SPEECH RECOGNITION ERRORS IN ASR ENGINES AND THEIR IMPACT ON LINGUISTIC ANALYSIS IN PSYCHOTHERAPIES

Martha Schubert[1], *Yamini Sinha*[1], *Julia Krüger*[2], *Ingo Siegert*[1]

[1]*Mobile Dialog Systems, IIKT,* [2]*Department of Psychosomatic Medicine and Psychotherapy, Medical Faculty, Otto von Guericke University Magdeburg, Germany*
martha.schubert@ovgu.de

Abstract: Modern intervention planning in psychotherapies may benefit from predicting process relevant psychotherapy constructs by automated speech analysis. One essential step is the extraction of relevant linguistic speech markers by ASR engines, which because of highly sensible data, work offline. We analyze transcription errors from NeMo, Whisper, and Wav2Vec2.0, focusing on their impact on linguistic markers usually requiring high quality transcripts. By utilizing part-of-speech tagging, we examine error occurrences among different word types. The Linguistic Inquiry and Word Count (LIWC) software aids in extracting markers. We highlight challenges in transcribing spontaneous speech, prevalent in therapy, and compare results with the Mozilla CommonVoice dataset, which features read speech.

1 Motivation

Modern psychotherapy strives for evidence-based situational decisions on interventions. Automated intra-session speech analysis has the potential to support this by gathering relevant psychotherapeutic process constructs [1, 2, 3]. For instance, in our pilot study, ASPIRE, we are working on the prediction of the quality of the psychotherapeutic alliance, known as the most relevant predictor of psychotherapy outcome [4], utilizable by automated analysis of speech content and prosodic-acoustic markers in patients' and therapists' speech. Especially for analyses of the speech content, the correctness of the transcription is crucial. Furthermore, the use of manual transcription is time-consuming and not feasible for the intended project. Using online cloud solutions is not appropriate for highly sensitive data from psychotherapies due to privacy issues. Therefore, a relevant step towards the development of reliable intra-session speech analysis is to extract relevant linguistic speech markers by offline Automated Speech Recognition (ASR) engines. Up to now, automatic transcription entails high error rates when recognizing spontaneous speech, although Automatic Speech Recognition (ASR) systems have undergone continuous evolution in recent years [5].

This success is attributed to the adoption of sophisticated deep network architectures, featuring efficient training methods for both acoustic and language modeling. Presently, deep convolutional neural network architectures are prevalent for acoustic modeling, while variants of long-short-term memory networks are employed for both acoustic and language modeling [6].

In the context of psychotherapeutic alliance, not only the pure Word Error Rate (WER) is of interest, but also the correctness of the linguistic analysis. To evaluate linguistic measures, we used the Linguistic Inquiry and Word Count (LIWC) software, which categorizes (written) speech. Those markers are influenced by the automatic transcription to varying degrees, which are explored in our work.

