Human-Evaluated Urdu-English Speech Corpus: Advancing Speech-to-Text for Low-Resource Languages

Humaira Mehmood
Fatima Jinnah Women University, Pakistan
humaira.mehmood111@gmail.com

Sadaf Abdul Rauf
Fatima Jinnah Women University, Pakistan
sadaf.abdulrauf@gmail.com

Abstract

This paper presents our contribution to the IWSLT Low Resource Track 2: "Training and Evaluation Data Track". We share a human-evaluated Urdu-English speech-to-text corpus based on Common Voice 13.0 Urdu speech corpus. We followed a three-tier validation scheme which involves an initial automatic translation with corrections from native reviewers, full review by evaluators followed by final validation from a bilingual expert ensuring reliable corpus for subsequent NLP tasks. Our contribution, CV-UrEnST corpus, enriches Urdu speech resources by contributing the first Urdu-English speech-to-text corpus. When evaluated with Whisper-medium, the corpus yielded a significant improvement to the vanilla model in terms of BLEU, chrF++, and COMET scores, demonstrating its effectiveness for speech translation tasks.

Keywords: Speech-to-text (S2T) translation, machine translation (MT), speech recognition (ASR).

1 Introduction

Speech translation (ST) is a key area of speech and natural language processing that involves translating spoken content across languages (Chen et al., 2024; Niehues et al., 2021). It typically integrates automatic speech recognition (ASR), machine translation (MT), and text-to-speech (TTS) capabilities in a pipeline. Early research adopted a cascade paradigm, where ASR, MT, and TTS operated in separate stages (Gaido, 2024; Iranzo-Sánchez et al., 2020). However, recent progress has shifted the focus toward end-to-end architectures that unify these components into a single, trainable model, reducing latency and error propagation between modules (Berard et al., 2016; Niehues et al., 2021; Chen et al., 2024; Gaido, 2024).

Speech-to-text translation (S2T), a specialized form of end-to-end ST, involves converting speech signal in the source language to textual output in the target language (Berard et al., 2016; Niehues et al., 2021; Chen et al., 2024). The success of S2T systems critically depends on the quality, size, and linguistic diversity of the training corpus, which underpins model generalization and robustness (Amrouche et al., 2023; Cattoni et al., 2021).

Historically, S2T corpora have evolved from task-specific datasets to large-scale multilingual resources that are essential for building performant translation systems (Cieri et al., 2004; Wang et al., 2020; Miller et al., 2021; Sikasote and Anastasopoulos, 2022; Sethiya et al., 2024). Despite this evolution, corpus creation for low-resource languages remains severely underdeveloped due to challenges such as dialectal diversity, limited written resources, and high annotation costs (Verdonik et al., 2024).

While, these advances have accelerated progress in ST for high-resource languages, low-resource languages continue to face substantial challenges (Shanbhogue et al., 2023; Bartelds et al., 2023; Court and Elsner, 2024). ASR, TTS, and MT models have shown impressive gains in well-resourced settings, but the lack of well-annotated, parallel speech-text corpora has hindered similar progress for underrepresented languages like Urdu. This data scarcity is a fundamental bottleneck not only for ST but also for downstream tasks like cross-lingual retrieval and multilingual dialogue systems (Magueresse et al., 2020; Singh et al., 2024; Farooq et al., 2019).

Urdu remains a low-resource language for speech translation, with only a few domain-specific corpora available (Qasim et al., 2016). Urdu-English is a moderately resourced language pair with existing corpora for TTS (Jamal et al., 2022), ASR (Arif et al., 2025) and machine trans-

Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025), pages 138–144

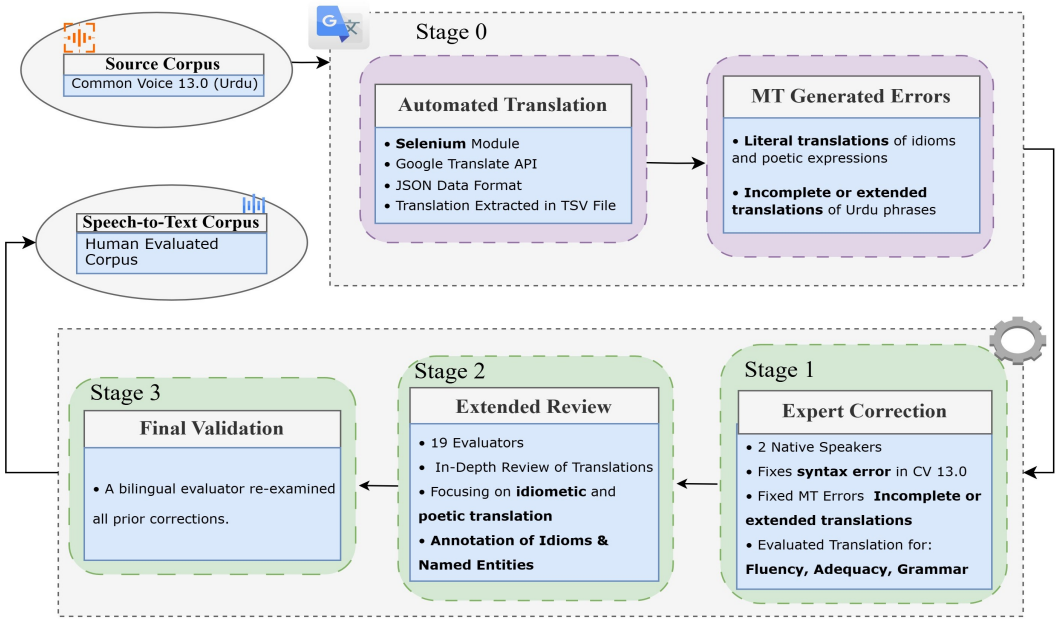

Figure 1: Steps involved in the corpus creation pipeline

lation (Abdul Rauf et al., 2020; Abdul Rauf and Hira, 2023) but speech to text corpus has been noticeably absent. This work is a first step in this direction, where we present and share first human-evaluated Urdu-English S2T corpus namely CV-UrEnST.

We worked on an Urdu subset of Mozilla Common Voice 13.0 (Ardila et al., 2020). Though, Mozilla Common Voice provides open-source Urdu speech, yet its original transcriptions are unvalidated and susceptible to crowd-sourced inconsistencies (Ardila et al., 2020). Since, the Urdu transcriptions underwent comprehensive human validation, they can be used as a gold-standard foundation for ASR tasks. In addition to transcription integrity, we ensured precise English translations that carefully preserve idiomatic structures, named entities, cultural references, and semantic intent. This positions the corpus as a valuable parallel text resource for Urdu-English machine translation and cross-lingual NLP applications.

We followed a three-tier validation scheme. Firstly, *Initial Translations* were generated automatically using the Google Translate API. This was followed by an *Expert Correction* phase, where two native Urdu-English bilinguals manually refined the translations to eliminate syntactic, semantic, and contextual errors. Finally, an *Extended Review and Final Validation* was performed. This multi-phase pipeline ensures high inter-annotator reliability, contextual fidelity, and translation accuracy, improving the datasets suitability for both speech-to-text and text-to-text modeling.

2 Related Work

Recent advances in multilingual low-resource speech datasets have led to innovative data collection and transcription strategies. For instance, Yang et al. (2024) introduced GigaSpeech 2, an ASR corpus for Thai, Indonesian, and Vietnamese via automated web crawling, transcription, and iterative refinement. Abraham et al. (2020) focused on Marathi ASR and emphasized diversity by sourcing speech from 36 speakers across rural and urban communities, yielding a 109-hour corpus that captures dialectal variance.

Community-driven initiatives are also central to low-resource dataset development. Butryna et al. (2020) presented 38 crowd sourced corpora spanning Asia, Africa, and the Americas, underscoring the role of open data in promoting global speech technology. Similarly, Guevara et al. (2024) released a 454 hour multilingual corpus across 10 Philippine languages, collected from domains like healthcare, education, and spontaneous speech demonstrating the value of domain and register diversity in corpus utility.

