Université François Rabelais de Tours

UFR Lettres et Langue

Master SLD – M2 L & R

Traitement Automatique des Langues

TRAVAUX PRATIQUES

Enseignant **Jean-Yves ANTOINE**

(Jean-Yves.Antoine AT univ-tours.fr)

Synthèse de parole et analyse de parole: PRAAT

1. Présentation

PRAAT est un outil d'analyse et de traitement du signal de parole qui a été développé par Paul Boersma et David Weenink, du département de phonétique de l'Université d'Amsterdam. Cet utilitaire est diffusé librement par ses auteurs et peut être récupéré à l'adresse WWW suivante : www.praat.org

PRAAT est un outil complet et très puissant comparé à ceux étudiés en Master 1ère année (*SFS* pour l'analyse du signal de parole et *Transcriber* pour la transcription orthographique de corpus oraux). Tout d'abord, il réunit la plupart des traitements qui peuvent être envisagés dans le domaine de la parole. Ensuite, il dispose de capacités de pilotage fines à l'aide d'un langage de script qui est très utile lorsqu'on veut mettre en place tout une chaîne de traitements à l'aide de PRAAT. Cette richesse du logiciel a bien entendu sa contrepartie : son utilisation est plus complexe que celle d'outils dédiés à une tâche précise. A titre d'exemple, *SFS* sera ainsi aussi utile que PRAAT pour calculer un simple spectrogramme. De même, il est recommandé d'utiliser plutôt *Transcriber*, très ergonomique, pour la transcription orthographique d'un corpus de parole. A l'opposé, PRAAT est à préférer pour une transcription plus fine de type phonétique ou micro-prosodique.

Au cours de cette année, nous allons étudier plusieurs fonctionnalités intégrées dans PRAAT:

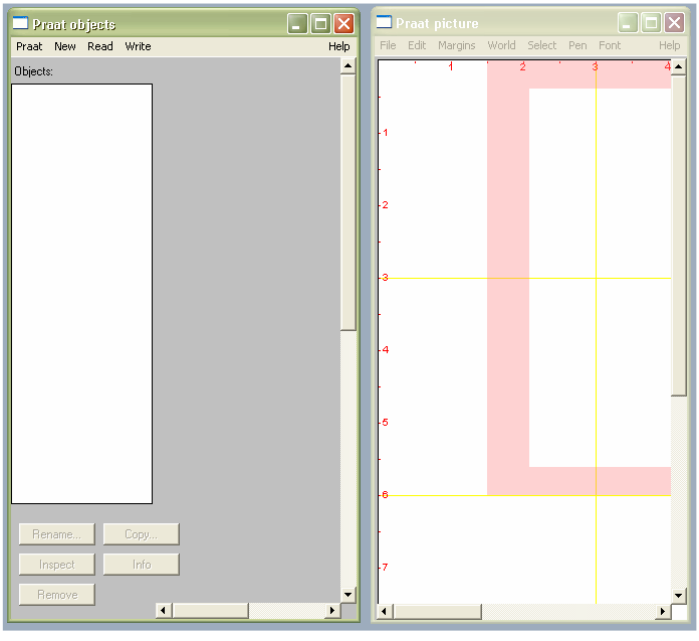
- **Traitement du signal : filtrage fréquentiel** – PRAAT permet d'effectuer des filtrages fréquentiels simples (filtre passe-bas ou passe haut) du signal de parole. Nous allons étudier l'influence d'un point de vue perceptif de ces filtres, fréquemment utilisés en télécommunications et en technologies de la parole, sur le signal résultant.
- **Synthèse de la parole** – PRAAT met en œuvre deux méthodes de synthèse à base de connaissances (modèle source-filtre et synthèse articulatoire). Ces techniques ne permettent pas de réaliser une synthèse de parole réellement naturelle. Elles restent par contre utiles pour comprendre le fonctionnement de la production de parole. C'est à cette fin que nous les étudierons.
- **Langage de script** – Nous étudierons rapidement les capacités de pilotage de PRAAT par langage de script.

Avant d'étudier ces différentes fonctionnalités avancées, il est nécessaire de maitriser un tant soit peu les fonctions de base de PRAAT. C'est ce que nous allons tout d'abord faire, ce qui nous permettra de réviser au passage ce que nous avions vu avec *SFS* au cours du Master 1ère année.

2. Prise en main du logiciel : quelques rappels sur l'analyse de signal de parole

2.1. Démarrer

Lancez le logiciel PRAAT. Deux fenêtres apparaissent à l'écran :



- A gauche, la **fenêtre des objets PRAAT** (*PRAAT objects*) recense tous les objets utilisés par PRAAT à un moment donné. Ces objets peuvent être bien entendu des fichiers de signal, mais également les fichiers d'annotation que vous avez effectués sur ces signaux, ou encore des scripts que vous auriez défini pour automatiser certaines tâches. Dans un premier temps, les seuls objets étudiés seront des fichiers de parole.
- A droite, la **fenêtre d'affichage PRAAT** (PRAAT picture) qui permettra d'afficher le signal de parole et tout affichage correspondant à un traitement réalisé par le logiciel. Notons que ces résultats graphiques peuvent être sauvegardés en format postscript encapsulable (EPS) ou en un méta-fichier Windows qui peut être ensuite récupéré par tout éditeur de texte.

Commençons par nous intéresser à la fenêtre des objets PRAAT. Sur le droite, le menu `Help` donne accès au manuel (très complet) du logiciel. Au centre, le menu `New` permet de créer de nouveaux objets. En particulier, il permet d'enregistrer directement un fichier sonore. Si l'on dispose déjà de fichiers de sons, ceux-ci peuvent être chargés à partir du menu `Read`. C'est précisément ce que nous allons faire.

