Toward Conversational Hungarian Speech Recognition: Introducing the BEA-Large and BEA-Dialogue Datasets

Máté Gedeon[*,†], **Piroska Zsófia Barta**[*,†], **Péter Mihajlik**[*,‡], **Tekla Etelka Gráczi**[‡], **Anna Kohári**[‡], **Katalin Mády**[‡]

[*]Department of Telecommunications and Artificial Intelligence,
Faculty of Electrical Engineering and Informatics
Budapest University of Technology and Economics, Hungary
[‡]ELTE Research Centre for Linguistics, Hungary
[†]SpeechTex Ltd.

{gedeonm, piroskazsofia.barta}@edu.bme.hu
{mihajlik.peter, graczi.tekla.etelka, kohari.anna, mady}@nytud.hu

Abstract

The advancement of automatic speech recognition (ASR) has been largely enhanced by extensive datasets in high-resource languages, while languages such as Hungarian remain underrepresented due to limited spontaneous and conversational corpora. To address this gap, we introduce two new datasets – BEA-Large and BEA-Dialogue – constructed from the previously unprocessed portions of the Hungarian speech corpus named BEA. BEA-Large extends BEA-Base with 255 hours of spontaneous speech from 433 speakers, enriched with detailed segment-level metadata. BEA-Dialogue, comprising 85 hours of spontaneous conversations, is a Hungarian speech corpus featuring natural dialogues partitioned into speaker-independent subsets, supporting research in conversational ASR and speaker diarization. We establish reproducible baselines on these datasets using publicly available ASR models, with the fine-tuned Fast Conformer model achieving word error rates as low as 14.18% on spontaneous and 4.8% on repeated speech. Diarization experiments yield diarization error rates between 12.46% and 17.40%, providing reference points for future improvements. The results highlight the persistent difficulty of conversational ASR, particularly due to disfluencies, overlaps, and informal speech patterns. By releasing these datasets and baselines, we aim to advance Hungarian speech technology and offer a methodological framework for developing spontaneous and conversational benchmarks in other languages.

Keywords: speech database, automatic speech recognition, spontaneous speech, evaluation

1. Introduction

The field of automatic speech recognition (ASR) has been fundamentally transformed by the availability of large-scale training datasets (Panayotov et al., 2015; Pratap et al., 2020). However, this revolution has predominantly benefited well-resourced languages, leaving low-resource languages like Hungarian significantly underrepresented in the modern ASR ecosystem (Besacier et al., 2014). While recent advances in self-supervised learning and multilingual models have shown promise for low-resource scenarios, the fundamental challenge remains: the scarcity of high-quality, diverse speech corpora that capture the full spectrum of natural human communication (Mihajlik et al., 2024).

Hungarian, with its morphologically rich and agglutinative nature, poses a significant challenge even for large multilingual models. To improve the results of these models via fine-tuning, or by training our own, having high-quality spontaneous speech as well as conversational datasets is vital (Mihajlik et al., 2010). BEA (Neuberger et al., 2014) is a large Hungarian speech database with the aim, among others, to provide material for research purposes in various fields. The complete 300-hour dataset consists of eight types of speech sessions, including repeated and spontaneous speech, from almost 500 speakers of varying age, gender, and educational background.

Although the BEA Base dataset (Mihajlik et al., 2022), comprising 140 speakers, has been released as a benchmark for ASR of spontaneous Hungarian, a considerable part of the BEA dataset has not yet been processed and subjected to benchmarking.

In this work, we leverage the remaining BEA recordings to create two new datasets and provide comprehensive baselines for Hungarian conversational ASR. The first dataset, BEA-Large, extends and refines BEA-Base by including additional training data and more extensive segment-level metadata. BEA-Large thus offers an extended corpus of spontaneous Hungarian, with fine-grained annotations to support research in ASR and related areas, while BEA-Dialogue contains dialogues mainly from the conversations that are also used for the construction of BEA-Large. Alongside the aforementioned datasets, we provide baseline ASR results for both, using publicly available models to ensure

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1562–1570
11–16 May 2026.

reproducibility. These resources not only advance Hungarian speech technology but also provide a methodological framework for similar efforts in other low-resource languages.

