NoiSLU: a Noisy speech corpus for Spoken Language Understanding in the Public Transport Domain

Mariano Frohnmaier[1], *Steffen Freisinger*[1], *Madeline Faye Holt*[2], *Munir Georges*[1,3]
[1]*Technische Hochschule Ingolstadt,* [2]*Georgia Institute of Technology,* [3]*Intel Labs, Germany*
mariano.frohnmaier@mailbox.org

Abstract: The use of local public transport requires the barrier-free purchase of a ticket. Travellers who are not proficient in the local language benefit from a multilingual human(ticket)machine voice interaction. This paper presents a nearly parallel audio dataset with 13218 annotated user queries from 20 speakers for English, German and Dutch. The domain-specific speech corpus can be understood as an evaluation dataset for future research in Spoken Language Understanding (SLU) and thus, it enables researches to improve the quality of human-machine interaction applications. Furthermore, we compare the SLU performance of different compositions of Automatic Speech Recognition (ASR) and Natural Language Understanding (NLU) models in baseline experiments on different test datasets.

1 Background

Public Transport (PT) Information Systems realized through Spoken Dialogue Systems (SDS) [1] are valued by blind people [2] and can serve non-native speakers to obtain information regarding PT [3]. In this work, we consider a voice-enabled ticket vending machine (TVM) installed in transport hubs such as bus or train stations for requesting PT information or buying a ticket. To make the use of PT more attractive and easier accessible for diverse customer groups, as encouraged by [4], [5] such a voice-enabled TVM should not only be able to adapt to different languages, but also to operate under domain-specific ambient noises like departing busses, arriving trains or chattering people. In addition, the SDS should be efficient, e.g. using low rank factorization methods or 6 hardware accelerators [7]. Finally, data privacy constraints needs to be applied, such as proposed by, e.g., 8. All together this makes the task challenging. To the best of our knowledge there exists no SLU corpus that allows multi-lingual voice interaction with a TVM, where user's speech is disturbed by environmental sounds. The contribution of this work is as follows:

- We publish[1] a multilingual SLU dataset with intent and slot annotations (incl. ASR transcriptions) for the PT domain where users interact with a TVM while being exposed to environmental noises, such as busses or trains passing by, and chattering people.
- We present baseline results of SLU experiments for different speech recording scenarios. The SLU architectures are a composition of Automatic Speech Recognition (ASR) models (variants of Whisper [9] and Wav2Vec2 [10]) and state-of-the-art Natural Language Understanding (NLU) models available in the OpenSLU framework [11] for joint Intent Detection (ID) and Slot Filling (SF).

The corpus can be used as an evaluation data set to investigate multi-lingual SLU for German and English, Dutch and Flemish. For the latter two languages, only intent labels are annotated, the slot labelling is in progress.

[1]More information about access to the dataset: `https://github.com/M4R14NO/NoiSLU`

 DOI: 10.35096/othr/pub-7084, URN: urn:nbn:de:bvb:898-opus4-70848

Table 1 – Number of spoken utterances for different recording types (Clean Speech, Loudspeaker Speech, Re-Recordings, Outside) and total number of recorded hours per language.

Language	CS	LS	RR	OUTS	**Total (hours)**
German	1414	307	678	59	2458 (2,7 h)
English	1437	242	635	484	2798 (3,1 h)
Dutch & Flemish	628	0	628	0	1256 (1,4 h)
Total	3479	549	1941	543	6512 (7,2 h)

Table 2 – Speaker Mother Tongue Statistics (n=20).

Female	Male	German	English	Dutch	Other
45%	55%	45%	15%	15%	25%

2 Method

The data collection was carried out in two stages: The goal of phase (1) was to generate an NLU text corpus $\mathscr{D}^{\text{text}}$ consisting of user utterances for natural interaction with a TVM in four different languages. In phase (2), we invited 20 volunteer speakers for speech recordings of a subset $\mathscr{D}^{\text{audio}}$ of the original text corpus $\mathscr{D}^{\text{text}}$ under different recording conditions.