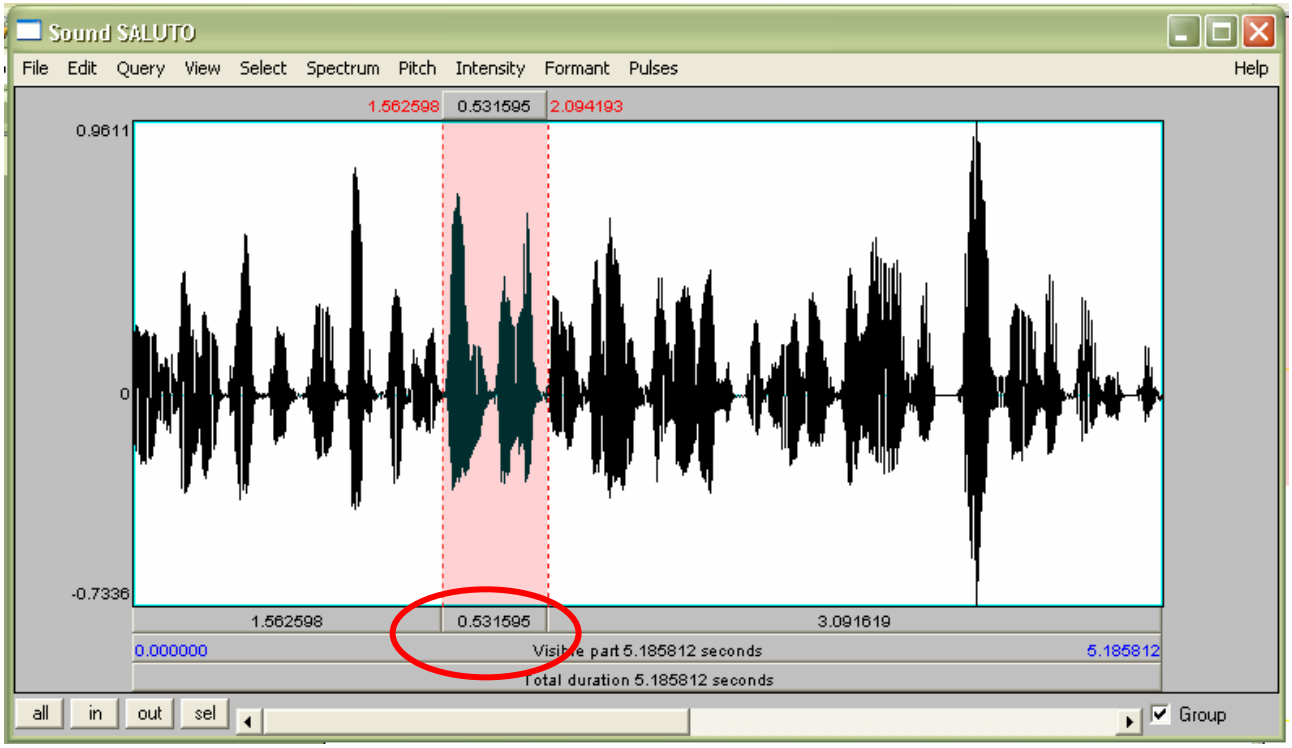
Téléchargez à partir du forum ou de ma page WWW (www.info.univ-tours.fr/~antoine) le fichier de parole `saluto.wav`. Sauvegardez ce fichier sur votre compte personnel. A l'aide du menu `Read/Read From File`, chargez ce fichier dans PRAAT. Le fichier apparait dans la fenêtre des objets, de même qu'un ensemble de boutons contextuels qui vont permettre de travailler sur ce fichier :

- En cliquant le bouton `Play`, écoutez le signal de parole correspondant,
- En cliquant le bouton `Edit`, affichez le signal dans une nouvelle fenêtre: on observe l'évolution de l'intensité du signal au cours du temps.

2.2. Sélection d'une zone temporelle de signal

Ce signal est assez long. Pour travailler confortablement, il est possible de ne visualiser qu'une partie du signal et de n'écouter que la portion de signal correspondante. Pour cela :

- Sélectionnez une zone temporelle en cliquant, dans la fenêtre d'affichage du signal, avec le bouton gauche de la souris pour marquer le début de la zone, et en maintenant ce bouton appuyé jusqu'à la fin de la zone à délimiter. Celle-ci s'affiche en rose clair.
- Vous remarquez qu'en dessous de l'affichage de l'intensité, la barre supérieure de défilement temporelle est désormais segmentée en plusieurs zones (avec durée en secondes). Il vous suffit de cliquer dans la zone d'intérêt considérée pour écouter le signal uniquement sur celle-ci.



- Pour limiter l'affichage à la seule zone considérée, il suffit de sélectionner le bouton `sel` en bas de la fenêtre. Les boutons `in` et `out` permettent de zoomer et dézoomer d'un facteur donné. Comme son nom l'indique, le bouton `all` dézoome pour visualiser l'ensemble du signal.

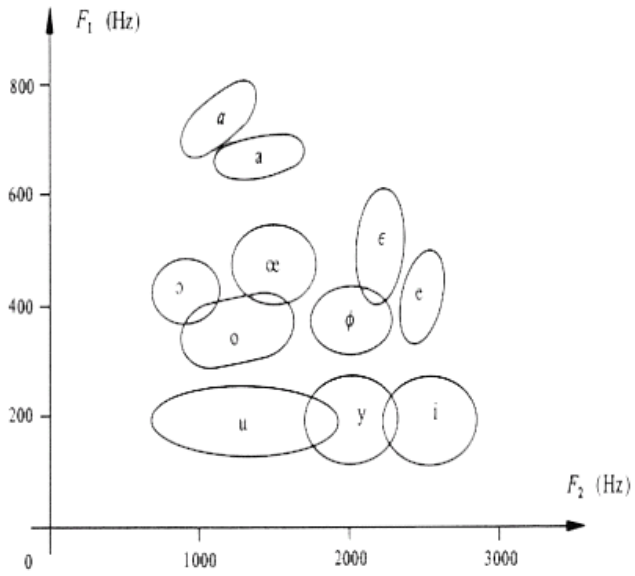
2.3. Traitements de base : spectrogramme, détection de formants et de pitch

Il est possible de réaliser plusieurs traitements de base pour visualiser les propriétés acoustiques du signal de parole :

- **Spectrogramme du signal**: option `Show spectogram` du menu `Spectrum` de la fenêtre d'affichage,
- **Formants**: détection des 5 premiers formats (par défaut) : option `Show Formants` du menu `Formants`,
- **Pitch**: détection du voisement et calcul de la fréquence fondamentale correspondante dans l'option `Show Pitch` du menu Pitch.
- **Pulse**: détection des pulsations successives des cordes vocales lorsque le signal est voisé. Cette information est affichée sur l'intensité, en sélectionnant l'option `Show Pulses` du menu `Pulses`.

Remarque : toutes ces visualisations peuvent être enregistrées dans un fichier image. Pour cela, choisissez l'option `Draw…` ou `Paint…` dans le menu correspondant.

Zoomez le signal pour ne visualiser que la partie correspondant à la prononciation de « *un saluto a tutti* ». Affichez tout d'abord le spectrogramme et les **formants** sur cet extrait de parole. La figure ci-contre donne la valeur des deux premiers formants pour les voyelles du français.



En observant le spectrogramme du signal, pouvez-vous dire si l'on retrouve en italien les mêmes valeurs de formants pour les voyelles /u/, /a/ et /i/ ?

Peut-on savoir, au vu du spectrogramme, si la première syllabe de *saluto* correspond à la réalisation d'un /a/ ou plutôt d'un /ɑ/ ? Quel phénomène explique ce résultat ?

