Université François Rabelais de Tours

UFR Lettres et Langue

Master SLD – M1 L & R

Outils linguistiques

TRAVAUX PRATIQUES

Enseignant **Jean-Yves ANTOINE**

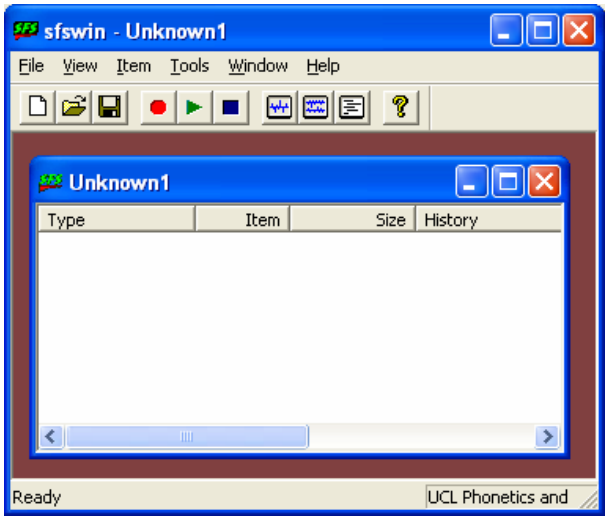
(Jean-Yves.Antoine AT univ-tours.fr)

Edition de signal de parole: SFSWin

Présentation

SFSWin (*Speech Filling System for Windows*) est un éditeur de signal développé par Mark Huckvale (University College, London) qui est dédié au traitement du signal de parole. Ce gratuiciel *(freeware*) permet en effet d'effectuer la plupart des traitements envisageables sur ce type très particulier de signal sonore. Dans le cadre de cet enseignement, nous étudierons les fonctionnalités les plus communes de SFSWin : calcul et affichage de spectrogramme, détection de fréquence fondamentale (pitch), suivi de formats ainsi que quelques opérations de filtrage. SFSWin peut être récupéré à l'URL suivante : http://www.phon.ucl.ac.uk/resource/sfs.

1 Prise en main du logiciel : quelques rappels de traitement du signal



Dans un premier temps, nous allons nous travailler sur des signaux synthétiques, n'ayant rien à voir avec la parole, pour découvrir les fonctionnalités de base de SFSWin et faire quelques rappels sur les propriétés spectrales des signaux.

Lancer SFSW. Une fenêtre vide (`Unknown1`) s'affiche à l'écran : c'est dans cette fenêtre de travail que s'affichent, au cours de toute session d'utilisation, l'ensemble des signaux qui peuvent être utilisés, soit en les créant ex nihilo (enregistrement ou synthèse de signal) soit en chargeant un fichier déjà enregistré.

Dans cette première partie, nous allons créer un signal artificiel pour faire des rappels de traitement du signal.

1.1 Création d'un signal artificiel

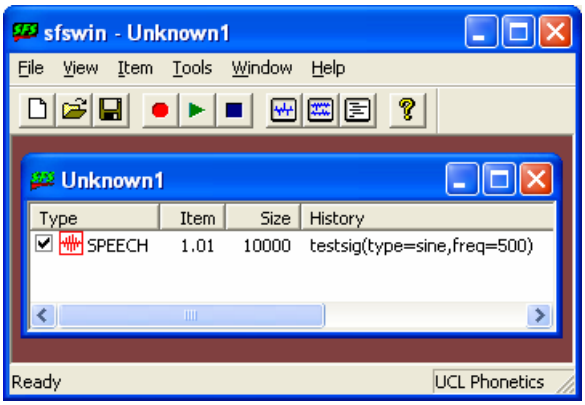
On désire créer un simple signal sinusoïdal de fréquence 500 Hz. Pour cela, faites les opérations suivantes :

- Sélectionner le menu l'option `Tools/Generate/Test signals`.
- Sélectionner les bonnes options dans la boîte de menu qui apparaît. Dans notre cas : signal sinusoïdal (`sine`) et fréquence de 500Hz.
- Validez vos choix (OK) : un item apparaît dans la fenêtre de travail, qui correspond au signal généré :

```
SPEECH 1.01 10000 testsig(type=sine,freq=500)
```

Cet élément est étiqueté comme étant de type `speech`. C'est par ce type que SfSWin réunit tous les signaux audio, qu'ils soient de parole ou non. SfSWin précise qu'il s'agit d'un fichier de taille 10 000. Il s'agit en fait du nombre d'échantillons de période générés (par défaut) par le logiciel. Nous reviendrons sur ce point.

