L'obtention et le placement de marques syntagmatiques en simulation de la parole

Eric Keller
Laboratoire d'analyse informatique de la parole (LAIP)
IMM - Lettres, Université de Lausanne
eric.keller@imm.unil.ch
Site: www.unil.ch/imm/docs/LAIP/LAIP_fr.html

Page Internet pour cette présentation (exemples):
www.unil.ch/imm/docs/LAIP/LAIPTTS_verif_fr.htm

Proposition

1. Je vous propose d'examiner un énoncé de parole spontanée de 15 secondes.
2. …d'en dériver les coupures majeures et mineures
3. …et de les simuler par notre synthèse de la parole.

Cet exemple fera un petit tour rapide de nos méthodes de *marquage syntagmatique*. En même temps, ceci démontrera ce qui est actuellement possible en simulation de *parole spontanée*.

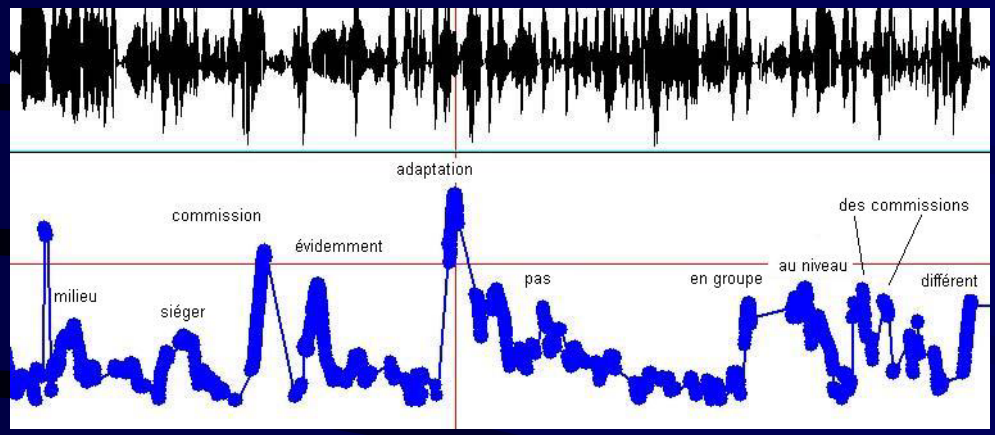
Un énoncé spontané: où sont les coupures?

- Sur le site Internet de RSR, nous avons trouvé ce reportage portant sur le remplacement des députés au milieu de législature au parlement jurassien (20.10.02). Claude Laville (CSI) offre ce commentaire (parole spontanée):

« Le député qui arrive au milieu de législature et qui doit siéger dans une commission évidemment il a un certain temps d'adaptation si au plenum ça n'joue pas véritablement son rôle parce que les décisions sont souvent prises en groupe au niveau des commissions alors c'est un tout petit peu différent | on a pu assister parfois où lorsque le nouveau député arrive en cours de législature pendant plusieurs séances il n'a pas le même poids parce qu'il ne connaît pas le dossier | ça se comprend effectivement et souvent le député n'arrive pas à s'affirmer assez rapidement contre le gouvernement et en cours de législature vous doutez bien que ça sert le gouvernement. » (| coupures semi-arbitraires imposées par les limites de notre synthèse. En jaune: le passage que nous allons analyser.)

Impression perceptive: l'intonation fournit les indices

- ***Marquage perceptif:*** **Le député qui arrive | au milieu de législature | et qui doit siéger dans une commission | évidemment | il a un certain temps d'adaptation | si au plenum ça n'joue pas véritablement son rôle parce que les décisions sont souvent prises en groupe | au niveau des commissions | alors c'est un tout petit peu différent. |**



- **Original**
- **"Hum" généré à partir de l'extraction de la fréquence fondamentale (Praat)**

