Praat: un outil pour l'annonation et l'analyse de la parole.

Praat et un outil pour l'étude de la parole. Plus précisément, dans ce TD, nous allons aborder l'annotation et l'analyse acoustique de la parole conversationnelle.

Préambule

Le corpus NCCFr

Le corpus utilisé dans cette étude est le corpus NCCFr (Nijmegen Corpus of Casual Speech) [1]. Il a été conçu pour mener des études sur la variation phonétique dans un registre familier intime. Il a été enregistré fin 2007 au Laboratoire de Phonétique et de Phonologie de Paris. Il comprend 35 h de parole produite par 46 locuteurs (24 femmes et 22 hommes). Il est composé de dialogues entre étudiants de la région parisienne qui se connaissent bien (environ 90 min de conversation pour chaque paire de locuteurs). La sélection des locuteurs en fonction de leur statut (étudiant) et de leur origine géographique (région parisienne) permet de relativement bien contrôler les variables socioprofessionnelles et régionales. Une partie des enregistrements des dialogues s'est faite en présence d'une troisième personne, également amie des autres locuteurs. Cette dernière avait pour rôle d'alimenter si nécessaire les échanges oraux entre les deux autres. Ses contributions restent donc limitées et n'ont pas été prises en compte dans cette étude.

Le LIMSI a coordonné la transcription orthographique manuelle du corpus, faite à l'aide du logiciel Transcriber [2]. Les transcripteurs avaient pour consigne de transcrire tous les événements audibles, y compris les disfluences, les autoréparations, les reprises, les amorces, etc. De même, ils pouvaient utiliser pour les transcriptions orthographiques des signes de ponctuation forte si cela leur semblait nécessaire. Notons cependant que les signes de ponctuation ont été retirés pour le traitement décrit dans cette contribution. Un exemple de transcription est présenté dans la figure 1. Le corpus NCCFr contient plus de 270 000 mots (occurrences ou tokens) et 8 700 mots distincts (types).

Dans ce TD, nous n'allons étudier qu'un extrait. Vous trouverez donc un fichier audio stéréo. Chaque canal correspond au micro cravate de l'un des deux participants. Sur le fichier stéréo, vous entendez donc les voix des deux participants.

Outil Praat

Praat[3] est un logiciel libre pour l'analyse, la manipulation et l'annotation de sons. Ces fonctionnalités en font un outil complet en particulier pour l'étude de parole. Il permet également de tracer des graphiques, construire des grammaires basées sur la théorie de l'optimalité, de faire de la synthèse articulatoire, de simuler des réseaux de neurones et de faire des analyses statistiques. Paul Boersma et David Weenink de l'Institute of Phonetic Sciences de l'Université d'Amsterdam ont créé Praat en 1996 et continuent activement de développer cet outil de manière très interactive avec la communauté des utilisateurs. Il a été conçu à la fois pour les non-experts en traitement de la parole grâce ses interfaces graphiques et menus simplifiés et pour les utilisateurs avancés grâce aux nombreuses possibilités de manipulations, d'analyses et de scripting.

Annotation avec Praat

Ouvrir un fichier audio `Read > Read from file`. Une fois le fichier chargé, un objet Sound apparait dans la liste, ainsi que des boutons à droite de la liste:

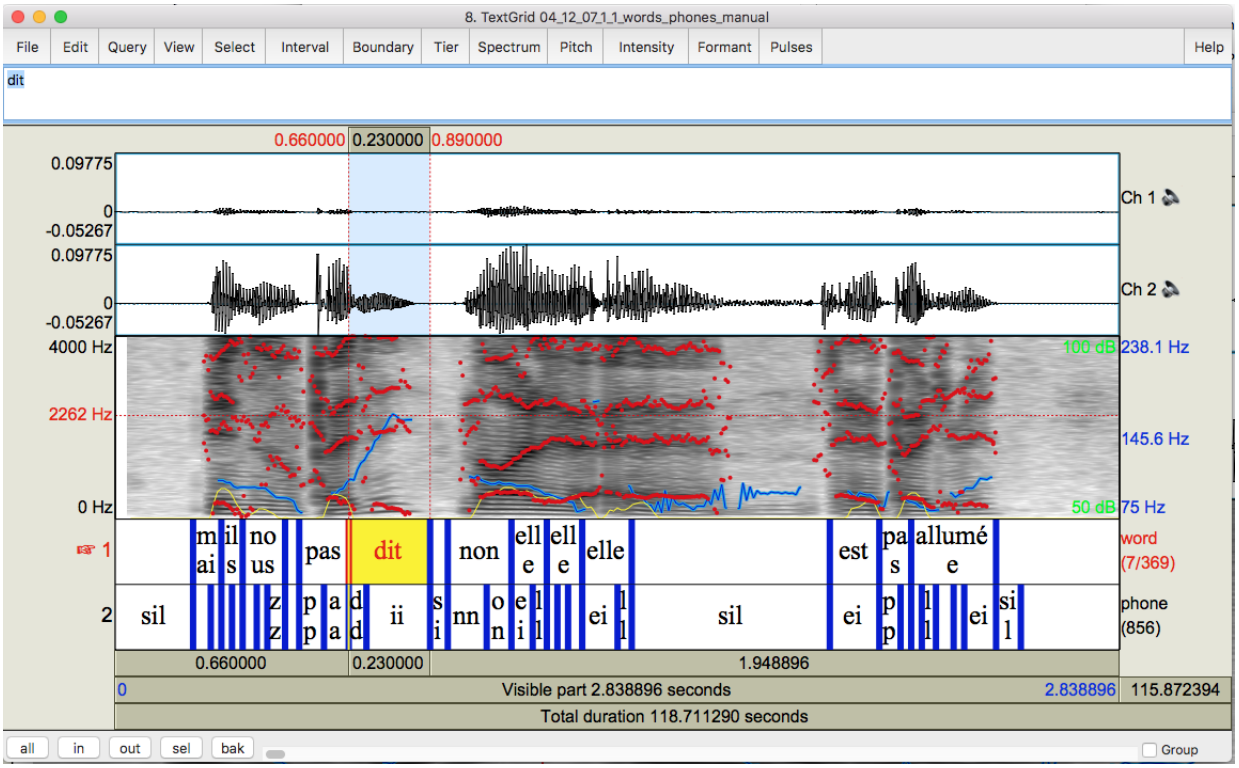

Figure 1: Visualisation des trois panneaux: signal temporel, analyse acoustique et annotation

- `Edit` ouvre une fenêtre pour visualiser le signal
- `Play` joue le son. Pour interrompre -> Echap.
- `Annotate -> To TextGrid` créé un objet d'annotation.

Vous avez normalement deux panneaux, le première pour la représentation de la forme d'onde temporelle (en haut), le second pour les analyses acoustiques (spectrogram, pitch, intensity...)

Annoter un fichier audio `Annotate -> To TextGrid` ouvre un formulaire demandant le nom du champ à remplir (Tier). Il faut définir si le champ correspond à des sons ponctuels, et l'annotation une série de points (PointTier) ou bien à des intervalles pour annoter des sons qui ont une durée (IntervalTier). Sélectionner ensuite dans la liste d'objet le TextGrid et le son et appuyer sur le bouton `Edit`.

La fenêtre d'édition permet de visualiser un troisième panneau (sous les analyses) et trois menus supplémentaires (Interval, Boundary, et Tier) pour l'annotation, voir figure 1. Pour l'annotation, vous pouvez:

- Ajouter une frontière (Boundary): cliquer sur le signal pour positionner le curseur à l'endroit voulu. Puis appuyer sur Entrée ou cliquer sur le petit cercle pour ajouter une frontière sur la tier active.
- Ajouter du texte dans un intervalle: sélectionner un intervalle en cliquant dessus. Il s'affichera dans l'intervalle mais aussi au dessus du signal dans une fenêtre blanche qui servira de fenêtre d'édition pour une modification ultérieure. Les fonctions Cut/Copy/Paste peuvent être utilisées.
- Supprimer une frontière: la sélectionner (apparait en rouge) et `Menu -> Remove` ou Alt+Backspace.
- Déplacer une frontière: la sélectionner et la bouger avec la souris.
- Sauvegarder l'annotation: `File -> Write TextGrid to text file`

Pour cette séance, voici les tâches à faire.

1. Ouvrez le fichier audio et créez deux Tier pour l'annotation des mots, et des phonèmes pour le locuteur 1 uniquement.
2. Quel type de Tier faudra-t-il choisir ? Interval ou Point ?
3. Annoter l'ensemble du fichier audio en mots.
4. Sélectionner une phrase et annoter en phonèmes en utilisant le dictionnaire de phonèmes CMU (adapté normalement à l'anglais)
5. Estimez le temps passé sur chacune des tâches d'annotation. Combien de temps passeriez-vous pour le corpus entier ?
6. Notez bien les choix avez à faire et les difficultés rencontrées, nous en discuterons.

