Création d'un corpus parallèle de styles de parole en mandarin via l'auto-transcription et l'alignement forcé

Jingyi Sun, Yaru Wu, Nicolas Audibert, Martine Adda-Decker

Création d'un corpus parallèle de styles de parole en mandarin via l'auto-transcription et l'alignement forcé

Jingyi Sun[1] Yaru Wu[1, 2, 3] Nicolas Audibert[1] Martine Adda-Decker[1, 3]
(1) Laboratoire de Phonétique et Phonologie (CNRS & Sorbonne Nouvelle), 4 Rue des Irlandais, 75005 Paris, France
(2) CRISCO/UR4255 (Université de Caen Normandie), Esp. de la Paix, 14000 Caen, France
(3) LISN (Univ. Paris-Saclay), Rue du Belvédère, 91405 Orsay, France
{jingyi.sun, nicolas.audibert, martine.adda-decker}@sorbonne-nouvelle.fr, yaru.wu@unicaen.fr

RÉSUMÉ

La technologie ASR excelle dans la transcription précise des discours lus préparés, mais elle rencontre encore des défis lorsqu'il s'agit de conversations spontanées. Cela est en partie dû au fait que ces dernières relèvent d'un registre de langage non préparé et informel, avec disfluences et réductions de parole. Afin de mieux comprendre les différences de production en fonction des styles de parole, nous présentons la création d'un corpus de parole conversationnelle, dont des extraits sont ensuite lus par leurs auteurs. Le corpus comprend 36 heures de parole en chinois mandarin avec leur transcription, réparties entre conversations spontanées et lecture. Nous avons utilisé WHISPER pour la transcription automatique de la parole et le *Montreal Forced Aligner* pour l'alignement forcé, résultant dans un corpus de parole transcrit avec annotations multi-niveaux incluant phonèmes, caractères/syllabes et mots. De telles productions de parole parallèles (en modes spontané et lu) seront particulièrement intéressantes pour l'étude des réduction temporelle.

ABSTRACT

Creating a Speaking Styles Parallel Corpus in Mandarin through Auto-transcription and Forced Alignment

ASR technology excels in accurately transcribing prepared read speech, but it still encounters challenges when dealing with spontaneous conversations. This is partly because the latter is an unprepared, casual language register with lots of disfluencies and speech reductions. In order to improve our knowledge about speech variations, we designed an oral corpus of 36 hours of Mandarin Chinese. Spontaneous conversations are automatically transcribedand selected excerpts of these are then read by their authors. We employed WHISPER for automatic speech transcription and the *Montreal Forced Aligner* for forced alignment, which represents an effective procedure from speech to text, then to multi-tier annotation within phones, characters/syllables, and words, particularly suitable for large-scale speech corpus construction. This enabled us to collect different speaking styles with partly overlapped content produced by the same speaker. Such parallel corpus are particularly helpful to investigate temporal speech reduction phenomena in spontaneous speech.

MOTS-CLÉS : corpus parallèle, style de parole, auto-transcription, alignement forcé.

KEYWORDS: parallel corpus, speaking style, auto-transcribe, forced aligment.

1 Introduction

Malgré les progrès considérables réalisés en ASR au cours des dernières décennies, la parole spontanée et, en particulier, informelle reste difficile. Différents facteurs, tels qu'une possible diminution de la qualité du canal, le bruit de fond et les chevauchements de parole, pourraient être mentionnés comme des explications possibles de la performance plus faible de l'ASR (Benzeghiba *et al.*, 2007; O'Shaughnessy, 2008). Au-delà des facteurs non linguistiques mentionnés ci-dessus, il semble que les changements dans le contenu linguistique lui-même pourraient également expliquer partiellement certains aspects de la diminution de la précision de l'ASR : les disfluences, l'hypoarticulation et la réduction temporelle de la parole, qui peuvent entraîner des variantes de prononciation inattendues ou moins documentées (Adda-Decker & Lamel, 2018). Ces caractéristiques de la parole sont particulièrement remarquables dans des styles de parole moins contrôlés.