- ***Marquage f-zéro:*** **Le député qui arrive au milieu de législature et qui doit siéger dans une commission évidemment il a un certain temps d'adaptation si au plenum ça n'joue pas véritablement son rôle parce que les décisions sont souvent prises en groupe au niveau des commissions alors c'est un tout petit peu différent.**
  - **Correspondances (commission, adaptation, groupe, différent): 4**
  - **Non-identifications de marques perçues (arrive, législature, évidemment, commissions): 4**
  - **Fausses identifications de syllabes non marquées (milieu, siéger, évidemment, pas, niveau, des, commissions): 7**

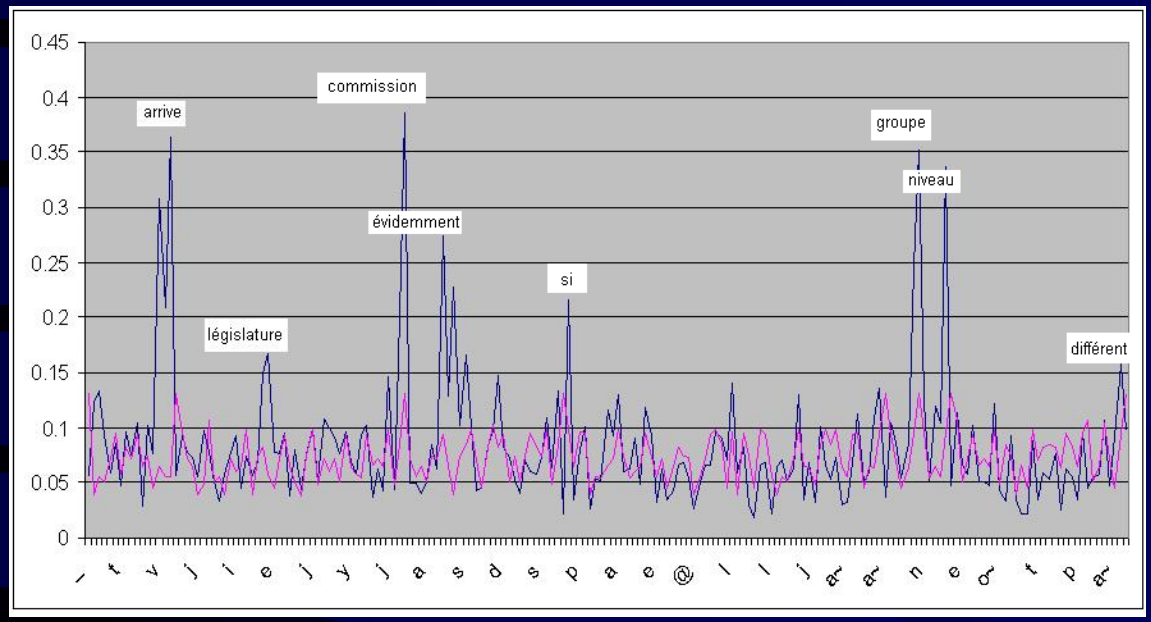
Première série d'observations

- Les non-identifications et les fausses identifications sont fréquentes.
- Les pics de f0 marquent souvent des syllabes au milieu ou en début de mots.

→ L'indice acoustique pour le mélodie (la fréquence fondamentale, ou f-zéro), pris tout seul, ne peut donc servir d'indice fiable pour l'identification des coupures perçues.

Est-ce que les durées signalent les coupures?

- ***Marquage perceptif:*** **Le député qui arrive | au milieu de législature | et qui doit siéger dans une commission | évidemment | il a un certain temps d'adaptation | si au plenum ça n'joue pas véritablement son rôle parce que les décisions sont souvent prises en groupe | au niveau des commissions | alors c'est un tout petit peu différent. |**



- En bleu: durées segmentales mesurées de M. Laville
- En rouge: durées segmentales moyennes dans un corpus de 12'000 segments (locuteur jurassien, Prof. A. Wyss)

- ***Marquage par durées segmentales:*** **Le député qui arrive au milieu de législature et qui doit siéger dans une commission évidemment il a un certain temps d'adaptation si au plenum ça n'joue pas véritablement son rôle parce que les décisions sont souvent prises en groupe au niveau des commissions alors c'est un tout petit peu différent.**

- ***Correspondances*** **(arrive, législature, commission, évidemment, groupe, différent): 6**
- ***Non-identifications*** **(adaptation, commissions): 2**
- ***Fausses identifications*** **(si, niveau): 2**

Deuxième série d'observations

- Les prolongations excessives affectent typiquement les syllabes finales, précédant les marques perçues.
- Les non-identifications et fausses identifications sont moins fréquentes.
- Les correspondances sont plutôt fréquentes.
- B. Zellner (1998, 2002) a montré que les durées syllabiques en excès d'un écart-type marquent de manière fiable les coupures majeures dans un texte "neutre".

