Preuve de concept d'un système de génération automatique en Langue française Parlée Complétée

Brigitte Bigi, Núria Gala

Preuve de concept d'un système de génération automatique en Langue française Parlée Complétée

Brigitte Bigi, Núria Gala
LPL, CNRS, Aix-Marseille Univ, 5 avenue Pasteur, 13100 Aix-en-Provence
brigitte.bigi@cnrs.fr, nuria.gala@univ-amu.fr

RÉSUMÉ

La Langue française Parlée Complétée (LfPC) est un système de communication développé pour les personnes sourdes afin de compléter la lecture labiale avec une main, au niveau phonétique. Il est utilisé par les enfants pour acquérir des compétences en lecture, en lecture labiale et en communication orale. L'objectif principal est de permettre aux enfants sourds de devenir des lecteurs et des locuteurs compétents en langue française. Nous proposons une preuve de concept (PoC) d'un système de réalité augmentée qui place automatiquement la représentation d'une *main codeuse* sur la vidéo pré-enregistrée d'un locuteur. Le PoC prédit la forme et la position de la main, le moment durant lequel elle doit être affichée, et ses coordonnées relativement au visage dans la vidéo. Des photos de mains sont ensuite juxtaposées à la vidéo. Des vidéos annotées automatiquement par le PoC ont été montrées à des personnes sourdes qui l'ont accueilli et évalué favorablement.

ABSTRACT

Toward an Automatic Cued Speech System for French Language

Cued Speech is a communication system developed for deaf people to complement speechreading at the phonetic level with hands. It is used by children to acquire skills in reading, in lip reading and oral communication. The main goal is to allow deaf children to become proficient readers and speakers of an oral language. We propose a Proof of Concept (PoC) of an augmented reality system that automatically places the representation of a coding hand on a video of a pre-recorded speaker. The PoC is predicting the key to be coded (shape and position), when it has to be coded relatively to the audio and its coordinates relatively to the face in the video. Photos of human hands are then juxtaposed to the video. Videos automatically encoded with this system have been shown to deaf people who have welcomed and positively evaluated.

MOTS-CLÉS : LfPC, automatisation, PoC, surdité, vidéo, annotation.

KEYWORDS: Cued Speech, automatic, PoC, deaf, video, annotation.

1 Introduction

Lorsque la LSF n'est pas utilisée, la lecture labiale est l'une des principales modalités visuelles qui permet l'accès à la parole pour les personnes sourdes ou malentendantes. Elle est utilisée en conjonction avec d'autres stratégies de communication, comme les aides auditives, et/ou des solutions visuelles. Parmi ces dernières, en 1966, R. Orin Cornett a inventé le « Cued Speech » (CS), un codage qui ajoute des informations visuelles sur les sons qui ne sont pas différentiables sur les lèvres (Cornett, 1967). Ce codage CS représente chaque son avec une forme de main pour une consonne et

une position autour du visage pour une voyelle. Leur combinaison forme une *clé*. Lorsque les sons se ressemblent sur les lèvres, ils sont codés différemment ; la combinaison entre forme labiale et clé implique un percept unique de ce qui est prononcé. Par exemple, "bi" et "mi" qui sont identiques sur les lèvres sont codés avec deux formes différentes de la main. Le CS est souvent utilisé dans les milieux éducatifs, en particulier pour les jeunes enfants ayant une déficience auditive, car il leur donne accès à la langue orale *via* l'information phonémique qu'ils pourraient manquer par des moyens auditifs traditionnels. L'objectif majeur du CS est ainsi de faire en sorte que les enfants sourds puissent accéder plus facilement à la communication orale. L'efficacité de ce codage pour améliorer la perception et la production de la parole a été démontrée dans un grand nombre d'études, notamment (Kaplan, 1975; Neef & Iwata, 1985; Leybaert *et al.*, 2010).

