for Polish in 2022

Evaluating selected ASRs on a corpus of customer support dialogs

University of Łódź

Piotr Pęzik[1], Michał Adamczyk

April 2022

[1] piotr.pezik@uni.lodz.pl

Motivation and Scope

Automatic Speech Recognition (ASR) also known as Speech-To-Text (STT) transcription and more specifically Large Vocabulary Continuous Speech Recognition (LVCSR) is a basic building block of many Natural Language Processing (NLP) solutions, such as voice-operated user interfaces, speech analytics applications and dialog systems. The last few years have seen a significant increase in the demand for the latter two types of systems both in Poland and worldwide. Large and medium-size companies, including banks, insurance firms and public institutions have implemented speech analytics solutions to digitize, archive and explore recordings of spoken interactions and gain analytical insights into customer support and sales processes. The intrinsic quality of ASR systems is a key prerequisite for the efficiency of such applications. Even a seemingly small difference in the quality of ASR may be critical in certain contexts. For example, the take up rate of a voice bot may directly depend on the word error rate of its underlying ASR engine. It may be difficult to successfully deploy a voice bot with an overall acceptable ASR rate which nevertheless consistently fails to recognize phone numbers or dates. More sophisticated Natural Language Understanding (NLU) modules and the general usefulness of speech analytics results also hinge upon ASR quality.

This report looks at the following commercially offered ASR engines for Polish:

- The ASR engine for Polish available as part of Microsoft's Azure Cognitive Services[2]
- VoiceLab ASR Service[3]
- Google ASR API[4]

Descriptions available for the non-English ASR models offered as part of those services tend to be scant. Some of them contain rather general claims about accuracy interlaced with occasional appeals to the vendor's overall reputation in the field of natural language processing. Even if such claims are generally reliable for some of the supported languages,

[2] See https://docs.microsoft.com/en-us/azure/cognitive-services/speech-service/language-support. Accessed in February/ March 2022.

[3] Model name: 8000_pl_PL. https://voicelab.ai/asr-api-web-socket-grpc-http/. Accessed in March 2022.

[4] See https://cloud.google.com/speech-to-text Accessed in February/ March 2022. We used the "default" model for Polish. Additional Google ASR models which seem to have become available for Polish since March 2022 were not tested in this report.

one can expect considerable differences in the performance of ASR for various combinations of models, languages and target domains of spoken language. In particular, none of the models available for Polish comes with documented intrinsic or extrinsic evaluation results on publicly available reference corpora which would represent a strictly defined topical domains, text genres or registers. The basic purpose of the work behind this report is to evaluate the out-of-the box accuracy of the above mentioned ASR systems on a subset of DiaBiz as a corpus which approximates phone-based customer support dialogs[5].

The DiaBiz Corpus

The evaluation of the ASRs compared in this report is based on a subset of DiaBiz,“a multimodal corpus of Polish telephone conversations conducted in varied business settings, comprising 4,036 call centre interactions from nine different domains, i.e. banking, energy services, telecommunications, insurance, medical care, debt collection, tourism, retail and car rental” at the time of writing this report (Pęzik et al., 2022). The dialogs were enacted by 5 ’agents’ and 191 ’customers’. The corpus is publicly available[6], which it makes it possible to run reproducible evaluations of ASR accuracy on a dataset which closely approximates the linguistic register of phone-based customer support center interactions. The average WER of the manually corrected transcriptions of DiaBiz recordings is currently estimated at less than 3 per cent. Although there are certain ecological validity considerations of using recordings of arranged conversations rather than a data set from real call-center, DiaBiz conversations are based on authentic scripts and they sound rather natural. In fact, at the level of prosodic features of speech such as intonation, fluency and timing, they may sound more natural than recordings of re-spoken transcripts of real conversations.[7].

Phone-call interactions were recorded using the Genesys PureCloud contact center application to closely mimic the quality of recordings collected in a real business environment. Recordings were exported as standard 16-bit, 8 kHz stereo WAV files. Participants were

[5] By out-of-the box performance we mean performance without any adaptations for a particular data set or domain. Although the acoustic and language models of ASRs can be fine-tuned for a specific domain or even dataset, such adaptations are either every limited or generally impossible in cloud-based services provided by the main vendors of such solutions.

[6] It can be purchased for research and commercial purposes. For a detailed description see (Pęzik et al., 2022).

[7] Samples of the DiaBiz corpus are available for assessment at https://clarin-pl.eu/dspace/handle/11321/887.

recorded independently in separate channels. It made it possible to maintain a clear-cut separation of the agent and client speech without employing error-prone speech diarization techniques.