DOI: 10.35096/othr/pub-7099, URN: urn:nbn:de:bvb:898-opus4-70999

2 Related Work

Regarding the WER, numerous studies have explored the performance of various cloud-based ASR systems using subsets of the Switchboard telephone speech dataset for English benchmarking [7]. Recent cloud-based ASR systems demonstrated WER values as low as 10%-5% WER: IBM Research reported a WER of 5.1% on the Switchboard Hub5 2000 evaluation test with their ASRU'17 system, incorporating unsupervised Language Model (LM) adaptation[8]. Similarly, Microsoft Research achieved a 5.1% WER on the NIST 2000 Switchboard task, employing ngram rescoring [6]. Google's cloud-based ASR system, although lacking Switchboard dataset results, achieved 6.7% WER on voice search and 4.1% WER for a dictation task using an internal traffic application dataset [9]. But it is apparent that all this data comprises command-style data or dictation-style data, usually very different from the data used in our ASPIRE study. Several studies also conducted research regarding the ability to identify the context on the ASR-genereted text. According to Wirth and Peinl, 9.40% of German ASR errors are deemed negligible, while 11.95% are noncontext-breaking. However, a significant majority, comprising 19.01% of errors, is identified as context-breaking. An additional 19.82% of errors involve names, anglicisms, or loan words, as reported in [10]. This aligns with the observation that words appearing at low frequencies in the language model corpus or those entirely outside the vocabulary are prone to misrecognition by ASR engines. Ma et al. also corroborate this finding, highlighting the substantial challenge posed by misrecognizing words of great importance, particularly proper names, within specific application contexts [11]. Interjections present another difficulty for ASR engines, since Siegert et al could only detect one German ASR engine (IBM) that could recognize nearly all of those filler words [12]. In the English language, monosyllable function words contribute most to the WER. Content words, however, which rarely appear in the training corpus, are also misrecognized most of the time [13]. It has also been shown that spontaneous speech is harder to recognize than read speech, Nakajima et al. detect a deterioration in detecting spoken Japanese when switching from read to conversational speech. For content words, they observe an even higher deterioration rate (mean 36.6%) than for function words (mean 28.0%). [14] Similar difficulties in detecting spontaneous speech have been observed by Silber-Varod et al. who report a higher WER for spontaneous dialogues than for frontal lectures [15].

3 Methods

We used 3 different ASR engines to transcribe the audio samples from two different databases, which will be elaborated upon in the following:

3.1 Datasets

Ulm State of Mind in Speech (USoMs) database: The first database we used is the USoMs database. It consists of approximately 1k recordings of spontaneous speech with psychotherapy context and their manually created transcriptions, which have been used as a gold standard. The speakers are not patients, but volunteers who tell short stories from their personal experience [16]. The corpus consists of recordings of younger and older people, but because the recordings of the older people were largely incomplete, we concentrated on the sub-corpus of the younger people and extracted a subset consisting of 70 of these recordings, which are fully transcribed.
Mozilla CommonVoice 7.0 (MCV7.0): The German Mozilla CommonVoice database, in contrast to the USoMs database, consists of read speech with no psychotherapeutical context. Despite the absence of fillers like "ähm" in the non-spontaneous speech, the sentences comprise partly everyday conversations and therefore contain words that are potentially hard to recognize

for ASR engines, such as proper names of places or people. The designated subset for ASR engine testing, labelled as "test," encompasses 30,569 different words. [17]

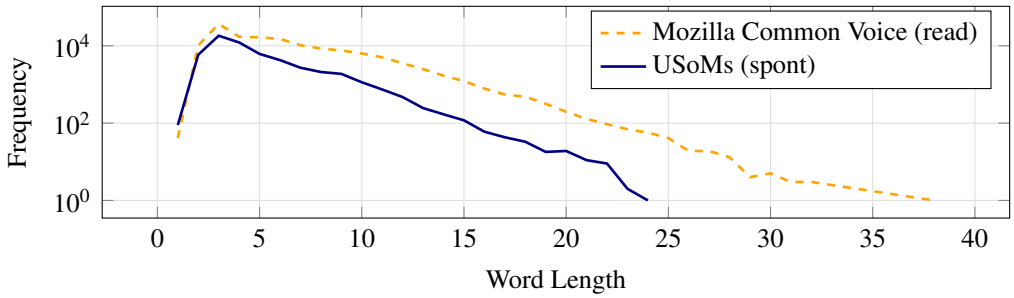

Figure 1 – Word Length of CommonVoice (read) and USoMs (spont) dataset

As depicted in Figure 1, the vast majority of the words in both corpora are about 5 letters in length, aligning with the principles of Zipf's Law of abbreviation, which states that shorter words appear more often than long ones. Additionally, it is evident that longer words are less frequently used in spontaneous speech and more prevalent in read speech.