1.2 Ecoute d'un signal



Nous allons maintenant écouter le signal que nous avons généré (attention aux oreilles pour ce signal mono-fréquentiel : n'utilisez pas votre casque si vous n'êtes pas sûr(e)s du volume sonore !). Pour ce faire, la procédure est la même quelle que soit le type de signal audio :

- Cocher le fichier de signal concerné (petite case de sélection dans le fenêtre de travail)
- Cliquer sur l'icône de lecture (triangle) dans la barre d'outil

Vous pouvez rejouer ce signal autant de fois que nécessaire en recliquant sur cet icône.

Question 1 — Générez un second signal sinusoïdal de fréquence 1000 Hz et écoutez-le (attention de bien sélectionner ce nouveau signal avant écoute). Vous vérifiez bien que la tonalité du signal est plus aiguë.

Générez maintenant un signal de fréquence 10 000 Hz.

Question 2 — Qu'observez-vous cette fois ? Nous allons expliquer plus loin cette observation.

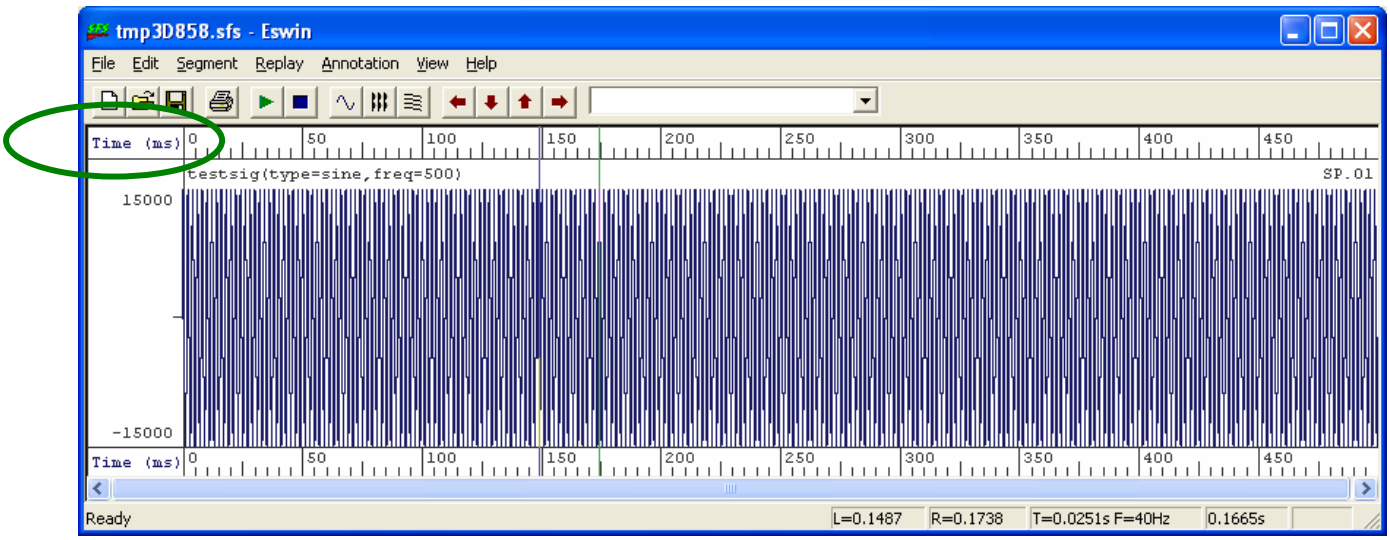
1.3 Edition d'un signal

Nous allons nous intéresser pour l'instant au premier signal sinusoïdal que nous avions généré (fréquence 500 Hz). Pour visualiser un signal (ou plusieurs), la procédure est toujours la même :

- Sélectionner le signal (ou les signaux) désirés dans la fenêtre de travail. Ce signal ne peut correspondre qu'à un item de type `Speech`.
- Choisir ensuite l'option d'édition dans le menu `Tools/Speech/Edit` ou cliquer directement sur l'icône "`Display checked items`" de la barre d'outils.

Le signal s'affiche, dans son intégralité, dans une nouvelle fenêtre.