For the purposes of this report we used a subset of 400 conversations, with 50 dialogs sampled from eight different domains and totalling 41 hours 14 minutes and 16 seconds of stereo recordings. The conversations recorded were held by 5 agents with 146 unique speakers acting as customers in different 180 scripts covering both incoming and outgoing scenarios. Table 1 summarizes the subset of DiaBiz used in this report.

Domain	Total time	Mean time	Median time	Dialogs	Scripts	Speakers	Agents
Banking	06:11:53	00:07:26	00:05:26	50	35	45	5
Car rental	04:31:19	00:05:25	00:05:08	50	12	38	5
Debt collection	05:29:56	00:06:35	00:05:13	50	14	46	5
Energy services	04:06:45	00:04:56	00:03:35	50	22	48	5
Insurance	04:44:41	00:05:41	00:05:02	50	24	43	5
Medical care	03:59:02	00:04:46	00:04:22	50	18	42	5
Telecommunications	03:16:15	00:03:55	00:03:07	50	29	37	5
Tourism	08:54:25	00:10:41	00:09:40	50	26	40	5
Total	**41:14:16**			**400**	**180**	**146**	**5**

Table 1: Aggregate properties of the data set

Whereas the subsets of dialogs representing the respective domains are evenly-sized, the average dialog duration is considerably larger for some of the domains such as tourism or banking. This difference reflects the natural variation in conversation length, which depends on the topic of the scripts used to conduct the dialogs.

Methodology and Data Preparation

Measuring the accuracy of an ASR system against an independent reference data set is not straightforward. One of the most commonly uesd ASR evaluation metrics is the (average) Word Error Rate (WER). It is defined as the total number of word-level recognition errors (substitutions, deletions and insertions) in the test set divided by the total number of words found in the reference set:

$$WER = \frac{S + I + D}{N} \tag{1}$$

However simple this formula may look, defining what constitutes an incorrectly recognized, substituted or omitted word can be problematic due to varying conventions of transcribing numbers, dates, abbreviations etc. The mapping of an ASR's output on the reference set transcription conventions is particularly challenging when comparing different speech recognition systems. This is illustrated in the example below. The first transcript is a sample of the DiaBiz corpus used in the evaluation process. The fragment in bold refers to "payments of 150 zlotys made with one card and then with another (second) one". Fully inflected forms of the numerals are used in the reference transcription and the output of the first ASR system shown in the second bullet below. In the output of the second ASR shown in the last bullet, the same numerals are differently formatted. The word "jedną" (one) is formatted in digits and the word "drugą" (the other) is interpreted as a time reference (two o'clock or "2:00"). On the other hand, although the amount of 150 zlotys is correctly formatted, it would needs to be mapped to the reference transcription convention in order to be considered correct.

1. DiaBiz: (...) suma transakcji musi być dla na jednej karcie czyli nie może być tak że wykona pani **sto pięćdziesiąt złotych jedną i sto pięćdziesiąt złotych drugą** niestety (...)
2. ASR_1: (...) suma transakcji musi być dla na jednej karcie czyli nie może być tak że wykona pani **sto pięćdziesiąt złotych jedną i sto pięćdziesiąt złotych drugą** niestety (...)
3. ASR_2: (...) suma transakcji musi być na jednej karcie czy nie może być tak że wykona pan **1 i 150 zł 2:00** niestety (...)

The problem illustrated by this example has led us to develop a transcription normalization script[8]. The script converts all numerals in ASR transcripts to digits and applies a set of other rules in order to equalise the chances of all ASR to produce acceptable transcriptions of frequent words which tend to be variably formatted. Since we do not have access to the formatting models used by the respective ASR systems, it is impossible to predict all such inconsistencies. Additionally, it should be stated that the transcriptions available in

[8] The script is named normalize_text.py and it is attached to the full version of this report.

the DiaBiz corpus were originally made by a different ASR model provided by Voicelab in the CLARIN-BIZ project before being corrected by human annotators using an independent set of transcription guidelines. This means that the results reported here may be slightly biased towards Voicelab's ASR transcription conventions. However, having analyzed the most frequent word errors, we estimate that such inconsistencies have a limited impact on the final WER calculations and that they certainly do not affect the ranking of the ASR engines. A different type of evaluation error introduced by this approach stems from the fact that numbers are inflected in Polish and the script converts them to digits. As a result, differences between number inflections are not factored in the WER calculations.

Results

Overall WER

Table 2 shows the WER scores of the systems tested, averaged over conversation transcripts. It needs to be stressed that the Voicelab model used is according to their documentation is dedicated to 8kHz recordings. It is not clear how the other two engines deal with the acoustic parameters of the submitted media files.

