INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

# What do self-supervised speech representations encode? An analysis of languages, varieties, speaking styles and speakers

*J. Linke*[1], *M. S. Kádár*[2,3], *G. Dobsinszki*[2,3], *P. Mihajlik*[2,3], *G. Kubin*[1], *B. Schuppler*[1]

[1]Signal Processing and Speech Communication Laboratory, Graz University of Technology
[2,3]Budapest University of Technology and Economics, Department of Telecommunication and Media Informatics; Hungarian Research Centre for Linguistics

linke@tugraz.at, mate.kadar1@gmail.com, dobsinszki.g@gmail.com, mihajlik@tmit.bme.hu, gernot.kubin@tugraz.at, b.schuppler@tugraz.at

## Abstract

Automatic speech recognition systems based on self-supervised learning yield excellent performance for read, but not so for conversational speech. This paper contributes insights into how corpora from different languages and speaking styles are encoded in shared discrete speech representations (based on wav2vec2 XLSR). We analyze codebook entries of data from two languages from different language families (i.e., German and Hungarian), of data from different varieties from the same language (i.e., German and Austrian German) and of data from different speaking styles (read and conversational speech). We find that – as expected – the two languages are clearly separable. With respect to speaking style, conversational Austrian German has the highest similarity with a corpus of similar spontaneity from a different German variety, and speakers differ more among themselves when using different speaking styles than from other speakers of a different region when using the same speaking style.

**Index Terms**: conversational speech, German varieties, Hungarian, self-supervised speech representations, wav2vec2

## 1. Introduction

Research on automatic speech recognition (ASR) is strongly domain dependent due to the diversity of its applications (e.g., keyword spotting, dictation, and human interaction with social robots). Usually, each application is trained with different task-specific data sets. For continuous ASR, mostly two speaking styles are distinguished, read (RS) and spontaneous speech [1]. Probably the best-known RS corpus is Librispeech [1], where ASR performance already converges to its limit (2.5%) [2]. Also for less spontaneous conversational speech (CS) (e.g., Switchboard corpus [3]), performance reaches benchmark limits (4.3%) [4]. Nevertheless, for more spontaneous CS (i.e., casual face-to-face conversations), performance ranges only between 16% and 33%, given high inter-speaker and inter-conversation variation [5, 6].

One of the reasons for why ASR performance degrades with increasing degree of spontaneity is the reduced spectral space [7, 8]. The same authors also state that one of the most important research issues is how to train and adapt statistical models for speech recognition. Modern ASR architectures have a strong focus on adaptation by developing self-supervised learning of speech representations, such as those provided within the wav2vec2 framework, which make use of large amounts of unlabeled multilingual data (e.g., XLSR [9, 10]). The experiments in [5] and [11] showed that ASR performance improves by finetuning the XLSR model with labeled data coming from a target domain (i.e., in both cases different varieties of German). For Hungarian conversational speech, [12] reached absolute WER improvements of approx. 12%. Furthermore, for telephone CS from low-resourced languages (BABEL) [13, 14], large WER improvements were reported on out-of-pretraining languages in comparison to baselines (e.g., absolute WER improvements of 9% on Swahili or 7.4% on Tagalog). The question arises what kind of information initial XLSR speech representations encode, as even out-of-pretraining languages seem to be well represented after finetuning. The aim of this work is to analyze initial XLSR speech representations to gain insights about how they encode data from different languages, their varieties, different speaking styles and different speakers. We aim at contributing to a better understanding of self-supervised speech representations, which is of interest not only to scientists in the field of ASR, but also to speech scientists interested in acoustic characteristics of different speaking styles.