Deux styles de parole typiques sont prédominants dans les ensembles de données d'entraînement couvrant plusieurs langues : la lecture attentive et soignée et la conversation informelle non préparée. Ces styles de parole représentatifs couvrent deux extrémités d'un continuum (Gabler *et al.*, 2023). La lecture attentive préserve les formes de prononciation relativement intactes des phonèmes et la coarticulation nécessaire (Farnetani & Recasens, 1997). Dans les dialogues spontanés, le débit de parole et l'accentuation sont très flexibles et parfois incorrects sur le plan grammatical. Les unités linguistiques non accentuées, prévisibles et de haute fréquence tendent à être prononcées rapidement et même avec une réduction extrême. De plus, cette parole contient de nombreuses pauses remplies, des bégaiements, des répétitions, des autocorrections, des hésitations et des marqueurs de discours, ainsi que des rires et des toux, tous difficiles à éviter tout en maintenant la naturalité de la parole. Ces caractéristiques sont également essentielles pour mesurer la spontanéité et l'informalité de la parole (Dufour *et al.*, 2009).

Le corpus parallèle mandarin que nous sommes en train de créer est une étape importante vers l'extraction des paramètres phonétiques entre le registre informel et formel. Il est cependant important de noter que le terme « parallèle » utilisé ici diffère à la fois de la définition de Baker (Baker, 1995), qui fait référence à des corpus contenant des textes et des traductions dans deux langues ou plus, et du concept de Johansson (Johansson *et al.*, 1998), qui fait référence à des corpus contenant des textes dans deux langues avec des relations universelles et comparables. En utilisant le style de parole comme seule variable de comparaison, nous avons créé un corpus parallèle mandarin avec la même langue, le même locuteur, aucun dialecte régional et un contenu linguistique égal, mais seulement dans des styles de parole différents.

2 Travaux Connexes

Le développement de bases de données de parole de conversation spontanée a connu une croissance rapide au cours des dernières décennies, avec des bases de données disponibles dans différentes langues, notamment l'anglais, le français, l'espagnol, l'allemand, l'italien et le mandarin taïwanais (Du Bois *et al.*, 2000; Torreira & Ernestus, 2010; Kohler, 1996; Mereu & Vietti, 2021; Tseng, 2019). Cependant, dans les corpus de parole disponibles publiquement et axés sur le chinois mandarin , le style prédominant reste la lecture scriptée. Des bases de données à grande échelle de mandarin provenant de différentes régions et groupes d'âge comprennent, par exemple, *Chinese Mandarin (South/North) database (ELRA)*, *Chinese Digital Speech Data by Mobile Phone (ELRA)*, *AISHELL*

Speech databases (Bu *et al.*, 2017) et *1997 Mandarin Broadcast News Speech* (Graff, 2002). Il y a également eu des progrès significatifs dans le développement de jeux de données de parole spontanée en mandarin. Des exemples notables incluent les données de conversation en mandarin du *Mandarin Conversational Speech Data du Primewords Chinese Corpus Set 1* (Primewords Information Technology Co., 2018) et le *Magic Data Chinese Mandarin Conversational Speech* (Yang *et al.*, 2022).

Les corpus interlinguistiques couvrant simultanément deux styles de parole différents sont néanmoins rares, et la plupart d'entre eux n'exigent pas un contenu linguistique identique. La création de corpus parallèles contenant des informations stylistiquement distinctes mais sémantiquement comparables, avec un alignement automatique sur plusieurs niveaux de texte pour permettre un lien direct avec le signal de parole, n'a été explorée que dans une quantité limitée de recherches (Barras *et al.*, 2004). La majorité des institutions de recherche ou des entreprises optent soit pour la diffusion de corpus ne comprenant qu'un style de parole, soit utilisent différentes méthodes pour recueillir les deux styles. La méthode la plus courante consiste à recueillir des discours spontanés de participants en réponse à des entretiens, puis à leur demander d'effectuer une tâche de parole spécifique, telle que donner des indications basées sur une carte (Thompson *et al.*, 1993; Ibrahim *et al.*, 2020) ou collaborer pour décorer un arbre de Noël (Ito & Speer, 2006). Enfin, les participants sont tenus de lire un texte standardisé. Cette stratégie de collecte de données avec des tâches indépendantes est largement utilisée, mais elle présente plusieurs inconvénients, notamment la possibilité que les réponses de différents individus à la même tâche soient très similaires ou excessivement simplistes.

