Exploration de systèmes end-to-end pour la reconnaissance automatique de la parole spontanée

Solène Evain, Solange Rossato, Benjamin Lecouteux, François Portet

Exploration de systèmes *end-to-end* pour la reconnaissance automatique de la parole spontanée

Solène Evain[1], Solange Rossato[1], Benjamin Lecouteux[1], François Portet[1]
(1) Univ. Grenoble Alpes, CNRS, Grenoble-INP, LIG, 38000 France
prenom.nom@univ-grenoble-alpes.fr

MOTS-CLÉS : Reconnaissance automatique de la parole, système end-to-end, parole spontanée.

KEYWORDS: Automatic speech recognition, end-to-end system, spontaneous speech.

Ces dernières années, les systèmes de Reconnaissance Automatique de la Parole (RAP) ont donné de très bons résultats sur les benchmarks de la communauté. Si ces résultats sont très bons sur la parole lue ou médiatique, les performances baissent considérablement pour la Reconnaissance de la Parole Spontanée (RAPS), notamment à cause de la faible disponibilité des corpus et de la difficulté de définir et de modéliser ce type de parole. Dans ce travail, nous souhaitons explorer l'utilisation d'un modèle neuronal pour la RAPS. En effet, l'optimisation *end-to-end* (de bout en bout) de ces modèles – sans modèle de langue *a priori* et en partie sans corpus annoté – offre non seulement des performances intéressantes, mais également l'opportunité d'étudier la modélisation de la parole spontanée uniquement à partir de données.

1 Vers une définition de la parole spontanée

Il est difficile de définir précisément ce qu'est la parole spontanée. D'une part, elle admet plusieurs dénominations : parole spontanée, *casual speech* (parole détendue) (Torreira et al., 2010), parole non scriptée (Llisterri, 1992), populaire (Guiraud, 1965), informelle, familière, non standard (Blanche-Benveniste, 1999), non préparée (Dufour et al., 2010), non planifiée (Veiga et al., 2012), non conventionnelle (Caron, 1992), conversation naturelle ou encore conversationnelle (Shriberg, 2005). D'autre part, Fujisaki (Fujisaki, 1997) parlait en 1997 d'un "continuum du degré de spontanéité", avec d'un côté la parole la plus préparée (la plus contrainte) et d'un autre la moins préparée (la plus libre). Derrière cette notion de continuum réside l'idée qu'il n'y a pas de limite franche entre parole préparée et parole spontanée. Enfin, certains articles font également état de niveaux de spontanéité, basés sur le nombre d'hésitations (Veiga et al., 2012) ou sur l'intelligibilité (Dufour et al., 2010). Le lien entre la situation d'énonciation, le degré d'intimité entre les locuteurs et le mode de communication (présentiel/distanciel, devant une foule etc) peut également permettre de déterminer un niveau de spontanéité. Ces trois points nous démontrent qu'il est difficile d'avoir une définition précise de ce qu'est la parole spontanée mais que l'on peut la lier au fait de parler sans contrainte (parole naturelle) et sans préparation (non préparée) et qui semble apparaître dans des contextes où peu d'attention est portée sur la forme du discours et où l'aspect conventionnel de la langue peut être omis. Luzzati (Luzzati, 2007) la définit comme le fait de "quitter l'univers de la phrase, celui de la langue préparée, celui de l'erreur qui s'efface et se rature, pour basculer du côté des énoncés non prémédités, dont l'émetteur est le premier auditeur, dans lesquels l'erreur se traduit par un allongement du message".

D'un point de vue caractéristique, la parole spontanée est notamment représentée par des disfluences (amorces de mots, hésitations, reprises, répétitions…) (Adda-Decker et al., 2004), une vitesse d'élocution assez rapide (Adda-Decker et al., 2012) ou à débit variable, une réduction temporelle fréquente (Wu and Adda-Decker, 2020), de nombreux allongements (Duez, 2001) et une prononciation peu soignée (hypoarticulation) (Dufour, 2008).

2 Les corpus de parole spontanée en français

Dans ce travail, nous avons recensé seize corpus du français comprenant de la parole spontanée, comme le montre le tableau 1. La durée totale de ces corpus est d'environ 1825 heures, ce qui est à