3.2 ASR-Engines

NeMo Conformer Transducer Model [18]: The autoregressive NeMo Conformer Transducer Model, developed by NVIDIA, combines convolutional neural networks and transformer models. It has been trained using supervised learning on a dataset comprising 2,300 hours of speech data. Table 1 shows the exact databases the network has been trained and evaluated with.
Wav2Vec2.0 [19]: Unlike the NeMo model, Wav2Vec2.0 (we utilized the facebook/wav2vec2-xls-r-1b model [20]) underwent training through self-supervised learning, without the use of annotated data. Annotated data is only used for fine-tuning on the German language. The applied datasets are listed in Table 1).
Whisper [21]: Whisper is a weakly supervised model, meaning that part of the training data has been transcribed, but the majority of the data used in the training process is unlabelled. The utilized Datasets are indicated in Table 1.

All models were evaluated on the Mozilla CommonVoice (MCV) data set, so that despite different versions, the type of data remains the same (read speech) and the results are therefore easily comparable.

Table 1 – Comparison of ASR Engines - Parameters, Databases and WER

	NeMo	Whisper	Wav2Vec2.0
Parameters	120M	769M (medium)	1B
Training Databases	MCV7.0 Multilingual LibriSpeech VoxPopuli	internet audios and transcripts 13,344k h of German audio	MCV8.0 Multilingual LibriSpeech VoxPopuli Multilingual TEDx
Evaluation Database	MCV7.0	MCV9.0	MCV8.0
WER	4.93%	6.4%	10.95%

4 Linguistic Evaluation

4.1 Linguistic Inquiry and Word Count - LIWC

The LIWC software simplifies the analysis of written language. Utilizing the German Dictionary, the text can be systematically categorized into 100 distinct categories. The value for a category indicates the proportion of words that correspond to this category [22]. These categories include both grammatical evaluations so that e.g. the proportion of certain word types or tenses can be calculated, as well as content-related evaluations with which the proportion of certain topics in the text can be determined [23].

In addition to this evaluation method, there are 4 summary variables which, in contrast, are not calculated as a percentage of the total number of words, but are based on standardized values from the developers' research. Percentiles between 1 and 99 are calculated from their comparison corpora. Those four summary variables are [24] :

- **Analytical Thinking:** The Analytic value serves as an indicator of the prevalence of analytical thinking in a speaker's language, reflecting the proportion of words employed that convey a sense of formality and logic. It provides insight into the articulate use of words that embody a structured and rational discourse.
- **Clout:** Clout is used to indicate the social status and self-confidence of the person speaking or writing.
- **Authenticity:** This value shows how honestly or authentically the person speaks: Individuals with higher values typically express themselves more spontaneously, while a lower Authentic score suggests a more thoughtful approach, where the speaker carefully considers their words and observes themselves.
- **Emotional Tone:** Tone provides information about the general emotional tone of voice and indicates whether it is more positive (for values above 50) or more negative (for values below 50).

4.2 Part-of-Speech (POS)-Tagging

For POS-Tagging we used the "de_core_news_md" model from the spaCy python library [25], which has an accuracy of 98.29%. Using a trained pipeline, the program predicts which POS-tag (see Table 2 for an overview) is most fitting in the context of the sentence.

Table 2 – Explanation POS-Tags

Tag	Explanation	Tag	Explanation	Tag	Explanation
INTJ	interjection	X	other	PROPN	proper noun
NUM	numeral	AUX	auxiliary	DET	determiner
PART	particle	PRON	pronoun	ADV	adverb
ADJ	adjective	VERB	verb	SCONJ	subordination
NOUN	noun	ADP	adposition		conjunction

5 Results

5.1 Word Error Rate

As depicted in Figure 2, the average WER (calculated using the jiwer python library) for each model significantly increases in spontaneous speech compared to read speech. This aligns with