Question 3 — En considérant l'échelle temporelle en haut d'affichage, donnez la durée du signal généré.



Question 4 — Nous avions choisi en question 1.1 de générer un fichier de parole comportant au total 10 000 échantillons. La durée du signal observé correspond donc à ces 10 000 échantillons. Pouvez-vous alors en déduire la période, puis la fréquence d'échantillonnage utilisée par SfSWin sur ce signal ?

Question 5 — Cette fréquence d'échantillonnage est la même pour tous les signaux générés par SfSWin. Pouvez-vous, en considérant le théorème de Shannon, expliquer les observations de la question 3 ?

Le signal qui a été affiché ne ressemble guère à une sinusoïde : il est en fait trop compressé sur une fenêtre d'observation temporelle aussi large. Pour le visualiser correctement, il va nous falloir le **zoomer** sur une longueur de signal plus courte. L'agrandissement d'un affichage sur une fenêtre temporelle précise se fait toujours de la même manière :

a) sélection d'une zone temporelle de signal

- Définissez le début de la zone en cliquant avec le bouton gauche de votre souris à l'endroit choisi : une barre verticale bleue apparaît sur le signal,
- Définissez la fin de la zone en cliquant avec le bouton droit de votre souris à l'endroit correspondant : une barre verticale verte apparaît sur le signal.

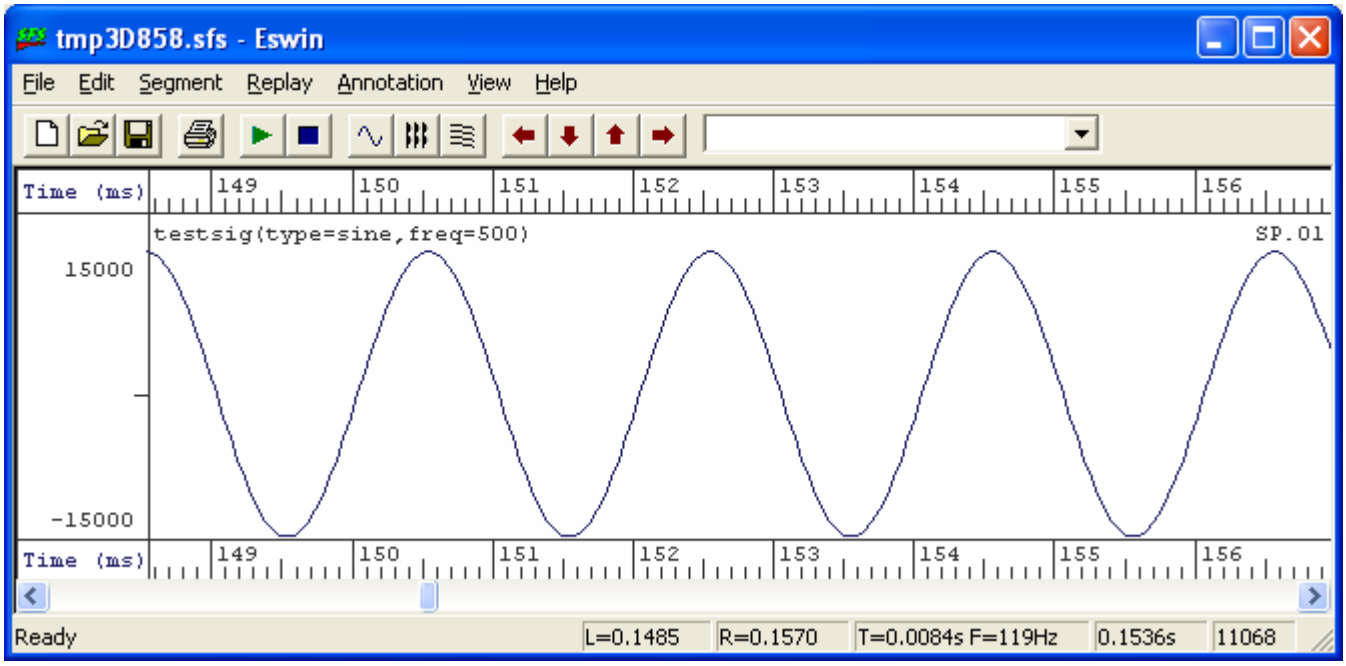
  Dans notre cas, faites en sorte que la zone ainsi sélectionnée corresponde à une fenêtre temporelle de 10 à 20 ms environ.

