Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

Adaptation de modèles de reconnaissance automatique de la parole pour le yiddish, langue sans standard oral

Keming Yi[1] Valentina Fedchenko[1]
(1) ERTIM, Institut national des langues et civilisations orientales (Inalco), 75007 Paris, France
{keming.yi, valentina.fedchenko}@inalco.fr

RÉSUMÉ

Ce travail explore l'adaptation de la reconnaissance automatique de la parole au yiddish dans un contexte marqué par une forte variabilité dialectale et l'absence de variété orale standardisée. Nous évaluons l'efficacité de modèles multilingues récents ainsi que différentes stratégies d'adaptation, en particulier celles liées à la sélection et à la présentation des données d'entraînement, afin de transcrire cette diversité linguistique. Les résultats mettent en évidence l'impact des stratégies de présentation des données dialectales lors de l'affinage de modèles préentraînés, ainsi que des ressources limitées disponibles, sur les performances globales du système. Cette recherche contribue ainsi au développement de modèles de reconnaissance de la parole plus robustes et inclusifs, capables de résister aux contextes multidialectaux.

ABSTRACT

Adapting Automatic Speech Recognition Models to Yiddish, a Language without Oral Standard

This work explores the adaptation of Automatic Speech Recognition to Yiddish in a context characterized by high dialectal variability and the absence of a standardized spoken variety. We evaluate the effectiveness of recent multilingual models as well as various fine-tuning strategies, with particular attention to data selection and presentation, in order to transcribe this linguistic diversity. The results highlight how different strategies for presenting dialectal data during the fine-tuning of pretrained models, together with the limited available resources, affect overall system performance. This study thus contributes to the development of more robust and inclusive speech recognition models capable of handling highly multidialectal contexts.

MOTS-CLÉS : yiddish, reconnaissance automatique de la parole, langue peu dotée, variabilité dialectale, adaptation de modèles acoustiques.

KEYWORDS: Yiddish, Automatic Speech Recognition, Low-resource language, Dialectal variability, Acoustic model adaptation.

1 Introduction

Bien que les avancées récentes en apprentissage profond aient considérablement amélioré les performances de la reconnaissance automatique de la parole (RAP), ces progrès profitent majoritairement aux langues bien dotées. L'adaptation des modèles de RAP aux langues peu dotées se heurte à deux obstacles majeurs : la forte variabilité dialectale, qui perturbe la modélisation acoustique, et le code-switching, particulièrement prévalente chez les locuteurs de ces langues souvent multilingues. Actuellement, les modèles multilingues à l'état de l'art, tels que Whisper (Radford *et al.*, 2022)

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 1 : articles scientifiques originaux, pages 53–64.
Cette œuvre est mise à disposition sous licence Attribution 4.0 International.

ou MMS (Pratap *et al.*, 2023), peinent encore à généraliser face à ces phénomènes lorsqu'ils sont confrontés à des variétés sous-représentées dans leurs corpus d'entraînement.

Le yiddish constitue un cas d'étude idéal pour explorer cette problématique. Langue germanique enrichie d'éléments hébraïques et slaves (Jacobs, 2005), elle se caractérise par une riche diversité dialectale et un manque critique de ressources numériques standardisées. De plus, il n'existe actuellement que très peu de travaux récents consacrés à la RAP pour cette langue.

Cet article vise donc à évaluer et à adapter des modèles multilingues récents (notamment Whisper et MMS) afin de mesurer leur capacité d'adaptation à une langue peu dotée présentant une forte variabilité dialectale. L'article est structuré comme suit : la section 2 introduit brièvement la langue yiddish ainsi que les avancées récentes en RAP pour les langues peu dotées. La section 3 présente les corpus mobilisés pour nos expérimentations, tandis que la section 4 détaille notre méthodologie et nos résultats. Enfin, la section 5 vient clore ce travail par une synthèse de nos résultats.

2 Contexte

Le yiddish est une langue germanique occidentale historiquement parlée par les Juifs ashkénazes, née au IXe siècle en Europe centrale (Jacobs, 2005). Bien que la population de locuteurs ait drastiquement diminué lors de l'Holocauste, elle compte aujourd'hui environ 600 000 locuteurs [1]. Une caractéristique majeure du yiddish est sa très forte variabilité dialectale.

La langue se divise fondamentalement entre le yiddish occidental et le yiddish oriental, avec des subdivisions nord et sud reposant sur d'importantes différences phonologiques, comme montré dans la figure 1. La principale division dialectale du yiddish est basée sur le système des voyelles accentuées. La méthode de classification la plus largement acceptée se fonde sur la prononciation de deux voyelles du Proto-yiddish (*ei_{24} et *ou_{44}) (Baviskar *et al.*, 1992). En prenant les mots « acheter » (*kojfn*$_{44}$) et « viande » (*flejš*$_{24}$) du yiddish standard comme exemples (Jacobs, 2005), cette distinction devient évidente. Concrètement, en yiddish de l'ouest (WY), ces termes se prononcent respectivement *ka :fn* et *fla :š*. Au sein des variantes orientales, le yiddish central (CY) utilise les formes *kojfn* et *flajš*, le yiddish du sud-est (SEY) se distingue par *kojfn* et *flejš*, tandis que le yiddish du nord-est (NEY) adopte les prononciations *kejfn* et *flejš*. De plus, il n'existe pas de norme orale unique, le yiddish parlé au quotidien reposant sur diverses variétés dialectales. À l'écrit, bien que le yiddish emploie traditionnellement l'alphabet hébraïque, le système de romanisation standardisé par l'Institut YIVO s'est imposé comme la norme de référence pour en faciliter la diffusion. Ce cadre normatif, qui privilégie une transcription phonétique cohérente, est celui que nous avons adopté pour les transcriptions au sein de nos expériences.

