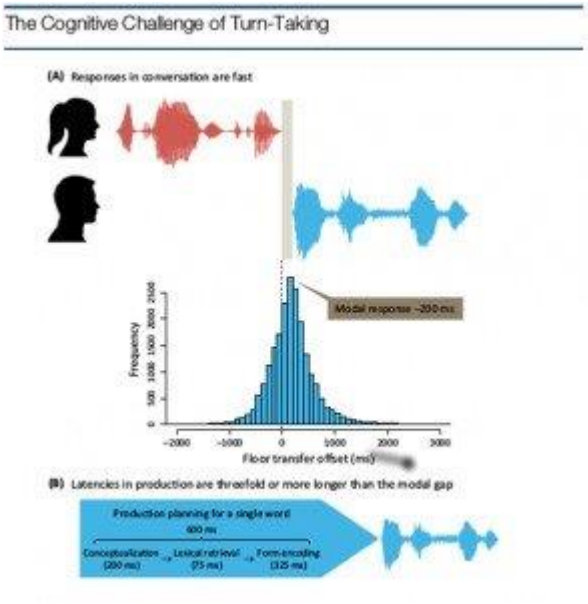
Stage MASTER 2018-2019

How long is the sentence?
prédiction incrémentale de la fin de la phrase

Contexte

Grâce au paradigme de déploiement progressif [1], le psycholinguiste François Grosjean [2] a démontré que nous sommes capables de prédire avec une relative bonne précision la longueur de la partie de phrase restant à énoncer étant donné l'écoute de son début plus ou moins tronqué. Il a montré l'importance des informations prosodiques (mélodie, rythme, etc) dans ce processus d'estimation. D'autres tâches de décision sur l'estimation de la fin du tour de parole [3], la valence émotionnelle [4] ou de l'attitude véhiculée par l'énoncé [5] [6] [7] ont montré notre capacité générale à prédire la fin d'un énoncé ou ses propriétés bien avant sa fin. Cette capacité est une des fonctions cognitives indispensables à la conduite de conversations (cf. figure extrait de [8]).



Cette prédiction a de multiples avantages: planifier une réaction, être prêts à prendre la parole - notamment par anticipation de la prise de souffle [9] [10] – ou signaler notre compréhension au cours de l'énoncé. Nous avons notamment montré [11] que la plupart des marqueurs phatiques (*backchannels*) sont produit immédiatement à la fin de l'énoncé.

Des modèles de prédiction de la fin de l'énoncé à partir des signaux acoustiques ont été proposés [12] [13]. Quelques travaux utilisant les dernières avancées en apprentissage automatique [14] [15] utilisant des centaines d'heures de dialogue ont montré qu'il est possible de détecter la fin d'un tour de parole 500ms avant sa réalisation effective avec une précision de l'ordre de 80% en utilisant à la fois les signaux bruts et les résultats d'une reconnaissance automatique incrémentale du contenu linguistique.

Objectifs

L'objectif de ce travail est d'explorer la capacité des réseaux neuronaux profonds (*DNN*) à prédire de manière robuste la fin de la phrase et comparer ces performances à des données expérimentales.

Ce travail se fera en collaboration avec Timo Baumann de l'institut de Technologies du Langage CMU, Pittsburg - USA

Sujet

Le travail consistera dès lors à :

- Effectuer une revue bibliographique sur le sujet
- Constituer une base de données d'énoncés spontanés issus de base de données parole existantes et segmenter ces énoncés en phrase intonatives (pauses > 400ms)
- Tester différentes représentations d'entrée (signaux bruts ou pré-traités, représentation de la mélodie) et de sortie (estimation de la durée, distribution sur des longueurs quantifiées, etc) et d'architectures de réseaux (CNN, RNN, LSTM, soft/hard attention, etc)
- Conduire des expériences subjectives de déploiement progressif ou de frappe sur quelques énoncés
- Evaluer la capacité des systèmes d'apprentissage à prédire les résultats expérimentaux

Possibilité de continuer en thèse

La thèse concernera le problème général de l'estimation en ligne de la fin de la phrase, du tour de parole et de l'attitude prosodique d'énoncés en langue maternelle et étrangère, afin d'estimer l'impact des connaissances hors-signal sur le traitement en ligne de la parole.