L'automatisation du « Cued Speech », c-à-d l'utilisation d'un système automatisé pour coder et/ou décoder les sons, concerne essentiellement deux domaines : la synthèse – codage, dont cet article fait l'objet, et la reconnaissance –décodage. Avec un système de codage automatique, toutes sortes de vidéos codées pourraient être élaborées et diffusées pour tous les types d'utilisations. Disposer d'outils permettant de s'entraîner à la pratique du code constituerait un bénéfice important pour les parents d'enfants sourds, ainsi que pour les centres d'éducation spécialisée, par exemple. Cela permettrait entre autres de réduire les inégalités d'accès à la LfPC sur le territoire, d'apporter une aide à l'acquisition de la langue orale par les enfants sourds, d'améliorer la communication entre les personnes sourdes ou malentendantes et les membres de leur famille entendants, ou d'aider à développer des compétences de lecture labiale. Dans ce domaine, le premier système *AutoCuer* avait été proposé par l'inventeur du codage (Cornett *et al.*, 1977). Par la suite, dans les années 1995-2000, plusieurs recherches ont été conduites au *Massachusetts Institute of Technology* (MIT) pour automatiser le codage (Bratakos, 1995; Sexton, 1997; Bratakos *et al.*, 1998; Duchnowski *et al.*, 2000). Ces travaux ont consisté à vérifier la faisabilité d'automatiser la génération des clés, c'est-à-dire à déterminer la séquence de clés qui doit être produite puis générer une vidéo augmentée d'une main codeuse. Un locuteur était filmé pendant qu'il parlait, sans coder. Un système de reconnaissance automatique de la parole permettait d'obtenir la séquence des phonèmes à partir desquels le système déterminait les clés à incruster dans la vidéo. Dans une autre pièce, se trouvait une personne qui devait décoder la vidéo ainsi générée en temps réel. Dans les versions successives de ce système, les mains étaient représentées par des *cliparts*. Les différentes évaluations ont toujours montré *a minima* un petit avantage du décodage avec ajout de l'image de la main codeuse par rapport à la lecture labiale seule.

Malgré les nombreux travaux démontrant ses avantages, et l'intérêt grandissant qu'il suscite, il n'existe aucun système de génération automatique des clés. Les études récentes relatives à l'automatisation du CS se concentrent, en effet, sur la reconnaissance (Sankar *et al.*, 2023). En outre, il n'existe que très peu d'études qui décrivent le fonctionnement du codage, notamment en ce qui concerne l'organisation temporelle et spatiale du code dans sa co-production avec la parole (Attina, 2005).

2 Méthodologie proposée

La figure 1 illustre le processus que nous proposons pour implémenter un système de codage automatisé, c-à-d un système qui permettra d'ajouter une main codeuse artificielle à la vidéo d'un locuteur. En amont, l'utilisateur doit préparer un enregistrement audio-vidéo et sa transcription orthographique, alignée dans des unités courtes, telles que les Unités Inter-Pausales (IPUs). Le

système utilise le logiciel libre SPPAS (Bigi, 2015) pour obtenir les annotations audio-vidéo requises en entrée. Pour chaque IPU, SPPAS détermine 1/ la séquence des phonèmes et leur alignement temporel avec l'audio, 2/ les coordonnées de 68 points spécifiques du visage du locuteur dans chacune des images de la vidéo. A partir de ces informations, l'objectif est de produire des annotations qui permettront d'augmenter la vidéo automatiquement avec la main codeuse.

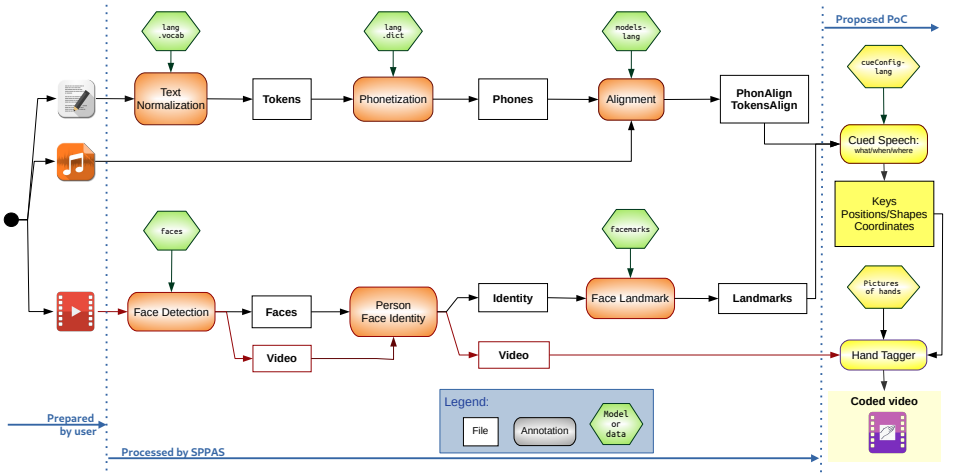

FIGURE 1 – Processus proposé pour la génération automatique du codage

Il est prévu que le système final soit implémenté avec des approches hybrides, donc partiellement basé sur des modèles à base de connaissance et partiellement sur des modèles issus de méthodes empiriques. Il sera développé pour le français ; dans ce cas, on utilise l'acronyme francophone LfPC, pour Langue française Parlée Complétée, plutôt que « Cued Speech ».