La dernière recherche spécifiquement dédiée à la RAP en yiddish est celle de (Ćavar *et al.*, 2016). Les auteurs y ont entraîné un aligneur forcé pour le yiddish et développé, à partir de celui-ci, un système de RAP dans le cadre du projet AHEYM [2]. Cependant, ils n'ont pas précisé les performances de ce système, et le site hébergeant les modèles n'est plus maintenu. Ainsi, nous avons consulté des travaux portant sur la RAP pour d'autres langues, en particulier celles qui sont peu dotées.

Face à un corpus de données limité, l'affinage de modèles préentraînés s'est imposé comme la straté-

1. https://jewishstudies.rutgers.edu/component/content/article/159-yiddish-faqs?Itemid=100120&catid=102#today
2. https://aheym.com

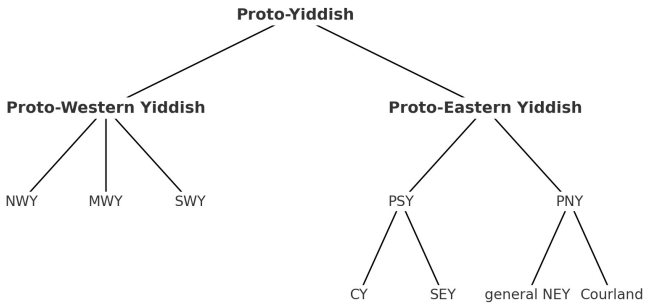

FIGURE 1 – Modèle arborescent des principaux dialectes du yiddish (Jacobs, 2005)

NWY : yiddish du nord-ouest ; **MWY :** yiddish du centre-ouest ; **SWY :** yiddish du sud-ouest ; **PSY :** proto-yiddish du sud ; **PNY :** proto-yiddish du nord ; **CY :** yiddish central ; **SEY :** yiddish du sud-est ; **NEY :** yiddish du nord-est.

gie de référence, évitant l'entraînement d'un système à partir de zéro ou le recours à des architectures coûteuses (Luo *et al.*, 2021; Ridoy *et al.*, 2025; Toyin *et al.*, 2023; Djanibekov *et al.*, 2025). Lors de l'apprentissage séquentiel de nouveaux dialectes, le modèle fait face au défi de l'oubli catastrophique. Pour y pallier, des méthodes d'apprentissage continu, telles que l'Experience Replay (ER) (Della Libera *et al.*, 2024) et l'optimisation ciblée du décodeur (Kwok *et al.*, 2024), se sont révélées efficaces. Enfin, sur le plan de l'efficacité calculatoire, l'intégration d'adaptateurs de faible rang comme LoRA (Hu *et al.*, 2022) permet d'adapter de grands modèles à de nouvelles langues tout en réduisant drastiquement les coûts en mémoire et en temps de calcul (Liu *et al.*, 2024).

3 Données d'entraînement

Notre expérience mobilise quatre corpus. Le premier est REYD (Webber *et al.*, 2022), qui comprend environ 8 heures de lecture de livres audio. Les données sont réparties entre trois locuteurs distincts, correspondant aux sous-corpus *pol1*, *lit1* et *lit2*. Les statistiques détaillées de ce corpus sont présentées dans la table 1. Comme la transcription utilise initialement l'alphabet hébraïque, nous avons procédé à sa romanisation YIVO.

Deuxièmement, nous avons intégré le sous-ensemble yiddish du jeu de données Common Voice [3] (CV), en nous appuyant sur la version du 25 juin 2025. Afin de garantir la qualité des données, nous n'avons conservé que les segments validés, ce qui représente une durée totale de seulement 48 minutes d'audio. Ce matériel est principalement représentatif du dialecte yiddish hassidique, une variété située entre le yiddish occidental et le yiddish oriental. De la même manière, les transcriptions ont également fait l'objet d'une procédure de romanisation YIVO.

