IMPROVING AUTOMATED SPEECH RECOGNITION USING RETRIEVAL-BASED VOICE CONVERSION

Anas Mohammed Alhumud, Muhammad AL-Qurishi, Yasser Omar Alomar, Ali Alzahrani & Riad Soussi
Research Center, Elm Company, Riyadh,12382, Saudi Arabia
{alahumud,mualqurishi,yaalomar,rsoussi}@elm.sa

ABSTRACT

This study examines the efficacy of voice conversion techniques in enhancing Automatic Speech Recognition (ASR) accuracy for non-native English speakers. Utilizing the OpenAI Whisper models, we analyzed transcription accuracy across various accents and countries. Significant reductions in Word Error Rates (WER) were observed, with the Whisper Large-v2 model showing the most pronounced improvements. Our findings indicate that advanced voice conversion can mitigate accent bias, promoting inclusivity and broadening the applicability of ASR technology to a more diverse user base.

1 INTRODUCTION

The advent of Automatic Speech Recognition (ASR) technology has revolutionized how humans interact with machines. From dictating texts to controlling smart home devices, ASR systems have become integral to our daily lives. However, despite their widespread use, these systems face significant challenges in accurately recognizing and transcribing speech from non-native English speakers. This discrepancy not only affects the efficiency of technological interaction but also raises concerns about accessibility and inclusivity Radzikowski et al. (2021); Dalmia et al. (2018).

ASR systems, traditionally optimized for native speech, often struggle with the phonetic and prosodic variations presented by non-native accents. This limitation leads to higher word error rates (WER) in transcription, resulting in misunderstandings and a diminished user experience. Previous research highlights the profound impact of accent variation on ASR performance Sisman et al. (2020); Chung et al. (2023), yet solutions to this issue have been limited and often not universally applicable. Addressing this critical gap, our study explores an innovative approach utilizing the Speech Accent Archive Kaggle (2019) alongside Whisper, an advanced ASR system developed by OpenAI Radford et al. (2023), and a Retrieval-based voice conversion technique. We hypothesize that converting non-native speech into a native speaker's voice before transcription can significantly reduce WER, thus enhancing the accuracy and reliability of ASR systems. This hypothesis stems from the assumption that ASR systems are more attuned to native speech patterns, and aligning non-native utterances to these patterns could mitigate recognition errors.

In this context, we incorporate and compare two advanced voice conversion techniques against our ASR-RVC model. First, VQMIVC Wang et al. (2021), an unsupervised method that employs Vector Quantization and Mutual Information to disentangle and manipulate components of speech for voice conversion. Second, a Diffusion-Based Voice Conversion model Popov et al. (2021) that innovatively combines a one-shot many-to-many conversion approach with an average voice encoder and a diffusion-based decoder, employing a Stochastic Differential Equations solver and maximum likelihood sampling for superior performance. Our research aims to not only quantify the improvement in transcription accuracy when applying voice conversion but also to analyze how this improvement varies across different countries and accents.

2 METHODOLOGY

We introduce a simple but effective architecture as shown in Figure 1. The system architecture for improving ASR through Retrieval-Based Voice Conversion involves processing audio inputs from both target (native English speaker) and source (non-native English speaker) through a ContentVec Qian et al. (2022) encoder to extract content vectors. These vectors from the target speaker form a database of target vectors. The HiFiGAN Kong et al. (2020) model, trained on these target vectors, is used to convert the source's voice characteristics to match the target. During inference, our system uses a combination of target vectors and source vectors. Specifically, we employ an index search to retrieve the closest matching target vectors utilizing FAISS Jégou et al. (2022) from the trained set of native English speaker. These vectors are then weighted according to their match score and combined with the source audio's content vectors. The combined features are processed through the HiFiGAN model to generate the output converted waveform that maintains the linguistic content of the source while adopting the voice characteristics of the target. Post voice conversion, we assess transcription accuracy using the Whisper ASR model. This involves transcribing both the original and converted speech samples using all 6 versions of Whisper.