L'amélioration de la précision et de la robustesse de la technologie de transcription automatique permet la construction de corpus parallèles relatifs à différents styles de parole. Nous utilisons le système de reconnaissance vocale multilingue *Whisper* (Radford *et al.*, 2023) pour obtenir rapidement la transcription de la conversation décontractée des locuteurs, qui est ensuite fournie aux locuteurs pour lecture après adaptation manuelle. Cela peut favoriser des variations de prononciation plus riches pour la modélisation des caractéristiques acoustiques et évaluer les performances de l'ASR tout en conservant un contenu linguistique cohérent.

La procédure peut également fournir des données linguistiques naturelles étendues pour soutenir l'étude des phénomènes phonétiques en mandarin, tels que la coarticulation, le sandhi tonal et la synérèse, en produisant un corpus mandarin couvrant à la fois la parole spontanée et la lecture. Ces informations permettent une évaluation minutieuse pour déterminer si des variations phonétiques spécifiques sont aléatoires ou résultent de processus phonologiques (Shih, 2005), éclairant les similitudes et les différences entre la coarticulation, la réduction de la parole et autres phénomènes propres à la parole connectée (Farnetani & Recasens, 1997). Une approche de recherche basée sur le corpus implique l'obtention de limites temporelles pour les voyelles et les consonnes individuelles par alignement forcé et annotation automatique, en se concentrant sur les correspondances erronées entre les instances dans le modèle acoustique et les phonèmes alignés, un phénomène particulièrement prévalent dans la parole spontanée. Ensuite, les distributions, les durées, les fréquences à l'intérieur et entre les catégories de sons sont comptabilisées séparément, et leurs distances acoustiques mesurées (Audibert *et al.*, 2015). De telles études servent de référence pour comprendre les causes et les tendances de la réduction phonétique. De plus, comme le ton porte une charge phonémique significative en chinois, les variations de ton possèdent une valeur théorique cruciale (Surendran *et al.*, 2006).

3 Protocole de Construction du Corpus

3.1 Locuteurs

Le corpus actuel comprend 40 locuteurs, répartis de manière équilibrée selon le genre (F :H=1 :1). Ils proviennent de 19 provinces de Chine, notamment Zhejiang, Henan, Shandong, Hunan, Anhui, Jiangsu, Yunnan, etc. Les locuteurs, âgés de 20 à 32 ans, sont parfaitement à l'aise en mandarin standard sans accent régional, et tous sont des étudiants universitaires en bonne santé ne présentant ni troubles du langage ou mentaux, ni pathologies des organes articulatoires.

Étant donné l'exigence de capturer une parole conversationnelle spontanée, les participants appariés doivent être familiers les uns avec les autres. Cela réduit les sentiments négatifs potentiels tels que l'anxiété et le malaise pendant les sessions d'enregistrement. Avant de commencer l'enregistrement, les participants ont reçu une description complète de la procédure de collecte de données. Nous leur avons présenté environ 20 sujets quotidiens portant sur l'hébergement, les études, la nourriture, les voyages, les loisirs, etc., tout en expliquant les précautions à prendre lors du processus d'enregistrement, les droits qu'ils ont de suspendre/retirer à tout moment et la sécurité des données de transcription de *Whisper*. Ensuite, ils ont été invités à signer le formulaire de consentement éclairé après avoir confirmé qu'ils n'avaient aucune préoccupation ou question. Chaque locuteur a reçu 15€ en espèces à la fin de la session d'enregistrement.

3.2 Paramètres d'Enregistrement

Nous avons utilisé un enregistreur de terrain Roland R-26 et deux microphones casques AKG C520. Le taux d'échantillonnage est de 48 kHz, tandis que la quantification est réglée sur 16 bits. De plus, nous avons mis en œuvre le système de transcription automatique de la parole open-source *Python*-basé, *Whisper*, sur un ordinateur exécutant un système Windows 11 équipé d'une carte graphique discrète NVIDIA.

Basé sur *Python* et *PyTorch*, *Whisper* est un modèle multitâche réalisé grâce à une supervision faible à grande échelle pour la reconnaissance de la parole, la traduction et l'identification de la langue. La précision et la vitesse de transcription varient en fonction de la langue (l'anglais, l'espagnol, le néerlandais et le coréen donnent les meilleurs résultats) et de la taille du modèle (avec cinq options : *tiny*, *base*, *small*, *medium*, *large*). Pour le chinois mandarin, le modèle de taille moyenne au minimum est recommandé pour la précision, les modèles plus grands atteignant une précision plus élevée mais des vitesses de reconnaissance plus lentes.