Troisièmement, nous avons mobilisé le projet CSYE (Bleaman & Nove, 2025) spécifiquement pour notre évaluation hors domaine. Ce corpus présente plusieurs caractéristiques singulières : il comporte du code-switching (principalement entre le yiddish et l'anglais) et se compose d'enregistrements d'entretiens spontanés, ce qui le distingue des contenus narratifs des autres corpus. De plus, il couvre trois dialectes du yiddish : CY, NEY et SEY. Nous avons rassemblé un total de 3 heures d'audio pour la phase de test, en extrayant les données de la première cassette de six entretiens (deux pour

3. https://commonvoice.mozilla.org/en

Sous-corpus	Locuteur	Genre	Durée
pol1 (CY)	Zylberberg	Homme	2 h 36 m
lit1 (NEY)	Blacher-Retter	Femme	2 h 42 m
lit2 (NEY)	Rubinov	Homme	2 h 36 m
Total			**7 h 54 m**

TABLE 1 – Détails du corpus REYD

Dialecte	Locuteur	Genre	Durée
CY	(Anafi, 1996)	Femme	30 m
	(Burekhovich, 1995)	Femme	30 m
NEY	(Abramson, 1995)	Homme	30 m
	(Dimantstein, 1996)	Homme	30 m
SEY	(Kopolovicz, 1998)	Femme	30 m
	(Luks, 1996)	Femme	30 m
Total			**3 h**

TABLE 2 – Détails du corpus CSYE

chaque dialecte), celle-ci offrant la plus forte densité de passages en code-switching. Les détails sont présentés dans la table 2.

Enfin, pour examiner si l'ajout de données d'entraînement en anglais améliore les performances des modèles sur le corpus CSYE, nous avons collecté 10 heures d'audio issues du corpus LibriSpeech (Panayotov *et al.*, 2015). Cette augmentation de données permet de tester la robustesse des modèles dans des contextes du code-switching.

Au total, le corpus comprend environ 22 heures d'audio, dont environ 9 heures exclusivement en yiddish, 3 heures contenant des segments en code-switching (principalement en yiddish et en anglais), et 10 heures uniquement en anglais.

Il convient de noter que l'ensemble des fichiers audio a été normalisé à une fréquence d'échantillonnage de 16 kHz. Parallèlement, les transcriptions ont subi un prétraitement consistant à convertir tous les caractères en minuscule et à supprimer la ponctuation. Nous avons réparti chaque corpus en ensembles d'entraînement (0,8), de validation (0,1) et de test (0,1) en fixant le random seed à 42. Pour les expériences, différents corpus ont été combinés, mais les jeux de test sont restés identiques pour assurer la comparabilité des résultats.

4 Méthode & Résultats

Cette section est consacrée à nos expériences ainsi qu'aux résultats des expériences. Elle débute par l'inférence des modèles préentraînés, suivie de sept expériences menées avec différentes configurations.

4.1 Inférence des modèles préentraînés

Avant de débuter les expérimentations, nous avons comparé les performances d'inférence de plusieurs modèles sur le corpus REYD afin de sélectionner ceux qui feraient l'objet d'un affinage lors des étapes suivantes (voir table 3). L'ensemble des expériences a été exécuté sur un GPU NVIDIA L40S doté de 45 Go de mémoire.

Pour les modèles auto-supervisés, nous avons utilisé le corpus REYD dans sa romanisation YIVO, à l'exception du modèle wav2vec2-xls-r-300m-hebrew. Il s'avère que MMS offre les performances

Modèle	WER	CER	Temps d'inférence
Modèles auto-supervisés			
facebook/wav2vec2-large-960h-lv60-self	1,0186	0,8909	53 s
facebook/wav2vec2-large-xlsr-53-german	0,9016	0,4373	**49 s**
imvladikon/wav2vec2-xls-r-300m-hebrew	0,9998	0,6631	1 m 36 s
facebook/mms-1b-all	**0,8605**	**0,3342**	1 m 20 s
facebook/hubert-large-ls960-ft	1,0105	0,9059	47 s
Modèles supervisés			
openai/whisper-large	2,2278	1,6946	8 h 39 m
openai/whisper-large-v3	**0,9537**	**0,5393**	3 h 56 m
openai/whisper-large-v3-turbo	1,0537	0,9960	**17 m 54 s**

TABLE 3 – Comparaison des performances d'inférence des différents modèles sur le corpus REYD

les plus élevées. Toutefois, le modèle wav2vec2-xls-r-300m présente, à l'instar de MMS, l'avantage d'avoir été préentraîné sur un corpus intégrant déjà du yiddish.

Pour ces raisons, nous avons choisi de conserver ces deux modèles, mms-1b et wav2vec2-xls-r-300m, afin d'évaluer plus finement leurs performances et leurs temps d'entraînement.

En ce qui concerne les modèles supervisés, Whisper est déjà entraîné sur le yiddish dans son écriture hébraïque, ce qui permet d'indiquer explicitement la langue avant d'effectuer la transcription. Toutefois, cette langue ne figure pas parmi celles recommandées par *OpenAI*, c'est-à-dire celles pour lesquelles le WER est inférieur ou égal à 50 %. Dans ces inférences, nous avons utilisé la version originelle du corpus REYD en lettres hébraïques.

Nous pouvons observer que whisper-large-v3 surpasse nettement whisper-large en termes de performance et de rapidité d'inférence. Cependant, le temps d'inférence des modèles Whisper diffère considérablement de celui des modèles basés sur wav2vec 2.0.