La zone temporelle ainsi définie servira désormais de zone de travail pour tout traitement ultérieur. Ecoutez ainsi la zone de signal ainsi

Question 6 — Essayez ainsi d'écouter la zone de signal sélectionnée. Le temps d'écoute est-il plus court ? Quelle autre remarque pouvez-vous faire à cette écoute ? Ce résultat est là encore une conséquence du théorème de Shannon…

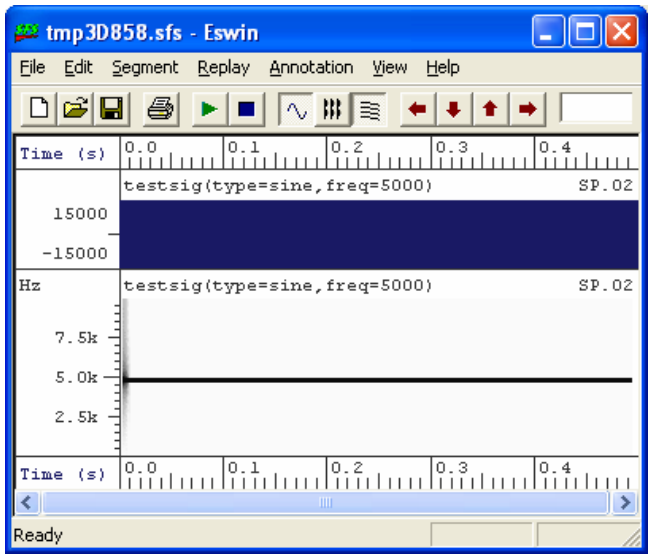
b) zoom sur la zone de signal sélectionnée

Revenons à notre objectif, qui était de zoomer le signal sur la zone considérée. Pour cela, il suffit de sélectionner dans la fenêtre d'affichage l'option `View/Zoom In` de la barre de menu. La visualisation est alors limitée à la zone temporelle définie : on observe bien une sinusoïde.



Question 7 — Quelle est la période (durée d'un cycle sinusoïdal) de ce signal ? Ce résultat était-il prévisible à partir des caractéristiques du signal généré ?

1.4 Spectrogramme d'un signal



Lorsqu'on travaille sur un signal de parole, la visualisation du signal en lui-même n'apporte guère de renseignements utiles. C'est l'analyse de sa distribution fréquentiel qui permet de caractériser ses propriétés phonétiques ou acoustiques. Cette analyse en fréquence correspond à la visualisation du **spectrogramme** du signal considéré. Ce spectrogramme peut-être calculé pour tout type de signal : nous allons donc découvrir cette fonctionnalité sur un signal artificiel, on considérant cette fois le signal sinusoïdal de fréquence 5000 Hz que nous avions généré précédemment.

Nous avons vu en cours qu'il était possible de calculer des spectrogrammes à bande étroite ou large.

Calculez et affichez tout d'abord le **spectrogramme à bande étroite** du signal en sélectionnant dans le menu l'option `Tools / Speech / Displays /Narrow Spectrogram`. Une fenêtre s'ouvre, qui affiche à la fois le signal audio (très compressé, en partie supérieure) et le spectrogramme en partie inférieure, comme le montre l'image à droite.

Ce spectrogramme ne ressemble en rien à ceux des signaux de parole étudiés en cours

Question 8 — Qu'observe-t-on ici? Ce résultat est-il prévisible au vu des caractéristiques du signal généré?

On remarquera également une zone particulière en tout début de signal : celle-ci ne correspond en rien aux caractéristiques du signal, mais à un effet de bord du à la méthode de traitement de signal employée pour calculer le spectrogramme. Nous ne chercherons pas à expliquer ce phénomène (pour les initiés : influence de la fenêtre de Hamming utilisée pour le calcul…). Retenons simplement que le tout début d'un spectrogramme est toujours perturbé par ce type de phénomène : c'est pourquoi il vaut toujours mieux garder une petite zone temporelle qui ne nous intéresse pas (silence, par exemple) en début d'observation.

Calculez et affichez maintenant le **spectrogramme à bande large** du signal. Pour cela, vous pouvez revenir à la fenêtre de travail (option `Tools / Speech / Displays / Wide Spectrogram`) ou sélectionner directement l'option `View / WideBand Spectrogram` dans la fenêtre d'affichage : le spectrogramme à large bande se rajoute au dessus de celui à bande étroite.

