Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Voice2Picto : un système de traduction automatique de la parole vers des pictogrammes

Cécile Macaire[1] Emmanuelle Esperança-Rodier[1] Didier Schwab[1]
Benjamin Lecouteux[1]
(1) Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP*, LIG, 38000 Grenoble, France
prénom.nom@univ-grenoble-alpes.fr

RÉSUMÉ

Nous présentons Voice2Picto, un système de traduction permettant, à partir de l'oral, de proposer une séquence de pictogrammes correspondants. S'appuyant sur des technologies du traitement automatique du langage naturel, l'outil a deux objectifs : améliorer l'accès à la communication pour (1) les personnes allophones dans un contexte d'urgence médicale, et (2) pour les personnes avec des difficultés de parole. Il permettra aux personnes des services hospitaliers, et aux familles de véhiculer un message en pictogrammes facilement compréhensible auprès de personnes ne pouvant communiquer via les canaux traditionnels de communication (parole, gestes, langue des signes). Dans cet article, nous décrivons l'architecture du système de Voice2Picto et les pistes futures. L'application est en *open-source* via un dépôt Git https://github.com/macairececile/Voice2Picto.

ABSTRACT

Voice2Picto : an automatic speech translation system into pictograms

We present Voice2Picto, a translation system that uses spoken language to propose a sequence of corresponding pictograms. Based on natural language processing technologies, the tool has two objectives : to improve access to communication (1) for people not speaking the local language in medical settings, and (2) for people having a cognitive disorder. It will allow practitioners and families to convey a message in pictograms that can be easily understood by people who cannot communicate via traditional communication channels (speech, gestures, sign language). In this article, we describe the architecture of the Voice2Picto system and the future directions. The application is open-source via a Git repository https://github.com/macairececile/Voice2Picto.

MOTS-CLÉS : Traduction Automatique, Reconnaissance de la Parole, Pictogrammes, Communication Alternative et Augmentée.

KEYWORDS: Automatic Translation, Speech Recognition, Pictographs, Alternative and Augmentative Communication.

1 Introduction

Lorsqu'une personne ne peut utiliser les canaux traditionnels de communication (parole, gestes, langue des signes) pour exprimer un message, une communication alternative et augmentée (CAA) peut être mise en place. On retrouve dans la CAA, l'utilisation de pictogrammes, image représentant un concept plus ou moins concret (un mot unique, une entité nommée, une expression polylexicale

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 5 : démonstrations, pages 10–13.

par exemple [1]). Actuellement, la prise en main d'outils de CAA (par exemple, des classeurs de communication sous forme de pictogrammes) est longue et difficile (Cataix-Nègre, 2017). En effet, un temps d'adaptation et d'apprentissage du fonctionnement de chaque outil est nécessaire. De plus, un manque de soutien institutionnel est présent, entre des équipes médicales non formées, et des lenteurs administratives pour leur mise en place (on parle ici des instituts médicalisés). Pourtant, l'utilisation de la CAA a un impact social positif pour les personnes en situation de handicap. La Croix-Rouge a identifié une réduction du stress, une amélioration de l'autonomie et de l'état de santé, ainsi qu'une meilleure sérénité et plaisir des personnes dans leur vie quotidienne (Croix-Rouge, 2021). L'accès à la communication par tous et pour tous reste donc un défi majeur. Proposer une séquence de pictogrammes à partir de la voix permettra aux personnes des services hospitaliers, et aux familles sans connaissances préalables de véhiculer un message en pictogrammes facilement compréhensible par les utilisateurs de CAA.

2 Présentation du système

Voice2Picto est un système cascade de traduction automatique de la parole vers des pictogrammes combinant différents domaines du traitement automatique du langage naturel. L'utilisation d'une telle architecture permet d'améliorer l'explicabilité du modèle, car nous pouvons évaluer les capacités et donc l'impact positif ou négatif d'une phase sur la suivante. Son architecture globale est présentée Figure 1 et s'articule entre quatre modules.