Our approach towards finding an answer to this question is inspired by the analysis on similarity matrices of XLSR codebook entries for 12 or 17 different languages by Conneau et al. [10]. Whereas that study demonstrated how the codebook entries group together related languages, in this paper we take the approach one step further by analyzing not only different languages, but also different language varieties, and individual speakers of different speaking styles (i.e., read, spontaneous-task oriented, and casual conversational speech). More concretely, we compare two languages from to different languages families, where one is an out-of-pretraining language (i.e., Hungarian) and one is an in-pretraining language (i.e., German). In addition, we perform a speaker-wise analysis, allowing us not only to study the distances of languages, styles and varieties, but also the distances between speakers, as well as the distance of speakers with themselves when producing different styles. We aim at answering the more general research question of whether the frequency usage of shared discrete speech representations (given by XLSR) encode acoustic properties/characteristics for different languages, varieties, speaking styles and speakers.

## 2. Materials

Our experiments are based on German (G), Austrian German (AG) and Hungarian (H) corpora (cf. Tab. 1), covering read speech (RS) and conversational speech (CS) of different degrees of spontaneity: $CS^{+}$ for topic-free casual conversations and $CS^{-}$ for task-oriented/task-restricted conversations.

### 2.1. GRASS

The Graz Corpus of Read and Spontaneous Speech (GRASS) [15, 16] contains 6h of read (GRRS) and 19h of conversational

[1]Note that we further distinguish between more restricted conversational speech (e.g., telephone speech or task-oriented speech) and *casual face-to-face* conversations without any topical restrictions.



10.21437/Interspeech.2023-951

speech (GRCS) from 38 Austrian speakers (19f/19m). GRRS and GRCS are spoken by the same 38 speakers. For GRCS, 19 pairs of speakers who have known each other for several years, were recorded for one hour. Chosen topics were not restricted leading to casual speech (thus classified as $CS^+$ in Tab. 1) with characteristics such as frequently occurring overlapping speech, laughter, and dialectal pronunciation [16]. After the conversation, speakers separately read short stories as well as selected isolated sentences. For the experiments with GRCS, chunks with artefacts, noise, whispering, foreign words and dialect lexemes were excluded, resulting in a total deletion of approx. 4h, leaving approx. 13.5h for our experiments. Then, filler labels were unified. We noticed long silence parts at the beginning of all GRRS chunks which could distort this analysis due to higher amounts of codebook usage relating to silence parts. Hence, we cut out 1.3s of audio at the beginning of each file.

### 2.2. GECO

The GECO corpus [17] contains 46 spontaneous dialogues of approx. 25 minutes between female speakers. The corpus introduces two settings: 1) a unimodal setting with 22 dialogues (GECO-Mono), where participants were separated by a solid wall and 2) a multimodal setting with 24 dialogues (GECO-Multi), with face-to-face conversations comparable to GRCS. The unimodal setting involves 12 speakers, of whom 7 returned for the multimodal setting meaning that some dialogue pairs are present in GEMO and GEMU. In both settings, speakers were able to freely talk about any topic they want (thus classified as $CS^+$ in Tab. 1). For our experiments, GEMO and GEMU were preprocessed similar as GRCS and almost all chunks were kept.

### 2.3. KIEL

The Kiel Corpus of Spoken German (KIEL) [18] contains approx. 5h of read and spontaneous speech produced by speakers from Northern Germany. The read speech (KIRS) contains sentences and stories from 53 speakers (26f/27m). The corpus has two spontaneous components: First, the "appointment-making-scenario" (KIVM), which contains approx. 4h of dialogues from 43 speakers (22f/21m) who were making appointments. In this scenario, speech was only recorded if participants were holding a button pressed which was also blocking the interlocutor's channel. Second, the "video-task-scenario" (KIVT) contains approx. 1h of dyadic conversations. In this scenario, manipulated video materials from a television series were presented separately to two subjects with the task to find the differences in the video they saw. As KIVM and KIVT contain task-oriented/topic-restricted dialogues, we classify them as $CS^-$ in Tab. 1. As for GRCS, also for KIVM and KIVT chunks with laughed speech and noise were excluded and filler annotations were unified. For KIRS, depending on given transcription material, we utilized already trimmed audio-files directly or trimmed the audio-files on the basis of the boundary markers of given annotations. In case of all GECO and KIEL corpus components we excluded resulting chunks with durations greater than 20s due to our limited computational infrastructure.