3.3 Processus d'enregistrement et d'alignement forcé

Le processus d'enregistrement et d'alignement forcé comprend cinq phases principales.

1. Conversation Spontanée

   Tout d'abord, nous demandons aux participants de s'engager dans des discussions ouvertes aussi longtemps que possible. Un groupe de locuteurs peut souvent discuter en continu pendant 40 à 70 minutes. Avant l'expérience, nous fournissons aux participants une variété de sujets

de référence portant sur de nombreux aspects de la vie quotidienne. L'audio des conversations spontanées de deux participants est enregistré en stéréo. Nous considérons que les données de conversation spontanée vraiment naturelles commencent environ cinq minutes après le début de la conversation. Une fois que les participants commencent à parler, notre enregistrement commence en même temps.

2. Transcription et Modification de Texte

Après que le locuteur a fini de parler, nous commençons la transcription avec *Whisper*. Selon nos tests, transcrire une conversation d'une heure en utilisant le modèle *large* prend environ 18 minutes. Étant donné que *Whisper* ne différencie pas entre les locuteurs et qu'il y a quelques erreurs de reconnaissance de la parole dans le texte transcrit, causées soit par des homophones soit par une segmentation incorrecte des mots, la vérifiation et la correction manuelle des transcriptions sont nécessaires. Ensuite, ces transcriptions corrigées font l'objet d'un processus de révision et de correction par les locuteurs pour obtenir la version finale écrite pour la relecture par les locuteurs.

Il convient cependant de noter que la structure grammaticale et sémantique de l'improvisation orale est informelle, avec de nombreuses hésitations telles que des répétitions ou des auto-corrections. Par conséquent, nous visons à maintenir ces hésitations et ces non-conformités grammaticales dans les traductions anglaises correspondantes de chaque phrase et à les effacer lors de l'adaptation au texte lu au style écrit. La parole de conversation spontanée comprend également de nombreuses phrases courtes, telles que des réponses brèves ou des déclarations interrompues. Dans ce cas, nous ne couvrons pas de manière exhaustive toutes les informations de la conversation spontanée. Au lieu de cela, nous sélectionnons des segments plus longs et relativement complets pour la réécriture. De plus, l'alternance rapide des tours de conversation peut entraîner des transcriptions modifiées incohérentes ou difficiles à comprendre lorsqu'elles sont extraites séparément. Pour éviter cela, nous pouvons, si nécessaire, incorporer des parties d'informations linguistiques du locuteur 1 dans le texte lu du locuteur 2, et vice versa.

Comme le montre la Figure 1, nous avons apporté les ajustements suivants pour organiser le texte au style écrit :

— Segmentation du discours non ponctué en phrases fluides à lire pour les locuteurs et qui ne semblent pas linguistiquement artificielles. Par exemple, nous transformons la phrase a. en deux phrases A., en utilisant la ponctuation pour séparer deux propositions indépendantes.
— Réorganisation des phrases fragmentées et incorporation des informations linguistiques de l'autre personne pour compléter l'énoncé. Par exemple, la phrase i. répond à la question de l'autre, nous combinons donc la phrase h. et la recréons en tant que phrase grammaticalement et sémantiquement bien formée H.+I.. En général, les énoncés appropriés pour la supplémentation et la fusion devraient inclure au moins deux des éléments suivants : sujet, prédicat et objet. Cela permet de sélectionner les parties manquantes à partir des informations linguistiques fournies par l'interlocuteur et de compléter l'énoncé. Cependant, si les deux parties ont fourni des informations importantes pour les phrases modifiées, alors les deux locuteurs seront invités à lire la phrase.
— Élimination des interjections et des particules modales de la transcription. Par exemple, nous retirons « 啊(ah) » dans la phrase e., une interjection, et « 嗯(um) » dans la phrase i., une particule modale.
— Réorganisation des phrases inversées incorrectes. Par exemple, dans la phrase F., nous avons déplacé l'adverbe « 先(d'abord) » devant le VP « 预约一下(prendre rendez-vous) ».