Pour ce faire, nous avons d'ores et déjà collecté un corpus, le Corpus de Lecture en LfPC (CLeLfPC) accessible sous licence libre (Bigi *et al.*, 2022). Il se compose de 4 heures d'enregistrements audio-vidéo, de 23 locuteurs codant en LfPC. Son enrichissement avec des annotations permettra les analyses requises pour la création de modèles de synchronisation audio-main basés sur des apprentissages supervisés, en répondant à quatre sous-problématiques que nous avons définies, et qui déterminent :

1. **quoi** : la séquence des clés à produire à partir des phonèmes,
2. **quand** : les moments de présentations et de transitions des formes et positions de la main,
3. **où** : les coordonnées, angle et taille de la main par rapport au visage, et,
4. **comment** : l'incrustation de représentations d'une main dans chaque image de la vidéo

Compte tenu du coût de la collecte et de l'annotation d'un corpus, avant d'aller plus avant dans la création de ce système, il nous a semblé indispensable de connaître l'intérêt réel que le système pourrait apporter, et d'en étudier la faisabilité.

Parallèlement à la création du corpus, nous avons donc élaboré une preuve de concept (PoC), qui est une phase déterminante pour **décider si le système peut et doit être implémenté**. La création des annotations du corpus, leur analyse et, d'une manière générale, l'ensemble des recherches liées à la mise en oeuvre des modèles et du système sont conditionnés **par l'approbation du PoC**. Il se compose de quatre modules, chacun impliquant de répondre à des problématiques spécifiques, afin de

déterminer la séquence de clés à produire à partir d'un signal d'entrée audio vidéo (lecture à haute voix) et de sa transcription orthographique, l'organisation temporelle entre la main et les phonèmes, le positionnement de la main par rapport au visage, et enfin le marquage de la vidéo avec la main.

3 Description de la preuve de concept

La preuve de concept est un système qui produit automatiquement des fichiers XML contenant les informations relatives au codage en LfPC. Ces fichiers contiennent les annotations indiquant quelles sont les formes et positions de la main qu'il faut intégrer, à quel moment et où il faut les placer dans la vidéo. D'autre part, le PoC crée un fichier vidéo augmenté avec la représentation de la main codeuse. En entrée, le PoC nécessite de connaître la séquence des phonèmes prononcés et leur position temporelle par rapport à l'audio, ainsi que les coordonnées du visage du locuteur pour chacune des images de la vidéo, comme indiqué dans la figure 1.

3.1 Quelles sont les clés ?

Afin de déterminer la séquence des clés à produire à partir des phonèmes, le PoC implémente un système à base de règles de productions, élaborées en collaboration avec des experts du codage.

Pour l'implémenter, dans un premier temps nous avons assigné un numéro à chacune des 8 formes de la main ainsi qu'à la forme neutre (voir figure 2), et nous avons assigné une lettre à chacune des 5 positions que compte la LfPC, autour du visage, et une lettre pour la position neutre sur la poitrine. Le système a donc pour tâche de proposer la séquence de clés qui correspond à la séquence de phonèmes, comme dans l'exemple suivant, dont les phonèmes sont codés en X-SAMPA :

```
entrée: 9~ d @ m i p o d H i l d @ k o k o
sortie: 5t.1s.5m.1s.1s.4m.6s.1s.2s.2s
```

Nous avons manuellement annoté une partie du corpus CLeLfPC (5 locuteurs) afin de déterminer les clés produites par les locuteurs, et nous les avons comparées aux clés prédites par le PoC (Auteur, 2023). Cette évaluation a permis de valider le modèle à base de règles que nous proposons. Cependant, la variabilité dans la production orale, notamment liée aux accents ou au contexte de la production, implique une difficulté dans cette tâche et des améliorations sont donc possibles et envisagées, par exemple en laissant le choix à l'utilisateur de modifier les règles de production.