2.1 Collection of Annotated Text Corpus

To identify an initial set of domain-specific user intents I_0, we examined the existing touch interaction possibilities of a TVM of Ingolstadt's local public transport company. To determine a first set of relevant slot values, the price list and timetable from the local service provider served as a reference. An iterative process of writing down potential user utterances in German and refining the set of intent labels has led to the current set I of intents, see Table 3. For labelling the utterances $\mathscr{D}^{\text{text}}$ with slots, we used the open-source tool Rubrix[2]. In total, there are 23 unique intents and 9 different slot labels. The slot tags are provided in the BIO format. The utterances were first collected in German and machine-translated to English. The obtained English dataset $\mathscr{D}^{\text{text}}_{\text{EN}}$ has been expanded by a native speaker. A subset of the German text data $\mathscr{D}^{\text{text}}_{\text{DE}}$ was translated to both Flemish and Dutch ($\rightarrow \mathscr{D}^{\text{text}}_{\text{NL}}$) by a Flemish mother tongue speaker, but these are not yet annotated with slots.

2.2 Audio Data Collection

A subset of audio samples for $\mathscr{D}^{\text{text}}$ was obtained in 3 steps: (1) we recorded ambient noises at Ingolstadt's main and north train station, followed by (2) speech recordings with 20 speakers in a semi-anechoic chamber. Three speakers participated in outside recordings. Finally, (3) we created re-recordings of previously collected clean speech and ambient noises. We categorize these recordings as follows:

1. **Ambient noise recordings.** We collected ambient noises at Ingolstadt's north train station using a Zoom H6 field recorder. At this highly frequented transport hub we obtained a wide range of background noises, ranging from passing trains and busses to chattering people next to TVMs.

[2]https://rubrix.readthedocs.io/en/v0.4.1/

2. **Speech Recordings in- and outside.** We recorded 20 volunteer speakers in a semi-anechoic chamber, where each participant was asked to read aloud a subset of utterances from $\mathscr{D}^{\text{text}}$. The Lombard Effect [12] was considered by playing back ambient noises in two different ways: In the first scenario, speakers listened to ambient noises over a headset while reading aloud utterances, resulting in clean speech (**CS**). In a second recording setup, background noises were played back over a pair of loudspeakers next to the participant's screen, resulting in loudspeaker disturbed speech (**LS**). Additionally, three speakers were recorded outside next to a TVM at the train station Ingolstadt Nord, resulting in outside speech (**OUTS**).

3. **Re-Recordings of Clean Speech and Ambient noises.** We gathered more noisy speech recordings as follows: Both, (CS) and environmental sounds from the previous steps were played back and recorded simultaneously in the semi-anechoic chamber to obtain noisy re-recordings (RR) at different ambient noise volumes.

Table 1 shows overall number for English, German and Dutch Recordings. However, for 659 German and 682 English recordings there exist recordings from 5 other microphones (4 of them from a microphone array), summing up to additional 6706 speech recordings (not included in the table statistics). In total, we obtained 13218 recordings from three different microphones.

Table 3 – General statistics about most of the spoken queris for selected intents. The displayed intents make up 89,6 percent of all English vocalizations.

Intent	**% of utterances**			**∅ slots/utterance**			**∅ tokens/utterance**		
	EN	DE	NL	EN	DE	NL	EN	DE	NL
SelectStation	18.17	13.74	6.00	1.09	1.09	0.0	5.99	4.35	5.36
SelectDate	16.22	15	5.74	1.48	1.6	0.0	3.58	3.01	2.98
GetTicketprice	13.17	13.52	3.48	1.63	2.37	0.0	8.26	6.36	6.26
SetTicketAmount	11.61	10.87	6.13	1.33	1.74	0.0	4.70	3.76	4.17
RequestTickets	9.96	9.3	2.52	2.88	2.9	0.0	7.83	6.81	7.62
ShowRouteToStation	5.65	3.35	1.78	1.01	1.21	0.0	7.65	6.19	6.00
FindConnection	4.74	4.57	2.22	1.98	1.85	0.0	5.79	5.07	4.92
AskForDiscount	4.09	4.30	2.70	1.07	0.91	0.0	6.15	5.41	5.13
GetDepartureTimes	3.22	1.61	0.48	2.35	2.14	0.0	7.24	6.70	6.91
BrowseTimetable	2.74	3.13	1.74	1.00	1.10	0.0	3.78	3.39	2.90

3 Evaluation

In this section, we present first SLU results on our dataset. The SLU models are a composition of different ASR and NLU models. We did not conduct ASR training or adaptation on the dataset. To detect intent and slots, we used 3 different state-of-the-art NLU models available in the open-source toolkit OpenSLU by Qin et al. [11].