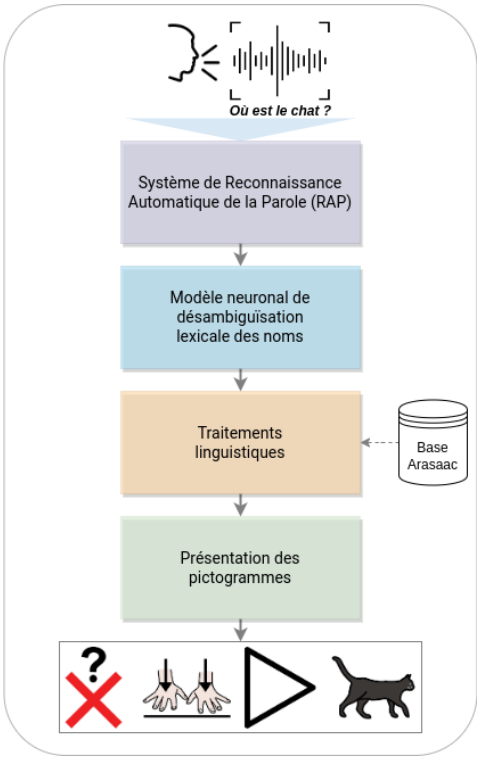

FIGURE 1 – Schéma de Voice2Picto.

Le premier module est un système de reconnaissance automatique de la parole (RAP), basé sur une boîte à outils de reconnaissance vocale Vosk [2]. L'API permet une réponse presque instantanée grâce à l'utilisation de modèles optimisés, améliorant ainsi les performances de l'application. Le modèle appelé a été entraîné avec la boîte à outils Kaldi (Povey *et al.*, 2011) et obtient 23.95% de taux d'erreurs au niveau des mots sur CommonVoice (Ardila *et al.*, 2020).

À partir de la transcription définie par le module de RAP, un modèle neuronal de désambiguïsation lexicale sur les noms est appliqué, proposé dans Vial *et al.* (2019). Récupérer le sens le plus probable a son importance. Certains termes n'ont pas la même signification compte tenu du contexte (par exemple, *café* peut désigner la graine, la boisson ou le lieu de consommation). Le système neuronal comporte, en entrée, des couches d'encodeur de type *transformers* et utilisent les vecteurs de mots contextuels pré-entrainés de type BERT (Devlin *et al.*, 2019). L'architecture se compose d'une couche de décodage correspondant à un classificateur softmax capable de classer un mot dans l'ensemble des synsets possibles de WordNet (Miller, 1995). Le sens WordNet le plus probable

1. Les pictogrammes sont disponibles sur Arasaac : https://arasaac.org/pictograms/search
2. https://github.com/alphacep/vosk-api

est alors prédit. Dans Voice2Picto, le modèle utilisé a été entrainé sur deux corpus SemCor et Wngt adaptés au français (Le *et al.*, 2020). Le score F1 sur le corpus d'évaluation SemEval 2013 est de 51.28%.

Un module de traitements linguistiques vient compléter ces deux modules. L'objectif ici est de "nettoyer" le texte, pour ensuite identifier les pictogrammes associés à chaque terme. Ainsi, une traduction pertinente sera proposée. Une phase de tokenisation, de lemmatisation, et d'étiquetage morphosyntaxique est appliquée en premier (à l'aide de la librairie spacy [3]). Puis, les expressions multi-mots et les entités nommées sont recherchées. Une table d'alignement entre lemmes, sens WordNet et identifiants ARASAAC permet de récupérer l'identifiant associé à chaque lemme ou au sens WordNet dans le cas d'un nom. Enfin, chaque image est récupérée dans la banque d'Arasaac grâce à l'identifiant, puis affichée.

L'interface de l'application est présentée Figure 2. Elle se veut simple d'utilisation, rapide et ergonomique. Les utilisations possibles sont nombreuses, voire infinies puisqu'elles peuvent rentrer dans de nombreuses situations de la vie quotidienne, telles que communiquer un besoin, une envie, une émotion, raconter une histoire ou poser une question à une personne.

