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, Solange Rossato, Benjamin Lecouteux, François Portet

Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France

MOTIVATIONS/CONTEXTE

Constat: très bons résultats des systèmes de **RAP** avec de la parole lue ou formelle journalistique. Mais grande variabilité des résultats avec de la **PS**.

Etude	Type parole	*WER*
(Mekki, 2020)	Lue	9,39%
(Desnous & al., 2018)	Journalistique	11,59%
(Elloumi, 2019)	Spontanée (radio/télé)	23,23% à 45,15%
(Tancoigne & al., 2020)	Spontanée (réunions)	88,4%

Problèmes: difficulté de modélisation & peu de données disponibles
Solution envisagée: nouveaux systèmes de **RAP** ***E2E***, sans **ML**, avec modèles pré-entraînés de façon auto-supervisée → des pistes prometteuses pour la **RAPS**.

DÉFINIR LA PAROLE SPONTANÉE

- Plusieurs dénominations: *casual speech*, conversation naturelle, informelle, non scriptée, populaire, familière, non conventionnelle...
- Continuum entre parole préparée et spontanée (Fujisaki, 1997)
- ≠ façons d'attester d'un niveau de spontanéité (nombre d'hésitations, intelligibilité)

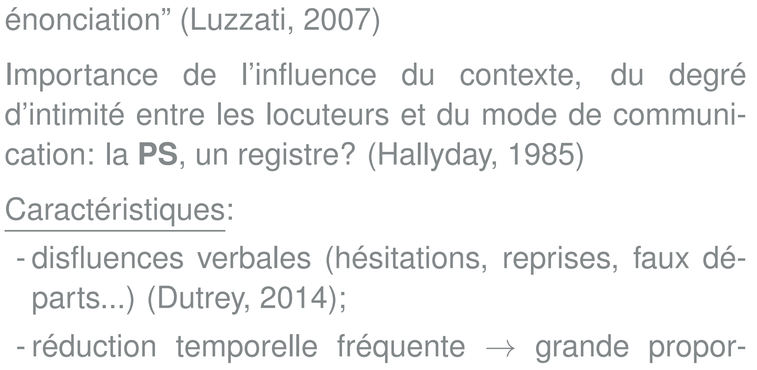
"Énoncés oraux conçus et perçus dans le fil de leur énonciation" (Luzzati, 2007)

Importance de l'influence du contexte, du degré d'intimité entre les locuteurs et du mode de communication: la **PS**, un registre? (Hallyday, 1985)

Caractéristiques:

- disfluences verbales (hésitations, reprises, faux départs...) (Dutrey, 2014);
- réduction temporelle fréquente → grande proportion de phones <30 ms observée (Wu Adda-Decker, 2020);
- allongement de la voyelle finale du groupe accentuel (Blanche-Benveniste, 2011);
- mots articulés avec moins de précision (Wu Adda-Decker, 2020);
- "agrammaticalité" des énoncés (Dufour, 2010).

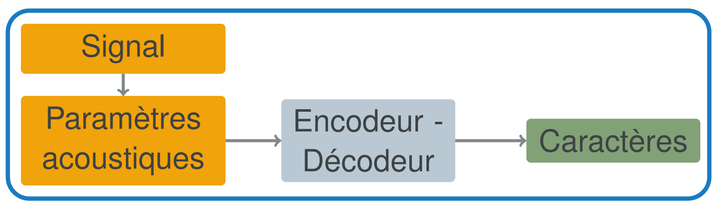
LIMITES D'UN SYSTÈME À BASE DE HMM



- Poids important accordé au modèle de langue
- Modèle de langue = +10/100k millions de mots
- Adaptation modèle acoustique = +10/100 heures de parole annotée
- Certaines variantes de prononciation ajoutées à la main dans le lexique

Présence de nombreuses variantes lexicales en **PS** qu'il serait coûteux de toutes représenter.
Grande proportion d'écrits de type "journal" ou transcriptions d'autres types de parole.

APPORTS D'UN SYSTÈME E2E



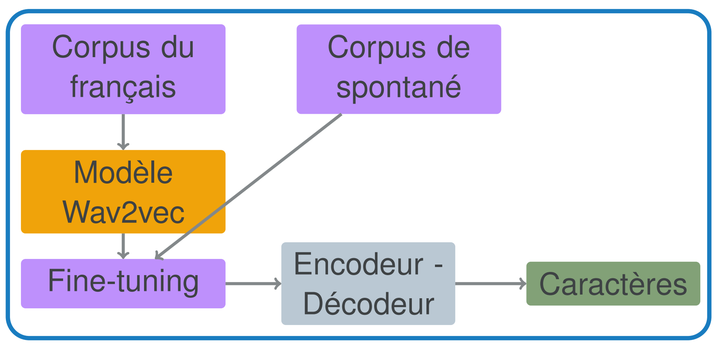
- Pas de lexique phonétisé
- Utilisation optionnelle d'un **ML** (sortie en caractères)

→ Moins d'*a priori* sur la langue injectés dans le système, il apprend ses propres représentations.