Indication: en cliquant sur un point précis du spectrogramme, vous pouvez visualiser la valeur précise de la fréquence ainsi pointée. Vous pouvez également avoir la valeur des formants à un instant précis en choisissant

Le spectrogramme qui est calculé par PRAAT correspond à un spectrogramme à **large bande**. Comme sur SFS, il est possible d'afficher un spectrogramme à **bande étroite**. Pour cela, ouvrez l'option de paramétrage `Spectrogram settings` du menu `Spectrum`. Vous observez que la fenêtre de calcul du spectrogramme a une longueur de 0.005 secondes. Cela signifie que pour calculer le spectre à un instant *t* donné, on ne considère le signal que sur la partie [*t* – 0,0025 sec. , *t* + 0,0025 sec.] du signal. Pour afficher un spectrogramme à bande étroite, augmentez la taille de la fenêtre à 0.03 secondes.

Qu'observez-vous sur l'affichage du spectrogramme ? Ce résultat correspond-il bien à ce qu'on attendait ? A quoi correspondent les bandes noires que l'on observe sur le spectre ?

Positionnez-vous à un endroit de votre choix dans le signal et affichez le **spectre du signal à cet instant précis** à l'aide de la commande `View spectral slice` du menu `Spectrum`. Faites de même mais en revenant à un paramétrage correspondant à un spectre à bande large. Observe-t-on là encore une différence attendue entre bande étroite et bande large ?

Au fait, pour afficher un spectre à bande étroite, nous avons choisi une fenêtre plus longue (et réciproquement). Cette démarche vous semble-t-elle logique ? Expliquez.

Sur le même extrait de signal, affichez maintenant les **pulsations du signal** et le **pitch** de celui-ci, quand la parole est voisée. Au regard de l'intensité du signal de parole, à quoi correspondent les pulsations successives ? La présence de pulsations est-elle corrélée à la détection d'un pitch.

Zoomez l'affichage sur une voyelle du signal. Quelle est la valeur du pitch au milieu de ce phonème ? Affichez maintenant le spectre à bande étroite du signal à cet instant précis. Retrouve-t-on la même valeur de fréquence fondamentale ? Montrez que les pics suivant du spectre correspondent bien aux fréquences harmoniques du voisement.

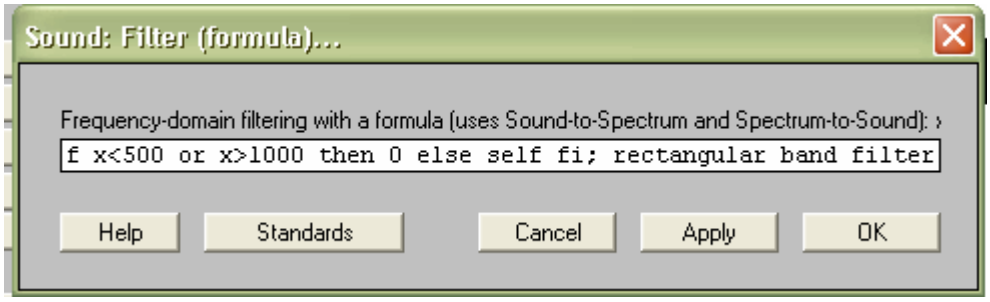
Affichez maintenant l'intégralité du signal de parole. Entre quels extrêmes la fréquence fondamentale du signal varie-t-elle au cours de la prononciation de la phrase ? Pour obtenir une estimation précise de cette variation, vous pouvez utiliser les options `Get minimum/maximum Pitch` du menu `Pitch`.

3. Filtrage d'un signal de parole : analyse perceptive d'un signal de parole

PRAAT permet de faire passer des filtres fréquentiels ou temporels sur un signal de parole. En particulier, il est possible de réaliser les traitements suivants :

- filtrage passe-bas, qui coupe toutes les composantes fréquentielles du signal qui sont supérieures à une fréquence seuil donnée. Ce filtrage laisse donc passer les basses fréquences, d'où le nom du filtre,
- filtrage passe-haut, qui ne laisse passer que les fréquences supérieures à une valeur seuil donnée.
- filtrage passe-bande, qui ne conserve que les fréquences comprises entre deux valeurs données,
- d'autres filtres moins exclusifs qui permettent de simplement renforcer certaines bandes de fréquence.

Nous allons utiliser ces capacités de filtrage pour étudier les fréquences utiles, d'un point de vue perceptif, dans un signal de parole. Les filtrages implémentés dans PRAAT sont accessibles dans la fenêtre objet à partir du bouton `Filter`. Au cours de ce TP, nous n'étudierons que l'outil `Filter(formula)` qui permet de définir manuellement tout type de filtre de base[1]. Sélectionnez cet outil. Une fenêtre s'affiche, qui donne la formule de filtrage utilisée :



Dans le cas de cette figure, la formule réalise un filtre passe-bande : les composantes fréquentielles qui ne sont pas comprises entre 500 et 1000 Hz sont annulées (`if x<500 or x >1000 then 0`) tandis que les autres composantes sont conservées en l'état (`else self fi`). La fenêtre de filtrage utilisée (`rectangular band filter`) est la plus sélective qui soit : filtrage tout ou rien des fréquences. Avec d'autres fenêtres, on aurait une annulation progressive des fréquences en dehors de la zone de conservation. Dans le cadre de ce TP, on conservera toujours ce type de fenêtre. Pour appliquer un filtre différent au signal, il suffit de modifier la formule et de sélectionner le bouton `Apply` : un nouveau signal filtré est généré. Il apparait alors dans la fenêtre des objets PRAAT. Pour quitter l'outil en conservant la dernière formule réalisée, sélectionner `OK`. Vous allez maintenant réaliser différents filtrages pour répondre aux questions à venir.

Filtrage passe-bas – On considère généralement que les deux premiers formants sont suffisants pour distinguer les voyelles du français. Pour vérifier cette affirmation sur le phonème /a/, on désire réaliser un filtre passe-bas qui ne garde que les composantes fréquentielles utiles à cette caractérisation. En considérant la figure de la page précédente, déterminez la fréquence de coupure qui permet de réaliser cette opération. Définissez le filtre correspondant dans PRAAT, qui sera appliqué au signal audio_aka.wav, correspondant à la prononciation /aka/.

Visualisez le signal filtré: le résultat obtenu est-il bien celui attendu ? Ecoutez ce signal. Le son /a/ est-il toujours perceptible ? Baissez maintenant progressivement la fréquence de coupure du filtre. A partir de quelle valeur la voyelle n'est-elle plus identifiable ?