Our primary contributions are:

- An extended dataset (BEA-Large) comprising 255 hours of speech from 433 speakers, substantially expanding the available Hungarian spontaneous speech training data with enriched metadata
- A specialized conversational speech dataset (BEA-Dialogue) featuring 85 hours of natural dialogues, addressing the critical shortage of Hungarian dialogue data for conversational ASR and speaker diarization research
- Systematic benchmarking using publicly available ASR models to ensure reproducibility and establish performance baselines

The structure of the paper is as follows. Section 2 reviews related work. Sections 3 and 4 describe the construction of the two datasets, followed by ASR baseline results in Sections 5 and 6, and diarization results on BEA-Dialogue in Section 7. Finally, Section 8 summarizes the findings and concludes the paper.

2. Related work

Concerning the international situation of spontaneous speech datasets, there are several resources for a range of languages and domains. However, a significant portion of these corpora remains unavailable for public use, especially for low-resource languages. Among the publicly accessible English datasets, SSSD (Sheikh et al., 2025) is designed for dialogue research and features 727 hours of spontaneous English conversations between randomly matched speaker pairs. CASPER (Xiao et al., 2025) provides over 100 hours of unscripted English dialogues.

Beyond English, the GRASS corpus (Schuppler et al., 2014)—the only existing resource for Austrian German for years (Linke et al., 2022)—which contains approximately 19 hours of dyadic speech; and ES-Port (García-Sardiña et al., 2018), comprising 40 hours of Spanish dialogues from technical support calls. The RAMC corpus (Yang et al., 2022) encompasses 180 hours of conversational speech recorded via telephone channels. The Verbmobil dialogue corpus (Weilhammer et al., 2002) covers multilingual data in German, English, and Japanese. The Kiel Corpus of Spoken German (Kohler et al., 2017) includes over 8 hours of spontaneous speech from 64 speakers.

For specialized domains, RescueSpeech (Sagar et al., 2023) focuses on the search and rescue context, containing approximately 2 hours of annotated German speech from simulated exercises. The SXUCorpus (Herms et al., 2016) offers spontaneous speech components in the Upper Saxon German dialect. The Portuguese CORAA corpus (Junior et al., 2021) comprises at least 189 hours of spontaneous speech from conversations, monologues, dialogues, and interviews. The CSJ corpus (Uchimoto et al., 2007), while primarily consisting of monologues, also features spontaneous speech. The Norwegian Parliamentary Speech Corpus (Solberg and Ortiz, 2022) includes unscripted parliamentary sessions, while the Korean Corpus of Spontaneous Speech (Yun et al., 2015) contains recordings from interviews with 40 speakers.

As discussed by Mihajlik et al. (2024), despite the relatively large cumulative volume of Hungarian speech data, inconsistencies in annotation practices and restricted accessibility hinder their simultaneous use—both for general-purpose ASR and for specialized tasks such as conversational ASR. Among the monolingual Hungarian resources, BUSZI (∼600h, 250 spk; (Kontra et al., 1997)) and SzöSzi (370h, 163 spk; (Kontra et al., 2016)) stand out as extensive sociolinguistic interview collections – unfortunately, their access is strongly limited. HuComTech (∼50h, 112 spk; (Pápay et al., 2011)) provides valuable audiovisual recordings for multimodal communication studies, while BEKK (20h, 56 spk; (Bodó et al., 2017)) contributes spontaneous, student-recorded conversations.

3. Construction of BEA-Large

The *BEA-Large* dataset is built upon the original *BEA-Base* corpus (Mihajlik et al., 2022) by extending the training data with an additional subset, *BEA-Extension* (referred to as *train-293*), where each segment corresponds to a single utterance of a given speaker. The *dev* and *eval* partitions remain identical to those in BEA-Base to preserve consistency and comparability across experiments.

	train-114	**train-293**
# speakers [f|m]	69 | 45	184 | 109
# segments	69,176	196,981
# words	555,322	1,622,151
# chars	3,310,493	9,550,276
duration [h]	67.95	177.4

Table 1: Metadata for the disjoint training sets of BEA-Large.

The additional training set re-introduces metadata attributes that were not available in BEA-Base, including speaker *age*, *gender*, *occupation*, and the

module indicating the communicative setting (e.g., summary, interview or discourse (Gósy, 2012)). In addition to the unique identifier of the target speaker associated with each recording, the dataset now explicitly labels each speaker's role according to the taxonomy outlined in Mády et al. (2024): *SPK* (target speaker), *EXP* (experiment leader), or *DP* (discourse partner). While SPK and EXP are represented in all modules, DP contributes only to the discourse module.