Quoi qu'il en soit, aucun modèle n'est réellement capable de reconnaître le yiddish. À la suite de ces comparaisons, nous avons décidé d'affiner trois modèles : wav2vec2-xls-r-300m, mms-1b et whisper-large-v3.

4.2 Expérience A1 : affinage sur le corpus REYD

Dans un premier temps, nous affinons les modèles sur le corpus REYD (comportant les dialectes NEY et CY) afin d'évaluer leur capacité de généralisation face au corpus hors-domaine Common Voice (yiddish hassidique). L'évaluation s'étend ensuite au corpus CSYE, dont la complexité réside dans le code-switching et le décalage de domaine, passant de la lecture de livres audio à des entretiens spontanés.

Concernant l'implémentation des modèles wav2vec2-xls-r-300m et mms-1b, nous avons ajouté une couche de projection linéaire associée à une fonction de perte CTC (Graves *et al.*, 2006). La construction du tokeniseur repose sur un vocabulaire global extrait du corpus, où le symbole « | » sert de séparateur lexical explicite. Les tokens spéciaux `[UNK]` et `[PAD]` ont été intégrés pour la gestion des caractères inconnus et le fonctionnement de la CTC. Enfin, l'extracteur de caractéristiques CNN a

Modèle	REYD (jeu de test)		CV		Moyenne		CSYE								Durée d'entraînement	Utilisation GPU (Go)
							NEY		CY		SEY		Moyenne			
	WER	CER	WER	CER	WER	CER	WER	CER	WER	CER	WER	CER	WER	CER		
wav2vec2-xls-r-300m	0,1118	0,0261	0,6437	0,2441	0,3778	0,1351	0,5418	0,2265	0,7580	0,3686	0,7085	0,3225	0,6802	0,3127	1 h 3 m	22
mms-1b	**0,0964**	**0,0237**	0,6357	0,2458	0,3661	0,1348	0,5278	**0,2258**	0,7672	0,3871	0,7084	0,3303	0,6799	0,3224	1 h 27 m	22
whisper-large-v3	0,1065	0,0370	**0,5526**	**0,2031**	**0,3296**	**0,1201**	**0,4861**	0,2263	**0,6713**	**0,3437**	**0,5920**	**0,2928**	**0,5936**	**0,2938**	3 h 6 m	33

TABLE 4 – Résultats de l'expérience A1

été figé afin de préserver les représentations acoustiques robustes acquises durant le préentraînement.

Pour la sélection des hyperparamètres, nous avons utilisé le modèle wav2vec2-xls-r-300m comme base de test. Après des tests, nous avons fixé le batch size à 4, le nombre d'époques à 50 (avec un early stopping de 5 de patience), le warmup ratio à 0,1 et le weight decay à 0,01. Concernant le learning rate, nos essais ont montré qu'une valeur de 5×10^{-5} offrait une meilleure convergence et des performances supérieures à 1×10^{-5}. Ces réglages ont été maintenus pour le modèle mms-1b.

En ce qui concerne whisper-large-v3, nous avons configuré le modèle pour la tâche de *transcribe* en yiddish. Bien que Whisper ait été initialement entraîné sur l'alphabet hébraïque, nous avons choisi d'utiliser la translittération latine YIVO afin de garantir la cohérence avec les modèles auto-supervisés et de faciliter la comparaison des résultats. Nos expérimentations confirment que le modèle s'adapte efficacement aux caractères latins sans être pénalisé par son entraînement d'origine. Pour Whisper, nous avons toutefois observé une tendance inverse concernant le learning rate : une valeur plus faible (1×10^{-5}) s'est révélée plus performante que 5×10^{-5}.

Analyse des résultats : La table 4 synthétise les performances des trois modèles sur les différents jeux de données. Pour le corpus CSYE, en raison du code-switching, la langue de génération de Whisper n'a pas été fixée au yiddish.

Nous observons que les trois modèles affichent des performances comparables sur le jeu de test du corpus REYD, confirmant l'efficacité de l'affinage pour la reconnaissance du yiddish. Néanmoins, whisper-large-v3 se distingue par une supériorité marquée sur le corpus hors-domaine Common Voice. Cette performance s'explique probablement par l'exposition massive du modèle à diverses données multilingues lors de son pré-entraînement. En contrepartie, l'utilisation de Whisper nécessite des ressources GPU plus importantes et un temps d'entraînement nettement plus long que les modèles MMS ou wav2vec 2.0.

Analyse phonémique qualitative : Afin d'approfondir l'évaluation, nous avons analysé les erreurs phonémiques générées par les modèles. Sur le corpus hors domaine (Common Voice), les résultats révèlent que les erreurs se concentrent massivement sur les voyelles. Les trois modèles présentent des profils de confusion phonémique très similaires, caractérisés par deux tendances majeures : une forte perturbation liée à l'alternance allophonique intradialectale (notamment la fluctuation entre [i] et [e]) et une vulnérabilité à la variabilité interdialectale (entraînant des confusions récurrentes telles que [ey] $\rightarrow$ [ay], [ay] $\rightarrow$ [a], ou [i] $\leftrightarrow$ [u]), comme mentionné dans la section 2. Ces similarités suggèrent l'existence d'une limite commune dans leur capacité de généralisation aux différents dialectes du yiddish, indépendamment du paradigme d'apprentissage employé.