The availability of *Urdu-English* speech-to-text corpora remains sparse compared to better-

Our Correction	Common Voice Transcription	Error Type
اس مسئلہ مت بناؤ اللہ کا واسطہ	اس مسئلہ اللہ کا واسطہ مہے بناؤ	**Orthographic**
ریاں مھو ایک ہے۔ گھر اکم پہ سوت سود	ریاں مھو ایک ہے گھر اکم پہ سوت سود	
استعمال لگاسرام	استعمال لگاسرام	
روپےواگ سے بھی ہے۔ ڈھل گیا	روپےواگ ہے ٹھیبے ڈھل پاغ هیں	
نے ہر گلس ناہ ہے بچے گئی سرناب	نے ہر گلس ناہ نہی سرنابیہ گئے	
کمی کے بچے نئے ٹارز کے استعمال سے پیداوار میں اضافہ ، مہنگے ٹارز سے	کمی کے بچے نئے ٹارز کے استعمال سے پیداوار ٹارز کے گنہمہ اضافہ میں سے	
بچے کی پیداوار میں	بچے کی پیداوار میں	
پس منظر	پس منظر	
وہ بیٹھے چاہے نہ کیو دکھ پو آجے اسٹارڈی نائٹ کپا	وہ وجہ اسٹارڈی نائٹ کپا اپ بیٹھے چاہے نہ کیو دکھ	
وکپ آپ کیسے ہیں۔ نظر منظر ہے ایہ	وکا پیہ کہ یسنے؟ نظر منظر ہے ایہ	**Morphological**
اھر وہ لیکن زلزلے سے ڈھ راشتہ والے زجیرے ارٹلس اور بیا میشیرنوڈنا کئی نیقیرو کو گوں ا ھی نہیں ھا۔ ت	وکھی لیکن زلزلی سی نوہ راشتہ والی زجیری ارٹلس اور ان بیا میشیرنوڈنہ کئی وگوں ا نیقیر	
بس، کسی شادی ہا ہائے	بس کس ہا ہائے شادی	**Punctuation**
وکل زنہ نمے اچھہ رامی ہاراکھاٹ شہ	لزنہ کونے اچھہ رہاکھاٹ شہ راھی	
ٹی ٹونٹی ورلڈکپ سے میرے کیریئر کا آخری ورلڈکپ ہو گا، ۔ یددیافر دھاشہ	ٹیی ٹونٹ ورلڈکپ سے میرے کیریئر کا آخری ورلڈ کپ ہو گا فرد ھاشہ یدی	**Named Entity**
سے اُر دازہہ شہا ہاتپ چلا دنر کرادر کی فنہ	دازہہ شہا تپ چلا دنر کرادر کی فنہ روئے	
سے وبابرمز کے خلاف ل یھ کسیں میں رارھ	بابرمز کے خلاف ل یھ کسیں میں رارھ	

Table 1: Category wise examples of transcription discrepancies in Mozilla Common Voice 13.0

resourced language pairs. While, efforts like the Urdu-English Parallel Corpus for Speech Translation offer foundational bilingual resources, they often lack rigorous human evaluation and speech alignment (Furqan et al., 2024; Amin et al., 2025). Mozilla Common Voice provides open-source Urdu speech, yet its original transcriptions are unvalidated and susceptible to crowd-sourced inconsistencies (Ardila et al., 2020).

In machine translation, domain-specific corpora have contributed meaningfully. For example, the LEGAL-UQA dataset addresses the legal Q&A domain using constitutional texts (Faisal and Yousaf, 2024), while the Urdu-English Religious Domain Corpus offers 18,426 sentence pairs for theological texts (Abdul Rauf and Hira, 2023). Although these datasets advance text-to-text MT, they lack paired audio components necessary for S2T applications.

3 Corpus Preparation Pipeline

Our corpus comprises of approximately 7k sentence pairs from the Urdu subset of Mozilla Common Voice 13.0 (Ardila et al., 2020). The focus was on creating high-quality, human-validated translations rather than maximizing scale. Common Voice was chosen for its open license, reproducibility, and established use in speech research. Future expansions will consider integrating other open-access Urdu speech resources, contingent on annotation capacity.