→ Un marquage basé sur la structure temporelle apparaît donc intéressant.

- Les fausses identifications:
  - "si" pourrait être récupéré, si on admettait les coupures *précédant* une prolongation excessive.
  - "niveau" pourrait être interprêté en tant qu'une *hésitation mineure* (tout en admettant que ceci pose la question de la définition d'une hésitation par rapport à une prolongation excessive).
- Une combinaison complexe entre indices temporels et mélodiques serait également envisageable.

Simulation à partir d'une identification temporelle des coupures

- ***Marquage par durées segmentales:*** **Le député qui arrive au milieu de législature et qui doit siéger dans une commission évidemment il a un certain temps d'adaptation si au plenum ça n'joue pas véritablement son rôle parce que les décisions sont souvent prises en groupe au niveau des commissions alors c'est un tout petit peu différent.**

- **Synthèse de la parole pour le français, développée à l'Université de Lausanne,** SpeechMill-LAIPTTS-F.

  - **Avec la sortie Mbrola:** 

Et si on partait du texte (mode TTS)?

- Jusqu'à ce point, nous avons effectué des copie-synthèses, c.-à-d., nous avons simplement réinséré dans la synthèse les indices observés.
- Serait-il possible de prédire les césures directement à partir du texte?
- Ceci est le mode d'utilisation typique en synthèse de la parole.

Que savons-nous sur les césures à partir du texte?

- Nous savons que les textes lus neutres montrent une assez grande régularité (Zellner, 1998):
  - *Aux virgules et aux points finaux*, des coupures majeures.
  - *Si les syntagmes dépassent les ±14 syllabes*, des coupures supplémentaires ("groupes de performance" [Grosjean, Zellner])
  - *Normalement, préservation du groupe G(G….)L(L…)*, où G = mot grammatical et L = mot lexical

    → Les coupures uniquement à l'interface L|G

- Ces régularités nous permettent de générer des phrases de texte lu avec un minimum de connaissances syntaxiques.
- Est-ce que la parole spontanée observe les mêmes conventions? Regardons notre énoncé.

Où se situent les coupures marquées par les prolongations excessives?

- Le député qui arrive
- au milieu de législature
- et qui doit siéger dans une commission
- évidemment
- il a un certain temps d'adaptation
- si au plenum ça n'joue pas véritablement son rôle ^ parce que les décisions sont souvent prises en groupe
- au niveau
- des commissions alors c'est un tout petit peu différent.

- GLGL 8 syllabes
- GLGL 8 syllabes
- GGGLGGL 10 syllabes

- L 4 syllabes
- GGGLLL 10 syllabes

- GGLGGLGLGLGGGLGLLGL
  26 syllabes (14 + 12)
- GL 3 syllabes
- GLGLGGGGLGL
  14 syllabes

Troisième série d'observations

- (Nous supposons qu'un ralentissement important marque une coupure empirique.)
- Une coupure empirique entre deux mots lexicaux est interprétée comme coupure finale (transition L-L, équivalente à un point final).
- Une coupure empirique entre un mot lexical et un mot grammatical est interprétée comme coupure majeure (transition L-G, équivalente à une virgule).
- Nous n'observons aucune coupure empirique entre un mot grammatical et un mot lexical (transition G-L).
- Les transitions L-G sont donc particulièrement "éligibles" à une césure.
- Nous observons un seul groupe dépassant les 14 syllabes. Cependant, une subdivision en deux groupes plus brefs (14 et 12 syllabes) ne serait pas aberrante dans ce cas. En effet, le dépassement des 14 syllabes pourrait être à la base de l'impression perceptive d'une longueur excessive de ce syntagme.