### 2.4. BEA

The original BEA ("BEszélt nyelvi Adatbázis" in Hungarian, meaning spoken language database) aimed at collecting studio quality speech data from 500 speakers, representative in age, sex, dialect, and educational background, primarily for linguistic research purposes [19]. For the experiments, we used the BEA-Base subset [12] of the database, specifically the read *Readtext* (BERS) and the conversational *Discourse* (BECS) modules of the "train-114" subset. Both, BERS and BECS included the same speakers while female and male participants were closely balanced. In case of BECS, each conversation was recorded approx. 45min and one experimenter guided the casual conversations between the speaker and an optional discourse partner on various random topics. The recordings were made in the same studio environment and were cleaned from ambiguous and parallel parts, similarly to the previous databases. The recordings containing the voices of the experiment leader or of a 3rd person were excluded from the investigations. Hence, conversations from BECS included recordings which relate to only one speaker which makes it possible to compare specific speakers between BECS and BERS but, different from GRCS, it is impossible to compare one speaker pair from BECS with respective speakers from BERS.

Table 1: *Overview of used data sets: Hungarian (H), German (G) and Austrian German (AG) corpora, containing read (RS) and conversational speech of different degrees of spontaneity (i.e., $CS^+$ for casual face-to-face conversations and $CS^-$ for task-oriented/task-restricted conversations).*

| Corpus | Abbr. | Style | Variety/ Lang. | Hours |
|---|---|---|---|---|
| BEA Discourse | BECS | $CS^-$ | H | 14.2 |
| BEA Readtext | BERS | RS | H | 3.8 |
| GECO-Multi | GEMU | $CS^+$ | G | 9.8 |
| GECO-Mono | GEMO | $CS^-$ | G | 8.92 |
| GRASS CS | GRCS | $CS^+$ | AG | 13.5 |
| GRASS RS | GRRS | RS | AG | 4.6 |
| KIEL-Verbmobil | KIVM | $CS^-$ | G | 3.72 |
| KIEL-Videotask | KIVT | $CS^-$ | G | 1.3 |
| KIEL RS | KIRS | RS | G | 2.8 |

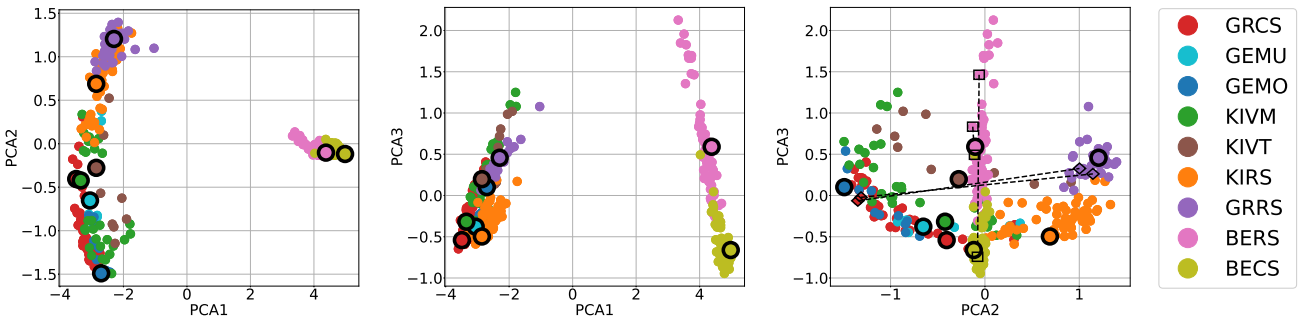
## 3. Analysis of self-supervised speech representations

We hypothesize that shared discrete speech representations of different corpora encode speaking styles and varieties. Here, we investigate this hypothesis by analyzing similarity matrices resulting from a comparison of normalized frequency usage of discrete XLSR speech representations (introduced by codebooks) from different data sets. The source code related to our analysis is publicly available and can be accessed on GitLab[2].