Filtrage passe-haut – Au contraire des voyelles dont l'énergie est avant tout concentrée dans les basses fréquences, les fricatives se traduisent par un son turbulent dans les fréquences moyennes et hautes. Chargez le fichier `TTS_ICP_93.au` qui correspond à un signal de parole obtenu par synthèse vocale. Filtrez toutes les fréquences inférieures à 2500 Hz. Qu'observez-vous à l'écoute ?

Ce fichier de signal numérique est assez ancien : il est échantillonné à 8000 Hz. Filtrez le signal pour ne garder que les composantes fréquentielles supérieures à 4000 Hz. Regardez le spectre et l'intensité du signal (n'hésitez pas à zoomer sur une petite zone du signal. Qu'observez-vous ? Pouvez-vous expliquer ce résultat ?

4. Synthèse de la parole : analyse perceptive et analyse de signal

Au cours de ce TP, nous allons étudier différentes méthodes de synthèse de parole à partir du texte. Afin d'avoir une idée de la difficulté de cette tâche, nous allons comparer différentes fichiers de parole :

[1] Pour plus de renseignement sur les autres types de filtres reconnus par PRAAT, on pourra consulter le tutoriel accessible à partir du sous-menu `Filtering tutorial`.

- `TTS_ICP_93.au` signal de parole obtenu par synthèse articulatoire
- `TTS_voix.mp3` synthèse de parole par sélection d'unités. Système commercial (ScanSoft)
- `TTS_FT_RD_2002.wav` synthèse de parole par sélection d'unités. Système recherche (Orange Labs)
- `audio_saluto.wav` parole naturelle.

Ecoutez ces différents fichiers. Classez-les suivant leur degré de naturalité. Nous allons essayer d'expliquer ces différences perceptives de qualité en réalisant l'analyse des signaux de parole correspondants.

Intéressons-nous tout d'abord aux voyelles dans ces fichiers. En prenant un ou deux exemples, comparez les formants des phonèmes ainsi synthétisés par les différents systèmes. Ces résultats peuvent-ils expliquer nos observations perceptives ?

Intéressez-vous maintenant à quelques fricatives dans ces fichiers. Etudiez leur spectre au milieu de leurs réalisations. Observez-vous des différences significatives entre ces différentes réalisations ?

En vous intéressant aux spectrogrammes des signaux (ou aux spectres à un moment donné), voire l'évolution de leur intensité, détectez-vous des différences dans les réalisations des différents systèmes ?

Cette étude devrait vous montrer la difficulté qu'ont les chercheurs et les ingénieurs pour arriver à une synthèse de parole naturelle. Vous allez avoir vous-même l'occasion de vous en rendre compte en pratique. Maintenant que nous maitrisons l'utilisation de PRAAT, nous allons en effet tenter de synthétiser quelques sons élémentaires…

5. Synthèse de la parole : synthèse par formants (modèle source-filtre)

Dans un premier temps, nous allons utiliser une synthèse par formants, qui a été une des premières approches envisagées par les chercheurs au tournant des années 1960-1970. Cette étude va nous permettre de mieux comprendre comment l'expulsion d'air à travail un conduit vocal nous permet de produire des sons de parole articulée. Dans ce modèle, il est nécessaire de générer tout d'abord une source, qui reproduit l'expulsion de l'air éventuellement modulée par une excitation des cordes vocales. Puis un filtre qui va rendre compte des fréquences de résonnances (formats) dépendant de la position des différents articulateurs le long de l'appareil phonatoire. Ces deux éléments vont correspondre à des objets différents. Dans PRAAT, on parle de couches (*tier*) différentes. Pour commencer, nous allons chercher à générer la simple tenue d'une voyelle /a/ sur une certaine durée.

Génération de la source (*Pitch Tier*) – Pour générer la couche correspondant à la source dans la fenêtre d'objets PRAAT, aller dans le menu `New > Tiers > Create Pitch Tier`. La fenêtre qui apparait vous demande de donner un nom à votre source, puis de préciser les limites temporelles d'activation de cette source. Choisissez par exemple de générer un signal de 2 s de long.

Pour le moment, vous n'avez pas encore créé de source sonore: vous n'avez fait que définir sa longueur ! Il va falloir modifier cette définition minimale pour préciser en particulier la fréquence fondamentale des cordes vocales. Pour cela, sélectionnez le bouton `Modify` qui est proposé si vous sélectionnez votre couche. Dans le menu qui apparait, choisissez l'option `Add Point`. Celle-ci permet de définir la courbe d'évolution du pitch point par point ! Pour le moment, nous nous contenterons de générer un signal de pitch constant (150 Hz par exemple). Il vous suffit donc de définir un point au début (t = 0.0 s) et à la fin (t = 1.0 s) du signal. PRAAT générera la courbe d'évolution du pitch (en fait, une droite…) en reliant ces deux points.

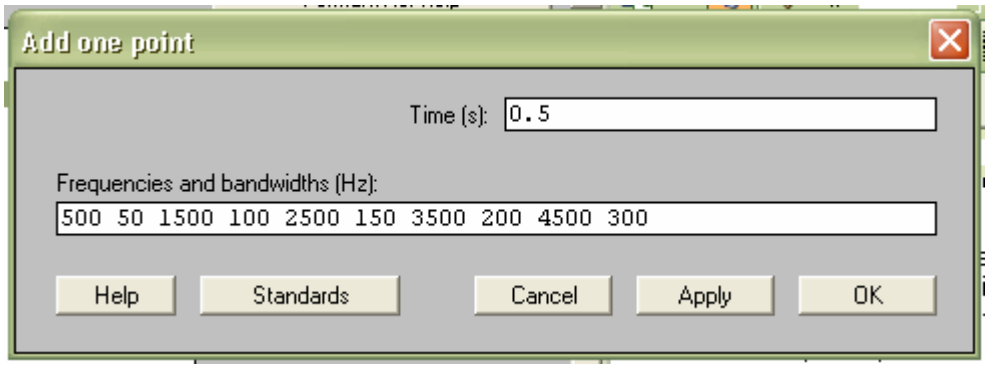
Maintenant, votre source est définie. Il ne vous reste plus qu'à générer le son correspondant à cette dernière. Plaçons-nous tout d'abord dans une situation artificielle : vos cordes vocales vibrent suivant une sinusoïde parfaite, et votre conduit vocal est tranché juste au-dessus d'elles (gloups !). Pour générer le signal, choisissez l'option `Synthetize > To Sound (sine).`

Visualisez et écoutez le signal correspondant. Qu'observez-vous ? Le spectre du signal correspond-il à nos attentes ?