Question 9 — Comparez le spectrogramme obtenu avec le précédent. Ce résultat correspond-il aux propriétés attendues pour ce type de spectrogramme.

1.5 Calcul et suivi de fréquence fondamentale

Enfin, un signal de parole se caractérise également, pour les sons voisés, par sa fréquence fondamentale. Si cette notion, qui correspond à la fréquence de vibration des cordes vocales, n'a de sens qu'en parole, elle peut être calculée sur tout type de signal. Dans ce cas, l'analyse en fréquence fondamentale cherchera à suivre au cours du temps la fréquence qui porte le plus d'énergie dans le signal. Essayons donc de retrouver cette fréquence sur le signal sinusoïdal à 500 Hz que nous avions généré.

Pour lancer le suivi de fréquence fondamentale, on procède comme suit :

- sélectionner le signal sur lequel on souhaite travailler,
- lancer le calcul en sélectionnant l'option : `Tools / Speech / Analysis / Fundamental Frequency / Fundamental Frequency Track`

Ce calcul entraîne la création d'un nouveau fichier dans lequel se trouve la variation de la fréquence fondamentale estimée au cours du temps : un nouvel item de type FX s'affiche ainsi dans la zone de travail:

FX 4.01 46 fxrapt(1.01) ⇨ montre que le calcul a été fait sur le signal 1.01

Pour visualiser ce fichier, par exemple en parallèle avec le signal de parole, il suffit de sélectionner les deux items concernés et de les afficher à l'aide de l'icône "Display checked items".

Question 10 — Une fois de plus, cet affichage ne correspond guère à celui observable sur un signal de parole. Pourquoi ce résultat est-il néanmoins conforme à nos attentes ?

1.6 Sauvegarder une session de travail

Si vous désirez sauvegarder l'ensemble des fichiers (signal, fréquence fondamentale…) générés au cours de cette session de travail, vous pouvez utiliser l'option `Save As` du menu `File`.

2 Etude de signaux de parole

Maintenant que nous connaissons les principales fonctionnalités du logiciel *SfSWin*, nous allons pouvoir l'utiliser comme éditeur et signal de parole. Nous allons tout d'abord travailler sur un signal de parole très court, correspondant à la prononciation du triphone /aka/.

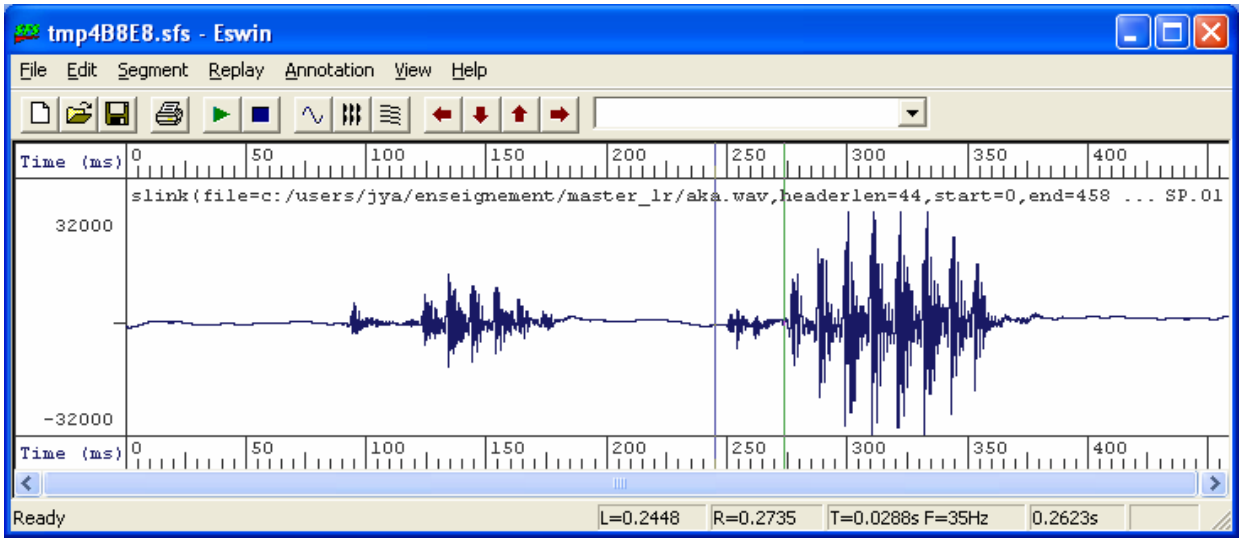
2.1 Chargement d'un signal de parole pré-enregistré

Téléchargez à partir de ma page WWW (www.info.blois.univ-tours.fr/~antoine) le fichier `aka.wav`, qui correspond à ce signal de parole. Sauvegardez ce fichier sur votre compte personnel.

