Comprendre les phénomènes permettant la gestion des tours de parole dans les contenus de médias audiovisuels

Rémi Uro[1,2,3], Marie Tahon[3], David Doukhan[1], Albert Rilliard[2,4]
[1] Institut National de l'Audiovisuel, [2] LISN-CNRS, [3] LIUM, [4] Universidade do Rio de Janeiro

GEM Gender Equality Monitor (ANR-19-CE38-0012)

Tours de parole et terminalité

Transition-Relevance Place (**TRP**)[1] : lieu probable pour un changement de locuteur ⇒ intéressant pour l'étude de la **gestion des tours de parole** et particulièrement des **interruptions**.

Une interruption survient lorsqu'un changement de locuteur·ice a lieu en dehors d'une **TRP**[2], i.e. à un endroit *non terminal*. Un modèle de prédiction de la terminalité a été mis en place, et une analyse DeepShap[3] et des modifications synthétiques permettent d'interroger les paramètres jouant sur cette décision.

Données

- 1991 échantillons de parole issus du coprus Allies[4] annotés en terminalité
- Issus de 23 émissions différentes
- Accord inter-annotateur: $\kappa = 0.75$
- 1151 terminaux, 1919 non-terminaux
  - Dont 839 terminaux et 1115 non-terminaux sans parole superposée

Modèle de classification de la terminalité

Modèle CNN architecture inspirée de [5] :

- Entrée : Mel-spectrogrammes (24 bandes), 2 s d'audio avant un changement de locuteur·ice
- Sortie : Classification binaire de la terminalité de l'extrait

Entraîné avec 1000 exemples, équilibrés en terminalité

Résultats:
Résultats de classification sur un jeu de test de 600 échantillons, équilibré en terminalité. Les échantillons proviennent d'émissions différentes de ceux du jeu d'entraînement.

label	**precision**	**recall**	**f1-score**
Terminal	0.69	0.71	0.70
Non Terminal	0.70	0.69	0.69

Valeurs de Shapley et interprétations

Exemples d'échantillons tirés du jeu d'entraînement :

Bleu = importance pour la décision « Terminal »
Rouge = importance pour la décision « Non Terminal »

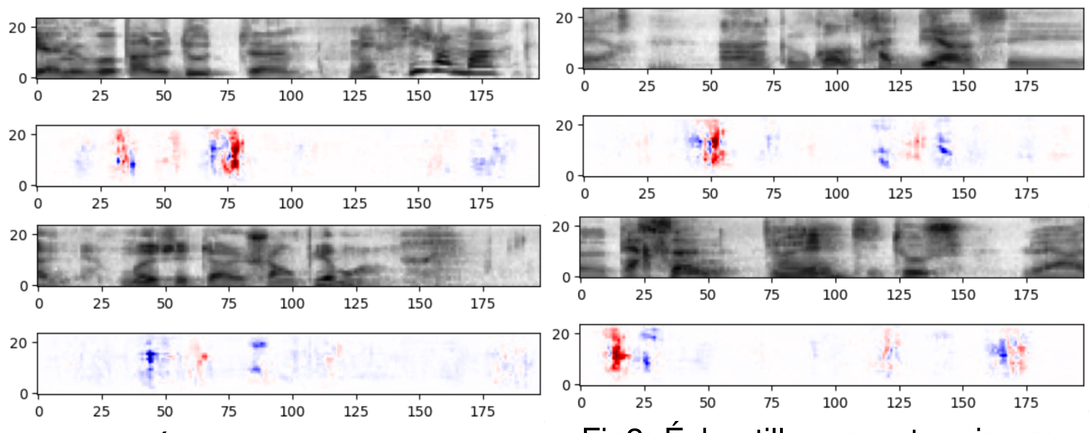

Fig1: Échantillons terminaux

Fig2: Échantillons non-terminaux

La décision semble être liée à la dynamique articulatoire sur des unités de l'ordre de la syllabe. Ceci renvoi à des informations liées à la perception du rythme (marqueur connu de la structure informationnelle et syntaxique [6]) et à la perception de la syllabe, avec notamment l'importance de la dynamique des mouvements articulatoire. [7], [8]