Analyse de la transcription

Nous avions au départ, la transcription manuelle en mots fournies par le LIMSI qui suivait le protocole détaillé fourni dans les documents. Nous avions besoin d'une segmentation précise en phonème. La démarche a été la suivante: nous avons fait un alignement forcé entre les mots annotés manuellement et la sortie d'un système de reconnaissance automatique. Cette alignement des phrases, nous a permis d'obtenir un alignement en phonème en récupérant les informations juste avant le modèle de langage. Nous avons donc les fichiers suivants:

- TextGrid 04-12-07_1_extrait : annotation manuelle du LIMSI
- TextGrid 04_12_07_1_1_words_phones_extrait: mots du locuteur 1 annotés manuellement et enrichis automatiquement afin de remplir tous les intervalles et phonèmes issues de l'alignement forcé, annotation manuelle en émotion (valence, activation et dominance) et en hésitation.

Nous allons nous étudier plus en détail ce qu'il se passe au temps $t_1 = 20,3$ sec. sur la phrase *non mais toute façon c'est enregistré* et au temps $t_2 = 39,73$ sec. sur la phrase *&ben j' ai j' ai j' ai enfin j' ai du mal g(roupe) avec le travail de groupe donc euh enfin.* Pour les deux passages, répondre aux questions suivantes.

1. Écouter le signal et vérifier si la transcription orthographique vous semble correcte.
2. Que se passe-t-il à la fin de la phrase ? Est-ce cohérent avec le guide d'annotation ?
3. Qu'est-ce qui est prévu dans le cas d'élisions , de troncations ou de répétitions (phénomènes très courants à l'oral) ? Quelles sont les conséquences de ces choix sur la segmentation phonétique ?

Analyse acoustique

Nous allons maintenant analyser d'un point de vue acoustique les extraits déjà abordés.P our cela afficher le spectre du signal,le pitch et les formants si ce n'est pas déjà fait.

Pour chacun des deux instants $t_1 = 20,3$ et $t_2 = 39,73$:

1. Calculer le débit en mot et en phonème à partir de la segmentation automatique. Qu'en pensez-vous ? Étant donné les erreurs possibles lors de la segmentation automatique, quelle marge d'erreur avez-vous sur des deux résultats ?
2. Calculer la $F0$ moyenne et écart-type. Pour cela vous pouvez récupérer les valeurs de Pitch: `Pitch -> Pitch Listing`. Notez les valeurs *undefined*. Qu'en pensez-vous ?
3. Calculer la $F0$ moyenne et écart-type pour l'autre locuteur. Est-ce différent significativement ?
4. Observez les annotations émotionnelles et hésitation données pour t_2. Qu'en pensez-vous ?

Nous allons maintenant faire une analyse spectrale sur l'extrait à $t_1 = 20,3$. Choisissez dans `spectrum settings` une fenêtre de 30 ms.

1. Pourquoi ce choix de 30 ms est-il pertinent ?
2. Observez le spectrogramme du premier phonème **oo**. Qu'observez-vous ?
3. Tracer le spectre moyenné sur ce phonème: sélectionner l'intervalle correspondant, `Spectrum -> View spectral slice` puis dans la liste des objets visualiser le spectre. Donner la valeur des 5 premiers pics, à quoi correspondent-ils ?
4. Imaginez que vous observez uniquement l'enveloppe spectrale et non, l'ensemble des variations. Relevez les valeurs des trois premières bosses, à quoi correspondent-elles ?
5. Observez sur la visualisation temporelle, l'évolution des formants entre 30,13 s et 30,30 s. Le locuteur prononce un **ei** suivant d'un **in**. Où est la frontière entre les deux phonèmes ?
6. Observez le spectre d'un phonème fricatif **ss** ou un **ff**. Qu'observez-vous ?

Enfin, si on a le temps, nous pouvons observer ce qu'il se passe à $t_3 = 44.3$ sec. Le locuteur 1 coupe la parole au locuteur 2. Comment analyseriez-vous cela d'un point de vue linguistique et acoustique ?

References

[1] Torreira, F., Adda-Decker, M. & Ernestus, M. (2010). The Nijmegen corpus of casual French. Speech Communication, 52, 201-212.

[2] Barras, C., Geoffrois, E., Wu, Z. & al. (2001). Transcriber: development and use of a tool for assisting speech corpora production. Speech Communication, 33 (1-2), 5-22.

[3] Paul Boersma & David Weenink (2018): Praat: doing phonetics by computer [Computer program]. Version 6.0.49, retrieved 14 March 2019 from https://www.praat.org/