### 3.1. From similarity matrix to PCA space

We used wav2vec2 [9] with fairseq [20] to compute discrete shared speech representations with a multilingual pre-trained model (XLSR) [21]. XLSR is pre-trained in self-supervision with 56000h of speech data coming from 53 languages including German but not Hungarian and comprising approx. 99% of read speech and 1% of spontaneous speech (BABEL). XLSR has 315M parameters containing 24 transformer blocks with model dimensions 1024, inner dimension 4096 and 16 attention heads. Given the pre-trained model, we computed latent speech representations and utilized the model's quantizer to obtain respective codebook indices of shared discrete representa-

[2]https://gitlab.tugraz.at/speech/speechcodebookanalysis




(a) *PCA1 and PCA2* (b) *PCA1 and PCA3* (c) *PCA2 and PCA3* (d) *Legend*

Figure 1: *Speaker-dependent codebook usage with respect to the considered German and Hungarian corpora in the* 3*-dimensional PCA space after transforming their similarity matrix which results from codebook frequency usage of XLSR. BECS (olive) and BERS (pink) as well as GRCS (red) and GRRS (purple) involve the same speakers and filled circles with black outlines indicate corpus centroids. Dashed line connections of black rectangles and diamonds in (c) illustrate distances between BECS and BERS referring to same speakers as well as GRCS and GRRS referring to same speakers of one GRCS conversation.*

tions. The quantizer is based on product quantization introducing $G = 2$ codebooks with each of them having $V = 320$ entries, resulting in a total number of 102400 possible codebook combinations.

In order to compare the frequency usage of speech representations coming from XLSR with respect to speakers, we quantized the utterances of each speaker of each preprocessed corpus and counted the utilized codebook entries. Then, we normalized each speaker's frequency usage with the total number of features per speaker, resulting in speaker-dependent prior distributions of codebook usage. Given these priors, we generated a similarity matrix by computing similarities of resulting distributions with a Jensen-Shannon divergence. Finally, the similarity matrix was transformed to a 3-dimensional PCA space.

### 3.2. Interpretation of three PCA dimensions

Fig. 1 shows 3 speaker-dependent scatter plots (PCA1/PCA2, PCA1/PCA3 and PCA2/PCA3) from the resulting 3-dimensional PCA-space. Speakers of each corpus are depicted in a different color. First thing we notice is that PCA1 describes language, where component values $> 0$ categorize Hungarian speech and component values $< 0$ (Austrian) German speech. Second thing we notice is that PCA2 separates the same GRASS speakers in different clusters based on speaking style. In general, we observe that PCA2 characterizes our degree of spontaneity within (Austrian) German where components $> 0$ visualize almost non-overlapping RS corpora. In the opposite direction, conversational components of higher spontaneity may overlap. Third thing we notice is that PCA3 distinguishes Hungarian speaking styles where components $> 0$ define Hungarian read speech and components $< 0$ Hungarian conversational speech.