Corpus (date)	**Type parole**	**Type d'interaction**	**Durée** (approx.)
ESLO1 (1968-71)	Spont.	repas, entretiens	318 h
ESLO2 (2008-)	Spont.	repas, entretiens	450 h
NCCFr (2010)	Spont.		36 h
MPF (2010-14 ?)	Spont.	entretiens	78 h
PFC (1999)	Lue/Spont.	lecture de liste de mots/textes, entretiens, conversations libres	>300 h ?
CFPP2000 (2005- ?)	Spont.	entretiens	38 h* / 58 h 40
CLAPI (1998-)	Spont./Prep.	conversations, réunions, visites guidées	16 h 30*
C-ORAL-ROM (2001-2003)	Spont./Prep.		22 h*
CRFP (1998-2002)	Spont./Prep.	souvenirs, théâtre, émissions, cours	34 h*
FLEURON (2009-2012)	Prep./Spont.	Interactions étudiants/administration	3 h 25*
OFROM (2008-2012)	Spont./Prep.	discussions, entretiens, communications	25 h 13*
TCOF (2005-2009)	Spont./Prep.	discussions, entretiens, réunions, débats	28 h40* / 61 h
TUFS (2005-2011)	Spont.	interviews, entretiens, discussions	52 h 40*
CFPB (2013-2015)	Spont.	entretiens	5 h* / 21 h 30
Réunions (2007-2008)	Prep./Spont.	réunions	18 h*
Valibel (1998-2008)	Spont./Prep.	interviews, discours, entretiens, journaux, souvenirs, conversations...	43 h 25* / 331 h
Total			1825 h 58

*Compris dans le corpus CEFC ; 'Spont.' : spontanée, 'Prep.' : préparée

TABLE 1 – Corpus de parole spontanée en français

première vue tout à fait acceptable pour pouvoir apprendre des modèles de RAP. Cependant, cette durée globale n'est pas la durée effective de parole spontanée disponible. En effet, neuf corpus sur seize comprennent à la fois de la parole spontanée et de la parole préparée et/ou lue, sans que l'on sache quelle est la proportion de chacun de ces types de parole. De plus, l'accès à certains de ces corpus n'est pas toujours aisé. Hormis certaines contraintes administratives (signature d'un contrat), il existe des corpus pour lesquels les sites ne permettent pas le téléchargement groupé ce qui oblige à aller chercher les fichiers un à un. Certains sont difficilement accessibles pour des raisons d'anonymisation et de traitement/transcription en cours. Enfin, dans des cas extrêmes, le contact avec les chercheurs ayant constitué les corpus est difficile, voire impossible, dû à peu de disponibilité de leur part ou à des départs (certains corpus sont assez anciens) ce qui enlève tout accès aux données. Par ailleurs, il convient de noter que pour beaucoup de corpus, des outils et des conventions de

transcription différentes ont été employés, ce qui rend leur fusion fastidieuse. Enfin, d'un point de vue plus technique, le matériel ou les conditions d'enregistrement, font qu'une partie des enregistrements sont inexploitables pour un objectif de RAP.

3 RAP et importance des données

L'accès à des données audio annotées est primordial pour l'entraînement d'un système de RAP. L'étude de (Lamel et al., 2002) montre que le *Word Error Rate* ou *WER* (taux d'erreur de mots) est dépendant de la quantité de données d'apprentissage. Or, l'accès à une grande quantité de données n'est possible que pour quelques dizaines de langues sur les 7 000 existantes. Les bonnes performances observées aujourd'hui sont donc plus représentatives de la reconnaissance de langues telles que l'anglais ou le français. La reconnaissance de la parole lue, préparée et spontanée ne sont pas non plus équivalentes (Tancoigne et al., 2020). Là aussi, le manque de données est problématique, notamment pour la parole spontanée. Enfin, lorsque l'apprentissage et le décodage se font sur des données issues du même corpus, il est difficile d'évaluer réellement la capacité de généralisation du système. En effet, lors d'un décodage sur des données issues d'un corpus différent de celui d'apprentissage, le *WER* augmente (Likhomanenko et al., 2021).

4 Les limites d'un système à base de HMM pour la RAPS

Pour un système de RAP de type *HMM-GMM/DNN*, l'objectif est de trouver la séquence de mots la plus vraisemblable étant donnée une séquence de paramètres acoustiques. Un tel système, s'appuie sur un modèle acoustique, un lexique phonétisé et un modèle de langue donnant la probabilité d'une séquence de mots. Le poids accordé au modèle de langue dans ce type d'architecture est important. La parole spontanée étant différente de la parole lue ou préparée, sa reconnaissance nécessite d'adapter un système de RAP directement sur ce type de parole. À la fin des années 2000, la quantité de données nécessaires pour construire un modèle de langue était de plusieurs dizaines à plusieurs centaines de millions de mots et de plusieurs dizaines à plusieurs centaines d'heures (Pellegrini, 2008) pour l'adaptation d'un modèle acoustique. Or, les corpus de parole spontanée annnotés étant assez peu nombreux, les modèles de langue sont alors souvent construits sur d'autres types de données textuelles comme des journaux ou des transcriptions de parole lue ou préparée. La parole spontanée comprenant des disfluences (répétitions, allongements...) dues à la construction du message au cours de sa réflexion n'est alors pas représentée dans le modèle de langue ayant pourtant un fort impact sur le processus de décodage de la parole. En ce qui concerne le lexique phonétisé, la difficulté pour la RAPS réside dans la présence de nombreuses variantes lexicales (dues notamment au phénomène d'hypoarticulation) qu'il serait coûteux de toutes représenter.

