Enhancing Automatic Speech Recognition of a Regional Dialect: A Pilot Study with Québécois French

Xinyi Zhang[*1], **Lucia Eve Berger**[†2], **Duc-Hoa Tran**[‡1], **and Rachel E. Bouserhal**[§1]
[1]École de technologie supérieure, Université du Québec, Canada
[2]MILA, Université de Montréal, Canada

1 Introduction

Automatic speech recognition (ASR) systems have been significantly improved by the implementation of deep neural networks (DNN) in natural language processing. However, the performance of these DNN-based systems largely depends on the amount of training data available [1]. When speech patterns deviate from the norm, they may not be sufficiently represented in the training data, which leads to lower performance in such cases. Therefore, making ASR systems more robust – not only to noise in the environment but also to the *implicit noise* within speech – is becoming more and more important [2]. This study examines such robustness in a state-of-the-art (SOTA) ASR system, Whisper, for the regional dialect of French spoken in Québec, Canada.

1.1 Whisper

Whisper is a multilingual and multitask ASR system developed by Radford et al. [3] with large-scale weak supervision. It uses sequence-to-sequence learning, which directly takes a sequence (in this case, audio) and outputs another sequence (in this case, text). After scaling its training data to 680k hours, Whisper was able to achieve robust zero-shot performance, approaching human-level robustness. Moreover, this excellent performance was attained without using self-supervision and self-training techniques, which are more complex but currently trending in the field.

Whisper is constructed with a transformer encoder-decoder architecture, where the encoder encodes the audio and the decoder decodes the output of the encoder into text. Within each transformer encoder block, there is a self-attention mechanism and multi-layer perceptron; it is also the case for the transformer decoder blocks, except additionally, there is a cross-attention mechanism between the encoder and the decoder.

1.2 Québécois French

Phonological, lexical, and syntactic differences are observed between Québécois French (QF) and Metropolitan French (MF). Reviewed in [4], QF and MF share the same consonant and glide inventories, although differences are seen in the realization of ”r”, the affrication of alveolar plosive before high vowels, and more. For vowels, QF has four more vowels than MF, resulting in fewer homophones. In addition, there is a phenomenon called the diphthongization of long vowels in QF. These movements are easily captured in a spectrogram, which is a time-frequency representation of speech and is usually given as input to neural networks.

In Canada, more than 7 million people speak French natively, and most of them speak QF. A strong demand for a speech-to-text model of QF arises in social services, second language learning for immigrants, smartphone applications, and so on. Despite the existing demand, there needs to be more research in this area. In 2020, Lancien et al. adapted the existing French lexicon, developed a QF-specific pronunciation dictionary, and created an acoustic model [4]. Other efforts made for this task include Gagnon et al. (2008), where a decision-based audiovisual fusion model was created for large-vocabulary speech recognition of French Canadian speech [5]. It is worth exploring the SOTA end-to-end deep neural network for the task of QF recognition.

2 Method

2.1 Data preparation

We searched for publicly available YouTube videos with human-transcribed closed-captioning and selected twelve videos that featured a diverse group of speakers (N = 83) with various sub-regional accents of QF and covered a variety of common topics. The selected videos consisted of casual conversations, interviews, scripted speech, and noises like music and background noise. See Table 1 for detailed statistics.

The audio files were converted into mono WAV format and down-sampled to 16 kHz. The VVT transcripts underwent preprocessing and then were used to extract plain text and timestamps. We used Montreal Forced Aligner (MFA) [6] to obtain precise, word-by-word timestamps. The audio files were then segmented into 5-sec chunks based on word-start and word-end timestamps. Since MFA's output was normalized, the text transcription of each segment was generated by matching MFA's pseudo-truth to the original text.

Table 1: Descriptive statistics of the QF data set videos

	Mean	SD	Sum
Duration (s)	1355.2	501.4	8809.0
Speech ratio	73.0%	2.8%	-
Speaker (F;M)	7.1 (2.4;4.7)	4.0 (1.5;3.8)	83 (29;54)

The audio segments were transformed into 80-channel log-Mel spectrograms that were computed on a 25-millisecond window with a stride of 10 milliseconds, and the text segments were tokenized in French and padded to the same length. We obtained 1568 segments, which were randomly split into three sets: 80% for the training set, 10% for the validation set, and 10% for the testing set.