3.1 ASR Experiment Setup and Baseline Results

The quality of the utterance transcription has a crucial influence on the downstream NLU task performance. To incorporate this dependency into our evaluation, we use different ASR model

setups to create transcriptions for the English and German data. We use five different Whisper models of sizes *tiny*, *base*, *small*, *medium* and *large*. Whisper is an end-to-end speech recognizer which was trained on 680,000 hours of multilingual and multitask data [9]. Since Whisper can handle various languages, we use the same models among both languages. We further employ XLSR-53 [13] models for English and German. Each of them was fine-tuned on the respective Mozilla Common Voice dataset and released by Grosman [14]. In addition, these two models were used with and without a language model. To determine the accuracy of the aforementioned ASR systems in the target domain, we evaluate them on our ticket machine recordings. Figure 1 gives an overview of the word error rates (WER) for the different model setups, showing that Whisper *large* provides the best transcription quality among both languages.

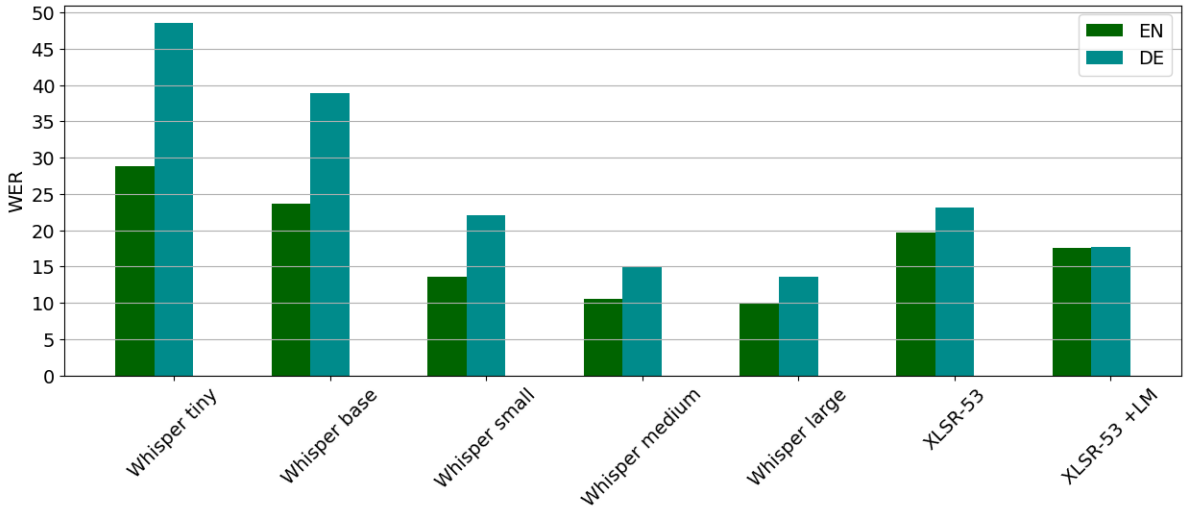

Figure 1 – Comparison of word error rates (WER) for every ASR model setup. The data which was used consists of all German and English speech recordings from Table 1.

3.2 NLU Experiment Setup and Baseline Results

We utilized the open-source framework *OpenSLU* by Qin et al. [11] to conduct baseline NLU experiments for joint Intent Detection and Slot Filling for English and German. In the experiments, we selected the existing recipes of 3 models where good results were obtained on the ATIS dataset: the first two models, DeBERTaV3 by He et al. [15] and JointBERT by Chen et al. [16] are pretrained models, whereas the third model, StackPropagation by Qin et al. [17], is non-pretrained. For NLU training with English data, we used the encoder models *bert-base-uncased*, Microsoft's *DeBERTa-v3-base*, and GloVe [18], respectively. For an NLU performance comparison between German and English, we used the JointBERT architecture with *dbmdz/bert-base-german-uncased* as German encoder model. The DeBERTaV3 and JointBERT decoder consist of linear classifiers for both intent and slot detection. The StackPropagation model uses auto-regressive Long-Short-Term-Memory (LSTM) classifiers for intent and slot recognition.
Evaluation metrics. The intent detection performance is evaluated by means of weighted average F1 score (f1), precision (prec) and recall (rec) scores due to class label imbalance. We further make use of a strict evaluation method for the slot filling task, namely the *strict* evaluation scheme implemented in the PyPi Package *nervaluate* from MantisAI [19]. The reason for this is the intended usage of the NLU output. The recognized intent and slots will be used to query the database from the local traffic provider. In this scenario, even a slight slot detection error can lead to an incorrect query and hereafter to bad user experience. According to this

