CROSSVOICE: CROSSLINGUAL PROSODY PRESERVING CASCADE-S2ST USING TRANSFER LEARNING

Medha Hira*, Arnav Goel* & Anubha Gupta
Indraprastha Institute of Information Technology
New Delhi - 110020, India
{medha21265,arnav21519,anubha}@iiitd.ac.in

ABSTRACT

This paper presents CrossVoice, a novel cascade-based Speech-to-Speech Translation (S2ST) system employing advanced ASR, MT, and TTS technologies with cross-lingual prosody preservation through transfer learning. We conducted comprehensive experiments comparing CrossVoice with direct-S2ST systems, showing improved BLEU scores on tasks such as Fisher Es-En, VoxPopuli Fr-En and prosody preservation on benchmark datasets CVSS-T and IndicTTS. With an average mean opinion score of 3.6 out of 4, speech synthesized by CrossVoice closely rivals human speech on the benchmark highlighting the efficacy of cascade-based systems and transfer learning in multilingual S2ST with prosody transfer.

1 INTRODUCTION

Transformer-based models (Vaswani et al., 2017) have revolutionized speech processing, leading to significant advancements in automatic speech recognition and text-to-speech technologies (Latif et al., 2023; Prabhavalkar et al., 2023). This shift towards end-to-end systems has opened new avenues in Speech-to-Speech Translation (S2ST) for translating speech across languages. Our work introduces *CrossVoice*, a cascade-based S2ST system utilizing the latest open-source automatic speech recognition (ASR), machine translation (MT), and text-to-speech (TTS) models unlike direct S2ST methods that bypass MT. It is evaluated against state-of-the-art (SOTA) direct S2ST systems for speech quality, cross-lingual prosody preservation, and translation accuracy using BLEU (BiLingual Evaluation Understudy) score (Papineni et al., 2002). Further, we investigate the performance of cascade-based vis-á-vis direct approaches in S2ST and demonstrate how transfer learning can enhance prosody transfer in cross-lingual settings.

2 RELATED WORK

Current open-source systems for direct-S2ST involve various techniques such as self-supervised learning (Lee et al., 2021b), using speech discrete units (Lee et al., 2021a), text modalities (Zhang et al., 2023) and linguistic decoders (Jia et al., 2022a). However, these systems often face challenges including lower translation accuracy and inferior audio quality, particularly, in cross-lingual prosody transfer (Bentivogli et al., 2021). In contrast, cascade-based S2ST systems that integrate separate ASR, MT, and TTS models (Nakamura et al., 2006) are criticized for high latency and subpar prosody transfer (Latif et al., 2021).

Recent advancements in transfer-learning, such as voice cloning (Jia et al., 2019) and transformer-based ASR and TTS, suggest the potential for more efficient and effective prosody transfer in cascade-based systems (Huang et al., 2023). Our study leverages these SOTA technologies in the proposed cascade-based framework, *CrossVoice*, and compares its performance with direct S2ST systems on prosody transfer and overall efficiency.

3 METHODOLOGY

CrossVoice integrates state-of-the-art ASR, MT, and TTS techniques to establish a baseline translation cascade: 1) Faster-Whisper[1] for ASR (comparision of other ASR models in A.2), which is a faster and batch-capable version of Whisper-Large (Radford et al., 2022; Moslem et al., 2022); 2)