Limite: système gourmand en données

FINE-TUNING AVEC MODÈLE PRÉ-ENTRAÎNÉ

SSL: apprendre des représentations globales de la langue de façon auto-supervisée.



Résultats de (Evain, 2021) pour **RAP** sur corpus ETAPE: système **E2E**, sans **ML**, avec modèle pré-entraîné sur 3 000 heures de parole + *fine-tuning*: 26.14% ***WER*** (vs système ***HMM-DNN***: 38.50%)

IMPORTANCE DES DONNÉES

- ***WER*** étroitement lié à la quantité de données d'apprentissage
- Grande quantité de données disponible que pour une dizaine de langues dans le monde

CORPUS DE PAROLE SPONTANÉE

Corpus (date)	**Type parole**	**Type d'interaction**	**Durée (approx.)**
ESLO1 (1968-71)	Spontanée	Repas, entretiens	318 h
ESLO2 (2008-)	Spontanée	Repas, entretiens	450 h
NCCFr (2010)	Spontanée	Conversations entre amis	36 h
MPF (2010-14?)	Spontanée	Entretiens	78 h
PFC (1999)	Lue, Spontanée	Entretiens, conversations libres	>300 h
CFPP2000 (2005-?)	Spontanée	Entretiens	38 h* / 58 h 40
CLAPI (1998-)	Spontanée	conversations, réunions, visites guidées	16 h 30*
C-ORAL-ROM (2001-03)	Spontanée, Préparée	Reportages, météo, conversations privées, interactions humain-machines, enseignement, interviews...	22 h*
CRFP (1998-2002)	Spontanée, Preparée	Souvenirs, théâtre, émissions, cours	34 h*
FLEURON (2009-12)	Spontanée, Préparée	Interactions étudiants/administration	3 h 25*
OFROM (2008-12)	Spontanée, Préparée	Discussions, entretiens, communications	25 h 13*
TCOF (2005-09)	Spontanée, Préparée	Discussions, entretiens, réunions, débats	28 h40* / 61 h
TUFS (2005-11)	Spontanée	Interviews, entretiens, discussions	52 h 40*
CFPB (2013-15)	Spontanée	Entretiens	5 h* / 21 h 30
Réunions (2007-08)	Spontanée, Préparée	Réunions	18 h*
Valibel (1998-2008)	Spontanée, Préparée	Interviews, discours, entretiens, journaux, souvenirs, conversations...	43 h 25* / 331 h
Total			**1825 h 58**

Problèmes rencontrés:

- 9/16 corpus sont mixtes;
- problèmes d'accès aux données;
- différents formats d'annotations;
- enregistrements de qualité variable.

PERSPECTIVES

Évaluer, pour la **RAPS**:

- l'influence du degré de spontanéité;
- l'influence du degré d'interaction;
- l'influence du degré d'intimité entre les locuteurs.

Pour l'étude des événements propres à la **PS**, une grille d'analyse:

- niveau prosodique (pour mesurer l'impact de la segmentation des données);
- niveau morphologique (pour analyser la gestion des nouveaux mots);
- niveau grammatical (pour analyser la gestion des suites de mots "non conventionnelles").

Les corpus: ESLO2, CRFP, Valibel

Caractéristiques des enregistrements:

- bonne qualité audio;
- 3 locuteurs maximum en interaction;
- données non utilisées pour modèles pré-entraînés.

Comparaison avec les résultats d'un système appris sur de la lecture suivi d'un décodage sur de la lecture (Commonvoice).

HMM: Hidden Markov Models • **RAP(S)**: Reconnaissance Automatique de la Parole (Spontanée) • **WER**: Word Error Rate • **E2E**: End-to-end • **PS**: Parole Spontanée • **ML**: Modèle de Langue • **SSL**: Self-Supervised Learning • **DNN**: Deep Neural Network

* Corpus partiellement compris dans le Corpus d'Etude pour le Français Contemporain (CEFC)

Contact: solene.evain@univ-grenoble-alpes.fr