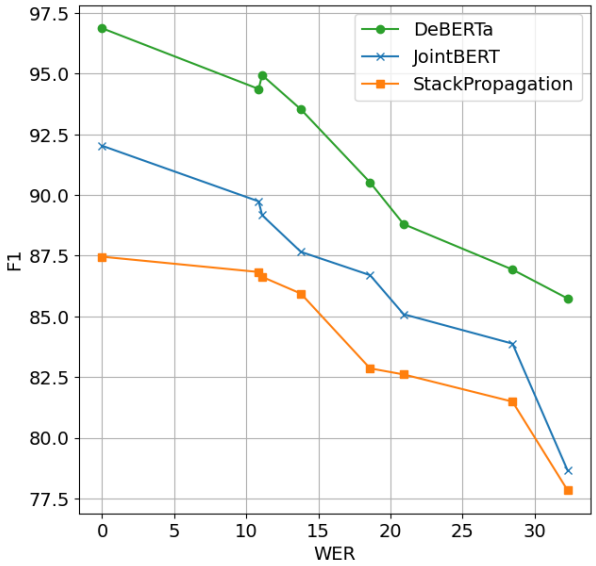

Figure 2 – Intent detection f1-score based on different WERs. Reported numbers are for all speech types (English).

Table 4 – Number of unique sentences (texts) and audio vocalizations for NLU train, dev and test splits with an average utterance duration of 4.1 seconds.

Split	# texts		# audios	
	EN	DE	EN	DE
train	610	616	1928	1704
dev	77	78	233	218
test	202	200	637	536
total	889	894	2798	2458

protocol, slot recognition is correct if both the actual string and the slot type match the reference. Table 4 shows general utterance statistics for the train/dev/test split used in the English and German NLU model training.

3.2.1 Influence of different ASR Models on Intent Detection

The influence of the transcription quality on the intent detection performance is illustrated in Figure 2. The plot clearly shows for all three models, that a higher WER has a negative impact on the NLU performance. This is as expected and highlights the importance of using an ASR component, which has a high transcription accuracy in the target domain.

3.2.2 SLU Results: Gold Transcriptions vs. ASR Input

Table 5 – Comparison of NLU model performance with and without preceding ASR model. Used ASR model: Whisper Large V2. The symbol x refers to gold transcriptions with zero word error rate (wer).

NLU Model	ASR	Intent results in %			Slot results in %		
	wer	pre	rec	f1	pre	rec	f1
StackPropagation	x	89.05	88.23	87.46	84.73	84.21	**84.47**
	10.86	88.67	87.60	86.83	75.98	75.89	75.93
DeBERTaV3	x	97.46	96.86	**96.87**	62.15	64.50	63.30
	10.86	95.23	94.51	94.37	56.38	58.38	57.37
JointBERT	x	92.24	92.62	92.03	52.48	67.44	59.03
	10.86	90.38	90.42	89.74	49.95	62.67	55.59

Table 5 compares the performance of NLU models (StackPropagation, DeBERTaV3, and JointBERT) using gold transcriptions and ASR outputs from Whisper Large V2. Our results

show that StackPropagation performs the best in slot filling, and DeBERTaV3 performs the best in intent recognition. Notably, all models have a drop in performance when WER>0 is introduced for both intent recognition and slot filling tasks. Slot filling tasks appear especially sensitive to the increase in WER due to the specific nature of slot filling where erroneous inputs have a greater impact than the broader task of intent recognition. Additionally, the decline in results across models when WER>0 further motivates the development of more error-tolerant NLU models, especially for situations where ambient noise exists such as a TVMs.

3.2.3 SLU Results for different Recording Types

Table 6 shows a high difference in slot filling performance between StackPropagation NLU model and the two pretrained models (DeBERTa, JointBERT). This is caused by different tokenization strategies from the respective NLU recipes. Unknown words are divided into subword tokens in the pretrained cases. Hence, the predicted slot label sequences are longer than the reference slot sequences for utterances with domain-specific vocabulary unknown to the pretrained embeddings. As we expected, $\mathrm{m(CS)} < \mathrm{m(OUTS)} < \mathrm{m(RR)}$ for metrics $\mathrm{m} \in \{\mathrm{WER}, \mathrm{Slot\text{-}f1}\}$. This means, the artificial re-recording test dataset is harder to recognize than our realistic outside recordings for both, the speech and slot recognition systems. But interestingly, the intent detection results for realistic outside speech recordings is not affected by the latter fact. In fact, the best intent detection results are obtained on the outside speech recording test data.