A l'aide du menu `File/Open`, chargez ce fichier. Une fenêtre de dialogue apparaît qui vous laisse plusieurs choix. Contentez vous de créer un item de type Speech, et de n'utiliser qu'un lien vers ce fichier.

2.2 Etude de la coarticulation

Ecoutez tout d'abord le signal, puis affichez le signal de parole. On distingue facilement les deux voyelles, fortement énergétiques, et la plosive qui les séparent : la phase de tenue (occlusion) se caractérise bien par la quasi absence de signal de parole. Ecoutez le signal de parole correspond à la seule de tenue du signal : vous vérifiez bien l'absence de son durant cette période temporelle.



Question 11 — Quelle est justement la longueur temporelle de cette période de silence ? La perçoit-on lors de l'écoute complète du signal ?

Nous percevons ici une des caractéristiques perceptives de l'audition : seules les transitions du signal de parole (co-articulations entre phonèmes) sont perçues par le système cognitif, d'où l'absence de détection de cette zone de silence pourtant significative. Etudions plus en avant cette capacité de masquage.

Question 12 — Ecoutez maintenant la zone correspondant uniquement à la phase d'explosion de l'occlusive /k/. Sans visualiser le spectogramme du signal, il est difficile à une personne non habituée de distinguer l'explosion de la réalisation de la voyelle suivante : la figure ci-dessous vous donne une indication de la zone concernée. Quelle est la durée temporelle approximative de cette zone de signal? Reconnaît-on le phonème /k/ à l'écoute de la zone d'explosion ?

Question 13 — C'est la coarticulation entre plosive et voyelle qui va donner du sens aux deux phonèmes: écoutez une zone du signal regroupant la phase d'explosion et la moitié de la voyelle : qu'entendez-vous ?

Question 14 — Réduisez la longueur d'écoute jusqu'à la limite de perception claire du diphone /&͡ɔ/. Quelle est la longueur temporelle du segment minimal ainsi obtenu ? Déduisez-en la longueur de la phase d'explosion de l'occlusive. Conclusion ?

2.3 Etude du spectrogramme : détection de formants

On désire maintenant étudiez les caractéristiques spectrales de ce signal de parole, et plus particulièrement s'intéresser aux formants des réalisations successives du phonème /a/. Pour visualiser ces formants, devez-vous afficher un spectrogramme à large bande ou à bande étroite ?

Question 15 — Comparez les structures formantiques des deux réalisation du /a/. Quelles sont les points communs et les variations entre ces deux réalisations ? A votre avis, à quoi sont dues ces variations, qui font qu'un phonème n'est jamais réalisé de la même manière dans deux contextes différents ?