*Equal Contribution
[1]https://github.com/SYSTRAN/faster-whisper

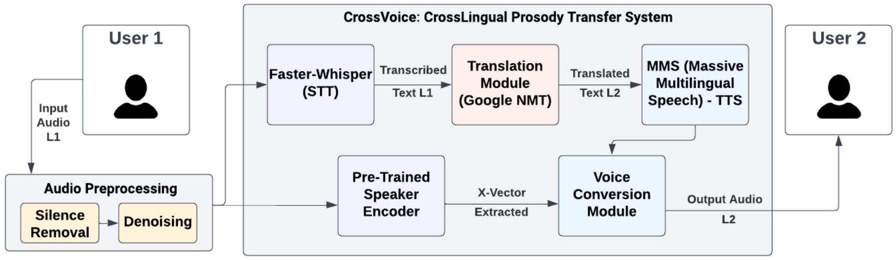

Figure 1: Proposed Architecture for CrossVoice

Google's NMT model (Wu et al., 2016) for MT, which is known to reduce error rates significantly; and 3) the Massive Multilingual Speech (MMS) model (Pratap et al., 2023) based on VITS-TTS (Kim et al., 2021) for TTS, which is capable of handling over 1000 languages with superior performance in linguistic diversity and speech synthesis. CrossVoice uses transfer learning on a voice cloning module (trained on the speaker identification task) for prosody preservation. For this, a pre-trained speaker encoder generates X-vector embeddings (Ravanelli et al., 2021; Snyder et al., 2018) and is coupled with FreeVC's (Li et al., 2023) voice conversion module to effectively transfer speaker prosody.

We conducted two sets of experiments to evaluate our system's performance in translation and speech synthesis. The first experiment evaluates synthesized speech quality on the CVSS-T (Jia et al., 2022b) and IndicTTS benchmark datasets (Kumar et al., 2023). We report mean opinion scores (MOS) from a survey of 40 respondents, rating on a four-point scale with a 95% confidence interval as per the protocol of Huang et al. (2023). MOS-h represents ratings for natural human speech (called as Ground Truth or GT here), MOS-v for baseline TTS audio without prosody transfer, and MOS-c for speech synthesized by CrossVoice[2]. The second experiment compares the BLEU performance of CrossVoice with recent direct-S2ST SOTA systems discussed in Section 2 on the translation tasks for which their superiority has been claimed over cascade-based systems.

4 Results

Table 1 tabulates results of the first experiment on five translation tasks. MOS-c score is almost the same as MOS-h (i.e., the GT) and also beats MOS-v scores of the vanilla TTS considerably, by almost 40% on each task. Figures 2 and 3 (**see A.5**) highlight high BLEU scores of CrossVoice that averaged to 33.4 over all the languages of the chosen benchmark datasets.

Table-2 lists the BLEU scores of CrossVoice (BLEU-c) and SOTA methods (BLEU-r). For calculating the BLEU scores, we employed Whisper (using the *temperature setting of one* and *greedy decoding*) for generating transcripts of the speech generated using CrossVoice and SOTA methods. We sourced BLEU scores from the original papers for the SOTA methods (reported as BLEU-r). CrossVoice surpasses the claimed superior performance of direct S2ST systems in their respective tasks, notably achieving almost a 19-point increase in BLEU score in the VoxPopuli S2ST Fr-En task. This significant performance boost is attributed to effective ASR and precise audio reconstruction through voice cloning.

Table 1: MOS comparison on S2ST quality

Translation Task	MOS-h (↑) (GT)	MOS-v (↑)° (Vanilla TTS)	MOS-c (↑)° (CrossVoice)
Spanish-English†	3.88	2.75 ± 0.12	$\mathbf{3.76 \pm 0.08}$
German-English†	3.83	2.64 ± 0.05	3.73 ± 0.11
Italian-English†	3.75	2.89 ± 0.01	3.53 ± 0.10
Hindi-English⋆	3.79	2.54 ± 0.07	3.63 ± 0.02
English-Hindi⋆	3.67	2.65 ± 0.03	3.34 ± 0.04

†CVSS-T, ⋆Indic-TTS, °mean±std

Table 2: Comparison on S2ST-BLEU

Task (reported in SOTA method)	BLEU-r (↑) (SOTA method)	BLEU-c (↑) (CrossVoice)
Fisher Es-En	42.9 (Jia et al., 2022a)	**45.6**
Fisher Es-En	39.9 (Lee et al., 2021a)	**45.6**
MuST-C En-De	30.2 (Zhang et al., 2023)	**39.7**
MuST-C En-Fr	40.8 (Zhang et al., 2023)	**46.5**
VoxPopuli Fr-En	20.3 (Lee et al., 2021b)	**39.6**

5 Conclusion and Future Work

CrossVoice effectively combines advanced ASR, MT, and TTS technologies, establishing itself as a highly proficient cascade-based S2ST system with strengths in cross-lingual prosody preservation and translation accuracy. Our comprehensive experiments reveal that CrossVoice outperforms existing direct S2ST systems, underscoring the effectiveness and reliability of cascade-based systems with transfer learning for direct speech translation across languages. Future work includes improving transfer of emphasis and intonation across languages as reported in A.4.