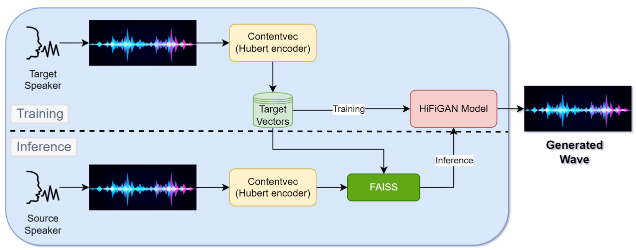


Figure 1: Our system pipeline

3 Results

In our analysis, we found a significant reduction in Word Error Rates (WER) when applying our method to non-native English accents and grouping results by country. The Whisper Large-v2 model exhibited the most substantial performance, with an average reduction in WER of 9.4% when grouping by country and 6.4% by accent. Notably, the maximum reduction in WER reached 72.5% for Large-v2 by country and 59.4% for both Large and Large-v2 by accent, showcasing the model's robustness. Across all models, the improvements affirm the potential of voice conversion technology to enhance ASR systems' inclusivity for a diverse range of speakers (Appendix A.1 and A.2). We compared our method with two distinct voice conversion models: Vector Quantization and Mutual Information-Based Unsupervised Speech Representation Disentanglement for One-shot Voice Conversion (VQMIVC), and Diffusion-Based Any-to-Any Voice Conversion (DiffCV) utilizing the Whisper Large V2. Our model dramatically reduces the Word Error Rate, achieving a WER of 0.0678% compared to 4.205% for VQMIVC and 1.356% for DiffCV, indicating a substantial improvement in accuracy and achieving a CER of 8.8% compared to 64.6% and 27.8%, respectively. To further illustrate our model's superior performance, we present comparisons in Figure 2a and 2b against the top 10 countries and accents where the weakest model ASR-VQMIVC model performs best (Appendix A.1). These comparisons clearly indicate that our model significantly reduces the word error rate across these challenging linguistic scenarios. Further details about our method generalization and limitation can be found in Appendix A.3.

4 Conclusion

The results of our study confirm that voice conversion can substantially mitigate accent bias in Automatic Speech Recognition (ASR) systems, as evidenced by the significant reduction in Word Error Rates (WER) across all tested models. The Whisper Large-v2 model, in particular, has proven to be exceptionally effective, indicating that more advanced models with larger capacities are better suited to handle the phonetic and prosodic variations of non-native English speech. This underscores the importance of continuing to develop and refine ASR technologies that are inclusive of global speech patterns. In conclusion, this study not only advances our understanding of the complexities involved in ASR systems but also opens avenues for more inclusive and universally accessible speech recognition technologies. Future work will focus on refining these voice conversion methods and exploring more hyperparameters in several real-world scenarios, potentially transforming how ASR systems are developed to serve a multilingual and multicultural user base.

URM STATEMENT

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2024 Tiny Papers Track.

REFERENCES

Hyelee Chung, Hosung Nam, et al. Zero-shot voice conversion with hubert. *Phonetics and Speech Sciences*, 15(3):69–74, 2023.

Siddharth Dalmia, Xinjian Li, Florian Metze, and Alan W Black. Domain robust feature extraction for rapid low resource asr development. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 258–265. IEEE, 2018.

Hervé Jégou, Matthijs Douze, Jeff Johnson, Lucas Hosseini, and Chengqi Deng. Faiss: Similarity search and clustering of dense vectors library. *Astrophysics Source Code Library*, pp. ascl–2210, 2022.

Kaggle. Speech accent archive. `https://www.kaggle.com/datasets/rtatman/speech-accent-archive/data`, 2019. Dataset.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033, 2020.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, Mikhail Sergeevich Kudinov, and Jiansheng Wei. Diffusion-based voice conversion with fast maximum likelihood sampling scheme. In *International Conference on Learning Representations*, 2021.

Kaizhi Qian, Yang Zhang, Heting Gao, Junrui Ni, Cheng-I Lai, David Cox, Mark Hasegawa-Johnson, and Shiyu Chang. Contentvec: An improved self-supervised speech representation by disentangling speakers. In *International Conference on Machine Learning*, pp. 18003–18017. PMLR, 2022.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pp. 28492–28518. PMLR, 2023.

Kacper Radzikowski, Le Wang, Osamu Yoshie, and Robert Nowak. Accent modification for speech recognition of non-native speakers using neural style transfer. *EURASIP Journal on Audio, Speech, and Music Processing*, 2021(1):1–10, 2021.