En contexte du corpus en domaine (REYD), bien que les performances globales soient nettement plus solides, un noyau d'erreurs communes persiste. Les confusions entre [i] et [e] ainsi que la simplification de certaines diphtongues (comme [ey] $\rightarrow$ [ay] et [ay] $\rightarrow$ [a]) restent fréquentes. Au-delà de quelques spécificités propres à chaque architecture (Whisper gère mieux la consonne [r] mais

Modèle	REYD+CV (jeu de test)		CV		CSYE								Durée d'entraînement	Utilisation GPU (Go)
					NEY		CY		SEY		Moyenne			
	WER	CER	WER	CER	WER	CER	WER	CER	WER	CER	WER	CER		
wav2vec2-xls-r-300m	0,1809	0,0437	0,3292	0,0989	0,6102	0,2468	0,7941	0,3707	0,7474	0,3323	0,7266	0,3225	1 h 20 m	22
mms-1b	0,1286	**0,0334**	0,3034	0,0924	0,5539	0,2404	0,7113	**0,3458**	0,6774	**0,3172**	0,6553	**0,3060**	1 h 36 m	22
whisper-large-v3	**0,1165**	0,0410	**0,2757**	**0,0861**	**0,4405**	**0,2069**	**0,6350**	0,3545	**0,6722**	0,3797	**0,5896**	0,3180	3 h 48 m	33

TABLE 5 – Résultats de l'expérience A2

peine sur les voyelles [oy] et [o], tandis que MMS montre des faiblesses sur les consonnes [m] et [p]), les résultats confirment que l'ambiguïté des frontières acoustiques entre voyelles voisines demeure le défi principal et commun à l'ensemble des systèmes évalués.

4.3 Expérience A2 : affinage conjoint sur le corpus REYD + Common Voice

À partir de cette expérience, sauf indication contraire, nous conservons les hyperparamètres et la configuration des modèles établis lors de l'expérience A1.

Conformément à (Della Libera *et al.*, 2024), nous utilisons l'affinage conjoint pour atteindre la limite supérieure d'adaptation. Notre but est de concevoir des modèles spécifiquement dédiés au yiddish, couvrant les dialectes NEY, CY et yiddish hassidique. Pour ce faire, nous affinons les modèles sur la totalité des corpus de type « monologue », en excluant rigoureusement le code-switching ainsi que les perturbations acoustiques (bruit de fond, chevauchements). Nous évaluons les modèles sur la combinaison des jeux de test REYD et Common Voice. Compte tenu de la taille du jeu de test Common Voice (seulement 5 minutes d'audio), nous avons choisi d'évaluer également ce corpus dans son intégralité de manière séparée.

Analyse des résultats : La table 5 indique une progression globale des trois modèles sur Common Voice par rapport à A1. Néanmoins, l'écart marquant de performance entre le corpus REYD et le corpus Common Voice met en évidence l'impact négatif du déséquilibre des données d'entraînement.

Par ailleurs, sur le corpus CSYE, si mms-1b et whisper-large-v3 affichent une légère amélioration, wav2vec2-xls-r-300m subit une baisse de performance. Ce dernier a par conséquent été écarté des expérimentations ultérieures.

4.4 Expérience A3 : affinage conjoint sur le corpus REYD + Common Voice (version équilibrée)

Afin de pallier le déséquilibre interdialectal observé précédemment, cette expérience évalue l'impact d'un équilibrage strict des données d'entraînement. Nous avons restreint chaque dialecte à une durée d'environ 48 minutes, correspondant au volume disponible pour le yiddish hassidique (Common Voice). Ainsi, pour le corpus REYD, 48 minutes ont été extraites pour le dialecte CY et 48 minutes pour le NEY (réparties équitablement entre les sous-corpus *lit1* et *lit2*).

L'objectif est double : déterminer si cette symétrie favorise la généralisation du modèle, et mesurer l'impact d'une réduction drastique du volume global d'entraînement. Afin d'établir une comparaison directe avec l'expérience A2, l'évaluation est maintenue sur les mêmes jeux de test REYD et Common

Modèle	REYD+CV (jeu de test)		CV		Durée d'entraînement	Utilisation GPU (Go)
	WER	CER	WER	CER		
mms-1b (A2)	**0,1286**	**0,0334**	0,3034	0,0924	1 h 36 m	22
mms-1b (A3)	0,3354†	0,0928	**0,2511**†	**0,0720**	46 m	22
whisper-large-v3 (A2)	**0,1165**	**0,0410**	0,2757	**0,0861**	3 h 48 m	33
whisper-large-v3 (A3)	0,1971†	0,0601	**0,2332**†	0,1151	2 h 27 m	33

† Indique une différence statistiquement significative par rapport à A2 ($p < 0,05$) pour le WER.

TABLE 6 – Comparaison des résultats de l'expérience A2 avec l'expérience A3

Voice.