5 L'apport des systèmes neuronaux

Étant donné le manque de données de parole spontanée transcrite en français et la complexité de traitement de ce type de parole, l'utilisation d'un système de RAP de type HMM-GMM/DNN ne semble pas le plus adéquat. Nous nous interrogeons donc aujourd'hui sur l'apport possible d'un

système *end-to-end* n'utilisant plus de lexique phonétisé et dont l'usage d'un modèle de langue externe devient optionnel : ceci permet d'injecter moins d'*a priori* sur la langue dans le système de RAP. De plus, l'association étant directe entre le signal en entrée et sa transcription en sortie (Chan et al., 2016) , le système apprend ses propres représentations. Néanmoins, ce type de système est très gourmand en données. Nous pensons étudier les systèmes pré-entraînés sur le français, tels que ceux appris pour LeBenchmark (Evain et al., 2021) pour pallier le manque de données d'apprentissage de parole spontanée. Cette technique consiste à apprendre des représentations globales de la langue de façon auto-supervisée, grâce à une grande quantité de données non-annotées. L'étude de (Baevski et al., 2020) montre que l'utilisation d'un modèle pré-entraîné suivi d'un ajustement du modèle sur 10 heures de parole annotées donne un meilleur WER (même sans modèle de langue) que l'utilisation d'un système HMM-DNN entraîné avec 100 heures de données annotées. Cette étude a été faite sur l'anglais, avec des modèles appris sur des corpus de lecture pour une tâche de RAP sur de la lecture. L'impact de l'utilisation de modèles pré-entraînés sur la parole spontanée en français reste donc à explorer et à analyser (Chung et al., 2020).

6 Perspectives

Ce travail a pour but d'explorer l'utilisation de systèmes de RAP *end-to-end* avec utilisation de modèles pré-entraînés sur le français pour la reconnaissance de la parole spontanée. Nous souhaitons évaluer l'influence du degré de spontanéité (ESLO2), du degré d'interaction (CRFP, Valibel, ESLO2) et du niveau d'intimité entre les locuteurs (ESLO2) sur la reconnaissance de cette parole. Pour cela, plusieurs corpus de test seront sélectionnés ou "composés" en respectant les critères suivants : premièrement, les données ne devront pas avoir été utilisées pour l'élaboration des modèles pré-entraînés. Ensuite, la qualité des fichiers audio doit être suffisante pour une tâche de RAP (nous ne nous focalisation pas sur la parole bruitée). Enfin, les enregistrements devront comprendre 3 locuteurs maximum, ceci afin d'éviter les situations de schismes interactionnels [1]. Par "composition" d'un corpus de test, nous entendons le rassemblement de fichiers audio de différents corpus. La grille d'analyse des résultats comprendra plusieurs niveaux : prosodique, morphologique et grammatical, afin d'étudier la gestion d'événements propres à la parole spontanée (impact de la segmentation des données en entrée, gestion des nouveaux mots et des suites de mots "non conventionnelles"). À des fins de comparaison, notre système sera également évalué sur un corpus de lecture (Commonvoice). En ce qui concerne le système de RAP, nous allons réutiliser l'architecture de type encodeur-décodeur utilisée dans la partie RAP de LeBenchmark (Evain et al., 2021) : *CRDNN (VGG-RNN-DNN) - Joint CTC/attention LSTM*, prenant en entrée des représentations de type Wav2vec.

7 Adéquation aux thématiques du GDR LIFT

La parole spontanée partage avec les langues peu dotées la caractéristique du manque de ressources. Nous espérons ainsi que les techniques d'analyse et d'apprentissage qui seront définies dans ce travail pourront ouvrir des perspectives nouvelles pour l'analyse linguistique, que ce soit pour collecter et annoter des données (systèmes de RAP) ou pour extraire ou vérifier des généralisations linguistiques (analyse des représentations apprises par le modèle).

1. Lorsqu'il y a plus de trois personnes en interaction, plusieurs conversations peuvent avoir lieu en parallèle.

8 Remerciements

Ce travail a été partiellement supporté par MIAI@Grenoble Alpes, (ANR-19-P3IA-0003).

Références

Adda-Decker, M., Fougeron, C., Gendrot, C., Delais-Roussarie, E., and Lamel, L. (2012). La liaison dans la parole spontanée familière : une étude sur grand corpus. *Revue francaise de linguistique appliquee*, Vol. XVII(1) :113–128. Bibliographie_available : 1 Cairndomain : www.cairn.info Cite Par_available : 1 Publisher : Publications linguistiques.

