Analyse des performances de systèmes de reconnaissance automatique de la parole spontanée après cancer oral ou oropharyngé

Mathieu BALAGUER[1]
Julien PINQUIER[1]
Virginie WOISARD[2,3]
Jérôme FARINAS[1]

[1] IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3, Toulouse, France
[2] Hôpital Larrey, Hôpitaux de Toulouse, France
[3] Laboratoire de Neuro-Psycho-Linguistique LNPL, Université Toulouse II, France

Contexte : L'analyse de la parole spontanée après cancer est essentielle en clinique, car il s'agit de la situation de production la plus « écologique » (Knuijt et al., 2017, Prins & Bastiaanse, 2004). Elle nécessite de pouvoir s'appuyer sur une transcription fiable. Actuellement, les systèmes de reconnaissance automatique de parole facilitant la transcription sont entraînés sur la parole typique, et prennent mal en compte les spécificités de la parole après cancer (Balaguer et al., 2021). Or, une reconnaissance lexicale plus juste permettrait des analyses plus précises du discours des patients, et l'accès à de nouvelles informations sur la dynamique psychosociale des patients.

Objectifs : Étudier les performances de systèmes de reconnaissance automatique de parole (RAP) au niveau du mot appliqués à la parole spontanée après cancer.

Méthodes : La parole spontanée de 12 patients traités pour un cancer oral ou oropharyngé a été enregistrée au cours d'un entretien. Une minute a été transcrite au niveau orthographique par un orthophoniste expert et deux sujets naïfs, sur l'hypothèse qu'un système entraîné sur de la parole typique aurait un comportement plus proche d'un auditeur naïf par réduction des effets d'expertise (Fex, 1992). Une transcription automatique au niveau du mot a été réalisée

par plusieurs systèmes de RAP (Tableau 1). La boîte à outils HTK (Young et al., 2015) a permis d'obtenir des scores de comparaison avec les transcriptions des experts puis avec les naïfs. Des tests de comparaison 2 à 2 (test de Wilcoxon) ont été menés pour comparer les performances des systèmes entre eux, puis les performances de chaque système selon l'expertise du juge humain (expert *vs* naïfs).

Tableau 1. Systèmes de RAP étudiés

Système	Architecture	Corpus d'entraînement (heures)
CRDNN	End-to-end Convolutional Recurrent Deep Neural Network (Heba et al., 2021)	LibriSpeech (960 h)
TDDNf	Time-Delay Neural Network factorisé – Hidden Markov Model (Gelin et al., 2021)	CommonVoice (147 h)
Whisper (Open AI)	Transformer Seq2seq (Radford et al., 2022)	Multilingue (117.113 h)

Résultats : Whisper présente une exactitude significativement meilleure (p=0,004) que le TDNNf bien que restant faible (taux d'erreurs mots = 76,47 % par rapport à l'expert !), et un pourcentage de mots corrects, un nombre de mots corrects et de délétions significativement meilleurs que les deux autres systèmes (p<0,05), tant en comparaison avec l'expert qu'avec les naïfs. En revanche, le nombre d'insertions et de substitutions est significativement plus élevé que pour le CRDNN (Tableau 2).

Tableau 2. Comparaison des systèmes de RAP selon le profil expert/naïf de l'examinateur (en gras, la meilleure valeur par métrique)

	Expert			Naïfs		
	CRDNN	TDNNf	Whisper	CRDNN	TDNNf	Whisper
Exactitude $Ex = \frac{100 \times (H-I)}{H+S+D}$	21,17 %	13,79 %	**23,53 %**	21,07 %	13,40 %	**27,43 %**
Pourcentage de mots corrects $P_{Corr} = \frac{100 \times H}{H+S+D}$	21,88 %	16,33 %	**37,87 %**	22,11 %	16,33 %	**36,44 %**
Nombre de mots correctement reconnus (H)	20	14,33	**32,83**	18,83	13,54	**29,42**
Nombre de délétions (D)	39,83	26,67	**15,92**	32,58	22,04	**10,25**
Nombre de substitutions (S)	**18,33**	37,17	29,42	**18,42**	37,50	31
Nombre d'insertions (I)	**0,5**	1,75	8	**0,75**	2,21	9,83

Les systèmes présentent des performances globalement similaires, que leur référence de comparaison soit un expert ou les naïfs. Seul Whisper semble avoir une transcription plus proche de celle des naïfs avec une exactitude significativement plus élevée comparativement aux experts (Tableau 3).

Tableau 3. Comparaison des performances de chaque système de RAP selon le profil expert/naïf de l'examinateur (p-value, test de Wilcoxon : en gras, les p-values significatives au seuil de 5 %, les valeurs de chaque système sont à retrouver dans le Tableau 2)

Système	CRDNN	TDNNf	Whisper
Exactitude	0,84	0,16	**0,049**
Pourcentage de mots corrects	0,5	0,75	0,18
Mots corrects	**0,02**	**0,02**	**0,003**
Délétions	**0,02**	**0,02**	**0,004**
Substitutions	0,28	0,24	0,14
Insertions	0,16	0,66	0,11

Conclusion : Cette étude exploratoire montre que l'architecture des systèmes de RAP et la taille de leur corpus d'entraînement (sur parole typique) influencent les performances de reconnaissance de la parole après cancer. Toutefois, l'exactitude reste faible. Cela montre la nécessité d'adapter les modèles acoustiques à la parole cancérologique pour permettre des analyses plus précises du discours informant sur la dynamique psychosociale des patients et l'ajustement des stratégies thérapeutiques. Ces résultats préliminaires devront être étudiés sur un échantillon plus large, notamment pour analyser l'effet de la sévérité du trouble de parole sur les performances de reconnaissance.

Références bibliographiques

Balaguer, M. (2021). Mesure de l'altération de la communication par analyses automatiques de la parole spontanée après traitement d'un cancer oral ou oropharyngé. Thèse de doctorat, Université Paul Sabatier Toulouse III. https://theses.hal.science/tel-03557511

Fex, S. (1992). Perceptual evaluation. *Journal of Voice*, 6(2), 155-158

Gelin, L., Daniel, M., Pinquier, J., & Pellegrini, T. (2021). End-to-end acoustic

modelling for phone recognition of young readers. *Speech Communication*, 134, 71–84. https://doi.org/10.1016/j.specom.2021.08.003

Heba, A. (2021). *Reconnaissance automatique de la parole à large vocabulaire : des approches hybrides aux approches End-to-End* [Université Paul Sabatier Toulouse III]. https://tel.archives-ouvertes.fr/tel-03616588

Knuijt, S., Kalf, J. G., van Engelen, B. G. M., de Swart, B. J. M., & Geurts, A. C. H. (2017). The Radboud Dysarthria Assessment: Development and Clinimetric Evaluation. *Folia Phoniatrica et Logopaedica*, 69(4), 143–153. https://doi.org/10.1159/000484556

Prins, R., & Bastiaanse, R. (2004). Analysing the spontaneous speech of aphasic speakers. *Aphasiology*, 18(12), 1075–1091. https://doi.org/10.1080/02687030444000534

Radford, A., Wook, J., Tao, K., Greg, X., Christine, B., & Ilya, M. (2021). Robust Speech Recognition via Large-Scale Weak Supervision. Openai.Com. https://doi.org/10.48550/arXiv.2212.04356

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X. (Andrew), Moore, G., Odell, J., Ollason, D., Pover, D., Ragni, A., Valtchev, V., Woodland, P., & Zhang, C. (2015). The HTK Book (for HTK Version 3.5, documentation alpha version) (Issue December). Cambridge University Engineering Department.