The *train-293* subset was compiled from recordings of 293 target speakers who do not appear in BEA-Base, ensuring that both training sets can be seamlessly combined without overlap. As summarized in Table 2, data were drawn from multiple modules detailed by Mihajlik et al. (2022), with the *repeat* and *readsent* modules – containing fixed sentences repeated or read by the SPK or EXP – excluded to maintain structural alignment with the train set of BEA-Base (*train-114*). Acoustic preprocessing followed the same protocol as in BEA-Base.

Module	SPK [%]	EXP [%]	DP [%]
repeat	–	–	–
readsent	–	–	–
interview	16.45	3.53	–
opinion	12.86	5.56	–
summhist	3.80	0.36	–
summplant	3.26	0.46	–
discourse	21.73	18.02	8.39
readtext	5.46	0.14	–

Table 2: Composition of the train-293 set (in percentage of total duration).

When combined with *train-114*, the resulting BEA-Large training corpus contains approximately three times as much data as BEA-Base. The duration of the segments in BEA-Large is shown in Figure 2, with approximately 99.6% of the segments in the database have a duration of less than 20 seconds. Figure 1 illustrates the age distribution of speakers in the BEA-Large training set compared to that of *train-114*.

4. Construction of BEA-Dialogue

A dedicated dialogue corpus, *BEA-Dialogue*, was derived from the recordings of 242 speakers that had not previously been included in BEA-Base.

In the original dataset, utterances were organized by the target speaker, with each module stored in a `TextGrid` file containing time-aligned transcriptions for all speakers appearing in that module. To create BEA-Dialogue, utterances were first extracted along with their timestamps and speaker labels (*SPK*, *EXP*, *DP*).

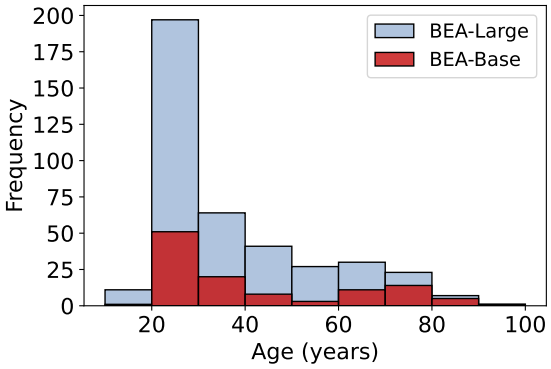

Figure 1: Comparison of the age distributions in the train sets of BEA-Large and Base.

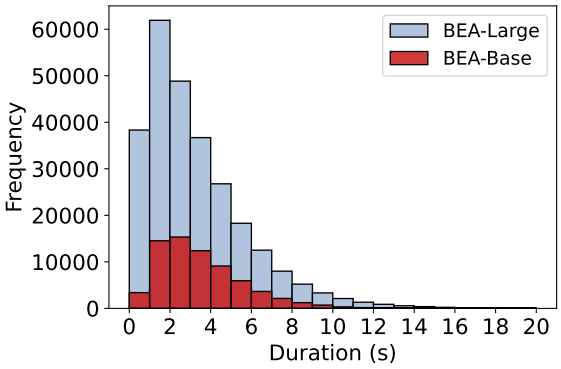

Figure 2: Duration of the segments in the train sets of BEA-Large and Base.

Candidate cut points for shorter dialogue segments were then identified by detecting silent intervals—regions that did not overlap with any utterance, except possibly at their boundaries. Using these silence-based boundaries, utterances from multiple speakers were grouped into coherent dialogue units according to the silent regions separating them. These smaller units were subsequently merged into larger dialogue segments, each with a target duration of 30 seconds. Figure 3 shows the histogram of the durations after the procedure.

Beyond the target speakers, the dataset includes the voices of five additional female speakers (*fem1–fem5*) and two male speakers (*male1–male2*), who served as experiment leaders and discourse partners. To construct *BEA-Dialogue*, which provides fully disjoint training, development, and evaluation sets across all speaker roles, the data were partitioned based on three experiment leaders: *fem1*, *fem3*, and *fem4*.

Since these experiment leaders also acted as discourse partners for different target speakers, we excluded dialogue segments from the discourse module in which the target speaker's discourse partner was one of the three experiment leaders.