— Élimination des phrases répétées ou des auto-corrections. Par exemple, le sujet de la phrase c. a fait l'objet d'une auto-correction, de « 我(je) » à « 我男朋友(mon petit ami) », nous ne conservons donc que cette dernière lors de la réorganisation de la phrase C.

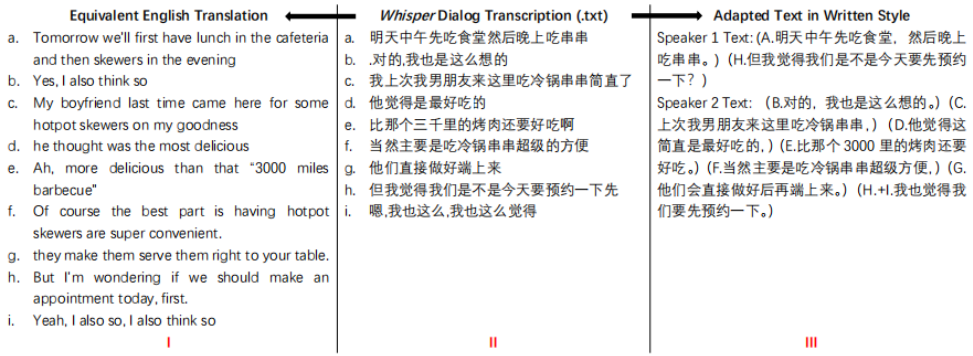

FIGURE 1 – Transcription du dialogue *Whisper* (II) au centre, la traduction anglaise correspondante (I) à gauche, et le texte adapté au style écrit (III) à droite

En nous basant sur notre expérience actuelle en matière d'édition de texte, le problème le plus difficile est de rectifier les inexactitudes dans la transcription, souvent dues aux homophones ou aux hésitations de la parole. Dans de tels cas, nous nous appuyons sur l'interprétation par le locuteur du contenu de l'interaction afin de finaliser le manuscrit du discours, car ils se souviennent généralement de ce qu'ils viennent de dire. La révision des textes des lectures des deux intervenants prend généralement environ 30 minutes au total.

3. Lecture de Texte

   Après avoir obtenu la transcription textuelle modifiée, nous demandons aux participants de la lire deux fois de manière émotionnellement neutre tout en enregistrant leur discours. Cela est destiné à obtenir autant de fluidité que possible dans la lecture de la parole typique, car la première lecture peut donner lieu à des hésitations en raison de la méconnaissance du matériel. Le discours lu est enregistré en mono, cette phase prenant généralement 10 à 15 minutes par locuteur.

4. Prétraitement de l'entrée pour l'alignement forcé

   Nous avons utilisé *Montreal Forced Aligner 3.0.6* (McAuliffe *et al.*, 2017) pour l'alignement automatique au niveau des phonèmes des mots, ce qui nécessite la préparation préalable de textes de transcription correctement formatés et d'audio sous-échantillonné. Le traitement des textes de transcription chinois nécessite une attention particulière, car ils sont écrits en chaînes continues, ce qui implique qu'il n'y a pas de repères de segmentation en dehors de la ponctuation. Ainsi, le package *spacy-pkuseg* est utilisé pour aider à la tokenisation automatique du texte chinois. L'alignement forcé est effecté en utilisant un lexique de prononciation chinois mandarin pré-entraîné et des modèles acoustiques dans MFA pour l'alignement forcé. Chaque heure de données audio nécessite environ 3 à 4 minutes de traitement pour l'alignement forcé.

5. Vérification manuelle et Réalignement

   La performance de Whisper utilisant le modèle large-v3 en mandarin a été évaluée par ses auteurs sur deux ensembles de données. Dans l'ensemble de données *Common Voice 15*, le