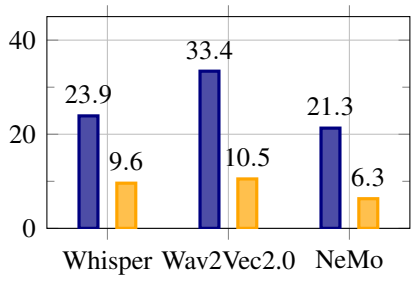

Figure 2 – Mean WER for each model on spontaneous speech ( ) and read speech ( )

		DEL	SUB	INS
spont	Whisper	7.83%	5.57%	1.59%
	Wav2Vec2.0	15.2%	17.7%	0.8%
	NeMo	7.6%	12.6%	1.1%
read	Whisper	0.74%	6.91%	1.18%
	Wav2Vec	0.89%	8.39%	0.76%
	NeMo	0.36%	4.27%	1.16%

Table 3 – Percentages of error types (excluding hits)

previous findings [14, 15] mentioned in section 2. Additionally, our findings reveal that NeMo is the best performing model, while Wav2Vec2.0 exhibits the least favourable performance, which also aligns with the WER provided by the developers (Table 1).

In Table 3 it is additionally shown that for Whisper and Wav2Vec2.0 the most prevalent error type on spontaneous speech are deletions (DEL) followed by substitutions (SUB) and insertions (INS) whereas for read speech substitutions are the most frequently occurring error type.

5.2 Influence of recognition mistakes on linguistic markers

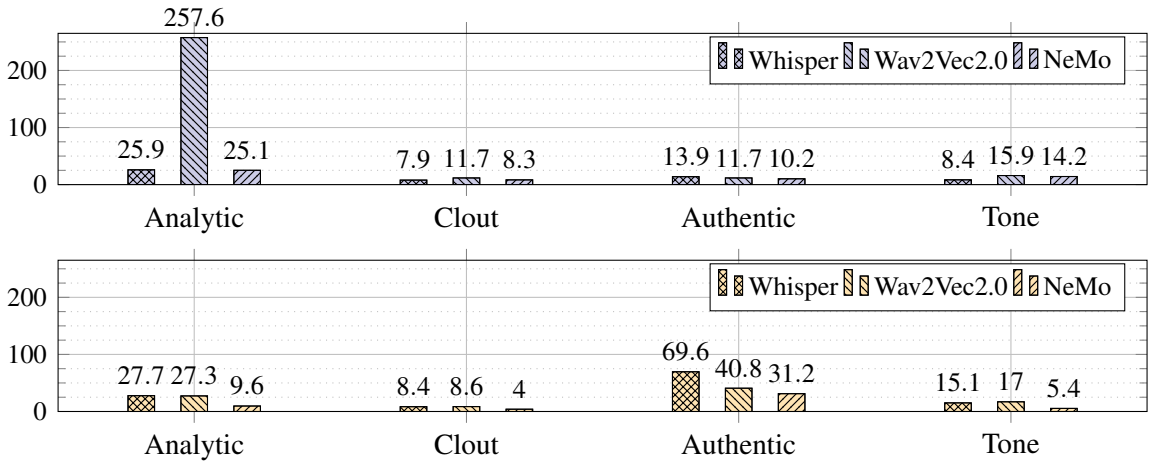

Figure 3 – Mean deviation of selected lingustic categories for each model on spontaneous speech (above) and read speech (below).

As shown in Figure 3, analysis conducted on the NeMo transcripts have the smallest deviations from the gold standard values. Interestingly, the performance is generally better on spontaneous speech than on read speech for Whisper and Wav2Vec2.0. This observation may be attributed to the characteristics of the datasets. Spontaneous speech transcripts are longer compared to those of read speech, which involve only one sentence each. Consequently, conducting LIWC analysis on spontaneous speech might yield more accurate results, as individual words and errors have less significance in the context of longer texts.

While most categories of spontaneous speech do not exceed a deviation of 26%, the Analytic category for the Wav2Vec2.0 model presents an exception with a deviation of 257.6%. This could be explained by the fact that spontaneous speech generally results in very low analytic values, and therefore even small deviations have a big impact in the overall category.