Model	Dataset	dev-repet		eval-repet		dev-spont		eval-spont	
		WER	CER	WER	CER	WER	CER	WER	CER
`whisper-large-v3`	*zero-shot*	***13.31***	***2.74***	***13.73***	***2.83***	***20.79***	***9.22***	***21.28***	***9.31***
`whisper-large-v2`	*zero-shot*	*18.56*	*4.47*	*21.26*	*4.46*	*34.71*	*19.67*	*32.56*	*18.61*
`whisper-medium`	*zero-shot*	*22.7*	*5.27*	*23.84*	*6.10*	*37.23*	*20.13*	*39.19*	*21.71*
`f_conformer_ctc`	train-114	7.01	1.54	7.05	1.82	18.26	6.04	19.22	6.37
	train-293	14.60	2.83	14.98	3.30	15.60	5.18	16.33	5.34
	train-114 + train-293	**5.21**	**1.16**	**4.80**	**1.20**	**13.52**	**4.41**	**14.18**	**4.56**

Table 3: WER and CER [%] for zero-shot and fine-tuned models on BEA-Large under different configurations.

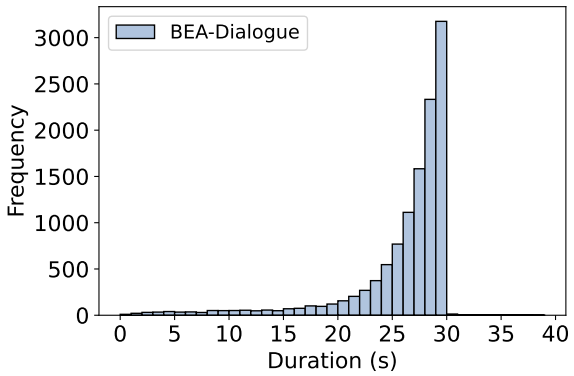

Figure 3: Duration of the segments in BEA-Dialogue.

The combinations of experiment leaders and discourse partners across all available data are shown in Table 4, along with the corresponding number of dialogue segments belonging to each pair.

		EXP		
		fem1	**fem3**	**fem4**
DP	**fem1**	–	286	852
	fem2	1751	–	122
	fem3	4415	–	546
	fem4	4418	–	–
	fem5	1662	367	816
	male1	983	88	588
	male2	46	–	–

Table 4: Number of discourse segments by experiment leader (EXP) and discourse partner (DP) across all dialogue segments.

The training set includes dialogue segments where the experiment leader is *fem1* and the discourse partner is either *fem2* or *male2*, as well as all segments from other modules featuring *fem1*. The development and evaluation sets contain dialogues where the experiment leaders are *fem3* and *fem4*, with discourse partners *fem5* and *male1*, respectively. Table 5 summarizes the metadata of BEA-Dialogue, which represents the largest set that can be derived from the data without violating speaker independence across subsets with respect to the three speaker roles (SPK, EXP, DP).

	BEA-Dialogue		
	Train	**Dev**	**Eval**
# Speakers [f|m]	121 | 67	3 | 6	29 | 16
# Segments	9,179	577	1,906
# Words	532,732	34,056	105,472
# Characters	3,217,617	206,740	641,628
Avg. # Speakers / Seg	1.77	1.92	1.61
Avg. # Utterances / Seg	10.99	8.68	9.74
Avg. Seg Duration [s]	26.23	26.31	26.09
Overlap Duration [h]	3.28	0.29	0.40
Total Duration [h]	66.87	4.22	13.81

Table 5: Metadata of the BEA-Dialogue dataset.

5. Baseline ASR results for BEA-Large

To establish reproducible baselines, we fine-tuned a potent yet relatively small (120M parameters) model publicly available in the NeMo toolkit (Kuchaiev et al., 2019) – `STT En Fast Conformer-CTC Large` [1] (`f_conformer_ctc`) – on three different training sets: *train-114 (BEA-Base)*, *train-293 (BEA-Extension)* and *train-114 + train-293 (BEA-Large)*. Performance was evaluated on subsets of BEA-Base – dev-repet and eval-repet containing repeated or read sentences, and dev-spont and eval-spont containing spontaneous speech – using two standard metrics: Word Error Rate (WER) and Character Error Rate (CER). The results are summarized in Table 3. Zero-shot benchmarks obtained using the Speechbrain toolkit (Ravanelli et al., 2021, 2024) with three Whisper models[2] — `whisper-large-v3`, `whisper-large-v2`, and `whisper-medium` (Radford et al., 2022) — are also presented in Table 3.

[1] https://huggingface.co/nvidia/stt_en_fastconformer_ctc_large

[2] https://github.com/openai/whisper

The Fast Conformer model fine-tuned on less than 250 hours of Hungarian speech consistently outperformed all Whisper models, across both the repetitive and spontaneous subsets of BEA-Large.