Analyse des résultats : La table 6 compare les résultats de cette expériences avec ceux d'A2, et nous pouvons y observer une nette amélioration des performances sur le corpus Common Voice. À l'inverse, les résultats déclinent sur REYD. Ce compromis indique que la réduction et l'équilibrage des données d'entraînement profitent aux dialectes les moins dotés, mais s'opèrent au détriment de ceux disposant initialement d'un volume plus important.

4.5 Expérience A4 : affinage conjoint sur le corpus REYD + Common Voice + Librispeech

Dans cette expérience, l'anglais est ajouté au jeu de données d'entraînement afin d'évaluer, lors du test sur le corpus CSYE et de la comparaison avec les résultats de l'expérience A2, si les modèles conservent leurs capacités en anglais ou s'ils présentent un phénomène d'oubli.

Modèle	REYD+CV (jeu de test)		CSYE								Durée d'entraînement	Utilisation GPU (Go)
			NEY		CY		SEY		Moyenne			
	WER	CER	WER	CER	WER	CER	WER	CER	WER	CER		
mms-1b (A2)	**0,1286**	**0,0334**	0,5539	0,2404	0,7113	0,3458	0,6774	0,3172	0,6553	0,3060	1 h 36 m	22
mms-1b (A4)	0,1410†	0,0342	**0,5369**	**0,2212**	**0,7022**	**0,3332**	**0,6516**	**0,2923**	**0,6389**†	**0,2879**	2 h 11 m	22
whisper-large-v3 (A2)	**0,1165**	**0,0410**	**0,4405**	**0,2069**	**0,6350**	**0,3545**	**0,6722**	**0,3797**	**0,5896**	**0,3180**	3 h 48 m	33
whisper-large-v3 (A4)	0,1369†	0,0461	0,7531	0,5586	0,8239	0,5873	0,6901	0,4761	0,7631†	0,5461	4 h 32 m	32

† Indique une différence statistiquement significative par rapport à A2 ($p < 0,05$) pour le WER.

TABLE 7 – Comparaison des résultats de l'expérience A2 avec l'expérience A4

Analyse des résultats : Afin d'évaluer la reconnaissance spécifique du yiddish, le corpus anglophone LibriSpeech a été exclu des évaluations de jeu de test. Pour le modèle whisper-large-v3, la langue de génération a été fixée au yiddish sur le jeu de test monolingue, mais laissée libre sur le corpus CSYE.

Le table 7, qui compare les expériences A2 et A4, montre une baisse des performances globales sur le jeu de test. Sur le corpus du code-switching CSYE, l'ajout de l'anglais profite à mms-1b, mais pénalise lourdement whisper-large-v3. Cette chute brutale ne traduit pas une faiblesse intrinsèque

du modèle étant donné que Whisper surpasse d'ailleurs MMS sur les données purement yiddish de l'expérience A2, mais résulte d'une confusion linguistique. Sans contrainte explicite, le décodeur de Whisper tend à interpréter certains segments comme appartenant à d'autres langues (arabe, hébreu ou russe).

En somme, l'augmentation des données anglophones ne se traduit pas par une amélioration systématique pour le yiddish, bien qu'elle s'avère bénéfique pour le code-switching dans le cas de MMS. Pour Whisper, les difficultés observées face au code-switching relèvent davantage d'une instabilité de l'identification de la langue lors du décodage que d'un oubli des connaissances en anglais.

4.6 Expériences A5-A7

Expérience A5 :affinage simple du modèle d'A1 sur le corpus Common Voice

(Della Libera *et al.*, 2024) indique que l'affinage simple représente la limite inférieure de performance.

Cette expérience a donc pour objectif d'affiner le modèle issu de l'expérience A1 sur le corpus Common Voice, afin de le comparer aux approches d'affinage avec Experience Replay (expérience A6) et d'adaptation par adaptateurs (expérience A7), et de vérifier si les méthodes proposées par (Della Libera *et al.*, 2024) permettent d'atténuer l'oubli catastrophique dans le cas du yiddish. Les résultats des expériences A5-7 seront comparés ensemble.

Expérience A6 : Experience Replay du modèle d'A1 sur le corpus Common Voice

Cette méthode (ER) est considérée comme la plus efficace pour atténuer l'oubli catastrophique, selon (Della Libera *et al.*, 2024). Ainsi, les modèles issus de l'expérience A1 sont repris, en ajoutant 10 % des données du corpus REYD au corpus Common Voice, avant d'affiner à nouveau les deux modèles sur ce jeu de données.

Expérience A7 : affinage du modèle avec adaptateurs

Dans cette expérience, nous évaluons l'affinage par adaptateurs. Pour le modèle MMS (point de contrôle mms-1b-all), nous utilisons son propre adaptateur linguistique (Pratap *et al.*, 2023). Toutes les couches du modèle sont figées, à l'exception de la couche d'adaptateur, qui est la seule mise à jour lors de l'entraînement. Lors de la phase de réglage, nous avons constaté qu'avec un learning rate fixé à 1×10^{-3}, une batch size de 4 permettait d'atteindre de meilleures performances qu'une batch size de 8.