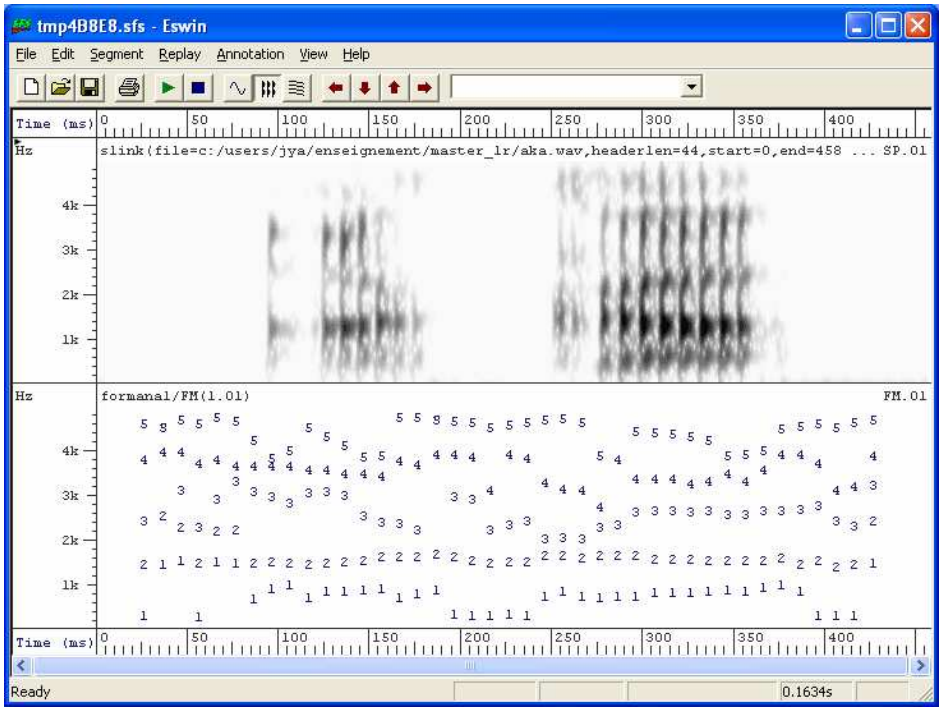
Question 16 — Quelles sont les valeurs moyennes (à 200 Hz près) des deux premiers formants observés sur le spectrogramme ? Ces valeurs sont-elles compatibles avec les valeurs moyennes citées en cours ?

SfSWin dispose de traitements de détection automatique de formants. Pour lancer le calcul d'un suivi de formant sur notre signal de parole, il faut sélectionner l'option *Tools/Speech/Analysis/Formants Estimation Tracks* dans le menu.

Une boîte de dialogue apparaît. Ne sélectionnez pas l'option "*Output synthetiser control data*", qui n'est utile pour la synthèse de parole (analyse / synthèse). Après validation ("OK"), *SfSWin* crée un item qui correspond aux résultats du calcul de suivi de formants. Celui-ci apparaît dans la fenêtre de travail avec le type `FM` :

```
Formants    12.01 41    formanal/FM(1.01)
```

Affichez à la fois le signal de parole et les formants détectés : les cinq premiers formants sont indiqués par un numéro d'ordre placé à la hauteur de leur fréquence estimée :



Question 17 — Pour le phonème /a/, les valeurs de formants estimées par le programme correspondent-elle à celles que vous aviez observé sur le spectrogramme ?

Question 18 — Les techniques de détection automatique de formant ne sont pas à l'abri d'erreurs d'estimation (ce qui, entre autres, explique par exemple la difficulté de la reconnaissance automatique de parole). Pouvez-vous me citer une erreur flagrante visible dans ce cas précis ?

2.4 Etude de la fréquence fondamentale : suivi de F0

On désire maintenant s'intéresser à la mélodie du signal de parole, et en particulier à l'évolution de sa fréquence fondamentale. Quel type de spectrogramme est-il préférable d'utiliser pour visualiser ce pitch ?

Question 19 — Le signal est-il voisé sur toute sa longueur ? La fréquence fondamentale est-elle constante sur l'ensemble du signal ? Pouvez-vous faire une estimation grossière de la valeur (à 50 Hz près) de la valeur de ce pitch.

Il est assez difficile d'étudier finement l'évolution de la fréquence fondamentale sur un spectrogramme aussi large (0 à 5000 Hz dans notre cas). Nous allons donc nous en remettre à une détection automatique, comme étudié au paragraphe 1.5.

Question 20 — Quelle est, cette fois de manière précise, la valeur moyenne de la fréquence fondamentale du signal ? L'outil de détection a-t-il été pris en erreur cette fois-ci ?

On observe que l'évolution du pitch est ici assez limitée. Cette situation est normale : le signal de parole prononcé étant sans signification, il y avait peu de chance d'y retrouver une quelconque marque d'insistance ou d'évolution prosodique. Ces variations vont au contraire être sensibles sur de la parole naturelle.

2.5 Parole naturelle et parole synthétique

Le premier fichier de parole naturelle auquel nous allons nous intéresser correspond à une prononciation en italien. Téléchargez à partir du forum ou de ma page WWW le fichier `audio_saluto.wav`. L'énoncé prononcé est le suivant :

Un saluto a tutti participante a la conference Eurospeech 99 a Budapest.