In both training partitions of BEA-Large, the voices of the seven individuals acting as experiment leaders and discourse partners are notably overrepresented. To obtain two speaker-independent subsets, we extracted utterances from the two training sets belonging exclusively to the target speakers. In addition to the training setups outlined in Table 3, we constructed setups from the two speaker-independent supplementary training sets: using each dataset individually, one at a time, and using both datasets merged. Notably, the WER and CER increased when the training sets were limited to the utterances of the target speakers.

Training exclusively on train-293 resulted in higher WER and CER on both dev-repet and eval-repet compared to models trained on train-114. This performance degradation can likely be attributed to differences in lexical overlap between the training and evaluation sets. Specifically, train-114 exhibits greater lexical overlap with the development sets, which is particularly consequential given the repetitive nature of the target data.

The best results were achieved by using the two full components – train-114 and train-293 – of the database yielding a WER of 14.18% and a CER of 4.56% on the most important spontaneous evaluation subset, indicating that using the voice of the experiment leaders and discourse partners does not have a negative impact on the evaluation results obtained on a speaker-independent evaluation set.

6. Baseline ASR results for BEA-Dialogue

For the *BEA-Dialogue* dataset, we trained a similar set of models to those used for *BEA-Large*, complemented with fine-tuned Fast Conformer models. During training, we employed Serialized Output Training (SOT) (Kanda et al., 2020), marking speaker transitions explicitly with a `<sc>` (speaker change) token. These tokens were inserted to reflect the structure of an ideal dialogue transcript. For example (English translations in parentheses):

> *Hogy vagy?* `<sc>` *Köszönöm, jól. És te?* `<sc>` *Én is, köszönöm.* `<sc>` *Örülök neki.* (*How are you?* `<sc>` *I'm good, thanks. And you?* `<sc>` *Me too, thanks.* `<sc>` *I'm glad.*)

The utterance of each speaker is maintained as an uninterrupted sequence, even when there is overlap in speech. This allows the model to learn to recognize speaker boundaries while preserving the linguistic integrity of each turn.

Evaluation followed the same criteria as in BEA-Large: *WER* and *CER*, complemented by the *concatenated minimum-permutation WER* (cpWER) and *cpCER*, which minimize errors across all possible permutations of speakers separated by `<sc>` tokens. Since some segments contained more than ten speaker changes, exhaustive permutation computation was infeasible. To address this, we adopted a hybrid strategy: computing all permutations for up to seven speaker changes and applying beam search to achieve near-optimal results for higher counts. For fine-tuned models, we additionally report the *speaker change accuracy* (*scAcc*) – the proportion of utterances in which the predicted number of speaker transitions matches the reference (i.e., occurrences of `<sc>` tokens).

Table 6 summarizes the obtained results.

As shown in Table 6, the fine-tuned `f_conformer_ctc` model achieves lower error rates on the *eval* set, demonstrating the benefits of domain-specific adaptation. Interestingly, however, on the *dev* set, the `Whisper-large-v2` model outperforms both its `v3` variant and the fine-tuned model. Nevertheless, the fact that an English model fine-tuned on fewer than 70 hours of Hungarian data can surpass state-of-the-art systems underscores the importance of corpora such as BEA-Dialogue for advancing Hungarian conversational ASR. These findings highlight the current limitations of state-of-the-art models when applied to spontaneous, real-world speech.

The ratio of character error rate (CER) to word error rate (WER) is notably higher than typically observed in non-conversational datasets, suggesting that fine-grained, character-level deviations occur more frequently in spontaneous dialogue.

Figure 4 shows the distribution of the number of speaker turns per segment for each split. As shown, most segments contain no or only one speaker change, making the task largely comparable to single-speaker spontaneous speech recognition. However, the ones with a higher number of turns substantially increase task complexity.

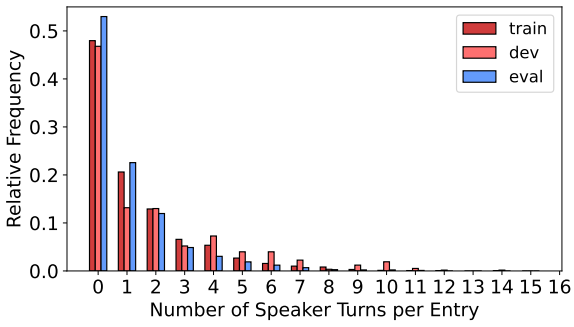

Figure 4: Distribution of speaker changes per segment in BEA-Dialogue.