Table 6 – SLU results for English Test data with preceding ASR component (Whisper Large V2) for different recording types.

Rec.	#	NLU	ASR	Intent results in %			Slot results in %		
Type	utts	model	wer	pre	rec	f1	pre	rec	f1
CS	171	StackProp	8.98	90.48	90.06	88.63	81.30	82.02	81.66
		DeBERTa		95.73	95.91	95.58	59.34	62.72	60.98
		JointBERT		92.00	91.81	91.29	52.76	67.11	59.07
OUTS	127	StackProp	10.50	91.04	90.55	89.64	75.28	75.71	75.49
		DeBERTa		98.85	97.64	97.8	58.66	59.32	58.99
		JointBERT		95.39	95.28	94.72	46.29	59.89	52.22
RR80	140	StackProp	16.69	83.68	85.00	83.56	75.84	75.42	75.63
		DeBERTa		90.04	90.71	89.45	55.38	57.54	56.44
		JointBERT		86.21	85.71	84.71	50.91	62.57	56.14

3.2.4 SLU Performance Comparison: English vs. German

From the results in Table 7 we can observe that the ASR model Whisper large yields lower word error rates for both recording scenarios (CS, RR). However, the final SLU performance in terms of intent detection and slot filling metrics is better for the German language, given a specific recording scenario. This is due to code switching: the English utterances contain lots of German slot values. The results of the loudspeaker disturbed speech recordings (referred to by LS in Table 1) are excluded as there are only 49 utterances in the test split. However, the numbers are close to the results presented for the OUTS speech type in Table 6.

Table 7 – Comparison of Jointbert NLU performance for different languages (DE, EN). Recording types are CS and RR. The preceding ASR component is Whisper large.

RecType	Language	ASR wer	Intent f1 (%)	Slot f1 (%)	Number of utts.
CS	English	8.98	91.29	59.07	171
	German	13.45	93.19	60.18	161
RR	English	16.69	84.71	56.14	140
	German	23.12	91.23	57.14	144

4 Discussion and Conclusion

The first baseline results indicate that a composition of state-of-the-art ASR and NLU models yield comparable results to OpenSLU baseline results on the ATIS dataset. Due to the fact that the pretrained NLU model recipes use a different tokenizer than StackPropagation, the slot filling results might be updated after a tokenization alignment[3]. The NLU corpus might be further extended, e.g. via Large Language Model paraphrasings for each intent. Also, models based on Generative Adversarial Networks could be used to generate fake user queries.
Overall, the NoiSLU corpus contains 13218 (nearly parallel) annotated utterances for German, English and Dutch/Flemish with intent and slot annotations for the PT domain where users interact with a TVM while being exposed to environmental noises.

References

[1] STRIK, H., A. RUSSEL, H. VAN DEN HEUVEL, C. CUCCHIARINI, and L. BOVES: *A spoken dialog system for the Dutch public transport information service*. *International Journal of Speech Technology*, 2(2), pp. 121–131, 1997. doi:10.1007/BF02208824.

[2] JURČÍČEK, F., O. DUŠEK, O. PLÁTEK, and L. ŽILKA: *Alex: A Statistical Dialogue Systems Framework*. In P. SOJKA, A. HORÁK, I. KOPEČEK, and K. PALA (eds.), *Text, Speech and Dialogue*, Lecture Notes in Computer Science, pp. 587–594. Springer International Publishing, Cham, 2014. doi:10.1007/978-3-319-10816-2_71.

[3] RAUX, A., B. LANGNER, D. BOHUS, A. BLACK, and M. ESKENAZI: *Let's Go Public! Taking a Spoken Dialog System to the Real World*. pp. 885–888. 2005. doi:10.21437/Interspeech.2005-399.

[4] BEUL-LEUSMANN, S., E.-M. JAKOBS, and M. ZIEFLE: *User-centered design of passenger information systems*. In *IEEE International Professonal Communication 2013 Conference*, pp. 1–8. 2013. doi:10.1109/IPCC.2013.6623931.

[5] CARMIEN, S., M. DAWE, G. FISCHER, A. GORMAN, A. KINTSCH, and J. F. SULLIVAN: *Socio-technical environments supporting people with cognitive disabilities using public transportation*. *ACM Transactions on Computer-Human Interaction*, 12(2), pp. 233–262, 2005. doi:10.1145/1067860.1067865.