3.2 Quand présenter les clés ?

Dans le codage CS, une clé correspond à un groupe spécifique de phonèmes (consonne + voyelle, consonne seule ou voyelle seule). Les mouvements de la main, forme pour la consonne et position pour la voyelle, doivent coïncider avec les phonèmes produits. Cette coordination précise est essentielle pour transmettre avec précision les nuances du langage parlé et combler le fossé entre la communication visuelle et auditive. (Cornett, 1967) avait déjà indiqué que les lèvres et les mouvements de la main n'apparaissent pas en même temps. Par la suite, (Bratakos *et al.*, 1998) indique que la main doit être en avance sur le son, qu'un retard de 33ms n'a que peu de conséquences et que le retard maximal acceptable est de 100ms. (Duchnowski *et al.*, 1998) démontrent ensuite que les scores de décodages sont meilleurs si la main est présentée 100ms avant le mouvement labial. Enfin, (Duchnowski *et al.*,

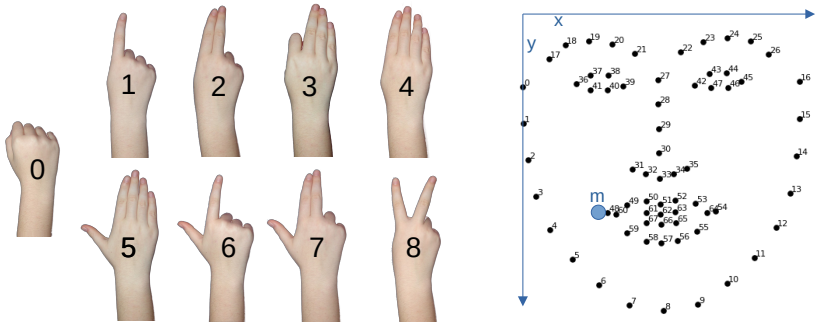

FIGURE 2 – Représentation du codage en LfPC

2000) indiquent qu'une transition de 150ms doit être opérée pour changer de position. D'autres études ont ensuite été menées pour la langue française, notamment dans (Cathiard *et al.*, 2003; Attina, 2005; Aboutabit, 2007) et ont abouti à la proposition de modèles de synchronisation main-lèvres-son, comme par exemple celui simplifié dans la figure 3. Cette figure représente le modèle que nous avons implémenté dans la preuve de concept. M1 indique le moment durant lequel la main commence à changer de position et M2 le moment où la main arrive à la position cible ; D1 indique le début du changement de forme de la main et D2 son accomplissement. Cependant, il ne couvre pas toutes les situations, et nous l'avons complété de règles à l'aide d'experts du codage. Entre-autres, nous avons traité les cas particuliers relatifs à la transition depuis la position neutre vers une position du visage (anticipée), et la transition depuis une position du visage vers la position neutre (retardée).

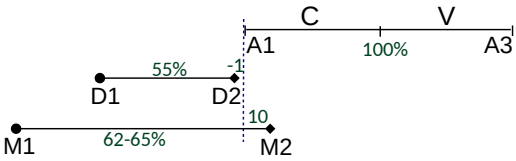

FIGURE 3 – Reproduction partielle de la synchronisation main-son, (Attina, 2005), figure 32 pp. 136. A1et A3 sont respectivement les moments de début et de fin des phonèmes consonne C, et voyelle V.

3.3 Où placer la main par rapport au visage ?

Il faut également modéliser la trajectoire de la main en proposant des modèles qui peuvent prédire à tout moment l'emplacement, l'angle et la taille de la main par rapport au visage du locuteur. Dans ce domaine, nous n'avons trouvé aucune étude antérieure publiée. Dans un premier temps, nous nous sommes appuyés sur les experts du codage pour placer les positions des voyelles par rapport à un visage théorique. Nous avons ensuite décrit ces positions par rapport aux 68 points de ce visage qui sont obtenus avec un système de "Face landmark". Par exemple, les coordonnées de la position 'm' qui représente les sons /i/, /a˜/ et /O˜/ et se situe proche du coin de la bouche, se calculent avec : $x = x_{48} - |(x_{54} - x_{48})/4|$, $y = y_{60}$. Cette position est illustrée dans la figure 2.