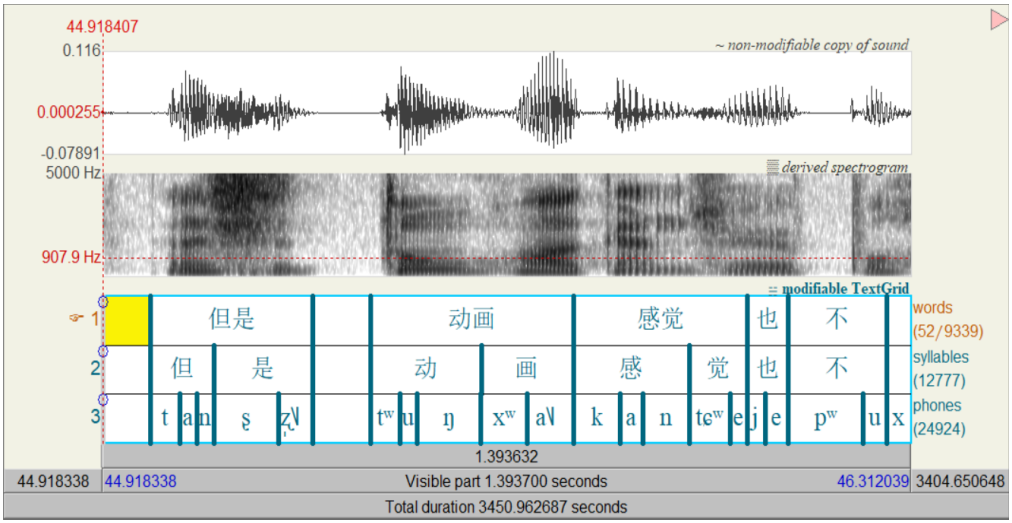

FIGURE 2 – Annotation à trois niveaux des Phones, des Caractères/Syllabes et des Mots

taux d'erreur de caractères (CER) était de 12,8%, tandis que dans l'ensemble de données *Fleurs*, le CER était de 7,7%. Cela indique que sa performance de reconnaissance varie en fonction de l'ensemble de données utilisé. Nous avons ensuite évalué *Whisper* sur la base de ce corpus parallèle de styles de parole et avons constaté que le CER pour la parole lue est de 4,18%, tandis que pour la parole de conversation spontanée, il est de 8,37%. Ces erreurs se reflètent également dans les résultats d'alignement forcé de la transcription provenant de *Whisper* et de l'audio. Par conséquent, après le premier alignement forcé, nous devons corriger les erreurs de transcription dans *Praat* (Boersma & Van Heuven, 2001) et ajouter du texte avec leurs limites pour correspondre à la parole pertinente dans les sections non reconnues. Contrairement à la correction précédente visant à faciliter la relecture par les locuteurs, cette correction de la transcription est effectuée en tenant compte autant que possible des hésitations et des autres disfluences. Cette démarche permettra d'obtenir une saisie de texte de transcription plus précise pour le réalignement.

Après vérification manuelle, un réalignement est nécessaire pour obtenir un nouveau fichier *textgrid* avec les frontières temporelles correctes pour chaque mot. De plus, étant donné que les mots ne sont pas l'unité la plus petite dans le système d'écriture chinois, nous pouvons également utiliser *Python* pour segmenter automatiquement le texte de transcription en caractères/syllabes chinois pour l'alignement. Les résultats alignés peuvent être superposés aux annotations obtenues au niveau des mots pour former des annotations à trois niveaux : phones, caractères/syllabes chinois et mots, comme illustré dans la Figure 2.

Cette approche permet la construction rapide de corpus de langage parlé spontané à grande échelle avec des données annotées. Le corpus comprend une durée totale de 36 :01 :53, comprenant 28 :59 :00 de parole de conversation spontanée et 7 :02 :53 de parole lue. Les statistiques préliminaires indiquent que la parole spontanée contient 586 554 caractères/syllabes chinois et 1 184 186 phonèmes, tandis que la parole lue contient 116 345 caractères/syllabes chinois et 245 605 phonèmes. Il est à noter que le nombre de phones dépend également du

modèle acoustique utilisé, dans ce cas, en utilisant le modèle pré-entraîné "mandarin_mfa", qui comprend 142 phones. Dans nos données de parole lue, les 28% des phones les plus représentés (40) représentent 72,11% de tous les phones. Les consonnes les plus fréquentes comprennent deux nasales, /n/ (7,75%) et /ŋ/ (4,60%), ainsi que des occlusives, des fricatives et des affriquées /t/ (4,35%), /ʂ/ (3,37%), /tɕ/ (2,24%), /ʐ/ (2,60%) et /x/ (1,61%). Les voyelles les plus fréquentes comprennent /o/ (7,16%), /a/ (6,46%), /i/ (4,88%) et /ə/ (3,46%). De plus, deux approximants, /w/ (4,32%) et /j/ (3,20%), sont également très fréquents. Un affinement supplémentaire du corpus permettra d'obtenir plus de différences statistiques entre la parole spontanée et lue en chinois mandarin, inspirant ainsi une exploration plus poussée des motifs de variation allophonique.