Model	dev					eval				
	WER	cpWER	CER	cpCER	scAcc	WER	cpWER	CER	cpCER	scAcc
whisper-large-v3 (zs)	21.19	21.04	12.74	12.56	–	22.21	22.13	12.27	12.18	–
whisper-large-v2 (zs)	**19.65**	**19.42**	9.84	9.58	–	24.50	24.42	13.13	13.05	–
whisper-medium (zs)	25.45	25.27	12.61	12.42	–	29.21	29.12	14.71	14.63	–
f_conformer_ctc (ft)	19.69	19.53	**7.95**	**7.78**	69.32	**20.56**	**20.44**	**9.11**	**9.00**	82.16

Table 6: Comparison of ASR accuracy (%) on the BEA-Dialogue dataset for zero-shot (zs) and fine-tuned (ft) models.

7. Baseline diarization results for BEA-Dialogue

We establish baseline diarization performance for the BEA-Dialogue dataset using its *dev* and *eval* splits. The goal of these baselines is to provide reference performance levels for future work on speaker segmentation.

Two state-of-the-art diarization systems were evaluated: *pyannote.audio* (Bredin, 2023) and *Sortformer* (Park et al., 2024). Both models were used in their pre-trained configurations without additional fine-tuning. The *pyannote.audio* model represents a hybrid neural approach that combines segmentation and clustering stages, while *Sortformer* adopts a transformer-based end-to-end architecture with a sorting-based loss function to better handle overlapping speech. The exact pre-trained checkpoints used are: `speaker-diarization-3.1`[3] and `diar_sortformer_4spk-v1`[4].

Table 7 reports the average Diarization Error Rate (DER) obtained by each model on the two dataset subsets.

Model	DER (%)	
	dev	eval
Pyannote	14.09	17.40
Sortformer	**12.46**	**15.11**

Table 7: Average DER comparison on BEA-Dialogue.

Figure 5 illustrates the distribution of DER values across individual dialogue segments in the *eval* subset. Both systems exhibit similar overall trends.

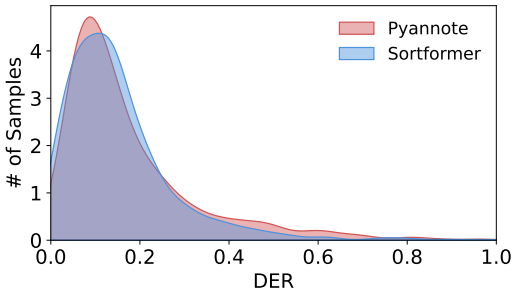

Figure 5: DER distribution on BEA-Dialogue (eval).

8. Conclusion

This work addresses the shortage of high-quality spontaneous and conversational speech data for Hungarian by introducing two substantial resources: BEA-Large and BEA-Dialogue. BEA-Large extends the original BEA-Base corpus with approximately 178 hours of additional speech from 293 speakers, enriched with segment-level metadata such as age, gender, occupation, and speaker role. This expansion nearly triples the amount of available training data while maintaining compatibility with the original benchmark splits.

BEA-Dialogue, comprising 85 hours of natural conversations, represents one of the largest Hungarian datasets explicitly designed for conversational ASR and speaker diarization research. The dataset includes predefined splits that ensure complete speaker independence across all conversational roles.

Our baseline experiments yield several key findings. Fine-tuned Fast Conformer models achieved WERs as low as 14.18% on spontaneous speech and 4.8% on repeated speech when trained on the combined datasets, highlighting the benefits of including data from both experiment leaders and discourse partners. In conversational ASR, the relatively high CER-to-WER ratio compared to monologue speech suggests that spontaneous dialogue introduces additional challenges—arising from phenomena such as overlapping speech, rapid turn-taking, and interaction-driven disfluency patterns that are not present in single-speaker settings. The speaker diarization baselines, with DERs ranging from 12.46% to 17.40%, establish solid reference points for future advancements.

By releasing these datasets together with reproducible baselines built on publicly available models, we aim to accelerate research in Hungarian

[3] https://huggingface.co/pyannote/speaker-diarization-3.1

[4] https://huggingface.co/nvidia/diar_sortformer_4spk-v1

speech technology and provide a methodological blueprint for similar initiatives in other low-resource languages. Future work will focus on advanced methods for overlap handling, speaker change detection, and context-aware modeling, with the goal of improving both ASR and diarization performance in conversational scenarios.