Nous n'avons introduit aucune variabilité dans l'estimation de ces coordonnées. Ainsi, si la main doit se trouver à cette position, le doigt cible (bout de l'index ou bout du majeur, selon la forme) est placé à ces coordonnées sur l'image correspondante dans la vidéo. Par ailleurs, une valeur d'angle de la main a été fixée pour chacune des positions par les experts du codage. Là aussi, nous n'avons introduit aucune variabilité : pour une position donnée, la main se place toujours avec le même angle. De même pour la taille de la main qui est proportionnelle à la hauteur du visage. Enfin, la trajectoire suivie par la main entre deux positions suit une ligne droite, à vitesse constante. Des analyses devront être conduites sur le corpus codé afin de rendre ce mouvement plus naturel.

3.4 Comment représenter le codage dans la vidéo ?

Une fois que le système a pu prédire quelle clé doit être codée, à quel moment et à quel endroit, le PoC peut augmenter la vidéo avec la main codeuse. Contrairement aux systèmes proposés par le MIT, pour le PoC, nous avons choisi d'utiliser des photos (figure 2), plutôt que des représentations imagées. Nous avons utilisé les fonctions de floutage (blur) et de transparence (fade in/fade out) pour indiquer respectivement les transitions de position et de forme.

4 Évaluations

4.1 Protocole

Pour évaluer la pertinence de la preuve de concept, nous avons mis en place un protocole d'évaluation en créant des séries de vidéos à décoder, *sans audio*. Le premier auteur de cet article a été filmé en lisant 4 sessions du corpus CLeLfPC, sans coder. Nous avons ensuite annoté le corpus avec SPPAS, en suivant le processus proposé dans la figure 1 : détection automatique des IPUs, transcription manuelle, segmentation automatique en phonèmes, détections des points du visage. Le PoC a ensuite généré les annotations et les vidéos codées automatiquement. Les vidéos des 4 sessions enregistrées ont été divisées en 4 séries différentes pour former un ensemble de 16 expériences avec des vidéos différentes. Chacune des 16 expériences se compose de 34 vidéos codées dont 10 servent de contrôle et 24 de test, avec un recouvrement des vidéos contrôle/test sur différentes expériences. Chaque vidéo ne contient qu'un mot ou une expression, que le participant ne peut visionner qu'une seule fois. Parmi les vidéos de contrôle, 8 ont été extraites du corpus CLeLfPC, codées par des codeurs professionnels. Les deux autres ont été codées automatiquement par le PoC et avaient été validées par des experts comme étant correctes.

Durant le stage annuel de l'Association pour la Langue française Parlée Complétée (ALPC), *des personnes sourdes connaissant le code se sont portées volontaires* pour décoder les vidéos. Pour participer, un texte de consentement devait être approuvé. Les expériences étaient anonymes, aucune donnée personnelle n'a été recueillie. Une vidéo qui décrit le protocole a été présentée à chaque participant pour s'assurer que tous ont reçu la même information. Durant l'expérience, pour chacune des 34 vidéos codées, le participant devant remplir les 3 champs suivants d'un formulaire :

- `J'ai décodé : ..............................`
- `J'ai correctement décodé:` *non ... peut-être ... oui* (barre de progression)
- `J'ai un commentaire sur cette vidéo (optionnel)`

4.2 Résultats quantitatifs

Les évaluations ont été réalisées avec l'outil Sclite, un programme inclut dans SCTK, le *Nist Scoring Toolkit*, habituellement utilisé pour estimer les résultats des systèmes de reconnaissance automatique de la parole. Cet outil permet de comparer des phrases de référence - les phrases à trouver, avec les phrases dites hypothèses - les phrases produites par le système automatisé. Il utilise un algorithme permettant d'estimer le pourcentage de mots correctement reconnus, ainsi que le taux de mots substitués, supprimés et insérés dans l'hypothèse. Parmi les 19 participants, 14 ont été sélectionnés après vérification du taux de décodage des vidéos de contrôle. Effectivement, nous avons estimé que si un participant n'est pas en mesure de décoder les mots du contrôle au moins à hauteur de 40%, il n'est pas suffisamment qualifié pour évaluer notre système. La table 1 résume les scores, obtenus sur les 14 réponses sélectionnées, dans les deux conditions (vidéos contrôles et vidéos du PoC). Les scores de décodage avec un codeur professionnel sont nettement meilleurs que ceux obtenus avec le PoC. Ils correspondent en fait au taux maximal qu'il est possible d'obtenir par les participants, dans la condition de test réalisée. Les résultats de décodage des vidéos lors du codage automatique avec le PoC s'en approchent et sont très prometteurs ; ils sont suffisamment corrects pour valider la méthodologie proposée.