In addition to the overall average, the WER scores reported here are calculated separately for the client and agent channels. As expected, agents' speech is better recognized due to more stable acoustic conditions and a generally more formulaic language model. The difference in the accuracy is important to recognize in various applications of ASRs. For example, the downstream performance of speech analytics systems which are mostly meant to search or extract information from customers' utterances should not be estimated on the basis of overall WER rates reported for the underlying ASR engine used. The best results in our benchmark were obtained for Microsoft's Azure service (10.51 WER for both channels), which is closely followed by Voicelab's ASR at 11.51 overall WER. Google's ASR service for Polish performed much worse on the DiaBiz dataset at 20.84 WER.

Table 3 shows a more detailed breakdown of WER values obtained for the ASRs tested across eight business domains represented in DiaBiz. Microsoft's Azure service outperforms the remaining ASR's in 8 out of 9 domains. Voicelab's results are slightly better for the dialogs representing telecommunications customer support lines.

Vendor	Total WER	Client WER	Agent WER
Microsoft Azure	10.51*	13.9 *	8.89*
Voicelab	11.51	14.86	9.92
Google	20.84	24.95	18.89

Table 2: Average WER per vendor for complete dialogs, client and agent channels respectively%. Lower is better. Best results are marked with an asterisk.

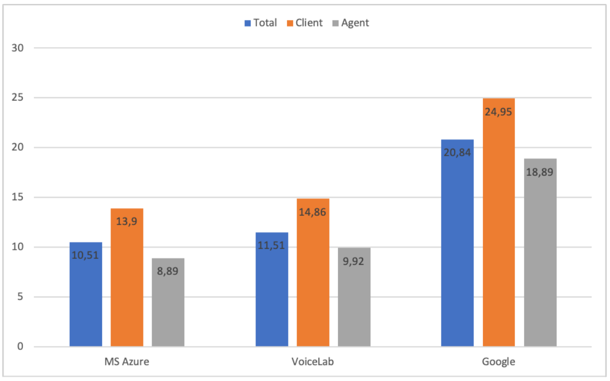

Figure 1: Overall WER results for the agent/client channels. (see Table 2).

Full Report

The full version of this report includes the ASR and reference transcriptions, the DiaBiz recordings and Google Colab scripts used to obtain the results and run more detailed analyses. For example, as shown in Tables 4 and 5 it is possible to generate frequency lists of (apparent) substitution or deletions for each of the engines tested. This in turn may reveal additional normalization rules which may be required to calculate a more accurate comparison of the systems tested. For example, some the most frequent deletion and substitutions found in Microsoft's Azure service are *okej* and *mhm*. The first of those words may be spelt as *ok* as well, which means that the results may need to be corrected in favour of this particular ASR engine. The same may apply for hesitation markers such as *mhm*, which may not have been included in this engine's dictionary.

As stated above, it is nearly impossible to account for all such discrepancies in a reasonable period of time. Nevertheless, we estimate that at least the ranking of the systems reported here is not affected by such remaining transcription normalization issues. For example, the most

Domain	WER	Vendor		
		Microsoft Azure	Voicelab	Google
Banking	Total	9.4 *	9.8	19.81
	Client	13.65*	14.02	25.22
	Agent	7.63*	8.05	17.57
Energy services	Total	10.23*	10.69	20.31
	Client	14.3 *	14.69	23.54
	Agent	8.5 *	8.98	18.95
Medical care	Total	11.97*	14.99	21.88
	Client	14.56*	16.58	25.74
	Agent	10.71*	14.22	20
Telecommunications	Total	10.97	10.92*	20.24
	Client	13.98	13.66*	23.88
	Agent	9.36*	9.47	18.31
Tourism	Total	11.22*	12.49	21.19
	Client	13.09*	14.49	22.82
	Agent	10.24*	11.45	20.34
Insurance	Total	11.34*	11.98	22.68
	Client	14.62*	15.49	27.64
	Agent	9.43*	9.95	19.81
Debt collection	Total	8.96*	9.79	19.71
	Client	14.27*	14.69	25.69
	Agent	6.66*	7.66	17.13
Car rental	Total	10.51*	11.86	21.19
	Client	13.61*	15.68	26.18
	Agent	9.17*	10.22	19.03

Table 3: Average WER per vendor and domain reported in %. Lower is better. Best results are marked with an asterisk.

We welcome ASR researchers and developers to contact us and look at our source data independently. Any substantial corrections will be published in new versions of this report.