Rapprochons nous maintenant de la réalité : les cordes vocales vibrent suivant un mouvement périodique mais non sinusoïdal que l'on sait approximer, et on peut également simuler l'effet de radiation de l'air à la sortie de vos lèvres. Seul problème, votre conduit vocal ressemble encore à un très beau tuyau de section constante et parfaitement droit : ce sera au filtre de corriger cette dernière approximation. Pour le moment, synthétisez donc le signal sonore correspondant à la source (cordes vocales + lèvres). Pour cela, choisissez l'option `Synthetize > To Sound (phonation).` Ne modifiez pas les paramètres prédéfinis, qui correspondent à une description relativement standard de la réalité.

Visualisez et écoutez le signal ainsi obtenu. Observe-t-on bien la fréquence fondamentale et ses harmoniques dans le spectre du signal ? Est on en présence de son articulé ? Pour cela, il nous reste en fait à définir le filtre correspondant aux différents articulateurs.

Articulation : génération du filtre (*Formant Tier*) – Définissez de même un filtre d'une longueur de 2 secondes à l'aide du menu `New > Tiers > Create Formant Tier`. Une fois encore, nous allons définir l'évolution du filtre point par point (`Modifiy > Add Point`). Une fenêtre apparait, qui décrit les 5 premiers formants et leur largeur (*bandwidths*) au point donné.



La formule ci-dessus se lit ainsi: le premier formant est à 500 Hz et a une largeur de 50 Hz, le second est à 1500 Hz est à une largeur de 100 Hz etc…

Modifiez donc votre couche de filtre pour réaliser un son /a/ « parfait », c'est-à-dire à formants constant tout au long de sa production. Il n'est pas nécessaire de définir 5 formats. Nous nous contenterons ici de F1, F2 et F3 en leur donnant les valeurs suivantes :

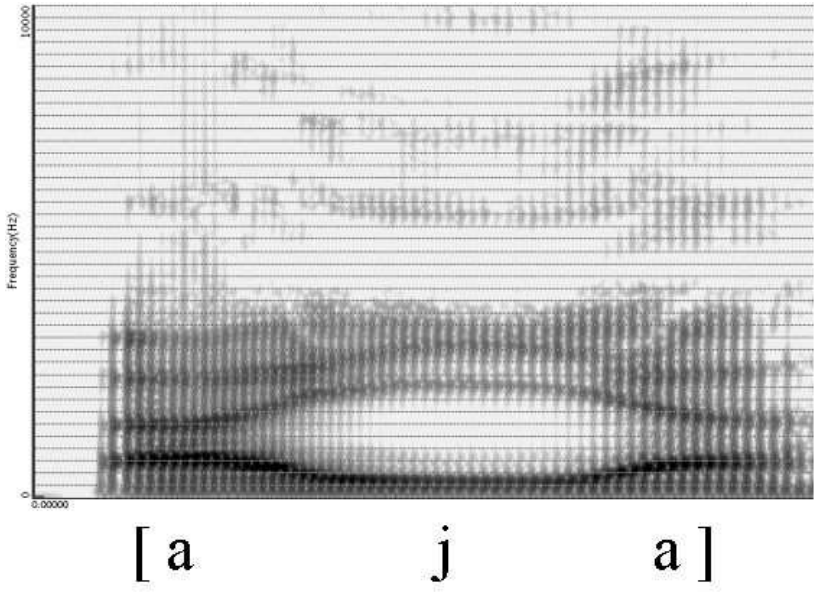
- F1 750 Hz largeur de 50 Hz
- F2 1200 Hz largeur de 100 Hz
- F3 2250 Hz largeur de 100 Hz

Créez deux points au début et à la fin du signal afin de créer au son aux contours formantiques constants.

Le filtre est défini, il ne nous reste plus qu'à le combiner avec la source. Pour cela, sélectionnez en même temps dans la fenêtre d'objet le signal glottal et le filtre et cliquez le bouton `Filter`. Un signal est généré : visualisez-le ainsi que son spectrogramme. Ecoutez maintenant ce signal : le résultat est très artificiel, mais vous devriez normalement reconnaître un /a/.

Ce premier exemple vous montre la difficulté de générer une parole naturelle par des approches à base de connaissances. Il est possible d'essayer d'améliorer le résultat. Par exemple, nous pouvons rajouter une couche intensité (`Intensity Tier`) pour modéliser une évolution plus progressive de cette dernière. Au final, 30 ans de recherche sur le sujet ont montré qu'il n'est pas possible d'atteindre une synthèse totalement naturelle à l'aide de cette approche.

Coarticulation – Contentons nous d'en rester à une qualité correspondant aux recherches du début des années … 1970 et étudions la coarticulation. Pour cela, il vous est demandé de synthétiser un fichier de parole reproduisant la transition /aia/ sur une durée de 3 secondes (1 seconde pour chaque phonème), en ne modélisant cette fois que les deux premiers formants. Parvenez-vous à un résultat intelligible ? Visualisez également l'intensité et le spectrogramme de votre signal. Comparez-le avec le spectrogramme ci-dessous, qui rend de la réalisation de la transition /aia/ (ou plutôt /aja/) par un locuteur réel. Observations ?



Vous pouvez encore modifier les points de votre filtre pour vous rapprocher de ce sinal réel : ajout des formants supérieurs, modélisation plus fine des transitions dues à la coarticulation. Cet exemple vous montre l'intérêt d'une synthèse par diphones, qui prend justement en compte ces transitions.

Pour terminer, nous allons chercher à générer la coarticulation entre une plosive et une voyelle en générant la syllabe /ba/. Pour cela :

- Créez une source avec un pitch 200 Hz sur une durée de 500ms, les 50 premières millisecondes étant non voisées (temps de mise en place du voisement). Pour générer cette zone non voisée, il vous suffit, dans `Modify`, de choisir l'option `Remove point between`. Synthétisez le son glottal correspondant et visualisez-le : le voisement est-il bien absent des 50 premières millisecondes ?
- Faites varier avec le filtre F1 de 100 à 750 Hz (largeur 50 Hz) entre 0 et 50 millisecondes, et F2 de 500 à 1200 Hz (largeur 100 Hz), les formants étant stables après cette zone de coarticulation entre le /b/ et le /a/. Modulez la source avec ce filtre, visualisez et écoutez le signal obtenu.
- Modélisez la phase d'occlusion de la plosive à l'aide d'une couche modélisant l'intensité : `New > Tiers > Create Intensity Tier`. Celle-ci suivra l'évolution suivante : 0.5 dB (autant dire rien…) d'intensité au début du signal, 60 dB entre 40ms et 50 ms (plosion), et enfin 80 dB au bout de 100 ms, cette intensité restant constant jusqu'à la fin de la réalisation de la voyelle, très énergétique. Sélectionnez cette couche avec le son précédent pour les multiplier (`Multiply`) et générer le son final. Ecoutez et visualisez le signal final. Satisfait(e)s ?