[2]Details about MOS calculations and the protocol are given in the appendix A

URM Statement

The authors acknowledge that all the authors of this work meet the URM criteria of ICLR 2024 Tiny Papers Track.

References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020.

Luisa Bentivogli, Mauro Cettolo, Marco Gaido, Alina Karakanta, Alberto Martinelli, Matteo Negri, and Marco Turchi. Cascade versus direct speech translation: Do the differences still make a difference?, 2021.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, October 2022. ISSN 1941-0484. doi: 10.1109/jstsp.2022.3188113. URL `http://dx.doi.org/10.1109/JSTSP.2022.3188113`.

Mattia A Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. Must-c: a multilingual speech translation corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2012–2017. Association for Computational Linguistics, 2019.

Wen-Chin Huang, Benjamin Peloquin, Justine Kao, Changhan Wang, Hongyu Gong, Elizabeth Salesky, Yossi Adi, Ann Lee, and Peng-Jen Chen. A holistic cascade system, benchmark, and human evaluation protocol for expressive speech-to-speech translation, 2023.

Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu. Transfer learning from speaker verification to multispeaker text-to-speech synthesis, 2019.

Ye Jia, Michelle Tadmor Ramanovich, Tal Remez, and Roi Pomerantz. Translatotron 2: High-quality direct speech-to-speech translation with voice preservation, 2022a.

Ye Jia, Michelle Tadmor Ramanovich, Quan Wang, and Heiga Zen. CVSS corpus and massively multilingual speech-to-speech translation. In *Proceedings of Language Resources and Evaluation Conference (LREC)*, pp. 6691–6703, 2022b.

Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pp. 5530–5540. PMLR, 2021.

Gokul Karthik Kumar, Praveen S V au2, Pratyush Kumar, Mitesh M. Khapra, and Karthik Nandakumar. Towards building text-to-speech systems for the next billion users, 2023.

Siddique Latif, Inyoung Kim, Ioan Calapodescu, and Laurent Besacier. Controlling prosody in end-to-end TTS: A case study on contrastive focus generation. In Arianna Bisazza and Omri Abend (eds.), *Proceedings of the 25th Conference on Computational Natural Language Learning*, pp. 544–551, Online, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.conll-1.42. URL `https://aclanthology.org/2021.conll-1.42`.

Siddique Latif, Aun Zaidi, Heriberto Cuayahuitl, Fahad Shamshad, Moazzam Shoukat, and Junaid Qadir. Transformers in speech processing: A survey, 2023.

Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Sravya Popuri, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, et al. Direct speech-to-speech translation with discrete units. *arXiv preprint arXiv:2107.05604*, 2021a.

Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Yossi Adi, Juan Pino, Jiatao Gu, et al. Textless speech-to-speech translation on real data. *arXiv preprint arXiv:2112.08352*, 2021b.

Jingyi Li, Weiping Tu, and Li Xiao. Freevc: Towards high-quality text-free one-shot voice conversion. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Yasmin Moslem, Rejwanul Haque, and Andy Way. Translation word-level auto-completion: What can we achieve out of the box?, 2022.

Satoshi Nakamura, Konstantin Markov, Hiromi Nakaiwa, Gen-ichiro Kikui, Hisashi Kawai, Takatoshi Jitsuhiro, J-S Zhang, Hirofumi Yamamoto, Eiichiro Sumita, and Seiichi Yamamoto. The atr multilingual speech-to-speech translation system. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2):365–376, 2006.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210. IEEE, 2015.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL `https://aclanthology.org/P02-1040`.

Matt Post, Gaurav Kumar, Adam Lopez, Damianos Karakos, Chris Callison-Burch, and Sanjeev Khudanpur. Fisher and callhome spanish–english speech translation. *LDC2014T23. Web Download. Philadelphia: Linguistic Data Consortium*, 2014.

Rohit Prabhavalkar, Takaaki Hori, Tara N. Sainath, Ralf Schlüter, and Shinji Watanabe. End-to-end speech recognition: A survey, 2023.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. Scaling speech technology to 1,000+ languages, 2023.

