Retour d'expérience : Whisper pour les langues régionales

Sam Bigeard, Panagiotis Tsolakis, Emmanuel Vincent, Vincent Colotte, Pascale Erhart, Slim Ouni

Retour d'expérience : Whisper pour les langues régionales

Sam Bigeard[1] Panagiotis Tsolakis[1] Emmanuel Vincent[1] Vincent Colotte[1]
Pascale Erhart[2] Slim Ouni[1]
(1) Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy
(2) Université de Strasbourg, LiLPa UR 1339, F-67000 Strasbourg, France
sam.bigeard@inria.fr

MOTS-CLÉS : Reconnaissance automatique de la parole, langues peu dotées.

KEYWORDS: Automatic speech recognition, low-resourced languages.

1 Contexte et objectifs

Notre objectif est de développer un système de reconnaissance automatique de la parole (ASR) de langues régionales qui soit utile aux locuteurs, pour sous-titrer des contenus audio ou traduire des échanges verbaux, et aux linguistes, pour transcrire et indexer des données de terrain. Pour cela, nous explorons la spécialisation ou l'adaptation de Whisper par affinage (*fine-tuning*).

Distribué par OpenAI, le système d'ASR Whisper (Radford *et al.*, 2022) est capable de transcrire 97 langues et de les traduire vers l'anglais. La connaissance qu'il a acquise des propriétés phonétiques, morphologiques et/ou syntaxiques de ces langues réduit la quantité de données audio transcrites nécessaires à l'affinage, ce qui ouvre la voie à l'intégration de langues peu dotées (Song *et al.*, 2024).

Dans cet article, nous présentons un retour d'expérience sur des travaux en cours dans deux langues : le basque et l'alsacien. Le basque est déjà pris en charge par Whisper, mais avec peu de données d'entraînement (21 h pour l'ASR et 168 h pour la traduction) qui se traduisent par un taux d'erreur sur les mots (WER) médiocre de 38.9% sur Common Voice V15. Sa prononciation et son orthographe normalisées et le volume important de données disponibles pour l'affinage dans Common Voice V18 (356 h issues de 10 707 locuteurs) en font un cas favorable. L'alsacien, en revanche, présente plusieurs difficultés. Il n'est pas présent dans le jeu de données d'entraînement de Whisper, mais il est proche de l'allemand standard qui est l'une des langues les mieux traitées (4ème meilleur WER sur Common Voice). Les jeux de données audio transcrites sont beaucoup plus petits et hétérogènes. En outre, c'est une langue non standardisée dans sa prononciation et son orthographe.

2 Expériences sur le basque

L'objectif est d'affiner le modèle sur une quantité croissante de données, afin de déterminer la quantité nécessaire pour obtenir un WER adapté aux usages. Le WER et le taux d'erreur sur les caractères (CER) sont évalués sur 50 phrases issues de l'ensemble d'évaluation de Common Voice V17. Les transcriptions sont normalisées en retirant la casse et les marques de ponctuation. Nous utilisons les modèles *whisper-base* et *whisper-medium*, qui présentent le meilleur équilibre entre vitesse d'entraînement et WER. Le Tableau 1 montre que moins de 3 h de parole basque suffisent à obtenir un WER inférieur à 30% et un CER de 5%, seuils en-deçà desquels une transcription par ASR suivie d'une correction manuelle est plus rapide qu'une transcription entièrement manuelle. Les expériences se poursuivent avec plus de données afin d'atteindre le seuil de 10% de WER nécessaire à un usage plus large.

Modèle	Affinage	WER	CER
whisper-base	non	94.1%	24.4%
whisper-medium	non	67.0%	13.2%
whisper-large	non	57.2%	12.1%
whisper-base	16 min	58.1%	13.0%
whisper-base	1 h 23	46.1%	9.7%
whisper-medium	2 h 46	28.2%	5.5%

TABLEAU 1 – Évaluation quantitative sur le basque.

En observant les résultats des meilleurs modèles, nous constatons que l'erreur la plus courante est un mauvais découpage des mots. Certaines consonnes semblent concentrer plus d'erreurs, notamment le *h*. Par exemple *Ura ezenuen* devient *Hura ez zenuen*, *Ez dugu laize* devient *Ez dugula ezer*, et *Patsa da* devient *Patxada*.

3 Expériences sur l'alsacien

Concernant l'alsacien, la plus grande partie de notre travail jusqu'ici s'est portée sur la constitution d'un corpus d'entraînement et la prise en compte des particularités de la langue.

Variation des données d'entraînement Peu de corpus oraux transcrits existent. Nous pouvons citer des enquêtes linguistiques : les ethnotextes recueillis dans l'Atlas Linguistique et Ethnographique de l'Alsace (Huck *et al.*, 2014) et Flars (Erhart, 2017). Ces corpus tendent à documenter la variation de la langue, avec des parlers anciens, particuliers à une petite zone géographique, etc. Cette variété peut être un obstacle pour l'apprentissage. Les émissions télévisées, comme la plateforme Panorama Grand Est transcrite par l'association OLCA, sont une autre source de données, dont la langue présente moins de variations. En revanche, elles incluent de la musique et du bruit de fond qu'il faut exclure des données pour les exploiter.