Thématiques abordées dans le stage

- Apprentissage automatique et techniques d'apprentissage profond
- Conception et réalisation de plan expérimental

Compétences requises

- Maîtrise de Python, Matlab ou R
- Forte motivation pour l'expérimentation et la méthodologie

Contacts

Gérard Bailly	GIPSA-Lab	04 76 57 47 11	gerard.bailly@gipsa-lab.fr

Indemnités de stage

Ce stage fait l'objet d'une indemnité fixée annuellement par décret ministériel (environ 540€ mensuels).

Références

[1] F. Grosjean, “Spoken word recognition processes and the gating paradigm,” *Perception & psychophysics*, vol. 28, no. 4, pp. 267–283, 1980.

[2] F. Grosjean, “How long is the sentence? Prediction and prosody in the on-line processing of language,” *Linguistica*, vol. 21, pp. 501–529, 1983.

[3] J.-P. De Ruiter, H. Mitterer, and N. J. Enfield, “Projecting the end of a speaker’s turn: A cognitive cornerstone of conversation,” *Language*, vol. 82, no. 3, pp. 515–535, 2006.

[4] B. Schuller and L. Devillers, “Incremental acoustic valence recognition: an inter-corpus perspective on features, matching, and performance in a gating paradigm,” presented at the Proc. INTERSPEECH 2010, Makuhari, Japan, 2010, pp. 801–804.

[5] S. Dalla Bella, I. Peretz, and N. Aronoff, “Time course of melody recognition: A gating paradigm study,” *Perception & Psychophysics*, vol. 65, no. 7, pp. 1019–1028, 2003.

[6] V. Aubergé, T. Grépillat, and A. Rilliard, “Can we perceive attitudes before the end of sentences? The gating paradigm for prosodic contours,” in *Proceedings of the European Conference on Speech Communication and Technology*, 1997, vol. 2, pp. 871–874.

[7] V. J. van Heuven, J. Haan, E. Janse, and E. J. van der Torre, “Perceptual identification of sentence type and the time-distribution of prosodic interrogativity marker in dutch,” in *ETRW Workshop on Prosody*, 1997, pp. 317–320.

[8] S. C. Levinson, “Turn-taking in human communication–origins and implications for language processing,” *Trends in cognitive sciences*, vol. 20, no. 1, pp. 6–14, 2016.

[9] D. H. McFarland, “Respiratory markers of conversational interaction,” *Journal of Speech and Hearing Research*, vol. 44, pp. 128–143, 2001.

[10] A. Rochet-Capellan, G. Bailly, and S. Fuchs, “Is breathing sensitive to the communication partner?,” in *Speech Prosody*, Dublin, Ireland, 2014, pp. 613–618.

[11] G. Bailly, F. Elisei, A. Juphard, and O. Moreaud, “Quantitative analysis of backchannels uttered by an interviewer during neuropsychological tests,” presented at the 17th Annual Conference of the International Speech Communication Association (Interspeech 2016), 2016.

[12] Y. Ishimoto, T. Teraoka, and M. Enomoto, “End-of-utterance prediction by prosodic features and phrase-dependency structure in spontaneous japanese speech,” presented at the Proceedings Interspeech, 2017, pp. 1681–1685.

[13] M. Atterer, T. Baumann, and D. Schlangen, “Towards incremental end-of-utterance detection in dialogue systems,” presented at the Proceedings of the 22nd International Conference on Computational Linguistics, 2008.

[14] A. Maier, J. Hough, and D. Schlangen, “Towards deep end-of-turn prediction for situated spoken dialogue systems,” *Proceedings of INTERSPEECH 2017*, 2017.

[15] M. Roddy, G. Skantze, and N. Harte, “Investigating Speech Features for Continuous Turn-Taking Prediction Using LSTMs,” *arXiv preprint arXiv:1806.11461*, 2018.