4 Conclusion et Travaux Futurs

Cette étude propose une méthode systématique pour créer un corpus vocal parallèle concernant différents styles de parole avec le même locuteur et les mêmes informations linguistiques, avec l'aide de *Whisper* et de *Montreal Forced Aligner*. Le modèle *large* de *Whisper* permet une récupération rapide et précise du texte de transcription hors ligne, qui est ensuite adapté pour être utilisé comme matériel au style écrit pour la tâche de lecture. Dans les cas où les mots et les phrases se chevauchent fortement, ce corpus peut être utilisé non seulement pour comparer les performances de reconnaissance vocale de différents styles de parole et localiser rapidement les différences, mais il peut également fournir des variantes de prononciation riches pour l'étude des mécanismes de réduction de la parole en chinois mandarin, de la coarticulation et du sandhi tonal. En particulier, l'étude des affriquées et de l'aspiration en chinois mandarin peuvent permettre de mieux documenter les phénomènes de réduction, tandis que l'interaction entre les tons et l'intonation, la focalisation et d'autres paramètres prosodiques dans la parole spontanée extensive peut également être étudiée, éclairant des caractéristiques multidimensionnelles dans la parole informelle et formelle (Chen & Yuan, 2007). Nos prochaines étapes impliquent l'annotation automatisée, la vérification manuelle du corpus et l'analyse des fréquences d'occurrence sur les syllabes, les phonèmes, les tons et d'autres aspects pertinents pour la caractérisation de la réduction en mandarin.

Remerciements

Ce travail a bénéficié du soutien financier du Laboratoire d'Excellence Empirical Foundations of Linguistics (LabEx EFL, ANR-10-LABX-0083), contribuant ainsi à l'IdEx Université de Paris (ANR-18-IDEX-0001), ainsi que du projet ANR-21-CE38-0019 DIPVAR. Jingyi Sun a été soutenue par une bourse du China Scholarship Council (Grant No. 202208410095).

Références

ADDA-DECKER M. & LAMEL L. (2018). Discovering speech reductions across speaking styles and languages. *Rethinking reduction : Interdisciplinary perspectives on conditions, mechanisms, and domains for phonetic variation*, **25**, 101. DOI : 10.1515/9783110524178-004.

AUDIBERT N., FOUGERON C., GENDROT C. & ADDA-DECKER M. (2015). Duration-vs. style-dependent vowel variation : A multiparametric investigation. In *18th International Congress of Phonetic Sciences (ICPhS'15)*.

BAKER M. (1995). Corpora in translation studies : An overview and some suggestions for future research. *Target. International Journal of Translation Studies*, **7**, 223–243. DOI : 10.1075/target.7.2.03bak.

BARRAS C., ADDA G., ADDA-DECKER M., HABERT B., DE MAREÜIL P. B. & PAROUBEK P. (2004). Automatic Audio and Manual Transcripts Alignment, Time-code Transfer and Selection of Exact Transcripts. In M. T. LINO, M. F. XAVIER, F. FERREIRA, R. COSTA & R. SILVA, Éds., *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, p. 877–880 : European Language Resources Association (ELRA).

BENZEGHIBA M., DE MORI R., DEROO O., DUPONT S., ERBES T., JOUVET D., FISSORE L., LAFACE P., MERTINS A., RIS C. *et al.* (2007). Automatic speech recognition and speech variability : A review. *Speech communication*, **49**, 763–786. DOI : 10.1016/j.specom.2007.02.006.

BOERSMA P. & VAN HEUVEN V. (2001). Speak and unspeak with praat. *Glot International*, **5**(9/10), 341–347.

BU H., DU J., NA X., WU B. & ZHENG H. (2017). Aishell-1 : An open-source mandarin speech corpus and a speech recognition baseline. http://www.aishelltech.com/kysjcp.