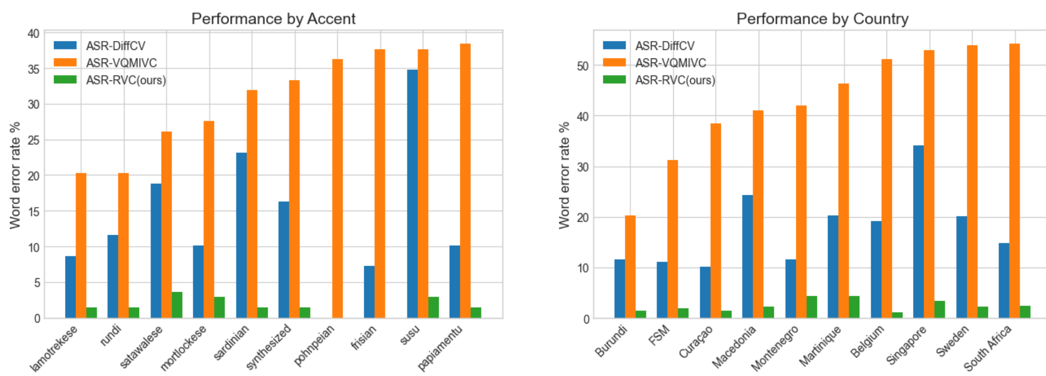
Berrak Sisman, Junichi Yamagishi, Simon King, and Haizhou Li. An overview of voice conversion and its challenges: From statistical modeling to deep learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:132–157, 2020.

Disong Wang, Liqun Deng, Yu Ting Yeung, Xiao Chen, Xunying Liu, and Helen Meng. VQMIVC: Vector Quantization and Mutual Information-Based Unsupervised Speech Representation Disentanglement for One-Shot Voice Conversion. In *Proc. Interspeech 2021*, pp. 1344–1348, 2021. doi: 10.21437/Interspeech.2021-283.

A APPENDIX

A.1 EXPERIMENT

For the training of the voice conversion model, we utilized the Crepe pitch extraction algorithm to preprocess the audio data. The model was trained over few hundreds epochs using a dataset comprising 30 minutes of the target speaker's voice. This training aimed to train the HiFiGAN model to accurately generate waveforms from content representations derived from the target speaker's voice features using ContentVec. During inference, the source speaker's audio is similarly processed using the Crepe algorithm. The audio is encoded using ContentVec to match the learned content representations. FAISS is then employed for vector search, retrieving the nearest vector from the target's database. The matched vector representation is subsequently fed into the HiFiGAN model, which generates the converted audio waveform. The sample rate for both training and inference phases is set at 16000 Hz to ensure consistency and high-quality audio output.Details on the training parameters and are provided in Table 1.



(a) Performance by Accent

(b) Performance by Country

Figure 2: Comparative Analysis of the ASR Performance with three models: ASR-DiffCV, ASR-VQMIVC and Ours (ASR-RVC)

Parameter	Value
GPU	2x3090
Epochs	300
Batch size	32
Seed	1234
Learning rate	1.00E-04
Sampling rate	16000
Filter length	2048
Hop length	480
Win length	2048
Number of mel channels	128
Number of Accent	199
Number of Countries	176

Table 1: The parameters being adopted for training

Models	Min %	Max %	Avg %
Tiny	-0.4	-55.1	-7.3
Base	-0.6	-24.6	-5.3
Small	-0.1	-36.2	-4.3
Medium	-0.1	-18.1	-2.4
Large	-0.2	-59.4	-7.9
Large-v2	-0.1	-59.4	-6.4

Table 2: WER reduction based on accent.

Models	Min %	Max %	Avg %
Tiny	-0.2	-30.5	-4.0
Base	-0.5	-15.4	-3.6
Small	-0.2	-26.8	-4.8
Medium	-0.1	-23.9	-2.8
Large	-0.1	-59.4	-5.3
Large-v2	-0.1	-72.5	-9.4

Table 3: WER reduction based on country.

A.2 Detailed Whisper Model Results

The following tables illustrate the performance of various Whisper models across different accents and countries, showing the Word Error Rate (WER) before and after voice conversion, and the percentage difference.

Table 4: Top 20 accent WER reduction using Whisper Tiny model.