Common Voice 13.0 features a community-driven validation system, where users vote on the correctness of audio-transcription pairs. While, this process ensures surface-level alignment, it does not address deeper syntactic or semantic inconsistencies common in Urdu, a morphologically rich language. Examples of such issues are presented in Table 1. Here, orthographic correspond to incorrect spellings, character substitutions, or missing graphemes. Semantic errors stem from misinterpretations of meaning. Named entity errors involve improper handling of proper nouns or technical terms. Punctuation and diacritic include inconsistencies that affect readability and disambiguation.

Machine Translation and Expert Correction MT often fails to preserve cultural complexity, idiomatic expressions, and emotional tone. Other common issues included incomplete renderings

	Machine Translation	Expert's Correction	Extended Review	Final Validation
Incomplete Translations:				
آبی جانور میں بطخ بگلا اور دُوسْرا آبی پرندہ شامل ہونا	Aquatic beasts include ducks and other aquatic birds	In aquatic animals, ducks and other aquatic birds are included	–	–
جو چاہے کر سکتے تھے۔	Whatever they wanted	–	–	They could do whatever they wanted.
ڈیرن سیمی کا نرالا انداز، کرتا زیب تن کرلیا	Darren Sammy's quirky wear	Darren Sammy has dressed up in a quirky style.	-	Darren Sammy's adopted unique style by wearing the kurta.
پہلا میل اتنا طویل لگا کہ قریب تھا کہ اسے سازش کا شاخسانہ قرار دے دیا جاتا	The first mail was so long that it was close	The first mile seemed so long, saying it was close, it was declared as a sign of conspiracy	-	The first interaction felt so prolonged that it was almost labeled as a result of a conspiracy
Poetic Translations:				
یار آشنا نہیں کوئی ٹکرائیں کس سے جام	Dude no one to collide with whom	–	No friend or lover is around, whom shall I toast with	–
ہر رگ خوں میں پھر چراغاں ہو	In every vein, then the light	Let there be lights in every vein again	To be fired up	Lets ignite the spark again
بیٹھا ہے بت آئنہ سیما مرے آگے	Sitting is an idol, Sima Murray in front of my mirror		The idol, with a mirror-like beauty, sits before me	The idol with a mirror-like visage is sitting before me
منزل کو نہ پہچانے، راہِ عشق کا راہی۔	Do not recognize the destination	-	-	The traveler on the path of love does not recognize the destination
Extended Translations:				
جسے سن کر عبدالقادر کی انکھیں	Hearing this, Abdul Qadir's eyes widened	–	Hearing this, Abdul Qadir's eyes	–
فلمیں دیکھنی ہو	I have to watch movies.			want to watch movies
Idiomatic Sentences:				
کبھی کبھار ھی خیالی پلاو بناتا ہوں	The traveler on the path of love does not recognize the destination	Sometimes I make air castle	–	Sometimes I build castles in the air.
بکرے کی ماں کب تک خیر منائے گی	How long will the goat's mother welcome	how long will the mother's prayers avail to save her kid	How long will you delay the inevitable?	-
جیسی نیت ویسی مراد	The same intention is Visi	As the intention, so is the outcome.	-	-
نہ رہے گا بانس نہ بجے گی بانسری	Will not be bamboo or pm	To deal with the issue at its root to prevent a more challenging problem.		If the bamboo is gone, the flute won't play
جو یَہاں کا جنگلات کےحسن کو چار چاند لگا دینا ہونا	To give four moons to the beauty of these forests	To enhance the beauty of these forests.	-	-

Table 2: Examples of translation in different validation phases for complex Urdu expressions

where parts of the original Urdu were missing, literal translations of idioms, and incorrect substitution of culturally specific terms. The reviewers refined these translations to ensure contextual fidelity and semantic precision. Each sentence was independently assessed across three linguistic dimensions: accuracy (semantic alignment), adequacy (completeness of meaning transfer), and fluency (naturalness and readability in English).

Consider the idiom جو یَہاں کا جنگلات کےحسن کو چار چاند لگا دینا ہونا, shown in Table 2, last row translated as "to give four moons to the beauty of these forests" a literal rendering of a metaphor for beautification. Similarly, the phrase بکاو چینل was mistranslated as "Baku channel", ignoring its intended meaning of "biased or corrupt media outlets."