Impact du volume :

Pourcentage d'échantillons classifiés comme « terminaux » en fonction de l'ajout d'une atténuation du signal au début ou à la fin d'un échantillon :

	Non modifié	Fin Brut	Début Brut	Décroissant	Croissant
%	50	72.5	48.2	85.2	81.5

Faible volume à la fin → classification « Terminal » plus probable, particulièrement lorsque l'atténuation se fait progressivement.
Une amplification progressive produit aussi une augmentation de la classification « Terminal » → besoin de travaux plus approfondis.

Dépendance au contexte :

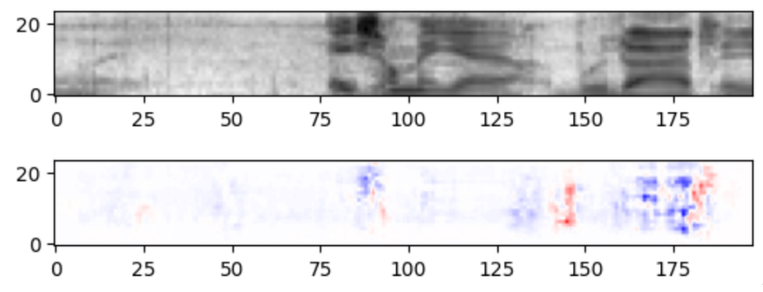

Fig3: Échantillon non modifié

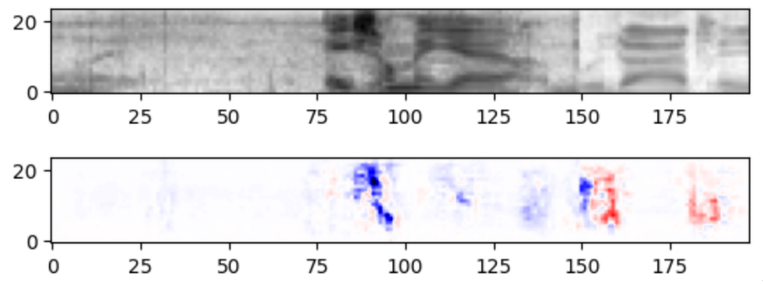

Fig4: Même échantillon avec une atténuation de la fin

Le contexte précédant et suivant joue un rôle important dans la décision : La différence entre les figures 3 et 4 montre qu'une atténuation de la fin du signal –une demi-seconde avant la fin– impacte l'importance accordée aux phénomènes bien avant l'atténuation, entre 750ms et 1000ms.
⇒ Changement de la classification de « Non-terminal » à « Terminal ».

Travaux futurs :

- Analyse de paramètres prosodiques et phonétiques aux endroits déterminants
- Synthèse et/ou transformations vocales pour tester ces hypothèses
- Inclusion explicite de paramètres prosodiques en entrée d'un modèle

[1] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, no. 4, p. 696, Dec. 1974.
[2] S. C. Levinson, *Pragmatics*, eng. Cambridge [England]; New York: Cambridge University Press, 1983.
[3] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, Curran Associates, Inc., 2017, pp. 4765–4774.
[4] A. Larcher and O. Galibert, "The allies evaluation plan for autonomous speaker diarization systems," en, p. 9, 2020.
[5] D. Doukhan, J. Carrive, F. Vallet, A. Larcher, and S. Meignier, "An open-source speaker gender detection framework for monitoring gender equality," in *ICASSP*, IEEE, 2018.
[6] P. A. Barbosa, "From syntax to acoustic duration: A dynamical model of speech rhythm production," *Speech Communication*, vol. 49, no. 9, pp. 725–742, Sep. 2007.
[7] M. Svensson Lundmark, "Rapid movements at segment boundariesa)," *The Journal of the Acoustical Society of America*, vol. 153, no. 3, pp. 1452–1467, Mar. 2023.
[8] O. Fujimura, "The c/d model and prosodic control of articulatory behavior," *Phonetica*, vol. 57, pp. 128–38, Apr. 2000.