Pour whisper-large-v3, la méthode LoRA (Hu *et al.*, 2022) est appliquée selon le même principe. En fixant le nombre d'époques à 50, la batch size à 8 et le rank à 32, nous avons observé qu'un learning rate de 5×10^{-4} s'est avéré nettement plus performant qu'une valeur plus conservatrice de 1×10^{-4}.

Contrairement aux travaux de (Pratap *et al.*, 2023) et (Liu *et al.*, 2024) qui conçoivent un adaptateur par langue, nous n'avons pas entraîné d'adaptateur distinct pour chaque variante dialectale du yiddish. En effet, comme l'a démontré l'expérience A3, la fragmentation d'un jeu de données déjà restreint dégrade significativement les performances. Par conséquent, les deux modèles ont été affinés conjointement sur la combinaison des corpus REYD et Common Voice (identique à la configuration de l'expérience A2).

Analyse des résultats : Prenant les modèles de l'expérience A2 comme référence (table 8), l'analyse des expériences A5 et A6 confirme l'impact de l'oubli catastrophique. Si l'affinage simple (A5)

Modèle	REYD+CV (jeu de test)		CV		Durée d'entraînement	Utilisation GPU (Go)
	WER	CER	WER	CER		
mms-1b (A2)	0,1286	0,0334	**0,3034**	**0,0924**	1 h 36 m	22
mms-1b (A5)	0,1837†	0,0434	0,3200†	0,1001	19 m	22
mms-1b (A6)	**0,1045**†	**0,0266**	0,3649†	0,1147	22 m	22
mms-1b-all (A7)	0,2152†	0,0499	0,4108†	0,1275	1 h 6 m	22
whisper-large-v3 (A2)	0,1165	0,0410	0,2757	0,0861	3 h 48 m	33
whisper-large-v3 (A5)	0,1733†	0,0534	**0,2406**†	**0,0708**	1 h 33 m	33
whisper-large-v3 (A6)	**0,1035**†	**0,0372**	0,2714†	0,0883	1 h 21 m	33
whisper-large-v3 (A7)	0,1142†	0,0413	0,2554†	0,0821	4 h 27 m	39

† Indique une différence statistiquement significative par rapport à A2 ($p < 0,05$) pour le WER.

TABLE 8 – Comparaison des résultats de l'expérience A2 avec les expériences A5 à A7

améliore le plus sur Common Voice, il dégrade significativement les acquis sur le jeu de test initial REYD. À l'inverse, la méthode d'ER (A6) garantit une stabilité supérieure : elle surpasse la référence A2 sur le jeu de test. Cependant, le déséquilibre des données entre REYD et Common Voice limite sa progression sur ce dernier corpus.

Concernant l'affinage par adaptateurs (A7), les dynamiques divergent fortement selon l'architecture. Le modèle MMS montre une chute drastique de sa capacité de généralisation. À l'opposé, l'utilisation de LoRA s'avère judicieuse pour whisper-large-v3 : elle dépasse également la référence A2, bien que le modèle A6 reste marginalement plus performant.

En conclusion, l'affinage améliore drastiquement les deux modèles par rapport à l'inférence initiale. L'ER (A6) s'impose comme la stratégie la plus robuste pour consolider un domaine spécifique sans oublier le précédent, tandis que l'approche LoRA (A7) appliquée à Whisper démontre également une forte capacité de généralisation. Bien que Whisper exige davantage de ressources computationnelles (GPU et temps d'entraînement), sa résilience justifie son utilisation, surtout en cas du code-switching.

5 Conclusion & Perspectives

Ces travaux démontrent la viabilité d'exploiter des modèles préentraînés (MMS et Whisper) pour développer un système de RAP performant pour une langue peu dotée et dialectalement fragmentée comme le yiddish, et avec des ressources limitées.

Les expériences mettent en évidence que l'affinage conjoint définit la limite supérieure des performances, et que l'ER s'impose comme la stratégie effectivement efficace pour contrer l'oubli des connaissances apprises précédemment. Par ailleurs, en matière d'adaptation efficace, l'approche par adaptateurs (LoRA) s'est révélée être une méthode également efficace pour optimiser le modèle Whisper.

À l'issue de cette étude, plusieurs axes d'amélioration se dessinent pour les travaux futurs. Tout

d'abord, concernant l'enrichissement et la fiabilisation des données, si les dialectes NEY et CY sont bien modélisés, le yiddish hassidique souffre encore d'un manque critique de ressources. L'application de techniques d'augmentation de données s'avère donc nécessaire. De plus, l'exploitation de l'intégralité du corpus CSYE, qui compte près de 278 heures d'enregistrements contre les 3 heures utilisées ici, permettrait également de consolider drastiquement la généralisation des modèles.

Ensuite, l'amélioration du décodage linguistique constitue un enjeu majeur. L'écart observé entre un WER perfectible et un CER relativement bas indique que les modèles reconnaissent les phonèmes avec précision, mais peinent à reconstituer les mots exacts face à la complexité dialectale. L'intégration d'un modèle de langue externe basé sur la norme YIVO, couplée à des méthodes de post-édition représenterait une solution très prometteuse pour corriger ces erreurs lexicales.