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

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. SpeechBrain: A general-purpose speech toolkit, 2021. arXiv:2106.04624.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5329–5333. IEEE, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Changhan Wang, Morgane Rivière, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation, 2021.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation, 2016.

Yuhao Zhang, Chen Xu, Bojie Hu, Chunliang Zhang, Tong Xiao, and Jingbo Zhu. Improving end-to-end speech translation by leveraging auxiliary speech and text data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 13984–13992, 2023.

A Appendix

A.1 Acronyms used

- MOS : Mean Opinion Score
- S2ST : Speech to Speech Translation
- ASR : Automated Speech Recognition
- MT : Machine Translation
- NMT : Neural Machine Translation
- TTS : Text to Speech
- GT : Ground Truth
- SOTA : State-Of-The-Art
- BLEU : Bilingual Evaluation Understudy

A.2 ASR Results

We compared various ASR models such as variants of Whisper, Wav2Vec2.0 (Baevski et al., 2020), WavLM (Chen et al., 2022) and Faster-Whisper on multilingual datasets: Librispeech-test-clean (English), IndicTTS (Indian accented English speech) (Kumar et al., 2023) and VoxPopuli - French, Spanish and German. Results are shown in Table 3 and 4. Faster-Whisper clearly performs very well on both WER and average latency metrics. We measured average latency as the weighted average of the time taken to transcribe each sample of the entire dataset.

Table 3: Results of different ASR Models on Librspeech-test-clean subset (Panayotov et al., 2015)

Model	WER (%)	Average Latency (s)
Whisper - Tiny	9.78	0.183
Whisper - Base	6.94	0.234
Whisper - Small	4.85	0.385
Whisper - Large	3.63	1.145
Wav2Vec2.0 - Large	3.20	0.415
WavLM - Large	2.80	0.525
Faster-Whisper	**4.23**	**0.152**

Table 4: WER benchmarking of models on various Datasets

Model	IndicTTS-en	VoxPopuli-French	VoxPopuli-Spanish	VoxPopuli-German
Wav2Vec2.0 - XLSR	15.65	25.34	21.34	24.73
WavLM - Large	14.25	23.21	18.65	20.56
Whisper - Tiny	10.74	31.53	19.63	25.24
Whisper - Base	8.63	21.34	15.32	19.75
Whisper - Small	5.28	13.24	12.18	13.32
Whisper - Large	3.85	10.56	7.82	9.75
Faster-Whisper	4.38	11.23	8.96	10.32

A.3 Translation Tasks

We benchmarked CrossVoice on 3 benchmark S2ST tasks and they are summarised as follows:

1. Fisher (Spanish-English) (Post et al., 2014): The Fisher Spanish dataset is a collection of telephone speech conversations in Spanish, primarily involving topics of daily life. It

contains over 160 hours of recorded conversations, involves more than 130,000 utterances and includes around 24,000 speakers.

2. MuST-C (English to German & English to French) (Di Gangi et al., 2019): It is a multilingual speech translation corpus with 273 hours of audio recorded for the English to German task and 236 hours of audio recorded for the English to French task.
3. VoxPopuli French-English (Wang et al., 2021): French and English segments of the VoxPopuli dataset are taken for translation with 211 and 543 hours of transcribed audio. Same text segments from the dataset are taken for the S2ST task.

For computing the BLEU-c scores, we randomly sampled 250 clips 10 times for each task and tested our system. The reported BLEU-c score is the average of these 10 iterations to ensure a fair and correct representation of our results.

A.4 MOS Calculation Methodology and Protocol

Following the protocol laid out by (Huang et al., 2023), a survey was conducted of 40 respondents, where each respondent was shown the same set of translated clips along with their clips in the source language. This set consisted of 15 voice clips of duration varying from 2 secs to 10 secs. The following questions were asked from the respondents:

1. Rate the similarity of the voice of the speaker to the original source clip : 1 - *Completely Different*, 2 - *Some similarities but more differences*, 3 - *Some differences but more similarities*, 4 - *Perfectly similar*.
2. Rate the quality and naturalness of the generated audio clip: 1 - *Extremely poor / robotic*, 2 - *Somewhat natural but more robotic / poor*, 3 - *Somewhat robotic/poor but more natural*, 4 - *Perfectly natural*.
3. Rate the similarity of the emphasis and intonation of the source clip and synthesised clip: 1 - *Completely Different*, 2 - *Some similarities but more differences*, 3 - *Some differences but more similarities*, 4 - *Perfectly similar*.