[6] GEORGES, M., J. HUANG, and T. BOCKLET: *Compact speaker embedding: lrx-vector*. In H. MENG, B. XU, and T. F. ZHENG (eds.), *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pp. 3236–3240. ISCA, 2020. doi:10.21437/INTERSPEECH.2020-2106. URL `https://doi.org/10.21437/Interspeech.2020-2106`.

[7] STEMMER, G., M. GEORGES, J. HOFER, P. ROZEN, J. G. BAUER, J. NOWICKI, T. BOCKLET, H. R. COLETT, O. FALIK, M. DEISHER, and S. J. DOWNING: *Speech recognition and understanding on hardware-accelerated DSP*. In F. LACERDA (ed.), *Interspeech 2017, 18th Annual Conference of the*

[3] E.g., with the tokenization alignment tool `https://github.com/explosion/tokenizations`

International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017, pp. 2036–2037. ISCA, 2017. URL https://www.isca-speech.org/archive/interspeech_2017/stemmer17_interspeech.html.

[8] GEORGES, M., S. KANTHAK, and D. KLAKOW: *Accurate client-server based speech recognition keeping personal data on the client*. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014*, pp. 3271–3275. IEEE, 2014. doi:10.1109/ICASSP.2014.6854205. URL https://doi.org/10.1109/ICASSP.2014.6854205.

[9] RADFORD, A., J. W. KIM, T. XU, G. BROCKMAN, C. MCLEAVEY, and I. SUTSKEVER: *Robust speech recognition via large-scale weak supervision*. 2022. doi:10.48550/ARXIV.2212.04356. URL https://arxiv.org/abs/2212.04356.

[10] BAEVSKI, A., Y. ZHOU, A. MOHAMED, and M. AULI: *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations*. In *Advances in Neural Information Processing Systems*, vol. 33, pp. 12449–12460. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html.

[11] QIN, L., Q. CHEN, X. XU, Y. FENG, and W. CHE: *OpenSLU: A Unified, Modularized, and Extensible Toolkit for Spoken Language Understanding*. In D. BOLLEGALA, R. HUANG, and A. RITTER (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 95–102. Association for Computational Linguistics, Toronto, Canada, 2023. doi:10.18653/v1/2023.acl-demo.9. URL https://aclanthology.org/2023.acl-demo.9.

[12] BOTTALICO, P., I. I. PASSIONE, S. GRAETZER, and E. J. HUNTER: *Evaluation of the starting point of the Lombard Effect*. *Acta acustica united with acustica : the journal of the European Acoustics Association (EEIG)*, 103(1), pp. 169–172, 2017. doi:10.3813/AAA.919043.

[13] CONNEAU, A., A. BAEVSKI, R. COLLOBERT, A. MOHAMED, and M. AULI: *Unsupervised cross-lingual representation learning for speech recognition*. *CoRR*, abs/2006.13979, 2020. URL https://arxiv.org/abs/2006.13979. 2006.13979.

[14] GROSMAN, J.: *Fine-tuned xlsr-53 large model for speech recognition in nglish*. 2021. URL https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english.

[15] HE, P., J. GAO, and W. CHEN: *DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing*. 2023. URL http://arxiv.org/abs/2111.09543. ArXiv:2111.09543 [cs].

[16] CHEN, Q., Z. ZHUO, and W. WANG: *BERT for Joint Intent Classification and Slot Filling*. 2019. URL http://arxiv.org/abs/1902.10909. ArXiv:1902.10909 [cs].

[17] QIN, L., W. CHE, Y. LI, H. WEN, and T. LIU: *A Stack-Propagation Framework with Token-Level Intent Detection for Spoken Language Understanding*. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2078–2087. Association for Computational Linguistics, Hong Kong, China, 2019. doi:10.18653/v1/D19-1214. URL https://www.aclweb.org/anthology/D19-1214.

[18] PENNINGTON, J., R. SOCHER, and C. MANNING: *GloVe: Global vectors for word representation*. In A. MOSCHITTI, B. PANG, and W. DAELEMANS (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543. Association for Computational Linguistics, Doha, Qatar, 2014. doi:10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162.

[19] BATISTA, D. and M. A. UPSON: *nervaluate*. 2020. URL https://github.com/MantisAI/nervaluate.