Enfin, les recherches futures devront s'orienter vers l'apprentissage continu et l'expansion dialectale en intégrant de nouvelles variétés au-delà des quatre étudiées (NEY, CY, SEY et hassidique). Pour accompagner cette mise à l'échelle sans intensifier l'oubli catastrophique ni faire exploser les coûts de calcul, il sera particulièrement pertinent d'explorer le préentraînement continu (Djanibekov *et al.*, 2025) ainsi que des approches avancées d'apprentissage continu (Della Libera *et al.*, 2024).

6 Remerciements

Ce travail est soutenu par l'Agence nationale de la recherche (ANR) et par le Ministère de l'Enseignement supérieur et de la Recherche de France (MESR).

Références

ABRAMSON C. (1995). Interview 8124. USC Shoah Foundation Visual History Archive.

ANAFI E. (1996). Interview 17228. USC Shoah Foundation Visual History Archive.

BAVISKAR V., HERZOG M. & WEINREICH U. (1992). *The Language and Culture Atlas of Ashkenazic Jewry : Historical and theoretical foundations*. The Language and Culture Atlas of Ashkenazic Jewry. M. Niemeyer.

BLEAMAN I. L. & NOVE C. R. (2025). The corpus of spoken yiddish in europe : Goals, methods, and applications. *Language Documentation & Conservation*, **19**.

BUREKHOVICH E. (1995). Interview 1227. USC Shoah Foundation Visual History Archive.

ĆAVAR M., ĆAVAR D., KERLER D.-B. & QUILITZSCH A. (2016). Generating a yiddish speech corpus, forced aligner and basic asr system for the aheym project. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 4688–4693.

DELLA LIBERA L., MOUSAVI P., ZAIEM S., SUBAKAN C. & RAVANELLI M. (2024). Cl-masr : A continual learning benchmark for multilingual asr. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

DIMANTSTEIN A. (1996). Interview 20327. USC Shoah Foundation Visual History Archive.

DJANIBEKOV A., TOYIN H. O., ALSHALAN R., ALATIR A. & ALDARMAKI H. (2025). Dialectal coverage and generalization in arabic speech recognition. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 29490–29502.

GRAVES A., FERNÁNDEZ S., GOMEZ F. & SCHMIDHUBER J. (2006). Connectionist temporal classification : labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, p. 369–376.

HU E. J., SHEN Y., WALLIS P., ALLEN-ZHU Z., LI Y., WANG S., WANG L., CHEN W. *et al.* (2022). Lora : Low-rank adaptation of large language models. *Iclr*, **1**(2), 3.

JACOBS N. G. (2005). *Yiddish : A linguistic introduction*. Cambridge University Press.

KOPOLOVICZ B. (1998). Interview 39085. USC Shoah Foundation Visual History Archive.

KWOK C. Y., YIP J. Q. & CHNG E. S. (2024). Continual learning optimizations for auto-regressive decoder of multilingual asr systems. *arXiv preprint arXiv :2407.03645*.

LIU Y., YANG X. & QU D. (2024). Exploration of whisper fine-tuning strategies for low-resource asr. *EURASIP Journal on Audio, Speech, and Music Processing*, **2024**(1), 29.

LUKS E. (1996). Interview 17055. USC Shoah Foundation Visual History Archive.

LUO J., WANG J., CHENG N., XIAO E., XIAO J., KUCSKO G., O'NEILL P., BALAM J., DENG S., FLORES A., GINSBURG B., HUANG J., KUCHAIEV O., LAVRUKHIN V. & LI J. (2021). Cross-language transfer learning and domain adaptation for end-to-end automatic speech recognition. In *Proceedings of 2021 IEEE International Conference on Multimedia and Expo (ICME)*, p. 1–6.

PANAYOTOV V., CHEN G., POVEY D. & KHUDANPUR S. (2015). Librispeech : an asr corpus based on public domain audio books. In *Proceedings of Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, p. 5206–5210 : IEEE.

PRATAP V., TJANDRA A., SHI B., TOMASELLO P., BABU A., KUNDU S., ELKAHKY A., NI Z., VYAS A., FAZEL-ZARANDI M., BAEVSKI A., ADI Y., ZHANG X., HSU W.-N., CONNEAU A. & AULI M. (2023). Scaling speech technology to 1,000+ languages.

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

RIDOY M. S. I., AKTER S. & RAHMAN M. A. (2025). Adaptability of asr models on low-resource language : A comparative study of whisper and wav2vec-bert on bangla. *arXiv preprint arXiv :2507.01931*.

TOYIN H. O., DJANIBEKOV A., KULKARNI A. & ALDARMAKI H. (2023). Artst : Arabic text and speech transformer.

WEBBER J., LO S. K. & BLEAMAN I. L. (2022). Reyd-the first yiddish text-to-speech dataset and system. In *Proceedings of INTERSPEECH*, p. 2363–2367.