Programmation prosodique – Pour terminer cette première approche de la synthèse par source/filtrage, nous allons voir qu'il est également possible de faire varier le pitch au cours du temps pour rendre compte de la programmation prosodique (évolution du timbre) au cours du temps. En jouant sur les différents points de votre couche source, faites prononcer à votre système une voyelle /a/ de 2 secondes de long pour laquelle le pitch évoluera de 100 à 200 Hz, puis redescendra à la valeur initial. Cela y est, votre synthétiseur est capable de chanter !

6. Synthèse articulatoire

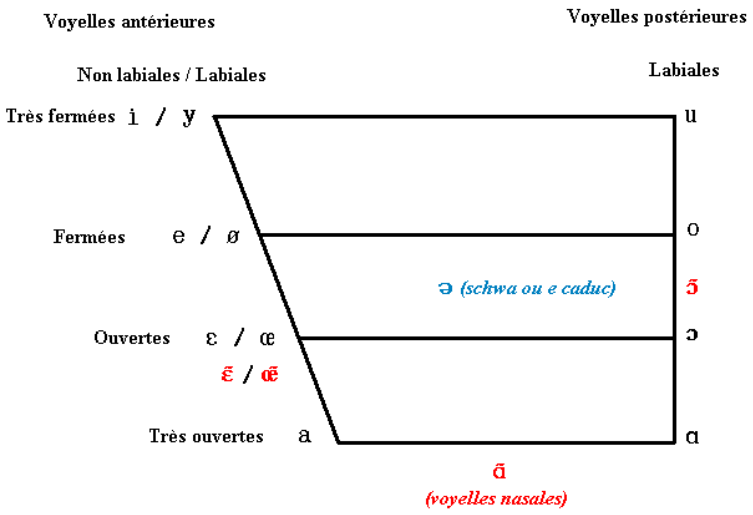
Jusqu'au début des années 1990, la synthèse de la parole a été dominée par les approches à bases de connaissances. A côté de la synthèse par formant, que nous avons étudié, certains chercheurs ont étudié une synthèse articulatoire dont l'intérêt était d'être proche des linguistes (phonéticiens mais aussi psycho-acousticiens). Plutôt que de définir la valeur des formants à un moment donner, l'idée était en effet de définir la position des différents articulateurs à cet instant. Un modèle mathématique complexe (modèle de Maeda, par exemple) était alors en charge de donner la « fonction de transfert » qui permettait de connaître la valeur des formats pour ces positions. L'autre intérêt du modèle est de prendre généralement en compte la coarticulation : si on définit par exemple les positions de la langue à deux instants donnés, le modèle peut inférer les positions intermédiaires de celles-ci. En pratique, cette méthode n'a pas donné les résultats escomptés. Certains chercheurs poursuivent toutefois cette voix de recherche, en cherchant à s'approcher du fonctionnement réel de l'appareil phonatoire à l'aide de cliché radiographiques ou de scanner. Nous allons faire une petite manipulation pour nous rendre compte des capacités de cette approche.

PRAAT intègre un modèle de synthèse articulatoire. Pour générer un son par synthèse articulatoire, il faut :

- définir un utilisateur `New > Articulatory Synthesis > Create Speaker`. Choisissez l'utilisateur de votre choix, que vous modéliserez avec 10 tubes glottaux (plus on découpe l'appareil phonatoire en tube fins, meilleure est la description de celui-ci),
- définir un objet de type *ArtWord* qui va décrire la succession des positions que vont prendre un ensemble d'articulateurs: `New > Articulatory Synthesis > Create ArtWork`

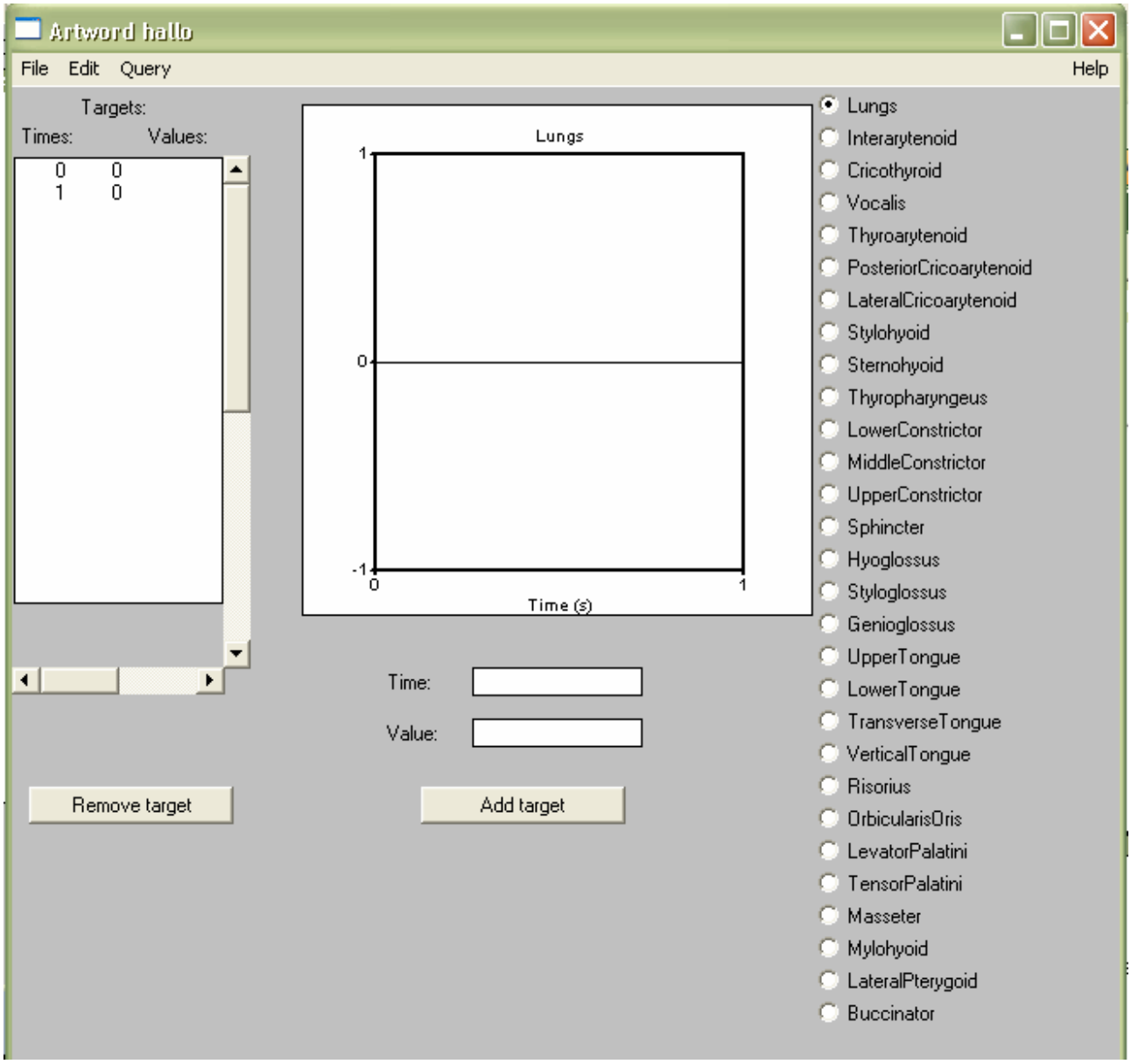
Nous allons essayer de décrire dans cet *ArtWord* la réalisation, pour une durée de 1 seconde du phonème /a/.