Variation de la langue Une particularité de l'alsacien est la variation orthographique. La graphie Orthal est utilisée par certaines collectivités pour des textes destinés au grand public. Mais elle ne normalise pas les variations locales de prononciation : par exemple la forme écrite *ich* peut être prononcée indifféremment [ɪx], [ɪʃ] ou [ɪç]. En outre, les locuteurs de l'alsacien ne sont pas forcément des scripteurs et, lorsqu'ils sont amenés à écrire, ils ne suivent pas de norme rigide, les usages parlés étant bien plus répandus que les usages écrits. Dans l'objectif de préservation de cette diversité, nous n'avons pas modifié les transcriptions fournies avec les corpus. Pour mieux couvrir cette diversité à l'avenir, nous planifions la création d'un corpus de phrases, transcrites en 5 variantes locales avec un système simple de règles et de dictionnaire, qui permettrait la collecte de données vocales à grande échelle par *crowdsourcing*. Une autre particularité retrouvée dans plusieurs corpus est l'alternance linguistique (*code switching*) fréquente avec le français. Whisper étant multilingue, cela ne constitue pas un obstacle en théorie. Toutefois, cela augmente le risque d'erreur.

Pré-traitements Des pré-traitements non-triviaux et partiellement manuels doivent être effectués pour pouvoir utiliser ces données. Cela inclut l'identification des segments parlés dans les émissions télévisées. Nous utilisons inaSpeechSegmenter (Doukhan *et al.*, 2018) pour cette tâche. Un autre pré-traitement important est le découpage des enregistrements et transcriptions en segments de 30 s maximum requis par Whisper. Pour cette tâche nous avons testé l'utilisation de Whisper non affiné, Audapolis et Montreal Forced Aligner (McAuliffe *et al.*, 2017) (MFA). Nous avons choisi d'utiliser MFA qui s'adapte bien aux mots inconnus mais produit parfois des décalages temporels importants.

Évaluation Pour l'évaluation, nous utilisons un segment d'une émission télévisée d'une durée d'1 min, choisi pour être dépourvu de musique, bruit de fond et parole superposée. Deux modèles *whisper-base* ont été affinés, l'un sur Flars, l'autre sur le corpus des ethnotextes.

Le Tableau 2 présente une évaluation qualitative des modèles, et le Tableau 3 un segment illustratif. Les WER et CER étant médiocres pour tous les modèles, il est difficile d'en tirer des conclusions à ce stade. Néanmoins, nous observons que le CER diminue plus rapidement que le WER, car dans de nombreux cas la différence entre un mot alsacien et le même mot en allemand standard se cantonne à quelques lettres. Les segments d'exemple montrent que les modèles non affinés prédisent des mots plus probables en allemand standard, mais plus éloignés de la réalité phonétique, tandis que les modèles affinés proposent un mot plus proche phonétiquement, bien que toujours incorrect.

Modèle	WER	CER
whisper-base	93.5%	59.8%
whisper-medium	92.2%	57.5%
whisper-base + Flars (9 h 11)	87.0%	37.2%
whisper-base + ethnotextes (3 h)	84.4%	36.1%

TABLEAU 2 – Évaluation quantitative sur l'alsacien.

Modèle	Transcription
vérité terrain	*wìll mer so üstellige gemàcht han*
whisper-base	*ich wollte mich so ausstelligen gemacht haben*
whisper-medium	*wir haben so ausstellungen gemacht ohne*
whisper-base + Flars	*will mir so üstdellige gmàcht han*
whisper-base + ethnotextes	*will mr so üställige gemàcht hàn*

TABLEAU 3 – Évaluation qualitative sur l'alsacien.

4 Conclusion

Quelle que soit la langue, nous observons que Whisper parvient à produire une transcription phonétiquement plausible après affinage sur quelques heures de parole transcrite seulement. En revanche, il peine à produire une transcription orthographiquement correcte par manque de connaissance du vocabulaire de la langue. À quantité égale de données d'affinage, le WER et le CER restent plus élevés pour les langues non standardisées et/ou non vues à l'apprentissage, telles que l'alsacien. Une piste d'amélioration possible est l'utilisation de corpus textuels, lorsqu'ils sont disponibles en plus grande quantité que les corpus oraux, afin d'enrichir ce vocabulaire et améliorer la *tokenisation*. Une mesure du WER et du CER invariante à la variation orthographique apparaît aussi nécessaire pour quantifier les résultats.

Références

DOUKHAN D., CARRIVE J., VALLET F., LARCHER A. & MEIGNIER S. (2018). An open-source speaker gender detection framework for monitoring gender equality. In *2018 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*.

ERHART P. (2017). Les effets de la frontière sur les pratiques linguistiques dans le Rhin supérieur. *Migration(s) et Langues ; Langues et Espace(s)*, **9**.

HUCK D., BOTHOREL-WITZ A., SPINDLER S., BEYER E., MATZEN R. & PHILIPP M. (2014). Atlas linguistique et ethnographique de l'Alsace. DOI : 10.34847/COCOON.A4B78743-D588-3C10-AADE-058FDD69FC46.

MCAULIFFE M., SOCOLOF M., MIHUC S., WAGNER M. & SONDEREGGER M. (2017). Montreal forced aligner : Trainable text-speech alignment using kaldi. In *Interspeech*.

RADFORD A., KIM J. W., XU T., BROCKMAN G., MCLEAVEY C. & SUTSKEVER I. (2022). Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv :2212.04356*.

SONG Z., ZHUO J., YANG Y., MA Z., ZHANG S. & CHEN X. (2024). LoRA-Whisper : Parameter-efficient and extensible multilingual ASR. In *Interspeech*.