Accent	Direct	Converted	Diff
Agni	97.1	42	-55.1
Edo	82.6	47.8	-34.8
Sundanese	85.5	50.7	-34.8
Nepali	62.2	28	-34.2
Sinhala	87	72.5	-14.5
Ife	29	18.8	-10.1
Nandi	36.2	27.5	-8.7
Filipino	15.2	8	-7.2
Lao	77.3	70.5	-6.8
Lamaholot	10.1	4.3	-5.8
Bambara	34.2	28.7	-5.5
Kru	14.5	10.1	-4.3
Moore	18.8	14.5	-4.3
Teochew	20.3	15.9	-4.3
Khmer	38.3	35.2	-3.1
Hainanese	18.8	15.9	-2.9
Tibetan	39.1	36.7	-2.4
Maltese	15.2	13	-2.2
Ngemba	25.4	23.2	-2.2
Chaldean	15.9	14.5	-1.4

Table 5: Top 20 accent WER reduction using Whisper Base Model.

Accent	Direct	Converted	Diff
Hadiyya	55.1	30.4	-24.6
Uyghur	66.2	45.4	-20.8
Hindi	23.8	8.8	-15.0
Fanti	67.1	52.7	-14.5
Amharic	39.1	28.8	-10.4
Ebira	46.4	37.7	-8.7
Croatian	15.9	7.4	-8.5
Jola	53.6	46.4	-7.2
Kiswahili	21.9	15.1	-6.8
Satawalese	12.3	5.8	-6.5
Taiwanese	47.5	41.5	-6.0
Bamun	23.2	17.4	-5.8
Yakut	11.6	5.8	-5.8
Tajiki	10.1	5.3	-4.8
Baga	46.4	42.0	-4.3
Ashanti	23.2	18.8	-4.3
Sesotho	21.7	17.4	-4.3
Taishan	10.1	5.8	-4.3
Tatar	4.3	0.0	-4.3
Yupik	7.2	2.9	-4.3

Table 6: Top 20 accent WER reduction using Whisper Small Model.

Accent	Direct	Converted	Diff
Jola	79.7	43.5	-36.2
Sylheti	87.0	63.8	-23.2
Bavarian	41.3	21.7	-19.6
Slovak	17.4	4.3	-13.0
Wolof	25.6	16.2	-9.4
Dari	33.9	25.8	-8.1
Hausa	16.4	10.6	-5.8
Somali	27.5	21.7	-5.8
Gedeo	11.6	5.8	-5.8
Kannada	11.6	5.8	-5.8
Tigrigna	24.6	19.4	-5.3
Kurdish	31.2	26.1	-5.1
Greek	14.3	9.3	-5.0
Bamun	17.4	13.0	-4.3
Hainanese	11.6	7.2	-4.3
Kabyle	8.7	4.3	-4.3
Lamaholot	7.2	2.9	-4.3
Taishan	7.2	2.9	-4.3
Tatar	4.3	0.0	-4.3
Arabic	22.7	18.9	-3.9

Table 7: Top 20 accent WER reduction using Whisper Medium Model.

Accent	Direct	Converted	Diff
Xiang	26.8	8.7	-18.1
Lithuanian	16.7	6.8	-9.9
Lao	47.3	41.5	-5.8
Faroese	10.1	4.3	-5.8
Konkani	7.2	1.4	-5.8
Taiwanese	41.7	36.2	-5.4
Mandarin	19.4	15.4	-4.0
Gujarati	13.7	10.6	-3.1
Amazigh	26.1	23.2	-2.9
Burmese	18.8	15.9	-2.9
Sinhala	65.2	62.3	-2.9
Ilonggo	11.6	8.7	-2.9
Kabyle	7.2	4.3	-2.9
Tajiki	7.2	4.8	-2.4
Polish	7.8	5.5	-2.3
Satawalese	9.4	7.2	-2.2
Greek	10.8	8.8	-2.0
Ukrainian	10.5	8.7	-1.8
Romanian	8.0	6.3	-1.7
Korean	16.8	15.3	-1.5

A.3 Generalization and Limitations

Our study also identified limitations in the application of our method to speakers of the native language. Specifically, for English, we found that speakers of Germanic languages exhibited little to no improvement in recognition accuracy. This suggests that our approach may have varying levels of effectiveness depending on the linguistic proximity to the target language. Furthermore, in an attempt to evaluate the generalization of our model across multiple languages, we extended our experiments to include the Arabic language. Participants from various West Asian nationalities,

Table 8: Top 20 accent WER reduction using Whisper Large V1.