Acknowledgements

Work on this report was supported by “CLARIN - Common Language Resources and Technology Infrastructure”, a project financed under the 2014-2020 Smart Growth Operational Programme, POIR.04.02.00-00C002/19.

word	frequency
okej	512
to	466
no	422
co	420
że	356
w	345
tak	253
pani	252
ja	227
i	221

Table 4: 10 most frequent substitution errors for Microsoft Azure

References

Pęzik P., Krawentek G., Karasińska S., Wilk P., Rybińska P., Peljak-Łapińska A., Cichosz A., Deckert M., and Adamczyk M. (2022). "DiaBiz – an Annotated Corpus of Polish Call Center Dialogs". *Language Resources and Evaluation Conference 2022.*

word	frequency
aha	1098
mhm	763
i	536
no	389
proszę	325
nie	309
wprowadzić	297
to	290
w	222
zatwierdzić	210

Table 5: 10 most frequent deletion errors for Microsoft Azure

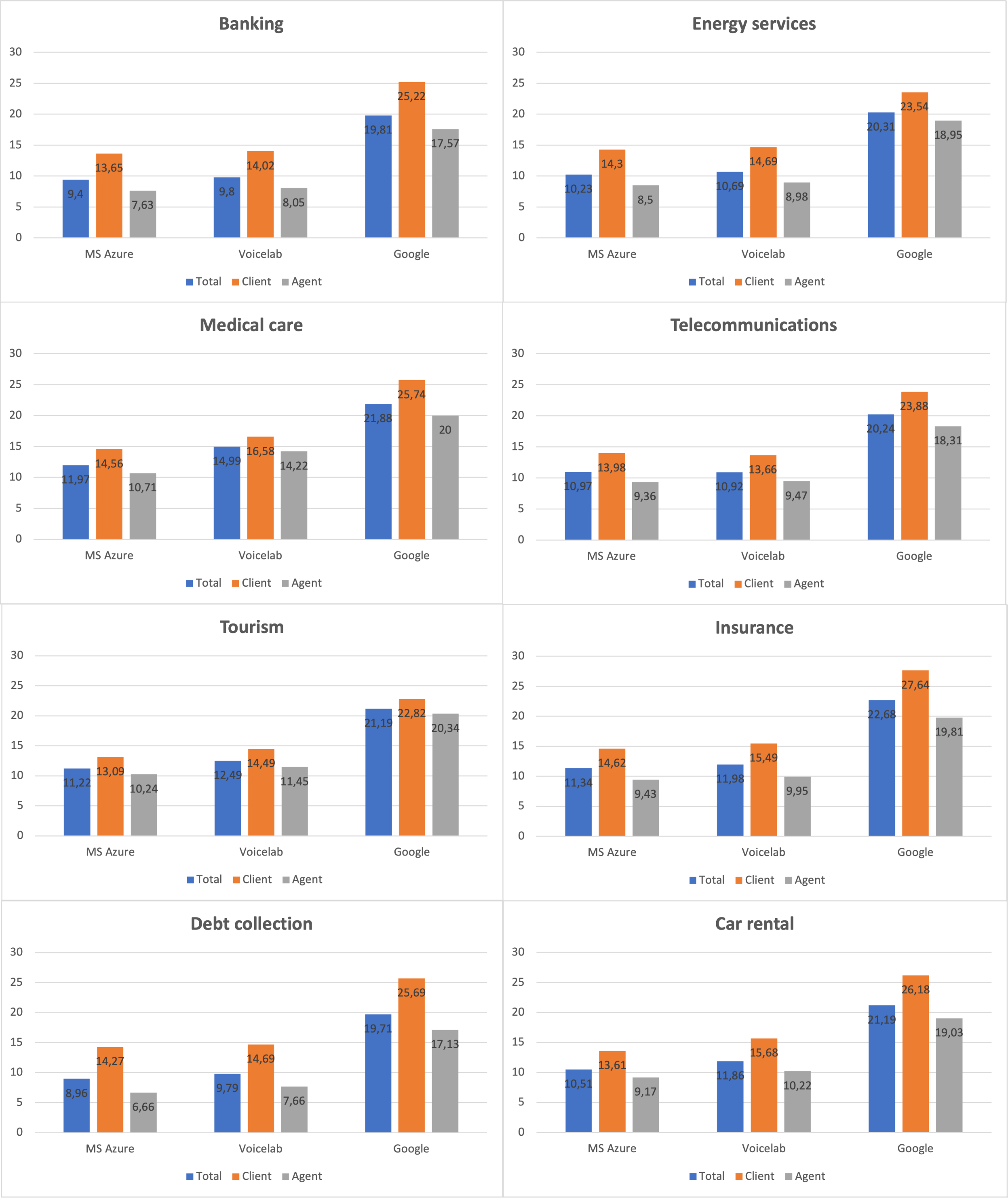

Figure 2: WER results for the agent/client channels across domains. (see Table 3).