Furthermore, on read speech the Authentic value generally shows higher deviations from the gold standard value than the other categories. This could be due to decreased spontaneity when

reading, which amplifies the impact of mistakes in the ASR transcripts on this overall category.

5.3 Misrecognized Words in POS-tagging

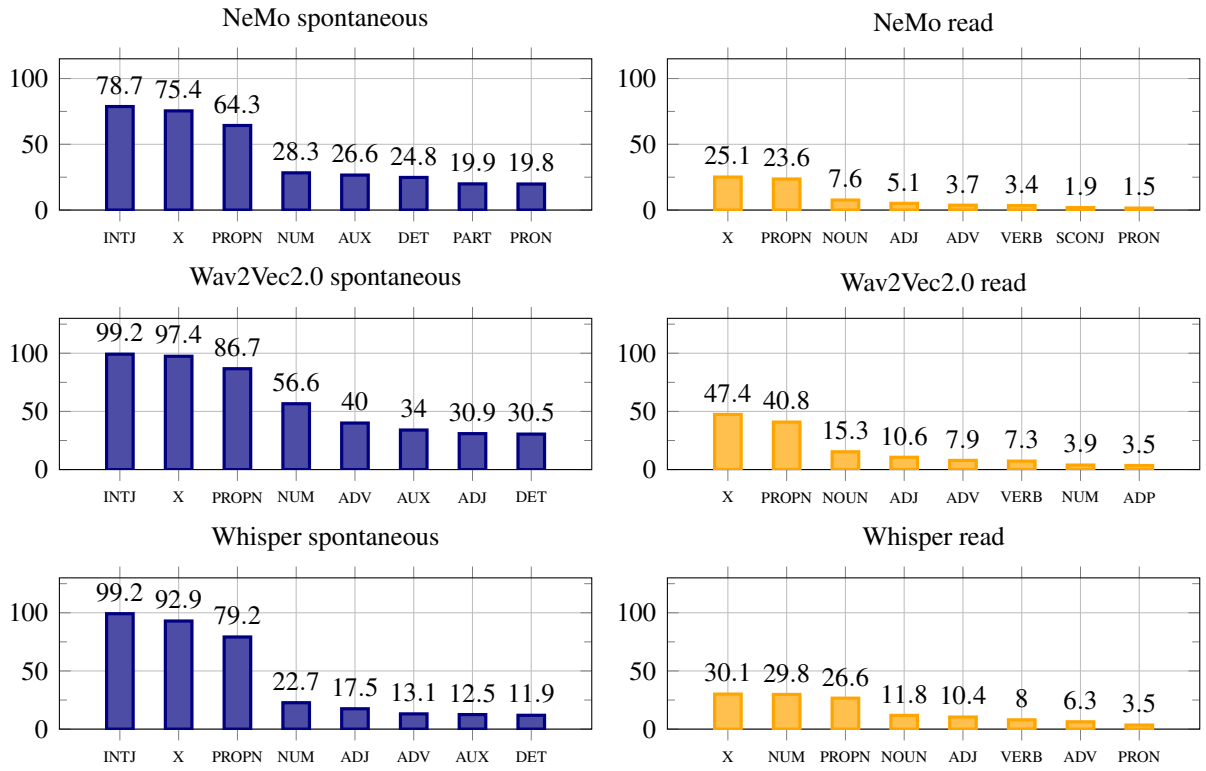

Figure 4 – Percentages of misrecognized words (selection) in automatically obtained transcripts

As shown in Figure 4 all word types show higher deviations for spontaneous speech than for read speech. Significantly, interjections (abbreviations provided in Table 2) present the most commonly misunderstood word types in spontaneous speech. This is attributed to the challenge faced by most ASR engines in recognizing filler words, as indicated by [12]. In contrast, read speech entirely lacks this category of words. This disparity can easily be explained by the characteristics of both speech types, since interjections are typically not used in written text.