CHEN Y. & YUAN J. (2007). A corpus study of the 3rd tone sandhi in standard chinese. In *Interspeech*, p. 2749–2752 : Citeseer.

DU BOIS J. W., CHAFE W. L., MEYER C., THOMPSON S. A. & MARTEY N. (2000). Santa barbara corpus of spoken american english. *CD-ROM. Philadelphia : Linguistic Data Consortium*. DOI : 10.35111/s2q7-gq73.

DUFOUR R., JOUSSE V., ESTÈVE Y., BÉCHET F. & LINARÈS G. (2009). Spontaneous speech characterization and detection in large audio database. *SPECOM, St. Petersburg*, **7**, 41–46.

FARNETANI E. & RECASENS D. (1997). Coarticulation and connected speech processes. *The handbook of phonetic sciences*, **371**, 404.

GABLER P., GEIGER B. C., SCHUPPLER B. & KERN R. (2023). Reconsidering read and spontaneous speech : Causal perspectives on the generation of training data for automatic speech recognition. *Information*, **14**, 137. DOI : 10.3390/info14020137.

GRAFF D. (2002). An overview of broadcast news corpora. *Speech Communication*, **37**, 15–26. DOI : 10.1016/S0167-6393(01)00057-7.

IBRAHIM O., ASADI H., KASSEM E. & DELLWO V. (2020). Arabic speech rhythm corpus : Read and spontaneous speaking styles. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 5337–5342, Marseille, France : European Language Resources Association.

ITO K. & SPEER S. R. (2006). Using interactive tasks to elicit natural dialogue. *Methods in empirical prosody research*, p. 229–257.

JOHANSSON S., EBELING S. O. & OKSEFJELL S., Éds. (1998). *Corpora and cross-linguistic research : Theory, method and case studies*. Rodopi.

KOHLER K. J. (1996). Labelled data bank of spoken standard german : the kiel corpus of read/spontaneous speech. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, p. 1938–1941 : IEEE.

MCAULIFFE M., SOCOLOF M., MIHUC S., WAGNER M. & SONDEREGGER M. (2017). Montreal forced aligner : Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, p. 498–502.

MEREU D. & VIETTI A. (2021). Dialogic italian : the creation of a corpus of italian spontaneous speech. *Speech Communication*, **130**, 1–14. DOI : 10.1016/j.specom.2021.03.002.

O'SHAUGHNESSY D. (2008). Automatic speech recognition : History, methods and challenges. *Pattern Recognition*, **41**, 2965–2979. DOI : 10.1016/j.patcog.2008.05.008.

PRIMEWORDS INFORMATION TECHNOLOGY CO. L. (2018). Primewords chinese corpus set 1. https://www.primewords.cn.

RADFORD A., KIM J. W., XU T., BROCKMAN G., MCLEAVEY C. & SUTSKEVER I. (2023). Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, p. 28492–28518.

SHIH C. (2005). Understanding phonology by phonetic implementation. In *Ninth European Conference on Speech Communication and Technology*.

SURENDRAN D., NIYOGI P. *et al.* (2006). Quantifying the functional load of phonemic oppositions, distinctive features, and suprasegmentals. *Amsterdam studies in the theory and history of linguistic science series 4*.

THOMPSON H. S., ANDERSON A. H., BARD E. G., DOHERTY-SNEDDON G., NEWLANDS A. & SOTILLO C. (1993). The hcrc map task corpus : Natural dialogue for speech recognition. In *Human Language Technology : Proceedings of a Workshop Held at Plainsboro, New Jersey*.

TORREIRA F. & ERNESTUS M. (2010). The nijmegen corpus of casual spanish. In *Seventh conference on International Language Resources and Evaluation (LREC'10)*, p. 2981–2985.

TSENG S.-C. (2019). Ilas chinese spoken language resources. *Proceedings of LPSS 2019*, p. 13–20.

YANG Z., CHEN Y., LUO L., YANG R., YE L., CHENG G., XU J., JIN Y., ZHANG Q., ZHANG P. *et al.* (2022). Open source magicdata-ramc : A rich annotated mandarin conversational (ramc) speech dataset. *arXiv preprint arXiv :2203.16844*. DOI : 10.48550/arXiv.2203.16844.