Acknowledgment

Project No. 2025-2.1.2-EKÖP-KDP-2025-00005 has been implemented with the support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, financed under the EKÖP_KDP-25-1-BME-21 funding scheme.

The work was also partially supported by the Hungarian NRDI Fund through the projects NKFIH K143075 and K135038, NKFIH-828- 2/2021(MILAB).

9. Bibliographical References

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100.

Hervé Bredin. 2023. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Interspeech 2023*, pages 1983–1987.

Naoyuki Kanda, Yashesh Gaur, Xiaofei Wang, Zhong Meng, and Takuya Yoshioka. 2020. Serialized output training for end-to-end overlapped speech recognition. In *Interspeech 2020*, pages 2797–2801.

Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Kriman, Stanislav Beliaev, Vitaly Lavrukhin, Jack Cook, Patrice Castonguay, Mariya Popova, Jocelyn Huang, and Jonathan M. Cohen. 2019. Nemo: a toolkit for building ai applications using neural modules.

Julian Linke, Philip N. Garner, Gernot Kubin, and Barbara Schuppler. 2022. Conversational speech recognition needs data? experiments with Austrian German. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4684–4691, Marseille, France. European Language Resources Association.

Katalin Mády, Anna Kohári, Tekla Etelka Gráczi, and P'eter Mihajlik. 2024. Revised annotation conventions in hungarian speech corpora. *BESZÉDTUDOMÁNY / SPEECH SCIENCE*, (1):185–202.

Peter Mihajlik, Katalin Mády, Anna Kohári, Fruzsina Sára Fruzsina, Gábor Kiss, Tekla Etelka Gráczi, and A. Seza Doğruöz. 2024. Is spoken Hungarian low-resource?: A quantitative survey of Hungarian speech data sets. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9382–9388, Torino, Italia. ELRA and ICCL.

Péter Mihajlik, Zoltán Tuske, Balázs Tarján, Bottyán Németh, and Tibor Fegyó. 2010. Improved recognition of spontaneous hungarian speech—morphological and acoustic modeling techniques for a less resourced task. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1588–1600.

Tae Jin Park, Ivan Medennikov, Kunal Dhawan, Weiqing Wang, He Huang, Nithin Rao Koluguri, Krishna C. Puvvada, Jagadeesh Balam, and Boris Ginsburg. 2024. Sortformer: A novel approach for permutation-resolved speaker supervision in speech-to-text systems.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Mirco Ravanelli, Titouan Parcollet, Adel Moumen, Sylvain de Langen, Cem Subakan, Peter Plantinga, Yingzhi Wang, Pooneh Mousavi, Luca Della Libera, Artem Ploujnikov, Francesco Paissan, Davide Borra, Salah Zaiem, Zeyu Zhao, Shucong Zhang, Georgios Karakasidis, Sung-Lin Yeh, Pierre Champion, Aku Rouhe, Rudolf Braun, Florian Mai, Juan Zuluaga-Gomez, Seyed Mahed Mousavi, Andreas Nautsch, Ha Nguyen, Xuechen Liu, Sangeet Sagar, Jarod Duret, Salima Mdhaffar, Gaëlle Laperrière, Mickael Rouvier, Renato De Mori, and Yannick Estève. 2024. Open-source conversational ai with speechbrain 1.0. *Journal of Machine Learning Research*, 25(333).

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. 2021. SpeechBrain: A general-purpose speech toolkit. ArXiv:2106.04624.

10. Language Resource References

C. Bodó, Z. Kocsis, and F. S. Vargha. 2017. A budapesti egyetemi kollégiumi korpusz: Elméleti és módszertani kérdések. In *Élőnyelvi kutatások és a dialektológia*, pages 169–177.

Laura García-Sardiña, Manex Serras, and Arantza del Pozo. 2018. ES-port: a spontaneous spoken human-human technical support corpus for dialogue research in Spanish. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Mária Gósy. 2012. Bea – a multifunctional hungarian spoken language data base. *The Phonetician*, pages 50–60.

Robert Herms, Laura Seelig, Stefanie Münch, and Maximilian Eibl. 2016. A corpus of read and spontaneous Upper Saxon German speech for ASR evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4648–4651, Portorož, Slovenia. European Language Resources Association (ELRA).