Further commonly misunderstood word types are unconventional terms labelled as X and proper nouns. This coincides with the previously mentioned challenge, ASR engines face when transcribe words that are rare or absent in the training dataset [10, 11].

6 Conclusion

Our findings reveal that offline ASR engines can be used to extract relevant linguistic markers in sensible, spontaneous speech. However, some ASR engines are more suitable than others: NeMo, which is the model with the lowest WER, also performs the best when extracting the LIWC-markers, whereas Wav2Vec2.0 leads to higher deviations. Interestingly, even though the WER on spontaneous speech is rather high, linguistic markers can still be extracted reliably, partly even with a higher accuracy than on read speech.

References

[1] KUČERA, D. and M. R. MEHL: *Beyond english: Considering language and culture in psychological text analysis. Frontiers in Psychology*, 13, 2022. doi:10.3389/fpsyg.2022.819543. URL http://dx.doi.org/10.3389/fpsyg.2022.819543.

[2] CHEKROUD, A. M., J. BONDAR, J. DELGADILLO, G. DOHERTY, A. WASIL, M. FOKKEMA, Z. COHEN, D. BELGRAVE, R. DERUBEIS, R. INIESTA, D. DWYER, and K. CHOI: *The promise of machine learning in predicting treatment outcomes in psychiatry. World Psychiatry*, 20(2), p. 154–170, 2021. doi:10.1002/wps.20882. URL http://dx.doi.org/10.1002/wps.20882.

[3] KRÜGER, J., I. SIEGERT, and F. JUNNE: *Künstliche Intelligenz für die Sprachanalyse in der Psychotherapie – Chancen und Risiken. PPmP - Psychotherapie · Psychosomatik · Medizinische Psychologie*, 72, pp. 395–396, 2022. doi:10.1055/a-1915-2589.

[4] FLÜCKIGER, C., A. C. DEL RE, B. E. WAMPOLD, and A. O. HORVATH: *The alliance in adult psychotherapy: A meta-analytic synthesis. Psychotherapy*, 55(4), p. 316–340, 2018. doi:10.1037/pst0000172. URL http://dx.doi.org/10.1037/pst0000172.

[5] DHANJAL, A. S. and W. SINGH: *A comprehensive survey on automatic speech recognition using neural networks. Multimedia Tools and Applications*, 2023. doi:10.1007/s11042-023-16438-y.

[6] XIONG, W., L. WU, J. DROPPO, X. HUANG, and A. STOLCKE: *The Microsoft 2017 Conversational Speech Recognition System*. In *Proc. of the IEEE ICASSP-2018*, pp. 5934–5938. Calgary, Kanada, 2018.

[7] GODFREY, J. J., E. C. HOLLIMAN, and J. MCDANIEL: *SWITCHBOARD: telephone speech corpus for research and development*. In *Proc. of the IEEE ICASSP-1992*, vol. 1, pp. 517–520 vol.1. San Francisco, CA, USA, 1992.

[8] KURATA, G., B. RAMABHADRAN, G. SAON, and A. SETHY: *Language modeling with highway LSTM*. In *Proc. of the IEEE ASRU*, pp. 244–251. Okinana, Japan, 2017. doi:10.1109/ASRU.2017.8268942.

[9] CHIU, C., T. N. SAINATH, Y. WU, R. PRABHAVALKAR, P. NGUYEN, Z. CHEN, A. KANNAN, R. J. WEISS, K. RAO, E. GONINA, N. JAITLY, B. LI, J. CHOROWSKI, and M. BACCHIANI: *State-of-the-art speech recognition with sequence-to-sequence models*. In *Proc. IEEE ICASSP-2018*, pp. 4774–4778. Calgary, Kanada, 2018.