Respondents were allowed to rate "exactly in-between" for intermediary cases. It was noted starkly that on the first two questions, a huge proportion of respondents rated the synthesised speech for the five languages as close to 4. However, on the last question, a lot of respondents rated the system between 2 and 3 indicating that while the speaker's voice characteristics and prosody are being transferred with quality, intonation and emphasis will need improvement.

For calculating MOS-v, we employed our MMS TTS without using any voice cloning. Similar surveys on a lesser number of clips were able to see the Vanilla TTS system getting lower ratings compared to CrossVoice on all the three questions. We referenced MOS-h scores from the official paper of (Jia et al., 2022b).

A.5 Results on CVSS-T and IndicTTS

We conducted experiments on 11 languages from the CVSS-T dataset and Hindi from IndicTTS dataset using CrossVoice. Figure 2 shows the results for these 12 languages when translated from any language $X \longrightarrow en$ (English), whereas Figure 3 shows the results for the 12 languages when translated from (English) $en \longrightarrow X$. Notably, our system shows higher BLEU scores on translating from English to any language because of low WER of Whisper on English and NMT being extensively pre-trained on $en \longrightarrow X$ tasks.

For calculating the results, we randomly took samples of 100 clips for each language and calculated results for one sample. We repeated this process for 10 iterations to check for biases. We report the average BLEU score for each language from the experiments. The standard deviation shown on all the tasks ranged between $\in (0.5, 1.5)$, thus, indicating lesser deviation.

A.6 Ethical Considerations

This study has been conducted and tested on standard open source datasets (that are appropriately cited in the paper), widely used in the literature.