Another poetic example, کہیں تو بہر خدا آج ذکر یار چلے , was rendered as "Somewhere else, God goes to Zikr today," which loses its figurative essence. A better translation "Let there be, for Gods sake, some talk of the beloved today" captures both semantic and emotional intent. Lastly, نہ اب رقیب نہ ناصح نہ غم گسار کوئی was translated as "No longer the rival nor Nasah nor the grief," where ناصح (meaning moral advisor) was poorly transliterated as "Nasah" .A faithful rendering would be: "Now, there remains no rival, no guide, and no comforter to ease the sorrow."

All such mistranslations were corrected during the extended evaluation phase, ensuring cultural fidelity and correct lexical choice in cultural and linguistic contexts.

Extended Review The second phase of validation involved 19 bilingual reviewers. These were graduate students in computer science, all native Urdu speakers with advanced academic proficiency in English. Each reviewer reviewed equal portion of the corpus and was instructed to focus on refining translation by checking for errors in idiomatic usage, named entities, and cultural refer-

ences.

As the corpus was partitioned into non-overlapping subsets, standard inter-annotator agreement metrics like Cohens Kappa could not be applied. To maintain annotation consistency, we provided comprehensive guidelines and examples to all annotators. In addition, a senior linguist performed a qualitative audit of randomly selected annotated pairs to verify adherence to syntactic, semantic, and cultural fidelity standards.

Employing a distributed review strategy offered several advantages. Crowdsourced evaluation, especially when conducted by native speakers with relevant academic backgrounds, has been shown to improve translation quality through consensus and error cross-checking (Zaidan and Callison-Burch, 2011). The diversity of reviewers helps to detect inconsistencies and ensures a more comprehensive assessment of the data. This phase was particularly valuable for capturing subtle sophistication that may have been overlooked in earlier stages.

Final Validation In the final stage of quality control, a senior bilingual evaluator fluent in both Urdu and English reassessed the outputs from the extended review phase. This validation focused specifically on the test set to ensure translation consistency, semantic constancy, and contextual appropriateness. Table 2 shows the representation of Idioms and named entities in the final corpus.

Data	Audio Count	Idioms	Equivalent Idioms	Named Entities
Test	4129	32	8	1152
Train	3304	17	3	648
Total	7433	49	11	1800

Table 3: Distribution of Annotated Idioms, Equivalent Idioms, and Named Entities in the Corpus

4 Model Building

To establish a performance reference, we fine-tuned OpenAIs `Whisper-medium`[1] model, a transformer-based encoder-decoder pretrained on multilingual speech data for direct speech-to-text translation.

Training was performed on a Google Colab A100 GPU using the AdamW optimizer with a learning rate of 1×10^{-5} with cosine annealing. The batch size of 16 was used and early stopping was based on the improvements in the BLEU score on the development set. All audio inputs were resampled to 16 kHz and converted into 80-bin log-Mel spectrograms. Zero-padding ensured uniform sequence lengths within batches.

We used BLEU (token-level accuracy), chrF++ (character-level fluency), and COMET (semantic adequacy) as evaluation metrics. BLEU scores measures token-level overlap with reference translations and reflects surface-level accuracy. chrF++ captures character-level fluency and recall, it is especially robust to morphological variation, whereas COMET evaluates semantic similarity using neural metrics, higher scores indicate better meaning preservation.

Scores We evaluated vanilla and fine-tuned `Whisper-medium` models on our test set. The original model, without domain-specific adaptation, showed minimal performance. In contrast, fine-tuning yielded substantial gains across all evaluation dimensions.

Metric	Original	Fine-tuned
BLEU	0.81	21.49
chrF++	6.31	46.22
COMET	0.414	0.731

Table 4: Evaluation results of Whisper-medium model before and after fine-tuning on our dataset.

These results demonstrate that the proposed dataset significantly enhances the Whisper models ability to produce fluent and semantically accurate translations, validating its utility for low-resource speech translation.

5 Conclusion

This study contributes a human evaluated Urdu-to-English speech-to-text corpus designed to advance NLP research in under-resourced linguistic domains. By integrating automated translation with systematic human validation, we address critical gaps in handling idiomatic and culturally specific content, producing translations which retain the cultural aspects.