Adda-Decker, M., Habert, B., Barras, C., Adda, G., de Mareüil, P. B., and Paroubek, P. (2004). Une étude des disfluences pour la transcription automatique de la parole spontanée et l'amélioration des modèles de langage. In *Actes des 25èmes Journées d'Etudes sur la Parole (JEP 2004)*, Fès, Maroc.

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0 : A Framework for Self-Supervised Learning of Speech Representations. In *Proceedings of the 34th conference on Neural Information Processing Systems (NeurIPS 2020)*, page 12, Vancouver, Canada.

Blanche-Benveniste, C. . a. (1999). "Français parlé - oral spontané". Quelques réflexions. *Revue française de linguistique appliquée*, IV(2) :21.

Caron, P. (1992). L'écriture de la noblesse vers 1680. In *Grammaire des fautes et français non conventionnels*. Paris, France, presses de l'ecole normale supérieure, rue d'ulm edition.

Chan, W., Jaitly, N., Le, Q., and Vinyals, O. (2016). Listen, attend and spell : A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4960–4964. ISSN : 2379-190X.

Chung, Y.-A., Tang, H., and Glass, J. (2020). Vector-Quantized Autoregressive Predictive Coding. In *Proceedings of Interspeech 2020*, pages 3760–3764, Shanghai, China. ISCA.

Duez, D. (2001). Signification des hésitations dans la productione t la perception de la parole spontanée. *Parole*, (17/18/19) :113–138.

Dufour, R. (2008). From prepared speech to spontaneous speech recognition system : a comparative study applied to French language. In *Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology*, CSTST '08, pages 595–599, New York, NY, USA. Association for Computing Machinery.

Dufour, R., Estève, Y., and Deléglise, P. (2010). Automatic indexing of speech segments with spontaneity levels on large audio database | Proceedings of the 2010 international workshop on Searching spontaneous conversational speech. In *SSCS '10 : Proceedings of the 2010 international workshop on Searching spontaneous conversational speech*, Firenze, Italy.

Evain, S., Nguyen, H., Le, H., Boito, M., Mdhaffar, S., Alisamir, S., Tong, Z., Tomashenko, N., Dinarelli, M., Parcollet, T., Allauzen, A., Estève, Y., Lecouteux, B., Portet, F., Rossato, S., Ringeval, F., Schwab, D., and Besacier, L. (2021). LeBenchmark : A Reproducible Framework for Assessing Self-Supervised Representation Learning from Speech. In *Proceedings of Interspeech 2021*, Brno, Czechia.

Fujisaki, H. (1997). Prosody, Models, and Spontaneous Speech. In *Computing Prosody*. Springer, New York, NY, sagisaka y., campbell n., higuchi n. (eds) edition.

Guiraud, P. (1965). *Le français populaire*. Number 1172 in "Que sais-je ?". Paris, France, presses universitaires de france edition.

Lamel, L., Gauvain, J.-L., and Adda, G. (2002). Unsupervised acoustic model training. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '02)*, pages I–877–I–880.

Likhomanenko, T., Xu, Q., Pratap, V., Tomasello, P., Kahn, J., Avidov, G., Collobert, R., and Synnaeve, G. (2021). Rethinking Evaluation in ASR : Are Our Models Robust Enough ? In *Proceedings of Interspeech 2021*, pages 311–315, Brno, Czechia. ISCA.

Llisterri, J. (1992). Speaking styles in speech research. Dublin, Ireland.

Luzzati, D. (2007). Le dialogue oral spontané : quels objets pour quels corpora. *Revue d'Interaction Homme-Machine*, 8(2).

Pellegrini, T. (2008). *Transcription automatique de langues peu dotées*. PhD thesis, Université Paris Sud - Paris XI, Paris, France.

Shriberg, E. (2005). Spontaneous Speech : How People Really Talk and Why Engineers Should Care. In *Proceedings of Interspeech 2005*, Lisbon, Portugal.

Tancoigne, E., Corbellini, J.-P., Deletraz, G., Gayraud, L., Ollinger, S., and Valéro, D. (2020). La transcription automatique : un rêve enfin accessible ? Technical report, MATE-SHS.

Torreira, F., Adda-Decker, M., and Ernestus, M. (2010). The Nijmegen Corpus of Casual French. *Speech Communication*, 52(3) :201. Publisher : Elsevier : North-Holland.

Veiga, A., Candeias, S., Celorico, D., Proença, J., and Perdigão, F. (2012). Towards Automatic Classification of Speech Styles. In *Proceedings of the 10th international conference on Computational Processing of the Portuguese Language*, Coimbra, Portugal. Pages : 426.

Wu, Y. and Adda-Decker, M. (2020). Réduction temporelle en français spontané : où se cache-t-elle ? Une étude des segments, des mots et séquences de mots fréquemment réduits. In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)*, volume Volume 1 : Journées d'Études sur la Parole, pages 627–635, Nancy, France.