Comme le rappelle la figure à votre droite, il s'agit d'une voyelle très ouverte, c'est-à-dire que la conduit vocal est largement ouvert lors de la production de cette voyelle. Elle est donc relativement simple à générer en synthèse articulatoire.



Une fois créé, l'objet *ArtWord* que l'on souhaite définir va être caractérisé à l'aide de la commande *Edit.* Une fenêtre s'ouvre, qui nous permet de préciser un certain nombre de paramètres contrôlant la production de sons articulés :

- le paramètre *Lungs* permet de définir la pression d'air envoyé par les poumons dans l'appareil phonatoire
- tous les autres paramètres figurant sur la figure (Interarytenoid, Cricothyroid, Vocalis…) correspondent à l'activation de tous les muscles qui contrôlent l'appareil phonatoire. Par défaut, ces muscles sont relâchés à la création d'un objet *ArtWord*.



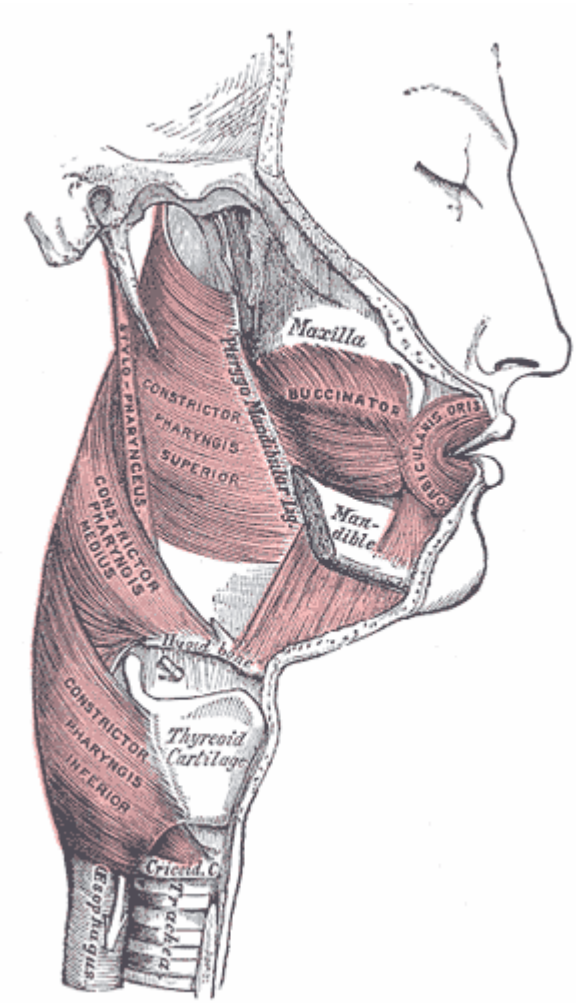
Pour le moment, nous avons donc créé un locuteur et une configuration musculaire totalement relâchée (un peu comme si vous dormiez la bouche ouverte !). Sélectionnez ces deux objets et cliquez sur le bouton *Movie* : vous avez un aperçu en tranche de l'appareil phonatoire. Générez le signal sonore correspondant à cette situation à l'aide de la commande `Synthetize > To Sound`. Ecoutez et visualisez le son obtenu. Ce résultat était-il prévisible ?

Afin de simuler l'expulsion d'air lors de la production, nous allons maintenant modifier le paramètre *Lungs* : à l'aide du bouton *Add target*, ajouter une pression de l'activité Lungs à 1 à l'instant t = 0s puis à l'instant t = 1s. Vous observez que PRAAT interpole l'évolution (ici stable) du paramètre entre ces deux points : ceci est très utile lorsqu'on souhaite décrire le passage progressif d'une position à une autre d'un articulateur donné. Générez une fois de plus ce nouveau son et écoutez-le. Observations ?

En fait, les phonèmes sont produits à l'aide d'une succession d'expulsions successives d'air pulmonaire (un enseignant peut ainsi apprendre à placer sa voix et son souffle pour ne pas trop fatiguer ses cordes vocales) et non par un flux continu. C'est ce que nous allons modéliser en modifiant comme suit le paramètre *Lungs*. Positionnez la pression de l'air à la sortie des poumons à 0.2 en début de phonation (t = 0), puis à 0 à t = 0.2

secondes. Cette pression nulle se maintiendra jusque la fin de la production du signal. Ecoutez et visualisez le signal obtenu correspondant.

Faisons maintenant bouger nos articulateurs au cours de la phonation. Prenez par exemple le paramètres `Masseter`. Celui-ci contrôle le mouvement de la mâchoire, et permet d'ouvrir plus ou moins la bouche au niveau des dents. Faites varier ce paramètre de 0 à l'initial jusque 0.7 à la fin du signal (vous pouvez aussi arriver à 0.7 au bout de 0.5 secondes et ensuite poursuivre en plateau). Observez le mouvement de l'appareil phonatoire à l'aide du bouton `Movie` puis écoutez le son généré. Vers quelle voyelle tend-on à évoluer ?



Cette voyelle est une voyelle fermée antérieure (le lieu de construction est à l'avant de l'appareil phonatoire). Essayer de modéliser au mieux ce sont en jouant sur les paramètres suivants (outre `Masseter`) :

- `Orbicularis Oris` qui est le muscle qui contrôle l'ouverture et l'arrondissement des lèvres (valeur entre 0 et 0.2 généralement),
- `Mylohyoid`, qui contrôle l'avancement et l'abaissement de la langue (valeur entre 0 et 0.5). Le muscle mylo-hyoidien est un muscle du cou (attaché à l'os hyoïde) qui contrôle l'abaissement de la mandibule et par conséquent de la langue.
- `Stylohyoid`, qui contrôle l'allongement du larynx et en conséquence l'arrondissement de la langue.
- Le *Levator Palatini* est le muscle qui contrôle la fermeture de l'accès au conduit nasal par la luette.