References

Sadaf Abdul Rauf, Syeda Abida, Noor-e Hira, Syeda Zahra, Dania Parvez, Javeria Bashir, and Qurat-ul-

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

ain Majid. 2020. On the exploration of English to Urdu machine translation. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 285--293, Marseille, France. European Language Resources association.

Sadaf Abdul Rauf and Noor e Hira. 2023. Development of Urdu-English religious domain parallel corpus. In *Proceedings of the Second Workshop on Corpus Generation and Corpus Augmentation for Machine Translation*, pages 14--21, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Basil Abraham, Danish Goel, Divya Siddarth, Kalika Bali, Manu Chopra, Monojit Choudhury, Pratik Joshi, Preethi Jyoti, Sunayana Sitaram, and Vivek Seshadri. 2020. Crowdsourcing speech data for low-resource languages from low-income workers. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2819--2826, Marseille, France. European Language Resources Association.

Muhammad Shahid Amin, Xiaoqiang Zhang, Luca Anselma, Alessandro Mazzei, and Johan Bos. 2025. Semantic processing for urdu: corpus creation, parsing, and generation. *Language Resources and Evaluation*.

Aissa Amrouche, Youssouf Bentrcia, Nabil Hezil, Khadidja Boubakeur, Nawel Behloul, Miloud Zalagh, Abed Ahcene, and Leila Falek. 2023. BAC TTS Corpus: Rich Arabic Database for Speech Synthesis. In *2023 International Conference on Electrical and Electronics Engineering (ICEEE)*, pages 189--193.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218--4222, Marseille, France. European Language Resources Association.

Samee Arif, Aamina Jamal Khan, Mustafa Abbas, Agha Ali Raza, and Awais Athar. 2025. WER we stand: Benchmarking Urdu ASR models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5952--5961, Abu Dhabi, UAE. Association for Computational Linguistics.

Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling. 2023. Making more of little data: Improving low-resource automatic speech recognition using data augmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 715--729, Toronto, Canada. Association for Computational Linguistics.

Alexandre Berard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. *Preprint*, arXiv:1612.01744.

Alena Butryna, Shan-Hui Cathy Chu, Isin Demirsahin, Alexander Gutkin, Linne Ha, Fei He, Martin Jansche, Cibu Johny, Anna Katanova, Oddur Kjartansson, Chenfang Li, Tatiana Merkulova, Yin May Oo, Knot Pipatsrisawat, Clara Rivera, Supheakmungkol Sarin, Pasindu de Silva, Keshan Sodimana, Richard Sproat, Theeraphol Wattanavekin, and Jaka Aris Eko Wibawa. 2020. Google crowdsourced speech corpora and related open-source resources for low-resource languages and dialects: An overview. *Preprint*, arXiv:2010.06778.

Roldano Cattoni, Mattia Antonino Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Must-c: A multilingual corpus for end-to-end speech translation. *Computer Speech Language*, 66:101155.

William Chen, Wangyou Zhang, Yifan Peng, Xinjian Li, Jinchuan Tian, Jiatong Shi, Xuankai Chang, Soumi Maiti, Karen Livescu, and Shinji Watanabe. 2024. Towards robust speech representation learning for thousands of languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10205--10224, Miami, Florida, USA. Association for Computational Linguistics.

Christopher Cieri, David Miller, and Kevin Walker. 2004. The fisher corpus: a resource for the next generations of speech-to-text. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Sara Court and Micha Elsner. 2024. Shortcomings of LLMs for low-resource translation: Retrieval and understanding are both the problem. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1332--1354, Miami, Florida, USA. Association for Computational Linguistics.

Faizan Faisal and Umair Yousaf. 2024. Legal-uqa: A low-resource urdu-english dataset for legal question answering. *Preprint*, arXiv:2410.13013.

Muhammad Umar Farooq, Farah Adeeba, Sahar Rauf, and Sarmad Hussain. 2019. Improving large vocabulary urdu speech recognition system using deep neural networks. In *Proceedings of Interspeech 2019*.