Question 21 — Affichez le spectrogramme et le signal de parole correspondant à la prononciation du début de l'énoncé *un saluto*. Comment se caractérise la fricative en début du nom *saluto*. A partir de quelles fréquences apparaît le bruit de friction ?

Effectuez maintenant une détection de fréquence fondamentale sur l'ensemble du signal de parole.

Question 22 — Entre quelles valeurs extrêmes se situe la fréquence fondamentale sur l'ensemble du signal? Cette observation vous montre quelles sont les capacités de programmation prosodique de notre appareil articulatoire au cours d'une prononciation relativement standard…

Question 23 — Comment est marqué l'accent prosodique sur le début du mot *saluto* ?

Question 24 — Au vu de cet exemple, comment est marqué prosodiquement en italien, la fin des énoncés déclaratifs ? Le français utilise la même programmation prosodique pour marquer ses fin de phrases.

Question 25 — A l'opposé, considérez l'énoncé anglophone du fichier `audio_pleased.wav`. Comment est marquée ici la fin de phrase affirmative ?

2.6 Pour aller plus loin : parole chantée

Les plus grands chanteurs et cantatrices ont des capacités vocales qui font que les caractéristiques spectrales de leur voix s'approche par bien des aspects plus d'un instrument de musique que de la parole ordinaire. Afin d'étudier ces propriétés vocales tout à fait particulières, nous allons nous intéresser à un extrait de mottet composé par Antonio Vivaldi. Téléchargez le fichier correspondant à cet extrait : `audio_vivaldi.sfs` (l'extension *sfs* marque un format de fichier sonore propre à *SfS*).

Question 26 — Visualisez le spectrogramme de ce signal de parole chantée (bande large et étroite) ainsi que le évolution de la fréquence fondamentale estimée. Quelles différences significatives observez-vous par rapport à de la parole naturelle ? Seul un très bon chanteur est capable d'utiliser ainsi sa voix : une parole chantée ordinaire se rapproche au contraire avant tout de la parole standard.

2.7 Pour aller plus loin : parole synthétique

Nous allons terminer cette étude par une comparaison entre la parole naturelle et de la parole artificielle obtenue par des technique de synthèse de parole (TTS : *Text To Speech Synthesis*). Plusieurs fichiers de parole synthétique sont à votre disposition sur le forum ou ma page WWW :

`TTS Limsi 1968.sfs`	Cet enregistrement est historique : il s'agit du tout premier signal de parole synthétique obtenu pour le français. Il a été réalisé en 1968 au laboratoire LIMSI à Orsay.
`TTS CNET 1976.sfs`	Cet enregistrement donne une idée des performances systèmes de synthèse au milieu des années 1970. A cette époque, la synthèse de parole nécessitait la réalisation de cartes de traitement de signal spécifiques, qui palliaient le manque de puissance des ordinateurs de l'époque. Il a été réalisé au CNET à Lannion, en 1976.
`TTS CNET 1989.sfs`	Réalisé également au CNET en 1989, cet enregistrement donne la mesure des progrès ultimes des techniques de synthèse de parole par concaténation d'unités. A cette époque, la synthèse logicielle (i.e. sans carte spécifique) commençait à se développer.
`FT RD 2002.wav`	Cet enregistrement illustre les dernières avancées obtenues en synthèse de parole. Ce fichier sonore est obtenu en recherchant les sous parties de l'énoncé les plus longues possibles qui ont déjà été prononcées et enregistrées dans une immense banque de sons (synthèse par sélection d'unités). Tout l'art de la synthèse consiste ensuite à concaténer de manière harmonieuse les différentes parties. Ces fichiers ont été réalisé dans le département de synthèse de parole de France Telecom R&D (ex-CNET, Lannion), sous la direction de Thierry Moudenc.

Ecoutez l'ensemble de ces fichiers afin de prendre la mesure des progrès réalisés en une trentaine d'année.

Question 29 — Quels reproches pourriez-vous adresser à chacun de ces enregistrements, en comparaison avec de la parole naturelle ?

Question 30 — Affichez maintenant ces signaux, leurs spectrogrammes à large bande et un suivi de leur fréquence fondamentale. Détectez-vous des caractéristiques qui permettraient d'expliquer cette différence perceptive avec des signaux naturels ? Ce résultat vous donne une idée de la difficulté à atteindre une synthèse de parole naturelle…