Accent	Direct	Converted	Diff
Chichewa	79.7	20.3	-59.4
Bafang	79.7	36.2	-43.5
Basque	75.4	43.5	-31.9
Sylheti	87.0	58.0	-29.0
Kikongo	58.7	33.3	-25.4
Bai	81.2	65.2	-15.9
Xiang	40.6	26.4	-14.1
Mandarin	30.6	17.2	-13.4
Kirghiz	36.2	26.1	-10.1
Somali	35.3	25.6	-9.7
Czech	13.5	5.6	-7.9
Khmer	31.7	24.4	-7.2
Amazigh	31.2	24.6	-6.5
Ukrainian	15.5	9.6	-5.9
Cantonese	17.1	11.3	-5.8
Hausa	13.5	8.2	-5.3
Taiwanese	41.1	36.2	-4.9
Teochew	10.1	5.8	-4.3
Mongolian	19.6	15.5	-4.2
Tigrigna	22.6	18.5	-4.2

Table 9: Top 20 accent WER reduction using Whisper Large V2.

Accent	Direct	Converted	Diff
Chichewa	79.7	20.3	-59.4
Jola	79.7	39.1	-40.6
Mauritian	39.9	3.6	-36.2
Hadiyya	51.4	15.9	-35.5
Burmese	42.0	12.3	-29.7
Malagasy	84.1	58.0	-26.1
Ilonggo	33.3	10.1	-23.2
Igbo	39.6	18.4	-21.3
Malayalam	20.7	0.7	-19.9
Bambara	38.8	26.1	-12.8
Kurdish	42.2	29.6	-12.6
Taiwanese	46.7	34.4	-12.3
Tibetan	45.4	33.3	-12.1
Ukrainian	19.5	8.7	-10.8
Tigrigna	27.4	16.8	-10.5
Mandarin	25.5	15.4	-10.1
Japanese	17.4	7.8	-9.6
Lithuanian	14.5	5.1	-9.4
Croatian	12.0	2.9	-9.1
Bosnian	13.5	4.7	-8.9

Table 10: Top 20 Country WER reduction using Whisper Tiny model.

Accent	Direct	Converted	Diff
Nepal	61.7	31.2	-30.5
Ivory Coast	55.6	36.2	-19.3
Colombia	32.9	21.4	-11.5
Isle Of Man	8.7	2.9	-5.8
Bahrain	7.2	1.4	-5.8
Cambodia	43.7	38.2	-5.6
Trinidad	10.1	5.8	-4.3
Togo	44.2	39.9	-4.3
Slovak Republic	10.6	6.3	-4.3
Liberia	25.1	21.3	-3.9
Ecuador	36.2	33.3	-2.9
The Bahamas	23.2	20.3	-2.9
Niger	7.2	4.3	-2.9
Us Virgin Islands	17.4	14.5	-2.9
Dominican Republic	27.4	24.8	-2.5
Indonesia	22.6	20.4	-2.2
Malta	15.2	13	-2.2
Libya	39.1	37	-2.2
Laos	49.6	47.5	-2
Sri Lanka	36.2	34.3	-1.9

Table 11: Top 20 Country WER reduction using Whisper Base Model.

Country	Direct	Converted	Diff
Croatia	23.5	8.1	-15.4
Tanzania	27.8	17.6	-10.1
United Arab Emirates	25.7	15.8	-10
Ethiopia	32.4	25.5	-7
Bosnia	34.8	30.4	-4.3
Niger	8.7	4.3	-4.3
Lesotho	21.7	17.4	-4.3
Somalia	34.1	30	-4.1
Slovak Republic	8.7	4.8	-3.9
Belarus	19.8	15.9	-3.9
Libya	39.1	35.5	-3.6
India	17.3	13.9	-3.4
Tajikistan	15.6	12.3	-3.3
Nepal	28.7	25.7	-3
Togo	38.4	35.5	-2.9
Cyprus	38.4	36.2	-2.2
Curacao	3.6	1.4	-2.2
Bolivia	21.4	19.8	-1.5
Faroe Islands	13	11.6	-1.4
Haiti	4.3	2.9	-1.4

including India, Pakistan, Sri Lanka, and Bangladesh, were involved in these experiments. While no significant improvements were observed in the Word Error Rate (WER), our method achieved unexpected reductions in the Character Error Rate (CER) with the Whisper Tiny and Whisper Small models, showing decreases of 23% and 33%, respectively. Our approach was particularly adept at recognizing challenging guttural sounds which are commonly misidentified by ASR models. This underscores the potential of voice conversion technology to enhance the performance of ASR systems, especially for models with smaller architectures.