	# Mots	Correct	Substitution	Supression	Insertion
contrôle	356	77,8 %	15,4 %	6,7 %	3,9 %
PoC	828	67,5 %	23,3 %	9,2 %	6,3 %

TABLE 1 – Taux d'erreur de décodage des mots

4.3 Résultats qualitatifs

Nous avons analysé les commentaires des participants sur les différentes vidéos. Dans quelques cas, des erreurs de clés ont été soulevées ; elles sont dues, soit à une erreur de conversion graphème-phonème qui a amenée à un mauvais choix de clé, soit à l'accent. D'autres commentaires indiquent que la clé est trop "rapide", ce qui sous-entend que le PoC a sur-estimé le temps de transition et donc sous-estimé le temps d'exposition. Les autres commentaires portent sur l'aspect de la main dans la vidéo : trop transparente et trop floue. En revanche, aucun commentaire n'a porté sur le côté non-naturel de la trajectoire de la main et son angle constant. Toutes ces informations permettront d'établir des priorités sur les actions à mener lors de l'élaboration du système. Ci-après, se trouvent quelques uns des commentaires :

- `difficulté à savoir si c'est "à six" ou "assis"`
- `dans la vidéo, le "è" final de "sorbet", est représenté par la clé du "é"`
- `dernière clé trop rapide`
- `main mal positionnée sur la pommette`
- `la main est un peu trop transparente`
- `main pas assez claire : trop de flou entre deux clés`

Enfin, nous avons recueilli les impressions des participants après leur passage de l'expérience et avons obtenu un retour très positif. Il semble que le système, lorsqu'il sera en phase finale, serait susceptible de trouver sa place dans la communauté. Enfin, le manque de ressources numériques pour la LfPC (vidéos codées notamment) a été mentionné par presque tous les participants.

5 Conclusion et perspectives

En combinaison avec les mouvements labiaux, le « Cued Speech » rend les phonèmes d'une langue parlée visuellement différents les uns des autres. Les clés du codage sont positionnées autour du visage, près des lèvres, ce qui facilite le suivi simultané des mouvements des lèvres et des mouvements de la main codeuse. Cet article a décrit une preuve de concept de génération automatique du codage en Langue française Parlée Complétée. La preuve de concept proposée est en mesure de prédire 1/ la clé à coder (position et forme de la main), 2/ les moments pour l'afficher, en changer et la déplacer, 3/ l'emplacement de la main, et 4/ d'augmenter la vidéo avec une main artificielle à partir de ces informations. Ce PoC a été bien accueilli par la communauté concernée, et les évaluations quantitatives ont révélé son fort potentiel, ce qui permet de le considérer comme étant approuvé.

Dans un futur proche, nous développerons un système basé sur l'analyse des annotations du corpus CLeLfPC. Pour la question du "quoi", le système restera à base de règles de productions. Pour les questions de "quand" et "où" placer la clé, nous pensons utiliser des techniques d'apprentissage automatique avec des modèles prédictifs appris à partir des données annotées, afin de rendre le système plus efficace dans les différentes prédictions. Quant à la question du "comment", seule une collaboration avec les personnes concernées permettra de l'améliorer. Avec un tel système de codage automatique, toutes sortes de vidéos codées pourront être élaborées et diffusées pour tous les types d'utilisations. Dans le contexte du présent projet, des textes sélectionnés sur différents thèmes (Gala *et al.*, 2024) seront lus par un acteur, afin de collecter des vidéos. Une fois codées automatiquement, évaluées et sélectionnées, ces vidéos seront assemblées pour créer des capsules pédagogiques destinées au grand public, aux débutants apprenant le code et aux enfants sourds.

6 Reproductibilité

Toutes les données et tous les codes sources mentionnés dans ce document respectent les principes de la science ouverte. Le code source de la preuve de concept est déposé sous les termes de la licence libre GNU Affero General Public License v3. Il fait partie du logiciel SPPAS. Les codes sources pour réaliser les expérimentations décrites dans cet article sont sous la licence GNU AGPL v3 et les données utilisées sont soumis aux termes des licences ODbL (Open Database License 1.0) et CC-By-NC-4.0. Nous avons utilisé SPPAS 4.11 https://sppas.org/, et SCTK 2.4.12 https://github.com/usnistgov/SCTK.