### 3.3. Centroids and their distances

At first, we compare resulting Hungarian centroids and (Austrian) German centroids in the 3-dimensional PCA-space (see filled circles with black outlines) with respect to Euclidean distances. In case of Hungarian centroids, we measured a distance of 1.3 between BECS and BERS which is mainly described by PCA3. In order to gain more insights into how speech representations differ between BECS and BERS, we randomly selected two speakers within BECS and measured their Euclidean distance to BERS resulting in 2.47 and 0.35 (see black dashed lines between olive and pink diamonds in Fig. 1). In general, mean and standard deviation of distances between same Hungarian speakers were $1.3 \pm 0.7$. In case of (Austrian) German centroids, we compared the resulting centroid of GRCS with the other 6 German-speaking centroids. We observe the smallest Euclidean distance between GRCS and GEMO (0.46) followed by distances with KIVM (0.53), GEMU (0.58), KIVT (1.19) and KIRS (1.77). The highest distance was between GRCS and GRRS (2.3), which is to some extent surprising as these two corpora contain speech from the same speakers. In order to gain more insights into how speech representations differ between GRCS and GRRS, we measured the Euclidean distance between a speaker pair within GRCS and to GRRS. We find that their distance in GRCS is approx. 0.07, whereas distances between the same speaker in GRCS and GRRS are considerably higher, i.e., approx. 2.56 and 2.4 (see black dashed lines between red and purple rectangles in Fig. 1). In general, mean and standard deviation of distances between same Austrian German speakers were $2.3 \pm 0.4$. Overall, when comparing the distances of all 19 speaker pairs, we found no correlation between GRCS and GRRS (Pearson Correlation Coefficient: $r \approx 0.02$, $p \approx 0.93$). These results show that the speech representations are more sensitive to the speech characteristics typical for read vs. conversational speech than to speaker specific characteristics. Finally, we compared the resulting centroid of BECS with German-speaking centroids and resulting centroid of GRCS with Hungarian-speaking centroids. We observe high Euclidean distances $> 7.1$ between BECS and the 6 (Austrian) German centroids with the smallest distance to KIVT (7.18) and the highest distance to GRCS (7.96). Overall, distances between GRCS and Hungarian centroids were $> 7.4$ since distance to BERS was 7.46.

### 3.4. Clustering of the 3-dimensional PCA space

Next, we performed k-Means clustering by using the resulting 3-dimensional PCA space with 6 clusters. This clustering enables classification by evaluating Euclidean distances to the 6 generated cluster centroids. We only measured the 2-dimensional distances with respect to projections in

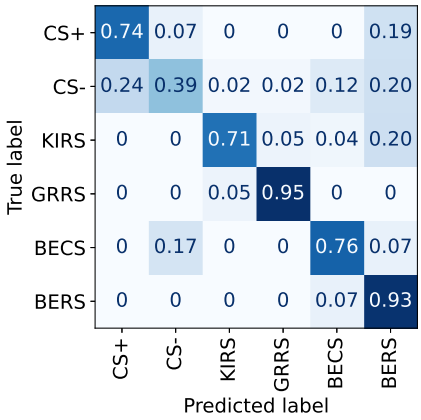



Figure 2: *Resulting confusion matrix when clustering the* 3*-dimensional PCA space of the speaker-dependent similarity matrix (see Fig. 1) with k-Means introducing* 6 *centroids.*

PCA2/PCA3, because those dimensions describe (Austrian) German (PCA2) and Hungarian (PCA3) speaking styles which is the focus of this study. Fig. 2 shows the resulting confusion matrix. The clusters correlate with the degree of (Austrian) German spontaneity ($CS^+$ and $CS^-$), correlate for (Austrian) German read speech with variety (GRRS and KIRS) and for Hungarian speech with speaking style (BERS and BECS). Interestingly, for German, clustering did not separate variety, but only the degree of spontaneity.

With respect to the confusions that occur, nearly all speakers from both (Austrian) German RS corpora were assigned correctly (KIRS: approx. 70%; GRRS: 90%), whereas only approx. 40% of speakers from $CS^-$ corpora were correctly assigned as $CS^-$, while approx. 20% of them were confused with $CS^+$, 2% of them were confused with KIRS and GRRS, 10% of them were confused with BECS and 20% of them were confused with BERS. In general, confusions of $CS^+$, $CS^-$ and KIRS with BEA (approx. 20% in case of BERS) can be explained by our analysis approach which compares only distances within the dimensions PCA2 and PCA3[3]. Likewise, assigning speakers from speaking style $CS^+$ was easier in general leading to a confusion with $CS^-$ of only approx. 7%. F1-scores of $CS^+$ and $CS^-$ were 0.77 and 0.45. In case of BECS approx. 80% of the speakers were correctly assigned, while approx. 20% of them were confused with $CS^-$ and approx. 7% of them were confused with BERS. Likewise, in case of BERS approx. 80% of the speakers were correctly assigned, while approx. only 7% of them were confused with BECS. Corresponding F1-scores of BECS and BERS were 0.78 and 0.74. These clustering results are in line with our earlier observation, as there is no confusion between GRCS ($CS^+$) and GRRS. Simultaneously, confusions between Hungarian speaking styles, namely BECS and BERS, were also small.