Table 12: Top 20 Country WER reduction using Whisper Small Model.

Country	Direct	Converted	Diff
Libya	57.2	30.4	-26.8
Slovak Republic	27.1	2.9	-24.2
Jordan	57	34.3	-22.7
Cyprus	51.4	29.7	-21.7
Qatar	53.6	33.3	-20.3
Portugal	23.2	8.7	-14.5
Eritrea	28.7	20	-8.7
Egypt	25.1	18.8	-6.3
Somalia	27.5	21.7	-5.8
Israel	13.3	7.9	-5.4
Singapore	8.7	3.6	-5.1
Iraq	28.3	23.8	-4.4
Martinique	10.1	5.8	-4.3
Sri Lanka	30.9	26.6	-4.3
Bolivia	18.6	14.4	-4.2
Algeria	6.5	2.5	-4
Colombia	15.3	11.7	-3.6
Saudi Arabia	24.4	20.9	-3.4
Nicaragua	44	40.8	-3.2
Senegal	27.8	24.8	-3

Table 13: Top 20 Country WER reduction using Whisper Medium Model.

Country	Direct	Converted	Diff
Cyprus	57.2	33.3	-23.9
Lithuania	14.7	7.7	-7
Venezuela	17.7	11.1	-6.6
Egypt	24.2	17.6	-6.5
Martinique	11.6	5.8	-5.8
Faroe Islands	10.1	4.3	-5.8
Tunisia	29	24.2	-4.8
Yemen	10.1	5.8	-4.3
Morocco	15.4	11.5	-4
Taiwan	29.5	25.9	-3.6
Romania	7.8	4.3	-3.4
Montenegro	7.2	4.3	-2.9
China	19.4	16.6	-2.8
Poland	7.8	5.5	-2.3
Ecuador	29.7	27.5	-2.2
Qatar	30.4	28.5	-1.9
Algeria	5.1	3.3	-1.8
South Korea	17.2	15.5	-1.7
Oman	5.8	4.3	-1.4
Madagascar	55.1	53.6	-1.4

Table 14: Top 20 Country WER reduction using Whisper Large V1.

Country	Direct	Converted	Diff
Malawi	79.7	20.3	-59.4
Qatar	49.8	30.4	-19.3
Chile	24.6	6.2	-18.4
Cameroon	34.5	17.9	-16.6
Ivory Coast	36.2	20.8	-15.5
Taiwan	38	25.4	-12.6
Angola	31.3	19.7	-11.6
Jordan	39.1	29.5	-9.7
Somalia	35.3	25.6	-9.7
Cambodia	36.2	27.3	-8.9
Puerto Rico	22.5	14.5	-8
Czech Republic	13.5	5.6	-7.9
DR of Congo	31.6	23.8	-7.8
USA	13.8	6.4	-7.3
Eritrea	26.8	20	-6.8
China	27.3	20.8	-6.5
Kyrgyzstan	20.3	13.8	-6.5
Mongolia	27.5	21.5	-6
Martinique	10.1	4.3	-5.8
Morocco	16.9	11.2	-5.7

Table 15: Top 20 Country WER reduction using Whisper Large V2.

Country	Direct	Converted	Diff
Guatemala	79.7	7.2	-72.5
Malawi	79.7	20.3	-59.4
Cyprus	81.2	31.2	-50
Ivory Coast	59.4	21.3	-38.2
Mauritius	39.9	3.6	-36.2
Libya	57.2	29	-28.3
Jordan	54.6	28	-26.6
Madagascar	84.1	58	-26.1
Qatar	49.8	30	-19.8
Egypt	32.1	16.2	-15.9
Tunisia	39.6	23.7	-15.9
Mali	28.6	12.7	-15.9
Armenia	25.7	11.2	-14.5
Eritrea	33.6	19.3	-14.3
Jamaica	21.4	7.8	-13.6
Angola	32.5	19.1	-13.3
Croatia	16.8	3.5	-13.3
Myanmar	27.9	15.6	-12.3
Ukraine	18.1	7.4	-10.7
Mongolia	31.4	20.8	-10.6