A chaque fois, vous pouvez contrôler la position des articulateurs à l'aide de la commande `Movie`. Si vous essayer d'autres articulateurs, vous verrez qu'aucune variation n'est visible. Ceci est du au fait que les mouvements correspondants ne se font pas dans le plan vertical mais horizontal (déplacement latéraux par exemple). Ces paramètres sont pourtant aussi importants en pratique pour réaliser une constriction. Ne vous étonnez donc pas que les sons que vous allez produire resteront imparfaits, un peu comme si vous aviez la langue un peu pâteuse ou endormie.

7. Synthèse de la parole : sélection d'unités et ajustement prosodique

Comme vous avez pu le constater, la synthèse de parole que nous avons obtenue très imparfaite. Par ailleurs, il est très lourd de définir, point par point, l'évolution du spectre d'un signal de parole, surtout si l'on veut décrire proprement la coarticulation. Enfin, une modélisation spectrale par la simple définition de quelques fréquences de résonnance (formants) et de leur largeur est insuffisante. Aussi a-t-on rapidement pensé à aller chercher l'information sur la réalisation spectrale de phonèmes (ou plutôt de diphones : synthèse PSOLA/MBROLA) puis sur des segments de parole enregistrés les plus longs possibles : c'est ce qu'on a appelé la synthèse par sélection d'unités.

Etudier en détail cette technique de synthèse demanderait du temps et des compétences d'ingénieur en traitement du signal. Dans ce TP, nous allons simplement faire une simulation (très approximative) des procédés utilisés dans ce type d'approche :

- Nous allons travailler sur des signaux existants pour avoir une description fréquentielle de segments de parole. Nous supposerons que ces signaux correspondent au résultat d'une sélection d'unité et qu'il nous reste à les mettre bouts à bouts pour synthétiser la phrase recherchée.
- A partir de ces signaux, nous allons extraire une information un peu plus précise que celle que nous définissions à la main sur les formats, leur largeur et surtout leur intensité. Pour cela, nous allons utiliser la technique de prédiction linéaire (LPC = *linear prediction coefficients*). Cette méthode donne une approximation, suivant une fenêtre temporelle donnée, des fréquences de résonnance d'un signal sonore sous la forme d'un petit nombre de pics fréquentiels (fréquence médiane, largeur et intensité). A partir de ces coefficients LPC, PRAAT est alors capable de créer un `Formant Tier` comme nous

l'aurions fait en le définissant point par point…

- L'intérêt de cette représentation formantique est qu'elle décrit la réalisation phonétique du signal indépendamment de sa prosodie (pitch, énergie, durée). Il n'est donc pas gênant que les fichiers sélectionnés aient été prononcés par des personnes différentes. C'est à la fin que nous allons créer de toute pièce un *Pitch Tier* et un *Intensity Tier* pour justement réaliser l'ajustement prosodique nécessaire à la synthèse d'un signal homogène.

On souhaite synthétiser la phrase suivante : « *bonjour, bienvenue en région parisienne* ». On suppose que l'on est arrivé à sélectionné dans notre base de signaux de parole deux unités qui pourraient couvrir totalement la phrase à générer :

- « *bonjour* », prononcé par une voix féminine — fichier `bonjour_16000.wav`
- « *bienvenue en région parisienne* » prnoncé par un homme — fichier `bienvenue_16000.wav`

Concaténation sans ajustement prosodique – Chargez les deux fichiers dans PRAAT, suivant leur ordre d'occurrence dans la phrase et concaténez-les. Pour cela, sélectionnez les deux objets PRAAT et faites combine `sounds > concatenate`. Visualisez le signal : quelle est sa durée totale ? Ecoutez le résultat : celui-ci est-il intelligible ? La prosodie est-elle acceptable ?

Concaténation avec ajustement prosodique (analyse LPC et pitch) – Afin d'améliorer la prosodie du signal de parole, on veut séparer la réalisation phonétique (formantique) des unités de leur prosodie. Pour cela, nous allons extraire l'évolution des formants par analyse LPC :

- Faites l'analyse LPC du signal de parole concaténé en choisissant dans la fenêtre d'objets PRAAT l'option `Formants&LPC > To LPC(autocorrelation)n`). Choisir un ordre de prédiction LPC égal à 10, ce qui signifie que l'on va garder 10 coefficients LPC, soit 5 formants.
- Générez l'évolution de ces formants à l'aide de l'option `To Formants` appliquée sur le *LPC Tier* obtenu.
- Générez enfin le filtre (*Formant Tier*) correspondant à ces formants.

Il nous reste maintenant à générer la source correspondant à ce signal. Créez donc une source correspondant à une voix d'homme (pitch de 150 Hz en moyenne) pour toute la longueur du signal. Vous pouvez garder le pitch constant sur toute la durée du signal, ou de préférence le moduler pour le rendre plus naturel. Par exemple :

- légère montée (175 Hz) sur la seconde syllabe de *bonjour* pour marquer l'insistance.
- légère baisse (125 Hz ou 100 Hz) continue pour marquer la fin de phrase.

Modulez la source par le filtre. Ecoutez le résultat. Distingue-t-on encore la concaténation entre la voix féminine et la voix masculine ? Visualisez le spectrogramme. Quels sont les phonèmes qui ont été le plus mal synthétisés ?

Concaténation avec ajustement prosodique (intensité) – La naturalité et l'intelligibilité du signal obtenu restent encore très perfectible (en réalité, la synthèse par concaténation utilise des outils moins violents qu'une re-synthèse de formats par LPC). Vous pouvez toutefois améliorer légèrement le résultat en ajoutant un paramètre prosodique lors de la synthèse : l'évolution de l'intensité du signal de parole.

Plutôt que de créer à la main un *Intensity Pitch*, nous allons l'extraire directement du signal concaténé. En effet, l'énergie ne dépend pas (ou peu) du locuteur. Pour cela

- extraire l'intensité à l'aide de l'option `To Intensity` de la fenêtre d'objets PRAAT.
- générer une Intensity Tier à partir de l'objet créé : DownToIntensityTier.
- il ne vous reste plus qu'à multiplier cette couche avec le signal synthétisé précédent : `Multiply` après avoir sélectionné ces deux objets.

Une fois encore, écoutez et visualisez le résultat. Celui-ci n'est pas parfait ? Pour arriver à cette perfection, il ne vous reste plus qu'à entamer un doctorat sur le sujet pour savoir piloter finement PRAAT (ou plutôt un toolkit de synthèse par concaténation) … ou plus simplement récupérer une synthèse de parole commerciale !