## 4. General Discussion and Conclusion

The main aim of this paper was to test the hypothesis that shared discrete speech representations from speakers of different corpora encode languages, varieties and speaking styles. To analyze this hypothesis, we performed a clustering experiment with XLSR codebook entries from the different data sets, demonstrating that, in addition to languages, read and spontaneous speaking styles are indeed also distinguished in this feature space. Based on a 3-dimensional PCA space, independent of language (PCA1) almost all speakers from the read speech corpora were assigned correctly to the corresponding clusters, for the spontaneous corpora, however, this was only the case with $CS^+$ and BECS with corresponding F1-scores of 0.77 and 0.78. We observed that speech representations of German spontaneous speaking style showed variety-independence, which we explain by the strongly varying speech representation usage. For read speech, we can distinguish between the German and Austrian German variety. In general, our findings are in line with those in the literature: The study by Conneau et al. [10] used similar methods to cluster discrete speech representations of multilingual pretrained wav2vec2 models, demonstrating the possibility of grouping related languages. Another study on dialect clustering with sentence vector representations based on character-based metrics also generated plausible clusters [22]. They found three emerging noticeable clusters in case of Japanese varieties, namely Tohoku dialect, Tokyo dialect and a combination of three Western dialects (Kansai, Chugoku and Kyushu).

Another focus of our analysis was on how the speech representations of the same speakers behave and whether they explain different degrees of spontaneity. We found that Austrian German speakers differ the most between different styles since mean distance of same Austrian German speakers was high (2.3). In contrast, mean distance of same Hungarian speakers was smaller (1.3). Furthermore, we found that Austrian German speakers also differ more from themselves within different styles, indicating speaker identity independence of the speech representations. Overall, our results indicate that speech representations vary the most among Austrian German speakers. Also Asami et al. [23] found that GMM supervectors based on utterances can discriminate read and spontaneous speech with less speaker-dependency. Simultaneously, the authors state that clustering spontaneous utterances is more difficult than read utterances.

To conclude, the results suggest that distance calculation based on shared quantized latent speech representations is also meaningful on a much finer granularity level (i.e., per speaker per speaking style) than it was introduced in [10] for languages. This may open new perspectives in speech data selection both for supervised and self-supervised learning, as speech sections matching the desired development set (or speaking style) could be collected at a relative low cost, requiring only a pre-trained wav2vec2 model but without the need of any additional information beyond the waveform. Furthermore, it may be worth exploring meaningful acoustic correlates that could shed more light on the nature of elusive self-supervised speech representations. We are going to extend our investigations in these directions in the future.

## 5. Acknowledgements

This work was partly funded by grant P-32700-NB from FWF and by grants K143075 and K135038 from NRDI.

[3]Note that we could easily implement a condition on PCA1 if the aim of our study would be a detection task

# 6. References


[1] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[2] Y.-A. Chung, Y. Zhang, W. Han, C.-C. Chiu, J. Qin, R. Pang, and Y. Wu, "W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," *ArXiv: abs/2108.06209*, 2021.

[3] J. Godfrey, E. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 1992, pp. 517–520 vol.1.

[4] Z. Tüske, G. Saon, and B. Kingsbury, "On the Limit of English Conversational Speech Recognition," in *Proc. Interspeech 2021*, 2021, pp. 2062–2066.