7 Remerciements

Nous tenons à remercier tous les participants à l'expérience qui ont accepté de donner de leur temps pour nous aider, lors du stage annuel organisé par l'ALPC https://alpc.asso.fr. Nous remercions également l'ALPC de nous avoir offert l'opportunité d'assister au stage et de présenter notre travail.

Les recherches présentées dans cet article ont été réalisées dans le cadre d'un projet financé par la FIRAH sous la référence APa2022_022 https://auto-cuedspeech.org/, en collaboration avec les associations Datha https://datha.io/ et AISAC https://www.academieinternationale.org/.

Références

ABOUTABIT N. (2007). *Reconnaissance de la Langue Française Parlée Complété (LPC) : décodage phonétique des gestes main-lèvres.* Thèse de doctorat, Institut National Polytechnique de Grenoble - INPG.

ATTINA V. (2005). *La Langue Française Parlée Complétée : Production et Perception*. Thèse de doctorat, Institut National Polytechnique de Grenoble - INPG.

BIGI B. (2015). SPPAS - Multi-lingual Approaches to the Automatic Annotation of Speech. *The Phonetician*, **111–112**, 54–69.

BIGI B., ZIMMERMANN M. & ANDRÉ C. (2022). Clelfpc : a large open multi-speaker corpus of french cued speech. In *Proceedings of The 13th Language Resources and Evaluation Conference*, p. 987–994, Marseille, France : European Language Resources Association. https ://hal.archives-ouvertes.fr/hal-03794830.

BRATAKOS M. S. (1995). *The effect of imperfect cues on the reception of cued speech*. Thèse de doctorat, Massachusetts Institute of Technology.

BRATAKOS M. S., DUCHNOWSKI P. & BRAIDA L. D. (1998). Toward the automatic generation of cued speech. *Cued Speech Journal*, **6**, 1–37.

CATHIARD M.-A., ATTINA V. & ALLOATTI D. (2003). Labial anticipation behavior during speech with and without cued speech. In *Proceedings of the 15th International Congress of Phonetic Sciences*, p. 1935–1938, Barcelona, Spain.

CORNETT R. O. (1967). Cued speech. *American annals of the deaf*, p. 3–13.

CORNETT R. O., BEADLES R. & WILSON B. (1977). Automatic cued speech. In *Research Conference on Speech Processing Aids for the Deaf*, p. 224–239, Gallaudet College (USA).

DUCHNOWSKI P., BRAIDA L. D., LUM D., SEXTON M., KRAUSE J. & BANTHIA S. (1998). Automatic generation of cued speech for the deaf : status and outlook. In *International Conference on Auditory-Visual Speech Processing*, Sydney, Australia.

DUCHNOWSKI P., LUM D. S., KRAUSE J. C., SEXTON M. G., BRATAKOS M. S. & BRAIDA L. D. (2000). Development of speechreading supplements based on automatic speech recognition. *IEEE transactions on biomedical engineering*, **47**(4), 487–496.

GALA N., BIGI B. & BAUER M. (2024). Automatically estimating textual and phonemic complexity for cued speech : How to see the sounds from french texts. In *The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, Turin, Italy.

KAPLAN H. (1975). *The effects of cued speech on the speech-reading ability of the deaf.* Thèse de doctorat, ProQuest Information & Learning.

LEYBAERT J., COLIN C. & HAGE C. (2010). *Cued speech and cochlear implants*, p. 107–125.

NEEF N. A. & IWATA B. A. (1985). The development of generative lipreading skills in deaf persons using cued speech training. *Analysis and intervention in developmental disabilities*, **5**(4), 289–305.

SANKAR S., BEAUTEMPS D., ELISEI F., PERROTIN O. & HUEBER T. (2023). Investigating the dynamics of hand and lips in French Cued Speech using attention mechanisms and CTC-based decoding. In *Interspeech 2023 - 24th Annual Conference of the International Speech Communication Association*, Dublin, Ireland.

SEXTON M. G. (1997). *A video display system for an automatic cue generator*. Thèse de doctorat, Massachusetts Institute of Technology.