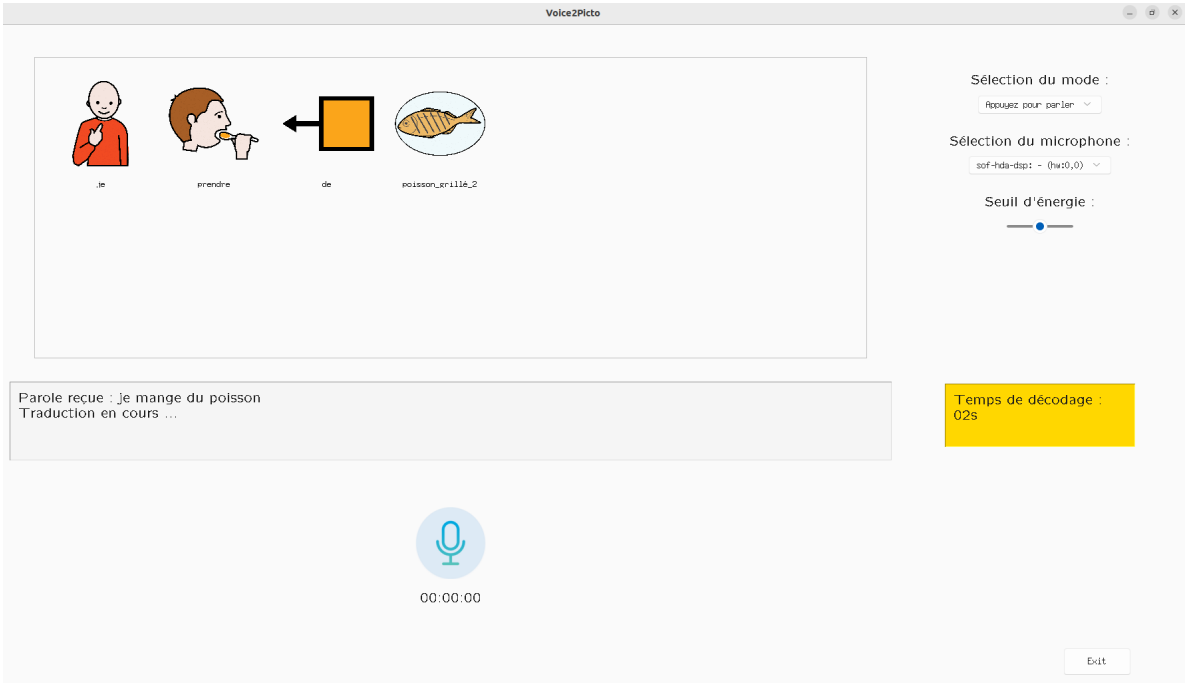

FIGURE 2 – Interface de Voice2Picto.

3 Conclusion & Perspectives

Voice2Picto facilite l'accès à la communication à l'aide de pictogrammes pour un public qui ne connaît pas ou n'utilise pas ce type de support. Les familles et les professionnels de la santé peuvent ainsi disposer d'un outil peu coûteux et facile à mettre en place pour aider les personnes qui utilisent les pictogrammes comme moyen de communication. L'outil étant libre, celui-ci pourra ainsi évoluer grâce

3. https://spacy.io/models/fr

aux retours reçus et aux avancées technologiques sur lesquelles il repose. Le code est disponible sur GitHub [4]. Les perspectives futures sont multiples : améliorer la traduction en définissant un vocabulaire précis de pictogrammes à utiliser (actuellement, un terme peut avoir plusieurs pictogrammes associés), gérer les termes non disponibles sous forme pictographique, utiliser une grammaire experte pour orienter les règles de traduction et évaluer la traduction et l'impact de chaque module auprès d'experts.

Remerciements

Ce travail a bénéficié d'un financement du Fond National Suisse (No. 197864) et de l'Agence Nationale de la Recherche, via le projet PROPICTO (ANR-20-CE93-0005). Ce travail a également bénéficié d'un accès aux moyens de calcul de l'IDRIS au travers de l'allocation de ressources 2022-AD011013625 attribuée par GENCI.

Références

ARDILA R., BRANSON M., DAVIS K., KOHLER M., MEYER J., HENRETTY M., MORAIS R., SAUNDERS L., TYERS F. & WEBER G. (2020). Common voice : A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 4218–4222, Marseille, France : European Language Resources Association.

CATAIX-NÈGRE E. (2017). *Communiquer autrement : Accompagner les personnes avec des troubles de la parole ou du langage*. APPRENDRE ET RÉAPPRENDRE. De Boeck Supérieur.

CROIX-ROUGE (2021). Communiquons autrement - déploiement de la communication alternative améliorée dans les établissements handicap de la croix-rouge française.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2020). FlauBERT : Unsupervised language model pre-training for French. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 2479–2490, Marseille, France : European Language Resources Association.

MILLER G. A. (1995). Wordnet : a lexical database for english. *Communications of the ACM*, **38**(11), 39–41.

POVEY D., GHOSHAL A., BOULIANNE G., BURGET L., GLEMBEK O., GOEL N., HANNEMANN M., MOTLICEK P., QIAN Y., SCHWARZ P. *et al.* (2011). The Kaldi Speech Recognition Toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, volume CONF : IEEE Signal Processing Society.

VIAL L., LECOUTEUX B. & SCHWAB D. (2019). Sense Vocabulary Compression through the Semantic Knowledge of WordNet for Neural Word Sense Disambiguation. In *Proceedings of the 10th Global Wordnet Conference*, Wroclaw, Poland.

4. https://github.com/macairececile/Voice2Picto