[5] J. Linke, P. N. Garner, G. Kubin, and B. Schuppler, "Conversational Speech Recognition Needs Data? Experiments with Austrian German," in *LREC*, 2022.

[6] J. Kim and P. Kang, "K-wav2vec 2.0: Automatic speech recognition based on joint decoding of graphemes and syllables," *ArXiv: abs/2110.05172*, 2021.

[7] S. Furui, M. Nakamura, T. Ichiba, and K. Iwano, "Why is the recognition of spontaneous speech so hard?" in *Text, Speech and Dialogue*, 2005, pp. 9–22.

[8] S. Furui, "Generalization problem in ASR acoustic model training and adaptation," in *2009 IEEE Workshop on Automatic Speech Recognition Understanding*, 2009, pp. 1–10.

[9] A. Baevski, 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, 2020, pp. 12 449–12 460.

[10] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised Cross-Lingual Representation Learning for Speech Recognition," in *Proc. Interspeech 2021*, 2021, pp. 2426–2430.

[11] A. Khosravani, P. N. Garner, and A. Lazaridis, "Modeling Dialectal Variation for Swiss German Automatic Speech Recognition," in *Proc. Interspeech 2021*, 2021, pp. 2896–2900.

[12] P. Mihajlik, A. Balog, T. E. Graczi, A. Kohari, B. Tarján, and K. Mady, "BEA-base: A benchmark for ASR of spontaneous Hungarian," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, Jun. 2022, pp. 1970–1977. [Online]. Available: https://aclanthology.org/2022.lrec-1.211

[13] H. Mary, "IARPA Babel Program." [Online]. Available: https://www.iarpa.gov/research-programs/babel

[14] M. Gales, K. Knill, A. Ragni, and S. Rath, "Speech recognition and keyword spotting for low-resource languages: Babel project research at cued," *Fourth International Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU-2014)*, pp. 16–23, May 2014. [Online]. Available: https://eprints.whiterose.ac.uk/152840/

[15] B. Schuppler, M. Hagmüller, J. A. Morales-Cordovilla, and H. Pessentheiner, "GRASS: the Graz corpus of Read And Spontaneous Speech," in *LREC*, 2014, pp. 1465–1470.

[16] B. Schuppler, M. Hagmüller, and A. Zahrer, "A corpus of read and conversational Austrian German," *Speech Communication*, vol. 94, pp. 62–74, 2017.

[17] A. Schweitzer, N. Lewandowski, D. Duran, and G. Dogil, "Attention, please! Expanding the GECO database," in *ICPhS*, 2015.

[18] K. J. Kohler, B. Peters, and M. Scheffers, "The Kiel Corpus of Spoken German—Read and Spontaneous Speech. New Edition, revised and enlarged," 2017. [Online]. Available: http://www.isfas.uni-kiel.de/de/linguistik/forschung/kiel-corpus/

[19] T. Neuberger, D. Gyarmathy, T. E. Gráczi, V. Horváth, M. Gósy, and A. Beke, "Development of a large spontaneous speech database of agglutinative Hungarian language," in *Text, Speech and Dialogue*, P. Sojka, A. Horák, I. Kopeček, and K. Pala, Eds. Cham: Springer International Publishing, 2014, pp. 424–431.

[20] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, Jun. 2019, pp. 48–53. [Online]. Available: https://aclanthology.org/N19-4009

[21] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve, and M. Auli, "Robust wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training," in *Proc. Interspeech 2021*, 2021, pp. 721–725.

[22] Y. Sato and K. Heffernan, "Dialect clustering with character-based metrics: in search of the boundary of language and dialect," in *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 985–990. [Online]. Available: https://aclanthology.org/2020.lrec-1.124

[23] T. Asami, R. Masumura, H. Masataki, and S. Sakauchi, "Read and spontaneous speech classification based on variance of GMM supervectors," in *Proc. Interspeech 2014*, 2014, pp. 2375–2379.