Arnaldo Candido Junior, Edresson Casanova, Anderson Soares, Frederico Santos de Oliveira, Lucas Oliveira, Ricardo Corso Fernandes Junior, Daniel Peixoto Pinto da Silva, Fernando Gorgulho Fayet, Bruno Baldissera Carlotto, Lucas Rafael Stefanel Gris, and Sandra Maria Aluísio. 2021. Coraa: a large corpus of spontaneous and prepared speech manually validated for speech recognition in brazilian portuguese.

Klaus J. Kohler, Benno Peters, and Michel Scheffers. 2017. The kiel corpus of spoken german – read and spontaneous speech. new edition, revised and enlarged. Accessed: 2025-10-12.

M. Kontra, T. Váradi, and Magyar Tudományos Akadémia. Nyelvtudományi Intézet. 1997. *The Budapest Sociolinguistic Interview: Version 3*. Working papers in Hungarian sociolinguistics. Linguistics Institute, Hungarian Academy of Sciences.

Miklós Kontra, Miklós Németh, and Balázs Sinkovics. 2016. *Szeged nyelve a 21. század elején*. Gondolat Kiadó, Budapest, Magyarország. Scientific monograph.

Peter Mihajlik, Andras Balog, Tekla Etelka Graczi, Anna Kohari, Balázs Tarján, and Katalin Mady. 2022. BEA-base: A benchmark for ASR of spontaneous Hungarian. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1970–1977, Marseille, France. European Language Resources Association.

Tilda Neuberger, Dorottya Gyarmathy, Tekla Gráczi, Viktória Horváth, Mária Gósy, and András Beke. 2014. Development of a large spontaneous speech database of agglutinative hungarian language. volume 8655.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. Mls: A large-scale multilingual dataset for speech research. In *Interspeech 2020*, pages 2757–2761.

Kinga Pápay, Szilvia Szeghalmy, and István Szekrényes. 2011. Hucomtech multimodal corpus annotation. *Argumentum*, 7:330–347.

Sangeet Sagar, Mirco Ravanelli, Bernd Kiefer, Ivana Kruijff Korbayova, and Josef van Genabith. 2023. Rescuespeech: A german corpus for speech recognition in search and rescue domain.

Barbara Schuppler, Martin Hagmueller, Juan A. Morales-Cordovilla, and Hannes Pessentheiner. 2014. GRASS: the graz corpus of read and spontaneous speech. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1465–1470, Reykjavik, Iceland. European Language Resources Association (ELRA).

Zaid Sheikh, Shuichiro Shimizu, Siddhant Arora, Jiatong Shi, Samuele Cornell, Xinjian Li, and Shinji Watanabe. 2025. Scalable Spontaneous Speech Dataset (SSSD): Crowdsourcing Data Collection to Promote Dialogue Research. In *Interspeech 2025*, pages 3963–3967.

Per Erik Solberg and Pablo Ortiz. 2022. The norwegian parliamentary speech corpus.

K. Uchimoto, Katsuya Takanashi, K. Takeuchi, C. Nobata, Ikuyo Morimoto, and A. Yamada. 2007. Construction of the corpus of spontaneous japanese and annotation techniques. 54:5–14.

Karl Weilhammer, Uwe Reichel, and Florian Schiel. 2002. Multi-tier annotations in the verbmobil corpus. In *Proceedings of the Third International*

Conference on Language Resources and Evaluation (LREC'02), Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Cihan Xiao, Ruixing Liang, Xiangyu Zhang, Mehmet Emre Tiryaki, Veronica Bae, Lavanya Shankar, Rong Yang, Ethan Poon, Emmanuel Dupoux, Sanjeev Khudanpur, and Leibny Paola Garcia Perera. 2025. Casper: A large scale spontaneous speech dataset.

Zehui Yang, Yifan Chen, Lei Luo, Runyan Yang, Lingxuan Ye, Gaofeng Cheng, Ji Xu, Yaohui Jin, Qingqing Zhang, Pengyuan Zhang, Lei Xie, and Yonghong Yan. 2022. Open source magicdata-ramc: A rich annotated mandarin conversational(ramc) speech dataset.

Weonhee Yun, Kyuchul Yoon, Sunwoo Park, Juhee Lee, Sungmoon Cho, Ducksoo Kang, Koonhyuk Byun, Hyeseung Hahn, and Jungsun Kim. 2015. The korean corpus of spontaneous speech. *Journal of the Korean society of speech sciences*, 7:103–109.