[*] xinyi.zhang.1@ens.etsmtl.ca
[†] lucia.eve.berger@umontreal.ca
[‡] duc-hoa.tran.1@ens.etsmtl.ca
[§] rachel.bouserhal@etsmtl.ca

2.2 Model Training and Evaluation

We fine-tuned the pre-trained Whisper model of the base and small versions to investigate the impact of model size on the zero-shot and fine-tuned performance. To avoid the potential of overfitting as the training data is limited, we maintained most of the pre-trained weights and only retrained a certain group of layers. We froze the encoder blocks and re-trained the decoder layers. In fine-tuning the base model, there were 52 M trainable parameters and 19.8 M non-trainable parameters. In fine-tuning the small model, there were 153 M trainable parameters and 87 M non-trainable parameters.

We used a learning scheduler, where after a warm-up period, the scheduler linearly decreases the learning rate to zero. We also AdamW optimizer [7] for weight decay regularization and a learning rate finder from Tune [8] to find the optimal initial learning rate. Every model was trained for 10 epochs. We used Python and PyTorch Lightning for our model implementation.

We used Word Error Rate (WER) as our performance evaluation metric, and the performance difference between zero-shot and fine-tuned is denoted as R-WER. Each fine-tuning experiment was run 10 times, each with a different random split of the data set. Means and standard deviations are reported.

3 Results

Table 2: Whisper's WER (%) on benchmarks reported in [3]

	base size	**small size**
Multilingual LibriSpeech	26.6	16.2
Common Voice 9	37.3	22.7
VoxPopuli	24.9	15.7
Fleurs	28.5	15.0
Mean	**29.3**	**17.4**

As shown in Table 2, Whisper's French transcription WER on four benchmarks was averaged to be 29.3% for the base size and 17.4% for the small size [3]. The zero-shot and fine-tuned performances for QF are shown in Table 3. After fine-tuning, the WER is reduced by 13.9% for the base version and 42.9% for the small version.

Table 3: WER (%) of the zero-shot and fine-tuned Whisper models transcribing QF

	zero-shot	**fine-tune**	**R-WER**
base - Mean (SD)	51.3 (6.9)	37.4 (5.5)	13.9 (5.6)
small - Mean (SD)	62.4 (6.8)	19.5 (2.8)	42.9 (8.5)

4 Discussion and Conclusion

There is a clear performance reduction of Whisper when transcribing QF in the zero-shot setting as compared to benchmarks that mostly consist of MF, motivating the need to fine-tune the Whisper model for this regional dialect. The QF zero-shot WER was almost twice as high as reported with MF, demonstrating the effect of the differences between QF and MF. Surprisingly, a higher WER was observed for the small model (62.4%) than the base model (51.3%). This unexpected result is unlikely to be attributed to the limited trial runs as the standard deviation was small and similar for both model sizes, but an explanation for it has not been found. Our fine-tuning reduced the WER by 13.9% for the base model and 42.9% for the small model, approaching Whisper's performance on benchmarks, especially the small model.

In conclusion, our study demonstrated that Whisper has the capability to be adapted for a regional dialect even with limited resources, and it could serve as a pre-trained model to enhance inclusivity and accessibility in voice-based technologies.

Acknowledgments

The authors would like to thank Yimei Yang for her contribution. We acknowledge financial support from the Natural Sciences and Engineering Research Council of Canada as well as the Fonds de recherche du Québec.

References

[1] Hardik B. Sailer, Ankur T. Patil, and Hemant A. Patil. Advances in low resource asr: A deep learning perspective. 2018.

[2] Meredith Moore. Speech Recognition for Individuals with Voice Disorders. In Troy McDaniel and Xueliang Liu, editors, *Multimedia for Accessible Human Computer Interfaces*, pages 115–144. Springer International Publishing, Cham, 2021.

[3] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. Technical report, Tech. Rep., Technical report, OpenAI, 2022.

[4] Mélanie Lancien, Marie-Hélène Côté, and Brigitte Bigi. Developing resources for automated speech processing of quebec french. In *12th Language Resources and Evaluation Conference*, pages 5323–5328. European Language Resources Association, 2020.

[5] L Gagnon, S Foucher, F Laliberte, and G Boulianne. A simplified audiovisual fusion model with application to large-vocabulary recognition of french canadian speech. *Canadian Journal of Electrical and Computer Engineering*, 33(2):109–119, 2008.

[6] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Proc. Interspeech 2017*, pages 498–502, 2017.

[7] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. 2017. Publisher: arXiv Version Number: 3.

[8] Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E Gonzalez, and Ion Stoica. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*, 2018.

[9] Yu Zhang, Daniel S Park, Wei Han, James Qin, Anmol Gulati, Joel Shor, Aren Jansen, Yuanzhong Xu, Yanping Huang, Shibo Wang, et al. Bigssl: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1519–1532, 2022.