[10] WIRTH, J. and R. PEINL: *Automatic speech recognition in german: A detailed error analysis*. pp. 1–8. 2022. doi:10.1109/COINS54846.2022.9854978.

[11] MA, X., X. WANG, and D. WANG: *Low-frequency word enhancement with similar pairs in speech recognition*. In *2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP)*, pp. 343–347. 2015. doi:10.1109/ChinaSIP.2015.7230421.

[12] SIEGERT, SINHA, JOKISCH, and WENDEMUTH: *Recognition Performance of Selected Speech Recognition APIs – A Longitudinal Study*, pp. 520–529. Springer International Publishing, Cham, 2020.

[13] HAHN, S., A. SETHY, H.-K. KUO, and B. RAMABHADRAN: *A study of unsupervised clustering techniques for language modeling*. pp. 1598–1601. 2008. doi:10.21437/Interspeech.2008-266.

[14] NAKAJIMA, H., I. HIRANO, Y. SAGISAKA, and K. SHIRAI: *Pronunciation variant analysis using speaking style parallel corpus*. In *Proc. 7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, pp. 65–68. 2001. doi:10.21437/Eurospeech.2001-15.

[15] SILBER-VAROD, SIEGERT, JOKISCH, SINHA, and GERI: *A cross-language study of selected speech recognition systems*. *The Online Journal of Applied Knowledge Management: OJAKM*, 9, pp. 1–15, 2021. doi:10.36965/OJAKM.2021.9(1)1-15. URL https://doi.org/10.36965/OJAKM.2021.9(1)1-15.

[16] RATHNER, E.-M., Y. TERHORST, N. CUMMINS, B. SCHULLER, and H. BAUMEISTER: *State of mind: Classification through self-reported affect and word use in speech*. In *Proc. Interspeech*. 2018.

[17] ARDILA, R., M. BRANSON, K. DAVIS, M. KOHLER, J. MEYER, M. HENRETTY, R. MORAIS, L. SAUNDERS, F. TYERS, and G. WEBER: *Common voice: A massively-multilingual speech corpus*. In *Proc. of the 12th Language Resources and Evaluation Conference*, pp. 4218–4222. ELRA, Marseille, France, 2020. URL https://aclanthology.org/2020.lrec-1.520.

[18] GULATI, A., J. QIN, C.-C. CHIU, N. PARMAR, Y. ZHANG, J. YU, W. HAN, S. WANG, Z. ZHANG, Y. WU, and R. PANG: *Conformer: Convolution-augmented transformer for speech recognition*. 2020. 2005.08100.

[19] BAEVSKI, A., H. ZHOU, A. MOHAMED, and M. AULI: *wav2vec 2.0: A framework for self-supervised learning of speech representations*. 2020. 2006.11477.

[20] GROSMAN, J.: *Fine-tuned XLS-R 1B model for speech recognition in German*. https://huggingface.co/jonatasgrosman/wav2vec2-xls-r-1b-german, 2022.

[21] RADFORD, A., J. W. KIM, T. XU, G. BROCKMAN, C. MCLEAVEY, and I. SUTSKEVER: *Robust speech recognition via large-scale weak supervision*. 2022. 2212.04356.

[22] PENNEBAKER, J., M. FRANCIS, and R. BOOTH: *Linguistic inquiry and word count (liwc)*. 1999.

[23] TAUSCZIK, Y. R. and J. W. PENNEBAKER: *The psychological meaning of words: Liwc and computerized text analysis methods*. *Journal of Language and Social Psychology*, 29(1), p. 24–54, 2009. doi:10.1177/0261927x09351676. URL http://dx.doi.org/10.1177/0261927X09351676.

[24] BOYD, R. L., A. ASHOKKUMAR, S. SERAJ, and J. W. PENNEBAKER: *The development and psychometric properties of LIWC-22*. University of Texas at Austin, Austin, TX, 2022.

[25] HONNIBAL, M. and I. MONTANI: *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*, 2017. To appear.