Muhammad Furqan, Rayan Bin Khaja, and Rameez Habeeb. 2024. Erupd - english to roman urdu parallel dataset. *arXiv preprint arXiv:2412.17562*.

Marco Gaido. 2024. Direct speech translation toward high-quality, inclusive, and augmented systems. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 2--3, Sheffield, UK. European Association for Machine Translation (EAMT).

Rowena Cristina L. Guevara, Rhandley D. Cajote, Michael Gringo Angelo R. Bayona, and Crisron Rudolf G. Lucas. 2024. Philippine languages database: A multilingual speech corpora for developing systems for low-resource languages. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages*

@ *LREC-COLING 2024*, pages 264--271, Torino, Italia. ELRA and ICCL.

Javier Iranzo-Sánchez, Adrià Giménez Pastor, Joan Albert Silvestre-Cerdà, Pau Baquero-Arnal, Jorge Civera Saiz, and Alfons Juan. 2020. Direct segmentation models for streaming speech translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2599--2611, Online. Association for Computational Linguistics.

Sahar Jamal, Sadaf Abdul Rauf, and Quratulain Majid. 2022. Exploring transfer learning for Urdu speech synthesis. In *Proceedings of the Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia within the 13th Language Resources and Evaluation Conference*, pages 70--74, Marseille, France. European Language Resources Association.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *Preprint*, arXiv:2006.07264.

Corey Miller, Evelyne Tzoukermann, Jennifer Doyon, and Elizabeth Mallard. 2021. Corpus creation and evaluation for speech-to-text and speech translation. In *Proceedings of Machine Translation Summit XVIII: Users and Providers Track*, pages 44--53, Virtual. Association for Machine Translation in the Americas.

Jan Niehues, Elizabeth Salesky, Marco Turchi, and Matteo Negri. 2021. Tutorial: End-to-end speech translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 10--13, online. Association for Computational Linguistics.

Muhammad Qasim, Sahar Rauf, Sarmad Hussain, and Tania Habib. 2016. Urdu speech corpus for travel domain. pages 237--240.

Nivedita Sethiya, Saanvi Nair, and Chandresh Maurya. 2024. Indic-TEDST: Datasets and baselines for low-resource speech to text translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9019--9024, Torino, Italia. ELRA and ICCL.

Akshaya Vishnu Kudlu Shanbhogue, Ran Xue, Soumya Saha, Daniel Zhang, and Ashwinkumar Ganesan. 2023. Improving low resource speech translation with data augmentation and ensemble strategies. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 241--250, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Claytone Sikasote and Antonios Anastasopoulos. 2022. BembaSpeech: A speech recognition corpus for the Bemba language. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7277--7283, Marseille, France. European Language Resources Association.

Deepanjali Singh, Ayush Anand, Abhyuday Chaturvedi, and Niyati Baliyan. 2024. IWSLT 2024 Indic track system description paper: Speech-to-text translation from English to multiple low-resource Indian languages. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 311--316, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.

Darinka Verdonik, Andreja Bizjak, Andrej gank, Mirjam Sepesy Mauec, Mitja Trojar, Jerneja ganec Gros, Marko Bajec, Iztok Lebar Bajec, and Simon Dobriek. 2024. Strategies for managing time and costs in speech corpus creation: Insights from the slovenian artur corpus. *Language Resources and Evaluation*.

Changhan Wang, Juan Pino, Anne Wu, and Jiatao Gu. 2020. CoVoST: A diverse multilingual speech-to-text translation corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4197--4203, Marseille, France. European Language Resources Association.

Yifan Yang, Zheshu Song, Jianheng Zhuo, Mingyu Cui, Jinpeng Li, Bo Yang, Yexing Du, Ziyang Ma, Xunying Liu, Ziyuan Wang, Ke Li, Shuai Fan, Kai Yu, Wei-Qiang Zhang, Guoguo Chen, and Xie Chen. 2024. Gigaspeech 2: An evolving, large-scale and multi-domain asr corpus for low-resource languages with automated crawling, transcription and refinement. *arXiv preprint arXiv:2406.11546*.

Omar F. Zaidan and Chris Callison-Burch. 2011. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1220--1229, Portland, Oregon, USA. Association for Computational Linguistics.