Spéculation et simulation

- Quel serait le résultat si on supposait un modèle simpliste qui impose…
  - Une coupure toutes les ±14 syllabes?
  - A la transition L-G la plus proche?

Le député qui arrive au milieu de législature | et qui doit siéger dans une commission évidemment | il a un certain temps d'adaptation si au plenum | ça n'joue pas véritablement son rôle parce que les décisions | sont souvent prises en groupe au niveau des commissions | alors c'est un tout petit peu différent. | 

- Pas mal, sauf pour l'absence de césure après "groupe" qui déforme le sens.

Pour les sceptiques: simulation du reste du passage selon les mêmes principes

- On a pu assister parfois où lorsque | le nouveau député arrive en cours de législature | pendant plusieurs séances il n'a pas le même poids | parce qu'il ne connaît pas le dossier.
- Ça se comprend effectivement et souvent | le député n'arrive pas à s'affirmer assez rapidement | contre le gouvernement et en cours de législature | vous doutez bien que ça sert le gouvernement.

Conclusions

- Nous ne prétendons pas avoir trouvé l'oeuf de Colomb! La sémantique, la pragmatique et la syntaxe imposeront des césures plus appropriées à certains endroits.
- Mais nous avons montré la plausibilité de trois concepts
  - La dominance temporelle pour le marquage empirique des césures
  - La sensibilité de la transition L-G
  - La limite approximative des ±14 syllabes pour les groupes de performance
- Nous estimons que ceci trace les lignes principales d'une structure phonologique sous-tendant l'imposition de césures à un énoncé spontané.

→ Analogie

Analogie

- La structure des "sensibilités phonologiques à la césure" est comme la forme d'un verre dans lequel on verse le "sens". Le verre impose sa forme au sens.
- En d'autres mots, le locuteur tentera d'exploiter au mieux les sensibilités de la structure phonologique de la langue, afin de transmettre son sens.
- Et pour cela, il utilise sensiblement les mêmes stratégies en parole spontanée qu'en parole lue.

Merci de votre attention

Références

- Page Internet pour cette présentation (exemples):

  www.unil.ch/imm/docs/LAIP/LAIPTTS_verif_fr.htm

- Références bibliographiques:

  - Grosjean, F., & Dommergues, J.Y. (1983). Les structures de performance en psycholinguistique. L'Année psychologique, 83. 513-536.
  - Keller, E., & Zellner, B. (1998). Motivations for the prosodic predictive chain. Proceedings of ESCA Symposium on Speech Synthesis. Paper 76, pp. 137-141. Jenolan Caves, Australia. Disponible: www.unil.ch/imm/docs/LAIP/Kellerdoc.html.
  - Monnin, P, & Grosjean, F. (1993). Les structures de performance en français: caractérisation et prédiction. L'Année Psychologique, 93. 9-30.
  - Siebenhaar-Rölli, B., Zellner Keller, B., & Keller, E. (2001). Phonetic and Timing Considerations in a Swiss High German TTS System. In E. Keller, G. Bailly, A. Monaghan, J. Terken & M. Huckvale (eds.). Improvements in Speech Synthesis (pp. 165-175). Wiley & Sons.
  - Zellner Keller, B. (2002). La simulation du rythme de parole. Travaux de l'Institut de Phonétique de Strasbourg. TIPS 31 (pp. 139-165). ISDN 0750-1315.
  - Zellner Keller, B. (2002). Revisiting the Status of Speech Rhythm. in Bernard Bel & Isabelle Marlien (eds.), 2002. Proceedings of the Speech Prosody 2002 conference, 11-13 April 2002.(pp. 727-730). Aix-en-Provence: Laboratoire Parole et Langage. ISBN 2-9518233-0-4.
  - Zellner, B. (1998). Caractérisation et prédiction du débit de parole en français. Une étude de cas. Thèse de Doctorat. Faculté des Lettres, Université de Lausanne. Disponible: www.unil.ch/imm/docs/LAIP/ZellnerKellerdoc.htm.