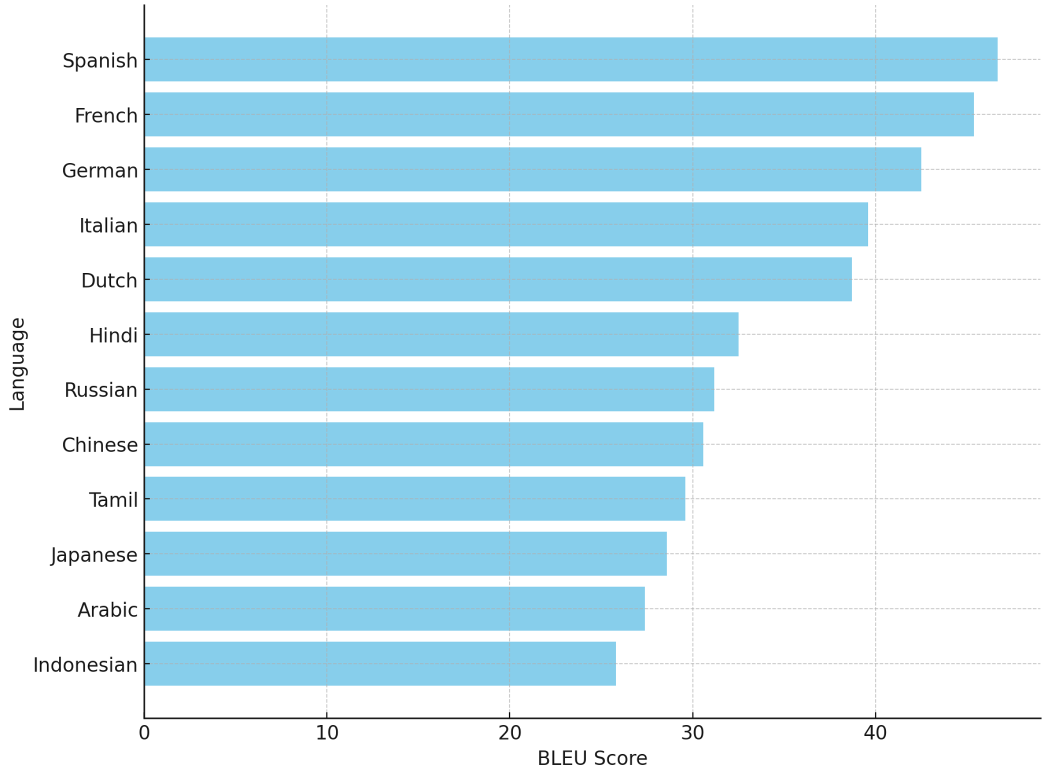

Figure 2: BLEU scores on CVSS-T for X-en Translation

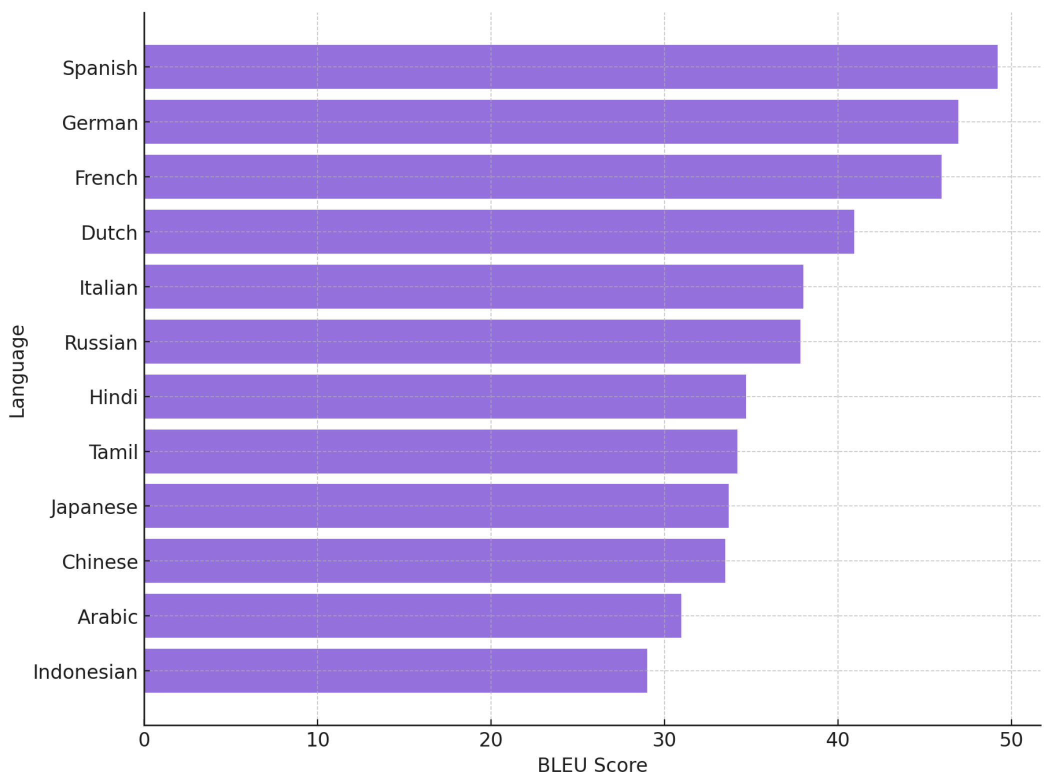

Figure 3: BLEU scores on CVSS-T for en-X Translation

We recognize that voice cloning has the potential to be used for malicious activities; however, the benefits of this technology may outweigh the negatives. Our system is designed to encourage inclusivity and transcend the language barrier in communication between individuals.

Further, we advocate for transparency in the use of voice cloning technology and users should always be informed when they are interacting with a cloned voice.

A.7 Limitations and Challenges

CrossVoice relies heavily on extensive datasets for training. Obtaining and processing large, high-quality, and diverse datasets that cover a wide range of languages and accents is a significant challenge and can limit the system's effectiveness and scalability. CrossVoice encounters challenges in accurately transferring prosody, like intonation and stress patterns, across different languages. This is a complex task due to the inherent differences in linguistic structures and prosodic features among languages. This lack of appropriate transfer of intonation and emphasis is also depicted by the MOS score protocol A.4